Chapter I

Preliminaries; Elementary Set Theory; Functions

Quotes for Chapter (I):

> (1) 'I shall not today attempt further to define [it], and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, ...'
>
> (Potter Stewart, Associate Justice of the U.S. Supreme Court, in his written opinion on the 1964 case of 'Jacobellis *vs* Ohio'.)
>
> Remark: The reader may wish to look up this case to find out what the 'it' was that Justice Stewart had such difficulty defining. '
>
> (2) 'I hate definitions.'
>
> (Spoken by Mrs. Felix Lorraine, a character in the novel 'Vivian Grey' by Benjamin Disraeli, later to become Prime Minister of the United Kingdom.)
>
> (3) '... first in war, first in peace, and first in the hearts of his countrymen.'
>
> (From the eulogy, by 'Light-Horse' Harry Lee, on the death of President George Washington.)
>
> (4) 'When we've been there ten thousand years, bright shining as the sun, we've no less days to sing God's praise then when we first begun'.
>
> (Anonymous text often included as an additional verse of the well-known hymn *Amazing Grace* by John Newton; it appears in the novel 'Uncle Tom's Cabin' by Harriet Beecher Stow.)

I.1 Preliminary Remarks

Basic Structure of This Work

(1) In what follows, we refer to this work as '***This Textbook***'; note the italics.

(2) The '**main body**' of *This Textbook*, consists of a number of chapters, enumerated by Roman numerals (I, II, III, etc.) In addition, there are several appendices, enumerated by capital letters (A, B, etc). Each chapter and appendix ends with exercises.

(3) The appendices expand on topics which appear in the main body of *This Textbook*. Many of these topics are of the type that every mathematician should know. However,

one can read the main body, and carry out the corresponding exercises, independently of the appendices.

(4) Inserted into the exposition are certain 'Side Comments', outside the strictly *logical* flow of the discussion, which attempt to give insight into the topic at hand. Many of these 'Side Comments' are historical in nature; some allow the author to express personal mathematical opinions which can be safely ignored by the reader; and some provide connections with pedagogy. (Those who study the type of analysis presented here often become teachers of calculus themselves.)

Remark The 'Side Comments' are indented and printed in a smaller font to make it easier for the reader to skip over them if so desired; in a sense, they play the role of footnotes. In particular, they are not needed for a purely logical treatment of the subject, and they are not referred to in the main body, the appendices to the main body, or the exercises of *This Textbook*. Nevertheless, it is recommended that they be read, at least when first learning the material.

Prerequisites for Reading *This Textbook*

(1) Anyone reading *This Textbook* should certainly know, from elementary-school arithmetic, the basic properties of the natural numbers $1, 2, 3, \ldots$, the integers $0, \pm 1, \pm 2, \pm 3, \ldots$, and the rational numbers (quotients of integers with nonzero denominators). For sake of completeness some of these properties are further discussed in appendices.

(2) In addition, the reader should also be familiar with the *real* numbers, again from elementary-school arithmetic. In contrast to the situation which holds for the integers and the rational numbers, however, to understand modern analysis it is necessary to delve much more deeply into the foundations of the real-number system than was possible in elementary arithmetic. Indeed, the body of Chapter (II) is devoted to a careful axiomatic description of the real-number system. As is usually the case with such axiomatic approaches, this description does not actually say what real numbers 'are', in any philosophical sense, only how to use them properly.

(3) It is assumed that the readers of *This Textbook* have already taken a standard course in elementary (single-variable) calculus; in particular, that they are familiar with the few simple proofs which normally appear in such a course, such as the derivations of the product and quotient rules for differentiation, and the proof via Rolle's Theorem of the standard formulation of the Mean-Value Theorem. In order to avoid simply repeating these proofs exactly as was already presented in elementary calculus, our Chapter (V) ('Differential Calculus') takes the opportunity to approach some of these topics in a way which provides different insights. For example, in Chapter (V) we of course prove the fact that that a function with positive derivative on an interval I is strictly increasing on I, but without using, or even mentioning, the Mean-Value Theorem; the resulting proof seems more direct than the standard one in calculus. Where appropriate, the standard approaches are also reviewed, but usually in the exercises.

(4) One who had not first taken elementary calculus could, in principle, understand the content of *This Textbook*, since all the definitions and theorems stated in the main body are carefully presented using ideas appearimg earlier here, and the corresponding proofs are

rigorous. However, such a reader would lack the motivating examples and applications which play such an important role in elementary calculus, but which traditionally are omitted from texts at our level. Indeed, a major goal of real analysis is to build a logical foundation under the already-known results of elementary calculus, so that errors which crept into calculus by the beginning of the nineteenth century can be avoided. On occasion it may be useful to refer to some topics from calculus, usually within clarifying examples, before they are 'officially' treated in the body of the text. A reader lacking the appropriate calculus background at that stage may simply ignore such examples.

Side Comment (on definitions, theoreme and proofs)

As is mentioned above, the exposition of analysis found in *This Textbook* attempts to be rigorous. This means that definitions should be carefully formulated, theorems should be stated fully, and proofs should be logically complete. Since *This Textbook* may be the first extensive encounter with rigorous mathematics for some readers, it is useful to make a few comments.

(1) The role of a 'definition' is to express the meaning of a concept in terms of concepts which are already understood. Normally these older concepts are themselves defined in terms of concepts which are older yet. It is clear, however, that at some point one must stop looking back, and simply accept some 'primitive' concepts as being 'already known', without providing definitions of them. Such a primitive concept is often handled along the lines of Justice Potter Stewart's famous confession; see Chapter Quote (1) at the start of this chapter.

Warning: Different authors may describe the same concept using different definitions.

(a) Example Consider the following statements:

'An even number is an integer of the form $2\,k$, where k is some integer.'

'An odd number is an integer which is not an even number.'

'An odd number is an integer of the form $2\,k-1$, where k is some integer.'

'An even number is an integer which is not an odd number.'

Each of these statements happens to be true; but *why* are they true?

Nearly everyone would say that the first statement is true 'by definition', since most authors *define* the meaning of 'even number' with this statement. From that point of view, one cannot *prove* the first statement, since it simply provides the meaning of the phrase 'even number'. Likewise, some authors use the second statement as the definition of 'odd number', so for such an author it is also true 'by definition'; in this case, the third statement would then be a theorem which needs to be proved. In contrast, some authors use the third statement as the definition of 'odd number', and for them *it* becomes true 'by definition', while the second statement becomes a theorem to be proved. Finally, an author who uses the third statement as the definition of 'odd number' might well use the fourth statement as the definition of 'even number', in which case the first statement would need to be proved as a theorem.

In particular, to give a logical treatment of a topic it is important to know which statements are taken as the definitions, and which are not: the former are automatically true ('by definition'!), while the latter require proof. Unfortunately, some authors are rather vague about which statements are their definitions, and which are their theorems. In *This Textbook* the convention is that when words or phrases are being defined they appear in boldface, and, most of the time, in a paragraph explicitly labeled as a definition. However, on occasion it is pedagogically useful to start with a 'primitive' or 'preliminary' definition of a concept, and then give an 'improved' definition later as circumstances change. In *This Textbook* auch situations are noted explicitly and explained.

The question of which formulation of a concept is chosen as the definition is largely a matter of the judgement and taste of the author. Thus, while one cannot argue that an author's definition is 'false', one *can* claim that the author used bad taste in choosing to use it and not another, or that the author's definition is 'nonstandard', or that the author's definition is not equivalent to one's own favorite.

Sometimes different authors use the same words for slightly different ideas. For example, some authors define the the smallest natural number to be 1, while others define it to be 0; see the discussion in Example (I.2.2) below.

(b) A related issue arises especially in the *construction* of important mathematical objects. Namely, authors may give the same name to – and treat the same way – objects which are of completely different types. For example, in elementary arithmetic the standard way to construct the integer -2 is to simply prefix the natural number 2 with the minus sign. The modern approach, in contrast, defines -2 to be a certain infinite collection of pairs of natural numbers; see Appendix B.

(2) A 'Theorem' is a mathematical statement which asserts that, under appropriate circumstances (namely, when certain 'hypotheses' are satisfied) then something else must also be true (namely the 'conclusions' of the theorem). Sometimes words such as 'Proposition', 'Lemma' or 'Corollary' are used in place of 'Theorem' to break the monotony, or to show the logical dependence of one statement on the other; but they are all theorems. When stating theorems, all hypotheses and conclusions should be formulated very precisely. Likewise, when applying a theorem to a later situation, it is equally important to check carefully that its hypotheses are satisfied and that its conclusions are being used correctly.

Another feature of theorems in mathematics is the frequent use of *names* for theorems, such as 'The Pythagorean Theorem' (in geometry) or 'The Mean-Value Theorem' (in elementary calculus) . In *This Textbook* we attempt to name as many of the important theorems as possible, especially those which have names that are in common use. The main reason is for ease of reference: it is more meaningful to say, for example, that a given proof uses 'the Mean-Value Theorem' than it is to say that the proof uses 'Theorem (V.6.14)'. Of course this use of names, instead of theorem numbers, becomes even more important in mathematical discussions outside the context of a given source. Unfortunately, as with definitions, mathematicians sometimes apply the same name to different theorems, so some caution is needed.

(3) Many texts include a brief discussion on 'Logic' or 'Proofs'. Such discussions often include the introduction of the formalism of 'symbolic logic', which can make 'logic' appear to be purely mathematical – even mechanical – and unrelated to ordinary experience. The approach taken in *This Textbook*, in contrast, treats 'logic' as primarily a linguistic phenomenon: what constitutes a correct logical argument in mathematics is really the same as for a logical argument in, say, history or the law. That is, the correctness of an argument consists primarily in using *words* properly and precisely, and not just symbols. In particular, as happens with learning ordinary language, one develops the skill for producing logical arguments by encountering – and ultimately mimicking – many examples of such arguments. When reading proofs in *This Textbook*, for instance, the beginner should go carefully through the arguments – don't simply 'scan' over them. To help, many of the early proofs are given in excruciating detail, so read each detail. With experience, less and less needs to be included in the printed proof, and the reader can be expected to fill the gaps. On occasion in *This Textbook* there will be a discussion, often within a Side Comment, about logical arguments, to help less experienced readers.

(4) The most common format for writing rigorous mathematical exposition nowadays is the so-called **Definition-Theorem-Proof style**: first give one or more definitions of important concepts; then state theorems about those concepts; finally prove those theorems. Such an exposition can be very clean, but it puts great burdens on readers to figure out where any of this comes from and what it means to them. For example, in real-life mathematics, 'definitions' usually appear on the scene only after the importance of the underlying idea is already clear, and all one needs is to express it clearly in words and give it a name. That is, 'definitions' normally do *not* appear first chronologically in the development of mathematics.

(5) It is the experience of many teachers of elementary calculus that most of their students believe that the *words* which appear in their calculus textbooks form an optional part of the course; this includes definitions and statements of theorems. For example, if, at

the end of such a course, the teacher asks the students to give the definition of 'derivative' – arguably the most fundamental concept of the subject, and one whose definition was actually used repeatedly throughout the course – the result is often a wall of blank stares.

I.2 Basic Set Theory

In the last hundred and fifty years or so mathematics has become increasingly abstract and axiomatic. This process has been greatly facilitated by the use of set theory as a common language. The present section is devoted to a quick review of the basics of that theory. The treatment here is informal – some authors use the word 'naive' – to contrast it with more formal axiomatic approaches that study the foundations of the subject. In particular, the 'set theory' developed in *This Textbook* is a tool for organizing the study of analysis, and not a goal in itself.

The primitive concept needed for our treatment of basic set theory is that of a **collection of objects**. In particular, we do not define here what is meant by either an *object* or by a *collection*; instead, we assume that these these are 'primitive concepts' which fall under Justice Stewart's dictum of 'I know it when I see it'.

I.2.1 Definition

(1) A **set** is a collection of objects, thought of as a single object in its own right.

Note The phrase 'thought of an object in its own right' is crucial: it implies that whatever the primitive concept of 'object' means, and whatever the primitive concept of 'collection' means, anything which is a 'collection of objects', i.e., a 'set', is itself an 'object', and therefore can be used in the formulation of new 'collections of objects'; in particular, sets of sets.

One frequently uses the word **family** instead of the word 'set'. In *This Textbook* the use of 'family' is normally restricted to sets whose elements are themselves sets of objects; thus, we may refer to a '*family* of sets' instead of a '*set* of sets'.

(2) If X is a set of objects, then the objects which form the collection X are called the **elements of X**; they are also called **points of X** and **members of X**. If b is a member of X then one also says that **b belongs to X** or that **b is in X**; the standard shorthand notation for any of these statements is $b{\in}X$. Likewise, if c is an object which is *not* a member of the collection X, one says that **c is not an element of X**, etc, and one writes $c \notin X$.

Side Comment (on set theory): The theory of sets, as a distinct branch of mathematics, was initially founded by the German mathematician Georg Cantor in the 1870's. It is worth noting that this very abstract theory grew out of Cantor's work on Fourier series, a subject which plays a vital role in quite concrete applied mathematics. Fortunately, one does not have to know anything about Fourier series to understand the main ideas of Cantor's set theory.

Some Nonmathematical Sets

(1) The set whose elements are the letters

$$A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z$$

is called the (English) **alphabet**. More precisely, it is the set of upper-case letters, to contrast it from the corresponding set of lower-case letters. Note that in mathematics one often uses letters from this alphabet, or from the alphabet of some other language, in particular Greek, as the names or symbols for certain mathematical objects. In that context one normally treats the upper-case and lower-case versions of a given letter as denoting different objects.

(2) The set consisting of all the United States senators as of January 1, 2000, is an object in its own right: the US Senate (as of that date), denoted here by the letter S. Its elements are the individual senators.

This example illustrates the significance of the phrase 'thought of as an object in its own right' appearing in Definition (I.2.1). Indeed, as an object in its own right, the Senate S can have properties appropriate to the type of object it is. For example, the Senate, as a (finite) set of objects, has a 'membership size': there are 100 members (senators). In contrast, an element of the set S (i.e., an individual senator) has 'party affiliation'; for example, 'independent'. Note that the property 'membership size' does not apply to an individual senator, nor does the property 'party affiliation' apply to the Senate as a whole.

(3) Consider the set whose elements are the Greek letter α, the country Australia, and the emperor Napoleon Bonaparte. Each of the three elements of this set is, individually, of considerable significance; but the set itself, thought of as an object in its own right, appears to be of little interest (except, possibly, as an example of a set of little interest).

I.2.2 Some Standard Mathematical Sets and Their Symbols

In contrast to the preceding 'nonmathematical' sets, the following sets *are* of considerable importance for us; the symbols used here are fairly standard in mathematics, and appear throughout *This Textbook*.

(a) The symbol $\mathbb{N}$ denotes the set of all **natural numbers**;. These are the numbers used in the process of 'counting' which one learns as children. Indeed, they are often called the **counting numbers**. In the body of *This Textbook* we treat the concept of these numbers, and the corresponding basic properties, as 'primitive' concepts: we know them when we see them, but do not define them further. However, we also follow the usual custom and identify intuitively whatever 'primitive' notion of 'natural number' we have with 'strings of finite decimals expressions': 1, 2, 3, ... 100, 101, and so on; keep in mind, however, that historically the decimal notation arose much later.

Warning Despite the antiquity of these numbers, the terminology used for them is far from universal even today. For example, in areas such as computer science it is customary to include 0 as the 'initial' natural number; what most mathematicians call 'natural numbers' (i.e., starting with 1) might in such areas be referred to as **counting numbers**. Good arguments can be made for or against either usage. (For example, try replacing the word 'first' with the word 'zero-th' throughout Chapter Quote (3) at the start of this chapter.) In any event, the choice of which usage to follow is mainly one of taste, convenience, and the conventions within the particular field of study.

(b) The symbol $\mathbb{Z}$ denotes the set of all integers, i.e., the numbers $0,\ \pm 1,\ \pm 2,\ \ldots$. There appears to be universal agreement on what numbers constitute this set, and nearly universal agreement on the use of the symbol $\mathbb{Z}$. (The use of this symbol comes from the German word '**Z**ahl', which means 'number'.)

(c) The symbol $\mathbb{Q}$ denotes the set of all rational numbers (i.e., **Q**uotients of integers).

(d) The symbol $\mathbb{R}$ denotes the set of all **R**eal numbers.

(e) It is convenient to list here the standard notations for a few more important sets of numbers.

If k is a natural number, then $\mathbb{N}_k$ denotes the set of all natural numbers m such that $1 \leq m \leq k$.

In contrast to the preceding, if k is a natural number, then $\mathbb{Z}_k$ stands for the set of integers m such that $0 \leq m \leq k-1$.

The symbols $\mathbb{Z}^+$, $\mathbb{Q}^+$ and $\mathbb{R}^+$ denote the sets of positive integers, positive rational numbers, and positive real numbers, respectively. In particular, one has $\mathbb{N} = \mathbb{Z}^+$.

Let a and b be real numbers such that $a < b$. The **closed interval with endpoints $\boldsymbol{a}$ and $\boldsymbol{b}$**, denoted $[a,b]$, is the set of all real numbers x such that $a \leq x \leq b$. The corresponding **open interval** is the set (a,b) whose members are the real numbers x such that $a < x < b$. The corrsponding **half open intervals** $[a,b)$ and $(a,b]$ are defined analogously.

Remark The symbols $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ are stylized versions of the ordinary upper-case letters N, Z, Q and R, respectively, and in *This Textbook* are used only to refer to the indicated fundamental sets of numbers. In contrast, the 'unstylized' versions, N, Z, Q and R, are also used frequently in *This Textbook*, but *never* to refer to these sets.

In Appendix A the interested reader can find an axiomatic development of the arithmetic of the natural numbers. However, the treatment found in that appendix is not needed to understand the main body of *This Textbook*. In particular, it is assumed that the reader is already familiar with the standard arithmetic (addition, multiplication and, where appropriate, division) of these numbers, as well as the their standard order properties (greater than, less than).

Side Comment (on the 'natural number' terminology): An obvious way to avoid the ambiguity of what which numbers ought to be called 'natural numbers' (see the 'Warning' above) is to use the standard symbol $\mathbb{Z}^+$, which stands for 'the set of positive integers', in place of the symbol $\mathbb{N}$, and to use the phrase 'positive integer' instead of 'natural number'. Indeed, many authors do just that, despite the fact that the former expression is longer than the latter: more letters, more syllables. We avoid the 'positive integer' terminology, at least at this early stage, because it introduces an anachronism into the discussion: if one wants to *define* $\mathbb{N}$ to be $\mathbb{Z}^+$, one needs to know in advance what the set $\mathbb{Z}$ is. Historically speaking, however, the 'counting' numbers $1,\ 2,\ 3,\ \ldots$ came long before the number 0 or the negative numbers, and thus long before it would make sense to talk about the '*positive* integers'. Of course for a long time 'number' simply meant what we now call 'natural number'. The need for the prefix 'natural' arose, one presumes, from the introduction of new numbers which were originally thought to be 'unnatural', such as negative numbers.

I.2.3 Remarks

A good deal of terminology and notation has developed in mathematics around the concept of 'set'.

(1) Frequently a set is denoted by explicitly listing its elements, or by providing some other more-or-less explicit description of its elements. For instance, if Y is the set consisting of all the even natural numbers, then one can write

$$Y = \{2, 4, 6, 8, \ldots\} = \{2\,k : k = 1, 2, 3, \ldots\} = \{2\,k : k{\in}\mathbb{N}\}.$$

In the first expression for Y, the reader is expected to figure out from the pattern that is setting up that the set consists of 'all even natural numbers'; likewise, in the second expression, the reader is expected to realize that the quantity k runs over all the natural numbers, so that the expression $2k$ runs over all the even natural numbers. The third expression is technically the clearest, although it does require that the reader remember the definition of the symbol $\mathbb{N}$; or, failing that, be willing to look up its definition.

Note that in this so-called '**set-builder notation**' the braces $\{$ and $\}$ are used as 'delimiters', in the sense that they tell where the description of the set begins and ends.

(2) If X is a set, then, as is stated in Definition (I.2.1), X can be viewed as a single object in its own right. As a consequence, it is possible to form new sets whose elements are themselves sets. For instance, suppose that $X = \{1, 2, 3, 4, 5\}$ while $Y = \{10, 11, 12\}$. Let Z be the set whose elements are the objects X and Y; that is,

$$Z = \{X, Y\} = \{\{1, 2, 3, 4, 5\}, \{10, 11, 12\}\}.$$

Notice that the set Z has precisely two elements, namely the set X and the set Y, thought of as objects in their own right. In particular, Z is *not* the set $\{1, 2, 3, 4, 5, 10, 11, 12\}$, which has precisely eight elements.

(3) The statement 'A set is a collection of objects' in Definition (I.2.1) may make it appear that every set must have more than one member, since the word 'objects' here is a plural noun. In reality, however, mathematicians allow the title 'set' to be applied to smaller collections as well:

(a) Let c be an object. Then the set $X = \{c\}$, whose only element is the object c, is said to be a **singleton set**. Likewise, if b and c are objects such that $b \neq c$, then one calls the set $\{b, c\}$ a **doubleton set**.

(b) Let X be defined as 'the set of all natural numbers whose square is negative'. Of course, there are no such numbers, so this 'set' has no elements at all! Nevertheless, it is useful to allow such a set into the theory, as will become clear. This set with no elements is called the **empty set** and is denoted by the symbol $\emptyset$. Note that this example may suggest that the empty set is a set of natural numbers; however, see Example (I.2.5) (3) below.

(4) The special case in which X a singleton set $\{c\}$, thought of as an object in its own right, is worth separate mention. Authors frequently treat the set $\{c\}$ and original object c interchangeably, as if they were the same object; otherwise stated, they 'identify' the set $\{c\}$ with the object c. Of course, equating these objects is incorrect, since the object $\{c\}$ is a set, whose only element is the object c, while the object c is not. Usually the context makes clear what is really meant, and no confusion results. Nevertheless, is is considered good manners to warn the reader when such an **abuse of notation** (as identifications like these are usually called) is being used.

In general, any set which arises 'naturally' in mathematics consists of objects that are, in some sense, 'mathematical objects' that are 'related' to each other; for example, each element of the set $\mathbb{Q}$ is a ratio of integers. Indeed, a major part of mathematics consists of studying the properties which arise from the relations which hold between elements of specific sets. Nevertheless, in set theory there is no requirement that the elements of a set be related to each other in any way other than have been grouped into the same set. The only purely set-theoretic issue is whether an object is in the set or not. This fact is sometimes expressed as the following principle:

I.2.4 Fundamental Principle (The Axiom of Extension)

A set is completely determined by its elements. More precisely: The set X and the set Y are equal, i.e., they are the same object, if, and only if, they have exactly the same members; that is, every object in X is also an element of Y, and every object in Y is also an element of X.

The preceding 'Fundamental Principle' is not a 'theorem' of set theory; instead, it is really just a clarification of the definition. The name 'Axiom of Extension' (or, sometimes, 'Axiom of Extensionality') comes from more formal treatments of set theory.

This princple may seem too obvious to need statement, but there is real content to it. For instance, it implies that the *manner* in which one describes a set is irrelevant. It also implies that *how one uses* the elements of a set is irrelevant to which set is involved.

I.2.5 Examples

(1) Let $X = \{1, 2, 3\}$, let $Y = \{3, 2, 1\}$, and let $Z = \{1, 3, 2, 2, 1, 1, 3\}$. The *lists* of numbers used to describe X, Y and Z are not the same – the order is different, and in the case of Z the list has repetitions; That is, the given *descriptions* of these sets are all different. Nevertheless, the sets X, Y and Z are the equal. For instance, the numbers 1, 2 and 3 are clearly in the set Z, and equally clearly no other object is in Z. Thus, Z has precisely the same elements as the set X, so $Z = X$. Likewise one sees that $X = Y = Z$.

(2) Let X be the set of all real numbers x such that $-1 \leq x \leq 1$, and let Y be the set of all numbers of the form $\sin x$ for x in $\mathbb{R}$. It should be clear to the reader that $X = Y$; of course this assumes that the reader recalls certain specific facts from high-school trigonometry.

(3) Let X be the set of all real numbers x such that $x^2 = -2$, and let Y be the set of all live dinosaurs in the San Diego Zoo as of January 1, 2000. It is clear from these definitions

that X is a set of numbers, while Y is a set of animals. Since no number is an animal and no animal is a number, at first glance it would appear that these sets could not possibly be equal. However, these sets do have exactly the same elements, in the sense that there is no element of X which is not in Y, and there is no element of Y which is not in X. Thus, $X = Y$; in fact, both X and Y equal the empty set $\emptyset$ discussed earlier. This illustrates the fact that there is only one 'empty set', which is why we can refer to *the* empty set and can use a fixed symbol, $\emptyset$, to denote it.

I.2.6 Definition

(1) Suppose that X and Y are sets, and suppose further that every element of X is also an element of Y. One then says that **X is a subset of Y**; in symbols, $X \subseteq Y$. This relationship between X and Y is also written $Y \supseteq X$; in words: '**Y is a superset of X**'.

(2) If $X \subseteq Y$ but $X \neq Y$ then one says that X is a **proper subset** of Y; the phrase 'proper superset' is defined analogously. For example, the empty set $\emptyset$ is a subset of *every* set X, and is a proper subset of every nonempty set.

Warning on Notation: Some math texts use the symbols $\subset$ and $\supset$ for what we write here as $\subseteq$ and $\supseteq$, respectively. Unfortunately, many other math texts use the same symbols, $\subset$ and $\supset$, to refer to *proper* subsets and supersets, respectively. Because of this, we avoid the use of $\subset$ and $\supset$ in *This Textbook.* If there is a need to indicate explicitly that X is a *proper* subset of Y, or, equivalently, Y is a *proper* superset of X, then we use the unambiguous notations $X \underset{\neq}{\subset} Y$ and $Y \underset{\neq}{\supset} X$.

It is convenient to formulate the 'Axiom of Extension' in terms of the 'subset' relation.

I.2.7 Theorem

Let X and Y be sets. A necessary and sufficient condition for these sets to be equal, i.e., for the relation $X = Y$ to hold, is that $X \subseteq Y$ and $Y \subseteq X$; equivalently, $Y \supseteq X$ and $X \supseteq Y$.

The (trivial) proof, which consists of recalling the meaning of the notations $\subseteq$ and $\supseteq$, is left to the reader.

Let us continue the very informal discussion of 'sets', given above, but now in the more formal 'Definition-Theorem-Proof' style. The first definition describes some standard ways of constructing new sets from old.

Two Notes on Language:

(1) In the following definition, as elsewhere in *This Textbook*, we follow the standard convention in mathematical writing and use the 'inclusive' sense of the word 'or'. For example, to say that 'an element c is a member of either the set X or the set Y' does *not* preclude the possibility that c could be a member of both sets. This contrasts the usage in ordinary (i.e., nonmathematical) English, in which the 'or' is often used in the 'exclusive' sense:

'Do your homework or you won't get to watch television.'

This usually implies that you *will* get to watch TV if, in fact, you do your homework.

(2) There is another problem in English usage which appears both within and outside of mathematics: the ambiguous indefinite article.

Example Consider the following questions:

(1) 'Does Kim have an apple?'

(2) 'Does the quadratic equation $x^2 - 4 = 0$ have a solution?'

In both questions the use of the indefinite articles '*an*' and '*a*' is ambiguous. For example, in Question (1), is the questioner asking whether Kim has *exactly* one apple, or whether Kim has *at least* one apple? Likewise, in Question (2), is the questioner asking whether the equation has *exactly* one solution, or whether it has *at least* one solution?

Experience suggests that in mathematical writing the meaning of the indefinite articles '*an*' and '*a*' is used is to mean 'at least one'. In any event, the obvious way to avoid such ambiguity is to append phrases such as 'at least one' or 'exactly one', as appropriate.

I.2.8 Definition

Let X and Y be sets.

(1) The **complement of $\boldsymbol{X}$ in $\boldsymbol{Y}$**, denoted $Y \backslash X$, is the set whose elements are precisely those members of Y which are *not* members of X; the symbol ' $\backslash$ ' is sometimes pronounced 'minus' in the present context.

(2) The **union of $\boldsymbol{X}$ and $\boldsymbol{Y}$**, denoted $X \cup Y$, is the set whose elements are precisely those objects which belong either to X or to Y; the symbol $\cup$ is pronounced 'union' or 'cup'. (As was stated in the 'Note' above, we use the 'inclusive or' here, so that an object which is an element of both X and Y is also an element of $X \cup Y$.)

Similarly, let k be a natural number, and let $X_1, X_2, \ldots X_k$ be sets. Then the 'union' of these sets is the set $X_1 \cup X_2 \cup \cdots \cup X_k$ whose elements are those objects which belong to *at least one* of the sets $X_1, \ldots X_k$. Such unions are also denoted by expressions such as $\bigcup_{j=1}^{k} X_j$.

(3) The **intersection of $\boldsymbol{X}$ and $\boldsymbol{Y}$**, denoted $X \cap Y$, is the set whose elements are precisely those objects which belong to *both* X and Y; the symbol $\cap$ is pronounced 'intersection' or 'cap'. If $X \cap Y = \emptyset$, then the sets X and Y are said to be **disjoint**

Similarly, if $X_1, X_2, \ldots X_k$ are sets then their 'intersection' is the set $X_1 \cap X_2 \cap \cdots \cap X_k$ whose elements are those objects which belong to *each* of the sets $X_1, \ldots X_k$. Such intersections are also denoted by expressions such as $\bigcap_{j=1}^{k} X_j$.

(4) More generally, let $\mathcal{A}$ be a nonempty family of sets; that is, $\mathcal{A}$ is a nonempty set such that each element of $\mathcal{A}$ is itself a set; we allow the empty set $\emptyset$ itself to be an element of the family $\mathcal{A}$. Then the **union of the family $\boldsymbol{\mathcal{A}}$**, denoted by $\bigcup \mathcal{A}$, is the set Z such that an object x is an element of Z if, and only if, x is an element of *at least one* set belonging to the family $\mathcal{A}$. Similarly, the **intersection of the family $\boldsymbol{\mathcal{A}}$**, denoted by $\bigcap \mathcal{A}$, is the set W such that an object x is an element of W if, and only if x is an element of *each* set belonging to the family $\mathcal{A}$.

It is possible to assign meanings to the union and intersection of the *empty* family of sets, but some subtle complications arise. We never need to consider the union or intersection of the empty family of sets in *This Textbook*, which allows us to ignore those complications.

Alternate Notation: The sets $\bigcup\mathcal{A}$ and $\bigcap\mathcal{A}$ are often denoted by symbols such as $\bigcup_{X\in\mathcal{A}} X$ and $\bigcap_{Y\in\mathcal{A}} Y$, in analogy to the notations $\bigcup_{j=1}^{k} X_j$ and $\bigcap_{j=1}^{k} X_j$ used in Parts (2) and (3) above.

Important Special Case: Suppose that $\mathcal{A} = \{X_1, X_2, \ldots X_j, \ldots\}$, where the quantity j ranges over the set $\mathbb{N}$ of all natural numbers. Then it is customary to denote the union and intersection of this family by expressions such as $\bigcup_{j=1}^{\infty} X_j$ and $\bigcap_{j=1}^{\infty} X_j$, respectively.

(5) Let $\mathcal{A}$ be a nonempty family of sets, as in Part (4) above. The sets in this family are said to be **mutually disjoint** provided that if X and Y are elements of the family $\mathcal{A}$ such that $X \neq Y$, then $X \cap Y = \emptyset$. In particular, this property certainly holds if the family $\mathcal{A}$ has exactly one element.

Note If $k = 1$, then, by convention, $\bigcup_{j=1}^{k} X_j$ and $\bigcap_{j=1}^{k} X_j$ both equal X_1.

The next results summarize a few of the properties associated with the preceding definitions. The proofs of these results are easy, so most of them are left as exercises.

I.2.9 Theorem

(a) Let X and Y be sets. Then
(i) $Y\backslash X$ is a subset of Y; written symbolically, $(Y\backslash X) \subseteq Y$.
(ii) Equality occurs in (i) (that is, one has $Y\backslash X = Y$) if, and only if, $X \cap Y = \emptyset$.

(b) Let X, Y and Z be sets. If $X \subseteq Y$ and $Y \subseteq Z$, then $X \subseteq Z$.

Note: Because of this last fact, it is convenient to use the notation $X \subseteq Y \subseteq Z$ as an abbreviation for the compound hypothesis '$X \subseteq Y$ and $Y \subseteq Z$' stated in this result.

(c) Let X be a set. Then

$$(i)\ \emptyset \subseteq X \subseteq X; \quad (ii)\ X \cup \emptyset = X; \quad (iii)\ X \cap \emptyset = \emptyset; \quad (iv)\ X\backslash\emptyset = X.$$

Furthermore, a necessary and sufficient condition for the statement '$X \subseteq \emptyset$' to be true is that $X = \emptyset$.

(d) Let X be a set. Then $X \cap X = X \cup X = X$.

(e) ('Commutative Laws for Union and Intersection') Let X and Y be sets. Then

$$X \cup Y = Y \cup X \text{ and } X \cap Y = Y \cap X.$$

(f) ('Associative Laws for Union and Intersection') Let X, Y and Z be sets. Then

$$(X \cup Y) \cup Z = X \cup (Y \cup Z) \text{ and } (X \cap Y) \cap Z = X \cap (Y \cap Z).$$

More precisely, one has

$$(X \cup Y) \cup Z = X \cup (Y \cup Z) = X \cup Y \cup Z \text{ and } (X \cap Y) \cap Z = X \cap (Y \cap Z) = X \cap Y \cap Z,$$

where $X \cup Y \cup Z$ and $X \cap Y \cap Z$ are as described in the preceding definition.

(g) Let X and Y be sets. Then $X \cup Y = Y$ if, and only if, X is a subset of Y. Likewise, $X \cap Y = X$ if, and only if, X is a subset of Y.

(h) Let X and Y be sets. Then $Y \backslash (Y \backslash X) = X \cap Y$. In particular, a necessary and sufficient condition for the equation $Y \backslash (Y \backslash X) = X$ to hold is that X be a subset of Y. Likewise, a necessary and sufficient condition for the equation $Y = (Y \backslash X) \cup X$ to be true is that X be a subset of Y.

Partial Proof All the parts of this theorem are easy to prove, usually by simply using the definitions. Let us carry out the proof for Part (a). To save some writing, let $Z = Y \backslash X$ throughout this proof.

Proof of Part (i) of (a): By the definition of 'complement' (see Part (a) of Definition (I.2.8)), the set Z consists of those elements of Y which are *not* in X. In particular, every element of Z is also an element of Y, and thus (by the definition of 'subset') Z is a subset of Y.

Proof of Part (ii) of (a): For convenience let us break up the proof into two parts.

The 'if' portion of Statement (ii): Suppose that $X \cap Y = \emptyset$, and let c be any element of Y. Clearly such c cannot be an element of X; for if it were, then it would be in both X and Y, hence (by definition of 'intersection') in $X \cap Y$, contradicting the hypothesis above that $X \cap Y = \emptyset$. Thus every such c is an element of Z; that is, $Y \subseteq Z$. Now combine this with the result of Part (i), together with Theorem (I.2.7), to conclude that that $Y = Z$, i.e., $Y = Y \backslash X$, as claimed.

The 'only if' portion of Statement (ii): Suppose that $Y = Y \backslash X$. It follows from this equation that if c is any element of Y then c is also an element of $Y \backslash X$, and thus, in particular, c is *not* an element of X. Thus no element of Y is also in X, hence $X \cap Y = \emptyset$, as claimed.

Side Comment (on proofs): The proofs given above are not difficult; but they do have some features that are worth a little extra discussion, especially for readers having limited experience with rigorous mathematics.

(1) The proof of Part (ii) of (a) is broken into the 'if' portion and the 'only if' portion. This reflects the fact that there are really two substatements which combine to give the full meaning of Statement (ii):

(ii)$_\alpha$ 'One has $Y \backslash X = Y$ **if** $X \cap Y = \emptyset$; and
(ii)$_\beta$ 'One has $Y \backslash X = Y$ **only if** $X \cap Y = \emptyset$'.

As the proof given above indicates, each of these substatements is a logical implication in which something is assumed ('the hypothesis'), from which something else is deduced ('the conclusion'):

(ii)$_\alpha$ 'If $X \cap Y = \emptyset$ (hypothesis), then $Y \backslash X = Y$ (conclusion)'
(ii)$_\beta$ 'If $Y \backslash X = Y$ (hypothesis), then $X \cap Y = \emptyset$ (conclusion)

Note that each of these substatements is simply the converse of the other. The complete Statement (ii) simply asserts that both of these substatements are simultaneously true:

(ii) 'If $X \cap Y = \emptyset$, then $Y \backslash X = Y$; and, conversely, if $Y \backslash X = Y$, then $X \cap Y = \emptyset$.'

Why would a mathematics text use the original 'if, and only if' phrasing of Statement (ii) when the second phrasing is less likely to cause confusion? Simple: The original phrasing takes up less space.

In mathematics the phrase 'if, and only if,' appears so frequently that it is treated almost as a single word: 'if-and-only-if'; indeed, many authors use the written abbreviation 'iff' (pronounced ‘if-and-only-if’, not ‘iffffff’) for this phrase. In *This Textbook*, however, we always write out the full phrase to emphasize that whenever this phrase appears it indicates that *two* separate statements are under consideration, and separate proofs are needed for each.

(2) In a similar manner, the phrase ‘necessary and sufficient condition’ appears frequently in mathematics. For instance, in Part (h) of the preceding theorem the presence of this phrase means that the statement there really breaks into two substatements:

‘A *necessary* condition for $Y\backslash(Y\backslash X) = X$ to hold is that X be a subset of Y.’

‘A *sufficient* condition for $Y\backslash(Y\backslash X) = X$ to hold is that X be a subset of Y.’

A complete proof of (h), then, would consist of proving both of these substatements. Note that the ‘condition’ referred to in each of these substatements is the phrase ‘X be a subset of Y’. (The actual condition would normally be stated by itself as ‘X is a subset of Y’; it is cast into the subjunctive voice in the statement of the theorem because of the rules of English grammar.)

Many students in math courses get almost as confused about the distinction between ‘necessary’ and ‘sufficient’ as they are about ‘if’ *vs* ‘only if’. They usually believe that the reason this confusion is because the context is a math course, and that mathematicians think differently from regular folks. In reality, however, the issue is one of understanding the use of such words in ordinary language. For example, consider the following statements:

‘A **necessary** condition to get your first driver’s license in California is to pass the written test.’

‘A **sufficient** condition to get your first driver’s license in California is to pass the written exam.’

Clearly, the second statement is not correct: passing the written exam is not enough; you also need to pass the driving test. In contrast, the first statement is true; indeed, passing the written exam is needed before you are even allowed to take the driving test.

(3) Some readers may wonder why the proof of Part (a) starts out by assigning the symbol Z to the set $Y\backslash X$; indeed, some might worry that there is something mystic going on which they are missing, because they would not have thought of starting that way. The explanation, however, is both benign and boring: The author realized, while doing the first draft of the proof, that the expression $Y\backslash X$ would need to be referred to repeatedly, so that the slight extra work of introducing the symbol Z in the final drafts would save enough typing to be worth the effort.

The main point of interest, however, is the fact that the proofs require the readers to know the meanings of all the words and phrases that appear in it! (You might be surprised how often it happens in math courses that students cannot do a problem simply because they don’t know what the words mean, and it does not occur to them to look up those meanings.) Sometimes this requirement, namely that the readers are expected to know the definitions of the terms, is made explicit; for example, in the proof of Part (i) of (a) the reader is told directly that the definitions of ‘complement’ and ‘subset’ are relevant.

In the rest of the proof of Part (ii) of (a), however, the requirement of knowing what the words mean is implicit. For instance, in the proof of the ‘only if’ portion of (ii) the argument that c is not in $X \cap Y$ uses the definition of ‘intersection’: since c can’t be in X, it can’t be in both X and Y simultaneously and thus (here’s where the definition of ‘intersection’ gets used) it can’t be in $X \cap Y$.

Normally the instruction, ‘Know the definitions of the words that appear!’, is left unstated; but it is always understood to hold.

Moral of this Story:

(a) You need to know the precise definition of everything you deal with in mathematics; in particular: ‘When in doubt, look it up’. Similarly, you are expected to know the precise statements of the theorems which you may need to use.

(b) You should expect to write up more than one draft of your work in math courses before handing it in to be graded. In doing so, you may find it useful to introduce new notation, or even new terminology, to simplify the job of your readers.

I.3 Some Properties of the Set $\mathbb{N}$ of Natural Numbers

As is mentioned earlier in this Chapter, a major goal in modern mathematics is to formulate the basic concepts and facts in terms of set theory. This includes even older results which were well understood before set theory and its terminology were introduced.

One of the most important facts about natural numbers is the classic Principle of Mathematical Induction. The discussion below formulates this Principle, as well as a couple of its near relatives, in terms of sets.

Side Comment (on the classic formulation of the Principle of Mathematical Induction): The Principle in question was classically formulated, without referring to sets, along the following lines:

For each natural number k let $S(k)$ denotes a statement which involves the natural number k. Suppose that:

(i) Statement $S(1)$ is true.

(ii) For every natural number k the truth of Statement $S(k)$ implies the truth of Statement $S(k+1)$.

Then $S(k)$ is true for *every* natural number k.

Remark The texts which use this statement of the Principle usually provide a few examples of what the phrase 'statement which involves the natural number k' signifies; a common example is the statement

'If k is a natural number, then the sum of the first k natural numbers equals $\dfrac{k\,(k+1)}{2}$.'

However, they usually don't define the meaning of 'a statement which involves the natural number k' any further, leaving it as a primitive concept for which 'I'll know it when I see it'.

I.3.1 Three Important Principles in $\mathbb{N}$

(a) **The Principle of Mathematical Induction** Suppose that A is a set of natural numbers such that the following conditions hold:

(i) (Initial Step) $1{\in}A$.

(ii) (Induction Step) If k is any element of $\mathbb{N}$ such that $k{\in}A$, then $(k+1){\in}A$.

Conclusion $A = \mathbb{N}$.

(b) **The Least-Natural-Number Principle** Suppose that B is a nonempty subset of $\mathbb{N}$. Then the set B has a least member. That is, there is a (unique) natural number m such that m is an element of B, and $m \leq k$ for every number k in B.

(c) **The Greatest-Natural-Number Principle** Suppose that C is a nonempty subset of $\mathbb{N}$ which is **bounded above in** $\mathbb{N}$, in the sense that there exists a number q in $\mathbb{N}$ such that $q \geq k$ for all k in C. Then the set C has a greatest element. That is, there exists a (unique) natural number n such that n is an element of C, and $n \geq k$ for all k in C.

Remark It is conventional to accept these Principles on faith, with primary emphasis given to Principle (a). However, many students of mathematics, even fairly advanced ones, find Principles (b) and (c) easier to understand than the original Induction Principle (a). In fact, it is not hard to show that each of these principles implies the other two. We illustrate one such proof below.

Proof that Principle (b) Implies Principle (a) We assume here that the basic properties of $\mathbb{N}$ known from elementary-school arithmetic.

Suppose that Principle (b) is true, and let A be a subset of $\mathbb{N}$ which satisfies Hypotheses (i) and (ii) of Principle (a). Now suppose, on the contrary, that this A does not also satisfy the conclusion of Principle (a); that is, suppose that $A \neq \mathbb{N}$, then the set $B = \mathbb{N} \setminus A$ is a nonempty subset of $\mathbb{N}$. Let m be the minimum element of B whose existence is guaranteed by Principle (b). Clearly Hypothesis (i) of Principle (a) implies that 1 is not in B, so $m > 1$. Then $m - 1$ is an element of $\mathbb{N}$ less than m so that $m - 1$ cannot be in B, hence $(m-1) \in A$. But then Hypothesis (ii) implies that $m = (m-1) + 1$ is also in A, contrary to the construction of m as the smallest element of B. That is, Principle (a) is true.

Remark Many authors refer to Principle (b) above as the **Well-Ordering Principle**, or sometimes the **Well-Ordering Principle in $\mathbb{N}$**. The Side Comment below explains the preference in *This Textbook* for using the nonstandard name ‘Least-Natural-Number Principle’.

> Side Comment (on preferring the name ’Least-Natural-Number Principle’): The name used in *This Textbook* for the principle in question is definitely not standard; a much more common name for it is ‘The Well-Ordering Principle’. The root of the latter name is a much deeper result, called ‘The Well-Ordering Theorem’, which arises in advanced set theory, but which normally is not described in this more elementary situation; roughly speaking, this deeper result states that if X is an arbitrary nonempty set, then there exists a notion of an ‘ordering’ $<.$ or ‘less than’, on X relative to which every nonempty subset of X has a least element; equivalently, there exists a notion of $>$, or ‘greater than’, on X relative to which every nonempty subset of X has a greatest element. Under such an ordering the set X is said to be ‘well ordered’. The principle under discussion then translates to say that the set $\mathbb{N}$, under its standard order of ‘less than’, is ‘well ordered’. Singling out this one example to assign the name ‘Well-Ordering Principle’, however, without explaining the origin of the phrase ‘well ordering’, seems pointless, especially when the name ‘Least-Natural-Number Principle’ actually reminds one of the content of Principle (b).
>
> A second reason for preferring the nonstandard name ‘Least-Natural-Number Principle’ is that it complements the name ‘Greatest-Natural-Number Principle’. The latter principle is *not* a special case of the ‘Well-Ordering Principle’, not even in the alternate formulation given above, because it requires the additional ‘bounded above’ hypothesis.

The next result is a useful variant of the Principle of Mathematical Induction, and could also be considered to be a ‘well-known result’, so that its proof would not be needed. We include a proof here to illustrate the use of the Greatest-Natural-Number Principle.

Reminder If $k \in \mathbb{N}$, then $\mathbb{N}_k$ is the set of all natural numbers j such that $1 \leq j \leq k$. In particular, $\mathbb{N}_1$ is the singleton set $\{1\}$.

I.3.2 Theorem (The Strong Principle of Mathematical Induction)

Suppose that A is a subset of $\mathbb{N}$ such that the following conditions hold:

(i) (Initial Step) $\mathbb{N}_1 \subseteq A$.
(ii) (Induction Step) If k is any element of $\mathbb{N}$ such that $\mathbb{N}_k \subseteq A$, then $(k+1)\in A$.
Conclusion $A = \mathbb{N}$.

Proof (by contradiction) Suppose that there is a subset A of $\mathbb{N}$ satisfying Conditions (i) and (ii) above, but $A \neq \mathbb{N}$. Let C be the set of all k in $\mathbb{N}$ such that $\mathbb{N}_k \subseteq A$. Since, by the Initial Step, one has $\mathbb{N}_1 \subseteq A$, it follows that $1\in C$ and thus $C \neq \emptyset$. Likewise, by the 'contradiction hypothesis' that $A \neq \mathbb{N}$, there must exist q in $\mathbb{N}$ such that $q \notin A$. It is clear that that if $k\in C$ then $k < q$ and thus the set C is bounded above in $\mathbb{N}$. It follows from the Greatest-Natural Number Principle that C has a greatest element; call it m. It follows from the definition of the (nonempty) set C that $\mathbb{N}_m \subseteq A$. Because of Condition (ii), it then follows that $(m+1)\in C$, and thus $\mathbb{N}_{m+1} = \mathbb{N}_m \cup \{m+1\} \subseteq A$, hence $(m+1)\in C$. This contradicts the fact that m is supposed to be the *largest* element of C.

Remarks (1) Since $\mathbb{N}_1 = \{1\}$, Condition (i) is simply a fancy way of stating that 1 is an element of the set A. In contrast, when $k \geq 2$, Condition (ii) is a stronger hypothesis than the corresponding Induction Step in the regular version of the Principle of Mathematical Induction.

(2) It is a useful exercise to show that the Strong Induction Principle implies the regular Induction Principle.

The terminology introduced next should be familiar from elementary-school arithmatic.

I.3.3 Definition

(1) A natural number k is said to be a **composite number** provided it can be expressed, in at least one way, as the product $k = i{\cdot}j$ of two natural numbers i and j such that $i \geq 2$ and $j \geq 2$.

(2) A noncomposite number greater than 1 is called a **prime number**.

Remarks (1) The requirements $i \geq 2$ and $j \geq 2$ on the factors i and j is to ensure that we consider only *nontrivial* factorizations of k; indeed, *every* natural number m admits the 'trivial' factorizations $m = m{\cdot}1 = 1{\cdot}m$. Note that one also has $i \leq k-1$ and $j \leq k-1$.

(2) Because of the trivial nature of the number 1 as a factor, it is conventional to separate the noncomposite numbers into two types: the number 1, and all other noncomposite numbers. One result of this is that the number 1 is, by convention, not considered to be a prime number, so that 2 is the smallest prime number. This convention makes the phrasing of some results about natural numbers a little easier; for example, Theorem (I.3.5) below.

I.3.4 Lemma

Every composite number can be expressed as the product of two or more factors, each factor being a prime number.

Proof Let A be the set of all natural numbers k such that at least one of the following statements is true:
(i) $k = 1$.

(ii) k is a prime number.

(iii) k is the product of two or more prime factors.

The statement to be proved is then equivalent to showing that $A = \mathbb{N}$. We use the Strong Principle of Mathematical Induction to prove this last equality.

Initial Step Note that $1 \in A$, by definition of A, hence $\mathbb{N}_1 \subseteq A$.

Inductive Step Suppose that k is an element of $\mathbb{N}$ such that $\mathbb{N}_k \subseteq A$. Clearly $k+1 = 1$ is an impossibility, so there are two possibiities:

(a) $k+1$ is a prime number. In this case $k+1$ is in A by the definition of A.

(b) $k+1$ is *not* a prime number. Then $k+1$ must be a composite number, so $k+1 = m \cdot n$ for some natural numbers m and n satisfying $2 \leq m \leq k$ and $2 \leq n \leq k$. In particular, by the induction hypothesis that '$\mathbb{N}_k \subseteq A$', one sees that both m and n are in A. Since $m > 1$ it follows that m is either a prime number or a product of primes; likewise n is either a prime or a product of primes. It then follows that $k+1 = m \cdot n$ is also a product of primes, as required, so $k+1$ is in A in this case as well.

Now the Strong Principle of Mathematical Induction implies that $A = \mathbb{N}$, as required.

Note The statement of this Lemma could have been accepted here as a 'well-known fact' and the proof omitted. The real reason for including a proof is to illustrate how to use the Strong Principle of Mathematical Induction. It is instructive to try to prove the preceding result using only the the original ('weak') form of the Principle of Mathematical Induction.

I.3.5 Theorem (Unique-Prime-Factorization Theorem)

If n is a natural number and $n \geq 2$ then either n is a prime, or n can be expressed, in exactly one way, as a product

$$n = p_1 \cdot p_2 \cdot \ldots \cdot p_m,$$

where $m \geq 2$ and $p_1, p_2, \ldots p_m$ are primes such that $1 < p_1 \leq p_2 \leq \ldots \leq p_m$.

The proof of this (well-known) result is left as an exercise. (It follows easily from Lemma (I.3.4).)

Remarks (1) This result is often called the **Fundamental Theorem of Arithmetic**.

(2) In the preceding result, if p is a prime number such that either $n = p$ (in the case n is a prime) or p is one of the factors $p_1, p_2, \ldots p_m$ (if n is not a prime), then one calls p a **prime factor of n**

The Unique-Prime-Factorization Theorem above has many consequences in arithmetic; for example, it can be used to characterize those natural numbers which are squares of *rational* numbers. More generally, one has the following result.

I.3.6 Theorem

Let n and m be natural numbers. If there is a positive *rational* number r such that $r^m = n$, then r is actually a natural number.

Proof Suppose that r is of the form $r = j/k$, where j and k are natural numbers, and $r^m = n$. Without loss of generality one can assume that the prime factorizations of j and k have no prime factors in common. (Indeed, by canceling out any such factors in the division j/k, one gets $r = j'/k'$ where j' and k' have no prime factors in common.) If $k = 1$ then $r = j$, a natural number, as claimed. Thus assume that $k \geq 2$, so that $j \geq 2$ as well, since $n \geq 1$. Then one has $j^m = r^m\, k^m = n\, k^m$. If $k \geq 2$, then either k is itself a prime number p or else k has at least one prime factor p. It follows easily from the Unique-Prime-Factorization Theorem that p is a prime factor of $n\, k^m$, hence of j^m, hence of j itself, contrary to the hypothesis that j and k have no prime factors in common. It follows that $k = 1$, and the desired result follows.

I.3.7 Examples

(1) There is no rational number r such that $r^2 = 2$. Indeed, if such r existed, it could be chosen so that $r > 0$, and thus it would follow from the preceding result that 2 is the square of some natural number, which is obviouly not true: note that $1^2 = 1 < 2$, while $j^2 > 2$ if $j \geq 2$.

(2) There is no rational number r such that $r^3 = 10$. Indeed, it is clear that any such r must be positive, and thus by the preceding theorem 10 would the cube of some natural number. But one computes that $1^3 = 1$, $2^3 = 8$ and $j^3 \geq 27$ for all $j \geq 3$. The claim follows.

I.4 Ordered Tuples; Cartesian Products of Sets

Introduction The process of strengthening the logical foundations of mathematics, begun in the late nineteenth century, has evolved into the modern situation in which everything ultimately is reduced to set theory. This reduction means that any deep foundational issues can be passed off to the set theorists and logicians to deal with.

One simple concept, used in many important constructions in mathematics, is that of an 'ordered pair' of objects. For example, in the elementary analytical geometry of the Euclidean plane, one identifies a (geometric) point P in that plane with an ordered pair (x, y) of real numbers, where the first number x denotes the abscissa of P and the second y denotes its ordinate (relative to given coordinate axes). (The use of the words 'first' and 'second' in this context reflects the 'order' of this ordered pair.) Likewise, in the next section we define the vital concept of 'function' in terms of sets of ordered pairs.

In order to comply with the spirit of 'reduce everything to set theory', the concept of 'ordered pair' should be also be defined purely in terms of sets. Such a definition was given by Wiener in 1912, and simplified by Kuratowski in 1921; we present the Kuratowski definition of 'ordered pair' in Appendix B. For the present section, however, we treat the concept of 'ordered pair' as a primitive concept: one knows it when one sees it.

In a similar manner, it is convenient to accept as primitive the concepts of ordered triples, ordered quadruples, ordered quintuples, ordered sextuples, ordered septuples, and so on. More generally, if k is a natural number, then we treat as 'primitive' the concept of an **ordered *k*-tuple** $(x_1, x_2, \ldots x_k)$ of objects, in which the 'order' is indicated here by the conventional left-to-right listing of the objects. (Of course, the concept of 'left-to-right listing' is taken here as primitive: you know it – literally – when you see it. It is closely related to the primitive concept of 'chronological order': one speaks the name x_1 first, then x_2, and so on.) For instance, in the ordered 3-tuple (triple) $(2, 4, -1)$ of numbers, the first entry is 2, the second is 4, and the third is -1. Likewise, in the ordered 3-tuple $(2, 3, 2)$ the first entry is 2, the second is 3, and the third is 2 again.

(2) In *This Textbook* the phrase 'ordered k-tuple' is often abbreviated to 'k-tuple'; likewise, if k is understood from the context, k-tuple' may be abbreviated to 'tuple'.

(3) The case $k = 1$ is special. The primitive concept of an ordered 1-tuple (x) is that it is simply the single object x itself; of course there is no real 'order' in this situation. In Appendix B, however, we develop the concept of 'tuple' from other, more primitive, concepts. With that treatment the 1-tuple (x) is technically *not* the same object as the original object x itself. Nevertheless, even in that more modern treatment one often ends up identifying the 1-tuple (x) with the original object x; no confusion seems to result.

I.4.1 Remarks

(1) The main feature of the primitive notion of 'ordered pair' is that two ordered pairs, (a, b) and (c, d), are equal if, and only if, both of the equations $a = c$ and $b = d$ hold. Note that if $a \neq b$ then this feature implies that $(a, b) \neq (b, a)$. In contrast, Principle (I.2.4), the 'Axiom of Extension', implies that the corresponding *sets*, namely $\{a, b\}$ and $\{b, a\}$ are equal. In particular the 'obvious idea' to identify the primitive concept of 'ordered pair (a, b)' with the set $\{a, b\}$, fails to reflect the 'main feature' described above and thus is unsuitable as a set-theoretic characterization of the primitve concept.

Likewise, the main feature of the primitive notion of 'ordered k-tuple' is that two such k-tuples of objects, $(x_1, x_2, \ldots x_k)$ and $(y_1, y_2, \ldots y_k)$, are equal if, and only if, $x_j = y_j$ for each $j = 1, 2, \ldots k$.

(2) Depending on the context, the notation (a, b) can be ambiguous. For example, it may refer to the ordered pair (a, b), formed by objects a and b, as above. But if a and b are real numbers such that $a < b$, then it may denote an open interval in $\mathbb{R}$, a very different idea. This ambiguity normally causes no difficulty in a given context, and it can be avoided easily by using clarifying phrases, such as 'the ordered pair (a, b)' or 'the open interval (a, b)' as needed.

Side Comment (on the 'tuple' terminology) The use of the ending 'tuple' in 'k-tuple' comes from ordinary English, where analogous words are 'singleton', 'pair', triple', 'quadruple', 'quintuple', 'sextuple', 'septuple', 'octuple' and so on; 'tuple' is the generic ending for such formations once the first few special cases are passed.

It turns out in modern mathematics that *collections* of ordered tuples are important.

I.4.2 Definition (Cartesian Products of Sets)

(1) Let (A,B) be an ordered pair of nonempty sets. The **Cartesian Product of *A* with *B*** is the set $A{\times}B$ consisting of all ordered pairs (a,b) with $a{\in}A$ and $b{\in}B$. That is, an object p is an element of the set $A{\times}B$ if, and only if, there exists an element a in A and an element b in B such that $p\,=\,(a,b)$.

(2) More generally, let k be a natural number such that $k \geq 2$, and let $(X_1, X_2, \ldots X_k)$ be an ordered k-tuple of nonempty sets. Then the corresponding **Cartesian Product $\boldsymbol{X_1{\times}X_2{\times}\ldots{\times}X_k}$** is the set of all ordered k-tuples of the form $(x_1, x_2, \ldots x_k)$ such that for each $j\,=\,1, 2, \ldots k$ the object x_j is an element of the set X_j. The sets X_1, $X_2, \ldots$ X_k are called the **factors** of this Cartesian product.

If any of the sets $X_1, \ldots$ X_k equals the empty set, one sets $X_1{\times}X_2{\times}\ldots{\times}X_k\,=\,\emptyset$.

(3) If there is a set Y such that $X_j\,=\,Y$ for all $j\,=\,1,2,\ldots k$, then one usually writes Y^k instead of $X_1{\times}X_2{\times}\ldots{\times}X_k$. For example, one usually writes Y^4 instead of $Y{\times}Y{\times}Y{\times}Y$.

Remarks (1) The symbol × is usually pronounced 'cross' in English; thus, for example, the expression $A{\times}B$ would be spoken '*A cross B*'.

(2) The use of the word 'product' in this context is because the 'cross' notation is also used in elementary arithmetic to denote the ordinary product of numbers. This also explains the 'exponential notation' Y^k used above: it mimics the exponential notation used in arithmetic. However, there is no further algebraic theory associated with the 'product' used in 'Cartesian product'.

(3) The word 'Cartesian' is capitalized here because it derives from a proper name; namely, from the second name of René Des**cartes**, the founder of modern analytic geometry.

I.4.3 Examples

(1) Let $A\,=\,\{1,2,3\}$ and $B\,=\,\{2,4\}$. Then it is clear that

$$A{\times}B\,=\,\{(1,2),(1,4),(2,2),(2,4),(3,2),(3,4)\}.$$

Indeed, each of the ordered pairs listed on the right side of the preceding equation is certainly in the set $A{\times}B$, since in each case the first entry is in the set A while the second is in B. Conversely, it is clear from the way these pairs are organized here that for every possible first entry (i.e., 1, 2 and 3) the two possible second entries have been used. Thus, no element of $A{\times}B$ is missing from the right side.

(2) The standard plane from high-school analytic geometry can be identified with $\mathbb{R}{\times}\mathbb{R}$, the set of all ordered pairs of real numbers. This set is usually written as $\mathbb{R}^2$.

(3) If A and B are nonempty sets such that $A \neq B$, then clearly $A{\times}B \neq B{\times}A$. For example, suppose that a is an element of A which is not in B. Let b be an element of B. Then the ordered pair (a,b) is in the set $A{\times}B$ but not in the set $B{\times}A$.

In contrast, if $A \neq B$ but one of the sets A or B is empty, then one does have $A{\times}B\,=\,B{\times}A$, since in this case both Cartesian products are, by definition, equal to the empty set.

I.5 Functions in Mathematics

It can be argued that the single most important concept in modern mathematics is that of 'function'. Although the word 'function' has many nonmathematical meanings in ordinary English, in *This Textbook* we refer to its role as a technical mathematical term.

The inception of the mathematical usage of this word is generally attributed to Leibniz in the late seventeenth century. Over the next couple of centuries the concept of 'function' evolved greatly, often through the requirements of applications of mathematics to the sciences. The study of this evolution can be found in many books, and provides many insights into how mathematics actually develops in the real world. In the discussion below we consider only the very final stages of this evolution.

Preliminary Discussion Let us remind ourselves of the somewhat informal approach to the function concept, as is taught nowadays in many elementary math courses, and then follow the evolution of this concept to its formal definition.

Near the beginning of most modern calculus texts one finds a definition that looks very much like this:

'A *function* is a rule which assigns to each element x of a set A a definite element y of a set B. If f is such a function and x is an element of A, then we denote by $f(x)$ the element y in B which the rule f assigns to x, and we call it the *value* of the function f at the element x; that is, $y = f(x)$.'

For most purposes this definition is quite adequate. The obvious issue, of what one means by a 'rule' and by 'assigns', is handled by Justice Stewart's dictum, 'I know it when I see it'. However, a deeper analysis of the meanings of these words can get fairly complicated. For example, consider the following 'rules' for defining the values of a pair of real-valued functions f and g of a real variable:

$$f(x) = (x+1)^3 - x^3 + 3\,x^2; \quad g(x) = 1 - 3\,x \quad (*)$$

If one thinks of the word 'rule' to mean a list of instructions for what to do to the number x to obtain the corresponding value of the function, then these are clearly different 'rules'; for example, the first 'rule' requires computing the cube of x while the second does not. Nevertheless it is easy to see, using basic laws of algebra, that $f(x) = g(x)$ for every number x. In calculus one treats f and g to represent 'the same function', even though they are not given by 'the same rule'. Note that this is not an isolated phenomenon: it is clear that *every* function which one encounters in calculus can be described in multiple ways using different 'rules'.

There is a simple cure for this ambiguity, also given in the calculus texts, which comes from the concept of the 'graph' of a function. Indeed, recall from calculus that if f is a real-valued function of a real variable, then one defines the *graph* of f to be the set of points (a,b) in $\mathbb{R}\times\mathbb{R}$ such that f is defined at a, and $b = f(a)$. Note that with this graphical interpretation the phrase 'assigns to each element a of a set A a definite element b of a set B' can be clarified: it means that if (a,b_2) and (a,b_2) are both points of the graph of f, then $b_1 = b_2$. Conversely one can identify directly those subsets of $\mathbb{R}\times\mathbb{R}$ which can be viewed as graphs of functions. More precisely, suppose that G is a nonempty subset of $\mathbb{R}\times\mathbb{R}$ which passes the following **Restricted Vertical-Line Test**:

If (a,b) and (a,c) are elements of G, then $b = c$.

Speaking geometrically: every vertical line in the $x\,y$-plane which intersects the set G does so in exactly one point; hence the name 'Vertical-line Test'. The function f associated with such a set G then can be described as follows:

(i) The domain A of f consists of all numbers x in $\mathbb{R}$ such that for some real number y the ordered pair $(x,y)\in G$.

(ii) The 'rule' for f is that for each x in the domain of f, $f(x)$ is the unique number y such that $(x,y)\in G$.

Remark The word 'restricted' is used here because it applies only to real-valued functions of a single real variable.

The formal definition of 'function' which follows incorporates the major features of the preceding discussion, except that the 'rule' formulation is replaces by the 'graph' concept, and the sets A and B can be sets of objects of any type, not just real numbers.

I.5.1 Definition

(1) Let A and B be nonempty sets of objects. Then a **function with domain *A* and values in the target *B*** is a subset f of the Cartesian product $A\times B$ which satisfies the following **Extended Vertical-Line Test**:

For every object x in the set A there is exactly one object y in the set B such that the ordered pair (x,y) is an element of the set f.

One calls y the **value of *f* at *x*** and says that ***f* assumes the value *y* at *x***. The standard notation for this (unique) y is $f(x)$, pronounced (in English) 'f-of-x' ('eff-of-eks'). One also says that the function f **maps the set *A* into the set *B***.

(2) A **function in the classical sense** is a set of ordered pairs of the type described in Part (1).

(3) Suppose that f is a function with domain A. If x is a point of the domain A, then one says that ***f* is defined at *x***. Likewise, if S is a nonempty subset of A, then one says that ***f* is defined on the set *S***

(4) Let f be a function with domain A and values in a set B, as above, and let S be a nonempty subset of A.

(a) One says that the function f is **one-to-one on *S*** provided whenever x_1 and x_2 are elements of S such that $f(x_1) = f(x_2)$, then $x_1 = x_2$.

(b) Likewise, one says that ***f* maps the** S **onto the set *B*** provided that every element y in B can be expressed in the form $y = f(x)$ for some element x in S.

Note Condition (a) can be reformulated as: for every y in B there is *at most* one x in A such that $f(x) = y$. Likewise, Condition (b) can be reformulated as: for every y in B there is *at least one* x in A such that $f(x) = y$.

I.5.2 Remarks

(1) If the set f is a function in the sense of Part (2) above, then it completely determines the points of the corresponding domain A:

$$A = \{x : \text{ there exists an ordered pair in the set } f \text{ whose first entry is } x\} \qquad (*)$$

In contrast, the set f, by itself, has only partial information about the target set B. More precisely, let $B_0 = \{y : \text{ there exists an ordered pair in the set } f \text{ whose second entry is } y\}$. Then any superset of B_0 can play the role of B.

The preceding observation raises the question of whether one needs to include explicit mention of the sets A and B in the definition of 'function'. Indeed, why not simply define a 'function' to be a nonempty set of ordered pairs of objects, with no restriction on the type of objects involved? The answer is that such an approach involves the concept of the 'set of all objects', a concept which leads to serious logical difficulties; see, for example, the treatment of 'Russell's Paradox' in Appendix B.

(2) Since functions are defined above to be certain types of sets of ordered pairs, it follows from the Axiom of Extension, i.e., Principle (I.2.4), that two functions f and g are the same function if, and only if, as sets they have exactly the same elements. When combined with Equation $(*)$ in the preceding remark, one obtains the following classic formulation for the equality of two functions f and g:

(i) f and g have the same domain; and

(ii) for every x in this common domain one has $f(x) = g(x)$.

This set-theoretic formulation avoids the ambiguity of the 'A function is a Rule such that ...' formulation. Nevertheless, in specific cases we may describe the function, i.e., the set f of ordered pairs, using an explicit 'rule' showing how, for each x in the domain of f, to obtain $f(x)$ from x.

(3) Depending on the context, mathematicians often use words such as **map**, **mapping**, **operation**, **operator** and **transformation** in place of 'function'. Also, the convention in *This Textbook* is that both A and B need to be nonempty; in other words, we don't admit the concept of the 'empty function'. This restriction is for convenience; but it should be noted that in certain other parts of mathematics the empty function *is* allowed.

(4) The statement 'f is defined on the set S', in Part (2) of the preceding definition, allows the possibility that f is also defined at points not in S; that is, it allows S to be a proper subset of the domain A of f. In contrast, some authors define this statement to mean that S is the (full) domain of f, so that if $x \notin S$, then $f(x)$ makes no sense. The situation is actually a bit more complicated: some of the authors who require that S be the (full) domain do, on occasion, and without stating that they are deviating from their usual convention, allow S to be a proper subset of that domain.

Many authors prefer an approach to 'functions' in terms of ordered pairs, as above, but in which the target B is also specified uniquely, not just the domain A; indeed, in certain parts of mathematics, such as Algebraic Topology, fixing the target of a function is of great importance. One obvious approach would be to always use the set B_0 described in Remark (1) above. However, that choice turns out to be much too restrictive, especially when considering simultaneously more than one function with the same domain.

The following well-known device allows one to include references to a specific target B without conflicting with Definition (I.5.1).

The Arrow Notation for Functions Assume that f is a function which is defined on a nonempty set S and has values in a set B, as described in Definition (I.5.1) above. (In particular, recall that 'is defined on S' means that the nonempty set S is a subset of the full domain A of f.) One often abbreviates this assumption in symbols using the following **arrow notation**:

$$f : S \to B;$$

in words: 'f maps the set S into B'.

The arrow notation provides a simple way to provide a more modern definition of 'function' in which the target set is completely specified.

I.5.3 Definition

(1) An expression of the form $f : S \to B$, where S, B and f are as above, is called a **function diagram**. The condition for two such diagrams, $f : S \to B$ and $g : T \to C$, to be equal as diagrams is that $S = T$, $B = C$, and $f(x) = g(x)$ for each x in S, where each equation is in the sense of Principle (I.2.4), the Axiom of Extension.

(2) A **function in the modern sense** is a function diagram $f : A \to B$, as in Part (1) above, in which the set A is the domain of f. The set B in this diagram is then called the **codomain** of the given function.

I.5.4 Remarks

(1) Many texts define 'function in the modern sense' to be an ordered pair (U, f) in which U is an ordered pair (A, B) of nonempty sets, and f is a subset of $A \times B$ which satisfies the Extended Vertical-Line Test. This approach is obviously equivalent to the 'function diagram' approach just described.

(2) If the given context makes clear which function diagram is under consideration, we shall often abbreviate 'the function diagram $f : A \to B$' to 'the function $f : A \to B$', or even more briefly to 'the function f'. In such a context it still makes sense to refer to the set B as 'the codomain of the function f', since the underlying function diagram $f : A \to B$ is understood.

(3) Similarly, the context should make it clear whether the word 'function' is being used in the 'classical sense' of Definition (I.5.1) or in the 'modern sense' of Definition (I.5.3). Note that some authors use the 'classical' definition, some the 'modern' definition, so it is useful to be familiar with both formulations. However, since most authors focus on only one of these formulations, they usually do not include the adjectives 'classical' or 'modern' as we do here.

(4) On occasion it is convenient to use the notation $S \xrightarrow{f} B$ as a substitute for the standard $f : S \to B$; for example, see the treatment of 'composition' below, in which one strings together a pair of function diagrams.

The 'function diagram' concept can be used to clarify some other terminology.

I.5.5 Definition

A function diagram $f : S \to B$ is said to be an **injection** provided the function f is one-to-one on the set S, in the sense of Part (4 a) of Definition (I.5.1).

Likewise, the diagram is said to be a **surjection** provided f maps the set S onto the set B, in the sense of Part (4 b) of the same definition.

A function diagram which is both an injection and a surjection is called a **bijection**. If $f : A \to B$ is a bijection, then one says that elements a in A and b in B **correspond under *f*** provided $b = f(a)$.

I.5.6 Some Examples of Functions

(1) The standard algebraic functions and transcendental functions used in calculus should be familiar. The 'algebraic' functions include polynomial functions and rational functions (i.e., ratios of polynomial functions). The 'transcendental' functions include the standard exponential, logarithmic and trigonometric functions.

(2) Less well known, perhaps, but even simpler, are the following real-valued functions, both with domain $A = \mathbb{R}$, named after mathematicians of the nineteenth century:

(i) The **Dirichlet Function** is the function $F_{\text{Diri}} : \mathbb{R} \to \mathbb{R}$ given by the rule

$$F_{\text{Diri}}(x) = \begin{cases} 1 & \text{if } x \text{ is a rational number} \\ 0 & \text{if } x \text{ is an irrational number} \end{cases}$$

(ii) The **Thomae Function** is the function $F_{\text{Thom}} : \mathbb{R} \to \mathbb{R}$ given by the rule

$$F_{\text{Thom}}(x) = \begin{cases} 1 & \text{if } x = 0 \\ 1/q & \text{if } x \text{ is a nonzero rational number } p/q, \text{ with } p \text{ in } \mathbb{Z} \text{ and } q \text{ in } \mathbb{N} \text{ being in lowest terms} \\ 0 & \text{if } x \text{ is an irrational number} \end{cases}$$

These function are used mainly as 'exotic examples' in various contexts.

(3) Let $f : A \to B$ be a function diagram. One says that ***f* is constant on *A*** if there exists an element c of B such that $f(x) = c$ for every x in A. If, in addition, A is the full domain of f, then one says that ***f* is the constant function on *A* with value *c***.

(4) Let A be a nonempty set, and let $f : A \to A$ be the function, with domain and codomain both equal to A, given by the rule $f(x) = x$ for all $x \in A$. (Equivalently, f is the subset of $A \times A$ consisting of all the ordered pairs of the form (x, x) with $x \in A$.) This function is called **the identity function on** A, and is usually denoted by the symbol I_A; if the set A remains fixed throughout a discussion, the notation may be simplified to I. It is clear that the map $I_A : A \to A$ is a bijection, in the sense of Definition (I.5.5).

Similarly, let A be a nonempty set, and let B be a superset of A; that is, $A \subseteq B$. Define the function diagram $\iota_{A;B} : A \to B$, with domain A and codomain B, by the rule

$$\iota_{A;B}(x) \;=\; x \text{ for all } x \text{ in } A.$$

This function is called the **inclusion of A into B**. (The symbol 'ι' is the lower-case Greek letter 'iota', which corresponds to the English letter i.)

Note that the domain of the function $\iota_{A;B}$ is the same as the domain of the identity function I_A considered above. In addition, $I_A \subseteq A \times A$ and $\iota_{A;B} \subseteq A \times B$ are the same sets of ordered pairs. Thus under the 'classical' formulation of 'function', one has $I_A \;=\; \iota_{A;B}$. However, under the modern 'function-diagram' formulation, they are different functions if A is a proper subset of B.

(5) Let A be a nonempty set, and let S be a subset of A, possibly empty. Then the **characteristic function associated with the subset S of A** is the function $\chi_{A;S} : A \to \{0,1\}$, with domain A and with values in the doubleton set $\{0,1\}$, given by the rule

$$\chi_{A;S}(x) \;=\; \begin{cases} 1 & \text{if } x \text{ in } A \text{ is an element of } S \\ 0 & \text{if } x \text{ in } A \text{ is } \textit{not} \text{ an element of } S \end{cases}$$

Remarks (i) The symbol 'χ' is the lower-case version of the ancient Greek letter 'chi', usually pronounced 'kai', and which corresponds (roughly) to the English 'k' sound.

(ii) If, as is often the case, the context makes clear which domain A is under consideration, it is customary to omit reference to it and write simply χ_S; indeed, some texts use this concept only when $A \;=\; \mathbb{R}$.

(iii) The phrase 'characteristic function' is used in certain other branches of mathematics with an unrelated meaning. Because of this, some authors use the phrase **indicator function** instead of 'characteristic function' for the concept defined here; and some use notations such as 1_S instead of χ_S.

(6) Let $P \;=\; (a,c)$ and $Q \;=\; (b,d)$ be points in $\mathbb{R} \times \mathbb{R}$, with $a \;<\; b$. Define $g : [a,b] \to \mathbb{R}$ to be the real-valued function, with domain the closed interval $[a,b]$, given by the rule

$$g(x) \;=\; c + \left(\frac{d-c}{b-a}\right)(x-a) \text{ for } a \leq x \leq b$$

In terms of the 'ordered pairs' definition of functions, g is the set of all ordered pairs (x,y) in $\mathbb{R} \times \mathbb{R}$ of the form $\left(x, c + \left(\frac{d-c}{b-a}\right)(x-a)\right)$ with $a \leq x \leq c$. Speaking geometrically, g is a function whose graph in the (x,y)-plane is the straight line segment joining the points P and Q. The function g described here is called the **linear interpolation between P and Q**.

(7) In elementary calculus one is often taught about **step functions**. The most common example of such a function given in calculus texts is the so-called **postage-stamp function**. For example, in the year 2013 the cost of mailing a first-class letter within the United States

is given by the following table:

Weight of Letter (in ounces)	Mailing Cost (in USD)
$0.0 < w \le 1.0$	\$0.46
$1.0 < w \le 2.0$	\$0.66
$2.0 < w \le 3.0$	\$0.86
$3.0 < w \le 3.5$	\$1.00

(Anything heavier falls into a different category of mail.)

Associated with this table is the function $f : (0, 3.5] \to \mathbb{R}$ given by the rule

$$f(x) = \begin{cases} 0.46 & \text{if } 0.0 < x \le 1.0 \\ 0.66 & \text{if } 1.0 < x \le 2.0 \\ 0.86 & \text{if } 2.0 < x \le 3.0 \\ 1.00 & \text{if } 3.0 < x \le 3.5 \end{cases}$$

The following terminology and notation is used throughout modern mathematics.

I.5.7 Definition

Suppose that the function $f : A \to B$ is a bijection. Then the corresponding **inverse function** $f^{-1} : B \to A$ is given by the following rule:

If $y \in B$, then $f^{-1}(y)$ is the unique element x in A such that $f(x) = y$.

Example Let A be the set of all real numbers $x \neq 1$, and let B be the set of all real numbers $y \neq 2$. Define the function f with domain A by the rule $f(x) = \dfrac{2x-1}{x-1}$ for all x in A. It is easy to see that f maps A onto B, and that the function diagram $f : A \to B$ is a bijection. Indeed, if y is a real number of the form $(2x-1)/(x-1)$, then

$$y\,(x-1) = 2x-1, \text{ hence } (y-2)\,x = y-1, \text{ and thus } x = \frac{y-1}{y-2}.$$

This implies that if $x \neq 1$ then $y = f(x) \neq 2$, so f maps A into B. Further, the calculation implies that if $y \neq 2$ then there is a unique x in A for which $y = f(x)$, so that the function f is one-to-one. Thus, f maps A onto B, so that $f : A \to B$ is a bijection. Finally, the calulation implies that the inverse $f^{-1} : B \to A$ is given by the formula $f^{-1}(y) = \dfrac{y-1}{y-2}$.

I.5.8 Theorem

Suppose that $f : A \to B$ is a bijection, and let $f^{-1} : B \to A$ be the corresponding inverse function. Then:

(a) The set $f^{-1} \subseteq B \times A$ is given by the rule

$$f^{-1} = \text{ the set of all ordered pairs } (b,a) \text{ such that } (a,b) \in f.$$

(b) The function diagram $f^{-1} : B \to A$ is also a bijection.

(c) The inverse of the bijection $f^{-1} : B \to A$ is the original bijection $f : A \to B$. In symbols: $(f^{-1})^{-1} = f$.

The simple proof is left as an exercise.

Associated with a function f from A to B are various subsets of A and B, along with corresponding notation and terminology.

I.5.9 Definition

Throughout this definition $f : A \to B$ is a function with domain A and codomain B.

(1) Let S be a subset of A. Then one associates, with the function f and set S, a subset of B, denoted $f[S]$, given by the rule

$$f[S] = \{f(x) : x \in S\}.$$

In words: the set $f[S]$ consists precisely of those elements y in B such that $y = f(x)$ for at least one element of S. One calls the set $f[S]$ the **image of the set S under the map f**. Note that if $S = \emptyset$ then $f[S] = \emptyset$.

The special set $f[A]$, i.e., the image of the (full) domain of f under f, is called simply the **image of the function f**; the elements of this set are precisely the values of the function f.

(3) Similarly, let U be a subset of B. Then one associates with f and U a subset of A, denoted by $f^{-1}[U]$ and given by the rule

$$f^{-1}[U] = \{x \in A : f(x) \in U\}.$$

In words: $f^{-1}[U]$ is the set of all x in A such that $f(x) \in U$. The set $f^{-1}[U]$ is called the **inverse image of U under f**; the word **preimage** is often used in place of the phrase 'inverse image'. Note, in particular, that $f^{-1}[\emptyset] = \emptyset$.

Side Comments (on function notation and terminology):

(1) Many texts write $f(S)$ instead of the notation $f[S]$ used above; that is, they surround the symbol S with parentheses, (and), instead of with brackets, [and]. Likewise, they write $f^{-1}(U)$ instead of the notation $f^{-1}[U]$ used above.

The most obvious problem with the 'parentheses' notation $f(S)$ is that it can conflict with the notation $f(x)$, used for the value of the function at a point x of the domain of f. Normally this 'abuse of notation' does not cause any confusion; but there are situations in which it could. For instance, let A be the doubleton set $\{1, \{1\}\}$ and let B be the singleton set $\{1\}$. Define $f : A \to B$ by the rule $f(1) = 1$, $f(\{1\}) = 1$. Note that the set $S = \{1\}$ is simultaneously an *element* of A and a *subset* of A. Viewing S as an *element* of A, one has, using the normal '$f(x)$' notation, $f(S) = 1$; viewing S as a *subset* of A, one has, using the bracket notation, $f[S] = \{\{f(1)\}\} = \{1\}$. However, using parentheses for both situations leads to the confusion of writing simultaneously $f(S) = 1$, when S is thought of as an element of A, and $f(S) = \{1\}$, when S is thought of as a subset of A.

(2) The notation for the inverse image has, under certain circumstances, a second ambiguity. Indeed, if $f : A \to B$ is a bijection, then the symbol f^{-1} is used to denote the inverse function associated with f; see Definition (I.5.7). In this case the notation $f^{-1}[U]$ makes sense for every subset U of B as the image under the function f^{-1} of the subset U of B. Fortunately, this set happens to be the same as the inverse image, in the sense of Part (2) of the preceding definition, of the set U under the original map f.

(3) Some authors use the terminology 'range of f' instead of the phrase 'image of f' used above. In contrast, some other authors use the word 'range' for what we refer to as 'target'. For that reason, we do not use 'range' in either sense in *This Textbook*.

I.6 New Functions from Old Functions

There are many ways of defining new functions in terms of other functions. One of the simplest is given next.

I.6.1 Definition

(1) Suppose that f is a function with domain A and values in a set B, as described in Definition (I.5.1) above. If S is a *nonempty* subset of A, then one associates with f and S a new function g, with domain S and values in B, called the **restriction of f to the subset** S. It is given by the rule

$$g(x) = f(x) \text{ for all } x \in S.$$

In this context one also refers to the function f as **an extension** of g to the set A.

(2) The standard notation for the restriction of the function f to the subset S is $f|_S$.

Remark We use the phrase '*an* extension of g to the set A' above because usually g can be viewed as the restriction to S of a function with domain A in more than one way. The exceptions are when $S = A$ or B is a singleton set.

The next construction appears frequently in elementary calculus, although usually in not so general form.

I.6.2 Definition (Composition of Functions)

Suppose that $f : A \to B$ and $g : C \to D$ are function diagrams such that $f[A] \subseteq C$; that is, for each x in A one has $f(x) \in C$ and thus in the domain of g. Then the **composition of g with f** is the function $g \circ h : A \to D$ gien by the rule

$$(g \circ h)(x) = g(f(x)) \text{ for all } x \in A \quad (*)$$

In symbols: $h = g \circ f$; the symbol $\circ$ is often pronounced 'of', or, sometimes, 'circle' in this context.

I.6.3 Remarks

(1) The notation '$\circ$' for the composition $g{\circ}f$ is used to make one think of a kind of 'multiplication' of the function g with the function f. In this case one refers to g as the 'left factor' and f as the 'right factor' in the expression $g{\circ}f$. In contrast to the 'product × used in the Cartesuan productThis 'multiplication'

The order in which the factors g and f appear in the expression $g{\circ}f$ is very important. Indeed, it is possible for the composition $g{\circ}f$ to be defined, while the composition $f{\circ}g$ is not. Even if $g{\circ}f$ and $f{\circ}g$ both make sense, they need not be equal. (That is, the operation of composition need not satisfy the 'Commutative Law'.) See the examples below for illustrations of these facts.

(2) Some texts - especially those used in elementary calculus – allow a slightly more general definition of 'composition'. The key is the defining equation $(g{\circ}f)(x) = g(f(x))$: these texts allow the domain of $g{\circ}f$ to be the set of all x for which this expression makes sense (provided this set is nonempty). More precisely:

Alternate Definition of Composition: Suppose that $f : A \rightarrow B$ and $g : C \rightarrow D$ are function diagrams. Let $S = \{x{\in}A : f(x){\in}C\}$; that is, $S = f^{-1}[B \cap C]$. If the set S is nonempty, then the composition of g with f is the function $p : S \rightarrow D$ given by $p(x) = g(f(x))$ for all $x{\in}S$.

In *This Textbook* we follow the original version given above (Definition (I.6.2)). However, most of the results concerning composition can be easily modified to work just as well with this alternate definition.

I.6.4 Examples

(1) Let $A = \{1,2,3\}$, $B = \{4,5,6\}$ and $C = \{40,50,60\}$. Define $f : A \rightarrow B$ by the rule $f(x) = x+3$ for $x{\in}A$, and define $g : B \rightarrow C$ by the rule $g(y) = 10y$ for $y{\in}B$. It is easy to see that the composition $g{\circ}f : A \rightarrow C$ is defined and is given by the rule

$$(g{\circ}f)(1) = g(1+3) = g(4) = 40; \quad (g{\circ}f)(2) = g(2+3) = g(5) = 50;$$

$$(g{\circ}f)(3) = g(3+3) = g(6) = 60$$

In contrast, the expression $f{\circ}g$ does not make sense here. (Do you see why?)

(2) Let $A = \{1,2,3\}$ as before, and let $f : A \rightarrow A$ and $g : A \rightarrow A$ be given by

$$f(1) = 2, \quad f(2) = 3, \quad f(3) = 1 \quad \text{and} \quad g(1) = 1, \quad g(2) = 3, \quad g(3) = 2.$$

Clearly $g{\circ}f : A \rightarrow A$ and $f{\circ}g : A \rightarrow A$ are both defined and have the same domain, namely A (because $A = B = C$). However, one readily computes that

$$(g{\circ}f)(2) = g(f(2)) = g(3) = 2, \text{ while } (f{\circ}g)(2) = f(g(2)) = f(3) = 1.$$

Since $g{\circ}f$ and $f{\circ}g$ do not have the same value at some point of their common domain, they cannot be the same function.

(3) The preceding example illustrates the following fact:

$$\text{the equation } g \circ f = f \circ g \text{ is not always correct.} \qquad (*)$$

Warning Some students go on to (mis)interpret Statement (∗) as meaning

$$\text{the equation } g \circ f = f \circ g \text{ is always not correct.} \qquad (**)$$

- note how the word 'always' has subtly shifted to the left of the 'not' in the second version. That is, they interpret the original statement as meaning that 'for all f and g, the function $g \circ f$ is *never* equal to $f \circ g$. This second interpretation is *not* what Statement (∗) says; indeed, the phrase 'not always correct' allows the possibility that the equation is 'sometimes correct, sometimes incorrect'. This second interpretation, Statement (∗∗), is wrong. For instance, let $A = \{1, 2, 3\}$ as above, but now let $f : A \to A$ and $g : A \to A$ be given by

$$f(1) = 2, \quad f(2) = 3, \quad f(3) = 1 \text{ and } g(1) = 3, \quad g(2) = 1, \quad g(3) = 2.$$

The reader is invited to verify that $g \circ f = f \circ g = I_A$ in this case.

Side Comment (on the left-to-right bias): This is a good place to discuss a phenomenon which pervades mathematics, but is rarely mentioned; namely, a bias in favor of reading mathematical expressions from left to right.

Examples

(1) Consider the statements

(a) 'George Washington was the first president of the United States.'

(b) 'The first president of the United States was George Washington.'

Superficially, it appears that these statements contain exactly the same information and thus are interchangable; and in a sense this is correct. However, there is a subtle difference: Statement (a) is 'about' George Washington, because 'George Washington' comes first, since English is read left-to-right; in contrast, Statement (b) is 'about' the office of the presidency. For example, Statement (a) would be the better response to the question 'Who was George Washington?', while Statement (b) would be the better response to the (very different) question 'Who was the first president of the United States?'

(2) Similarly, consider the inequalities $\sqrt{2} < 2$ and $2 > \sqrt{2}$. Technically speaking, they contain exactly the same information about the relation between the numbers 2 and $\sqrt{2}$. Psychologically speaking, however, there is a subtle difference in focus: because of the left-to-right nature of written English, the first inequality is a statement 'about' the number $\sqrt{2}$ – written as an English sentence, it says 'the square root of 2 is less than 2'; while the second is a statement 'about' the number 2, namely '2 is greater than $\sqrt{2}$'. It is possible that mathematicians felt the need to introduce *two* symbols for essentially the same idea, namely $<$ and $>$, in order to allow symbolically for this subtle difference of focus. Similar remarks can be made concerning the 'duplicative' notations $\subseteq$ and $\supseteq$.

(3) Sometimes mathematicians need to be a little devious to write down exactly what they mean. For instance, the (compound) inequality $1 < \sqrt{2} < 2$ has the same mathematical content as the sentence '$\sqrt{2}$ is between 1 and 2'; but the sentence form makes it clearer that the object of interest is the square root. In contrast, the symbolic form is really an abbreviation of the pair of inequalities $1 < \sqrt{2}$ and $\sqrt{2} < 2$. The first 'about' the number 1, because we read '1' first; likewise, the second is 'about' the number $\sqrt{2}$. Some mathematicians get around this, while still using symbolism, by writing $\sqrt{2} \in (1, 2)$, to emphasize the fact that

desired (compound) statement is ‘about’ the number $\sqrt{2}$. Of course this formulation requires the reader to decode the ‘member of’ symbol $\in$ and the ‘open interval’ notation $(1,2)$.

Note that the ambiguity in the ‘focus’ illustrated here does not disappear by simply replacing $<$ with $>$.

(4) The ‘left-to-right’ bias appears in equations. For instance, the so-called Quadratic Formula, from elementary algebra, for the solutions of equations of the form $ax^2+bx+c = 0$, is always written

$$x = \frac{-b \pm \sqrt{b^2-4a\,c}}{2\,a},$$

never

$$\frac{-b \pm \sqrt{b^2-4a\,c}}{2\,a} = x,$$

even though the equations have the same content. The first equation tells us, because of the left-to-right bias, that the equation is ‘about’ the unknown x.

(5) Sometimes the ‘left-to-right’ bias causes true difficulties. For instance, in multivariable calculus one encounters expressions such as $\frac{\partial^2 z}{\partial y \partial x}$. If one asks students to carry out the partial derivatives ‘in the order indicated’, many will do the y-derivative first, because it appears on the left. Of course, the notation really means that the x-differentiation should be carried out first. It is fortunate for such students that normally ‘mixed partials are equal’.

Incidently, notice that the widely-used ‘subscript’ notation for partial derivatives is in accordance with the ‘left-to-right’ bias, and thus does not share this problem:

$$\frac{\partial^2 z}{\partial y \partial x} = z_{xy}$$

(6) A similar problem occurs with the definition of ‘composition’. Thus, consider functions $A \xrightarrow{f} B$ and $B \xrightarrow{g} C$. The use of the ‘right arrow’ $\rightarrow$ here reflects the ‘left-to-right’ bias; indeed, one rarely encounters the notation $B \xleftarrow{f} A$. It becomes even clearer if one writes the corresponding function diagram

$$A \xrightarrow{f} B \xrightarrow{g} C$$

Many students look at this and then write the corresponding composition as $f \circ g$, to match the ‘left-to-right’ order seen above, instead of the (correct) order $g \circ f$. Unlike the ‘mixed partials’ situation described in the preceding example, however, the error here is much more serious, since usually it is *not* the case that $f \circ g = g \circ f$.

I.6.5 Theorem (Basic Facts about Composition)

(a) Let $f : A \rightarrow B$ be a map with domain A. Then the compositions $f \circ I_A$ and $I_B \circ f$ are both defined, and

$$f \circ I_A = I_B \circ f = f.$$

(The symbols I_A and I_B denote the identity functions on the sets A and B, respectively.)

(b) Let $f : A \rightarrow B$, $g : B \rightarrow C$, and $h : C \rightarrow D$ be maps with domains A, B and C, respectively. Then the compositions $(h \circ g) \circ f : A \rightarrow D$ and $h \circ (g \circ f) : A \rightarrow D$ are both defined, and they are equal:

$$(h \circ g) \circ f = h \circ (g \circ f).$$

Stated briefly: 'The Associative Law for Composition is valid'.

(c) Suppose that $f : A \to B$ is a bijection, and let $f^{-1} : B \to A$ be the inverse map; see Definition (I.5.7). Then $f^{-1} \circ f = I_A$ and $f \circ f^{-1} = I_B$.

(d) A necessary and sufficient condition for a map $f : A \to B$ to be a bijection from A onto B is that there exist maps $g : B \to A$ and $h : B \to A$ such that

$$g \circ f = I_A \text{ and } f \circ h = I_B.$$

If such maps g and h exist, then $g = h = f^{-1}$.

(e) If $f : A \to B$ and $g : B \to C$ are bijections from A onto B and from B onto C, respectively, then their composition $g \circ f : A \to C$ is a bijection from A onto C. Furthermore, the inverse map $(g \circ f)^{-1} : C \to A$ is given by the formula

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

Partial Proof:

(a) The reader should be able to do this.

(b) The fact that both compositions are defined, and both have domain A and have values in C, follows directly from the definition of 'composition'. To see that the equation is satisfied, note first that if x is in A then

$$((h \circ g) \circ f)(x) \overset{(1)}{=} (h \circ g)(f(x)) \overset{(2)}{=} h(g(f(x))) \overset{(3)}{=} h((g \circ f)(x)) \overset{(4)}{=} (h \circ (g \circ f))(x).$$

Indeed, Equation (1) follows from the definition of the composition $p \circ f$, where $p = h \circ g$. Likewise, Equation (2) reflects the definition of $h \circ g$, Equation (3) uses the definition of $g \circ f$, and Equation (4) comes from the definition of $h \circ q$ where $q = g \circ f$. Thus the maps $(h \circ g) \circ f$ and $h \circ (g \circ f)$ not only have the same domain, but they also assume the same value as each other for each point of that domain. In other words, they are the same map, as claimed.

(c) The simple proof is left as an exercise.

(d) The 'Sufficient' Part of (d): Suppose that maps g and h with the indicated properties exist.

(i) Since $g \circ f = I_A$, it follows that f is one-to-one on A. Indeed, suppose that x_1 and x_2 are points of A such that $f(x_1) = f(x_2)$. Apply the function g to each side of this last equation to get $g(f(x_1)) = g(f(x_2))$. From the definitions of 'composition' and 'identity map', plus the hypothesis $g \circ f = I_A$, one then sees that

$$x_1 = I_A(x_1) = (g \circ f)(x_1) = g(f(x_1)) = g(f(x_2)) = (g \circ f)(x_2) = I_A(x_2) = x_2.$$

That is, if $f(x_1) = f(x_2)$ then $x_1 = x_2$, which means that f is one-to-one, as claimed.

(ii) Since $f \circ h = I_B$ it follows that f maps A *onto* B. Indeed, let y be any element of B, and note that

$$y \overset{(1)}{=} I_B(y) \overset{(2)}{=} (f \circ h)(y) \overset{(3)}{=} f(h(y)).$$

Equation (1) simply repeats the definition of the identity map I_B; Equation (2) restates the hypothesis $f \circ h = I_B$; Equation (3) uses the definition of 'composition'. In any event, one now has $y = f(x)$ for at least one $x \in A$, namely $x = h(y)$. It follows that f is a surjection of A onto B.

Since, as has just been shown, f is maps A one-to-one onto B, it follows that f is a bijection, as claimed. To see that $g = h = f^{-1}$, note first that the equation $g \circ f = I_A$ implies

$$f^{-1} \overset{(1)}{=} I_A \circ f^{-1} \overset{(2)}{=} (g \circ f) \circ f^{-1} \overset{(3)}{=} g \circ (f \circ f^{-1}) \overset{(4)}{=} g \circ I_B \overset{(5)}{=} g,$$

so that $g = f^{-1}$, as claimed. Indeed, Equation (1) follows from Part (a) of this theorem, Equation (2) uses the hypothesis about g, Equation (3) uses the 'Associative Law for Composition', Equation (4) uses the results of Part (c) of this theorem, and Equation (5) uses Part (a) of this theorem again.

A similar argument shows that $h = f^{-1}$.

The 'Necessary' Part of (d): If f is a bijection, simply set $g = h = f^{-1}$. The fact that g and h have the desired properties follows easily from the results of Part (c) of this theorem.

(e) This result can be easily proved directly from the definitions of the various concepts which appear in it; the reader is encouraged to carry out such a proof.

It frequently happens that the most convenient way to describe a function is in terms of its restrictions (in the sense of Definition (I.6.1)) to a suitable family of subsets of its domain. The next result makes this precise.

I.6.6 Theorem (The Union-of-Functions Theorem)

Let A and B be nonempty sets, and let $\mathcal{S}$ be a family of nonempty subsets of A whose union is A. For each set S in the family $\mathcal{S}$ let $f_S : S \to B$ be a function with domain S and with values in B. Let $\mathcal{F} = \{f_S : S \in \mathcal{S}\}$ be the corresponding family of functions. Then a necessary and sufficient condition for there to exist a function $g : A \to B$ such that for each S in $\mathcal{S}$ one has $f_S = g|_S$ is the following **consistency condition**:

If S_1 and S_2 are elements of $\mathcal{S}$ such that $S_1 \neq S_2$ and $S_1 \cap S_2 \neq \emptyset$, then $f_{S_1}(x) = f_{S_2}(x)$ for all x in $S_1 \cap S_2$.

When this consistency condition is satisfied, the resulting function g is unique; more precisely, g is the union of the family $\mathcal{F}$. (Recall that, by definition, g is a subset of $A \times B$ while each f_S is a subset of $S \times B \subseteq A \times B$, so this union makes sense.)

The simple proof is left as an exercise.

I.6.7 Definition

The function g obtained above is called the **union of the functions in the family $\mathcal{F}$**.

Remark The name 'Union-of-Functions Theorem' for this result is not standard.

I.6.8 Examples of Functions Described as Unions

(1) The function $g : \mathbb{R} \to \mathbb{R}$ given by the rule

$$g(x) = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x \leq 0 \end{cases}$$

is called the **Absolute-Value Function**; the corresponding quantity $g(x)$ is usually denoted by the symbol $|x|$, although many computer languages and programs (e.g., spreadsheets) use the notation abs$\,(x)$ instead.

Note The construction of this function is of the type described in the preceding theorem. Indeed, let $S_1 = \{x{\in}\mathbb{R} : x \geq 0\}$ and $S_2 = \{x{\in}\mathbb{R} : x \leq 0\}$, and define $f_{S_1} : S_1 \to \mathbb{R}$ and $f_{S_2} : S_2 \to \mathbb{R}$ by the rules

$$f_{S_1}(x) = x \text{ for all } x{\in}S_1, \quad f_{S_2}(x) = -x \text{ for all } x{\in}S_2$$

Note that $S_1 \cap S_2 = \{0\}$ and that $f_{S_1}(0) = f_{S_2}(0) = 0$, so the preceding result applies with $\mathcal{S} = \{S_1, S_2\}$ and $\mathcal{F} = \{f_{S_1}, f_{S_2}\}$ to get the function $g = $ abs.

Of course, this is not the only way to express the function g as the union of simpler functions. For instance one could chose instead $S_1 = \{x{\in}\mathbb{R} : x > 0, S_2 = \{x{\in}\mathbb{R} : x < 0\}$ and $S_3 = \{0\}$; now define $f_{S_j} : S_j \to \mathbb{R}$, for $j = 1, 2, 3$, by

$$f_{S_1}(x) = x \text{ for } x{\in}S_1, \quad f_{S_2}(x) = -x \text{ for } x{\in}S_2, \quad f_{S_3}(0) = 0.$$

The corresponding family $\mathcal{S} = \{S_1, S_2, S_3\}$ automatically satisfies the consistency condition since, in this case $S_i \cap S_j \neq \emptyset$ only when $i = j$.

Remark It would have been slightly simpler to use the notation f_j instead of f_{S_j}. Henceforth we shall normally make such simplifications when they cannot cause confusion.

(2) The Dirichlet function F_{Diri} (see Example (I.5.6) (3) above) is the union of two constant functions. More precisely, let S_1 be the set of all rational numbers and let S_2 be the set of all irrational numbers. Define $f_1 : S_1 \to \mathbb{R}$ to be the constant function on S_1 with value 1, and let $f_2 : S_2 \to \mathbb{R}$ be the constant function on S_2 of value 0. Then F_{Diri} is the union (in the sense of Definition (I.6.7)) of f_1 and f_2. (Note that $S_1 \cap S_2 = \emptyset$, so the consistency condition is automatically satisfied.)

(3) Let $I = [a, b]$ be a closed interval in $\mathbb{R}$. For some m in $\mathbb{N}$ let $P_0 = (a_0, b_0)$, $P_1 = (a_1, b_1), \ldots P_{m-1} = (a_{m-1}, b_{m-1})$, $P_m = (a_m, b_m)$ be points in $\mathbb{R}{\times}\mathbb{R}$ such that $a = a_0 < a_1 < a_2 < \ldots < a_m = b$. For each $j = 0, \ldots, m-1$ let $S_j = [a_j, a_{j+1}]$. Now define $g_j : S_j \to \mathbb{R}$ to be the linear interpolation between P_j and P_{j+1} (see Example (I.5.6) (4)). It is easy to verify that the family $\mathcal{G} = \{g_1, g_2, \ldots g_k\}$ satisfies the consistency condition, so that the union of these functions is also a function; call it g. The function g is called the **piecewise-linear interpolation through the points $\boldsymbol{P}_0$, $\boldsymbol{P}_1, \ldots$ $\boldsymbol{P}_m$**.

Special Case: Suppose that $f : [a, b] \to \mathbb{R}$ is a real-valued function with domain $[a, b]$, and that the numbers b_0, b_1m$\ldots$ b_m above satisfy $b_j = f(a_j)$ for each $j = 0, 1, \ldots m$; that is, $P_j = (a_j, f(a_j))$. Then one calls the corresponding piecewise-linear function g described as above a **piecewise-linear interpolant of the function $\boldsymbol{f}$**.

The $m+1$ points a_0, $a_1, \dots P_m$ described above the **nodes** associated with this interpolation construction. These nodes are said to be **equally spaced** provided the differences $a_1 - a_0$, $a_2 - a_1, \dots a_m - a_{m-1}$ are equal; more precisely, provided $a_{j+1} - a_j = (b-a)/m$ for each $j = 0, 1, \dots m-1$.

Remarks (1) Suppose that $f : [a,b] \to \mathbb{R}$ is a real-valued function with domain $[a,b]$. Then it is an easy exercise to show that for each m in $\mathbb{N}$ there is exactly one piecewise-linear interpolant of f on $[a,b]$ having exactly $m+1$ equally spaced nodes.

We finish this section by discussing briefly the functions of main interest in real analysis, namely functions **real-valued functions**; that is, functions $f : X \to \mathbb{R}$ from some domain X (which need not be part of $\mathbb{R}$) with values in $\mathbb{R}$. The special nature of the target $\mathbb{R}$ allows these functions to be combined algebraically to form new functions.

I.6.9 Definition

Let $f : X_1 \to \mathbb{R}$ and $g : X_2 \to \mathbb{R}$ be real-valued functions whose domains are the sets X_1 and X_2, respectively. Let $X_3 = X_1 \cap X_2$, and assume that $X_3 \neq \emptyset$.

(1) The **sum of *f* and *g*** is the function $f+g : X_3 \to \mathbb{R}$ given by the rule

$$(f+g)(x) = f(x) + g(x) \text{ for all } x \text{ in } X_3.$$

Similarly, the **difference of *f* and *g*** is the function $f - g : X_3 \to \mathbb{R}$ given by the rule

$$(f-g)(x) = f(x) - g(x) \text{ for all } x \text{ in } X_3.$$

(2) The **product of *f* and** g is the function $f \cdot g : X_3 \to \mathbb{R}$ given by the rule

$$(f \cdot g)(x) = (f(x)) \cdot (g(x)) \text{ for all } x \text{ in } X_3$$

(3) Let X_4 be the set of all points x in X_3 at which $g(x) \neq 0$, and assume $X_4 \neq \emptyset$. Then the **quotient of *f* by *g*** is the function $f/g : X_4 \to \mathbb{R}$ given by the rule

$$\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)} \text{ for all } x \text{ in } X_4.$$

I.7 The Cardinality of Sets

One of the major advances made by Cantor in his theory of sets is his analysis of the concept of the 'number of elements of an infinite set'. The special case of a set have equally many elements as the set $\mathbb{N}$ of natural numbers plays a particularly important role in analysis, so we focus on that case in the main body of *This Textbook*. In Appendix B we consider Cantor's theory in more generality.

Side Comment (on sets having equally many elements):

One aspect of Cantor's theory is to establish under what circumstances two sets A and B have equally many elements. The solution to this problem seems very easy if the sets A and B are both finite: simply count the number of elements in each set, and if the numbers agree, then the sets have equally many elements. However, in modern mathematics one is forced to study sets which have *infinitely many* elements, and in such cases 'counting' may not even make sense. Fortunately, there is a way of determining when two sets have equally many elements which, even in the case of finite sets, is older than the idea of 'counting', and this older method is what Cantor uses.

Example Anthropologists tell us that there were primitive societies whose concept of 'number' was restricted to 'one', 'two' and 'many' (or, perhaps, 'several'). Suppose that the king of such a society needs to know whether there are as many spears available in storage as there are warriors; could he do so? The answer is 'Yes': he could tell the warriors to each take one spear from storage; if, after this, every warrior has one spear, and there are no spears left over, then the king knows that there are equally many spears as warriors. Note that this process does not involve 'counting' the number of spears and the number of warriors; indeed this king could not count numbers that high. It is simply a matter of 'pairing off' spears and warriors, so that each spear is carried by one warrior, and each warrior holds one spear.

The preceding suggests a way of determining whether two sets A and B, possibly infinite, have equally many elements: if it is possible to 'pair off' the elements of the sets, so that in this process each element of A gets paired with exactly one element of B and each element of B gets paired with exactly one element of A, then the sets have equally many elements. This primitive analysis is made more precise in the following definition.

I.7.1 Definition

Let (A,B) be an ordered pair of nonempty sets.

(1) An **exact pairing of *A* with *B*** is a subset F of the Cartesian product $A{\times}B$ which has the following properties:

Pairing Condition (i) For each element a in A there is a unique element b in B such that $(a,b){\in}F$;

Pairing Condition (ii) For each b in B there is a unique a in A such that $(a,b){\in}F$.

If an ordered pair (a,b) is an element of an exact pairing F, then one also says that ***a* is paired with *b* by *F***.

(2) One says that ***A* has equally many elements as *B***, or, using more modern terminology, that ***A* has the same cardinality as *B***, provided there exists at least one exact pairing of A with B.

(3) Let k be a natural number, and recall that $\mathbb{N}_k$ denotes the set of all natural numbers m such that $1 \leq m \leq k$. One says that a set A is a **finite set** provided either $A = \emptyset$ or there exists a natural number k such that A has the same cardinality as $\mathbb{N}_k$. All other sets are said to be **infinite sets**.

I.7.2 Remarks

(1) Note that the statement 'A has the same cardinality as B' does *not* mean the same as the statement 'B has the same cardinality as A'. Indeed, the former statement concerns

a subset of $A{\times}B$, while the latter statement concerns a subset of $B{\times}A$, a set which need not equal $A{\times}B$. Nevertheless, the two statements are logically equivalent. More precisely, suppose that A has the same cardinality as B, and let F be an exact pairing of A with B. Associate with F a corresponding subset G of $B{\times}A$ by the rule

$$G = \{(b,a){\in}B{\times}A : (a,b){\in}F\}.$$

It is clear that G is an exact pairing of B with A, so that B has the same cardinality as A. One calls G the **reverse exact pairing associated with *F***. It is clear that F is then the reverse exact pairing associated with F.

(2) The previous discussion makes precise the concept of two sets having 'equally many elements'. It is then natural to consider pairs of sets which do *not* have 'equally many elements', and the question of whether one of these sets must have 'fewer' elements than the other. The analysis of this question is fairly simple for finite sets; indeed, it falls under the heading of 'well-known facts about natural numbers'. However, for general sets the situation is somewhat unintuitive. Fortunately, we don't need this general theory in the main body of *This Textbook*, so it is relegated to Appendix B.

The formulation of the concept of 'same cardinality' given above, in terms 'exact pairings' is used here for historical reasons: the basic idea of pairing up objects is quite ancient. From here on, however, we use the conventional modern formulation in terms of bijections.

I.7.3 Theorem

Let A and B be nonempty sets. Then A has the same cardinality as B, in the sense of Definition (I.7.1), if, and only if, there exists a bijection of A onto B.

More precisely, if F is an exact pairing of A with B, then F, viewed as a subset of $A{\times}B$, is a bijection $F : A \to B$ of A onto B. Conversely, if $F : A \to B$ is a bijection, then F satisfies the two pairing conditions of Definition (I.7.1), and thus is an exact pairing of A with B. Furthermore, the reverse exact pairing G associated with F is the inverse $G = F^{-1} : B \to A$ of the bijection F.

The simple proof is left as an exercise.

The reformulation of 'exact pairings' in terms of 'bijections' given above allows one to use known properties of the latter concept to prove some simple facts about 'equal cardinality'.

I.7.4 Theorem

(a) Let A be a set. Then A has the same cardinality as A.

(b) Let A and B be sets, If A has the same cardinality as B then B has the same cardinality as A.

(c) Let A, B and C be sets. If A has the same cardinality as B, and B has the same cardinality as C, then A has the same cardinality as C.

(d) Suppose that A, B, C and D are nonempty sets such that A has the same cardinality as C and B has the same cardinality as D. Then $A{\times}B$ has the same cardinality as $C{\times}D$.

(e) Suppose that A and B are nonempty sets. Then $A{\times}B$ and $B{\times}A$ have the same cardinality.

Outline of Proof

(a) Use the bijection $I_A : A \rightarrow A$. (Recall that I_A is the 'identity map on A'.)

(b) If $F : A \rightarrow B$ is a bijection of A onto B, then $F^{-1} : B \rightarrow A$ is a bijection of B onto A.

(c) If $F : A \rightarrow B$ and $G : B \rightarrow C$ are bijections, then $G{\circ}F : A \rightarrow C$ is a bijection of A onto C.

(d) and (e): The simple proofs are left as an exercise.

I.7.5 Examples

(1) Let $A = \{p, q, r, s, t\}$ be the set consisting of certain (lower case) letters of the English alphabet, and let $B = \{\text{Oh, say, can, you, see}\}$ be the set consisting of the first few words of the US national anthem. There are many different bijections of A onto B; here are two examples, viewed as subsets of the Cartesian product $A{\times}B$:
(a) $F_1 = \{(p, \text{Oh}), (q, \text{say}), (r, \text{can}), (s, \text{you}), (t, \text{see})\}$.
(b) $F_2 = \{(r, \text{you}), (p, \text{say}), (t, \text{can}), (q, \text{Oh}), (s, \text{see})\}$.

(2) Let A be the set $\mathbb{N}$ of all natural numbers, and let B be the set of all perfect squares of natural numbers. That is,

$$A = \{1, 2, 3, 4, 5, \dots\} \text{ and } B = \{1, 4, 9, 16, 25, \dots\}$$

There is an obvious bijection of A onto B, namely the function $F : A \rightarrow B$ given by $F(k) = k^2$ for all k in A.

I.7.6 Remark

Example (2) is originally due to Galileo, in his famous book *Dialogues Concerning Two New Sciences*. The fact it describes is usually called the 'Galileo Paradox'. The 'paradox' for Galileo, expressed in modern terminology, is the counter-intuitive fact that a set can the same cardinality as one of its *proper* subsets.

The following statements about finite sets fall under the heading of 'well-known facts from arithmetic', and thus are accepted here without proof. They are listed explicitly here for ease of reference. Their proofs are relegated to Appendix B.

I.7.7 Theorem

(a) Let X be a nonempty finite set. Suppose that there are natural numbers k and m such that X has the same cardinality as $\mathbb{N}_k$ and the same cardinality as $\mathbb{N}_m$; then $k = m$.

Stated otherwise: if X is a nonempty finite set, then there is a unique natural number k such that X has the same cardinality as $\mathbb{N}_k$.

Remark The number k described here is called the **number of elements of of** X. One sometimes abbreviates this number as $\#(X)$. It is also convenient to write $\#(\emptyset) = 0$.

(b) If Y is a subset of a finite set X, then Y is a finite set, and $\#(Y) \leq \#(X)$. Moreover, the only time one gets $\#(Y) = \#(X)$ is when $Y = X$. In particular, X cannot have the same cardinality as one of its proper subsets.

(c) Suppose that $\{X_1, X_2, \ldots X_n\}$ is a finite collection of finite sets. Then the union $X_1 \cup X_2 \cup \ldots \cup X_n$ is also a finite set. More precisely,

$$\#(X_1 \cup X_2 \cup \ldots \cup X_n) \leq \#(X_1) + \#(X_2) + \ldots + \#(X_n).$$

One gets equality in this last relation if, and only if, the sets are mutually disjoint, in the sense that $X_i \cap X_j = \emptyset$ whenever $i \neq j$.

Bijections of a finite set onto itself have a special terminology.

I.7.8 Definition

Let X be a finite nonempty set. Then a bijection $f : X \to X$ is called a **permutation** of X.

Side Comments (on the counting process):

(1) Part (a) of Theorem (I.7.7) contains the key to the standard process of 'counting' the number of elements of a nonempty finite set. The phrasing of this result makes it appear, however, that one must first come up with the natural number k and then find an exact pairing F. In practice the method is different.

For example, suppose that we wish to 'count out' a large, but finite, collection A of marbles which sit inside an urn. (Why an urn, and not a large box? Basically: It's a tradition in mathematics to use urns – often Greek urns – for counting problems.) Reach into the urn, pull out a marble while saying 'one', and set that marble aside (outside the urn, of course). Then reach into the urn again, pull out another marble while saying 'two', and set the new marble aside. Keep doing this until the marbles in the urn run out. The last number k spoken in this process is the number of marbles originally in the urn, and the process of pulling out marbles establishes the complete pairing of A with $\mathbb{N}_k$. In particular, the value of k is established only at the end of the 'counting' process, as is the corrresponding complete pairing F.

(2) The counting process described above uses the infinite set $\mathbb{N} = \{1, 2, 3, \ldots 99, 100, 101, \ldots\}$ as a 'standard comparison set for counting'. In contrast, a couple of millenia ago, Julius Caesar probably would have used the set $\{I, II, III, \ldots XCIX, C, CI, \ldots\}$ as his 'standard comparison set' for counting counting out the Roman legions. In any event, every such 'standard comparison set' for counting enjoys the following properties:

(a) The 'standard comparison set' has a natural ordering and a unique initial element relative to that ordering. For instance, in the set $\mathbb{N}$ the ordering is the usual one, and the initial element is 1.

(b) There is a systematic procedure for going from one element of the comparison set to the next higher element. For instance, in the set $\mathbb{N}$ the procedure can be stated simply: 'Add 1 to the given element to get the next higher one'. To see just how systematic, and well-known, this procedure is, consider the following *extremely* large element of $\mathbb{N}$:

$$37182946822901028333009155404045373779925111110382238549871172499.$$

If this is the first time you are reading this *Side Comment*, then the probability is high that you have never encountered this gigantic number before. Indeed, it is so large that relatively few people could actually *speak* its name. (In the United States its name would start '37 vigintillion . . . '; in Europe it would start '37 decilliard . . . .) Nevertheless, what you learned as a child provides all you need to know to write down the next higher natural number; try to do so before reading further.

You should get

$$37182946822901028333009155404045373779925111110382238549871172500.$$

(This is one of the great features of the Arabic numerals. Had the original number been expressed in Roman numerals, the result might well have been less simple.) The process used in obtaining the successor of the given giant number can be thought of as 'adding 1'. However, it is likely that most people would simply look at the digits of the original number and write down the answer directly, without really doing 'addition' as such.

(c) By starting with the initial element discussed in (a), and repeating the procedure discussed in (b), one eventually obtains every element of the comparison set.

It turns out that these properties allow one to uniquely construct the other properties of the standard comparison set; see Appendix A.

I.8 Countable and Uncountable Sets

I.8.1 Definition

A set X is said to be **countable** provided either X is finite or X has the same cardinality as $\mathbb{N}$. If the latter situation holds, then X is said to be **countably infinite**, or, sometimes, **denumerable**. A set which is not countable is said to be **uncountable**.

I.8.2 Example

Let $X = \mathbb{Z}$, the set of all integers. Then $\mathbb{Z}$ is countable; indeed, it is countably infinite.

Indeed, define $F : \mathbb{N} \to \mathbb{Z}$ by the rule

$$F(1) = 0; \quad F(2\,k) = k \text{ for all } k \text{ in } \mathbb{N}; \quad F(2\,k-1) = -k \text{ for all } k \text{ in } \mathbb{N}.$$

It is easy to see that the map $F : \mathbb{N} \to \mathbb{Z}$ is a bijection.

I.8.3 Remarks

(1) The usage of the terms 'countable', 'countably infinite' and 'denumerable' here is quite common in the mathematical literature, but many authors follow a slightly different usage. For example, some use 'countable' for what we call 'countably infinite' or 'denumerable'; for such authors a finite set is not countable. In contrast, some authors use 'denumerable' for what we call 'countable'; they would then need to write 'denumerably infinite' for what we call 'countably infinite'.

(2) As has already been pointed out, in this chapter we treat the set $\mathbb{N}$ of natural numbers as a 'primitive concept' related to 'counting'. It is clear, however, that any other countably infinite set would work just as well for the purpose of 'counting'. In Chapter (II) we replace our 'primitive' notion of $\mathbb{N}$ with a certain countable subset of the real numbers. This subset can be used not just for counting, but also for interacting directly with other real numbers; see Theorem (II.2.11).

I.8.4 Theorem

Let A be a subset of $\mathbb{N}$. Then A is countable. More precisely, A is a finite set if it is bounded in $\mathbb{N}$, while it is countably infinite if it is unbounded in $\mathbb{N}$. (As usual, the empty set is viewed as being a bounded subset of $\mathbb{N}$.)

Proof: The result is obviously true if $A = \emptyset$, so without loss of generality assume that A is nonempty.

Case 1 Suppose that A is a bounded subset of $\mathbb{N}$, so that there exists a natural number m such that $k \leq m$ for all k in A. Then clearly $A \subseteq \mathbb{N}_m$ and thus, by Part (b) of Theorem (I.7.7), A is a finite set, and thus A is countable.

Case 2 Suppose that A is an unbounded subset of $\mathbb{N}$. The 'obvious' choice of a bijection $F : \mathbb{N} \to A$ is given by the following rule:

$$F(m) = \text{ the } m\text{-th smallest element of the set } A.$$

What follows simply makes this a bit more precise.

For each element m in A let A_m denote the set of all natural numbers k in A such that $k \leq m$; that is, $A_m = A \cap \mathbb{N}_m$. Note that A_m is nonempty because, by definition, m itself is an element of A_m. Also, A_m is a finite set since it is a subset of the finite set $\mathbb{N}_m$. Now define a function $G : A \to \mathbb{N}$ by the rule $G(m) = \#(A_m)$; that is, $G(m)$ is the number of elements of the finite set A_m.

It is clear that the map $G : A \to \mathbb{N}$ is a surjection. Indeed, let $B = G[X]$. Note that if m is the smallest element of A, which exists by the Least-Natural-Number Principle, then $A_m = \{m\}$, so that $G(m) = 1$; that is, $1 \in B$. Firthermore, suppose that $k \in B$, so that there exists m in A such that $k = \#(A_m)$. Let n be the smallest element of the nonempty set $A \setminus A_m$. Then clearly $A_n = A_m \cup \{n\}$. Since $n \notin A_m$, it follows that $\#(X_n) = \#(A_m) + 1 = k + 1$; that is, $k + 1 = G(n)$, so $(k+1) \in B$. It now follows from the Principle of Mathematical Induction that $B = \mathbb{N}$, so that the function $G : A \to \mathbb{N}$ is a surjection.

Likewise, it is clear that the function $G : A \to \mathbb{N}$ is an injection. Indeed, suppose that m and n are elements of A with $m \neq n$; without lose of generality, assume that $m < n$. Then it is clear that A_m is a *proper* subset of A_n, so that $\#(A_m) < \#(A_n)$; that is, $G(m) < G(n)$.

It follows that the map $G : A \to \mathbb{N}$ is a bijection. Let $F = G^{-1} : \mathbb{N} \to A$. Then $F : \mathbb{N} \to A$ is also a bijection, and it follows that A is a countable set, as required; indeed, A is countably infinite.

I.8.5 Remarks

(1) It is clear that F is strictly increasing on $\mathbb{N}$, in the sense that $F(i) < F(j)$ whenever i, j in $\mathbb{N}$ satisfy $i < j$. Conversely, $F : \mathbb{N} \rightarrow A$ is the only bijection with this property. For that reason, in *This Textbook* we refer to the map F just constructed as the **strictly increasing map of $\mathbb{N}$ onto A** ; we denote this bijection by Ψ_A. (The latter notation is not standard.) Clearly the quantity $\Psi_A(k)$ agrees with one's intuitive concept of the k-th smallest element of the subset A.

(2) It is an easy exercise to show that if $F : \mathbb{N} \rightarrow \mathbb{N}$ is a strictly increasing function with values in $\mathbb{N}$, then for each k in $\mathbb{N}$ one has $F(k) \geq k$. Futhermore, if $F(n) = n$ for a particular n in $\mathbb{N}$, then $F(k) = k$ for all k in $\mathbb{N}$ such that $1 \leq k \leq n$.

(3) If A is an infinite subset of $\mathbb{N}$, then it is possible to express A as $A = \{k_1, k_2, \ldots k_m, \ldots\}$, with each k_m being an element of $\mathbb{N}$. However, the set notation allows the possibility that the numbers k_1, k_2 and so on are not in increasing order; indeed, it allows the possibility that the same number may appear more than once in this list. Of course one could simply append the phrase 'where $k_1 < k_2 < \ldots < k_m < \cdot$' to the equation; but this gets tedious when done repeatedly. Another shorter solution is to add the phrase 'where $k_m = \Psi_A(m)$'. However, the most common solution is to write instead $A = \{k_1 < k_2 < \ldots k_m < \ldots\}$. In *This Textbook* we usually follow the last approach.

I.8.6 Corollary

If one removes a finite subset from $\mathbb{N}$, what remains still has the same cardinality of the original set $\mathbb{N}$.

Proof This follows from the preceding theorem by noting that removing a bounded subset from $\mathbb{N}$ leaves an unbounded subset of $\mathbb{N}$.

Remarks:

(1) One tends to feel that a finite set which has a very, very large number of elements is 'a nearly infinite set'. The preceding corollary says that this feeling, although perhaps natural, is very, very misleading.

(2) Note that Chapter Quote (4) for this chapter provides a poetic illustration of the content of the preceding corollary: 'removing' the first $10,000$ numbers from $\mathbb{N}$ does not diminish the 'size' of the set of numbers remaining. Because of the source of this quote, we refer to this corollary as the **Amazing Grace Property for $\mathbb{N}$**.

(3) The situation becomes more complicated if one removes an *infinite* set from $\mathbb{N}$. Indeed, the resulting set may have the same cardinality as $\mathbb{N}$ (e.g., remove all the odd numbers), or it may be a finite set (e.g., remove all the natural numbers greater than 5).

The next result gives analogs, for general countable sets, of parts of Theorem (I.8.4) and Corollary (I.8.6); the simple proof is left to the reader.

I.8.7 Theorem

(a) Every subset of a countable set is also countable.

(b) If Y is an infinite subset of a countably infinite set, then Y is countably infinite.

(c) ('Amazing Grace' Property of Countably Infinite Sets) If one removes a finite subset from a countably infinite set, then what remains is a countably infinite set.

The next theorem provides a useful tool for proving that a given nonempty set is countable without actually needing to find a bijection.

I.8.8 Theorem

Suppose that Y is a nonempty set. Then a necessary and sufficient condition for Y to be countable is that there exist a surjection of $\mathbb{N}$ onto Y.

Proof Suppose that there exists a function $f : \mathbb{N} \to Y$ of $\mathbb{N}$ onto Y. Define a corresponding function $h : Y \to \mathbb{N}$ by the rule that if $y \in Y$, then $h(y)$ is the smallest element of the subset $f^{-1}[\{y\}]$. (This subset of $\mathbb{N}$ is nonempty because f maps $\mathbb{N}$ onto Y. The existence of a unique *smallest* element of this set then follows from the Least-Natural-Number Principle.) Let $W = h[Y]$. It is clear that $h : Y \to W$ is a bijection of Y onto W. Indeed, h automatically maps Y onto $h[Y]$. Furthermore, the fact that h is one-to-one follows from the fact that $f(h(y)) = y$ for each y in Y, so that $h(y_1) = h(y_2)$ implies $y_1 = y_2$ because f is a function and thus has a unique values at each point. It follows from the proceding that Y has the same cardinality as a nonempty subset of $\mathbb{N}$. It then follows from Theorem (I.8.4) and Theorem (I.7.4) that Y is countable, as claimed.

The proof of the converse, namely that if Y is countable then there exists a surjection of $\mathbb{N}$ onto Y, is obvious and is left to the reader.

I.8.9 Corollary

Suppose that Y is a nonempty set. Then a necessary and sufficient condition for Y to be countable is that for every countably infinite set X there exist a surjection of X onto Y.

The trivial proof is left as an exercise.

Remark By expressing the preceding results in terms of surjections, and not bijections, we are able to prove some theorems about 'countability' without needing to always separate into separate 'finite' and 'countably infinite' cases. An important tool for doing that is the following result.

I.8.10 Important Example

There is a well-known fact about natural numbers which one learns in elementary arithmetic: If $m \in \mathbb{N}$, then there are unique natural numbers k and n such that $m = 2^{k-1}(2n-1)$. That is, m can be expressed, in exactly one way, as the product of a (nonnegative) power of 2 multiplied by an odd natural number. For instance:

$$36 \;=\; 2{\cdot}18 \;=\; 2^2{\cdot}9 \;=\; 2^2{\cdot}(2{\cdot}5-1).$$

Thus, in this case $k \;=\; 3$ and $n \;=\; 5$.

Now for each k in $\mathbb{N}$ let X_k denote the set of all natural numbers of the form $2^{k-1}(2n-1)$ for n in $\mathbb{N}$. That is,

$$X_k \;=\; \{2^{k-1}{\cdot}1, 2^{k-1}{\cdot}3, 2^{k-1}{\cdot}5, \ldots 2^{k-1}{\cdot}(2n-1), \ldots\}$$

Note that these sets form a countably infinite family $\{X_1, X_2, \ldots X_n, \ldots\}$ of subsets of $\mathbb{N}$ such that

(i) Each set X_k is countably infinite; indeed, one obvious bijection of $\mathbb{N}$ with X_k is given by the rule $f_k(n) \;=\; 2^{k-1}(2n-1)$ for each n in $\mathbb{N}$.

(ii) If $k \neq l$ then $X_k \cap X_l \;=\; \emptyset$; that is, the sets in this family are pairwise disjoint;

(iii) $\bigcup_{k=1}^{\infty} X_k \;=\; \mathbb{N}$.

The next result illustrates how one can use this decomposition.

I.8.11 Theorem

Suppose that S is a set which can be expressed as the union of a nonempty countable family of countable nonempty sets. Then S is itself a countable set.

Proof: Let $\mathcal{F}$ be a nonempty countable family of nonempty countable sets such that $S \;=\; \bigcup \mathcal{F}$. It follows from Theorem (I.8.8) that there exists a surjection $f : \mathbb{N} \to S$ of $\mathbb{N}$ onto $\mathcal{F}$. As in Example (I.8.10) above, for each k in $\mathbb{N}$ let X_k be the countably infinite subset of $\mathbb{N}$ consisting of all natural numbers of the form $2^{k-1}\,(2\,n-1)$ with n in $\mathbb{N}$. For each k in $\mathbb{N}$ let $g_k : X_k \to S$ be a surjection of the countably infinite X_k onto the countable subset $f(k)$. Since the sets of the form X_k, with k in $\mathbb{N}$, are mutually disjoint, it follows from Theorem (I.6.6), the 'Union-of-Finctions Theorem', that there is a function $g : \mathbb{N} \to S$ such that for each k in $\mathbb{N}$ one has $g_k \;=\; g|_{X_k}$. It is clear that $g : \mathbb{N} \to S$ is a surjection, and thus, by Corollary (I.8.9), the set S is countable, as claimed.

I.8.12 Corollary

The following sets are all countably infinite.

(a) The Cartesian product $\mathbb{N}\times\mathbb{N}$.

(b) The set $\mathbb{Z}$ of all integers.

(c) The set $\mathbb{Q}^+$ of all positive rational numbers.

(d) The set $\mathbb{Q}$ of *all* rational numbers.

Proof: Note first that each of the sets mentioned above is obviously an infinite set, so we need only prove they are all countable.

(a) For each k in $\mathbb{N}$ let Y_k be the set of all ordered pairs of the form (k, m) for m in $\mathbb{N}$, and let $\mathcal{F}$ be the family of all sets of the form Y_k with k in $\mathbb{N}$. It is obvious that $\mathcal{F}$ is a countably infinite family of countable sets, and that the union of this family is $\mathbb{N}\times\mathbb{N}$. It follows from Theorem (I.8.11) that $\mathbb{N}\times\mathbb{N}$ is countable.

(b) Define a map $f : \mathbb{N}\times\mathbb{N} \to \mathbb{Z}$ by the rule $f(m,n) = n-m$ for all pairs (m,n) in $\mathbb{N}\times\mathbb{N}$. It is clear that f is a surjection of the countable set $\mathbb{N}\times\mathbb{N}$ onto the set $\mathbb{Z}$. It follows from Theorem (I.8.8) that $\mathbb{Z}$ is countable. (Note that we gave a more direct proof of this fact above in Exampl (I.8.2).)

(c) Define a function $g : \mathbb{N}\times\mathbb{N} \to \mathbb{Q}^+$ by the rule $g(m,n) = m/n$ for each pair (m,n) in $\mathbb{N}\times\mathbb{N}$. By definition of 'rational numbers', it is clear that g is a surjection of the countable set $\mathbb{N}\times\mathbb{N}$ onto the infinite set $\mathbb{Q}^+$. It follows as before that $\mathbb{Q}^+$ is countable.

(d) The proof of this is left as an exercise.

The fact that $\mathbb{Q}$ has the same cardinality as $\mathbb{N}$ (see Corollary (I.8.12) above) should appear counter-intuitive: Geometrically speaking, the natural numbers are scattered discretely along the positive real axis; in particular, any interval of length less than one has at most one natural number in it. In contrast, the real axis is 'densely populated' by rationals; more precisely, every interval in the real axis of positive length, no matter how small, contains infinitely many rational numbers.

Once one accepts the fact that $\mathbb{Q}$ is countably infinite, it then is natural to conjecture that *all* infinite sets are countably infinite. If this conjecture were valid, then the 'Galileo Paradox' (see Remark (I.7.6)) could be interpreted to say that no infinite set is 'larger' than any other infinite set. The next example shows that this conjecture is not correct.

I.8.13 Example (Existence of Uncountable Sets)

Let Y be the set of real numbers y in the closed interval $[0,1]$ such that y admits a decimal representation of the form $y = 0.d_1\,d_2\,\ldots d_n\,\ldots$, where each decimal digit d_n is either 0 or 9. It is clear from well-known properties of the decimal representation of real numbers that if $y = 0.d_1\,d_2\,\ldots d_n\,\ldots$ and $y' = 0.d'_1\,d'_2\,\ldots d'_n\,\ldots$ are such decimal representations of numbers y and y', respectively, then $y = y'$ if, and only if, $d_n = d'_n$ for each number n in $\mathbb{N}$.

Claim If $F : \mathbb{N} \to Y$ is a function with domain $\mathbb{N}$ and with values in the set Y, then it is not a surjection. In particular, it follows from Theorem (I.8.8) that the set Y is uncountable.

Proof of Claim Define a number $z = 0.c_1\,c_2\,\ldots c_n\,\ldots$ in Y by the following rule: if the n-th decimal digit of $F(n)$ is 0, then $c_n = 9$; but if the n-th decimal digit of $F(n)$ is 9, then $c_n = 0$. It is clear that z is in Y, and that for each n the n-th decimal digit of z differs from the n-th decimal digit of $F(n)$. It follows that for each n one has $z \neq F(n)$. In particular, z is not in the image of the function F, hence F does not map $\mathbb{N}$ onto Y, as claimed

Remarks (1) The description of the set Y above uses familiar properties of the decimal representation of real numbers. In the current chapter we take such properties for granted, but later we prove them rigorously using the axioms for $\mathbb{R}$ in Chapter (II).

(2) The set Y constructed above is sometimes called the **Cantor Middle-Eight-Tenths Set**; see the exercises for an explanation for this name. This set is just one of a family of

similar subsets of $\mathbb{R}$, called **Cantor sets**, which play an important role in analysis; such sets are studied in more detail in Appendix B.

I.9 Sequences and Subsequences

Preliminaries The intuitive concept of an 'infinite sequence' is quite ancient, and should be familiar to everyone. The usual presentation of this idea is that it is an 'infinite ordered list of objects': $(x_1, x_2, \ldots x_k, \ldots)$. For example, associated with the rational number 1/12 is the repeating decimal representation $0.08333\ldots$. This decimal expression actually encodes the sequence $(0.0, 0.08, 0.083, 0.0833, \ldots)$ of all the truncated decimal approximations of 1/12. The view of a 'sequence' being an ordered list of objects also matches the use of this word in ordinary English.

It is assumed above that the reader already understands what it means to be an 'ordered list of objects'; that is, it is treated as a 'primitive concept': 'You know one when you see one'. In practice this viewpoint causes no logical difficulties, and indeed much of the time we find it the most useful way to interpret this concept. However, one of the goals of modern mathematics is to formulate all important concepts ultimately in terms of set theory. The standard approach, which is presented below, is based on the following observation: in an ordered list of the type we are considering, there is a first object in the list, followed by the second object, then the third object, and so on; the process never stops. The next definition formulates this primitive concept in a a modern way using the 'function' concept.

I.9.1 Definition

Let X be a nonempty set of objects.

(1) A function $\xi : \mathbb{N} \to X$, whose domain is $\mathbb{N}$ and whose values are all in X, is said to be an **infinite sequence in *X***. If $j \in \mathbb{N}$ then the point $\xi(j)$ is called the ***j*-th term of the sequence** ξ.

(2) The set $\{x_1, x_2, \ldots x_k, \ldots\}$, whose elements are the terms of the sequence ξ, is called the **term-set** of the sequence ξ, and in *This Textbook* is denoted by S_ξ.

Notes (1) We shall use both the 'function' interpretation of sequences, as given in the preceding definition, as well as the classic 'ordered list' interpretation. It is always easy to reformulate the one in terms of the other.

(2) In the context of sequences, we often refer to a natural number which appear as a subscript, such as the j in the expression x_j, as an **index** (plural: 'indices').

I.9.2 Examples

(1) Let X be a nonempty set, and let c be any element of X. Suppose that $\xi : \mathbb{N} \to X$ is a constant function, with $\xi(k) = c$ for all k in $\mathbb{N}$. Then ξ corresponds to the 'constant

ordered list' $(c, c, \ldots c, \ldots)$. We often refer to such a sequence as a **constant sequence**. The term-set for this sequence is the singleton set $S_\xi = \{c\}$.

(2) Consider the list $(1, 0, 1, 0, 1, 0, \ldots)$; the pattern should be easy to discern. This list can be described by a function $\xi : \mathbb{N} \to \mathbb{Z}$ using the following rule:

$$\xi(k) = \begin{cases} 1 & \text{if } k \text{ is odd} \\ 0 & \text{if 0 if } k \text{ is even} \end{cases}$$

The reader is encouraged to show that the function ξ can also be described by a single formula:

$$\xi(k) = \frac{1 - (-1)^k}{2} \text{ for all } k \text{ in } \mathbb{N}.$$

In contrast, consider a similar ordered list $(0, 1, 0, 1, 0, 1, \ldots)$. It is described by the function $\zeta : \mathbb{N} \to \mathbb{Z}$ whose formula can be written as follows:

$$\zeta(k) = \frac{1 + (-1)^k}{2} \text{ for all } k \text{ in } \mathbb{N}.$$

It is clear that the functions ξ and ζ are not equal to each other; for example, $\xi(1) = 1$ while $\zeta(1) = 0$. That is, the two sequences here, although they are very similar, do not equal each other. (See Remark (1) below as well.)

Note that the sequences ξ and ζ have the same (doubleton) term-sets:

$$S_\xi = \{1, 0\} \text{ while } S_\zeta = \{0, 1\} = \{1, 0\}.$$

(3) The sequence $\left(1, \frac{1}{2}, \frac{1}{3}, \ldots \frac{1}{k}, \ldots\right)$, whose k-th term is the number $1/k$, is called the **harmonic sequence**.

Similarly, the sequence $\left(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots (-1)^{k-1}\frac{1}{k}, \ldots\right)$ is called the **alternating harmonic sequence**, since it is obtained by alternating the signs of the terms of the harmonic sequence.

(4) A sequence of the form $(A, A\,r, A\,r^2, \ldots A\,r^n, \ldots)$ is called a **geometric sequence with initial term $\boldsymbol{A}$ and common ratio $\boldsymbol{r}$**. If either A or r equals zero, the correspond sequence is of little interest: $(0, 0, \ldots 0, \ldots)$ or $(A, 0, 0 \ldots 0)$. In the more interesting case that $A \neq 0$ and $r \neq 0$, one has $r = (A\,r^{n+1})/(Ar^n)$ for each n. This explains the 'r is the common ratio' terminology. Of course, it is clear why A is called the 'initial term' of this sequence.

Remarks

(1) Many authors would abbreviate the notation $(x_1, x_2, \ldots x_k, \ldots)$ for a sequence ξ by the expression $\{x_k : k \in \mathbb{N}\}$, or even by just $\{x_k\}$ with the tacit understanding that the index k ranges over the set $\mathbb{N}$. In particular, they would use the 'braces' { and }, instead of our use of the parentheses (and) as the delimiters.

Unfortunately, the notation $\{x_k : k \in \mathbb{N}\}$ is also a common abbreviation for the *set of values* $\{x_1, x_2, \ldots x_k, \ldots\}$ of the function ξ; that is, the term-set S_ξ as described above. As

Example (2) above illustrates, however, two different sequences can have the same term-set, so this 'braces' notation can lead to confusion. To avoid such confusion is why we prefer to use the 'parentheses' notation for sequences in *This Textbook*.

(2) In *This Textbook* we often follow the common custom of abbreviating the phrase 'infinite sequence' to just 'sequence'. The analogous concept of 'finite sequence' has already been described in the discussion of 'ordered tuples'; see the beginning of Section (I.4), where this idea is treated as a 'primitive concept'.

Subsequences of an Infinite Sequence

There is an important technique for constructing new sequences from a given sequence.

Preliminary Example Let $\xi = (x_1, x_2, \ldots x_k, \ldots)$ be the Alternating Harmonic Sequence described above; thus, $\xi(k) = x_k = (-1)^{k-1}/k$ for each natural number k. The corresponding ordered list is

$$\xi = (x_1, x_2, \ldots x_k, \ldots) = \left(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots\right) \qquad (*)$$

If one deletes from this list the terms x_k with index k even, and the retains, in their original order, the terms x_k with index k odd, one obtains an infinite ordered 'sublist' $\zeta = (z_1, z_2, \ldots z_m, \ldots)$ of the original list ξ:

$$\zeta = \left(1, \frac{1}{3}, \frac{1}{5}, \ldots \frac{1}{2\,m-1} \ldots\right);$$

in the 'function' viewpoint, one has $\zeta(m) = 1/(2\,m-1)$ for each m in $\mathbb{N}$.

The key step in forming the 'sublist' ζ from the original list ξ is choosing the set A of which indices k to retain; in this example, $A = \{1, 3, 5, \ldots (2\,m-1), \ldots\}$.

Now let us repeat this 'deletion/retention' process to form a third ordered list $\tau = (t_1, t_2, \ldots t_n, \ldots)$. More precisely, let τ be the list obtained from the subsequence ζ by retaining the terms z_m for which $2\,m-1$ is a multiple of 3 and deleting the others. One sees easily that the retained terms are z_m with m of the form $m = 2 + 3\,(n-1)$ with n in $\mathbb{N}$; that is, with $m = 2, 5, 8, 11, \ldots$

$$\tau = \left(\frac{1}{3}, \frac{1}{9}, \frac{1}{15}, \ldots\right)$$

Note that τ can be interpreted simultaneously as a sublist of ζ, and thus a *sub*sublist of the original list ξ, and also directly as a sublist of the original list ξ. In the former interpretation, the list τ is obtained by retaining the terms z_m for which m is in the set $B = \{2, 5, 8, \ldots\}$. In the latter interpretation, the list τ is obtained from the original list ξ by retaining the terms x_k for which k is in the set $C = \{3, 9, 15, \ldots\}$. That is, one can obtain the subsublist τ from the original list ξ in two stages: first, delete the indices k with k even, retaining the indices in the set A described above; then delete even more indices k from A to get the retained indices forming the set C.

These ideas are extended and formalized in the following definition. However, instead of using words such as 'sublist' or 'subsublist', we follow the usual custom and use words such as 'subsequence' and 'subsubsequence'.

I.9.3 Definition

Let X be a nonempty set, and let $\xi = (x_1, x_2, \ldots)$ be an infinite sequence in X.

(1) Suppose that $A = \{k_1 < k_2 < \ldots < k_m < \ldots\}$ is an infinite subset of $\mathbb{N}$. (See Part (3) of Remark (I.8.5) for an explanation of this notation.) Then the sequence $\zeta = (x_{k_1}, x_{k_2}, \ldots x_{k_m}, \ldots)$ is called the **subsequence of $\boldsymbol{\xi}$ determined by the set $\boldsymbol{A}$**. Equivalently, $\zeta = \xi \circ \Psi_A : \mathbb{N} \to X$, where $\Psi_A : \mathbb{N} \to A$ is the strictly increasing map of $\mathbb{N}$ onto A described in Part (1) of Remark (I.8.5); that is, $\Psi_A(m)$ is the m-th smallest element of the set A. In *This Textbook*, the subsequence ζ arising from ξ this way is denoted ξ_A. (The notation ξ_A is not standard, but it is convenient.)

(2) More generally, let m be a natural number. Suppose that $\mathcal{A} = (A_1, A_2, \ldots A_m)$ is an ordered m-tuple of infinite subsets of $\mathbb{N}$ such that $A_j \subseteq A_{j+1}$ for each $j = 1, 2, \ldots m-1$. One calls $\mathcal{A}$ a **subsequence structure of order $\boldsymbol{m}$**. Furthermore, the sequence ξ_{A_m} is called the **$\boldsymbol{m}$-th order subsequence of $\boldsymbol{\xi}$ associated with the subsequence structure $\mathcal{A}$**.

Note A subsequence of order 1 is simply a subsequence. Subsequences of order 2 are often called *sub*subsequences; however it becomes tedious to replace 'order m' by m copies of the syllable 'sub' when $m \geq 3$. Sometimes the original sequence ξ is referred to as a 'subsequence of order 0' of itself.

I.9.4 Examples

(1) Consider a sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$, thought of as an ordered list. Let n be a natural number, and let $A = \{n, n+1, \ldots\}$; that is, A is the (infinite) subset of $\mathbb{N}$ obtained by omitting the first $n-1$ natural numbers from $\mathbb{N}$. (Note that if $n = 1$ then $A = \mathbb{N}$.) The resulting subsequence ζ of ξ, viewed as an ordered list, is $(x_n, x_{n+1}, \ldots)$; viewing ζ as a function, one has $\zeta(m) = \xi(m+n-1)$ for every m in $\mathbb{N}$. One calls any subsequence of ξ obtained this way a **tail of $\boldsymbol{\xi}$**; specifically, the sequence ζ constructed here is the **$\boldsymbol{n}$-tail of** ξ.

(2) It is clear that every sequence can be viewed as a subsequence of itself. Indeed, note that $\Psi_{\mathbb{N}} = I_{\mathbb{N}}$, the identity map on $\mathbb{N}$, and thus

$$\xi = \xi \circ I_{\mathbb{N}} = \xi_{\mathbb{N}}.$$

By repeating this argument, one also sees that for every m in $\mathbb{N}$ the sequence ξ can be viewed as a subsequence of order m of itself; namely, it is the subsequence of order m associated with the subsequence structure $\mathcal{A} = (A_1, A_2, \ldots A_m)$, where $A_j = \mathbb{N}$ for each j.

(3) More generally, suppose that A is an infinite subset of $\mathbb{N}$ and that $\zeta = \xi_A$. Then for every m one can view ζ as being the m-th order subsequence of ξ associated with the associated with the subsequence structure $\mathcal{A} = (A_1, A_2, \ldots A_m)$, where $A_j = A$ for each j.

I.9.5 Remark

The 'subsequence structure' terminology is not standard, but the underlying concept is present in all analysis texts, often in hidden form. The standard way of dealing with subsequences of order 2 or higher is by the use of multiply-subscripted indices.

I.9.6 Example

In Chapter (III) one learns about an important property, call it Property X, which pertains to infinite sequences of real numbers. It is not important for the purposes of this example to know what this property is; one needs to know only that some sequences have it, some don't. Furthermore, if a given sequence has Property X, then so does every subsequence of it. In the same chapter there is also a major theorem which can be phrased here as follows:

Major Theorem If $\xi = (x_1, x_2, \dots x_k, \dots)$ is a sequence of real numbers, then at least one subsequence of ξ has Property X.

There is an extension of this theorem, *not* stated in Chapter (III), which applies to sequences of points in Euclidean space:

Corollary of Major Theorem Suppose that $\sigma = (P_1, P_2, \dots P_k, \dots)$ is a sequence of points in $\mathbb{R}^3$, the standard Euclidean space of dimension 3 from high-school analytical geometry. Write $P_k = (x_k, y_k, z_k)$ where x_k, y_k and z_k are all real numbers. Let $\xi = (x_1, x_2, \dots x_k, \dots)$, $\eta = (y_1, y_2, \dots y_k, \dots)$, and $\zeta = (z_1, z_2, \dots z_k, \dots)$ be the corresponding 'component' real sequences.

Claim There exists a subsequence of the sequence σ for which each of the component real sequences ξ, η and ζ has Property X.

Note At first this sounds obvious. Indeed, it is given that the sequence ξ has a subsequence with Property X, and likewise for the sequences η and ζ. It is tempting to express this in the following index form:

There is a subsequence $(x_{k_1}, x_{k_2}, \dots x_{k_m}, \dots)$ with Property X. Likewise, there is a subsequence $(y_{k'_1}, y_{k'_2}, \dots y_{k'_m}, \dots)$ with Property X, and a subsequence $(z_{k''_1}, z_{k''_2}, \dots z_{k''_m}, \dots)$ with Property X.

The use of the 'primes' above reflects the fact that different sequences of reals may well require different choices of indices to obtained a subsequence which has Property X. What the Corollary requires, however, is a single choice of indices which works simultaneously for all three sequences ξ, η and ζ. At this point one might even begin to doubt the correctness of the claim; but in fact it is true.

First Proof of Claim (using indices) Let $(x_{k_1}, x_{k_2}, \dots x_{k_m}, \dots)$ be a subsequence of the real sequence ξ which has Property X, and let $(P_{k_1}, P_{k_2}, \dots P_{k_m}, \dots)$ be the corresponding subsequence of σ. Note that $P_{k_m} = (x_{k_m}, y_{k_m}, z_{k_m})$ for each index m. In particular, the choice of a subsequence of ξ with Property X leads to choices of subsequences of η and ζ as well; but the latter subsequences need not have Property X. However, since the Major Theorem applies to *every* real sequence, it implies that the subsequence $(y_{k_1}, y_{k_2}, \dots y_{k_m}, \dots)$ of η has itself a subsequence with Property X. Let $(y_{k_{m_1}}, y_{k_{m_2}}, \dots y_{k_{m_n}}, \dots)$ be such a subsubsequence of η. This in turn determines a subsubsequence of the original point sequence σ, namely

$$\left(P_{k_{m_1}}, P_{k_{m_2}}, \ldots P_{k_{m_n}}, \ldots\right).$$

The n-th term of this subsubsequence of σ is $P_{k_{m_n}} = \left(x_{k_{m_n}}, y_{k_{m_n}}, z_{k_{m_n}}\right)$. That is, choosing the appropriate subsubsequence of η to have Property X leads to a corresponding subsubsequence of σ, which in turn leads to corresponding subsubsequences of ξ and ζ. The first of these subsubsequences, being a subsequence of a subsequence ξ having Property X, also has Property X, by the fact given above that if a sequence has Property X, then so does every one of its subsequences. Unfortunately, there is no reason to expect that the subsubsequence of ζ constructed here has Property X. However, it does have itself have a subsequence, which would then be a subsubsubsequence of the original sequence ζ, with Property X. As is the custom at this stage of such an argument, the remaining details are left to the reader; but notice that the final answer involves expressions such as $P_{k_{m_{n_q}}}$ whose stacked indices are hard to read.

Second Proof of Claim (using infinite subsets of $\mathbb{N}$) Let A be such an infinite subset of $\mathbb{N}$ for which the subsequence ξ_A has Property X; such A exists by applying the Major Theorem to ξ. Next let B be an infinite subset of A for which η_B has Property X; once again, the fact that η_A has a subsequence with Property X follows from the Major Theorem; the fact that a subsequence of η_A can be expressed in the form η_B for some infinite subset B of A has already been noted. Similarly, let C be an infinite subset of B such that ζ_C has Property X. (Note that the ordered triple (A, B, C) is a subsequence structure of order 3.) It follows easily from the fact that if a sequence has Property X, then so does each of its subsequences, that $\sigma_C = (\xi_C, \eta_C, \zeta_C)$ is a subsequence of σ with the required property.

I.9.7 Remarks

(1) If A and B are infinite subsets of $\mathbb{N}$ such that $\xi_A = \xi_B$, it need not be the case that $A = B$. For example, if ξ is a constant sequence then clearly $\xi_A = \xi_B = \xi$ for every such pair of subsets A and B. However, if, when viewed as a function $\xi : \mathbb{N} \rightarrow X$, the sequence ξ is one-to-one, then it is easy to see that $\xi_A = \xi_B$ implies $A = B$.

(2) The difference between the two proofs above is one of notation. For example, the notation $(x_{k_1}, x_{k_2}, \ldots x_{k_m}, \ldots)$ simply lists out 'explicitly' the elements of the set A in increasing order; indeed, it is clear that $k_m = \Psi_A(m)$; the m-th smallest element of A. However, only the existence of a suitable infinite subset A of $\mathbb{N}$ is used, and not the fact that its terms can be written in a particular order. The introduction of the subscripted indices k_m does nothing except to complicate the notation. Nevertheless, from time to time in *This Textbook* we shall use subscripted indices, simply because that is the most common notation in analysis books and papers, so one must become familiar with it as well.

Subsequence structures of *infinite* order also are important in analysis. However, the way they arise and are used is different from the situation with subsequence structures of finite order.

I.9.8 Definition

A **subsequence structure of infinite order** is an infinite sequence $(A_1, A_2, \ldots A_m, \ldots)$ of infinite subsets of $\mathbb{N}$ such that for each k in $\mathbb{N}$ one has $A_{k+1} \subseteq A_k$.

I.9.9 Examples

(1) For each m in $\mathbb{N}$ let $A_m = \{m, m+1, \ldots\}$. Then $\mathcal{A} = (A_1, A_2, \ldots A_m, \ldots)$ is also subsequence structure of infinite order. Note that the intersection $\bigcap_{m=1}^{\infty} A_m$ of the sets A_m is empty.

(2) For each m in $\mathbb{N}$ let A_m be the infinite subset of $\mathbb{N}$ given by

$$A_m = (1, 2^{m-1}{\cdot}2, 3, 2^{m-1} \cdots 4, 5, \ldots);$$

the pattern is clear: $\Psi(k) = k$ if k is odd, while $\Psi(k) = 2^{m-1}{\cdot}k$ if k is even. It is clear that $A_{m+1} \subseteq A_m$ for each natural number m, so that $\mathcal{A} = (A_1, A_2, \ldots A_m, \ldots)$ is a subsequence structure of infinite order. It is also clear that the intersection $\bigcap_{m=1}^{\infty} A_m$ of the sets A_m is the infinite subset $B = \{1, 3, 5, \ldots\}$ of $\mathbb{N}$.

Let $\mathcal{A} = (A_1, A_2, \ldots A_m, \ldots)$ be a subsequence sructure of infinite order, and let $\xi : \mathbb{N} \to X$ be a sequence with values in a nonempty set X. The structure $\mathcal{A}$ determines for each m a corresponding subsequence of ξ, namely ξ_{A_m}, such that for each m in $\mathbb{N}$ the sequence $\xi_{A_{m+1}}$ is itself a subsequence of ξ_{A_m}. It is natural to ask whether there exists a subsequence ζ of ξ which is simultaneously a subsequence of ξ_{A_m} for each m. If there is, ζ must be of the form $\zeta = \xi_B$ for some infinite subset B of $\mathbb{N}$ such that $B \subseteq A_m$ for each m. That is, B must be an infinite subset of $\bigcap_{m=1}^{\infty} A_m$.

Example (1) above shows that such an infinite subset B of $\mathbb{N}$ may fail to exist. However, a slight weakening of the requirement that B be a subset of each A_m allows one to procede in a useful manner.

I.9.10 Definition

Let A and B be infinite subsets of $\mathbb{N}$. One says that ***B* is eventually a subset of *A*** provided there exists a natural number N such that if $n \in B$ and $n \geq N$ then $n \in A$; equivalently, if there are at most finitely many elements of B that are *not* in A.

I.9.11 Example

Let $\mathcal{A} = (A_1, A_2, \ldots A_m, \ldots)$ be a subsequence stucture of infinite order. Then there exists an infinite subset B of $\mathbb{N}$ such that for each m in $\mathbb{N}$ the set B is eventually a subset of A_m. For instance, let n_1 be any element of the set A_1. Then let n_2 be any element of A_2 such that $n_2 > n_1$; such n_2 exists because A_2 is an infinite subset of $\mathbb{N}$. Continuing this way, suppose that n_1, $n_2, \ldots$ n_m have been defined so that $n_1 < n_2 < \ldots < n_m$, and $n_j \in A_j$ for each $j = 1, 2, \ldots m$. Then define n_{m+1} to be an element of A_{m+1} such that $n_{m+1} > n_m$. It is easy to see that the set $B = \{n_1, n_2, \dots n_m, \dots\}$ is an infinite subset of $\mathbb{N}$ such that for each m the set B is eventually a subset of A_m. Indeed, it is clear that, for each m in $\mathbb{N}$, the set $\{n_m, n_{m+1}, n_{m+2}, \dots\}$ is a subset of A_m.

Special Case Choose n_1 to be the *least* element of A_1, choose n_2 to be the *least* element of A_2 such that $n_2 > n_1$, and so on. These explicit choices are possible because of the Least-Natural-Number Principle.

The construction described in the preceding example is used frequently enough in analysis to deserve its own terminology.

I.9.12 Definition

Let $\mathcal{A} = (A_1, A_2, \dots A_m, \dots)$ be a subsequence structure of infinite order. An infinite subset $B = \{n_1 < n_2 < \dots < n_m < \dots\}$ of $\mathbb{N}$ is said to be a **cross section of $\boldsymbol{\mathcal{A}}$** provided that for each m in $\mathbb{N}$ one has $n_m \in A_m$.

If, in addition, the numbers n_m are chosen as in the Special Case in the preceding example, then the set B is called the **minimal cross section of $\boldsymbol{\mathcal{A}}$**.

Remarks (1) Let $\mathcal{A} = (A_1, A_2, \dots A_m, \dots)$ be a subsequence structure of infinite order, as above. Suppose that $B = \{n_1 < n_2 < \dots < n_m < \dots\}$ is the minimal cross section of $\mathcal{A}$, and that $C = \{l_1 < l_2 < \dots < l_m < \dots\}$ is an arbitrary cross section. It is an easy exercise to show that $n_m \leq j_m$ for all m.

(2) Suppose that the intersection of the sets A_m, for m in $\mathbb{N}$, is an *infinite* subset C of $\mathbb{N}$. Then certainly C is a cross section of $\mathcal{A}$, but it need not be the minimal cross section.

(3) Given a subsequence structure $\mathcal{A} = (A_1, A_2, \dots A_m, \dots)$, it may be possible to construct an infinite subset D of $\mathbb{N}$ which is eventually a subset of each set A_m, but which is *not* a cross section of $\mathcal{A}$; for instance, D might include a natural number which is smaller than the smallest element of A_1. However, it is an easy exercise to show that any such D has an infinite subset B which *is* a cross section of $\mathcal{A}$. In practice this means there is no loss by dealing only with cross sections of the given subsequence structure.

I.9.13 Example

Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be a sequence of real numbers such that ξ is unbounded above, in the sense that for every number M there exists at least one index k such that $x_k > M$.

Claim The sequence ξ has a subsequence $\zeta = (z_1, z_2, \dots z_m, \dots)$ such that $z_m > m$ for each index m.

Proof of Claim For each natural number m let A_m be the set of all indices k such that $x_k > m$. The hypothesis that the original sequence ξ is unbounded above clearly implies that each set A_m is an infinite subset of $\mathbb{N}$. Also it is clear that for each m one has $A_{m+1} \subseteq A_m$, so that $\mathcal{A} = \{A_1, A_2, \dots\}$ is an infinite-order subsequence structure. Let $B = \{k_1 < k_2, \dots < k_m < \dots\}$ be a cross section of $\mathcal{A}$, as described above. Set $z_m = x_{k_m}$ for each

m in $\mathbb{N}$. Then it is clear that the subsequence $\zeta = (z_1, z_2, \ldots z_m, \ldots)$ of ξ has the required property. Indeed, one has $z_m = x_{k_m}$. However, $k_m \in A_m$ by construction, so $x_{k_m} > m$, by the definition of the set A_m. That is, $z_m > m$ for each index m, as required.

Let us finish this subsection with a bit of terminology that will be useful later on.

I.9.14 Definition

A sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ is said to **eventally have a certain property** provided there exists a natural number N such that if $k \geq N$ then the subsequence $(x_{N+1}, x_{N+2}, \ldots x_{N+k}, \ldots)$ has the given property.

Example Let $\xi = (-1, 3, 4, 7, \ldots 7, 7 \ldots)$ whose first three terms are $x_1 = -1$, $x_2 = 3$, $x_3 = 4$, and whose terms x_k for $x \geq 4$ are all equal to 7. Then the sequence ξ is not constant, but it is *eventually* constant.

I.10 Binary Operations on a Set

In this section a new class of functions is introduced. Examples of such functions have been studied from ancient times, and new ones arise frequently in many branches of modern mathematics. The discussion of this topic is placed here because this type of function plays an important role in the next chapter.

I.10.1 Definitions

Let A be a nonempty set.

(1) Suppose that $f : A^2 \to A$ is a function whose domain is the Cartesian product of a nonempty set A with itself, and whose values are in A. Then one says that f is a **binary operation on *A***.

Note The word **operator** is often used in place of 'operation' in this context.

(2) A binary operation $f : A^2 \to A$ is said to be **commutative** if

$$f(x, y) = f(y, x) \text{ for all } x \text{ and } y \text{ in } A.$$

It is said to be **associative** if

$$f(f(x, y), z) = f(x, f(y, z)) \text{ for all } x,\ y \text{ and } z \text{ in } A.$$

I.10.2 Examples

(1) The 'Sum Function' $S : \mathbb{N}^2 \to \mathbb{N}$ and the 'Product Function' $P : \mathbb{N}^2 \to \mathbb{N}$, given by the rules

$$S(x, y) = x + y \text{ and } P(x, y) = x{\cdot}y \qquad \text{for all } (x, y) \text{ in } \mathbb{N}^2,$$

are familiar binary operations on $\mathbb{N}$. There are analogous binary operations of 'sum' and 'product' on $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$. It is well known that each one of these operations is both commutative and associative, although the 'S' and 'P' formulation may obscure this fact. For example, when the associative law for S, namely

$$S(S(x,y),z) \;=\; S(x,S(y,z)),$$

is reformulated using the plus sign, one gets the more familiar looking

$$(x+y)+z \;=\; x+(y+z)$$

(2) Closely related to sums and products in algebra are differences and quotients. Unfortunately, the situation now becomes less pleasant.

For example, one can define the 'difference' $D(x,y)$ for all (x,y) in $\mathbb{N}^2$, but the values of the function D need not be in $\mathbb{N}$. Since the values of a binary operator on a set A must all lie in A, it follows that this function is not a binary operator on $\mathbb{N}$. This is not an issue in $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$, so in each of these cases the well-known 'difference' function is a true binary operator. However, in each of these cases the operator is neither commutative nor associative.

In the case of 'quotients', once again there is the issue that the quotient of two elements of the set may fail to be in the given set. More serious is that in all these sets division by 0 is not allowed, so in none of these cases is there a true binary operator.

(3) In Example (1) above we saw binary operations which are both commutative and associative, while in Example (2) we saw binary operations which enjoy neither of these properties. It is also possible for a binary operation to have one of these properties but not the other.

(i) Define the so-called 'Absolute Difference' function $\hat{D} : \mathbb{Z}^2 \to \mathbb{Z}$ by the rule $\hat{D}(x,y) = |x-y|$. It is obvious that this binary operation on $\mathbb{Z}$ is commutative, and it is easy to check that it fails to be associative.

(ii) Let X be a nonempty set and let A be the set of all functions $f : X \to X$ with domain X and having values in X.

Define the **Composition Operation** C on A to be the binary operation $C : A^2 \to A$ given by the rule $C(f,g) = f{\circ}g$, where the latter expression is described in Definition (I.6.2). It is clear that the binary operation C is associative no matter which nonempty set X is used; see Remark (I.6.3) and Theorem (I.6.5). However, if the set X has more than one element, then C is not commutative.

Remark The preceding examples point out a notational difficulty associated with defining 'binary operations' as functions. For instance, the 'functional' notation for sums is $S(x,y)$, with the symbol S for ths operation located to the left of both numbers being summed. In contrast, the common notation for the sum of two numbers x and y is $x+y$, with the symbol for this operation, the 'plus' sign $+$, placed between the numbers. Roughly speaking, the difference between these two perfectly valid notational choices corresponds to the difference between pronouncing them as 'the sum of x and y' or 'x plus y'. Similarly, one can say 'the product of x and y', or 'x times y'; 'the difference between x and y', or 'x minus y; 'the composition of f with g,' or 'f circle g'.

Most people prefer to use the 'symbol-between-the-inputs' notation instead of the 'function' notation (e.g., $x + y$ instead of $S(x, y)$). The official name of this preferred notation is the **infix representation** of the binary operation. To accomodate this notational preference, it is common when discussing binary operations 'in the abstract' to choose a 'generic' symbol, such as $*$, to act as the (infix) symbol of a 'generic' binary operation $f : A^2 \to A$, much as the plus sign $+$ is the infix symbol of the operation S. With such a choice, one then normally writes $x{*}y$ instead of $f(x, y)$. Likewise, one often refers to 'the binary operation $*$' instead of the more proper usage, 'the binary operation f'. Note that with the infix notation the commutative and associative laws take the more familiar forms

$$x{*}y \;=\; y{*}x \text{ and } (x{*}y){*}z \;=\; x{*}(y{*}z), \text{ respectively.}$$

Expressions which involve repeated applications of a binary operator, but involving no parentheses, appear frequently in mathematics; for instance, in arithmetic one might encounter the expression $1-2-3$, which involves two applications of the 'difference' operation. (Try computing the value of this expression before reading further.)

More generally, if $*$ is a binary operator on a set A, the question arises as to how should one interpret a parenthesis-free expression of the form $x_1{*}x_2{*}x_3{*}\ldots{*}x_n$. Of course, the issue is that technically such expressions should be meaningless, since a binary operation can act on only a pair of objects.

For simplicity, first consider the case $n = 3$. There are two obvious interpretations of the expression $x_1{*}x_2{*}x_3$:

Interpretation 1 Carry out the operations successively 'from left to right'. More precisely, define $x_1{*}x_2{*}x_3$ to mean $(x_1{*}x_2){*}x_3$. That is, first perform the $*$ on the left, and then the $*$ on the right.

Interpretation 2 Carry out the operations successively 'from right to left'. More precisely, set $x_1{*}x_2{*}x_3 \;=\; x_1{*}(x_2{*}x_3)$.

It is clear that if the operation $*$ is associative, then these two interpretations are in agreement, but not so if $*$ is not associative.

Example Suppose that $*$ is the binary operation of 'subtraction' on the set $\mathbb{Z}$ of all integers. It is easy to check that $1 - 2 - 3 \;=\; -4$ under the 'left-to-right' computation, but $1 - 2 - 3 \;=\; 2$ under the 'right-to-left' computation. (Compare this with the value for this expression you obtained above.)

In *This Textbook* we follow the standard convention and use the 'left-to-right interpretation' in extending this discussion to expressions of the form $x_1{*}x_2{*}x_3{*}\ldots{*}x_n$. More precisely:

I.10.3 Definition

Let $f : A^2 \to A$ be a binary operation on a nonempty set A, with corresponding 'infix' symbol $*$. The **left-to-right extension of the operation $*$ to k-tuples**, where $k \geq 3$, is given recursively so that the following laws hold:

If $k \;=\; 3$ then $x_1{*}x_2{*}x_3 \;=\; (x_1{*}x_2){*}x_3$.

If $k \;>\; 3$, then $x_1{*}x_2{*}\ldots{*}x_k{*}x_{k+1} \;=\; (x_1{*}x_2{*}\ldots{*}x_k){*}x_{k+1}$.

Remarks (1) The corresponding 'right-to-left extension' is easy to define, but we do not use it in *This Textbook*.

(2) See the Side Comment after Example (I.6.4) for a possible explanation for preferring the 'left-to-right' interpretation.

Side Comments (on the representations of binary operations):

(1) There are some places in mathematics in which the 'right-to-left interpretation' of extensions of binary operations are in common use. For example, consider the usual 'exponentation' binary operator ^, defined on natural numbers by the rule x^$y = x^y$. This is often used in place of the standard 'raised exponent' notation in scientific calculators, where the use of raised symbols is inconvenient. For example, calculus texts typically interpret the expression 10^{2^3} to mean $10^{(2^3)} = 10^8$; that is, 10^2^3 = 10^(2^3), which is the 'right-to-left' extension of the operator ^. In contrast, scientific calculators would say that $10^{2^3} = (10^2)^3 = 10^{(2\cdot 3)} = 10^6$; that is, 10^2^3 = (10^2)^3, which is the left-to-right interpretation. As always, the easy way to avoid any ambiguity is to insert parentheses as needed.

(2) The 'infix representation' of a binary operation $f : A^2 \to A$ described above is so called because it locates ('fixes') the symbol $*$ for the operation between ('inside') the two inputs, as in $x{*}y$. In contrast, the functional representation for the same quantity places the symbol f for the operation before the inputs x and y: $f(x, y)$. This is essentially what computer scientists call the 'prefix representation' of the binary operation, except they would write $f\,x\,y$ in place of $f(x, y)$. They also use the so-called 'postfix representation': $x\,y\,f$. These notations are also known as the 'Polish' and 'Reverse Polish' notations, respectively, in honor of the Polish logician Jan Lukasiewicz, who introduced such notations as a way to completely avoid the use of parentheses in complicated expressions built out of binary operations. For example, the 'prefix' version of the expression $x + a\,(y - z\,(b + c))$ is $S\,x\,P\,a\,D\,y\,P\,z\,S\,b\,c$; the 'postfix' version is $b\,c\,S\,z\,P\,y\,D\,a\,P\,x$. Note that in each of these expressions the order in which the binary operations are evaluated is completely determined: in the prefix notation one evaluates the operations from right to left, while in the postfix notation the order is from left to right. (It is likely that this last fact explains why the postfix notation is more widely used than the prefix; see the Side Comment on the 'left-to-right bias'.)

The algebra for repeated applications of a binary operator becomes much simpler if the operator is associative. It becomes simpler still if the operator is also commutative. In what follows, all the expressions are defined as in Definition (I.10.3).

I.10.4 Theorem

Let $f : A^2 \to A$ be a binary operation on a nonempty set A, and let $*$ be the corresponding infix symbol of this operation.

(a) Assume that the operation $*$ is associative. Let k and m be natural numbers, and let $(x_1, x_2, \ldots x_k, x_{k+1}, \ldots x_{k+m})$ be an element of A^{k+m}. Then

$$(x_1{*}x_2{*}\ldots{*}x_k)*(x_{k+1}{*}x_{k+2}{*}\ldots{*}x_{k+m}) = x_1{*}x_2{*}\ldots{*}x_k{*}x_{k+1}{*}\ldots{*}x_{k+m}.$$

(b) Assume that $*$ is both commutative and associative. Then the the operation $*$ has a stronger version of commutivity:

Let k be a natural number, and let $p : \mathbb{N}_k \to \mathbb{N}_k$ be a permutation of the finite set $\mathbb{N}_k$ (see Definition (I.7.8)). Then for every k-tuple $(x_1, x_2, \ldots x_k)$ in A^k one has

$$x_{p(1)}*x_{p(2)}*\ldots*x_{p(k)} = x_1*x_2*\ldots*x_k.$$

Outline of Proof The details of the proof are left as an exercise. What follows are hints:

(a) Use mathematical induction on the index m.

(b) Use mathematical induction on the index k; break into the special cases $p(k) = k$ and $p(k) \neq k$. (I assume that you let the induction hypothesis be that the statement holds for $k-1$.)

I.10.5 Remark

The conclusion of Part (b) of the preceding theorem, namely that the binary operation $*$ enjoys the stronger form of 'commutivity' given there, does not remain true if the hypothesis that $*$ be associative is omitted. That is, ordinary commutivity of $*$ is not enough by itself to guarantee the stronger form of commutivity. This contrasts with the situation in Part (a) for which assuming ordinary associativity for $*$ does guarantee the stronger form of associativity.

The next definition is encountered more often in the algebraic branches of moderm mathematics than in analysis, but it does get used here too. It makes precise the concept of two binary operations being 'equivalent', in the sense that every statement which can be expressed purely in terms of the first operation corresponds to a fact about the second.

I.10.6 Definition (Isomorphism)

Let $(A, *)$ and $(A', *')$ be binary operations on the sets A and A', respectively.

(1) An **isomorphism of $(A, *)$ with $(A', *')$** is a bijection $\varphi : A \to A'$ of A onto A' such that $\varphi(a_1*a_2) = (\varphi(a_1))*'(\varphi(a_2))$ for each ordered pair (a_1, a_2) in A^2.

Note If, in a given context, the binary operations $*$ and $*'$ are clear, one often refers more briefly to 'the isomorphism $f : A \to A'$.

(2) Two binary operations are said to be **isomorphic (to each other)** if there exists an isomorphism of one with the other.

I.10.7 Examples

(1) Let $\varphi : \mathbb{Z} \to \mathbb{Z}$ be the bijection given by the rule $\varphi(k) = -k$ for each integer k. It follows from the equation $-(k+m) = (-k)+(-m)$, valid for integers, that φ is an isomorphism of the operation of addition on $\mathbb{Z}$. However, since $-(1{\cdot}1) = -1$, while $(-1){\cdot}(-1)) = 1$, the map φ is *not* an isomorphism of multiplication of integers.

(2) Let $A = \mathbb{N}$, and let $A' = \{2, 2^2, \ldots 2^n, \ldots\}$, the set of all positive powers of 2. Let $*$ denote the usual binary operation of 'addition of natural numbers' on A. Likewise,

let $*'$ denote the restriction to the set A' of the usual binary operation of 'multiplication of natural numbers'; it is clear from the standard laws of powers of natural numbers that $*'$ is a binary operation on the set A'. Let $\varphi : A \to A'$ be given by the rule $\varphi(k) = 2^k$. It follows from the usual rules for exponents that φ is an isomorphism of $(A, *)$ with $(A', *')$.

(3) It is an easy exercise to show that if $\varphi : A \to A'$ is an isomorphism of $(A, *)$ with $(A', *')$, then $\varphi^{-1} : A' \to A$ is an isomorphism of $(A', *')$ with $(A, *)$.

(4) Suppose $\varphi : A \to A'$ is a bijection of the set A onto the set A'. If $(A', *)$ is a binary operation on the set A', then there is a unique binary operation $(A, *)$ on A for which φ is an isomorphism. It is given by the rule

$$a_1 * a_2 = \varphi^{-1}[\varphi(a_1) *' \varphi(a_2)] \text{ for each ordered pair } (a_1, a_2) \text{ in } A^2.$$

It is called the **binary operation induced on** A **by the bijection** φ **from the binary operation** $(A', *')$. Intuitively speaking, one uses the bijection φ to relate the pair (a_1, a_2) in A^2 to the corresponding pair (a_1', a_2') in A'^2; then one uses the operation $*'$ to get a third element $a_3' = a_1' *' a_2'$ in A'. Finally, $a_1 * a_2$ is the element a_3 in A which corresponds under the bijection φ to the element a_3'.

The significance of φ being an isomorphism of the binary operation $(A, *)$ with $(A', *')$ is that these operations then have exactly the same algebraic properties. More precisely, if an equation involving only the operation $*$ and certain elements a_1, $a_2, \ldots$ a_n of A is valid, then the equation one obtains by replacing each occurance of $*$ by $*'$, and each occurance of a_j by $a_j' = \varphi(a_j)$ for each $j = 1, 2, \ldots n$, is also valid. For example, if $(A, *)$ is an associative operator, then so is $(A', *')$. Or if there exists an element u in A such that $u * a = a$ for each a in A, then $u' *' a' = a'$ for all a' in A', where $u' = \varphi(u)$.

Note that the objects in sets A and A' need not be of the same type for there to exist an isomorphism.

Example Let $A = \{0, 1\}$ and let $A' = \{c, d\}$; thus A consists of natural numbers, while A' consists of letters, objects of different type. Define binary operations $(A, *)$ and $(A', *')$ as follows:

$$a_1 * a_2 = 0 \text{ for all } a_1, a_2 \text{ in } A \ ; \qquad a_1' *' a_2' = c \text{ for all } a_1', a_2' \text{ in } A'$$

It is clear that the map $\varphi : A \to A'$ given by the rule $\varphi(0) = c$, $\varphi(1) = d$ is an isomorphism.

The usual way to prove that two binary operations are isomorphic is to produce an isomorphism of one with the other, as in the preceding example. In contrast, the usual way to prove that two such operations are *not* isomorphic is to find an example of an algebraic property which holds for one but not the other. For example, the 'absolute difference' operator on $\mathbb{Z}$ is commutative, but the 'difference' operatior on $\mathbb{Z}$ is not. It follows that these binary operators are not isomorphic to each other.

I.11 EXERCISES FOR CHAPTER I

I - 1 Prove Parts (b) and (c) of Theorem (I.2.9).

I - 2 Prove Parts (d) and (e) of Theorem (I.2.9).

I - 3 Prove Parts (f), (g) and (h) of Theorem (I.2.9)

I - 4 Prove or disprove the following statements:

(a) For all sets A, X and Y, one has $(A - X) \cap (A - Y) = A - (X \cap Y)$.

(b) For all sets A, X and Y, one has $(A - X) \cup (A - Y) = A - (X \cup Y)$.

(c) Same as (a), but with 'all sets' replaced by 'some sets'.

(d) Same as (b), but with 'all sets' replaced by 'some sets'.

I - 5 The Kuratowski definition of 'ordered pair' (see Definition (B.1.1)) is only one of several set-theoretic definitions of this concept which have been proposed over the years. Of course the goal of all these definitions is that they should allow one to distinguish between the 'first' element x and the 'second' element y of the pair (x, y) purely in terms of set theory. Determine which, if any, of the following definitions of the ordered pair (x, y) satisfies this goal. (Explain your answer.)

(a) $(x, y) = \{x, y\}$

(b) $(x, y) = \{\{x, 1\}, \{y, 2\}\}$

(c) $(x, y) = \{\{\{x\}, \emptyset\}, \{\{y\}\}\}$

(d) $(x, y) = \{x, \{y\}\}$

I - 6 There are two 'obvious' ways to define the concept of 'ordered triple' using the Kuratowski definition of 'ordered pair':

Method (1) Define (x, y, z) to equal $((x, y), z)$.
Method (2) (x, y, z) to equal $(x, (y, z))$.

Determine whether these definitions give the same value for (x, y, z).

I - 7 Let X be a nonempty set. Recall that a map $f : X \times X \rightarrow X$ is called a *binary operation on* X. Such an operation is said to be *commutative* if $f(x_1, x_2) = f(x_2, x_1)$ for all x_1, x_2 in X. It is said to be *associative* if $f(f(x_1, x_2), x_3) = f(x_1, f(x_2, x_3))$ for all x_1, x_2, x_3 in X.

Problem: Show that if f is an associative and commutative binary operation on X, then

$$f(f(x_1, x_2), f(x_3, x_4)) = f(f(x_3, x_2), f(x_1, x_4)) \text{ for all } x_1, x_2, x_3, x_4 \text{ in } X.$$

(Hint: Try it out on some familiar case first.)

I - 8 Suppose that f is a function whose domain is a set A and whose values lie in a set B.

(a) Prove the following statement: If X is a subset of A and Y is a subset of B, then

$$X \subseteq f^{-1}[f[X]] \text{ and } Y \supseteq f[f^{-1}[Y]].$$

Also, determine whether it could happen that X is a *proper* subset of $f^{-1}[f[X]]$; likewise, determine whether it could happen that Y is a *proper* superset of $f[f^{-1}[Y]]$.

(b) Let X_1 and X_2 be subsets of A. Prove the following:
(i) $f[X_1 \cup X_2] = f[X_1] \cup f[X_2]$ and $f[X_1 \cap X_2] \subseteq f[X_1] \cap f[X_2]$.
(ii) Either find an example of f, A, B, X_1 and X_2 for which one has $f[X_1 \cap X_2] \neq f[X_1] \cap f[X_2]$, or else prove that no such example exists.

(c) Let Y_1 and Y_2 be subsets of B. Prove the following:
(i) $f^{-1}[Y_1 \cup Y_2] = f^{-1}[Y_1] \cup f^{-1}[Y_2]$ and $f^{-1}[Y_1 \cap Y_2] = f^{-1}[Y_1] \cap f^{-1}[Y_2]$.
(ii) $f^{-1}[B \setminus Y] = A \setminus f^{-1}[Y]$.

I - 9 Let $f : X \to Y$ be a function from a set X into a set Y. Recall that one says that a function $g : Y \to X$ is a **left inverse of** f provided $g \circ f = I_X$. Likewise, one says that a function $h : Y \to X$ is a **right inverse of** f provided $f \circ h = I_Y$.

(a) Show that a necessary and sufficient condition for f to be a bijection of X onto Y is that f admit both a left inverse $g : Y \to X$ and a right inverse $h : Y \to X$. In addition, if this occurs then $g = h = f^{-1}$.

(b) Prove or Disprove If $f : X \to Y$ has at least one left inverse $g : Y \to X$ but no right inverse, then f has more than one such left inverse.

I - 10 Prove or Disprove: If f and g are maps from a set X onto X, and if the composition $f \circ g : X \to X$ is one-to-one, then f and g are also one-to-one. (Compare with Part (e) of Theorem (I.6.5).)

I - 11 Let X be a set which has exactly k elements, where k is a natural number. Let Y be the set of all bijections of X onto itself. Prove that Y has exactly $k!$ elements.

I - 12 Let W be a set, and let A be the set of all subsets of W; that is, A is the power set $\mathcal{P}(W)$ of W.

Definition: A function $f : A \to \mathbb{R}$ is said to be an **additive function on** A provided $f(X_1 \cup X_2) = f(X_1) + f(X_2)$ for every pair of mutually disjoint subsets of W.

(a) Suppose that W and A are as above, and that $f : A \to \mathbb{R}$ is an additive real-valued function defined on A. Prove that if X_1 and X_2 are *arbitrary* subsets of W then

$$f(X_1 \cup X_2) = f(X_1) + f(X_2) - f(X_1 \cap X_2).$$

(b) Suppose that $W = \{1, 2\}$, so that W has exactly two elements. Determine all of the corresponding additive functions.

I - 13 (a) Prove Part (b) of Theorem (I.7.7). You may use the result of Part (a) of that theorem.

(b) Prove Part (c) of Theorem (I.7.7). You may use the results of Parts (a) and (b) of that theorem.

(c) Prove Part (d) of Theorem (I.8.4). You may assume the results of Parts (a), (b) and (c).

I - 14 Determine whether there exists a nonempty binary relation R on some set X such that R satisfies both symmetry and transitivity, but not reflexivity.

I - 15 Let $P_1 = (-3, 2)$ and $P_2 = (2, 5)$. Determine the value at $x = 1$ of the linear interpolation between P_1 and P_2.

I - 16 Prove or Disprove: Let A and B be nonempty sets. If $f : A \to B$ and $g : B \to A$ are surjections, then A and B have the same cardinality.

I - 17 In each part of this exercise, find an explicit example of a bijection of X with Y. ('Explicit': Give a formula using standard functions that would be familiar to students in elementary calculus. You may use the well-know properties of such functions, but make it clear which such properties you are using.)

(a) $X = [3, 5]$, $Y = [2, 10]$.

(b) $X = \mathbb{R}$, $Y = (-\pi, +\infty)$.

(c) $X = \mathbb{R}$, $Y = (-1, 1)$.

I - 18 Give a direct proof the following version of the 'Amazing Grace' effect: If X is an infinite set, and B is any finite subset of X, then the set $X\backslash B$ has the same cardinality as X.

NOTE: By a 'direct proof', it is meant that you are not allowed to use the conclusions of Corollary (I.8.6) or Theorem (**??**) in your solution.

I - 19 In Set Theory one often encounters the following notation: Let X and Y be nonempty sets. Then Y^X is the set of all functions $f : X \to Y$ with domain X and with values in Y.

(a) Show that if $Y = \{a, b\}$ is a doubleton set, so that $a \neq b$, then Y^X has the same cardinality as $\mathcal{P}(X)$, the power set of X.

(b) Prove or Disprove If $Y = \{a, b, c\}$ is a set with exactly three distinct elelments a, b and c, and X is an infinite set, then Y^X has the same cardinality as $\mathcal{P}(X)$.

I - 20 Prove Parts (b) and (c) of Theorem (I.7.7).

I - 21 Let X and Y be nonempty sets, and suppose that $F : X \to Y$ is a bijection of X onto Y.

(a) Let $\beta : Y \times Y \to Y$ is a binary operation on Y. Show that there exists a unique binary operation $\alpha : X \times X \to X$ such that

$$F(\alpha(x_1, x_2)) = \beta(F(x_1), F(x_2))$$

(b) Let α and β be as in Part (a). Prove that α has an identity element if, and only if, β has an identity element.

(c) Let α and β be as in Part (a). Prove that α satisfies the associative law if, and only if, β satisfies the associative law.

I - 22 Suppose that X and Y are finite sets. Prove that $X \times Y$ is a finite set, and express $\#(X \times Y)$ in terms of $\#(X)$ and $\#(Y)$.

I - 23 Definition A real number c is said to be an**algebraic number** provided there exists a polynomial function of the form $p(t) = t^k + a_{k-1}t^{k-1} + \ldots + ta_1 + a_0$, where k is a natural number and the coefficients a_{k-1}, $a_{k-2}, \ldots$ a_1, a_0 are rational numbers, such that $p(c) = 0$. A real number which is not algebraic is said to be a **transcendental number**.

Problem Prove that the set of algebraic real numbers is a countably infinite set. You may use – without needing to prove – standard facts from high-school algebra concerning the roots of polynomials. NOTE: In light of Corollary (**??**), it follows from the conclusion of this problem that in every interval $[a, b]$ in $\mathbb{R}$ there exist uncountably many transcendental numbers.

I - 24 Let X and Y be sets. Let S be a subset of X, and let T a subset of Y.

Prove or Disprove If X and Y have the same cardinality as each other, and S and T have the same cardinality as each other, then $X \backslash S$ and $Y \backslash T$ have the same cardinality as each other.

Chapter II

Axioms for the Real Number System

Quotes for Chapter (II):

> (1) ‘Familiarity breeds contempt.’
> (From Aesop’s *Fable of The Fox and the Lion*)
>
> (2) ‘If you should put even a little on a little, and should do this often, soon this too would become big.’
> (From Hesiod’s ‘*Works and Days*’, c. 700 BC)
>
> (3) ‘How often have I said to you that when you have eliminated the impossible whatever remains, *however improbable*, must be the truth?’
> (Spoken by Sherlock Holmes to Dr. Watson in *The Sign of the Four*, by Sir Arthur Conan Doyle.)
>
> (4) ‘Chacun à son goût’
> (Title of an aria in the operetta ‘Die Fledermaus (The Bat)’, by Johann Strauss. It translates roughly as ‘To each according to his own taste’.)

Introduction

In Chapter (I) it is assumed that the reader is already familiar with the natural numbers, the integers, and the rational numbers, as well as their main algebraic properties. Appendix A provides the interested reader with a deeper study of the foundations of those numbers; for example, it provides an axiomatic treatment of the natural numbers using the Dedekind-Peano axioms for $\mathbb{N}$, and develops the the integers and the rational numbers, and many of their properties, from the properties of the natural numbers. However, reading that treatment is not needed to understand the main body of *This Textbook* if one simply accepts the basic properties of these numbers.

Likewise, in Chapter (I) it is assumed that the reader is already familiar with basic properties of the system of *real* numbers. However, experience shows that to do analysis rigorously one needs to discuss the properties of real numbers in greater depth. The main goal of the present chapter is to provide an axiometic treatment of the real-number system. This is a list of primitive statements (‘axioms’) about real numbers that everyone will normally accept as being true, without demanding further proof; other properties of real numbers are then deduced rigorously from these axioms. For the interested reader, Appendix C outlines a rigorous development of the real numbers, in terms of the rational numbers, from which the axioms given below can be deduced from the known properties of the rational numbers. One does not need to read that Appendix in order to understand the main body of *This Textbook.*

The properties of the Real Number System to be developed in this chapter split naturally into three classes: the 'Algebraic Properties', the 'Order Properties', and the 'Completeness Property'. We devote a separate section to each class of properties.

II.1 Algebraic Properties of the Real Numbers

The set $\mathbb{R}$ of real numbers comes with a pair of well-known binary operations, called 'addition' and 'multiplication'. In terms of Definition (I.10.1), these are functions $\mathrm{Add}_{\mathbb{R}} : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and $\mathrm{Mult}_{\mathbb{R}} : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$. We follow convention and denote the infix representations of these operations by using the symbols $+$ and $\cdot$, respectively:

$$\mathrm{Add}_{\mathbb{R}}(x,y) = x+y, \quad \mathrm{Mult}_{\mathbb{R}}(x,y) = x\cdot y \text{ for all } x,y \text{ in } \mathbb{R}.$$

Note that we assume that the reader is already familiar with these operations; we do not further define them here.

We assume that the binary operations $+$ and $\cdot$ satisfy the following **Algebraic Axioms for $\mathbb{R}$**:

(**Axiom A0**) (The Closure Laws for Addition and Multiplication) If x and y are real numbers, then

$$x+y \text{ and } x\cdot y \text{ are both real numbers.}$$

The real numbers $x+y$ and $x\cdot y$ are called the **sum of x with y** and the **product of x with y**, respectively.

(**Axiom A1**) (The Commutative Laws for Addition and Multiplication) If x and y are real numbers, then

$$x+y = y+x \text{ and } x\cdot y = y\cdot x.$$

(**Axiom A2**) (The Associative Laws for Addition and Multiplication) If x, y and z are real numbers, then

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

Axiom A3 (The 'Identity Element' Laws for Addition and Multiplication) There is a unique real number, denoted 0, such that

$$0+x = x+0 = x \text{ for every real number } x.$$

Likewise, there is a unique real number, denoted by 1, such that

$$1\cdot x = x\cdot 1 = x \text{ for every real number } x.$$

The numbers 0 and 1 are called the **additive identity** and the **multiplicative identity**, respectively.

Axiom A4 (The 'Inverse Element' Laws) For each real number x there is a unique real number u such that

$$x+u = u+x = 0$$

This unique u is called the **negative of** x, and is denoted by $-x$; thus the preceding equation takes the form

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

(As usual, 0 is the additive identity described in Axiom (4).)

Likewise, for each real number y such that $y \neq 0$, there is a unique real number v such that

$$y{\cdot}v \;=\; v{\cdot}y \;=\; 1.$$

This unique v is called the **reciprocal of** y, and is denoted $\dfrac{1}{y}$; thus the preceding equation takes the form

$$y{\cdot}\left(\frac{1}{y}\right) \;=\; \left(\frac{1}{y}\right){\cdot}y \;=\; 1.$$

(As usual, in this axiom 0 and 1 are the additive and multiplicative identities, respectively, described in Axiom (4).)

Note: Some texts refer to $-x$ as the **additive inverse of** x. Likewise, they often refer to $1/y$ as the **multiplicative inverse of** y; they also may use the 'exponent' notation y^{-1} in place of $1/y$.

Axiom A5 (The Distributive Laws for Addition and Multiplication) If x, y and z are real numbers, then

$$x{\cdot}(y+z) \;=\; (x{\cdot}y)+(x{\cdot}z) \text{ and } (x+y){\cdot}z \;=\; (x{\cdot}z)+(y{\cdot}z).$$

Axiom A6 (The Nontriviality Law) The numbers 0 and 1 described in Axiom A4 are distinct. That is,

$$1 \neq 0$$

II.1.1 Remarks

(1) Axioms A0–A6 imply that the real number system, with its usual notions of 'addition' and 'multiplication', forms what in modern abstract algebra is called an **algebraic field**, or, more briefly, a **field**. For that reason, we refer to these axioms as the **field axioms for** $\mathbb{R}$.

(2) The field axioms A0–A6 do not come close to completely characterizing the real number system. For example, if in these axioms one replaces the phrase 'real number' throughout by 'rational number', and if now $+$ and $\cdot$ denote the usual operations on rational numbers, and 0 and 1 denote the usual elements of $\mathbb{Q}$ with these symbols, then it is clear that the resulting axioms are satisfied by $\mathbb{Q}$; that is, $\mathbb{Q}$ is also a field. Of course $\mathbb{Q}$ and $\mathbb{R}$ are very different fields; for example, the set $\mathbb{Q}$ is countable but $\mathbb{R}$ is not, and $\mathbb{R}$ has an element x such that $x^2 \;=\; 2$ but $\mathbb{Q}$ does not.

Indeed, the field axioms for $\mathbb{R}$ are not even strong enough, by themselves, to show that $\mathbb{R}$ is an infinite set. For example, suppose that one replaces the phrase 'real number' in these axioms by the 'element of the doubleton set $\{0,1\}$', and define the binary operations $+$ and $\cdot$ on this set as follows:

$$0+0 \;=\; 1+1 \;=\; 0, \quad 0+1 \;=\; 1+0 \;=\; 1; \quad 0{\cdot}0 \;=\; 0{\cdot}1 \;=\; 1{\cdot}0 \;=\; 0,\; 1{\cdot}1 \;=\; 1.$$

Then the resulting system is also a field, but it has only two elements. Many authors denote this field by the symbol $\mathbb{Z}_2$, pronounced '$\mathbb{Z}$-mod 2'. Note that in this field one has $-1 \;=\; 1$, since, by definition, $1+1 \;=\; 0$.

(3) The field axioms refer to the special elements 0 and 1 and the operations $+$ and $\cdot$ on $\mathbb{R}$. These symbols are widely used in algebra; but, as is clear from the preceding remark, their meanings can vary depending on the context. For example, the roles of the symbols $+$, 0 and 1 in the context of $\mathbb{R}$ are very different from their roles in the context of the two-element field $\mathbb{Z}_2$: one has $1+1 \neq 0$ in the former case but $1+1 \;=\; 0$ in the latter.

Likewise, in the preceding remark one finds the familiar equation $x^2 = 2$. A subtle point about this equation is that on the left side the symbol '2' refers to the number of factors in the product $x{\cdot}x$; that is, this '2' is a natural ('counting') number. In contrast, the number '2' appearing on the right side is shorthand for the quantity $1+1$, with the '1' and '+' referring to the field in question, namely $\mathbb{Q}$ or $\mathbb{R}$.

If it might be unclear which role such a symbol is playing in a given context, one should do whatever is needed to make that role clear. For instance, one could write $1_{\mathbf{R}}$, $1_{\mathbb{Q}}$, $+_{\mathbf{R}}$ $2_{\mathbf{N}}$, etc for the various roles played by the symbol 1; or one could refer in words to 'the real multiplicative unit 1', etc. See Remark (II.1.6) (3) below for further discussion on this issue.

<u>Side Comments</u> (on alternate axioms for $\mathbb{R}$)

(1) Many authors express the axioms for $\mathbb{R}$ slightly differently. For example, some write the 'Additive Identity' law (see Axiom A3) as '$x + 0 = x$', and leave it as an exercise for the reader to show that $0 + x = x$ follows from it in combination with the Additive Commutative Law. The advantage of that approach is that it provides a list of axioms which are weaker; that is, the axioms assume (slightly) less. Many authors find such 'weakness' to be an aesthetically pleasing feature of an axiom system, and prefer a list of axioms that assume as little as possible. The disadvantage, of course, is that the reader pays for the author's aesthetic pleasure by needing to prove some minor details from these weaker axioms, such as $0+x = x$. This merely adds unnecessary clutter to the discussion.

In contrast, some authors prefer to use axioms that are even *stronger* than the ones given above. For instance, they might use, in place of our Axiom A4, the following statement:

For every pair of real numbers x and z, there exists a unique real number u such that $x+u = z$; and for every pair y and w of real numbers with $y \neq 0$, there exists a unique v such that $y{\cdot}v = w$.

Clearly these new statements reduce to our Axiom A4 in the special cases $z = 0$ and $w = 1$, respectively. Note that these new statements can be combined into a single statement:

For every triple of real numbers a, b and c with $a \neq 0$, there is a unique real number x such that $a{\cdot}x + b = c$.

(2) Most authors weaken Axioms A3 and A4 by omitting the word 'unique' from the statements; the uniqueness is later deduced using the weaker statements of these axioms together with the earlier axioms. The problem with omitting the word 'unique, however, is what usually results is a pair of axioms that display a serious ambiguity. For example, the 'additive' portions of these axioms are normally written in the following revised form:

A3′ There is a real number, denoted 0, such that

$$0+x = x+0 = x \text{ for every real number } x.$$

A4′ For each real number x there is a real number u such that

$$x+u = u+x = 0$$

It is not assumed in A3′ that the number 0 described here is unique; indeed, the phrase 'a real number' really means 'at least one real number'. This fact then makes the meaning of A4′ ambiguous: it is not clear whether A4′ claims that there is a single u which works simultaneously for each 0 satisfying the first statement, or that for each such 0 there exists a corresponding u, depending on the choice of 0, which works. Of course if the first statement were to be followed immediately by the proof that the quantity 0 is unique, the ambiguity would disappear. Unfortunately, most authors do not prove this uniqueness until after the second statement.

The simplest way around this issue is to simply include 'uniqueness' in the axiom itself, as we do in our Axioms A3 and A4 above. This has the added benefit of reducing 'clutter': one does not need to separately state, then prove, this uniqueness.

(3) The two Distributive Laws expressed in Axiom A5 are usually written

$$x{\cdot}(y+z) \;=\; x{\cdot}y + x{\cdot}z \text{ and } (x+y){\cdot}z \;=\; x{\cdot}z + y{\cdot}z;$$

note that in this version the right sides of the equations are missing the parentheses which appear in Axiom A5 above. This formulation is acceptable because there are notational conventions in high-school algebra for the 'order of precedence' of algebraic operations. For instance, in the expression $a{\cdot}b+c$, it is understood, because of these conventions, that the multiplication is carried out first, then the addition. That is,

$$a{\cdot}b+c \;=\; (a{\cdot}b)+c, \text{ NOT } a{\cdot}(b+c).$$

Unless there is a need for extra clarity, we normally follow such conventions in order to minimize the use of parentheses.

(4) Axiom A6 is sometimes replaced by the statement 'There is more than one real number'. It is easy to see that this formulation, in conjunction with the other axioms, is equivalent to Axiom A6.

For the rest of this section everything is formulated in terms of the field $\mathbb{R}$ of real numbers, since that is the most important example for us. However, since only the field axioms are used, the results (and proofs) are valid for any field.

II.1.2 Definition

(1) If x and y are real numbers, then the **difference between** x **and** y, denoted $x-y$, is the number $x+(-y)$. The function $D:\mathbb{R}{\times}\mathbb{R}\to\mathbb{R}$ given by the rule $D(x,y) \;=\; x-y$ is called the **difference function**, or the **subtraction**.

(2) If x and y are real numbers, and $y\neq 0$, then the **quotient of** x **by** y, denoted $\dfrac{x}{y}$ or x/y, is the number $x{\cdot}\dfrac{1}{y}$. The function $Q:\mathbb{R}{\times}(\mathbb{R}\backslash\{0\})\to\mathbb{R}$ given by the rule $Q(x,y) \;=\; x/y$ is called the **quotient function**, or **division**.

The next theorem states several familar algebraic facts; indeed, some are so familiar that it may seem that they are as obvious as the axioms, so that no proofs should be needed. However, the purpose for axiomatizing a subject is to list a small number of facts that can be taken as true (the axioms), and then to deduce all other facts from them. The results given here are to be deduced from the field axioms A0–A6.

II.1.3 Theorem

(a) If x is a real number then $0{\cdot}x \;=\; x{\cdot}0 \;=\; 0$, $-x \;=\; (-1){\cdot}x$ and $-(-x) \;=\; x$. In particular, $(-1){\cdot}(-1) \;=\; 1$.

(b) If x and y are real numbers then $-(x{\cdot}y) \;=\; (-x){\cdot}y \;=\; x{\cdot}(-y)$, and $(-x){\cdot}(-y) \;=\; x{\cdot}y$.

(c) If y is a real number such that $y\neq 0$, then $1/y\neq 0$, and $1/(1/y) \;=\; y$.

(d) If y and z are real numbers such that $y\neq 0$ and $z\neq 0$, then $y{\cdot}z\neq 0$. Furthermore, $\dfrac{1}{y{\cdot}z} \;=\; \left(\dfrac{1}{y}\right)\cdot\left(\dfrac{1}{z}\right)$.

(e) Suppose that x, y and z are real numbers, with $z \neq 0$. Then

$$\frac{x+y}{z} = \frac{x}{z} + \frac{y}{z}$$

Likewise,

$$\frac{x{\cdot}y}{z} = \left(\frac{x}{z}\right){\cdot}y = x{\cdot}\left(\frac{y}{z}\right).$$

(f) If x, y and z are real numbers then

$$z - x = (z - y) + (y - x).$$

(g) If x and y are real numbers such that $x{\cdot}y = 0$, then either $x = 0$ or $y = 0$.

Proof

(a) Note that

$$0{\cdot}x \overset{(1)}{=} (0+0){\cdot}x \overset{(2)}{=} 0{\cdot}x + 0{\cdot}x$$

Indeed, Equation (1) reflects the fact that $0 = 0 + 0$ (Axiom A3), while Equation (2) comes from the Distributive Laws (Axiom A5). Now add $-(0{\cdot}x)$ to both sides of the equation $0{\cdot}x = 0{\cdot}x + 0{\cdot}x$ to obtain

$$0 \overset{(3)}{=} 0{\cdot}x + (-(0{\cdot}x)) \overset{(4)}{=} (0{\cdot}x + 0{\cdot}x) + (-(0{\cdot}x)) \overset{(5)}{=} 0{\cdot}x + (0{\cdot}x + (-(0{\cdot}x))) \overset{(6)}{=} 0{\cdot}x + 0 \overset{(7)}{=} 0{\cdot}x.$$

In the preceding string of equalities, Equations (3) and (6) use Axiom A4; Equation (4) reflects the result of adding $-(0{\cdot}x)$ to both sides of the previously obtained equation $0{\cdot}x = 0{\cdot}x + 0{\cdot}x$; Equation (5) comes from the Associative Law for Addition; and Equation (7) uses Axiom A3. The desired result $0 = 0{\cdot}x$ now follows easily. The fact that $0 = x{\cdot}0$ then follows from the preceding together with the Commutative Law for Addition.

Next notice that

$$0 \overset{(1)}{=} 0{\cdot}x \overset{(2)}{=} (1 + (-1)){\cdot}x \overset{(3)}{=} 1{\cdot}x + (-1){\cdot}x \overset{(4)}{=} x + (-1){\cdot}x.$$

In this last string of equalities, Equation (1) follows from the result just proved above, Equation (2) comes from the fact that $1 + (-1) = 0$ (i.e., Axiom A4), Equation (3) uses the Distributive Law, and Equation (4) uses the equation $1{\cdot}x = x$ (i.e., Axiom A3). However, Axiom A4 states that there is only one element of $\mathbb{R}$ which when added to x yields the value 0, namely $-x$. Since, as has just been shown, $(-1){\cdot}x$ also has this property, it follows that $(-1){\cdot}x = -x$, as claimed.

Similarly, note that, by Axioms A4 and A1 one has

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

The uniqueness of the additive inverse of $-x$ guaranteed by Axiom A4 then implies that $x = -(-x)$, as claimed.

Finally, apply the preceding results to the case $x = -1$ to obtain

$$1 = -(-1) = (-1){\cdot}(-1),$$

as claimed.

(b) Notice that

$$-(x{\cdot}y) = (-1){\cdot}(x{\cdot}y) = ((-1){\cdot}x){\cdot}y = (-x){\cdot}y.$$

In this string of equalities one uses the fact that $-u = (-1){\cdot}u$, proved in Part (a), for two values of u; namely $u = x{\cdot}y$ and $u = x$; and one uses the Associative Law for Multiplication. A similar argument shows that $-(x{\cdot}y) = x{\cdot}(-y)$.

Finally, note that by the Associative and Commutative Laws for Multiplication, together with the results of Part (a) and Axiom A3, one gets

$$(-x){\cdot}(-y) = ((-1){\cdot}x){\cdot}((-1){\cdot}y) = (((-1){\cdot}x){\cdot}(-1)){\cdot}y =$$

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

as claimed.

(c) Notice that if $1/y$ were to equal 0 then it would follow from Part (a) that $y{\cdot}(1/y) = y{\cdot}0 = 0$. However, by Axiom A4 one has $y{\cdot}(1/y) = 1$, so it would then follow that $0 = 1$. Since this contradicts Axiom A6, one sees that $1/y$ cannot equal 0. Since this is the case, Axiom A4 then implies that $1/y$ also has a unique multiplicative inverse; that is, there is a unique real number u such that $(1/y){\cdot}u = 1$. However, by the definition of $1/y$, together with the Commutative Law, one has

$$1 = y{\cdot}\left(\frac{1}{y}\right) = \left(\frac{1}{y}\right){\cdot}y,$$

so that y and u must be the same. That is, y is the reciprocal of $1/y$, so that

$$y = \frac{1}{(1/y)},$$

as claimed.

(d), (e), (f) and (g): The proofs of these parts are left as exercises.

Definition (I.10.3) describes the 'left-to-right extension', to A^k, of a general binary operation $*$ defined on a set A. Theorem (I.10.4) then states properties enjoyed by this extension if the operation $*$ is associative or, better yet, associative and commutative. Since, by Axioms A1 and A2 above, the binary operations '$+$' and '$\cdot$', addition and multiplication of real numbers, are both commutative and associative on $\mathbb{R}$, those earlier results apply to them as well. It is convenient to reformulate those results here, and even to expand them slightly, in terms of the $+$ and $\cdot$ notation.

II.1.4 Definition

Let k be a natural number such that $k \geq 3$, and let $(x_1, x_2, \ldots x_k)$ be an ordered k-tuple of real numbers.

(1) The **(finite) ordered sum** $x_1+x_2+\ldots+x_k$ associated with this ordered k-tuple is defined recursively by the rule

$$x_1 + x_2 + \ldots x_{k-1} + x_k = (x_1 + x_2 + \ldots + x_{k-1}) + x_k.$$

The parenthetical adjective 'finite' is included here because in Chapter (IX) an analogous concept of *infinite* ordered sum is introduced. Until then, however, usually the word 'finite' is omitted.

(2) In a similar manner, the **(finite) ordered product** $x_1{\cdot}x_2{\cdot}\ldots{\cdot}x_{k-1}{\cdot}x_k$ is defined by the rule

$$x_1{\cdot}x_2{\cdot}\ldots{\cdot}x_{k-1}{\cdot}x_k = (x_1{\cdot}x_2{\cdot}\ldots{\cdot}x_{k-1}){\cdot}x_k$$

(3) Special Cases: If $x_1 = x_2 = \ldots = x_k = c$, then often one writes $k\,c$ in place of the expression $x_1 + x_2 + \ldots + x_k$. Likewise, often one writes c^k in place of $x_2{\cdot}x_2{\cdot}\ldots{\cdot}x_k$.

(4) Frequently the expression $x_1 + x_2 + \ldots x_{k-1} + x_k$ is written more briefly as, say, $\sum\limits_{i=1}^{k} x_i$; the symbol $\sum$ is the upper-case Greek letter 'Sigma', which corresponds to the 'S' sound (as in '**S**um').

Likewise, one can write the expression $x_1{\cdot}x_2{\cdot}\ldots{\cdot}x_{k-1}{\cdot}x_k$ more briefly as $\prod\limits_{i=1}^{k} x_i$; the symbol $\prod$ is the upper-case Greek letter 'Pi', which corresponds to the 'P' sound (as in '**P**roduct').

The reformulation of Theorem (I.10.4) in terms of addition and multiplication of real numbers is easy to carry out. The result is the following extension of the commutative and associative laws for addition and multiplication.

II.1.5 Theorem (Extended Associative/Commutative Laws in $\mathbb{R}$)

(a) Let k and m be natural numbers. Then for every ordered k-tuple $(x_1, x_2, \ldots x_k)$ in $\mathbb{R}^k$ and every ordered m-tuple $(x_{k+1}, x_2, \ldots x_{k+m})$ in $\mathbb{R}^m$ one has

$$(x_1 + x_2 + \ldots + x_k) + (x_{k+1} + x_{k+2} + \ldots + x_{k+m}) = x_1 + x_2 + \ldots + x_k + x_{k+1} + x_{k+2} + \ldots + x_{k+m}.$$

Likewise, one has

$$(x_1{\cdot}x_2{\cdot}\ldots{\cdot}x_k)\cdot(x_{k+1}{\cdot}x_{k+2}{\cdot}\ldots{\cdot}x_{k+m}) = x_1{\cdot}x_2{\cdot}\ldots{\cdot}x_k{\cdot}x_{k+1}{\cdot}x_{k+2}{\cdot}\ldots{\cdot}x_{k+m}.$$

In terms of the $\sum$ and $\prod$ notations, these equations take the forms

$$\left(\sum_{i=1}^{k} x_k\right) + \left(\sum_{j=1}^{m} x_{k+j}\right) = \sum_{r=1}^{k+m} x_r \text{ and } \left(\prod_{i=1}^{k} x_k\right) \cdot \left(\prod_{j=1}^{m} x_{k+j}\right) = \prod_{r=1}^{k+m} x_r.$$

(b) Let k be a natural number, and let $p : \mathbb{N}_k \to \mathbb{N}_k$ be a permutation of the finite set $\mathbb{N}_k$ (see Definition (I.7.8)). Then for every ordered k-tuple $(x_1, x_2, \ldots x_k)$ in $\mathbb{R}^k$ one has

$$x_{p(1)} + x_{p(2)} + \ldots + x_{p(k)} = x_1 + x_2 + \ldots + x_k \text{ and } x_{p(1)}{\cdot}x_{p(2)}{\cdot}\ldots{\cdot}x_{p(k)} = x_1{\cdot}x_2{\cdot}\ldots{\cdot}x_k.$$

In terms of the $\sum$ and $\prod$ notations, these equations take the forms

$$\sum_{i=1}^{k} x_{p(i)} = \sum_{i=1}^{k} x_i \text{ and } \prod_{i=1}^{k} x_{p(i)} = \prod_{i=1}^{k} x_i.$$

Proof Replace the symbol $*$ in Theorem (I.10.4) throughout by $+$ to obtain the formulas involving addition. Likewise, replace $*$ by $\cdot$ throughout to obtain the formulas involving multiplication.

II.1.6 Remarks

(1) The last several results have been formulated in terms of real numbers, but it is clear that they are valid in any field.

(2) The results of Theorem (II.1.3) and Theorem (II.1.5), combined with the distributive laws for addition and multiplication (see Axiom A5), allow one to work with algebraic expressions in the usual manner, as carried out in ordinary high-school algebra, without needing to explicitly refer back to Axioms A0 through A6 time after time.

Example One computes

$$(x+y)^2 \;=\; (x+y){\cdot}(x+y) \;=\; (x+y){\cdot}x + (x+y){\cdot}y \;=\; x{\cdot}x + y{\cdot}x + x{\cdot}y + y{\cdot}y \;=\; x^2 + 2\,x{\cdot}y + y^2$$

Normally such calculations are left to the reader to carry out, sometimes without comment, sometimes with a vague hint such as ‘**by the usual algebraic manipulations** .

(3) The role played by ‘2’ in the terms x^2, y^2 and $2\,x{\cdot}y$ above is that of a natural number (i.e., a ‘counting’ number), and *not* that of an element of the field; for example, $2\,x{\cdot}y$ and x^2 are shorthands for $x\,y + x\,y$ and $x{\cdot}x$, respectively. Likewise, in the expressions $k\,c$ and c^k described in Part (3) of Definition (II.1.4), the role of k is as a counting number, that is, an element of $\mathbb{N}$, not an element of the field. For example, the expression $3\,c$ is simply a shorthand for the repeated sum $c+c+c$, which has three terms, and *not* ‘multiplication of 3 with c in the field’. This distinction becomes clearer in case the field is $\mathbb{Z}_2 \;=\; \{0,1\}$, the field with two elements described in Remark (II.1.1) (2): since the (infinite) set $\mathbb{N}$ is not a subset of the (finite) set $\mathbb{Z}_2$, the expression $3\,c$, cannot mean ‘multiplication in the field $\mathbb{Z}_2$ of 3 with c’, since the natural number 3 is not an element of the field $\mathbb{Z}_2$.

In the next section, however, after introducing more axioms for the field $\mathbb{R}$, we are able to replace our ‘primitive’ notion of $\mathbb{N}$ with an improved version which forms a subset of $\mathbb{R}$. For these new counting numbers the equation $k\,c \;=\; k{\cdot}c$, with $\cdot$ now denoting multiplication of real numbers, does hold.

II.2 Order Properties of the Real Numbers

The field axioms discussed in the preceding section involve both ‘plus signs’ and ‘minus signs’. Despite that suggestive notation, no reasonable concepts of ‘positive number’ or ‘negative number’ can come out of that discussion. For that one needs to state some additional axioms.

The set $\mathbb{R}$ has a distinguished subset, the set of **positive numbers**, that satisfies the following **Order Axioms**:

Axiom O1 For every number z exactly one of the following statements is true:

‘z is positive’; ‘$z \;=\; 0$’; $-z$ is positive’

Axiom O2 (The Closure Laws for Positive Numbers) The sum of two positive numbers is positive, and the product of positive numbers is positive.

II.2.1 Remarks

(1) The existence of a distinguished subset of $\mathbb{R}$ that satisfy Axioms O1 and O2 cannot be deduced from the field axioms studied in the preceding section. Indeed, if the existence of such a subset could be deduced from the field axioms alone, then *every* field would have such a subset. Consider, however, the field $\mathbb{Z}_2$ described in Remark (II.1.1) (2), and suppose it has such a subset.

In that field there are exactly two distinct elements, denoted 0 and 1, and one has $-1 = 1$. By Axiom O1 exactly one of the following statements is true:

$$\text{‘1 is positive’;} \quad \text{‘}1 = 0\text{’;} \quad \text{‘}-1 \text{ is positive’}$$

Since in the field $\mathbb{Z}_2$ one has $-1 = 1$ and $1 \neq 0$, it follows that either none of these statements is true or two of them are true. In either case, Axiom O1 cannot hold in this field.

(2) The real number system is not the only field which satisfies all the axioms considered so far. For example, the set $\mathbb{Q}$ of rational numbers, together with its usual operations of addition and multiplication, and with its usual notion of ‘positive’, satisfy these axioms. Any such system is called an **ordered field**. In particular, every property obtained in this section remains valid in any ordered field.

(3) An ordered field is a set S with a pair of binary operations $+$ and $\cdot$, together with a distinguished subset set P of ‘positive’ elements, which satisfies the axioms A0-A6, O1, O2. If S' is a second such ordered field, with corresponding operations $+'$ and $\cdot'$, and set of positives P', then one says that they are **isomorphic as ordered fields** provided there exists a bijection $f : S \to S'$ which is an isomorphism , in the sense of Definition (I.10.6), of $+$ with $+'$ and of $\cdot$ with $\cdot'$, and in addition f maps P onto P'.

Since $\mathbb{Q}$ is a countable set while $\mathbb{R}$ is uncountable, it is clear that these ordered fields are not isomorphic to each other. Examples of ordered fields which are isomorphic to neither $\mathbb{Q}$ nor $\mathbb{R}$ are given in the exercises.

II.2.2 Definitions

(1) The set of positive real numbers is denoted by the symbol $\mathbb{R}^+$.

(2) An element z in $\mathbb{R}$ is said to be a **negative number** provided $-z \in \mathbb{R}^+$. The set of all negative numbers in $\mathbb{R}$ is denoted $\mathbb{R}^-$.

(3) If c is a real number, then the **absolute value of *c*** is the number $|c|$ given by the rule

$$|c| = \begin{cases} c & \text{if } c \text{ is positive} \\ 0 & \text{if } c = 0 \\ -c & \text{if } c \text{ is negative} \end{cases}$$

Similarly, the **absolute-value function** is the function $\text{abs} : \mathbb{R} \to \mathbb{R}$ given by the rule

$$\text{abs}\,(x) = |x| \text{ for every real number } x$$

II.2.3 Theorem

(a) Suppose that x_1, $x_2, \ldots x_k$ are positive numbers. Then the sum $x_1 + x_2 + \ldots + x_k$ and product $x_1 \cdot x_2 \cdot \ldots \cdot x_k$ are also positive. (The repeated sums and products used here are defined, as in Definition (I.10.3), using ‘left-to-right extensions’.)

(b) If $z \in \mathbb{R}^+$ then $(-z) \in \mathbb{R}^-$.

(c) If $z, w \in \mathbb{R}^-$ then $(z + w) \in \mathbb{R}^-$ but $z \cdot w \in \mathbb{R}^+$. That is, the sum of two negative numbers is negative, but the product of two negative numbers is positive.

(d) If $z{\in}\mathbb{R}^+$ and $w{\in}\mathbb{R}^-$ then $z{\cdot}w{\in}\mathbb{R}^-$. That is, the product of a positive number and a negative number is negative.

(e) If z is a nonzero number then z^2 is positive and $-(z^2)$ is negative. In particular, the multiplicative identity 1 is positive, while -1 is negative.

(f) If $z{\in}\mathbb{R}^+$ then $1/z{\in}\mathbb{R}^+$. Likewise, if $z{\in}\mathbb{R}^-$ then $1/z{\in}\mathbb{R}^-$.

Partial Proof: The proofs of Parts (a), (b), (d) and (e) are left as simple exercises.

(c) Suppose that $z, w{\in}\mathbb{R}^-$. Then, by Part (a), $-z$ and $-w$ are in $\mathbb{R}^+$, and thus, by Axiom O2, $(-z) + (-w)$ and $(-z){\cdot}(-w)$ are both in $\mathbb{R}^+$. However, by Part (a) of Theorem (II.1.3), together with the Distributive Law, it follows that $(-z)+(-w) = -(z+w)$, so $-(z+w){\in}\mathbb{R}^+$. By definition of $\mathbb{R}^-$, it follows that $(z + w){\in}\mathbb{R}^-$, as claimed. Likewise, from Part (b) of Theorem (II.1.3) one sees that $z{\cdot}w = (-z){\cdot}(-w){\in}\mathbb{R}^+$, so $(z{\cdot}w){\in}\mathbb{R}^+$, as claimed.

(f) Note that if $z \neq 0$ then $1/z \neq 0$ (by Part (c) of Theorem (II.1.3)) and $z{\cdot}\dfrac{1}{z} = 1$; that is, by Part (d), the product of z with $1/z$ is in $\mathbb{R}^+$. It then follows from Part (c) that one cannot have z in $\mathbb{R}^+$ and $1/z$ in $\mathbb{R}^-$ or *vice versa.* Thus either both z and $1/z$ are in $\mathbb{R}^+$ or both are in $\mathbb{R}^-$, as claimed.

Remark The careful reader will notice that in the preceding paragraphs the word 'negative' is referred to in two different senses. For example, in Part (a) the statement '$(-z){\in}\mathbb{R}^{-1}$', when expressed in words, says that 'the negative of z (in the sense of Axiom A6) is a negative number (in the sense of Part (2) of Definition (II.2.2))'.

Axioms O1 and O2 are called the 'Order Axioms' for $\mathbb{R}$ because they lead directly to the standard notion of 'ordering of numbers', as follows:

II.2.4 Definition

Let x and y be real numbers.

(1) One says that ***x* is less than *y*** (in symbols: $x < y$) if $y - x$ is positive (in the sense of Definition (II.2.2)). As in ordinary English, one can also say that ***y* is greater than *x*** (in symbols: $y > x$) to mean the same thing.

(2) If either of the statements $x < y$ or $x = y$ is true, then one says that ***x* is less than, or equal to, *y***; in symbols: $x \leq y$. In this case one can also say that ***y* is greater than, or equal to, *x*** (in symbols: $y \geq x$).

The next result summarizes the basic facts about the relations $<$ and $\leq$ between real numbers. The reader should formulate the analogous facts for the relations $>$ and $\geq$.

II.2.5 Theorem

Throughout the statement of this theorem the quantities u, v, x, y and z are real numbers.

(a) If z is any real number, then $z > 0$ if, and only if, z is positive. Likewise, $z < 0$ if, and only if, z is negative.

(b) (Trichotomy Property of Order) The numbers x and y satisfy exactly one of the following conditions:

$$x < y; \quad x = y; \quad y < x.$$

(c) (Transitivity Properties of Order) If the numbers x, y and z satisfy the conditions that $x < y$ and $y < z$, then $x < z$.

More generally, let n be a natural number such that $n \geq 3$, and suppose that x_1, x_2, x_3, $\ldots x_n$ are real numbers such that each of the following inequalities hold:

$$x_1 < x_2; \quad x_2 < x_3; \ \ldots x_{n-2} < x_{n-1}; \quad x_{n-1} < x_n \quad (*)$$

Then for each pair of indices i and j in $\mathbb{N}_n$ such that $i < j$ one has $x_i < x_j$.

<u>Note</u> One normally abbreviates the string of inequalities listed in $(*)$ by the expression

$$x_1 < x_2 < \ \ldots \ < x_{n-1} < x_n.$$

(d) (Addition Properties of Order) If $x < y$, then $x + u < y + u$. Similarly, if $x < y$ and $u < v$, then $x + u < y + v$.

(e) (Negation Property of Order) If $x < y$, then $-y < -x$.

(f) (Multiplication Properties of Order) Suppose that $x < y$.
(i) If u is positive then $u\,x < u\,y$
(ii) If u is negative then $u\,y < u\,x$.
(iii) If $u = 0$ then $u\,x = u\,y = 0$.

(g) (Reciprocal Properties of Order) Let x and y be nonzero numbers, so that the reciprocals $1/x$ and $1/y$ exist. Suppose that $x < y$. Then the following statements hold:
(i) If x and y are both positive, then their reciprocals are both positive and $\frac{1}{y} < \frac{1}{x}$.
(ii) If x and y are both negative, then their reciprocals are both negative and $\frac{1}{y} < \frac{1}{x}$.

(h) If, in Parts (c), (d), (e), (f) and (g) above, one replaces each occurance of the symbol $<$ by the symbol $\leq$, the resulting statements remain true.

<u>Note</u> Only the symbol $<$ is to be replaced; the *words* ‘positive’ and ‘negative’ are not to be changed.

<u>Partial Proof</u>:

The simple proofs of Parts (a), (f), (g) and (h) are left as exercises.

(b) Consider the number $z = y - x$. By Axiom O1 exactly one of the following must hold:

$$z \in \mathbb{R}^+; \text{ or } z = 0; \text{ or } -z \in \mathbb{R}^+.$$

However, $-z = -(y - x) = x - y$, so by Definition (II.2.4) the three possibilities for z imply that exactly one of the following must hold:

$$x < y; \text{ or } x = y; \text{ or } y < x,$$

as claimed.

(c) Note that, by Part (f) of Theorem (II.1.3), one has $z - x = (z - y) + (y - x)$. However the hypothesis that $x < y$ and $y < z$ implies (by the definition of ‘$<$’) that $z - y$ and $y - x$ are both positive. It now follows from Axiom O2 that their sum is positive; that is, $z - x > 0$, and thus $x < z$, as claimed. The more general version follows readily from this by using Mathematical Induction; the details are left as an exercise.

(d) Note that, by Part (f) of Theorem (II.1.3), one has

$$y - x \,=\, (y - (-u)) + ((-u) - x) \,=\, (y + u) - (u + x) \,=\, (y + u) - (x + u)$$

But, by the hypothesesis that $x \,<\, y$ one knows that $(y - x) \in \mathbb{R}^+$, hence $((y + u) - (x + u)) \in \mathbb{R}^+$, hence $x + u \,<\, y + u$, as claimed.

If, in addition, $u \,<\, v$, then the preceding result also implies that $y + u \,<\, y + v$. Now use Transitivity to get $x + u \,<\, y + v$.

(e) Apply Part (d) above, with $u \,=\, -(x + y)$, to get

$$x - (x + y) \,<\, y - (x + y),$$

which after the obvious algebraic simplification becomes $-y \,<\, -x$.

II.2.6 Remarks

(1) In light of the 'Trichotomy Property', Part (b) of the preceding theorem, the statement '$x \leq y$' is equivalent to 'x is not greater than y; in symbols, $x \not> y$. Likewise, the statement '$y \geq x$' is equivalent to 'y is not less than x'; in symbols, $y \not< x$. Normally we use these equivalences without explicitly saying so.

(2) Sometimes one refers to inequalities of the form $x \,<\, y$ or $y \,>\, x$ as **strict inequalities**, while those of the form $x \leq y$ or $y \geq x$ are called **weak inequalities**.

(3) The careful reader will observe a notational 'anachronism' in Part (c) of the preceding theorem. Indeed, the theorem concerns properties of the newly-defined notion of the order '<' in the real numbers. However, the same symbol '<' also appears here in the expression $i \,<\, j$ in Part (c), indicating the relation between the natural numbers i and j. The meaning here of the symbol '<' (and its near relative '≤') in the context of natural numbers is assumed to be already known from our primitive notions of natural numbers. In contrast, the use of the symbol '<' (and the symbol $\leq$), in the context of real numbers, is in the process of being developed here. Such ambiguities occur frequently in mathematics. Normally the context makes it clear which meaning is intended.

II.2.7 Corollary

Suppose that x and y are real numbers such that $0 \,<\, x \leq y$. Let k be an element of $\mathbb{N}$. Then $0 \,<\, x^k \leq y^k$. Furthermore, one has $x^k \,=\, y^k$ if, and only if, $x \,=\, y$.

Proof Left as an exercise.

II.2.8 The Infinity Symbols

In analysis, as in calculus, it convenient to introduce the 'infinity' symbols, $+\infty$ and $-\infty$. We use the infinity symbols in *This Textbook*, but subject to the following rules:

(i) The symbols $+\infty$ and $-\infty$, pronounced '**plus infinity** and **minus infinity**, are distinct objects; that is, $-\infty \neq +\infty$. Also, neither symbol is a real number. As a class they are referred to as **the infinities**.. We often abbreviate the symbol $+\infty$ to ∞.

(ii) The meaning of the phrase 'x is less than y' (and thus of the expression $x < y$), previously restricted to the case x and y are both real numbers, is extended to include the possibility that x or y might be $-\infty$ or $+\infty$ in accordance with the following rules:

$$-\infty < +\infty; \quad -\infty < x \text{ and } x < +\infty \text{ for all } x \text{ in } \mathbb{R}.$$

The latter statement is sometimes abbreviated to: '$-\infty < x < +\infty$ for all real x'.

(iii) To say that a quantity z is an **extended real number** means that either z is a real number or z is one of the infinities $+\infty$ or $-\infty$.

(iv) Following the standard usage, we allow a restricted **arithmetic of infinities**. More precisely, one is allowed to write:

$-\infty = -(+\infty) = -(\infty)$ and $\infty = +\infty = -(-\infty)$.
$+\infty + c = +\infty$ and $-\infty + c = -\infty$ for every real number c.
$(+\infty) + (+\infty) = +\infty$ and $(-\infty) + (-\infty) = -\infty$.
$(+\infty)\cdot c = +\infty$ and $(-\infty)\cdot c = -\infty$ for every extended real number $c > 0$.
$(+\infty)\cdot c = -\infty$ and $(-\infty)\cdot c = +\infty$ for every extended real number $c < 0$.
$c/(+\infty) = c/(-\infty) = 0$ for every real number c.
$|-\infty| = |+\infty| = +\infty$.

In contrast, normally we are *not* allowed to use certain expressions, such as $\infty - \infty$, $0\cdot\infty$, ∞/∞, or the corresponding expressions in which the symbol ∞ is replaced by either $+\infty$ or $-\infty$. Indeed, such expressions are called **indeterminate forms**, and any exceptions to this proscription will be justified by the specific context.

Geometrically speaking, one thinks of the object $+\infty$ as being 'to the right of every real number x on the x-axis'. Likewise, one thinks of $-\infty$ as being 'to the left of every real number x'.

One of the distinctive features of analysis is the type of techniques it frequently uses to prove the equality of two quantities. Unlike the situation in algebra, where one generally proves equality of, say, A and B directly, in analysis often such equality is proved indirectly, usually by showing that the statement $A \neq B$ cannot possibly be true. Many students initially find such a 'proof by contradiction' less convincing than a direct proof. Perhaps they should consult with Sherlock Holmes; see Chapter Quote (4) at the beginning of this chapter.

Among the most oldest of these indirect methods is the following simple result.

II.2.9 Theorem (The Principle of Eudoxus)

(1) A necessary and sufficient condition for real numbers A and B to be equal is that the following condition hold:

$$|B - A| < \varepsilon \text{ for every real number } \varepsilon > 0.$$

(2) Similarly, necessary and sufficient condition for real numbers A and B to be equal is that the following condition hold:

$$|B - A| \leq \varepsilon \text{ for every real number } \varepsilon > 0.$$

Proof (1) If $A = B$, then $|B - A| = 0$, hence $|B - A| < \varepsilon$ for every positive ε.

Conversely, suppose that $A \neq B$, so that $|B - A| > 0$. Choose $\varepsilon = |B - A|$. Then it is not the case that $|B - A| < \varepsilon$.

(2) Left as a simple exercise.

The Principle of Eudoxus is so obvious that it may appear impossible that it could be the basis of anything significant. Nevertheless, we use it many times in *This Textbook*. Frequently it arises in conjunction with one of the results in the next theorem.

II.2.10 Theorem

(a) (The Magnitude–Interval Inequalities) Let M be a positive real number.

(i) A necesary and sufficient condition for real numbers x and c to satisfy the inequality $|x-c| \leq M$ is that they satisfy the inequalities $-M+c \leq x \leq M+c$. Furthermore, the case $|x-c| = M$ occurs if, and only if, either $x = M+c$ of $x = -M+c$.

(ii) A necessary and sufficient condition for x and c to satisfy the inequality $|x-c| \geq M$ is that either $x \leq -M+c$ or $x \geq M+c$.

(iii) The statements, obtained from (i) and (ii) by replacing all occurances of the 'weak inequality' symbols, $\leq$ and $\geq$, by the corresponding strict inequality symbols, $<$ and $>$, are also true. That is, $|x-c| < M$ if, and only if, $-M+c < x < M+c$; and $|x-c| < M$ if, and only if, $-M+c < x < M+c$.

Note In the important special case in which $c = 0$ one has $x - c = x$, and the preceding statements simplify a bit: $|x| \leq M$ if, and only if, $-M \leq x \leq M$, and so on.

(b) (The Basic Triangle Inequality) Suppose that x and y are real numbers. Then

$$|x+y| \leq |x| + |y| \tag{II.1}$$

and

$$|x-y| \leq |x| + |y| \tag{II.2}$$

Furthermore, the case of 'equality' holds in Inequality (II.1) if, and only if, x and y are *not* of mixed sign; that is, provided that either both are nonnegative or both are nonpositive. Likewise, the equality holds in (II.2) if, and only if, x and y *are* of mixed sign; equivalently, either $x \geq 0$ and $y \leq 0$ or $x \leq 0$ and $y \geq 0$.

(c) (The Extended Triangle Inequality) Suppose that $x_1, x_2, \ldots x_k$ are real numbers. Then

$$|x_1 + x_2 + \ldots + x_k| \leq |x_1| + |x_2| + \ldots + |x_k|. \tag{II.3}$$

Furthermore, the case of 'equality' holds in Inequality (II.3) if, and only if, the numbers $x_1, x_2, \ldots x_k$ are *not* of mixed sign; equivalently, either all of them are nonnegative, or all of them are nonpositive.

(d) (The Reverse Triangle Inequality) Suppose that x and y are real numbers. Then

$$||x| - |y|| \leq |x-y| \tag{II.4}$$

Furthermore, a necessary and sufficient for the case of equality to hold is that x and y are not of mixed sign, in the sense that either $x \geq 0$ and $y \geq 0$, or $x \leq 0$ and $y \leq 0$.

Proof

(a) These inequalities follow directly from the definition of 'absolute value'; the details are left as an exercise.

(b) Notice that, by Part (a), one has

$$-|x| \leq x \leq |x| \text{ and } -|y| \leq y \leq |y|$$

Thus, by Part (g) of Theorem (II.2.5), one gets

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

Since $-|x|-|y| = -(|x|+|y|)$, it then follows from Part (a) again that $|x+y| \leq |x|+|y|$, as claimed. Moreover, it also follows from Part (a) that equality occurs in Inequality (II.1) if, and only if, either $x + y = |x| + |y|$ or $x + y = -(|x| + |y|)$. Since $x \leq |x|$ and $y \leq |y|$, it follows from Part (g) of Theorem (II.2.5) again that $x + y = |x| + |y|$ precisely when $x = |x|$ and $y = |y|$; that is, when $x \geq 0$ and $y \geq 0$. Likewise, since $-|x| \leq x$ and $-|y \leq y|$, the only way to have $-(|x|+|y|) = x+y$ is if $x = -|x| \leq 0$ and $y = -|y| \leq 0$.

(c) This follows easily from Part (b) by using Mathematical Induction on k.

(d) Note that, by the Basic Triangle Inequality, $|x - y| + |y| \geq |(x - y) + y| = |x|$. Thus, $|x-y| \geq |x|-|y|$. Interchanging the roles of x and y then yields $|y-x| \geq |y|-|x|$. Since one of the numbers $|x| - |y|$ or $|y| - |x|$ equals $||x| - |y||$, the desired inequality follows. The case of 'equality' is left as an exercise.

Interpreting $\mathbb{N}, \mathbb{Z}$ and $\mathbb{Q}$ as Subsets of $\mathbb{R}$

The axiomatic treatment of $\mathbb{R}$ being developed in this chapter describes various important properties of real numbers. However, it says nothing about the nature of these numbers; that is, what type of objects they are. In particular, the axioms do not specify the relation between 'real number' and, say, the primitive notion of 'natural number' used in Chapter (I). Experience from elementary calculus demonstrates the usefulness of being able to think of natural numbers, as well as integers and rational numbers, as being examples of real numbers. For example, in dealing with the function $f(x) = \sqrt{2}\,x^3$, one wants to treat the '3' in x^3 as a natural (i.e., counting) number: x^3 is the product $x{\cdot}x{\cdot}x$ with 3 factors. In contrast, one wants the '2' in the expression $\sqrt{2}$ to be a real number so that extracting its square root makes sense.

There is a simple solution, which does not require changing one's primitive notion of 'natural number', that works in any ordered field. Indeed, every such field has a naturally defined countably infinite subset which inherits, from the field operations of + and $\cdot$ and the field ordering $<$, binary operations and an order structure 'isomorphic' to the corresponding ones in the primitive notion of $\mathbb{N}$. This subset is defined below. Since the ordered field of greatest interest for us is the field $\mathbb{R}$ of real numbers, we formulate everything in terms of that field; but of course the construction works for every ordered field.

Temporary Notation

In the theorem below we use the standard notation 0, 1, +, $\cdot$, $<$ when dealing with elements of $\mathbb{N}$; but in $\mathbb{R}$ we temporarily append the subscript 'r' (for 'real'): 0_r, 1_r, $+_r$, $\cdot_r$, $<_r$.

II.2.11 Theorem

Let $f : \mathbb{N} \to \mathbb{R}$ be given by the rule $f(k) = k\,1_r$ for all k in $\mathbb{N}$; that is, $f(k) = 1_r +_r 1_r +_r +_r \ldots +_r 1_r$, where the quantity on the right of this last equation is the sum in $\mathbb{R}$ of k copies of the real number 1_r; see Definition (II.1.4). Let $\mathbb{N}_r$ denote the image in $\mathbb{R}$ of the function f; that is,

$$\mathbb{N}_r = \{1_r, 1_r +_r 1_r, \ldots\} = \{k\,1_r : k {\in} \mathbb{N}\}$$

(a) The function f preserves the relation 'less than', in the sense that if k and n are elements of $\mathbb{N}$ such that $k < n$ then $f(k) <_r f(n)$. In particular, the function $f : \mathbb{N} \to \mathbb{N}_r$ is a bijection, hence the set $\mathbb{N}_r$ is countably infinite.

(b) The bijection f is an isomorphism of the binary operations addition and multiplication, in the sense of Definition (I.10.6); that is,

$$f(j+k) = f(j) +_r f(k) \text{ and } f(j \cdot k) = f(j) \cdot_r f(k) \text{ for all } j \text{ and } k \text{ in } \mathbb{N} \qquad (*)$$

Proof

(a) Suppose that k and n are natural numbers such that $k < n$. As a notational convenience, for each $j \in \mathbb{N}_n$ let $x_j = 1_r$, and let $m = n - k$, so that $m \in \mathbb{N}$ (since $k < n$). Then one has

$$f(n) = n\,1_r = x_1 + x_2 + \ldots + x_k + x_{k+1} + \ldots + x_{k+m} = (x_1 + \ldots + x_k) + (x_{k+1} + \ldots + x_{k+m}),$$

where the last equation comes from Part (a) of Theorem (II.1.5). That is, $f(n) = f(k) + m\,1_r$. Since $m\,1_r$ is positive in the ordered field $\mathbb{R}$, it follows that $f(k) <_r f(n)$, as required.

It is now clear from the definition of the set $\mathbb{N}_r$ that the function $f : \mathbb{N} \to \mathbb{N}_r$ is a surjection. It is equally clear that it is also an injection, and thus a bijection, since if $i, j \in \mathbb{N}$ with $i \neq j$, then either $i < j$ of $j < i$; in either case, it follows from what has just been proved that $f(i) \neq f(j)$.

(b) The 'additive fact' $f(j+k) = f(j) +_r f(k)$ follows easily from the Extended Associative Law for Addition, as above.

Concerning the equation $f(j \cdot k) = f(j) \cdot_r f(k)$, let A denote the set of natural numbers k such that $f(j \cdot k) = f(j) \cdot_r f(k)$ for each k in $\mathbb{N}$. It is clear that $1 \in A$, since 1 and 1_r are the multiplicative identities in $\mathbb{N}$ and $\mathbb{R}$, respectively. The fact that k being in A implies that $(k+1) \in A$ follows easily using the Distributive Laws in $\mathbb{N}$ and $\mathbb{R}$, together with the 'additive fact' proved above; the details are left as an exeercise.

II.2.12 Remark

The bijection $f : \mathbb{N} \to \mathbb{N}_r$ allows one to use the countably infinite set $\mathbb{N}_r$ to carry out any 'counting' activites for which one would ordinarily use the 'primitive' set $\mathbb{N}$. However, since $\mathbb{N}_r$ is a subset of $\mathbb{R}$ which is closed under addition and multiplication of reals, it also interacts usefully with other real numbers; see Remark (I.8.3) (2). For example, we can now think of the '2' and the '3' in the example $\sqrt{2}\,x^3$ mentioned above as both being real numbers, with the 'counting' interpretation of the '3' arising from the counting activity assigned to the countable set $\mathbb{N}_r$.

It is easy to form from the subset $\mathbb{N}_r$ above, together with the operations operations of the ordered field $\mathbb{R}$, subsets of $\mathbb{R}$ which behave algebraically exactly like the 'primitive' notions of the integers and the rational numbers.

II.2.13 Definition (The Integers and the Rational Numbers as Subsets of R)

Let $\mathbb{N}_r$ be the subset of $\mathbb{R}$ described above.

(1) A real number k is said to be a (real) **integer** provided one of the following cases holds:

(i) k is in $\mathbb{N}_r$;

(ii) $-k$ is in $\mathbb{N}_r$

(iii) $k = 0_r$.

The set of all such (real) integers is denoted by $\mathbb{Z}_r$.

Note that the 'negation' indicated in (ii) is the usual negation from the axioms for the real numbers.

(2) A real number x is said to be a (real) **rational number** provided it can be expressed as a ratio $x = j/k$ where j and k are (real) integers, as defined in Part (1), with $k >_r 0_r$. The set of all such real rational numbers is denoted $\mathbb{Q}_r$.

Note that the division indicated by the expression j/k is as described in Definition (II.1.2).

The subsets $\mathbb{N}_r$ and $\mathbb{Q}_r$ described above inherit naturally from the ordered field $\mathbb{R}$ binary operations of addition and multiplication, as well as an order relation. The inherited 'order relation' is the obvious one; for example, if x and y are in $\mathbb{Q}_r$, then they are in $\mathbb{R}$, where the relation $x < y$ already makes sense. The inherited addition and multiplication requre a bit more work, since (by definition) a binary operation must satisfy the 'closure' requirement.

II.2.14 Theorem

Let $\mathbb{Z}_r$ and $\mathbb{Q}_r$ be the sets of (real) integers and (real) rationals, respectively, described above. Then these sets are closed under the (real) binary operations $+_r$ and $\cdot_r$. That is, if x and y are in $\mathbb{Z}_r$, then $x +_r y$ and $x \cdot_r y$ are in $\mathbb{Z}_r$; similarly, if x and y are in $\mathbb{Q}_r$, then so are $x +_r y$ and $x \cdot_r y$. Otherwise stated, the restrictions of these operations to $\mathbb{Z}_r \times \mathbb{Z}_r$ and to $\mathbb{Q}_r \times \mathbb{Q}_r$ are binary operations on the sets $\mathbb{Z}_r$ and $\mathbb{Q}_r$, respectively.

II.2.15 Remarks

(1) It is not hard to see, using the bijection $f : \mathbb{N} \to \mathbb{N}_r$ described above in Theorem (II.2.11), that the sets $\mathbb{Z}_r$ and $\mathbb{Q}_r$ defined here, together with the operations on them arising from the real $+_r$ and $\cdot_r$, form systems isomorphic with the corresponding 'primitive' systems $\mathbb{Z}$ and $\mathbb{Q}$, respectively. The same holds for the inherited 'order' relations.

(2) In light of the preceding results, it makes sense to use, in place of the 'primitive' concepts of $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$, the 'new and improved' versions $\mathbb{N}_r$, $\mathbb{Z}_r$ and $\mathbb{Q}_r$ constructed above. We do exactly that for the rest of the main body of *This Textbook*. In particular, from now on the terms 'natural number', 'integer' and 'rational number' refer to elements of the sets $\mathbb{N}_r$, $\mathbb{Z}_r$ and $\mathbb{Q}_r$ unless indicated explicitly to the contrary.

Likewise, since the 'primitive' versions of $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ are no longer in use, we simplify notations and henceforth drop the subscript 'r' from $\mathbb{N}_r$, $\mathbb{Z}_r$ and $\mathbb{Q}_r$, as well as from $+_r$, $\cdot_r$ and $<_r$.

(3) This is also a good place to introduce the customary abbreviation of omitting the infix 'dot' for real multiplication and using 'juxtaposition. Thus from now on we normally write, say, $x\,y$ instead of $x \cdot y$; but if clarity needs it, e.g., $3\,2$ *vs* $3 \cdot 2$, we'll include the dot.

The final results in this section use the fact, developed above, that we can think of natural numbers and rational numbers as types of real numbers.

II.2.16 Theorem

(a) Let u be a real number and let k be a natural number. Then the following formula holds:

$$1 + u + \ldots + u^k = \begin{cases} k+1 & \text{if } u = 1 \\ \dfrac{1 - u^{k+1}}{1-u} & \text{if } u \neq 1 \end{cases}$$

(b) More generally, if m is any natural number, then

$$u^m + u^{m+1} + \ldots + u^{m+k} = \begin{cases} k+1 & \text{if } u = 1 \\ \dfrac{u^m - u^{k+m+1}}{1-u} & \text{if } u \neq 1 \end{cases}$$

(c) The following formula holds:

$$1 - u^{k+1} = (1-u)(1 + u + u^2 + \ldots + u^k) \text{ for all } u \text{ in } \mathbb{R}.$$

Proof

(a) The case when $u = 1$ is trivial. Thus, suppose that $u \neq 1$. Let A be the set of natural numbers k for which the statement is true.

It is clear that $k = 1$ in A. Indeed, when $k = 1$ the statement to be proved reduces to

$$1 + u = \frac{1-u^2}{1-u};$$

note that the hypothesis $u \neq 1$ means that the indicated division is allowed. This last equation follows easily from the observation that $1 - u^2 = (1-u)(1+u)$ and thus, by Part (e) of Theorem (II.1.3), one has

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

Next, suppose that k is in A. Then the statement to be proved for $k+1$ is

$$1 + u + u^2 + \ldots + u^k + u^{k+1} = \frac{1-u^{k+2}}{1-u}.$$

Notice that by the induction hypothesis (i.e., k is in A), together with the definition of 'Repeated Addition' and Parts (e) and (f) of Theorem (II.1.3), one sees that

$$1+u+u^2+\ldots+u^k+u^{k+1} = (1+u+u^2+\ldots+u^k)+u^{k+1} = \frac{1-u^{k+1}}{1-u}+u^{k+1} = \frac{(1-u^{k+1})+u^{k+1}(1-u)}{1-u} =$$

$$\frac{(1-u^{k+1})+(u^{k+1}-u^{k+2})}{1-u} = \frac{1-u^{k+2}}{1-u}.$$

Thus $k+1$ is also in A; so, by Mathematical Induction, one has $A = \mathbb{N}$ and hence the claimed result is true.

(b) Multiply both sides of the result in Part (a) by u^m and simplify.

(c) The result follows by multiplying both sides of the equations obtained in Part (a) by the quantity $(1-u)$.

II.2.17 The Principle of Ingenious Cancellations

Students learning real analysis often find proofs in the texts to be mysterious; in too many cases, the authors write down complicated expressions, seemingly without motivation, that magically work to solve the problem at hand. Although on occasion an author really does perform a mathematical miracle, usually there is ample motivation for what the author does; but frequently the reader is simply not told about that motivation. The purpose of this brief discussion is outline one important idea that often lies behind the 'miracles' found in analytical proofs.

Typical Situation in Analysis Much of real analysis is devoted, directly or indirectly, to the issue of approximating a quantity A by a second quantity B. In such circumstances it is usually important to have an idea of the magnitude $|A-B|$ of the error in that approximation; for example, this is the case in situations in which one wants to use the Principle of Eudoxus to show equality of two quantities. Thus, many proofs in analysis contain a step that takes the following form. (In what follows the words in italics are supposed to represent the thoughts of the person who is doing the problem.)

Stating the Problem *'I have to show that the error $|B-A|$ in the approximation $A \approx B$ is smaller than a certain given amount $\varepsilon > 0$; that is,*

$$|A-B| < \varepsilon \qquad (*)$$

How can I do that?'

Of course, any proof of such a fact must be based on what one already knows. Thus a key step in proving an inequality such as $(*)$ is this:

Reminding Myself What I Already Know *'Out of the enormous mass of mathematics that I have already learned, combined with the given hypotheses concerning A and B, what facts seem relevant to the problem of showing Inequality $(*)$?'*

In real analysis the 'relevant known facts' often look something like this:

Relating What I Know to the Problem at Hand *'Hey, in Theorem X on Page 2348 it was proved that there are quantities C and D such that we know a lot about the differences $A-C$, $C-D$ and $D-B$. Thus, if I could somehow relate these 'known' quantities to the quantity $A-B$ under study, then I might get somewhere. But look:*

$$A-B = A-C+C-D+D-B = (A-C)+(C-D)+(D-B) \qquad (**)$$

That is, by adding and subtracting just the right quantities, C and D, in just the right way, and regrouping cleverly, I can relate the quantity of interest, $A-B$, to quantities about which I know something, namely $A-C$, $C-D$ and $D-B$.

But wait: the problem actually asks about the magnitude $|A-B|$, not the simple difference $A-B$. However, I just learned about the Extended Triangle Inequality; maybe I can use that. Applying it to Equation $(**)$ above, I get

$$|A-B| = |(A-C)+(C-D)+(D-B)| \leq |A-C|+|C-D|+|D-C| \qquad (***)$$

This looks promising!'

The next step is easy to state:

Reduce the Problem to Simpler Problems *'If I can use my knowledge of the differences $A-C$, $C-D$ and $D-B$ to show that $|A-C| < \varepsilon/3$, $|C-D| < \varepsilon/3$ and $|D-B| < \varepsilon/3$, then Inequality $(***)$ can be used to show that*

$$|A-B| \leq |A-C|+|C-D|+|D-C| < \frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3};$$

that is, $|A - B| < \varepsilon$, as required.'

Remark Choosing the factors of ε above to all be 1/3, 1/3, 1/3 is somewhat arbitrary. Any choice of factors a, b and c such that $a, b, c > 0$ and $a + b + c = 1$ could work just as well.

Unfortunately, in many problems completing this last step is hardest. Indeed, it may be that you *cannot* prove that, for example, $|A - C| < \varepsilon/3$: perhaps it is simply not true, or because you can't figure out how to prove it (but someone else could). Should this occur, you may have to go back to an earlier step and try something else. (You don't see this happen in the proofs in textbooks, of course: authors normally don't publish their failed attempts!)

It is convenient, for future reference, to give a name to the idea outlined above, namely to simplify problems by adding and subtracting ingeniously chosen quantities. In *This Textbook* we use the name **The Principle of Ingenious Cancellations**. (This definitely *not* standard terminology in analysis.)

Note The preceding discussion, and especially Equation ($**$), involves repeated use of the famous 'Add-and-Subtract Trick' from high-school algebra: one simplifies an algebraic expression by adding and subtracting the same quantity, then regrouping. In high-school algebra one also encounters a similar 'Multiply-and-Divide Trick', which involves multiplying and dividing a quantity of interest by the same nonzero quantity. Both tricks are used extensively in analysis.

II.3 Some Terminology and Notation Based on Order

The order properties of $\mathbb{R}$ lead to frequently used terminology and notation. As usual, it also makes sense in a general ordered field.

II.3.1 Definition (Intervals in R; Partitions of Intervals)

(1) Let a and b be real numbers with $a < b$. A nonempty subset I of $\mathbb{R}$ is said to be a **bounded interval in R with endpoints a and b** if I equals one of the four sets (a, b), $[a, b]$, $[a, b)$, $(a, b]$ given as follows:

(i) $(a, b) = \{x \in \mathbb{R} : a < x < b\}$; (ii) $[a, b] = \{x \in \mathbb{R} : a \le x \le b\}$;
(iii) $(a, b] = \{x \in \mathbb{R} : a < x \le b\}$; (iv) $[a, b) = \{x \in \mathbb{R} : a \le x < b\}$.

More precisely, the interval (a, b) is the **open interval with endpoints a and b**, while $[a, b]$ is the **closed interval with left endpoint a and right endpoint b**. Intervals of the form $(a, b]$ or $[a, b)$ are said to be **half-open** or **semiopen**.

(2) A nonempty subset J of $\mathbb{R}$ is said to be an **unbounded interval in R** if either

(a) $J = \mathbb{R}$; or

(b) there exists a real number c such that J is one of the four sets $(-\infty, c)$, $(-\infty, c]$, $[c, +\infty)$, $(c, +\infty)$ given as follows:

(i′) $(-\infty, c) = \{x \in \mathbb{R} : x < c\}$ (ii′) $(c, +\infty) = \{x \in \mathbb{R} : c < x\}$
(iii′) $(-\infty, c] = \{x \in \mathbb{R} : x \le c\}$ (iv′) $[c, +\infty) = \{x \in \mathbb{R} : c \le x\}$

Unbounded intervals of the form in (i′) and (ii′) are sometimes called **open half-lines**; likewise, those of the form in (iii′) and (iv′) are called **closed half-lines**. The point c is then called **the endpoint of** J.

It is sometimes convenient to use the expression $(-\infty, +\infty)$ as alternate notation for the set $\mathbb{R}$.

(3) A nonempty subset I of $\mathbb{R}$ is said to be an **interval in $\mathbb{R}$** provided it is either a bounded interval or an unbounded interval; that is, if it is one of the nine types of sets described in Parts (1) and (2) above.

A real number x is said to be an **interior point** of an interval I provided $x \in I$ and x is not an endpoint of I.

(4) Let I be a closed bounded interval $\mathbb{R}$; that is, $I = [a, b]$ for some real numbers a and b with $a < b$. Let k be a natural number. A **partition of the interval I into k subintervals** is set P, consisting of exactly $k + 1$ points of I, such that the endpoints a and b are in P. Thus, P can be written in the form $P = \{a, x_1, x_2, \ldots, x_{k-1}, b\}$.

It is customary, although not required, that the interior points x_1, $x_2, \ldots\, x_{k-1}$ of $[a, b]$ in the set P be labeled so that if $i < j$ then $x_i < x_j$. If that is done, then it becomes convenient to set $x_0 = a$ and $x_k = b$, and to write $P = \{a = x_0 < x_1 < x_2 < \ldots < x_{k-1} < x_k = b\}$. When this notational convention is in effect, the number x_j is called the **j-th partition point associated with P**; also, the subinterval $[x_{j-1}, x_j]$, whose endpoints are consecutive partition points, is called the **j-th subtinterval of I determined by the partition P**. With this notation, one often writes the length $x_j - x_{j-1}$ of the j-th subinterval by Δx_j; note that $\Delta x_j > 0$. Finally, the **norm of P**, also called the **mesh of P**, is the quantity $||P||$ given by

$$||P|| = \max\{\Delta x_1, \Delta x_2, \ldots \Delta x_k\}.$$

A **refinement of a partition P of I** is a partition P' of I such that $P' \subseteq P$. The set of all partitions of the interval $I = [a, b]$ is denoted by $\mathcal{P}_I$.

II.3.2 Examples

(1) Suppose that c is a real number and that $R > 0$. Then the set of all x in $\mathbb{R}$ such that $|x - c| < R$ is an open interval. Indeed, by Part (h) of Theorem (II.2.10) the inequality above is equivalent to

$$-R < x - c < R.$$

By adding c to each term one then gets the equivalent statement

$$c - R < x < c + R.$$

In other words the desired set consists precisely of the numbers x in the open interval $(c-R, c+R)$.

Similarly, the set of all x in $\mathbb{R}$ such that $|x - c| \leq R$ is a closed interval, namely the interval $[c - R, c + R]$.

(2) The set of all x in $\mathbb{R}$ such that $|5 - 2x| < 3$ is an open interval in $\mathbb{R}$. Indeed, by Part (h) of Theorem (II.2.10), the inequality above is equivalent to

$$-3 < 5 - 2x < 3.$$

By subtracting 5 from each term, one sees that this is equivalent to

$$-8 < -2x < -2.$$

Now multiply both sides by the negative quantity $-1/2$, and recall that multiplication by negative numbers reverses inequalities, to get the equivalent

$$1 < x < 4.$$

That is, the given set is the open interval $(1,4)$.

(3) The set A of all x in $\mathbb{R}$ such that $1 - x^2 > 0$ is an interval in $\mathbb{R}$. Indeed, it is clear that if $-1 < x < 1$ then $x^2 < 1$, so the given set A contains the open interval $(-1,1)$ as a subset. Moreover, if x is *not* in $(-1,1)$ then either $x \geq 1$ or $x \leq -1$. In either case it is clear that $x^2 \geq 1$, so such x cannot be in A. Thus, $A = (-1,1)$.

(4) Consider now the set B of all x in $\mathbb{R}$ such that $2 - x^2 > 0$. At first glance it would appear that an argument similar to that used in the preceeding example would show that this is also an interval in $\mathbb{R}$, namely the open interval $(-\sqrt{2}, \sqrt{2})$. However, this argument presupposes that there exists a real number whose square equals 2. Unfortunately, the existence of such a number does *not* follow from the axioms currently at our disposal; for example, the rational numbers also satisfy these axioms, yet there is no rational number whose square equals 2. In other words, from the axioms given so far we cannot determine whether B is an interval (i.e., whether B is of one of the nine forms given above) or not.

Remark In the notations (a,b) and $[a,b]$ given above for open and closed bounded intervals in $\mathbb{R}$, it is required that $a < b$. This requirement is nearly universal in analysis. In particular, most text books do not permit one to write $(3,-1)$ as an alternate notation for the open interval $(-1,3)$, even though the meaning seems clear. Likewise, most texts do not allow one to write, for instance, $[2,2]$ as a shorthand for $\{x{\in}\mathbb{R} : 2 \leq x \leq 2\}$, i.e., for the singleton set $\{2\}$.

These notational restrictions can be annoying. For example, suppose that one is studying a real-valued function $f : [2,5] \to \mathbb{R}$ defined on the closed interval $[2,5]$. Then it would be natural to ask about the 'interval' whose endpoints are $f(2)$ and $f(5)$. However, it is possible that one might have $f(2) = f(5)$, in which case the 'interval with endpoints $f(2)$ and $f(5)$' would be a 'degenerate' interval consisting of a single point, namely the common value of $f(2)$ and $f(5)$. Or, it could be that $f(2) > f(5)$, so that there is an honest interval with endpoints $f(2)$ and $f(5)$, but in this case it should be denoted $[f(5), f(2)]$. Unfortunately, it often happens that one does not actually know the values of $f(2)$ or $f(5)$.

It turns out that there is a simple solution which does not require violating the '$a < b$' requirement for intervals.

II.3.3 Definition (Segments in R)

Let a and b be real numbers.

(1) The **segment in R** determined by the numbers a and b is the set $\mathrm{Seg}\,[a,b]$ consisting of all real numbers x such that either $a \leq x \leq b$ or $b \leq x \leq a$. The numbers a and b are called the **endpoints of the segment Seg $[\boldsymbol{a},\boldsymbol{b}]$**; note that if $a = b$ then the segment has only one endpoint. An element x of the segment $\mathrm{Seg}\,[a,b]$ is said to be **weakly between $\boldsymbol{a}$ and $\boldsymbol{b}$**. ('Weakly' because one allows the possibility that $x = a$ or $x = b$.)

(2) Suppose that $a \neq b$, so that either $a < b$ or $b < a$. Then any number x such that $a < x < b$ or $b < x < a$, depending on whether $a < b$ or $b < a$, is called an **interior point of the segment Seg $[\boldsymbol{a},\boldsymbol{b}]$**. Such points are said to be **strictly between the endpoints** a **and** b.

(3) If the endpoints of a segment are the numbers a and b, then the **length** is the number $|b-a|$; equivalently, $|a-b|$.

Remarks (1) The definition of 'segment' extends the concept of 'closed bounded interval'. One could similarly extend the concepts of 'open interval', 'half-open interval' and 'unbounded interval' as well, but we do not do so, since there is no need in *This Textbook* for those extensions.

(2) It is a simple exercise to show that every interval in an ordered field is a convex set.

The following theorem list some basic facts about segments in $\mathbb{R}$.

II.3.4 Theorem

Suppose that a and b are real numbers.

(a) For each real number c one has $\mathrm{Seg}\,[c,c] \,=\, \{c\}$.

(b) If $a \,<\, b$ then $\mathrm{Seg}\,[a,b]$ equals the closed interval $[a,b]$, while if $b \,<\, a$, then $\mathrm{Seg}\,[a,b] \,=\, [b,a]$. More concisely, if $a \neq b$ then

$$\mathrm{Seg}\,[a,b] \,=\, [\min\{a,b\}, \max\{a,b\}]. \tag{II.5}$$

(c) For every pair of real numbers a and b, equal or not, one has $\mathrm{Seg}\,[a,b] \,=\, \mathrm{Seg}\,[b,a]$. In other words, the order in which one writes down the numbers a and b does not affect the resulting segment.

(d) A necessary and sufficient condition for x to be an element of $\mathrm{Seg}\,[a,b]$ is that $|x-a|+|b-x| \,=\, |b-a|$.

Partial Proof

The proofs of Parts (a), (b), and (c) are left as simple exercises.

Proof of (d): Note first that for *every* real x one has $b-a \,=\, (x-a)+(b-x)$. Thus, by Part (2) of Theorem (II.2.10), the Triangle Inequality, one has

$$|x-a| + |b-x| \leq |b-a| \qquad (*)$$

with equality in $(*)$ if, and only if, the numbers $x-a$ and $b-x$ are either both nonnegative or both nonpositive.

Suppose first that $x \in \mathrm{Seg}\,[a,b]$; then either $a \leq x \leq b$ or $b \leq x \leq a$. In the first case $x-a$ and $b-x$ are both nonnegative, while in the second case $x-a$ and $b-x$ are both nonpositive. In either case, the conditions for equality in $(*)$ are satisfied, so the desired equation holds.

Conversely, suppose that the desired equation holds, and thus one gets equality in $(*)$. It then follows, by reversing the previous argument, that $x \in \mathrm{Seg}\,[a,b]$, as required.

II.3.5 Corollary

Suppose that a and b are real numbers. Then a necessary and sufficient condition for a number x to be an element of $\mathrm{Seg}\,[a,b]$ is that there exist a number t in the closed interval $[0,1]$ such that

$$x \,=\, t\,a + (1-t)\,b \qquad (*)$$

Proof Suppose that Equation $(*)$ holds for some t in $[0,1]$. Using the fact that $t \geq 0$ and $1-t \geq 0$, together with the usual order properties, one easily computes that

$$|x-a|+|b-x| \,=\, |(t\,a+(1-t)\,b)-a|+|b-((t\,a)+(1-t)\,b)| \,=\, |(1-t)\,(b-a)|+|t\,(b-a)| \,=\, (1-t)\,|b-a|+t\,|b-a| \,=\, |b-a|$$

It follows from Part (d) of the preceding theorem that $x \in \mathrm{Seg}\,[a,b]$.

Conversely, suppose that $x \in \mathrm{Seg}\,[a,b]$. If $x = a$, then clearly $x = 1 \cdot a + (1-1) \cdot b$, which is Equation $(*)$ with $t = 1$. Likewise, the case $x = b$ corresponds to $t = 0$ in Equation $(*)$. Finally, suppose that x is an interior point of $\mathrm{Seg}\,[a,b]$, so that $|x-a| > 0$, $|b-x| > 0$, and of course $|b-a| > 0$. As is pointed out in the proof of Theorem (II.3.4), the nonzero quantities $x-a$, $b-x$ and $b-a$ must all be of the same sign. Thus, if one sets $t = (b-x)/(b-a)$, then $t > 0$ and $1-t = (x-a)/(b-a) > 0$, so $0 < t < 1$. It is a straight-forward calculation to verify that $x = t\,a + (1-t)\,b$.

II.3.6 Remarks

(1) Many authors use the results of the preceding corollary as the *definition* of 'segment'. One major advantage is that it makes sense in Euclidean spaces of arbitrary dimension, where there is no natural concept of 'order'. In contrast, our Definition (II.3.3) uses strongly the concept of 'order' that is valid in $\mathbb{R}$, and thus is easier to use in that context.

(2) The case $t = 1/2$ in the preceding corollary is important enough to have its own terminology. Indeed, when $t = 1/2$ one has $x = t\,a + (1-t)\,b = (a+b)/2$. This point is called the **midpoint of Seg** $[a,b]$.. It follows easily from the preceding corollary that if x is the midpoint of $\mathrm{Seg}\,[a,b]$, then $x \in \mathrm{Seg}\,[a,b]$ and $|x-a| = |b-x| = |b-a|/2$. Speaking geometrically, the distance of the midpoint from either endpoint equals half the length of the segment. (Of course if $a = b$, then all three points coincide.)

II.3.7 Definition (Convex Sets in $\mathbb{R}$)

A nonempty subset X of $\mathbb{R}$ is said to be **convex** if, for every pair of points x and y in X, one has $\mathrm{Seg}\,[x,y] \subseteq X$; that is, if x and y are in X then so is every number that is weakly between x and y.

II.3.8 Examples

(1) Every singleton subset $\{c\}$ of $\mathbb{R}$ is a convex set in $\mathbb{R}$. Indeed, $\{c\} = \mathrm{Seg}\,[c,c]$.

(2) It is easy to see that every interval in $\mathbb{R}$, whether closed, open, bounded or unbounded, is a convex subset, as is every segment in $\mathbb{R}$.

The question of whether there are any other convex subsets of $\mathbb{R}$ cannot be answered using only the algebra and order axioms. Indeed, in the ordered field $\mathbb{Q}$ let S be the set of all (rational) numbers x such that $x^2 \leq 2$. It is easy to see that this set is bounded in $\mathbb{Q}$; for example, if $x \in S$, then clearly $|x| < 4$. Nevertheless, it can be shown that this set is not an interval or segment in $\mathbb{Q}$. A full treatment of this question requires use of the 'Completeness' property of the real numbers; see Theorem (II.4.33).

Functions Which Preserve the Order

The presence of the order relation on $\mathbb{R}$ makes it natural to single out those real-valued functions that, in some sense, 'preserve' the order.

II.3.9 Definition

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty subset X of $\mathbb{R}$. (The set X need not be the full domain of the function f.)

(1) One says that f is **strictly increasing on** X provided that, for each pair of elements x_1 and x_2 in X with $x_1 < x_2$, one has $f(x_1) < f(x_2)$. Likewise, one says that f is **nondecreasing on** X, or that f is **monotonic up on** X, provided that for each pair of elements x_1 and x_2 in X with $x_1 < x_2$ one has $f(x_1) \leq f(x_2)$.

(2) One says that f is **strictly decreasing on** X provided that, for each pair of elements x_1 and x_2 in X with $x_1 < x_2$, one has $f(x_1) > f(x_2)$. Likewise, one says that f is **nonincreasing on** X, or that f is **monotonic down on** X, provided that for each pair of elements x_1 and x_2 in X with $x_1 < x_2$ one has $f(x_1) \geq f(x_2)$.

(3) One says that f is **monotonic on** X if either f is monotonic up on X or f is monotonic down on X. Likewise, one says f is **strictly monotonic on** X if either f is strictly increasing on X or f is strictly decreasing on X.

> Side Comment (on the 'increasing/decreasing' terminology) (1) Many authors use the words 'increasing' and 'decreasing' to mean the same as the phrases 'strictly increasing' and 'strictly decreasing', respectively, found in the preceding definition. Unfortunately, many other authors use the words 'increasing' and 'decreasing' to mean the same as the words 'nondecreasing' and 'nonincreasing', respectively. This conflict of usage can easily cause confusion, especially if the author neglects to make clear which usage is intended.
>
> (2) In contrast, the meanings given here of the words 'nondecreasing' and 'nonincreasing' are standard; there appears to be no disagreement among authors. However, students often find these terms confusing. The main reason may be because the statement 'f is a nondecreasing function on X' does *not* mean the same as 'f is not a decreasing function on X'. For example, consider the 'squaring function' $f : \mathbb{R} \to \mathbb{R}$, given by the rule $f(x) = x^2$ for all x in $\mathbb{R}$. This function is certainly not a decreasing function on $\mathbb{R}$; for instance, $0 < 1$ but $f(0) < f(1)$. However, it is also not a nondecreasing function on $\mathbb{R}$, since $f(-1) > f(0)$.
>
> A second source of this confusion may be that the word 'nondecreasing' involves both 'non' – a negative word – and 'decreasing' – a word which connotes, roughly speaking, 'becoming less' – to describe a concept which tries to indicate 'getting larger': that is, it is a type of 'cognitive dissonance'.
>
> (3) In *This Textbook* we generally use the terminology with the least ambiguity: strictly increasing and strictly decreasing; monotonic up and monotonic down. On occasion we may use the 'nonincreasing/nondecreasing' terminology, mainly to remind reader that these words do appear in the mathematical literature.

The concepts of 'strictly increasing/decreasing' and 'monotonic up/down' also apply to sequences of real numbers; indeed, they correspond to the case $X = \mathbb{N}$. It is convenient, however, to repeat the definitions in the context of sequences since the custom is to write the functions values using subscripts.

II.3.10 Definition

Let $\xi = (x_1, x_2, \ldots)$ be an infinite sequence of real numbers.

(1) One says that ξ is a **strictly increasing sequence** provided $x_{k+1} > x_k$ for each k in $\mathbb{N}$. It is said to be a **monotonic-up sequence** provided $x_{k+1} \geq x_k$ for each k in $\mathbb{N}$.

(2) One says that ξ is a **strictly decreasing sequence** provided $x_{k+1} < x_k$ for each k in $\mathbb{N}$. It is said to be a **monotonic-down sequence** provided $x_{k+1} \leq x_k$ for each k in $\mathbb{N}$.

(3) One says that ξ is a **monotonic sequence** if either it is monotonic up or it is monotonic down. Likewise, it is said to be a **strictly monotonic sequence** if either it is strictly increasing or strictly decreasing.

(4) The sequence ξ is said to be **eventually strictly increasing** provided there exists M in $\mathbb{R}$ such that $x_{k+1} > x_k$ for all $k \geq M$. The concepts of 'eventually monotonic up', 'eventually strictly decreasing' 'eventually monotonic down', 'eventually monotonic' and 'eventually strictly monotonic' are defined in an analogous manner.

II.3.11 Examples

(1) A constant sequence is both monotonic up and monotonic down.

(2) Let $\xi = (x_1, x_2, \dots)$ be the sequence given by the rule

$$x_k = k(k-3)(k-6) \text{ for each } k \text{ in } \mathbb{N}.$$

Clearly $x_1 = 10$, $x_2 = 8$, $x_3 = 0$, $x_4 = -8$, $x_5 = -10$, $x_6 = 0$; in particular, the sequence ξ is neither monotonic up nor montonic down. It is left as an exercise to verify that if $k \geq 6$ then $x_{k+1} > x_k$; that is, the sequence ξ is *eventually* strictly increasing.

(3) Let $\tau = (t_1, t_2, \dots)$ be given by the rule

$$t_k = \frac{1 + (-1)^k}{k}.$$

It is easy to see that this sequence is not eventually monotonic.

The simple concepts of 'bounded' and 'unbounded' subsets of $\mathbb{R}$ are important in analysis.

II.3.12 Definition

Let A be a subset of $\mathbb{R}$.

(1) A real number M is said to be an **upper bound of** A if $M \geq x$ for all x in A; equivalently, if there does *not* exist $x \in A$ such that $x > M$. Likewise, a real number m is said to be a **lower bound of** A if $m \leq x$ for all x in A; equivalently, if there does *not* exist $x \in A$ such that $x < m$.

(2) The set A is said to be **bounded above**, or sometimes **bounded on the right**, if A has an upper bound. If no such upper bound exists, then A is said to be **unbounded above**, or **unbounded on the right**

Similarly, the set A is said to be **bounded below**, or **bounded on the left**, if A has a lower bound. If no such lower bound exists then A is said to be **unbounded below** or **unbounded on the left**.

Remark The 'on the right' and 'on the left' phrasing corresponds to the standard geometric interpretation of the set $\mathbb{R}$ as a horizontal straight line with negative numbers to the left of 0 and positive numbers to the right.

(3) A subset A of $\mathbb{R}$ is said to be a **bounded set in $\mathbb{R}$** if it is bounded above and bounded below; equivalently, there exist real numbers m and M such that $m \leq x \leq M$ for every x in A. If A is not a bounded set then A is said to be an **unbounded set**.

(4) Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty set X. Let $A = f[X]$ denote the image of X under f; that is, $A = \{y \in \mathbb{R} : y = f(x) \text{ for at least one } x \text{ in } X\}$. One says that the function f is **bounded above on** X provided the set A is bounded above, in the sense of Part (2) above. Likewise, f is said to be **bounded below on** X if A is bounded below. Finally, f is said to be **bounded on** X if A is a bounded set, in the sense of Part (3).

The next result is simple, and 'obviously true'. It is included mainly for future reference.

II.3.13 Theorem

Let A be a subset of $\mathbb{R}$.

(a) If A is a finite set, then A is is bounded in $\mathbb{R}$.

(b) If A is bounded above in $\mathbb{R}$, then every subset of A is bounded above in $\mathbb{R}$.

(c) Suppose that A can be expressed as the finite union $A = A_1 \cup A_2 \cup \ldots \cup A_n$ of subsets of A. If each of these subsets is bounded above in $\mathbb{R}$, then A is bounded above in $\mathbb{R}$.

(d) Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty set X. Suppose that X can be expressed as the finite union $X = X_1 \cup X_2 \cup \ldots \cup X_n$ of subsets of X. If f is bounded above in $\mathbb{R}$ on each of the subsets $X_1, X_2, \ldots X_n$, then f is bounded above in $\mathbb{R}$ on X.

(e) If, in Parts (b), (c) and (d) above, each occurance of 'bounded above' is replaced by 'bounded below', then the resulting statements are true. Likewise, if each such occurance is replaced by 'bounded', then the resulting statements are true.

The simple proof is left as an exercise.

II.3.14 Remarks

(1) The phrasing used in Part (1) of the preceding definition makes it clear that if $A = \emptyset$, then *every* real number M is an upper bound of A. Indeed, the 'equivalent' formulation given there of M being an upper bound of A is that there not exist x in A with a given property, namely that $x > M$. But if $A = \emptyset$, then there does not exist x in A, with or without the given property, so M is an upper bound of A. A similar argument shows that if $A = \emptyset$, then *every* real number m is a lower bound of A. In particular, the empty set is a bounded subset of $\mathbb{R}$.

(2) Some authors phrase the definition of a set A being a bounded subset of $\mathbb{R}$ as follows: There exists a real number $B \geq 0$ such that $|x| \leq B$ for every x in A. The proof that this is equivalent to the definition given above is left as a simple exercise.

II.3.15 Corollary

(a) Let A be a subset of $\mathbb{R}$, and suppose that B is a subset of A such that $A \setminus B$ is a finite set. If B is bounded above in $\mathbb{R}$, or B is bounded below in $\mathbb{R}$, or B is bounded in $\mathbb{R}$, then A has the corresponding boundedness property.

(b) Likewise, let $f : X \rightarrow \mathbb{R}$ be a real-valued function defined on a nonempty set X. Suppose that there exists a subset Y of X such that $X \setminus Y$ is a finite set. If f is bounded above in $\mathbb{R}$ on Y, or f is bounded below in $\mathbb{R}$ on Y, or f is bounded in $\mathbb{R}$ on Y, in the sense of Part (4) of Definition (II.3.12) above, then f has the corresponding boundedness property on X.

Proof (a) This follows using Parts (c), (d) and (e) of the preceding theorem by noting that $A = B \cup (A \setminus B)$.

(b) This follows easily from Part (a).

Part (4) of Definition (II.3.12) includes the special case $X = \mathbb{N}$; that is, the case of sequences of real numbers. The following result often simplifies the determination of whether a given real sequence is has one of the 'boundedness' properties.

II.3.16 Theorem

Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be a sequence of real numbers.

(a) A necessary and sufficient condition for the sequence ξ to be bounded above is that there exist a real number M such that

$$x_k \leq M \text{ for all but finitely many values of the index } k;$$

equivalently, the sequence is eventually bounded above by M. Similarly, a necessary and sufficient condition for the sequence ξ to be bounded below is that there exist a real number m such that

$$x_k \geq m \text{ for all but finitely many values of the index } k;$$

equivalently, the sequence is eventually bounded below by m.

(b) A necessary and sufficient condition for ξ to be a bounded sequence (i.e., to be bounded above and below) is that there exist a real number $M > 0$ such that $|x_k| \leq M$ for all but finitely many values of k.

Proof (a) This follows directly from Part (b) of Corollary (II.3.15).

(b) See Part (b) of Remark (II.3.14) above.

Notation: In *This Textbook* it convenient to denote the set of all upper bounds of a nonempty subset A of $\mathbb{R}$ by the symbol U_A. Likewise the set of all lower bounds is denoted L_A. (This notation is not standard in mathematics.) Then stating that A bounded above is equivalent to the statement $U_A \neq \emptyset$, and similarly for 'bounded below'.

II.3.17 Examples

(1) Let a and b be real numbers such that $a < b$. Then the closed interval $[a, b]$ and the open interval (a, b) are both bounded sets. Indeed, for each set it is clear that a is a lower bound and b an upper bound.

Note that $U_{[a,b]} = U_{(a,b)} = [b, +\infty)$; likewise, $L_{[a,b]} = L_{(a,b)} = (-\infty, a]$. In particular, the number b is simultaneously the right-hand endpoint of each of the sets $[a, b]$ and (a, b), and the least element of the upper bounds of the sets $[a, b]$ and (a, b). Similarly, the number a is simultanously the left-hand endpoint and the greatest of the lower bounds for the sets $([a, b])$ and (a, b).

(2) The set $\mathbb{R}^+$ of all positive real numbers is bounded below but unbounded above. Indeed, it is clear that 0 is a lower bound for $\mathbb{R}$. In contrast, there is no upper bound for $\mathbb{R}^+$; for if M were such a number then one would have $M \geq x$ for all x in $\mathbb{R}^+$. In particular one would have $M \geq M + 1$, which is impossible. But then $M + 1$ would also be in $\mathbb{R}^+$, so that M would have to satisfy $M \geq M + 1$, which is impossible.

Note that $U_{\mathbb{R}^+} = \emptyset$ while $L_{\mathbb{R}^+} = (-\infty, 0]$.

(3) The set $\mathbb{N}$ of all natural numbers, viewed as a subset of $\mathbb{R}$, is bounded below. Indeed, it is clear that $L_{\mathbb{N}}$ is the interval $(-\infty, 1]$. In contrast, the axioms for an ordered field are not enough to guarantee that the subset $\mathbb{N}$ of $\mathbb{R}$ is unbounded above in $\mathbb{R}$.

The condition for a subset of $\mathbb{R}$ to be a bounded set is frequently described in the following equivalent ways.

II.3.18 Theorem

(a) A necessary and sufficient condition for a nonempty subset A of $\mathbb{R}$ to be a bounded set in $\mathbb{R}$ is that there exist a nonnegative real number B such that $|x| \leq B$ for all x in A.

(b) A necessary and sufficient condition for a nonempty subset A of $\mathbb{R}$ to be a bounded set in $\mathbb{R}$ is that there exist a real number c and a nonnegative real number B such that $|x - c| \leq B$ for all x in A.

Equivalently, a necessary and sufficient condition for a nonempty subset A of $\mathbb{R}$ to be a bounded set in $\mathbb{R}$ is that for *every* real number c there exist a nonnegative real number B such that $|x - c| \leq B$ for all x in A.

The simple proof is left as an exercise.

> Side Comment (on complicated definitions) The conditions stated in Part (b) of the preceding theorem may seem strange. However, these formulations of the concept of 'bounded subset of $\mathbb{R}$' have a major advantage which is invisible in the context of the main body of *This Textbook*: they easily extend to the much more general context of 'bounded subsets of a metric space'. ('Metric space' is an important concept from advanced analysis; it is not needed for *This Textbook*.)
>
> It happens frequently in mathematical writing that an author chooses a more complicated formulation of a concept, in preference to a version which is simpler or more intuitive, because the complicated version extends more easily to a more general context. Whether such a choice is pedagogically wise depends on the circumstances.

The concept of a subset of $\mathbb{R}$ being 'bounded above' or 'bounded below' gives rise to some natural questions:

Suppose that A is a nonempty subset of $\mathbb{R}$ such that A is bounded above. Does A have a 'maximum element'? Likewise, if A is bounded below, does A have a 'minimum element'?
The study of these questions gives rise to some concepts that play an important role in analysis.

II.3.19 Definition

Let A be a nonempty subset of $\mathbb{R}$.

(a) A real number M is said to be **the maximum element of** A, and one writes $M = \max A$, provided M satisfies the following conditions:

(i) M is an element of A.

(ii) If $x \in A$, then $x \leq M$.

In other words, M is an upper bound of the set A which is also an element of A.

(b) Likewise, a real number m is said to be **the minimum element of** A, and one writes $m = \min A$, provided m satisfies the following conditions:

(i') m is an element of A.

(ii') If $x \in A$, then $x \geq m$.

In other words, m is a lower bound of the set A which is also an element of A.

(c) Important Special Case: Suppose that $f : X \to \mathbb{R}$ is a real-valued function defined on a nonempty set X, and let $A = f[X]$; see Definition (I.5.9). If the set A has maximum element M then one calls M the **maximum value of the function *f* on the set *X***. Similarly, if A has minimum element m then one calls m the **minimum value of the function *f* on the set *X***. The numbers M and m are called the **extreme values of *f* on *X***.

Since the minimum and maximum values m and M decribed above are actual *values* of f on the set X, there must exist at least one element c of X such that $f(c) = m$, and at least one element d of X such that $f(d) = M$. One then says that ***f* assumes its maximum value for *X* at *c*, and it assumes its minimum value for *X* at *d*.** . Likewise, one says that ***f* has a local minimum at *c*** provided there is an open interval I containing c such that f has its minimum value for the set $X \cap I$ at c. The concept of **local maximum** is defined similarly..

II.3.20 Remarks

(1) Conditions (ii) and (ii') are obviously equivalent to the following conditions:

Modified (ii) If $x \in A$ and $x \neq M$, then $x < M$.

Modified (ii') If $x \in A$ and $x \neq m$, then $x > m$.

Although these formulations may come closer to the intuitive concept of M being bigger (or m being smaller) than every other number in the set, the original formulation turns out to be easier to use in practice.

(2) It is clear from the order properties that a nonempty subset can have at most one maximum and at most one minimum, so the use of the word 'the' in the phrases '*the* maximum element' and '*the* minimum element' is justified. For example, if M_1 and M_2 are elements of A such that $M_1 \neq M_2$, then the smaller of these numbers cannot also satisfy Condition (ii). Of course there are many bounded nonempty subsets of $\mathbb{R}$ which have neither a maximum element nor a minimum element; see the examples below.

(3) It is incorrect to write statements such as $\max(\mathbb{R}) = +\infty$ or $\min(\mathbb{R}) = -\infty$. By definition, the max and min of a set A of real numbers must be elements of that set; but $+\infty$ and $-\infty$ are not real numbers and thus can't be elements of the set A.

(4) Suppose that the set A is a finite set, so that $A = \{x_1, x_2, \ldots x_n\}$ for some numbers x_1, $x_2, \ldots\, x_n$. Then one often writes $\max\{x_1, x_2, \ldots x_n\}$ in place of $\max A$, and $\min\{x_1, x_2, \ldots x_n\}$ in place of $\min A$.

(5) On occasion one uses the words 'greatest' or 'largest' instead of 'maximum'; likewise, one uses 'least' or 'smallest' instead of 'minimum'.

II.3.21 Examples

(1) Let a and b be real numbers with $a < b$, and let A be the closed interval $[a, b]$ in $\mathbb{R}$. Thus $A = \{x \in \mathbb{R} : a \leq x \leq b\}$. Clearly b satisfies Condition (i) of Definition (II.3.19), since $x = b$ is a special case of $x \leq b$. Likewise, Condition (ii) is satisfied. That is, the closed interval $[a, b]$ has a maximum element, and it is the right endpoint b of this interval. Similarly, the left endpoint a is the minimum element of $[a, b]$.

(2) Let a and b be as in the preceding example, but now let A be the *open* interval (a, b). It is clear that the right endpoint b still satisfies Condition (ii) for being the maximum of A, but it fails Condition (i). Indeed, the open interval (a, b) has neither a maximum element nor a minimum element.

(3) Let A be the set of all real numbers of the form $1 - \dfrac{1}{n}$ for n in $\mathbb{N}$; that is,

$$A = \left\{0, \frac{1}{2}, \frac{2}{3}, \ldots \frac{n-1}{n}, \ldots\right\}.$$

(Recall that we interpret rational numbers as being elements of the set $\mathbb{R}$.)

It is clear that the set A has a minimum element, namely the number 0. Indeed, 0 is in the set A, since it corresponds to the case of $n = 1$ in the definition of A. And it is clear that $0 \leq 1 - 1/n$ for all n in $\mathbb{N}$, since $n \geq 1$ implies $0 < 1/n \leq 1$.

In contrast, the set A has no maximum element. In fact, if a number x satisfies Condition (i), then it cannot satisfy Condition (ii). More precisely, such x must be of the form $x = 1 - 1/n$ for some n in $\mathbb{N}$; but clearly $1 - 1/(n+1)$ is an element of A larger than x.

(4) Let $A = \{-2, 3, 7, 1, 0, -10\}$. It is clear by inspection that the smallest element of this set is the number -10, while the largest is the number 7. That is, $\min A = -10$, $\max A = 7$.

(5) Let $f : \mathbb{R} \to \mathbb{R}$ be the function given by the equation

$$f(x) = x\,(1-x) \text{ if } |x| < 1; \qquad f(x) = 0 \text{ if } |x| \geq 1.$$

It is clear that if $0 < x < 1$, then $0 < 1 - x < 1$, hence $0 < x\,(1-x) < 1$. In particular, $f(x) \geq 0$ for all x in $\mathbb{R}$, and f assumes its minimum value of 0 for $\mathbb{R}$ at every x such that $|x| \geq 1$. As for the maximum value of f, recall the well-known method of 'Completing the Square', from high-school algebra, to get

$$x\,(1-x) = x - x^2 = \frac{1}{4} - \frac{1}{4} + x - x^2 = \frac{1}{4} - \left(x - \frac{1}{2}\right)^2 \qquad (*)$$

It is clear that the right side of $(*)$ assumes its maximum value of $1/4$ when $x = 1/2$.

II.4 Completeness of the Real Numbers

It has been noted several times that the results obtained so far in this chapter apply to every ordered field. In particular, these results by themselves cannot distinguish between the ordered fields $\mathbb{Q}$ and $\mathbb{R}$. Among the differences between just these two fields are these:

(i) In Chapter (I) it was proved that the set $\mathbb{Q}$ is countable, while the set $\mathbb{R}$ is uncountable.

(ii) It is well known that the algebraic equation $x^2 = 2$ has a solution in $\mathbb{R}$, but does not have one in $\mathbb{Q}$. (This is usually expressed as 'the number $\sqrt{2}$ is irrational'.)

The aim of the current section is to formulate additional properties which distinguish $\mathbb{R}$ from all other ordered fields. For various reasons the issues considered in this section fall under the general heading of the **Completeness of the Real Number System $\mathbb{R}$**.

Remark: Note that the axioms considered so far, namely A0-A6, O1 and O2, enjoy several pleasant properties: they are easy to state; the properties of $\mathbb{R}$ that they describe are familiar; and everyone believes that $\mathbb{R}$ has these properties. Even the minor differences in presentation of these axioms that one can find in various texts are mainly cosmetic.

In contrast, the 'Completeness' aspects of $\mathbb{R}$ considered in the current section are much less straight forward than the algebraic and order properties considered before. For one thing, there are several quite different approaches to 'Completeness' in the mathematical literature, and at first glance these approaches may not appear to be at all closely related. In addition, most of these approaches requires a fair amount of preparation before the formulation can be carried out. Finally, to a greater or lesser degree each of these approaches is based on properties of real numbers which are *not* especially familiar. Indeed, many readers will not have considered the issue of 'Completeness' before, so whatever we end up adding to our axiom system for $\mathbb{R}$ in order to obtain 'completeness' can hardly be considered 'axiomatic', in the common meaning of that word as being 'obviously true'.

> Side Comment(on motivating the concept of 'Completeness') As motivation for the approach to 'Completeness' to be followed here, it helps to look at the situation in geometric terms. Thus, think of the real number system $\mathbb{R}$ as forming the standard x-axis of analytical geometry; that is, as a straight line. Within this line there lies the subset $\mathbb{Q}$ of all rational numbers. To the 'naked eye' the sets $\mathbb{Q}$ and $\mathbb{R}$ look very much alike: each consists of lots and lots of 'dots' that are spread uniformly throughout the line.
>
> Of course one knows intellectually that these sets are quite different – as mentioned above, $\mathbb{Q}$ is countable, $\mathbb{R}$ is not – but cardinality is not a precise enough concept for this discussion. The true issue is that in this geometric interpretation the set $\mathbb{Q}$ has lots of 'holes'; for example, there is a 'hole' in the rationals where the number $\sqrt{2}$ ought to appear. (Each such 'hole' in $\mathbb{Q}$ is infinitely small – only a single point wide – but there are uncountably many of them.) In contrast, the straight line (i.e., $\mathbb{R}$) is, in our intuition, 'continuous' and thus has no such 'holes'. Otherwise stated, there is no need to adjoin any new points to the standard set $\mathbb{R}$ in order to make it 'complete', i.e., free of 'holes'. Otherwise stated, $\mathbb{R}$ is 'complete'.

Remark In the various approaches to completeness given below, several candidates for additional axioms, called 'Principles', are proposed. Each of these 'Principles' makes sense in any ordered field; in particular, each makes sense in both $\mathbb{Q}$ and $\mathbb{R}$, the ordered fields of primary interest in *This Textbook.* (Of course, 'making sense' in any ordered field is not the same as 'being true'; for example, all of the Principles expounded below make sense in $\mathbb{Q}$, but none are valid in $\mathbb{Q}$.) Thus in the statements of these 'Principles', unless explicitly stated otherwise, the word 'number' can refer to an element of any given ordered field; but there is no harm in restricting the meaning of 'number' to either 'rational number' or 'real number'. Several of these approaches are grouped together for efficiency since they are so similar to each other.

Approach #1: Using Bolzano's Endpoint Principles

In the year 1817, Bernhard Bolzano published a pamphlet on the foundations of calculus. In it he singled out for special attention a certain property which he claimed to hold for the real number system. The reformulation of that property given here is in terms of certain subsets of $\mathbb{R}$, a mode of expression that is natural today, but which was not common in 1817. The formulation given here differs a bit from Bolzano's, but not in any essential way.

Using the terminology developed in the preceding section, Bolzano's approach to 'Completeness' can be described as clarifying the relation between the concepts of 'convexity' and 'interval'. In doing so, Bolzano finds a way to distinguish between the general ordered field and the specific case of the real numbers. There is one case, however, in which this relation is the same for all ordered fields:

II.4.1 Proposition

If S is a convex set of numbers which is unbounded both above and below, then S is the set of all numbers; in particular, S is the interval $(-\infty, +\infty)$.

Proof Let z be any number. Since, by hypothesis, S is unbounded both above and below, there exist numbers x and y in S such that $x < z < y$. It then follows from the convexity hypothesis for S that $z \in S$ as well. That is, every number is in S, as claimed.

Bolzano's original approach focuses on convex sets S of numbers which are bounded above but not bounded below. Specifically, the question is whether such a set must be an interval, in the sense of Definition (II.3.1). More precisely, the question is whether there exists a number B such that S is one of the intervals $(-\infty, B]$ or $(-\infty, B)$. (If S is to be an interval, than the hypothesis that S be unbounded below requires that it extend on the left to $-\infty$.)

Key Example Let S be the set consisting of all numbers $x \geq 0$ such that $x^2 < 2$, together with all negative numbers. It is clear from the order properties of multiplication that the set S is convex and bounded above; for example, the number 3 is an upper bound. Obviously it is also unbounded below. The question therefore is whether this convex set is actually an interval in the given ordered field.

At first the answer appears obvious: clearly S is the open interval $(-\infty, \sqrt{2})$. However, the existence of a number whose square equals 2 cannot be deduced from the axioms for an ordered field. Indeed, if that existence could be so deduced, then since the rational numbers also form an ordered field, it would follow that there would exist a *rational* number whose square equals 2, which of course is not the case. Thus we need one or more additional axioms for $\mathbb{R}$ which allow us to deduce facts such as the existence of square roots.

Bolzano's analysis of this situation suggests three equivalent principles which would hold for $\mathbb{R}$, although not for $\mathbb{Q}$.

II.4.2 Bolzano's Right-Endpoint Principle

If S is a convex set of numbers which is bounded above but not bounded below, then there exists a number B such that S is an unbounded interval with right endpoint B. That is, either $S = (-\infty, B)$ or $S = (-\infty, B]$.

II.4.3 Bolzano's Left-Endpoint Principle

If S is a convex set of numbers which is bounded below, but not bounded above, then there exists a number A such that S is an interval with left endpoint A. That is, either $S = (A, +\infty)$ or

$S = [A, +\infty)$.

II.4.4 Bolzano's Two-Endpoints Principle

If S is a bounded convex set of numbers which has more than one element, then there exist numbers A and B, with $A < B$, such that S is an interval with left endpoint A and B . That is, one of the following statements holds: $S = (A, B)$, $S = [A, B)$, $S = (A, B]$ or $S = [A, B]$.

Remark It is easy to see that each of these principles implies the others, so only one is needed. Bolzano explicitly formulated the right-endpoint principle (although in slightly different form; see below.) However, it appears that he implicitly accepted the others as well.

Side Comment (on Bolzano's formulation of the Right-Endpoint Principle)
Bolzano's formulation of the principle in question essentially as follows:

Suppose that a certain property P of real numbers is not enjoyed by all numbers x. Suppose, in addition, there exists a number u such that the Property P is satisfied by all $x < u$. Then there exists a *largest* number U such that Property P is satisfied by all $x < U$.

It is easy to show that Bolzano's formulation of the principle is equivalent to the convex-set version given in Definition (II.4.3); that is, each formulation implies the other.

First, suppose that Bolzano's version of the principle holds. Let S be a set which satisfies the hypotheses of the convex-set version of the principle, and define P to be the property 'is less than or equal to at least one element of the set S'; note that since convex sets are, by definition, nonempty, the property P is satisfied by at least one number. To see that P satisfies Bolzano's first hypothesis, let M be an upper bound for the set S; by hypothesis, such M exists. Clearly, if $y > M$, then y does not enjoy Property P. Next, let u be any element of the set S. The definition of P implies that if $x < u$ then x satisfies Property P. That is, Property P satisfies Bolzano's second hypothesis. Let U be the number whose existence Bolzano's version of the principle now guarantees. It is easy to check that $B = U$ satisfies the conclusion of the convex-set version of the principle.

Conversely, suppose that the convex-set version of the principle holds. Let a property P be given which satisfies Bolzano's hypotheses, and let S be the set consisting of all numbers u such that if $x < u$, then x enjoys Property P. Bolzano's second hypothesis implies that S is nonempty. Furthermore, let M be a number which does not enjoy Property P; such M exists by Bolzano's first hypothesis. It follows that if $u \in S$, then M cannot satisfy $M < u$, for otherwise M would satisfy Property P. Thus $M \geq u$; that is, S is bounded above by M. To show that S is convex, suppose that u and v are in S, and let $w = \max\{u, v\}$. If $w = u$ then, by definition of the set S, every $x < w$ satisfies Property P; the same conclusion holds if $w = v$. It follows that if $z \leq w$ then every $x < z$ satisfies $x < w$ and therefore lies in S. It follows eaily that $\mathrm{Seg}\,[u, v] \subseteq S$, as required. Thus the convex-set version implies that there is a number B such that S is an interval with right endpoint B. Clearly the number $U = B$ satisfies Bolzano's conclusion.

Remark Bolzano's formulation of the principle is, in a sense more general than the convex-set version, since the set of x which satisfy Property P need not form a convex set at all. For example, let $f : \mathbb{R} \to \mathbb{R}$ be given by the formula $f(x) = 2 - x^2$ for eac x, and let P be the property 'x is a number such that $f(x) < 0$'. Property P clearly satisfies Bolzano's hypotheses; indeed, $x = 0$ does not satisfy Property P, while $x = -2$ can be used as Bolzano's u (as can many other values of x). Nevertheless, the set of x which satisfy Property P is not convex, since clearly $x = -2$ and $x = +2$ both satisfy it, but $x = 0$ does not.

In this case it is easy to see that the quantity U referred to in Bolzano's formulation is the number $-\sqrt{2}$; that is, U is the smaller of the two solutions of the quadratic equation $f(x) = 0$. Indeed, it was the analysis of such equations that led Bolzano to his formulation of the principle.

The 'Bolzano' approach to completeness described above has been largely replaced in modern texts by the following.

Approach #2: Using the Supremum Principle or the Infimum Principle

Bolzano's Right-Endpoint Principle asserts, in effect, that if S is a certain type of convex set, then S is an interval with a right endpoint. It is easy to use Bolzano's idea to extend the concept of 'right endpoint' to *every* nonempty subset X of numbers which is bounded above. Indeed, associate with such X the set S_X of all numbers y such that $y \leq x$ for at least one element x in X. It is clear from the hypotheses on X, and the usual order properties, that S_X is a nonempty superset of X which is convex, bounded above and not bounded below. If S_X has a right endpoint B in the sense of Bolzano's Principle, then one calls B the right endpoint of the original set X; and of course if S_X does *not* have a right endpoint in the sense of Bolzano's Principle, then one says that X also has no right endpoint.

Remark Similar statements hold for Bolzano's Left-Endpoint Principle; but since the two principles are so closely related, we restrict our attention for now to the 'Right-Endpoint' case, and include the 'Left-Endpoint' case in Definition (II.4.6).

It is useful to characterize the 'right endpoint' of X described above directly, without needing to introduce the auxiliary convex set S_X.

II.4.5 Lemma

Let the sets X and S_X be as above.

(1) Suppose that S_X has a right endpoint in the sense of Bolzano; that is, there is a number B such that either $S_X = (-\infty, B)$ or $S_X = (-\infty, B]$. Then the number B satisfies the following conditions:

(i) The number B is an upper bound of the original set X;
(ii) For every number $y < B$ there exists at least one x in X such that $y < x$.

(2) Conversely, if there exists a number B which satisfies Conditions (i) and (ii), then B is the right endpoint of S_X in the sense of Bolzano.

Proof Part (1) is obviously true, so let us prove only Part (2). Indeed, let S_X be the set of all numbers y such that $y \leq x$ for at least one x in X, so that the set S_X is a convex superset of X which is unbounded below. If $y \in S_X$ then, by definition of S_X, Condition (i), and the transitivity of '$<$', it follows that $y \leq B$; that is, the convex set S_X is bounded above, by B. Furthermore, if $y < B$ then, by Condition (ii) there exists x in X such that $y < x$, so in particular $y \leq x$ and thus $y \in S_X$. It follows that the open interval $(-\infty, B)$ is a subset of S_X. If $S_X = (-\infty, B)$, we are done. If, instead, $S_X \neq (-\infty, B)$, then there must be at least one number z such that $z \in S_X$, and thus $z \leq B$, but $z \notin (-\infty, B)$, and thus $z \geq B$. In this situation it follows that $z = B$, so that $S_X = (-\infty, B) \cup \{B\} = (-\infty, B]$, as required.

The preceding result suggests calling the number B satisfying (i) and (ii) to be the 'right endpoint' of the set X, and doing so would be fine. Likewise, the obvious modification of the preceeding discussion would allow one to use Bolzano' 'Left-Endpoint Principle' to define the left endoint of any nonempty set which is bounded below. However, the custom in modern analysis is to restrict the 'right endpoint' and 'left endpoint' terminology to the special case of intervals. For sets of more general type the following Latin-based terminology is preferred.

II.4.6 Definition (Supremum, Infimum)

Let X be a nonempty subset of numbers.

(a) Suppose that there exists a number B satisfying the following conditions:

Condition (i) The number B is an upper bound of X; that is, if $x \in X$, then $x \leq B$.
Condition (ii) For every number $y < B$ there exists at least one x in X such that $y < x \leq B$

If such a number B exists, it is called the **supremum of $\boldsymbol{X}$**, and is denoted by $\sup X$.

(b) Likewise, suppose that there exists a number b satisfying the following conditions:

Condition (i′) The number b is a lower bound of X; that is, if $x \in X$, then $x \geq b$.
Condition (ii″) For every number $y > b$ there exists at least one x in X such that $b \leq x < y$.

If such a number b exists, it is called the **infimum of $\boldsymbol{X}$**, and is denoted by $\inf X$.

The following are the analogs to Bolzano's 'Right-Endpoint Principle' and 'Left-Endpoint Principle'.

II.4.7 The Supremum Principle

If X is a nonempty set of numbers such that X is bounded above, then X has a supremum.

II.4.8 The Infimum Principle

If X is a nonempty set of numbers such that X is bounded below, then X has an infimum.

II.4.9 Remarks

(1) Condition (ii) in Part (a) Definition (II.4.6) is called the **Approximation Property of the Supremum**. Likewise, Condition (ii′) in Part (b) of the same definition is called the **Approximation Property of the Infimum**. The reason for this terminology will be make clear in an exercise.

(2) The use of the phrase '*the* supremum' suggests that if a set X has a supremum, it is unique; likewise for '*the* infimum'. It is easy to show that these suggestions are correct.

(3) The Latin nouns 'supremum' and 'infimum' translate into ordinary English as the phrases 'the greatest one' and 'the least one', respectively. Note that the Latin nouns 'maximum' and 'minimum' also translate to exactly the same phrases. In mathematics, however, there is a subtle technical difference which does not hold in ordinary English usage.

Example If one writes that 'the number B is the maximum of the set S', it is implied that B is actually an element of the set S. In contrast, to say that 'the number B is the supremum of the set S' does *not* imply that B is an element of S; consider, for instance, the case in which S is the open interval $(1, 2)$, so that $\sup S = 2$, but 2 is not an element of the set. In particular, the Supremum Principle does *not* state that a nonempty set S which is bounded above must have a maximum element. Similar comments hold for the (subtle) disinction between 'minimum' and 'infimum'.

(4) Because of the Latin roots of the nouns 'supremum' and 'infimum', the custom in mathematics is to use the Latin formations of their plurals: 'suprema' and 'infima' respectively. A

minority of authors, however, use the English versions of the plurals: 'supremums' and 'infimums'. Similarly, one usually writes 'maxima' and 'minima', not 'maximums' and 'minimums'.

(5) One pronounces the 'su' in the expression '$\sup X$' the same as the 'su' found in the word 'super', *not* as the 'su' found in the word 'submarine'. Likewise, the 'in' found in the expression '$\inf S$' is pronounced the same as the 'in' found in the word 'into'.

Approach #3: Using the Least-Upper Bound Principle or the Greatest-Lower-Bound Principle

The next pair of principles require no further preparation for their statements.

II.4.10 The Least-Upper-Bound Principle

Let X be a nonempty set of numbers which is bounded above. Then X has a *least* upper bound, denoted $\operatorname{lub} X$. More precisely, let U_X denote the set of all the upper bounds of X (so that U_X is nonempty by the hypothesis that X is bounded above.) Then the set U_X has a minimum (i.e., least) element.

II.4.11 The Greatest-Lower-Bound Principle

Suppose that X is a nonempty set of numbers which is bounded below. Then X has a *greatest* lower bound, denoted $\operatorname{glb} X$. More precisely, let L_X denote the set of all the lower bounds of X (so that L_X is nonempty by the hypothesis that X is bounded below.) Then L_X has a maximum (i.e., greatest) element.

Remark The expressions 'lub' and 'glb' are both pronounced to rhyme with the English word 'flub'.

II.4.12 Examples

Suppose the ordered field under consideration is $\mathbb{Q}$, the ordered field of rational numbers. Then some bounded sets have both a greatest lower bound and a least upper bound in $\mathbb{Q}$, some have neither, and some have one but not the other.

(1) Let $S_1 = \{x \in \mathbb{Q} : 0 \leq x < 1$. It is easy to see that S_1 has both a greatest lower bound and a least upper bound, namely $b = 0$ and $B = 1$, respectively. Note that $\operatorname{glb} S_1$ is an element of S_1 while $\operatorname{lub} S_1$ is not.

It is also instructive to check directly from the definition of 'supremum', Definition (II.4.6) (a), that $\sup S_1 = 1$. Indeed, it is obvious that $B = 1$ satisfies Part (i) of that definition. As for Part (ii), let y be any number such that $y < 1$. If $y < 0$ then *every* x in S_1 satisfies $y < x$. And if $0 \leq y < 1$ then $x = (y+1)/2$ is in S_2 and satisfies $y < x$. A similar argument shows directly from Part (b) of the same definition that $\inf S_1 = 0$.

(2) Let $S_2 = \{x \in \mathbb{Q} : x \geq 0 \text{ and } x^2 < 2\}$. Obviously S_2 is bounded above (by 3, for example), and clearly this set has a greatest lower bound, namely $b = 0$. Suppose that S_2 has a least upper bound B in $\mathbb{Q}$. It may appear to be an 'obvious fact' that such B must satisfy the equation $B^2 = 2$; this would then contradict the well-known fact that there is no rational number whose square equals 2. Unfortunately, the 'obvious fact' in question is not especially easy to prove directly. Such facts will be handled in a much more general manner in Chapter (IV); thus there is no harm

in simply accepting this 'obvious fact', and therefore the conclusion that the ordered field $\mathbb{Q}$ does not satisfy the Least-Upper-Bound Principle, on faith for now.

Warning Many texts in analysis do not distinguish between the concepts of 'supremum of X' and 'least upper bound of X' as we do in *This Textbook*. More precisely, they *define* the expression '$\sup X$' to mean the least upper bound of the set X (when such exists). In other words, in these texts the Latin word 'supremum' is simply an abbreviation for – and a bad translation of – the English phrase 'least upper bound of X'.

There are several difficulties with this usage. The most glaring issue is that there is a well-established shorthand for 'least upper bound of X', namely '$\mathrm{lub}\, X$'; this notation is as brief as the 'sup' notation, and it is certainly more natural.

Another problem is that this definition does not mention the key feature of 'supremum; namely Condition (ii) of Definition (II.4.6), what we refer to as the 'Approximation Property of the Supremum'. These texts eventually do prove that this feature follows from the 'least upper bound' property, and of course they use it repeatedly. Some texts even refer to this feature as the 'Approximation Property', but many give it no name at all.

Perhaps the most serious difficulty with this common usage, however, is that it forces a cognitive dissonance into the discussion. More precisely, the notation $C = \sup X$ in those texts refers to the property of C being the *least* number of a certain set – and not even the given set X itself – whereas the root meaning of the Latin word 'supremum' is *greatest*, which is much closer to the intuition of 'right endpoint of X'.

In *This Textbook* we avoid that cognitive dissonance by carefully maintaining the conceptual distinction between C being a right endpoint, i.e., the supremum, of X, and the same C simultaneously being a least element, i.e., the least element of the set U_X of upper bounds of X. Otherwise stated, sometimes it is useful to look at a given number C 'from the right', while sometimes it is more useful to look at C 'from the left'.

The next several results clarify and extend the concepts just introduced.

II.4.13 Lemma

Let X be a nonempty set of numbers, and let $Y = \{y : y = -x$ for some number x in X. Let U_X be the set of upper bounds of X; likewise, let L_Y be the set of lower bounds of Y.

(a) The set L_Y is nonempty if, and only if, U_X is nonempty; that is, Y is bounded below if, and only if, X is bounded above. Furthermore, if this happens, then $L_Y = \{l : l = -u$ for some number u in $U_X\}$.

(b) Suppose that such boundedness occurs. Then X has a least upper bound if, and only if, Y has a greatest lower bound. Furthermore, in this case one has $\mathrm{glb}\, Y = -\mathrm{lub}\, X$. Of course, this is equivalent to $\inf Y = -\sup X$.

(c) If in Part (b) 'least upper bound' and 'greatest lower bound' are replaced throughout by 'supremum' and 'infimum', respectively, then the resulting statements remain true.

Proof (a) Suppose that $u \in U_X$, so that $u \geq x$ for all x in X. Then $-u \leq -x$ for all x in X, hence $l = -u$ is a lower bound for Y. Similarly, if $l \in L_X$ then $l \leq y$ for all y in Y; that is, $l \leq -x$, and thus $-l \geq x$, for all x in X. The desired result now follows.

(b) This follows by a similar argument.

(c) This is left as an exercise.

II.4.14 Theorem

(a) Let X be a nonempty set of numbers. If X has both a greatest lower bound and a least upper bound, then $\mathrm{glb}\, X \leq \mathrm{lub}\, X$, with equality if, and only if, X is a singleton set.

(b) Let X and Y be nonempty sets of numbers such that $X \subseteq Y$. If X and Y both have least upper bounds, then $\mathrm{lub}\, X \leq \mathrm{lub}\, Y$. Likewise, if both X and Y have greatest lower bounds, then $\mathrm{glb}\, X \geq \mathrm{glb}\, Y$.
Warning Note the reversal of the order which occurs in the 'glb' case.

(c) Let X and Y be nonempty sets of numbers such that $x \leq y$ for all x in X and all y in Y. If X has a least upper bound and Y has a greatest lower bound, then $\mathrm{lub}\, X \leq \mathrm{glb}\, Y$.

(d) If in the previous parts 'least upper bound' and 'greatest lower bound' are replaced throughout by 'supremum' and 'infimum', respectively, then the resulting statements remain true.

Proof The simple proof is left as an exercise.

It is easy to extend the concepts of 'supremum and 'infimum, and likewise the concepts of 'least upper bound' and 'greatest lower bound', to include *unbounded* sets.

II.4.15 Definition

(1) Let X be a nonempty set of numbers. If X is unbounded above, then one sets $\sup X = +\infty$ and $\mathrm{lub}\, X = +\infty$. Likewise, if X is unbounded below, then one sets $\inf X = -\infty$ and $\mathrm{glb}\, X = -\infty$.

(2) Let Y be a nonempty set of numbers. One says that ***Y* has a supremum** if either Y has a supremum in the sense of Definition (II.4.6) (in which case Y must be bounded above) or Y is unbounded above (in which case $\sup Y = +\infty$).
One defines the meaning of 'Y has an infimum' analogously.

(3) One defines the meaning of 'Y has a least upper bound' and 'Y has a greatest lower bound' by replacing 'supremum' and 'infimum' above by these phrases in the usual manner.

II.4.16 Remarks

(1) The Supremum, Infimum, Least-Upper-Bound and Greatest-Lower-Bound Principles described above are all formulated for nonempty sets which are bounded either above or below, depending on the specific case. With the extension of the concepts just given, one could reformulate these principles more briefly; for example, with the extended concept of 'supremum' the original Supremum Principle is equivalent to the following statement:

'Every nonempty set of real numbers has a supremum.'

Despite the temptation to use these new formulations, in *This Textbook* we elect to follow convention and stay with the original formulations above.

(2) The 'Approximation Properties' for Supremum and Infimum, see Part (1) of Remark (II.4.9), hold for the extended notions of 'supremum' and 'infimum' described above: replace the numbers B and b in those conditions by $+\infty$ and $-\infty$, respectively.

(3) Some authors extend the concepts of 'supremum' and 'infimum' (and likewise 'least upper bound' and 'greatest lower bound') to apply to the empty set, while others use these concepts only

in the context of nonempty sets. In *This Textbook* the preference is to NOT apply these concepts to the empty set.

The main reason is that there is no such extension of these concepts for which all the standard properties also remain correct. For instance, since $\emptyset$ is a subset of every set of numbers, to extend the concept of 'least upper bound' to $\emptyset$ one would want, by Part (b) of the preceding theorem, $\mathrm{lub}\,\emptyset \leq \mathrm{lub}\,X$ for every set X which has a least upper bound. For this to hold one would need $\sup\emptyset = -\infty$. By a similar argument one would need $\inf\emptyset = +\infty$. However, with these values Part (a) of the same theorem would fail to hold.

Here are some standard facts about suprema and infima.

II.4.17 Theorem

Let X and Y be nonempty sets of numbers, and let W be the set of all numbers w of the form $x+y$ for some x in X and some y in Y.

(a) If each of the sets X and Y has a finite supremum, then W also has a finite supremum, and $\sup W = \sup X + \sup Y$.

(b) If each of the sets X and Y has a finite infimum, then W also has a finite infimum, and $\inf W = \inf X + \inf Y$.

(c) If in Parts (a) and (b) the words 'supremum' and 'infimum' are replaced by 'least upper bound' and 'greatest lower bound', respectively, then the resulting statements remain true.

Proof (a) For simplicity let $A = \sup X$ and $B = \sup Y$, and set $C = A + B$. Note that if $x{\in}X$ and $y{\in}B$, then $x \leq A$ and $y \leq B$, hence $x+y \leq A+B$. It follows that $A+B$ is an upper bound of the set Z. Next, let $\varepsilon > 0$ be any positive number. It follows from the properties of 'supremum' that there exist x in X and y in Y such that $A - \varepsilon/2 < x$ and $B - \varepsilon/2 < y$. Thus

$$\left(A - \frac{\varepsilon}{2}\right) + \left(B - \frac{\varepsilon}{2}\right) < x+y;\ \text{that is, } C - \varepsilon < x+y.$$

Since $w = x+y$ is in W, and ε is arbitrary, it follows that $C = \sup W$, as required.

(b) This follows from Part (a) by using Lemma (**??**)

(c) This is obvious.

The approaches to 'Completeness' described above clearly have a common origin, namely the Bolzano Right-Endpoint Principle. The final approach to be considered here is quite different in spirit.

Approach #4: The Bisection Principle

Background For thousands of years people have had the need to find rational approximations to solutions of polyomial equations; for example, to compute the square root of 2 is equivalent to solving the quadratic equation $x^2 - 2 = 0$. The procedure given below is perhaps the most obvious method, and probably has been been known since ancient times; it was certainly known by both Bolzano and Cauchy, and is called *Bolzano's Method* by many authors.

II.4.18 Example: The Square Root of 2

Motivation It is well-known that there is no rational number whose square equals 2. For the moment let us assume that such a *real* number does exist, and try to determine it as accurately as possible.

The obvious procedure for generating approximations of $\sqrt{2}$ starts by choosing positive rational numbers a_1 and b_1 such that $a_1^2 < 2$ and $b_1^2 > 2$; for instance, one might set $a_1 = 1$ and $b_1 = 3$, so that $a_1^2 = 1 < 2$ and $b_1^2 = 9 > 2$. Thus a_1 is too small to be $\sqrt{2}$, while b_1 is too big. By simple order properties it follows that $\sqrt{2}$ – assuming it exists – must lie in the interval $[a_1, b_1]$.

Now cut the interval $[a_1, b_1]$ at its midpoint $c_1 = \dfrac{a_1 + b_1}{2} = 2$ into equal halves $[a_1, c_1]$ and $[c_1, b_1]$; see Remark (II.3.6). That is, 'bisect' the original interval. Note that c_1 is also a positive rational number, hence $c_1^2 \neq 2$ (since $\sqrt{2}$ cannot be rational). Thus $\sqrt{2}$ must lie in one of the subintervals $[a_1, c_1]$ or $[c_1, b_1]$. More precisely, it lies in $[a_1, c_1]$ if $c_1^2 > 2$ and it lies in $[c_1, b_1]$ if $c_1^2 < 2$. In either case, let $[a_2, b_2]$ denote the half of $[a_1, b_1]$ containing $\sqrt{2}$.

By similar reasoning this **bisection procedure** can be used to cut $[a_2, b_2]$ into equal halves, one of which must contain $\sqrt{2}$. If one continues this bisection procedure indefinitely, one obtains an infinite sequence of intervals $[a_1, b_1]$, $[a_2, b_2], \dots [a_k, b_k], \dots$, where each a_k and b_k is rational, such that:

(i) $a_j^2 < 2 < b_j^2$ for each j in $\mathbb{N}$;

(ii) $b_{j+1} - a_{j+1} = \dfrac{b_j - a_j}{2}$ for each j in $\mathbb{N}$. That is, each interval in this sequence (after the first) is half the length of its predecessor.

One's geometric intuition about numbers then pictures these intervals as 'squeezing in' towards the desired number $\sqrt{2}$. Here is the calculation up to the case $k = 10$, based on the initial choices $a_1 = 1$, $b_1 = 3$ suggested above:

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a_k	1	1	$\frac{2}{2}$	$\frac{5}{4}$	$\frac{11}{8}$	$\frac{22}{16}$	$\frac{45}{32}$	$\frac{90}{64}$	$\frac{181}{128}$	$\frac{362}{256}$	$\frac{724}{512}$	$\frac{1448}{1024}$	$\frac{2896}{2048}$	$\frac{5792}{4096}$	$\frac{11585}{8192}$
b_k	3	2	$\frac{3}{2}$	$\frac{6}{4}$	$\frac{12}{8}$	$\frac{23}{16}$	$\frac{46}{32}$	$\frac{91}{64}$	$\frac{182}{128}$	$\frac{363}{256}$	$\frac{725}{512}$	$\frac{1449}{1024}$	$\frac{2897}{2048}$	$\frac{5793}{4096}$	$\frac{11586}{8192}$

The reader can easily verify that $a_j^2 < 2 < b_j^2$ for each $j = 1, 2, \dots 15$. The final column of the table tells one that the location of the number $\sqrt{2}$ has been specified with error no bigger than 1/8192; in particular, the error is less than 0.0002. Clearly the accuracy can be improved indefinitely by simply repeating the bisection procedure.

Note that in the preceding discussion it is *assumed* that there is a number x such that $x^2 = 2$; in other words, one assumes that $\sqrt{2}$ exists. However, a major reason that one actually *believes* that such a number does exist is because of the very existence such methods for approximating it to any desired degree of accuracy. Indeed, the intuition formed by computation like the preceding is this: The number that one seeks, $\sqrt{2}$, is the unique real number that lies in each of the intervals $[a_k, b_k]$, $k = 1, 2, 3, \dots$, constructed above. If there were no such square root, then these intervals would be squeezing in on a 'hole' in the real line; but we have been taught (probably in analytical geometry) that no such 'holes' exist. In terms of set-theoretic notation, the intersection $\bigcap_{k=1}^{\infty} [a_k, b_k]$ is a singleton set; namely, the set $\{\sqrt{2}\}$.

Another way to look at the preceding calculation, without mentioning 'holes in a line', is illustrated by expressing a_{15} and b_{15} as decimals to eight places:

$$a_{15} = \frac{11585}{8192} = 1.41418457; \quad b_{15} = \frac{11586}{8192} = 1.41430664$$

From this one sees that the number we seek, $\sqrt{2}$, has been determined through the third place to the right of the decimal point; it seems obvious that with sufficient time the procedure could be extended produce as many of the decimal digits of the desired number as one might want. Thus if one believes what one was taught in grade-school arithmetic, namely that every infinite decimal corresponds to a real number, then it seems clear that this process really is producing a real number with the desired property.

The preceding example suggests another candidate for an axiom that is enjoyed by the real numbers but not by the rationals.

II.4.19 Definition

Let $(X_1, X_2, \ldots X_n, \ldots)$ be an ordered sequence of sets. One says that this is a **nested sequence** provided for each index n one has $X_{n+1} \subseteq X_n$.

II.4.20 Examples

(1) Let $(I_1, I_2, \ldots I_n, \ldots)$ be an ordered sequence of closed bounded intervals in $\mathbb{R}$. Thus, each set I_k is of the form $[a_n, b_n]$ with $a_n < b_n$. It is easy to see, from the definition of 'interval, that this sequence is nested if, and only if, for every index n one has $a_n \leq a_{n+1}$ and $b_{n+1} \leq b_n$. Furthermore, since $a_1 \leq a_2 \leq a_n < b_n \leq b_{n-1} \leq \ldots \leq b_1$, it follows easily that $a_i < b_j$ for each pair of indices i and j.

(2) The nested sequence of intervals described above is said to be a **bisection sequence** provided that for each index n the interval $[a_{n+1}, b_{n+1}]$ is one of the halves of the preceding interval $[a_n, b_n]$. That is, if one sets $c_n = (a_n + b_n)/2$, so c_n is the midpoint of $[a_n, b_n]$, then $[a_{n+1}, b_{n+1}]$ is one of the two subintervals $[a_n, c_n]$ or $[c_n, b_n]$. (Equivalently, either $a_{n+1} = a_n$ and $b_{n+1} = c_n$, or $a_{n+1} = c_n$ and $b_{n+1} = b_n$.) One then says that this bisection sequence is **based on** the original interval $[a_1, b_1]$.

II.4.21 The Bisection Principle

Suppose that $[a_1, b_1]$, $[a_2, b_2]$, $\ldots [a_n, b_n], \ldots$ is a bisection sequence, in the sense of Definition (II.4.19). Then the intersection $\bigcap_{k=1}^{\infty} [a_k, b_k]$ of this family of intervals is a set with exactly one element. If c is that number, then one says that one obtains c by the **bisection procedure** or the **bisection method** starting with the initial interval $[a_1, b_1]$.

It is an instructive exercise to return to Example (II.4.18) above to show that the Bisection Principle does *not* hold in the or dered field $\mathbb{Q}$. Note that this exercise is not as simple as it may seem at first.

All seven of the major 'Principles' listed above – Right-Endpoint and Left-Endpoint, Supremum and Infimum, Least-Upper-Bound and Greatest-Lower-Bound, Bisection – are encountered frequently in analysis, so it is important to know them all.

The next theorem says, in effect, that all these 'Principles' are equivalent in any ordered field.

II.4.22 Theorem

The seven principles stated above, namely Bolzano's Right-Endpoint and Left-Endpoint Principles, the Supremum and Infimum Principles, the Least-Upper-Bound and Greatest-Lower-Bound Priniples and the Bisection Principle, are equivalent. That is, if one of these principles holds in a given ordered field, then all the others also hold in that field.

Proof The equivalence of the Right-Endpoint Principle and the Left-Endpoint Principle is easy to check and is left as an exercise. The same is true for the equivalence of the Supremum and Infimum Principles and for the equivalence of the Least-Upper-Bound and Greatest-Lower-Bound Principles. Also, the equivalence of the Supremum Principle with the Right-Endpoint Principle follows from Lemma (II.4.5). Thus, it suffices to show the equivalence of the Supremum Principle with the Least-Upper-bound Principle and with the Bisection Principle.

II.4.23 Lemma

The Supremum Principle implies the Least-Upper-Bound Principle. Moreover, in this situation one has $\sup X = \operatorname{lub} X$ for every nonempty set X which is bounded above.

Proof Suppose that X is a nonempty set which is bounded above. Then the Supremum Principle implies that X has a supremum; call it B. Condition (i) of Definition (II.4.6) then implies that B is in the set U_X of upper bounds of X. Furthermore, suppose that $y < B$. Then by Condition (ii) of the same definition there exists $x{\in}X$ such that $y < \leq x$; in particular, if $y < B$ then y is *not* an upper bound of X. Thus B is the *least* element of U_X. That is, the existence of $\sup X$ implies the existence of $\operatorname{lub} X$. Moreover, the fact that B is the least upper bound of X translates to the equation $\sup X = \operatorname{lub} X$, as required.

II.4.24 Lemma

The Least-Upper-Bound Principle implies the Bisection Principle.

Remark We actually prove the contrapositive statement: if the Bisection Principle fails to hold, then so does the Least-Upper-Bound Principle.

Proof Suppose that the Bisection Principle is *not* valid. Let $[a_1, b_1]$, $[a_2, b_2], \ldots$ be a bisection sequence for which the conclusion of this principle, namely that there exists exactly one number which lies in each of the intervals $[a_n, b_n]$, fails to hold. This failure occurs either when there is no number which lies in each interval or when there is more than one such number.

Case 1 Suppose that there does not exist any number which lies in all of the intervals in the given bisection sequence. If this happens, then for every number u there exists an index m such that $u \notin [a_m, b_m]$. For this m one then has either $u < a_m$ or $u > b_m$. Based on this observation, let X be the set of all numbers x such that $x < a_k$ for at least one index k. Clearly every number less than a_1 is in X, so $X \neq \emptyset$. Likewise, $b_n > a_k$ for every k and n so clearly each b_n is an upper bound for X. Let c be any upper bound for X. Then certainly $c \geq a_n$ for every n; indeed, if $c < a_k$ for some k then $c < x < a_k$ for, say, $x = (c + a_k)/2$, and clearly $x{\in}X$, contrary to c being an upper bound of X. But by the observation above, with $u = c$, it follows that there must exist m

such that $c > b_m$, and therefore c is not the least upper bound of X. In other words, no upper bound of X is the least upper bound of X, so in this case the Least-Upper-Bound Principle fails.

Case 2 Suppose that the set of numbers which lie in each interval $[a_n, b_n]$ has more than one element. Let c and d be such numbers with $c \neq d$; without loss of generality we may assume that they are labeled so that $c < d$. Then one has $a_n \leq c < d \leq b_n$ for each index n. It follows from the definition of 'bisection sequence' that

$$0 < d - c \leq b_n - a_n = \frac{b_1 - a_1}{2^{n-1}} \text{ for each index } n.$$

Let Y be the set of numbers of the form $(b_n - a_n)/2^{n-1}$ with n in $\mathbb{N}$, so that $d - c$ is a positive lower bound for the set Y. Now let L be any positive lower bound for Y, so that $0 < L \leq (b_1 - a_1)/2^{n-1}$ for each n. Then $0 < 2\,L \leq (b_1 - a_1)/2^n$ for each n in $\mathbb{N}$. Since $1/2^n < 1/2^0 = 1$ for every n in $\mathbb{N}$, it follows that $2\,L$ must also be a lower bound for Y. Since $2\,L > L$, because $L > 0$, it follows that Y does not have a greatest lower bound. Thus in this case the Greatest-Lower-Bound Principle fails to hold.

In summary, if the Bisection Principle fails to hold, then either the Least-Upper-Bound Principle must fail (Case 1) or the Greatest-Lower-Bound Principle must fail (Case 2); in either case, *both* of these principles must fail since they are equivaent.

II.4.25 Lemma

The Bisection Principle implies the Supremum Principle.

Proof Let X be a nonempty set of numbers which is bounded above. To simplify the exposition a bit, let S denote the set of all numbers y such that $y < x$ for at least one element x in X, and let U_X denote the set of upper bounds of X; the hypotheses on X ensure that both these sets are nonempty and that every number is an element of exactly one of these sets. Let a_1 be an element of S and b_1 an element of U_X; clearly $a_1 < b_1$. Let m_1 be the midpoint of the interval $[a_1, b_1]$, and note, as above, that m_1 is in exactly one of the sets S or U_X. If $m_1 \in S$, then define $[a_2, b_2]$ to be the half interval $[m_1, b_1]$; while if $m_1 \in U_X$, define $[a_2, b_2]$ to be $[a_1, m_1]$. It is clear that $a_2 \in S$ and $b_2 \in U_X$. Continuing this way one obtains a bisection sequence $[a_1, b_1]$, $[a_2, b_2]$, $\ldots$ $[a_n, b_n] \ldots$ such that $a_n \in S$ and $b_n \in U_X$. The Bisection Principle implies that there is exactly one number c which lies in each of these intervals.

Claim 1 The number c satisfies Condition (i) of Definition (II.4.6). That is, c is an upper bound of the set X.

Proof of Claim 1 Suppose not. Then there exists a number x in X such that $c < x$. Since c is the only number which lies in each of the intervals $[a_n, b_n]$, it follows that there exists an index m such that $x \notin [a_m, b_m]$. Since $a_n \leq c$ for all n, it follows that that $a_n < x$ for all n, including $n = m$. Thus the only way x could not be in $[a_m, b_m]$ is if $x > b_m$. Since, by construction, $b_m \in U_X$, it follows that $z \leq b_m < x$ for every element z of X, including $z = x$, which is impossible. Thus, no such x exists, so c is an upper bound of X, as claimed.

Claim 2 The number c satisfies Condition (ii) of Definition (II.4.6), i.e., the Approximation Property for the Supremum. That is, if $y < c$ then there exists x in X such that $y < x$.

Proof of Claim 2 If $y < c$ then, by the fact that c is the only number which lies in each interval $[a_n, b_n]$, it follows that there must be an index m such that $y \notin [a_m, b_m]$. Since $y < c \leq b_n$ for all n, it follows that y cannot satisfy $y > b_m$, and therefore $y < a_m$. However, $a_m \in S$ by construction, so there exists x in X such that $a_m < x$ and thus $y < x$, so that Condition (ii) is satisfied.

It follows that X has a supremum, namely c, and the Supremum Principle is satisfied, as required.

The fact that the Supremum Principle implies the Least-Upper-Bound Principle, the Least-Upper-Bound Principle implies the Bisection Principle, and the Bisection Principle implies the Supremum Principle, guarantees that all three of these principles are equivalent. Combined with the other equivalences obtained earlier, it follows that all seven principles are equivalent, as claimed.

In light of the preceding theorem, the following makes sense.

The Completeness Axiom

At least one of the seven preceding 'Principles' holds in the ordered field $\mathbb{R}$; equivalently, *each* of the preceding 'Principles' holds in the ordered field $\mathbb{R}$. One abbreviates this by saying that $\mathbb{R}$ is a **complete ordered field**.

Remark Adding the Completeness Axiom allows one to distinguish between the ordered fields $\mathbb{R}$ and $\mathbb{Q}$. It may come as a surprise that no further axioms for $\mathbb{R}$ are needed. More precisely, *every* pair of complete ordered fields are essentially the same, in a sense to be made clearer in Appendix C.

Side Comment (on Dedekind's treatment of $\sqrt{2}$)

The proof to be given below that $\mathbb{Q}$ does not satisfy the Least-Upper-Bound Principle is based on a clever observation made by Richard Dedekind in his famous pamphlet of 1872 *Stetigkeit und irrationale zahlen* (Continuity and Irrational Numbers).

Dedekind's Observation Let x be a positive rational number; it is well known that that $x^2 \neq 2$. Now let $y = \dfrac{x\,(x^2+6)}{3\,x^2+2}$; note that clearly y is also a positive rational number. Then it is a simple exercise, which the reader is encouraged to carry out, to show that y^2 lies strictly between x^2 and 2. That is, if $x^2 > 2$ then $x^2 > y^2 > 2$, while if $x^2 < 2$ then $x^2 < y^2 < 2$.

The application of Dedekind's observation to the issue at hand is simple. Indeed, let S be the set of rational numbers x such that $x > 0$ and $x^2 < 2$; This set is nonempty; for example $1 \in S$. It is also bounded above in $\mathbb{Q}$ by, say, 3. Let B be any positive number in $\mathbb{Q}$. It is impossible that $B^2 = 2$, so either $B^2 > 2$ or $B^2 < 2$. Suppose first that $B^2 < 2$. Then by Dedekind's observation there exists a rational number y such that $0 < B < y$ and $y^2 < 2$. In particular, $y \in S$, so B is not an upper bound of the set S. Next, suppose that $B^2 > 2$. Then by Dedekind's observation there exists a rational number y such that $B < y$ and $2 < y^2 < B^2$. Since $y^2 > 2$, it follows that y is an upper bound of S. In this case, B is also an upper bound of S, but because $y < B$ it is clear that B is not the *least* such upper bound. In particular, no rational nnumber can be the least upper bound of S.

Remark It is worth pondering how someone might come up with the formula for y used by Dedekind.

II.4.26 Remarks

(1) In light of the fact, mentioned above, that the Bisection Principle does *not* hold in the ordered field $\mathbb{Q}$, it follows that $\mathbb{Q}$ is *not* a complete ordered field.

(2) Most modern textbooks choose the Least-Upper-Bound Principle as their choice of the 'official' Completeness Axiom for $\mathbb{R}$, and then treat all the other principles as theorems to be proved using this axiom (combined, of course, with the axioms for an ordered field).

The obvious advantage is simplicity: the Least-Upper-Bound Principle can be stated with virtually no preparation other than knowing the meaning of 'least element of a set' and 'upper bound of a set', both simple ideas.

One obvious disadvantage, however, is that most readers first encounter the phrase 'least upper bound' about three paragraphs before the statement of the corresponding axiom, so it is hard to argue that this axiom's claim is already 'clearly true'. In particular, the existence of a certain number, which this axiom claims, is not obvious to the beginners that have never considered the issue before.

A less obvious difficulty, but perhaps even more important for beginners, is that this principle involves sets of numbers which can be of arbitrary complexity. However, most beginning students of analysis have little experience with complicated sets of numbers. Any axiom which involves sets of such variety can hardly be considered to be 'obviously true'. (The same objections could be raised about choosing the Greatest-Lower-Bound Principle as the Completeness Axiom; but hardly any text uses that Principle as its version of the Completeness Axiom, so that issue is moot.)

(3) The Bolzano Principles, along with Supremum and Infimum Principles, also have the defect of not being 'obviously true'. In addition, they require substantially more preparation before one can even state them. Finally, all these principles are comparatively recent: Bolzano's is the oldest, and it dates from the early nineteenth century.

(4) The Bisection Principle is, of course, considerably more complicated to state than the other principles. However, the statement involves approximation techniques which have been in common use for millenia, and thus really are familiar. In addition, the sets which appear in this principle are simple and well known: intervals. Even a beginner should be willing to accept this statement about real numbers as being 'obviously true'; that is, 'axiomatic'. The statement of this Principle also reflects the fact that 'Completeness' should support the intuition that there are no 'holes' in the real numbers; a 'hole' would correspond to the case in which the intersection of the nested intervals in question is the empty set.

(5) There are other principles that could be used as the 'Completeness Axiom'. For example, the most elementary approach would be to base it on the fact that every real number has a representation in terms of the decimal notation, where here 'elementary' refers to the fact everybody learns the relevant facts about this representation in elementary-school arithmetic. Some of these alternate approaches are discussed later.

The next several results show some properties of $\mathbb{R}$ which do not follow directly from the axioms of an ordered field, and thus require 'Completeness'.

II.4.27 Theorem

The subset $\mathbb{N}$ of $\mathbb{R}$ is unbounded above in $\mathbb{R}$.

Proof: For each n in $\mathbb{N}$ let $a_n = 0$ and $b_n = 1/2^{n-1}$. Then the intervals $[a_1, b_1]$, $[a_2, b_2]$, $\ldots$ $[a_n, b_n]$, $\ldots$ obviously form a bisection sequence of intervals in $\mathbb{R}$. It is also obvious that the number 0 is an element of each of these intervals. Since, by the Bisection Principle, there is exactly one number with this property, it follows that for every number $u > 0$ there exists an index n such that $u \notin [0, 1/2^{n-1}]$, and thus $u > 1/2^{n-1}$. In particular, let $u = 1/B$ for some positive number B. Then one has $B < 2^{n-1}$. In particular, no number can be an upper bound for $\mathbb{N}$.

II.4.28 Corollary (The Principle of Archimedes)

Let ε and M be positive real numbers.

(a) There exists a natural number k such that $k\,\varepsilon > M$.

(b) Equivalently, there exists a natural number k such that $M/k < \varepsilon$.

(c) More generally, there exists a real number B such that if k is a natural number such that $k \geq B$, then $k\,\varepsilon > M$; equivalently, $M/k < \varepsilon$.

The simple proof is left as an exercise. Note that Part (c) is the formulation of the Principle of Archimedes which is most often used in practice.

II.4.29 Remarks

(1) The 'Archimedes' referred to in the name 'Archimedean Principle' is the most celebrated Greek geometer and scientist of ancient times, Archimedes of Syracuse (c. 250 BC). To understand what Archimedes means, think of ε as 'very small, but positive', and M as 'enormously large'. Then the property states that by repeatedly adding a small quantity to itself, one can obtain quantities which are arbitrarily large. (Compare this with Chapter Quote #2, made hundreds of years before Archimedes.) Another way of paraphrasing this property is that 'There are no infinitely small positive quantities.'

(2) One can easily prove directly, without invoking completeness, that the ordered field $\mathbb{Q}$ has the Archimedean Principle 'within itself'. Equivalently, one can prove that the sets $\mathbb{N}$ and $\mathbb{Q}$ are unbounded above (and below) in the ordered field $\mathbb{Q}$; compare with Theorem (I.8.4). In contast, one cannot prove that either of the sets $\mathbb{N}$ or $\mathbb{Q}$ is unbounded above in $\mathbb{R}$ using only the axioms from Sections (II.1) and (II.2). Indeed, there exist examples of ordered fields for which the sets $\mathbb{N}$ and $\mathbb{Q}$, viewed as subsets of the field, are bounded above in that field. Such fields are called *non-Archimedean fields* in light of the preceding corollary. We have no need to study such fields in *This Textbook*, but it is good to know that they exist.

The Archimedien Principle implies the following useful result.

II.4.30 Theorem (Density of $\mathbb{Q}$ in R)

Let c be any real number. Then for every $\varepsilon > 0$ there are infinitely many rational numbers r such that $|r - c| < \varepsilon$.

Proof For simplicity, assume that $c > 0$; the other cases can be reduced to this and are left to the reader. Let n_0 be a natural number such that $1/n_0 < \varepsilon$; such n_0 exists by the Principle of Archimedes. Then, also by the Principle of Archimedes, there exists an integer k, necessarily positive, such that $k/n_0 > c$. Let m_0 be the smallest such k; see (I.3.1). Then $(m_0 - 1)$ is an integer such that $(m_0 - 1)/n \leq c < m_0/n_0$. One computes that

$$0 < \frac{m_0}{n_0} - c \leq \frac{m_0}{n_0} - \frac{m_0 - 1}{n_0} = \frac{1}{n_0} < \varepsilon$$

Next, apply the same argument to the same c, but now with the original ε replaced by $\varepsilon_1 = m_0/n_0 - c$, to get natural numbers m_1 and n_1 such that $0 < c - m_1/n_1 < \varepsilon_1$. Clearly $m_1/n_0 < m_0/n_0$. Continuing this way, one gets a sequence of rational numbers m_j/n_j, $j = 0, 1, 2, \ldots$ such that $0 < c - m_j/n_j < \varepsilon$. The rational numbers $r_j = m_j/n_j$ have been constructed to be be distinct, so there are infinitely many of them.

Remarks (1) For obvious reasons, the preceding theorem is often abbreviated to the statement that 'the rational numbers form a dense subset of $\mathbb{R}$'. We use this formulation as the basis for a more general concept of 'denseness' later.

(2) Many texts define the concept of 'denseness of $\mathbb{Q}$ in $\mathbb{R}$' as follows:

If a and b are real numbers such that $a < b$, then there exists a rational number r such that $a < r < b$.

It is a simple exercise to prove that these formulations are equivalent.

The next results are closely related to the Bisection Principle.

II.4.31 Theorem The Nested-Intervals and Nested-Segments Theorems in $\mathbb{R}$

(a) (The Nested-Intervals Theorem) Let $(I_1, I_2, \dots I_k \dots)$ be a sequence of closed bounded intervals in $\mathbb{R}$.

Claim 1 If the sequence $(I_1, I_2, \dots I_k, \dots)$ is nested (see Definition (II.4.19)), then the intersection $J = \bigcap_{k=1}^{\infty} I_k$ of these intervals is a nonempty subset of $\mathbb{R}$.

More precisely, write each $I_k = [a_k, b_k]$, where a_k and b_k are real numbers such that $a_k < b_k$, and let $\alpha = (a_1, a_2, \dots a_k, \dots)$ and $\beta = (b_1, b_2, \dots b_k, \dots)$ be the corresponding sequences of endpoints of these intervals. As usual, let S_α and S_β denote the term-sets $\{a_1, a_2, \dots a_k, \dots\}$ and $\{b_1, b_2, \dots b_k, \dots\}$ of these sequences, and let $A = \sup S_\alpha$ and $B = \inf S_\beta$. Then A and B are both finite, and $J = \mathrm{Seg}\,[A, B]$.

Claim 2 If, in addition, for every $\varepsilon > 0$ there exists an index k such that $|b_k - a_k| < \varepsilon$, then $A = B$ and J is a singleton set $\{c\}$, where $c = A = B$.

(b) (The Nested-Segments Theorem) Replace 'interval' throughout Part (a) by 'segment' and '$[a_k, b_k]$' by '$\mathrm{Seg}\,[a_k, b_k]$', and drop the requirement $a_k < b_k$. The resulting claims are still true.

Proof

(a) Claim 1 It is shown in Example (II.4.20) (1) that the sequence α is monotonic up, while the sequence β is monotonic down. It is also shown that α is bounded above (by b_j for each index j, while β is bounded below (by a_i for each i). It follows from basic properties of 'supremum' and 'infimum' that A and B are both finite and that $A \leq B$.

Since A is an upper bound for the set S_α it follows that $a_k \leq A \leq B$ for each index k; similarly, $A \leq B \leq b_k$ for each k. More precisely, if x is any real number such that $A \leq x \leq B$, then $a_k \leq A \leq x \leq B \leq b_k$ for each index k. That is, every element of the set $\mathrm{Seg}\,[A, B]$ is an element of each of the intervals $[a_k, b_k]$, which implies that $\mathrm{Seg}\,[A, B] \subseteq J$.

Conversely, suppose that $y \notin \mathrm{Seg}\,[A, B]$. Then either $y < A$ or $y > B$. Suppose that the former situation holds. Then, by the Approximation Property for Suprema, there exists an element a_k in the set S_α such that $y < a_k$. It follows that for this k one has $y \notin [a_k, b_k]$, and thus $y \notin J$. By a similar argument, using the Approximation Property for Infima, one sees that if $y > B$ then $y \notin J$. It follows that $J \subseteq \mathrm{Seg}\,[A, B]$.

Combining these results implies that $\mathrm{Seg}\,[A, B] = J$, as claimed.

Remark One must write $J = \mathrm{Seg}\,[A, B]$, and not $J = [A, B]$, because it is possible that $A = B$ instead of $A < B$; the latter situation would be required to use the interval notation $[A, B]$.

Claim 2 From the string of inequalities $a_k \leq A \leq B \leq b_k$, valid for each k, it follows that

$$0 \leq |B - A| \leq |b_k - a_k| \text{ for each index } k.$$

Now let $\varepsilon > 0$ be given. From the hypothesis in this Claim, there exists k such that $|b_k - a_k| < \varepsilon$. It follows that $0 \leq |B - A| < \varepsilon$ for every $\varepsilon > 0$. It then follows from the Principle of Eudoxus that $A = B$, as claimed.

(b) The simple modifications of the preceding proof needed to prove the desired result are left to the reader.

Note The absolute values which appear in the statement and proof of Claim 2 in Part (a) are obviously not really needed. However including them makes the transition from Part (a) to Part (b) a bit simpler.

II.4.32 Remarks

(1) Some texts apply the name 'Nested-Intervals Principle' only to the result stated in Claim 1 of the preceding theorem. Also, some authors use the name 'Cantor Intersection Theorem' in place of 'Nested-Intersection Theorem'.

(2) In his famous treatment of analysis, *Foundations of Modern Analysis*, French mathematician Jean Dieudonné used Claim (1) of Part (a) of the preceding theorem, together with the Archimedean Principle, as his 'Completeness Axiom'. It is easy to show that this version of 'Completeness' is equivalent to all the other versions discussed in this chapter.

Reminder Definition (II.3.1) describes the concept of 'interval in $\mathbb{R}$' by writing down nine different cases. Assuming that $\mathbb{R}$ is complete allows one to sometimes avoid slogging through these nine cases.

II.4.33 Theorem

Let X be a subset of $\mathbb{R}$. Then the following statements are equivalent.

(i) The set X is an interval in $\mathbb{R}$, in the sense of Definition (II.3.1).

(ii) The set X has at least two points, and X is a convex subset of $\mathbb{R}$, in the sense of Definition (II.3.3).

Proof The fact that Statement (i) implies Statement (ii) has already been pointed out; see Example (II.3.8).

Conversely, suppose that Statement (ii) holds. Let $a = \inf X$ and $b = \sup X$. These quantities certainly exist, since the set X is nonempty. Moreover, the fact that X has at least two elements implies that $a \neq b$. Of course, it is possible that $a = -\infty$ or $a \in \mathbb{R}$, and it is possible that $b \in \mathbb{R}$ or $b = +\infty$. Now consider the various possibilities:

(a) If $a \in X$ and $b \in X$, then set $I = [a, b]$.

(b) If $a \in X$ and $b \notin X$, then set $I = [a, b)$.

(c) If $a \notin X$ and $b \in X$, then set $(a, b]$.

(d) If $a \notin X$ and $b \notin X$, then set $I = (a, b)$.

It is easy to show that $X = I$; in particular, X is an interval. For example, suppose that (d) holds, so that $I = (a, b)$. Then $I = \{u \in \mathbb{R} : a < u < b\}$. If $x \in X$ then $a \leq x \leq b$, since a is a lower bound for X and b is an upper bound for X. Moreover, $a \neq x$ and $b \neq u$, since Condition (d) holds. Thus, $a < x < b$. That is, $X \subseteq I$.

Conversely, suppose that $u \in I$, so that $a < u < b$. By the properties of 'inf' and 'sup', there must exist elements x_1 and x_2 of X such that

$$a \leq x_1 < u < x_2 \leq b.$$

In particular, $u \in \text{Seg}\,[x_1, x_2]$. Since X is convex, it follows that $\text{Seg}\,[x_1, x_2] \subseteq X$. In particular, since $u \in \text{Seg}\,[x_1, x_2]$, it follows that $u \in X$. That is, $I \subseteq X$. The fact that $X = I$ now follows.

The proof in Cases (a), (b) and (c) that one also has $X = I$ is similar, and is left to the reader.

II.5 Base-N Reprsentations of Real Numbers

Any axiom system for the real number system should be able to reproduce all the familiar features of that system. One of the most important of those features – important because it is in constant use by almost everyone – is the decimal representation of real numbers; that is, representation to base 10. Of course representations using other bases are also important; for example, in computer programming representations in base-2 ('binary') and base-16 ('hexadecimal') are widely used. The next several results clarify and justify the use of such representations, at least for numbers in the closed unit interval $[0, 1]$. (In *This Textbook* we use these representations only for such numbers; the extension of the theory to *all* real numbers is then easy to carry out.)

II.5.1 Definition

Let N be a fixed natural number such that $N \geq 2$.

(1) The **base N digits** are the nonegative integers k such that $0 \leq k \leq N - 1$.

Remark In *This Textbook* the only bases that are used are $N = 2, 3, 10$. In these cases the 'base-N' terminology is normally replaced by '**binary**', '**ternary**' and '**decimal**', respectively.

(2) Let $(d_1, d_2, \ldots d_k)$ be a k-tuple of base-N digits. The corresponding **base-N fraction** is the rational number $0 \overset{(N)}{\cdot} d_1\, d_2\, \ldots d_k$ given by the rule

$$0 \overset{(N)}{\cdot} d_1\, d_2 \ldots d_k = \frac{d_1}{N} + \frac{d_2}{N^2} + \ldots + \frac{d_k}{N^k}.$$

The the symbol $\overset{(N)}{\cdot}$ in this expression is called the **base-N point**; if $N = 10$, the '(N)' above the dot is usually omitted, and the symbol is simply called the **decimal point**. Also, the leading zero before this point is optional; it is usually included however for ease of reading.

Remark Since $0 \leq d_j \leq N - 1$ for each index j, it follows from Part (a) of Theorem (II.2.16), using $u = 1/N$ in that theorem, that $0 \overset{(N)}{\cdot} d_1\, d_2\, \ldots\, d_k \leq$

$$\frac{N-1}{N} + \frac{N-1}{N^2} + \ldots + \frac{N-1}{N^k} = \left(\frac{N-1}{N}\right)\left(1 + \frac{1}{N} + \ldots + \frac{1}{N^{k-1}}\right) =$$

$$\left(\frac{N-1}{N}\right)\left(\frac{1-(1/N)^k}{1-(1/N)}\right) = \left(1 - \frac{1}{N^k}\right) < 1.$$

That is, every base-N fraction lies in the interval $[0, 1)$. Also it is easy to show that a number x in $[0, 1)$ is a base-N fraction if, and only if, it can be expressed as a fraction k/N^m where k and m are nonnegative integers and $k < N^m$.

(3) Let $\sigma = (d_1, d_2, \dots d_n, \dots)$ be an infinite sequence of base-N digits, a so-called **base-N sequence**. Associate with σ the set $X_\sigma = \{0 \overset{(N)}{\cdot} d_1, 0 \overset{(N)}{\cdot} d_1\, d_2, \dots 0 \overset{(N)}{\cdot} d_1\, d_2, \dots d_k, \dots\}$, consisting of all the base-N fractions which arise from these digits in the given order. It is clear, from the preceding 'Remark', that this set is bounded above (by 1) and thus, by 'Completeness', has a finite supremum. The number $\sup X_\sigma$ is called **the real number with base-N representation determined by the sequence** σ; it is denoted by the infinite expression $0 \overset{(N)}{\cdot} d_1\, d_2 \dots d_n \dots$.

It follows from what was said above that for each base-N sequence $\sigma = (d_1, d_2, \dots)$ the corresponding number $0 \overset{(N)}{\cdot} d_1\, d_2 \dots$ lies in the interval $[0, 1]$. The next result says, among other things, that *every* number in $[0, 1]$ has a base-N representation of this type.

II.5.2 Theorem

(a) Let N be any natural number such that $N \geq 2$, and let x be any number in the closed interval $[0, 1]$ in $\mathbb{R}$. Then there exists at least one base-N sequence $\sigma = (d_1, d_2, \dots d_n, \dots)$ such that $x = 0 \overset{(N)}{\cdot} d_1\, d_2 \dots d_n \dots$.

(b) Suppose that x and y are real numbers in the interval $[0, 1]$ with base-N representations $x = 0 \overset{(N)}{\cdot} d_1\, d_2 \dots$ and $y = 0 \overset{(N)}{\cdot} c_1\, c_2 \dots$. Suppose futher that $d_n \leq c_n$ for each index n. Then $x \leq y$, with equality if, and only if, $d_n = c_n$ for each n. More precisely, if for some index k one has $d_k < c_k$, then $x \leq y - \left(\dfrac{c_k - d_k}{N^k}\right)$.

Extreme Cases The number $x = 0$ corresponds to the base-N sequence $\sigma_0 = (0, 0, \dots 0, \dots)$, and to no other base-$N$ sequence. Likewise, the number $x = 1$ corresponds to the base-N sequence $\sigma_{N-1} = (N - 1, N - 1, \dots N - 1, \dots)$, and to no other base-$N$ sequence.

Remark Recall that the only base-N representations under consideration here start with zero to the left of the decimal point. In particular, the familiar base-N representation $1 \overset{(N)}{\cdot} 0\, 0 \dots 0 \dots$ for the number 1 is not allowed here.

(c) Suppose that $0 < x < 1$. Then a necessary and sufficient condition for x to correspond to more than one base-N sequence is that x be a base-N fraction; see Part (a) of Definition (II.5.1). More precisely, suppose that x equals the base-N fraction $0 \overset{(N)}{\cdot} d_1\, d_2 \dots d_k$. Since, by hypothesis, $x > 0$, it follows that at least one of the digits in this expression must be positive. Since any zero digits beyond the final nonzero digit contribute nothing to the value of the number x, one can omit all such 'trailing zero digits' and assume that the final digit d_k is positive. Then x corresponds to the base-N sequences $\sigma = (d_1, d_2, \dots d_{k-1}, d_k, 0, 0, \dots 0, \dots)$ and $\tau = (d_1, d_2, \dots d_{k-1}, (d_k - 1), N - 1, N - 1, \dots N - 1, \dots)$, and to no other such sequences.

(d) If x is given by the base-N representation $x = 0 \overset{(N)}{\cdot} d_1\, d_2 \dots d_n \dots$ then for each k one has

$$|x - 0 \overset{(N)}{\cdot} d_1\, d_2 \dots d_k| \leq \frac{1}{N^k}.$$

Furthemore, one has equality for some k if, and only if, $d_m = 0$ for all $m > k$; that is, if, and only if, x is a base-N fraction.

(e) Suppose that x and y are numbers in the interval $[0, 1]$ which are given by the base-N representations $x = 0 \overset{(N)}{\cdot} d_1\, d_2 \dots d_n \dots$ and $y = 0 \overset{(N)}{\cdot} c_1\, c_2 \dots c_n \dots$. If for some natural number k one has $d_j = c_j$ for $1 \leq j \leq k$, then $|x - y| \leq 1/N^k$.

(f) Suppose that x and y are as in Part (e) above, with $x \neq y$. Let m be the smallest natural number such that $d_m \neq c_m$. Then $|x - y| \geq ||d_m - c_m| - 1|/1/N^m$. In particular, if the digits d_m and c_m differ by at least 2, then $|x - y| \geq 1/N^m$.

Proof Most of the proof simply reviews ideas from grade-school arithmetic.

(a) Recursively define a sequence $d_1, d_2, \ldots$ of base-N digits as follows:

<u>Initial Step</u> Note that there exists at least one base-N digit d such that $d/N \leq x$; for example, $d = 0$ has this property. Define d_1 to be the maximal such digit. It is clear that $\dfrac{d_1}{N} \leq x \leq \dfrac{d_1 + 1}{10}$. Indeed, if $d_1 < N-1$ then d_1+1 is also a base-N digit, so that the stronger inequality $\dfrac{d_1+1}{N} > x$ follows from the maximality condition on d_1. If $d_1 = N-1$, so that $1 + d_1 = N$ is not a base-N digit, then $\dfrac{1+d_1}{N} = \dfrac{N}{N} = 1$, so the required inequality follows from the hypothesis $x \leq 1$.

<u>Recursive Step</u> Suppose that digits d_1, d_2, $\ldots$ d_k have been defined so that

$$\frac{d_1}{N} + \ldots + \frac{d_k}{N^k} \leq x \leq \frac{d_1}{N} + \ldots + \frac{1+d_k}{N^k};$$

that is,

$$0 \overset{(N)}{\cdot} d_1\, d_2 \ldots d_k \leq x \leq 0 \overset{(N)}{\cdot} d_1\, d_2 \ldots d_k + \frac{1}{10^k} \qquad (*)$$

If the inequality on the left side of $(*)$ is actually an equation, then define $d_{k+1} = 0$; it is then easy to show that one will have $d_m = 0$ for all $m \geq k+1$. If, instead, the inequality on the left side of $(*)$ is strict, then let d_{k+1} be the greatest of the base-N digits d such that $\dfrac{d}{N^{k+1}} \leq x - \left(\dfrac{d_1}{N} + \ldots + \dfrac{1+d_k}{N^k}\right)$. Then, as before, the maximality condition on d_{k+1} implies

$$\frac{d_1}{N} + \ldots + \frac{d_k}{N^k} + \frac{d_{k+1}}{N^{k+1}} \leq x \leq \frac{d_1}{10} + \ldots + \frac{d_k}{10^k} + \frac{1+d_{k+1}}{10^{k+1}};$$

that is,

$$0 \overset{(N)}{\cdot} d_1\, d_2 \ldots d_k\, d_{k+1} \leq 0 \overset{(N)}{\cdot} d_1\, d_2 \ldots d_k\, d_{k+1} + \frac{1}{10^{k+1}}.$$

Let $\sigma = (d_1, d_2, \ldots d_n, \ldots)$ be the resulting decimal sequence.

<u>Claim</u> The given number x satisfies $x = 0.d_1\, d_2 \ldots d_n \ldots$.

<u>Proof of Claim</u> Let X_σ be the set of base-N fractions $\{0 \overset{(N)}{\cdot} d_1, 0 \overset{(N)}{\cdot} d_1\, d_2, \ldots\}$. It is clear from the left side of $(*)$ than x is an upper bound of the set X_σ. Furthermore, if $\varepsilon > 0$ is given, then, by the Archimedes Principle, there exists k such that $\dfrac{1}{N^k} < \varepsilon$. It follows from this, when combined with the right side of $(*)$, that

$$x - \varepsilon < x - \frac{1}{N^k} \leq 0 \overset{(N)}{\cdot} d_1\, d_2 \ldots d_k.$$

Since the right side of the last inequality is an element of X_σ, and $\varepsilon > 0$ is arbitrary, it follows from Definition (II.4.6) that $x = \sup X_\sigma$, as required.

(b) This follows easily from the definition of base-N representation together with Part (a) of Theorem (II.2.16).

(c) Suppose that x is a positive base-N fraction, so that x can be written in the form $0 \overset{(N)}{\cdot} d_1\, d_2 \ldots d_k\, 0\, 0 \ldots$, with $d_k > 0$. Then it is easy to show, using Part (a) of Theorem (II.2.16) that

$x = 0 \overset{(N)}{\cdot} d_1 \dots d_{k-1}\,(d_k - 1)\,(N-1)\,(N-1)\dots$. Reversing this argument, one sees that if x has a base-N representation which ends entirely in the base-N digit $N-1$, then it has a second representation which ends entirely in the base-N digit 0, and thus is a base-N fraction.

Now suppose that x cannot be expressed in either of the preceding forms; that is, it is not possible to express x with a base-N representation that ends entirely in the base-N digit 0 or ends entirely in the base-N digit $(N-1)$. Let $0 \overset{(N)}{\cdot} d_1\, d_2 \dots d_n \dots$ and $0 \overset{(N)}{\cdot} c_1\, c_2 \dots c_n \dots$ be base-N representations of x. Suppose, first, that $d_1 \neq c_1$; without loss of generality, assume that $d_1 < c_1$. Then clearly

$$0 \overset{(N)}{\cdot} 0\, d_2 \dots d_n \dots = \frac{c_1 - d_1}{N} + 0 \overset{(N)}{\cdot} 0\, c_2 \dots c_n \dots$$

Arguing as above, one sees that

$$0 \overset{(N)}{\cdot} 0\, d_2 \dots d_n \dots \leq 0 \overset{(N)}{\cdot} 0\,(N-1) \dots (N-1) \dots = \frac{1}{N}$$

with equality if, and only if, $d_k = N-1$ for each $k \geq 2$. By hypothesis one has $c_1 > d_1$. Also $0 \overset{(N)}{\cdot} 0\, c_2 \dots c_n \dots \geq 0$, with equality if, and only if, $c_k = 0$ for every $k \geq 2$. It follows that $c_1 - d_1 = 1$, $0 \overset{(N)}{\cdot} 0\, d_2 \dots d_n \dots = 1$, and $0 \overset{(N)}{\cdot} 0\, c_2 \dots c_n \dots = 0$. Thus, $c_1 > d_1$ conflicts with the hypothesis that x does not admit either of the forms above.

(d), (e) and (f) The proofs of these statements are left as exercises.

Part (b) of the preceding theorem says, in effect, that if two numbers in $[0,1]$ have base-N representations which are the same for the first k digits, then the numbers differ by at most $1/N^k$. Unfortunately, the converse is very far from true: numbers which are close to each other may fail to agree in any of their digits; see, for example, the decimal expressions for the number 1/10, $0.1\,0\,0 \dots 0 \dots$ and $0.0\,9\,9 \dots 9 \dots$. For the applications of base-N representations used in *This Textbook*, the following partial result is sufficient.

II.5.3 Theorem

Let x and y be numbers in $[0,1]$ with base-N representations $\overset{(N)}{\cdot} d_1\, d_2 \dots$ and $0 \overset{(N)}{\cdot} c_1\, c_2 \dots$, respectively. Suppose that there is an index k such that $d_j = c_j$ for each j such that $1 \leq j \leq k$, but $d_{k+1} \neq c_{k+1}$. Let $m = |d_{k+1} - c_{k+1}|$, so that $m \geq 1$. Then one has

$$|x - y| \geq \frac{m-1}{N^{k+1}}$$

In particular, if the first unequal digits d_{k+1} and c_{k+1} differ by at least 2 units, then $|x-y| \geq 1/N^{k+1}$.

The simple proof is left as an exercise.

CONSTRUCTION OF THE REAL NUMBERS FROM THE RATIONALS USING DECIMALS

Definition A **positive decimal string** is a function $\alpha : \mathbb{Z} \to \mathbb{Z}_9$ with the following properties:

(i) There exists $N \in \mathbb{Z}$ such that $\alpha(n) = 0$ for all $n < N$, and

(ii) $\alpha(k) \neq 0$ for at least one k in $\mathbb{Z}$.

The set of all such strings is denoted D^+.

Notation

Example

II.6 EXERCISES FOR CHAPTER II

PART I – THESE EXERCISES REQUIRE ONLY THE FIELD AND/OR ORDER AXIOMS FOR $\mathbb{R}$; DO NOT USE ANY RESULTS BASED ON 'COMPLETENESS' IN THE SOLUTIONS.

II - 1 (a) Prove Parts (d) and (e) of Theorem (II.1.3).

(b) Prove Parts (f) and (g) of Theorem (II.1.3) directly from the field axioms, Axioms A0–A6.

II - 2 Prove the **Cancellation Law**: Let a, b and c be real numbers such that $a \neq 0$. Then there is exactly one real number x such that $a\,x + b = c$.

II - 3 NEED NEW EXERCISE

II - 4 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by the formula $f(x, y) = |x - y|$ for all (x, y) in $\mathbb{R}^2$.

(a) Prove that the binary operation f commutative but not associative.

(b) Determine whether the operation f satisfies the 'Extended Commutative Law'; see Theorem (II.1.5).

II - 5 Let $D' : (\mathbb{R}\backslash\{0\})\times(\mathbb{R}\backslash\{0\}) \to \mathbb{R}$ be the restriction to $\mathbb{R}\backslash\{0\}$ of the 'division' function D described in Definition (II.1.2).

(a) Explain why D' is a binary operation but D is not.

(b) Determine whether the binary operation D' satisfies the Associative Law.

II - 6 NEED NEW EXERCISE

II - 7 NEED NEW EXERCISE

II - 8 Use the Extended Commutative and Associative Laws for Addition (ThemB10.32) to prove

$$x_1+((x_2+x_3)+x_4)+x_5+((x_6+(x_7+x_8))+x_9)+x_{10} = x_{10}+x_7+x_9+x_6+x_1+x_5+x_4+x_3+x_2+x_8$$

for all real numbers x_j, $j = 1, 2, \ldots 10$.

II - 9 NEED NEW EXERCISE

II - 10 NEED NEW EXERCISE

II - 11 NEED NEW EXERCISE

II - 12 (a) Prove Theorem (**??**)

(b) Find a number $B > 0$ such that if $M > B$ then $\dfrac{1}{M} < \dfrac{1}{2} < M$.

II - 13 Prove Corollary (II.2.7)

II - 14 Prove that if X is a finite nonempty set of real numbers then the set X has both a maximum element and a minimum element. That is, show that there are real numbers m and M such that

(i) m and M are elements of X.

(ii) If x is any number in X such that $x \neq m$ and $x \neq M$, then $m < x < M$.
Also, show that the numbers m and M with these properties are unique.

II - 15 (a) Prove Parts (e) and (f) of Theorem (II.3.4).

(b) Prove Part (g) of Theorem (II.3.4).

(c) Prove Part (h) of Theorem (II.3.4).

II - 16 Suppose that f, g and h are real-valued functions defined on a nonempty set X. Assume that there are positive numbers A, B and C such that $|f(x)| \leq A$, $|g(x)| \leq B$ and $|h(x)| \geq C$ for all x in X. Show that $\left|\frac{f(x)-3g(x)}{h(x)}\right| \leq \frac{A+3B}{C}$ for all x in X.

II - 17 (a) Prove Parts (a) and (b) of Theorem (II.2.14).

(b) Prove Part (c) of Theorem (II.2.14).

(c) Prove Part (d) of Theorem (II.2.14).

II - 18 NEED NEW EXERCISE

II - 19 Prove Parts (a), (b), (c) and (d) of Theorem (**??**)

II - 20 Prove Parts (e), (f), (g) and (h) of Theorem (**??**)

II - 21 Prove Corollary (II.2.7)

II - 22 Prove **Bernoulli's Inequality**: If c is a real number such that $c > -1$, then $(1+c)^k \geq 1+kc$ for each natural number k.

II - 23 Let $p : \mathbb{R} \to \mathbb{R}$ be a quadratic polynomial function; that is, there are numbers a, b and c in $\mathbb{R}$ such that $p(t) = at^2 + bt + c$ for all t in $\mathbb{R}$. To simplify the discussion, assume that $a > 0$.

(a) Prove that that the function p assumes a minimum value on $\mathbb{R}$. More precisely, prove that there exists a unique number t_0 such that $p(t_0) \leq p(t)$ for all t in $\mathbb{R}$. (Hint: Does the phrase 'Complete the Square' ring a bell?)

(b) Give simple criteria, in terms of the coefficients a, b and c of the polynomial p, for the minimum value of the polynomial p to be positive, negative or zero.

II - 24 This exercise has as its goal to obtain an important criterion for a pair of vectors $\overrightarrow{x} = (x_1, \ldots x_n)$ and $\overrightarrow{y} = (y_1, \ldots y_n)$ in $\mathbb{R}^n$ to be linearly dependent. Since this property is trivially true if either of the vectors is the zero vector, throughout most of this exercise we assume that $\overrightarrow{x}$ and $\overrightarrow{y}$ are nonzero vectors.

Recall from elementary linear algebra that a necessary and sufficient condition for a pair of nonzero vectors in $\mathbb{R}^n$ to be linearly dependent is that each of them is a multiple of the other. Using the notation established above, this means that the vectors $\overrightarrow{x}$ and $\overrightarrow{y}$ are linearly dependent if, and only if, there is a number t such that $x_i = ty_i$ for each $i = 1, 2, \ldots n$.

(a) Let $\overrightarrow{x}$ and $\overrightarrow{y}$ be as above, and let $p : \mathbb{R} \to \mathbb{R}$ be the function given by the rule

$$p(t) = (x_1 - ty_1)^2 + (x_2 - ty_2)^2 + \ldots + (x_n - ty_n)^2 \text{ for all } t \text{ in } \mathbb{R},$$

Show that the vectors $\overrightarrow{x}$ and $\overrightarrow{y}$ are linearly dependent if, and only if, the equation $p(t) = 0$ has a real solution.

(b) Apply the results of Exercise **II - 20** above to prove that for all vectors $\overrightarrow{x}$ and $\overrightarrow{y}$ in $\mathbb{R}^n$ as above one has

$$(x_1y_1 + x_2y_2 + \ldots + x_ny_n)^2 \leq \left(x_1^2 + x_2^2 + \ldots + x_n^2\right) \cdot \left(y_1^2 + y_2^2 + \ldots + y_n^2\right) \qquad (*)$$

Moreover, one gets equality in $(*)$ if, and only if, the vectors $\overrightarrow{x}$ and $\overrightarrow{y}$ are linearly dependent.

(c) In the preceding it was always assumed that neither $\overrightarrow{x}$ nor $\overrightarrow{y}$ equals the zero vector. Determine the status of the inequality $(*)$, and its relation to linear independence, when one of the vectors equals $\overrightarrow{0}$.

Remark Inequality $(*)$ above is usually called the Cauchy-Schwarz Inequality; or, more precisely, the Cauchy-Schwarz Inequality for n-tuples. It is due to Cauchy. There is also an analogous result, but for definite integrals, which is due to Schwarz. It is also known as the Cauchy-Schwarz Inequality, but for integrals. Cauchy was, one presumes, unaware of this integral form, since he was already dead when it was published. It seems likely that Schwarz's name is added to the original Cauchy result simply to distinguish it from the myriad of other results in mathematics also called 'Cauchy's Theorem'. Finally, it turns out that Bunyakowski proved the same result as Schwarz, so some authors add his name to the list. The order in which the names are listed in texts and research papers seems dependent on the nationality of the writer.

II - 25 Prove Parts (e), (f), (g) and (h) of Theorem (II.3.4)

II - 26 Suppose that $x_1,\, x_2, \ldots\, x_k$ are real numbers which satisfy a string of inequalities of the following form:

$$x_1\, L_1\, x_2\, L_2\, x_3\, L_3\, \ldots x_{k-1}\, L_{k-1}\, x_k,$$

where each of the symbols $L_1,\, L_2, \ldots\, L_{k-1}$ stands for one of the relations '$<$' or '$\leq$'. Let i and j be indices such that $1 \leq i < j \leq k$.

(a) Show that if at least one of the expressions $L_i,\, L_{i+1}, \ldots\, L_{j-1}$ equals '$<$', then $x_i < x_j$.

(b) Show that a necessary and sufficient condition for the equation $x_i = x_j$ to hold is that $x_i = x_{i+1} = \ldots = x_{j-1} = x_j$. Make it clear which part of your solution corresponds to the 'necessary' portion of the statement, and which corresponds to the 'sufficient' portion.

II - 27 Let X be the set of all real numbers x such that $x^2 \leq 2$. Show that X is a convex set.

II - 28 Let X be a nonempty set of real numbers, and let $\mathcal{F}_X$ be a nonempty family of convex subsets Y of $\mathbb{R}$ such that $X \subseteq Y$. Then the intersection of the family $\mathcal{F}_X$ is a convex subset of $\mathbb{R}$ which has X as a subset.

II - 29 NOTE: The purpose of this exercise is to show that there are several equivalent ways to formulate a suitable notion of 'completeness' for $\mathbb{R}$. The choice of which statement to take as the 'official' Completeness Axiom is largely a matter of taste and pedagogy.

Consider the following statements:

(1) If $[a_1, b_1],\, [a_2, b_2], \ldots$ is a bisection sequence in $\mathbb{R}$, then the intersection $\bigcap_{k=1}^{\infty}[a_k, b_k]$ is a singleton set.

(2) If X is a nonempty subset of $\mathbb{R}$ which is bounded above, and if Y is the set of all upper bounds of X, then Y has a minimum element.

(3) If X is a nonempty subset of $\mathbb{R}$ which is bounded below, and if Y is the set of all lower bounds of X, then Y has a maximum element.

(4) Every convex subset of $\mathbb{R}$ with at least two elements is an interval.

(5) For every closed interval $[a, b]$ in $\mathbb{R}$ the following is true: if $\mathcal{F}$ is a family of open intervals in $\mathbb{R}$ whose union contains $[a, b]$ as a subset, then there is a finite subfamily $\mathcal{F}'$ of $\mathcal{F}$ whose union also contains $[a, b]$ as a subset.

The simplest way to label the various parts of this exercise is with *pairs* of integers i and j, with $1 \leq i, j \leq 5$ and $i \neq j$:
Problem (i, j) Show that Statement i implies Statement j.

II - 30 Let $(x_1, x_2, \ldots x_k)$ be a k-tuple of real numbers in an interval $[a, b]$ in $\mathbb{R}$, with $k \geq 2$. Prove that there exist indices i and j, with $1 \leq i, j \leq k$ and $i \neq j$, such that $|x_i - x_j| \leq \dfrac{b-a}{k-1}$.

PART II – EXERCISES WHICH MAY INVOLVE ALL THE AXIOMS, INCLUDING 'COMPLETENESS'

II - 31 Prove or Disprove A necessary and sufficient condition for two real numbers A and B to be equal is that $|A - B| < 1/k$ for all k in $\mathbb{N}$.

II - 32 Prove Theorem (**??**).

II - 33 Prove Parts (a), (b) and (c) of Theorem (**??**).

II - 34 Prove Parts (d), (e) and (f) of Theorem (**??**).

II - 35 Let A be a nonempty set of real numbers.

(a) If c is a real number such that $\inf A < c$, then there exists a real number x in A such that $\inf A \leq x < c$; equality is possible if, and only if, A has a minimum element; namely, $x = \inf A$.

(b) If d is a real number such that $d < \sup A$, then there exists a real number y in A such that $d < y \leq \sup A$; equality is possible if, and only if, A has a maximum element; namely $y = \sup A$.

II - 36 Suppose that A and B are nonempty subsets of $\mathbb{R}$, and let $C \subseteq \mathbb{R}$ be defined by the rule

$$C = \{c \text{ in } \mathbb{R} : c = a + b \text{ for some choice of } a \text{ in } A \text{ and } b \text{ in } B\}.$$

Problem (a) Show that if either of the sets A or B is unbounded above, then so is C, so $\sup C = +\infty$. Likewise, if either of the sets A or B is unbounded below, then so is C, and $\inf C = -\infty$.

(b) Show that if A and B are both bounded above, then so is C, and $\sup C = \sup A + \sup B$. Likewise, if both A and B are bounded below, then so is C, and $\inf C = \inf A + \inf B$.

II - 37 For each of the sets given below, determine the supremum and infimum of the set. Also, determine whether the set has a maximum or minimum element.

Note Resist any temptation to use 'limits'; they are not introduced until Chapter (III).

(a) $X = \{x : x = 1 - (-1)^k \frac{1}{k} \text{ for all } k \text{ in } \mathbb{N}\}$

(b) $Y = \{x : x = (-k)^k \text{ for } k \text{ in } \mathbf{Z}\}$

(c) Z is the set of all the digits which appear in the decimal expansion of the number $1/7$. (You may use the usual rules for decimals here, even though they won't be proved until later.)

II - 38 Definition Let A be a nonempty subset of $\mathbb{R}$. Define subsets $\mathcal{L}_A$, $\mathcal{R}_A$ and $\mathcal{I}_A$ of $\mathbb{R}$ by the rules

$$\mathcal{L}_A = \{x : x \leq a \text{ for some element } a \text{ of } A\}; \quad \mathcal{R}_A = \{x : x \geq a \text{ for some element } a \text{ of } A\}; \quad \mathcal{I}_A = \mathcal{L}_A \cap \mathcal{R}_A$$

(a) Show that if A is a nonempty subset of $\mathbb{R}$, then $\mathcal{I}_A$ is the intersection of the family of all convex subsets of $\mathbb{R}$ which contain A as a subset.

(b) Prove or Disprove If X and Y are nonempty subsets of $\mathbb{R}$, then $\mathcal{L}_{X \cup Y} = \mathcal{L}_X \cup \mathcal{L}_Y$.

(c) Prove or Disprove If X and Y are nonempty subsets of $\mathbb{R}$, then $\mathcal{R}_{X \cup Y} = \mathcal{R}_X \cup \mathcal{R}_Y$.

Chapter III

Limits of Real Sequences – Basic Theory

Quotes for Chapter (III):

> (1) 'A philosopher of imposing stature doesn't think in a vacuum. Even his most abstract ideas are, to some extent, conditioned by what is or is not known in the time when he lives.'
>
> (From 'Dialogues of Alfred North Whitehead')
>
> (2) 'In contrast, consider the following ancient algorithm, attributed to Heron of Alexandria, for approximating square roots:
>
> To approximate the square root of a positive number X,
>
> – Make a guess for the square root of X
>
> – Compute an improved guess as the average of the guess and X divided by the guess.'
>
> (From the published majority opinion of the U.S. Ninth Circuit Court of Appeals in the case of Bernstein vs U. S. Department of Justice *et al*, 1999.)

Reminder of 'Limit' as Taught in Elementary Calculus

In elementary calculus one learns that the concept of 'limit' plays a vital role, especially in formulating the main concepts of the subject. Specifically, two types of limit processes appear in calculus:

(1) Limits of the form $\lim\limits_{x \to c} g(x)$; that is, limits of functions defined on an interval I. The quantity c can be either a real number or one of the symbols $+\infty$ or $-\infty$. The quantity x is a real variable, and is allowed to vary over all numbers in I except at c itself (if c is itself in I).

The most important example of such a limit in elementary calculus arises in the definition the definition of the 'derivative' of a function f at a number c:

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c}.$$

In this case, $g(x)$ is the ratio $(f(x) - f(c))/(x - c)$, which ratio, of course, is not defined at c.

One also encounters such limits in L'Hôpital's Rule. More precisely, under suitable circumstances one has

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

(2) Limits of the form $\lim_{k\to\infty} x_k$; that is, limits of sequences of real numbers. The numbers x_1, x_2, $\ldots$ form an infinite sequence, and the variable k – usually called the *index* – is allowed to vary over all natural numbers.

The most important example of such a limit in elementary calculus occurs in the definition of the definite integral:

$$\int_a^b f(x)\,dx = \lim_{k\to\infty} R_k,$$

where $R_k = f(c_1)\Delta x_1 + f(c_2)\Delta x_2 + \ldots + f(c_k)\Delta x_k$ is a typical Riemann sum for a partition of the interval $[a,b]$ into k parts.

In *This Textbook* the limits of the second type, i.e., limits of sequences, play the dominant role, so we study those limits first.

Remarks (1) If you've forgotten the definitions of 'derivative' and 'definite integral', or if you have only misty memories of these concepts, fear not: they are not needed in the rest of this chapter, and we do study them in great detail later in *This Textbook*.

(2) In typical courses of elementary calculus many of the results of this section do not appear at all. In particular, this is the case for results directly involving the completeness of $\mathbb{R}$ – 'supremum', 'infimum', and so on. Moreover, those that do appear in elementary calculus are often phrased more informally there and without full proofs.

III.1 The Limit of a Sequence of Real Numbers

The intuitive idea behind the statement that 'The number L is the limit of the sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$' is this: if k is large, then x_k is close to L. More precisely, by merely choosing k sufficiently large, one can be sure that x_k is as close to L as one wants, or even closer. Of course, qualitative phrases such as 'as close as one wants' and 'sufficiently large' are difficult to incorporate into rigorous proofs; one wants a more quantitative formulation. The next definition provides such a formulation.

III.1.1 Definition

Let $\xi = (x_1, x_2, \ldots x_k, \ldots)$ be a sequence of real numbers; that is, ξ is a function with domain $\mathbb{N}$ and values in $\mathbb{R}$ (see Definition (I.9.1)).

(1) One says that the sequence ξ **converges to a real number L** provided the following statement is true:

For every real number $\varepsilon > 0$ there exists a positive real number B such that if k is any index such that $k \geq B$, then $|x_k - L| < \varepsilon$. In this case one writes $\lim_{k\to\infty} x_k = L$.

(2) The sequence ξ is said to be **convergent** if there exists a real number L such that $\lim_{k\to\infty} x_k = L$, in the sense of Part (1).

(3) If the sequence ξ fails to be convergent in the sense of Part (2), then one says that ξ **is a divergent sequence**, or, more briefly, that ξ **diverges**.

Side Comment (on the limit definition and intuition) The relation between this formal definition and the intuitive view of 'limit' stated above is not hard to see. Indeed, the intuitive version can be reformulated as follows: You tell me how close you want x_k to be to L, and I'll tell you how large is large enough to guarantee, from the size of k alone, that x_k is that close to L, or even closer.

In the formal statement of the definition, the measure of 'how close x_k is to L' is, as usual, the nonnegative number $|x_k - L|$. The degree of closeness that you want is the positive number ε that you give me, so what you want is that $|x_k - L| < \varepsilon$. The 'large enough' refers to the number B, which depends on ε (and, of course, on the choice of the sequence ξ under discussion, but that is usually understood in the given context): if k is at least as large as B, then k is large enough to guarantee that $|x_k - L| < \varepsilon$. Note that it is irrelevant that there might also be values of k *smaller* than B for which one has $|x_k - L| < \varepsilon$.

Remark Some authors use minor variations of Definition (III.1.1) (1) for the meaning of the statement that a given sequence ξ converges to a given number L. The next result gives some examples of such variations which are equivalent to Definition (III.1.1) (1). It also provides an example of a similar variation which is *not* equivalent to that definition.

III.1.2 Theorem

(1) (a) Supppose that in the statement of Definition (III.1.1) (1) one makes one or more of the following changes:

(i) replace the hypothesis 'positive real number B' by 'positive integer B'

(ii) replace the hypothesis '$k \geq B$' by '$k > B$'

(iii) replace the hypothesis '$|x_k - L| < \varepsilon$' by '$|x_k - L| \leq \varepsilon$'.

Then the resulting modified definition is equivalent to Definition (III.1.1) (1), in the sense that if the sequence ξ converges to the number L according to one definition, then it does so according to the other.

(b) In contrast, if the hypothesis '$\varepsilon > 0$' in Definition (III.1.1) (1) is replaced by the hypothesis '$\varepsilon \geq 0$', then the resulting definition is *not* equivalent to Definition (III.1.1) (1).

The proofs of these assertions are left as exercises. In what follows we occasionally use the definitions modified as in Part (a) of this theorem, often without further comment.

The next result provides a more significant reformulation of Definition (III.1.1) (1) which is often useful.

III.1.3 Theorem

The following statements are equivalent:

(i) The sequence $\xi = (x_1, x_2, \dots)$ converges to the real number L, in the sense of Definition (III.1.1) (1).

(ii) For every $\varepsilon > 0$ there are only finitely many indices j such that $|x_j - L| \geq \varepsilon$. That is, for each $\varepsilon > 0$, the sequence $(|x_1 - L|, |x_2 - L|, \dots |x_k - L|, \dots)$ is eventually less than ε.

Proof Suppose that Statement (i) is true, and let $\varepsilon > 0$ be given. Let N in $\mathbb{N}$ be large enough that if the index k satisfies $k \geq N$, then $|L - x_k| < \varepsilon$ (see Theorem (III.1.2) (a) (i)). It follows that if j is an index such that $|x_j - L| \geq \varepsilon$, then $1 \leq j < N$. In particular, the set of all such indices j

is bounded above in $\mathbb{N}$ (by N), and thus is a finite set (by Theorem (I.8.4)). The reformulation in terms of the concept of 'eventually' now follows from Definition (I.9.14).

Conversely, suppose that Statement (ii) is true, and let $\varepsilon > 0$ be given. Let A be the set of all indices j such that $|x_j - L| \geq \varepsilon$. Then, by the hypothesis that Statement (ii) holds, the set A is a finite subset of $\mathbb{N}$, and thus (by Theorem (I.8.4) again) is bounded above by some natural number N. It follows that if $k \geq N + 1$ then k is *not* in A and thus $|x_k - L| < \varepsilon$. Statement (1) now follows.

The preceding result provides a useful method for proving that a given sequence is convergent. Also important, however, is the ability to prove that a given sequence is divergent. The next result is the analog to Theorem (III.1.3) for that type of question.

III.1.4 Theorem

The following statements are equivalent:

(i) The sequence $\xi = (x_1, x_2, \dots)$ is divergent, in the sense of Definition (III.1.1) (3).

(ii) For every real number L there exists $\varepsilon > 0$ such that the inequality $|x_j - L| \geq \varepsilon$ holds for infinitely many indices j.

Proof Suppose that Statement (i) is true, so that ξ is divergent. This means, by Definition (III.1.1) (3), that for each real number L the sequence ξ does not converge to L. That is, for each real L Statement (i) of Theorem (III.1.3) is false. Since that statement is equivalent to Statement (ii) of Theorem (III.1.3), this implies that for each real L the latter statement is also false. It follows that there exists $\varepsilon > 0$ such that $|x_j - L| \geq \varepsilon$ for infinitely many indices j; that is, Statement (ii) of the current theorem is also true. The proof that our Statement (ii) implies our Statement (i) follows similarly.

Remark The argument given here is typical of what one finds in research papers and in texts at the level of *This Textbook*. In particular, it assumes that the reader is already used to dealing with the negations of mathematical statements. Some readers may not yet be comfortable dealing with such negations. The following Side Comment is intended for such readers.

> Side Comment (on negating mathematical statements) A common occurance in constructing logical arguments in mathematics (as well as in other areas of life) is the need to consider the negation of a given statement; that is, to consider the new statement which asserts that the given statement is not true. Such negations occur in so-called 'Proofs by Contradiction': a given statement must be true because assuming that its negation is true leads to a contradiction to a known fact; see, for example, the proof of the Strong Principle of Mathematical induction (see Theorem (I.3.2)). (Students often are queasy about using such arguments: they prefer to prove directly that a given statement is true, not indirectly by showing that it can't be false. This allows them to avoid considering the negation of the given statement. Once they get the hang of arguments 'by contradiction', however, they sometimes go to the other extreme and use them in all cases, even when a direct proof might be easier and clearer.) Negations also arise as in Theorem (III.1.4) above, where one tries to prove directly the truth of the negation of a given statement.
>
> In any event, the process of negating a given mathematical statement, and then using that negation in a useful manner, often confuses beginning math students. What follows illustrates, in excruciating detail, this process in the special case of proving that a given sequence is divergent.
>
> The statement to be proved is that a certain sequence is divergent. Thus, the first step would be to recall, or, if needed, to look up, the meaning of the phrase 'divergent sequence'

as it appears, say, in Definition (III.1.1) (3). (Note: Experienced math teachers will tell you that surprisingly many students do *not* automatically think of 'Look up its definition' as the obvious response to the statement 'I don't know what this word means'.)

The definition of 'divergent sequence' is short and simple – it means 'it is not a convergent sequence'. That is, the statement to be proved presents itself as the negation of the statement that the given sequence *is* convergent. Beginners often encounter difficulties with formulating the negation of this last statement, mainly because the meaning of 'convergent sequence' is complicated:

'A sequence ξ of real numbers is convergent if there exists a real number L such that, for every $\varepsilon > 0$, there exists N in $\mathbb{N}$ such that if k in $\mathbb{N}$ such that $k \geq N$, then the inequality $|x_k - L| < \varepsilon$ holds.'

The definition of $\xi = (x_1, x_2, \ldots x_k, \ldots)$ being convergent then boils down to the following condition:

<u>Condition 0</u> 'There exists a real number L such that for every $\varepsilon > 0$ there exists N in $\mathbb{N}$ such that for every k in $\mathbb{N}$ such that $k \geq N$ then the inequality $|x_k - L| < \varepsilon$ holds.'

Condition 0 is of the form 'There exists a real number L such that a certain complicated condition holds', where that condition is:

<u>Condition 1</u> 'For every $\varepsilon > 0$ there exists N in $\mathbb{N}$ such that if k in $\mathbb{N}$ satisfies the inequality $k \geq N$, then the inequality $|x_k - L| < \varepsilon$ holds.'

The assertion that ξ is *not* convergent is then the *negation* of Condition 0, which then takes the form

'There does not exist a real number L such that Condition 1 holds'.

Otherwise stated, 'For every real number L, Condition 1 is not true'. This in turn requires understanding what it means for Condition 1 to not be true; that is, what is the meaning of the negation of Condition 1. However, Condition 1 is of the form 'For every $\varepsilon > 0$ a certain condition holds', where this 'certain condition' is

<u>Condition 2</u> 'There exists N in $\mathbb{N}$ such that if k in $\mathbb{N}$ satisfies $k \geq N$, then $|x_k - L| < \varepsilon$.'

The negation of Condition 1 then takes the form 'It is not the case that for every $\varepsilon > 0$ Condition 2 holds'. That is, there exists $\varepsilon > 0$ for which Condition 2 fails to hold. Now one needs to understand the negation of Condition 2. This condition is of the form 'There exists N in $\mathbb{N}$ such that a certain condition holds', where this next condition is

<u>Condition 3</u> 'If k in $\mathbb{N}$ satisfies $k \geq N$, then $|x_k - L| < \varepsilon$.'

The negation of Condition 2 then takes the form 'It is not the case that there exists N in $\mathbb{N}$ such that Condition 3 holds'. That is, for every N in $\mathbb{N}$, Condition 3 must fail to hold. This leads one to consider the negation of Condition 3. But Condition 3 is of the form 'If k in $\mathbb{N}$ satisfies $k \geq N$, then a certain condition holds'. This last condition is

<u>Condition 4</u> '$|x_k - L| < \varepsilon$.'

Thus, to say that Condition 3 fails to hold takes the form 'It is not the case that if k in $\mathbb{N}$ satisfies $k \geq N$, then Condition 4 holds'. In other words, there exists k in $\mathbb{N}$ such that $k \geq N$ but Condition 4 fails to hold. This leads one, finally, to consider the negation of Condition 4, which is 'It is not the case that $|x_k - L| < \varepsilon$'.

Up to now, the analysis has been essentially all linguistic. Indeed, the analysis has consisted in two observations:

(i) If a condition can be written in the form 'For every object of a certain type, a certain simpler condition must be true', then the negation of the original condition can be written in the form 'There exists an object of that type for which the simpler condition fails to be true'.

(ii) If a condition can be written in the form 'There exists an object of a certain type for which a certain simpler condition must be true', then the negation of the original condition can be written in the form 'For every object of that type the simpler condition fails to be true'.

One does not really need to understand the mathematics to carry out this linguistic analysis.

The exception is Condition 4: It is not of Type (i) or Type (ii). In this example, Condition 4 is where mathmatics actually begins to play a role. Indeed, the statement 'It is not the case that $|x_k - L| < \varepsilon$' does not admit further linguistic analysis along the lines followed above. However, this statement is known, from the order properties of $\mathbb{R}$, to be equivalent to '$|x_k - L| \geq \varepsilon$'.

If we combine all these results we get the following condition for the sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ to *not* be convergent:

'For every real number L, there exists $\varepsilon > 0$ such that for every N in $\mathbb{N}$ there exists k in $\mathbb{N}$, with $k \geq N$, such that $|x_k - L| \geq \varepsilon$.'

This last statement implies that the set of indices k for which $|x_k - L| \geq \varepsilon$ is not bounded above by a number in $\mathbb{N}$. By Theorem (I.8.4), this last statement can be reformulated as

'For every real number L, there exists $\varepsilon > 0$ such that there exist infinitely many indices k in $\mathbb{N}$ such that $|x_k - L| \geq \varepsilon$.'

The preceding analysis seems lengthy, but that is mainly because all the details are included. In mathematical writing it is assumed that the reader can carry out such an analysis without help, and can do it fairly rapidly (and accurately). With a little experience in carefully reading correct proofs in mathmatical texts, one finds that this assumption becomes reality.

One reason for the complexity of this analysis is that it reverts to the original definition of 'convergence'. Often, however, the same result can be proved more easily later on by using structural theorems; for instance, see Example (III.2.3).

Final Warning Consider the following statement: 'All automobiles are green'

The correct negation of this statement is 'There is at least one automobile which is not green.' (This is often written more informally as 'Some automobiles are not green'; but the use of the plural here may make it appear that one is saying that more than one automobile is not green, which is not what is intended.)

Unfortunately, some beginners would state the negation of the original statement as 'All automobiles are not green'; equivalently, 'No automobile is green'. One simple way to avoid this common error is to formulate the negation of a statement, as was done above, by simply adding the phrase 'It is not the case that'. The correct negation is usually easy to formulate from that.

III.1.5 Examples

(1) Let c be a real number, and let $\xi = (c, c, c, \ldots)$ be the constant sequence with value c; that is, ξ is the function from $\mathbb{N}$ to $\mathbb{R}$ such that $\xi(k) = c$ for all k. It seems 'obvious' that the sequence ξ is convergent, and $\lim_{k\to\infty} x_k = c$. To verify this directly from Definition (III.1.1), suppose that $\varepsilon > 0$ is given. Clearly $|x_k - c| = 0$, so that $|x_k - c| < \varepsilon$, for every $k \geq 1$. In particular, one can choose $B = 1$ in the definition of 'convergence'; although any larger value of B would work just as well.

More generally, suppose that $\xi : \mathbb{N} \to \mathbb{R}$ is *eventually* the constant c; that is, there exists m in $\mathbb{N}$ such that $\xi(k) = c$ for all $k \geq m$. Then it is easy to see that $\lim_{k\to\infty} x_k = c$.

(2) (a) Suppose that $x_k = 1/k$ for each k in $\mathbb{N}$. It seems 'obvious' that this sequence converges to 0. To verify this directly from Definition (III.1.1), let $\varepsilon > 0$ be given, and let N be a natural number such that $N > 1/\varepsilon$. (That such N exists follows from the Principle of Archimedes.) If $k \geq N$, so that $k > 1/\varepsilon$, then clearly

$$|x_k - 0| = |x_k| = \frac{1}{k} < 1/(1/\varepsilon); \text{ that is, } |x_k - 0| < \varepsilon \text{ if } k \geq N.$$

Thus, $L = 0$ satisfies the requirements for being the limit of the sequence ξ.

(b) More generally, suppose that there is a real number a such that $x_k = a/k$ for all k in $\mathbb{N}$.

Case 1: Suppose that $a = 0$. Then $x_k = 0$ for each k, so the result follows from Example (1) above, with c in that example taken to equal 0.

Case 2: Suppose, instead, that $a \neq 0$. Then by an argument similar to that used in Part (a) it follows that if N in $\mathbb{N}$ satifies the inequality $N > |a|/\varepsilon$, then for each index k such that $k \geq N$ one has

$$|x_k - 0| = \frac{|a|}{k} \leq \frac{|a|}{N} < |a| \frac{\varepsilon}{|a|} = \varepsilon.$$

As before, this implies that $L = 0$ is the limit of the given sequence.

Note Handling the case $a = 0$ separately from the case $a \neq 0$ is not a waste of time: The proof in the case $a \neq 0$ involves division by $|a|$, a process that is not possible in the case $a = 0$.

(3) Consider the sequence ξ whose k-th term x_k is given by

$$x_k = \frac{3k^2 + k + 5}{6k^2 + 1} \text{ for } k \text{ in } \mathbb{N}.$$

In the previous examples there was an 'obvious' candidate to play the role of the number L in Definition (III.1.1). However, in the present example the candidate for L is not at all obvious, especially for one inexperienced with such problems. Thus, let us compute x_k for several values of k to try to get some insight. One gets (after rounding off the decimals)

k	1	5	10	20	50	100	1000
x_k	1.28	0.5629	0.5241	0.5102	0.5036	0.50174	0.50017

As k gets larger and larger, it seems plausible that x_k is approaching $1/2$. Thus let us see what happens if we try $L = 1/2$. Using standard algebra, such as putting fractions over a common denominator and then simplifying the results, and some simple facts about inequalities, one computes:

$$\left|x_k - \frac{1}{2}\right| = \left|\left(\frac{3k^2 + k + 5}{6k^2 + 1}\right) - \frac{1}{2}\right| = \left|\frac{-(6k^2 + 1) + 2(3k^2 + k + 5)}{2(6k^2 + 1)}\right| = \frac{2k + 9}{12k^2 + 2} < \frac{2k + 9}{12k^2} \leq \frac{2k + 9k}{12k^2} = \frac{11}{12k}.$$

Now let $\varepsilon > 0$ be given. Then, using the same reasoning as in Example (2 b) above, let N in $\mathbb{N}$ be chosen so that $N > \varepsilon/a$, where $a = 11/12$. It is clear that if $k \geq N$ then $|x_k - 1/2| < \varepsilon$. Thus, the sequence ξ is convergent, and $\lim_{k \to \infty} x_k = 1/2$, as expected.

Note The approach taken in this example tries to determine convergence using Definition (III.1.1) directly. This requires knowing – or at least guessing correctly – the limit $L = 1/2$ in advance. Indeed, the sequence used here was chosen precisely because the 'guessing' method would work. However, it is not hard to construct examples of sequences for which the exact value of L can not be predicted, so some other method must be used. In Section (III.3) the present example is analysed again using such a method.

(4) Let $\xi = (2, -2, 2, -2, \dots)$ be the sequence whose k-th term x_k equals $(-1)^{k-1}{\cdot}2$ for each k in $\mathbb{N}$. Then ξ is divergent.

Indeed, suppose that L is any real number. If $L \geq 0$, then clearly for all even values of k one has $|x_k - L| = |-2 - L| = L + 2 \geq 2$; in particular, this holds for infinitely many values of the index k. It follows from Theorem (III.1.4), with $\varepsilon = 2$, that ξ does not converge to L if $L \geq 0$.

A similar argument shows that if $L \leq 0$, then $|x_k - L| = 2 + |L| \geq 2$ for all the (infinitely many) *odd* k, and thus the sequence still does not converge to L.

(5) Let $\zeta = (1^2, 2^2, 3^2, \ldots k^2, \ldots)$ be the sequence whose k-th term z_k is given by $z_k = k^2$ for each k in $\mathbb{N}$. Let L be any real number.

If $L \leq 0$, then clearly $|x_k - L| = k^2 + |L| \geq k^2 \geq 1$ for every index k. If, instead, $L > 0$, then there exists a positive integer N such that $L \leq N$. It follows that if k is any index such that $k \geq N+1$, then $k^2 \geq k \geq N+1 \geq L+1$; that is, $|x_k - L| = x^2 - L \geq 1$. In particular, for every L there are infinitely indices k such that $|x_k - L| \geq 1$, so it follows from Theorem (III.1.4), with $\varepsilon = 1$ in that result, that ξ is divergent.

Remark The results obtained for Examples (4) and (5) above can be verified much more quickly once we prove Theorem (III.2.1) below; indeed, somewhat more can be proved.

Infinite Limits of Sequences The sequences in Example (III.1.5) (4) and (5) above are both divergent, but they are divergent in very different ways. Indeed, the sequence in Example (4) 'bounces around', with no definite 'goal' as the index k gets larger. In contrast, the sequence in Example (5) does have, in an obvious sense, a 'goal'; namely, its terms 'tend towards' $+\infty$. This type of divergence occurs so often that a special terminology has been developed to describe it.

III.1.6 Definition

Let $\xi = (x_1, x_2, \ldots)$ be a sequence of real numbers.

(1) One says that the sequence ξ **diverges to** $+\infty$, and one writes $\lim_{k\to\infty} x_k = +\infty$, if, for every positive real number M, there is a real number B such that if $k \geq B$ then one has $x_k \geq M$; equivalently: for every real $M > 0$ the sequence is eventually greater than M.

Similarly, one says that ξ **diverges to** $-\infty$, and one writes $\lim_{k\to\infty} x_k = -\infty$, if for every negative real number M, there is a real number B such that if $k \geq N$ then one has $x_k \leq M$; equivalently: for every real $M < 0$, the sequence is eventually less than M.

(2) One says that the sequence ξ **has a limit**, or that $\lim_{k\to\infty} x_k$ **is defined**, or that $\lim_{k\to\infty} x_k$ **exists** if one of the following three possibilities holds:

(i) ξ is convergent to some real number L;

(ii) ξ diverges to $+\infty$;

(iii) ξ diverges to $-\infty$.

If either (ii) or (iii) holds, one can say that ξ **has an infinite limit**. If, instead, case (i) holds, one can say that ξ has a **finite limit**, although that gives no more information than simply saying that ξ is convergent, since (as we shall see) it may well be that the value of the liimit L may not be known.

III.1.7 Remarks

(1) The preceding definition allows one to make sense of an equation of the form $\lim_{k\to\infty} x_k = L$ whenever L is an extended real number; that is, either L is an ordinary real number, or L is one of the infinities, $+\infty$ or $-\infty$.

(2) Sometimes, in place of the notation $\lim_{k\to\infty} x_k = L$, authors use the notation $x_k \to L$ as $k \to \infty$. The latter symbolism is read 'x_k approaches the limit L as k approaches infinity'.

(3) It is tempting to translate an equation such as $\lim_{k\to\infty} x_k = +\infty$ as 'the sequence ξ *converges* to $+\infty$'. This is improper usage; if ξ has an infinite limit, one should say that ξ *diverges* to that limit.

(4) It is common in calculus texts to write down expressions and only afterwards ask whether the expression makes sense. For instance, a typical 'limit' problem in such a text mught take the form

'Determine whether $\lim_{k\to\infty} x_k$ exists'

where x_k is given by a specific formula. A better formulation would be

'Determine whether the sequence $(x_1, x_2, \ldots x_k, \ldots)$ has a limit.'

In *This Textbook* we attempt to follow the latter model. However, be aware that many authors use the former model freely. Indeed, there are situations in which that mode of expression is so ingrained in standard usage that trying to avoid it would cause confusion.

III.1.8 Examples

Note: The statements made in the examples here follow easily from Definition (III.1.6), and their proofs are left as simple exercises.

(1) Let $\xi = (1, 2, 3, \ldots)$, so that $x_k = k$ for all n. Then $\lim_{k\to\infty} x_k = +\infty$. Likewise, the sequence $-\xi = (-1, -2, -3, \ldots)$ has the limit $-\infty$.

(2) The sequence $(1, -2, 3, -4, \ldots)$ does not have a limit.

(3) The sequence $\xi = (0, 4, 0, 8, 0, 12, \ldots)$ for which $x_k = k + (-1)^k k$.

Remark The careful reader will have noticed that, depending on whether L is finite or one of the infinities, the description of '$\lim_{k\to\infty} x_k = L$' assumes different forms. In Definition (III.1.1) and what follows, one focuses on the expression $|x_k - L|$. This expression describes the distance between x_k and L if L is finite, and thus it is adequate to describe the intuitive requirement that 'x_k is close to L if k is large'. In constrast, if $L = \pm\infty$, the expression $|x_k - L|$ equals $+\infty$ for each index k (see Remark (II.2.8)), and thus provides no information about how 'close' the real number x_k is to the (infinite) quantity L. Because of this dichotomy, one is frequently forced to provide separate proofs for results involving an equation $\lim_{k\to\infty} x_k = L$ depending on whether L is finite, $+\infty$ or $-\infty$. Although providing such separate proofs is normally not difficult, it can be annoying. The next result can sometimes be used to give unified proofs which work regardless of whether L is finite or infinite.

III.1.9 Theorem

Suppose that $\xi = (x_1, x_2, \ldots)$ is a sequence of real numbers, and that L is an extended real number. Then a necessary and sufficient condition for the equation $\lim_{k\to\infty} x_k = L$ to hold is that both of the following statements are correct:

Statement A If y is any real number such that $y < L$, then $y < x_k$ for all but finitely many indices k.

Statement B If z is any real number z such that $z > L$, then $x_k < z$ for all but finitely many indices k.

Proof Case 1: L is finite

Suppose that $\lim_{k\to\infty} x_k = L$. Let y be any number such that $y < L$, and let $\varepsilon = L - y$, so that $\varepsilon > 0$, and $y = L - \varepsilon$. The convergence hypothesis implies that $|L - x_k| < \varepsilon$, that is, $L - \varepsilon < x_k < L + \varepsilon$, for all but finitely many indices k. In particular, since $L - \varepsilon = y$, Statement A holds. A similar argument shows that Statement B holds.

Conversely, suppose that both Statement A and Statement B hold. Let $\varepsilon > 0$ be given, and set $y = L - \varepsilon$ and $z = L + \varepsilon$. Then the truth of these statements implies that one has $L - \varepsilon < x_k$ and $x_k < L + \varepsilon$, that is, $|x_k - L| < \varepsilon$, for all but finitely many indices k. Thus, ξ converges to L, as required.

Case 2 L is infinite; that is, $L = +\infty$ or $L = -\infty$.

In this case a similar analysis works; indeed, the analysis is slightly easier. For example, if $L = +\infty$, then Statement B is automatically true, because there are no real numbers z such that $z > +\infty$. Thus in this case one needs to check only that Statement A is true. Likewise, if $L = -\infty$, then one needs to check only that Statement B is true.

Extension of the Definition of Limit of a Sequence This is a convenient place to discuss one feature of the definition of 'infinite sequence' which is a source of mild irritation in practice. Specifically, Definition (I.9.1) defines an infinite sequence to be a function $\alpha = (x_1, x_2, \dots x_k, \dots)$ whose domain is $\mathbb{N}$, the set of *all* natural numbers. This form of the definition is widely used, and it corresponds exactly to the intuition that 'sequence' is a fancy word for 'ordered list'.

The 'mild irritation' referred to above is this: in dealing with an expression such as $\lim_{k\to\infty} x_k$, one really needs to consider only terms x_k with k large. More precisely, given any positive integer N, one can ignore all the terms x_k with $k \leq N$. Whether the sequence has a limit L, and the value of L if it exists, does not change even if one changes the values of x_k for $k \leq N$. Indeed, it is not important for such 'limit' questions whether x_k is defined for $k \leq N$.

For example, one would like to say that $\lim\limits_{k\to\infty} \dfrac{k}{k-2} = 1$. Indeed, one computes that

$$\frac{k}{k-2} = \frac{k}{k\,(1-(2/k))} = \frac{1}{1-(2/k)};$$

and when k is large then $-2/k$ is almost equal to 0, and so on.

Unfortunately, the fraction $k/(k-2)$ is not defined for *all* k in $\mathbb{N}$: there is a 'division-by-zero' issue when $k = 2$. In this case the notation $\lim\limits_{k\to\infty} k/(k-2)$ is technically not even defined, since the 'limit' notation assumes that one is dealing with a true sequence. One can get around this by introducing a true sequence $\sigma = (s_1, s_2, \dots)$ by the rule

$$s_k = \begin{cases} \dfrac{k}{k-2} & \text{if } k \neq 2 \\[2ex] b & \text{if } k = 2 \end{cases}$$

where the number b can be any real number, and then study the limit of this modified (but 'legal') sequence.

The obvious problem with this solution is its artificiality, made evident by the fact that the choice of the value b is irrelevant to the issue; it seems excessively fussy for what it accomplishes.

The next definition uses the results of Theorem (III.1.9) as motivation to extend the definition of '$\lim_{k\to\infty} x_k = L$.

III.1.10 Extended Definition of the Limit of a Real Sequence

Suppose that ξ is a real-valued function whose domain is a subset of $\mathbb{N}$ of the form $\mathbb{N}\backslash S$, where S is a finite subset of $\mathbb{N}$. Let $x_k = \xi(k)$ for all k in $\mathbb{N}\backslash S$, and let L be an extended real number.

One says that $\lim_{k\to\infty} x_k = L$ (in the extended sense) provided both of the following statements hold:

Statement A For each y in $\mathbb{R}$ such that $y < L$, there are only finitely many values of the index k in $\mathbb{N} \setminus S$ such that $x_k < y$.

Statement B For each z in $\mathbb{R}$ such that $z > L$, there are only finitely many values of the index k in $\mathbb{N} \setminus S$ such that $x_k > z$.

Note It is clear that this definition is equivalent to the usual one in the special case $S = \emptyset$ so that $\mathbb{N} \setminus S = \mathbb{N}$. Thus there is no harm in using it in that case as well.

III.2 Basic Theory of Limits of Sequences

III.2.1 Theorem

Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be a sequence of real numbers.

(a) (**Uniqueness-of-Limits Principle**) If the sequence ξ has a limit, then its limit is unique. More precisely, suppose that L is an extended real number such that the equation $\lim_{k\to\infty} x_k = L$ is true. If L' is an extended real number such that $L' \neq L$, then the equation $\lim_{k\to\infty} x_k = L'$ is *not* true.

(b) Suppose that ξ has limit L. Then:

(i) If L is a real number, so that ξ is convergent, then ξ is a bounded sequence.

(ii) If $L = +\infty$, so that ξ diverges to $+\infty$, then the sequence ξ is bounded below but unbounded above.

(iii) If $L = -\infty$, so that ξ diverges to $-\infty$, then ξ is unbounded below but bounded above.

(c) Suppose that $\xi = (x_1, \dots x_k, \dots)$ is a sequence of real numbers with limit L. Let $\zeta = (z_1, z_2, \dots)$ be a subsequence of ξ. Then $\lim_{j\to\infty} z_j = \lim_{k\to\infty} x_k = L$.

Alternate Phrasing: A sufficient condition for a subsequence ζ of a sequence ξ to have the limit L is that ξ have the limit L.

Proof

(a) Suppose that $\lim_{k\to\infty} x_k = L$, and let L' be any extended real number such that $L' \neq L$. Assume first that $L' < L$, and let c be a real number such that $L' < c < L$. By Statement A in Theorem (III.1.9), there are only finitely many values of the index k such that $x_k < c$. In other words, $x_k \geq c > L'$ for all but finitely many values of k. Now Statement B of the same theorem implies that the sequence ξ does *not* have L' as a limit. By a similar argument one proves that $L' > L$ implies that ξ does not have L' as a limit. The desired result follows.

(b) (i) Let y be a real number such that $y < L$; for example, let $y = L - 1$. (This subtraction makes sense because, by hypothesis, L is a real number.) By Statement A in Theorem (III.1.9) one has $y < x_k$ for all but finitely many indices k. By Part (b) of Corollary (II.3.15), with $f = \xi : \mathbb{N} \to \mathbb{R}$ in that result, it then follows that the sequence ξ is bounded below. A similar argument shows that ξ is also bounded above, and thus ξ is a bounded sequence, as claimed.

(ii) Let y be any real number; of course one automatically has $y < +\infty$. Then, by Statement A of Theorem (III.1.9) again, it follows that $y < x_k$ for all but finitely many indices k; therefore, by Part (b) of Corollary (II.3.15) again, the sequence ξ is bounded below. Furthermore, since y can be any real number, and $\mathbb{N}$ is an infinite set, it follows that the sequence ξ is *un*bounded above.

(iii) This follows much as in (ii) above.

(c) Let $A = \{k_1, k_2, \ldots k_m, \ldots\}$ be an infinite subset of $\mathbb{N}$, with $k_1 < k_2 < \ldots < k_m < \ldots$, such that ζ is the subsequence of ξ determined by A; see Definition (I.9.3). Suppose that y is any real number such that $y < L$. Then since L is the limit of the sequence ξ, one has $x_k < y$ for at most finitely many indices k. It follows that $x_{k_m} < y$ for only finitely many m. A similar proof shows that if z is a real number such that $z > L$, then there are at most finitely many m such that $x_{k_m} > z$. It now follows from Theorem (III.1.9) that L is also the limit of the subsequence ζ, as claimed.

III.2.2 Remark

We have already seen examples in which a subsequence ζ of a given ξ has limit L, but ξ does not; that is, the 'sufficient condition' described above in the Alternate Phrasing of Part (c) of Theorem (III.2.1) is not necessary for ζ to have limit L. However, it is a simple exercise to show that if ζ is a *tail* of the sequence ξ (see Example (I.9.4)), then this condition is also necessary.

III.2.3 Example

In Example (III.1.5) (4) it was proved, directly from the definition of 'convergence', that the sequence $\xi = (2, -2, 2, \ldots (-1)^{k-1}\, 2, \ldots)$ is divergent. In light of Part (c) of the preceding theorem, this proof can be made simpler and more intuitive; actually, even more can be proved. Indeed, the odd-order terms of ξ form a constant sequence with limit 2, while the even-order terms form a constant sequence with limit -2. Since these subsequences do not have the same limit, the original sequence ξ cannot have a limit, finite or infinite. In particular, ξ is divergent, as previously claimed.

Similarly, the preceding theorem can provide a simpler proof of the result in Example (III.1.5) (5) given earlier. Indeed, the sequence $\zeta = (1^2, 2^2, 3^2, \ldots k^2, \ldots)$ is clearly unbounded, so by Theorem (III.2.1) (b) (i) the sequence ζ is not convergent. Indeed, it is easy to prove that this sequence has $+\infty$ as its limit.

III.2.4 Remark

The converses of Statements (i), (ii) and (iii) in Part (b) of the previous theorem are not true. More precisely, knowing that a sequence is bounded on one or both sides does not imply that it has a limit. Examples illustrating this fact have already been presented.

III.2.5 Theorem

Recall from Chapter (I) that if $\zeta = (z_1, z_2, \ldots z_k, \ldots)$ is a sequence in a nonempty set X, then S_ζ denotes the 'term-set' of ζ; that is, the range $\{z_1, z_2, \ldots z_k, \ldots\}$ of the function $\zeta : \mathbb{N} \to X$.

(a) **Monotonic-sequences Principle** Suppose that $\alpha = (a_1, a_2, \dots a_k, \dots)$ is a real sequence which is monotonic up; that is, $a_{k+1} \geq a_k$ for all k in $\mathbb{N}$ (see Definition (II.3.10)). Then $\lim_{k\to\infty} a_k$ exists, and equals the supremum of the term-set S_α. In particular, this limit is finite (and the sequence α is convergent) if α is bounded above, while the limit is $+\infty$ if α is unbounded above.

Likewise, suppose that $\beta = (b_1, b_2, \dots b_k, \dots)$ is a real sequence which is monotonic down. Then β has a limit. This limit is finite, and β is convergent to it, if β is bounded below; while the limit is $-\infty$ if β in unbounded below.

(b) **Squeeze Theorem for Sequences** Suppose that $\alpha = (a_1, a_2, \dots)$ and $\beta = (b_1, b_2, \dots)$ are real sequences, each having the same limit L. Suppose further that $\xi = (x_1, x_2, \dots)$ is a real sequence such that $x_k \in \mathrm{Seg}\,[a_k, b_k]$ for all but finitely many indices k. Then ξ also has the same limit L.

(c) Let $\xi = (x_1, x_2, \dots)$ be a real sequence, and let S_ξ be the term-set of ξ. Assume that the sequence ξ has limit L. Then

$$\inf S_\xi \leq L \leq \sup S_\xi \qquad (*)$$

More generally, let m be a lower bound for the sequence ξ, and let M be an upper bound for ξ. Then

$$m \leq L \leq M \qquad (**)$$

Proof

(a) Let $L = \sup S_\alpha$, and let y be any real number such that $y < L$. By the Approximation Property for the supremum, there must exist an element u of the set S_α such that $y < u \leq L$. By the definition of the set S_α, this implies that there exists an index m such that $u = x_m$ and thus $y < x_m \leq L$. By the 'monotonic-up' hypothesis, this in turn implies that if $k \geq m$ then $y < x_m \leq x_k \leq L$. That is, Statement A of Theorem (III.1.9) is satisfied. Since L is an upper bound for the set S_α, it follows that if z is a number such that $z > L$, then there are no indices k such that $x_k > z$, and thus Statement B of Theorem (III.1.9) also holds, and thus $\lim_{k\to\infty} x_k = L$, as required.

The proof in the case of monotonic-down sequences is similar.

(b) Let y be any number such that $y < L$. Then by Theorem (III.1.9) (and the hypothesis that the sequences α and β both have limit L), it follows that $a_k, b_k \geq y$ for all but finitely many indices k. It follows that, for all but finitely many k, every element x of $\mathrm{Seg}\,[a_k, b_k]$ satisfies $x \geq y$. In particular, $x_k \geq y$ for all but finitely many indices k. In a similar manner, if $z > L$ then one has $x_k \leq z$ for all but finite many indices k. It follows from Theorem (III.1.9) again that the sequence ξ also has limit L, as required.

(c) The simple proof is left as an exercise.

III.2.6 Remarks

(1) In Theorem (II.4.31), the 'Nested-sequences/Nested-segments Theorem', Claim 2 of that theorem was phrased in terms of suprema, and proved using the Eudoxus Principle. The Monotonic-sequences Principle proved above allows an alternate approach to that result. Indeed, in Theorem (II.4.31) one proves that $J = \mathrm{Seg}\,[A, B]$, where $A = \sup S_\alpha$ and $B = \inf S_\beta$ (see Theorem (II.4.31) for the definition of these quantities). The Monotonic-sequences Principle allow one to state that $A = \lim_{k\to\infty} a_k$, and $B = \lim_{k\to\infty} b_k$, which is the usual formulation of the Nested-intervals Theorem. A similar result holds for the Nested-segments Theorem.

(2) In Part (c) of the preceding theorem it is clear that Inequality $(**)$ never provides an estimate of L which is more precise than that given by Inequality $(*)$. Thus, it may seem pointless to even mention Inequality $(**)$. In practical cases, however, computing the exact values of $\inf \xi$ and $\sup \xi$ may be very difficult, while determining values of m and M which are sufficiently accurate for the needs of the given situation may be easy. We encounter several examples of this phenomenon later.

(3) The word 'squeeze' which appears in the name of Part (b) of the preceding theorem comes from the fact that the hypotheis $x_k \in [a_k, b_k]$ means that the number x_k is 'squeezed' between the numbers a_k and b_k.

(4) In many texts the theorem called the 'Squeeze Theorem for Sequences' would have an stronger hypothesis such as that $a_k \leq b_k$ for each index k. It is a simple exercise to show that the resulting theorem is equivalent to the one obtained above.

III.2.7 Theorem

Suppose that S be a nonempty set of real numbers. Then there exists a sequence $\xi = (x_1, x_2, \dots x_k, \dots)$ of numbers in S such that ξ is monotonic up and $\lim_{k\to\infty} x_k = \sup S$. More precisely, if $\sup S \in S$, so that S has a maximum element, then ξ can be chosen to be a constant sequence. If, instead, $\sup S \notin S$, then ξ can be chosen to be strictly increasing. Likewise, there exists a monotonic-down sequence $\zeta = (z_1, z_2, \dots z_k, \dots)$ of numbers in S such that $\inf S = \lim_{k\to\infty} z_k$. This sequence can be chosen to be constant if $\inf S \in S$, or chosen to be strictly decreasing if $\inf S \notin S$.

The straight-forward proof is left as an exercise.

III.2.8 Theorem

Let $\xi = (x_1, x_2, \dots)$ be a sequence of real numbers, and suppose that there are are subsequences $\zeta = (z_1, z_2, \dots)$ and $\tau = (t_1, t_2, \dots)$ of ξ which have limits L_1 and L_2, respectively, with $L_1 \neq L_2$. Then the sequence ξ does not have a limit.

Proof: Suppose, to the contrary, that $\lim_{k\to\infty} x_k = L$ for some extended real L. Since $L_1 \neq L_2$, then at least one of the quantities L_1 or L_2 must not equal L. Thus at least one of the subsequences ζ or τ must fail to have the limit L. This would contradict Part (c) of Theorem (III.2.1).

The next several results apply mainly to questions of convergence.

III.2.9 Theorem

(a) Suppose that $\xi = (x_1, \dots x_k, \dots)$ is a real sequence, and L is a real number. Let $z_k = x_k - L$ for each k in $\mathbb{R}$. Then the following statements are equivalent:

(i) $\lim_{k\to\infty} x_k = L$.
(ii) $\lim_{k\to\infty} z_k = 0$.
(iii) $\lim_{k\to\infty} |z_k| = 0$.

(b) Suppose that $(x_1, x_2, \dots x_k, \dots)$ is a real sequence such that $\lim_{k\to\infty} x_k = 0$. Let $\zeta = (z_1, z_2, \dots)$ be a bounded real sequence, and define $t_k = z_k u_k$ for each k. Then the sequence $\tau = (t_1, t_2, \dots t_k, \dots)$ also converges to 0.

Proof

(a) Suppose that Statement (i) holds; that is, $\lim_{k\to\infty} x_k = L$. If $\varepsilon > 0$, then there is B in $\mathbb{R}$ such that $k \geq B$ implies that $|x_k - L| < \varepsilon$. Thus, since $z_k - 0 = z_k = x_k - L$, it follows that $k \geq B$ implies $|z_k - 0| = |x_k - L| < \varepsilon$. Thus, $\lim_{k\to\infty} z_k = 0$. That is, Statement (i) implies Statement (ii).

Now suppose that Statement (ii) holds, and let $\varepsilon > 0$ be given. Then there exists B in $\mathbb{R}$ such that if $k \geq B$ then $|z_k| = |z_k - 0|, < \varepsilon$. But $||z_k| - 0| = ||z_k|| = |z_k|$, so if $k \geq B$ then $||z_k| - 0| \leq \varepsilon$. Thus, $\lim_{k\to\infty} |z_k| = 0$. That is, Statement (ii) implies Statement (iii).

Finally, suppose that Statement (iii) holds, and let $\varepsilon > 0$ be given. Then there exists B in $\mathbb{R}$ such that if $k \geq B$ then $||z_k| - 0| < \varepsilon$; that is, $|z_k| < \varepsilon$, which can be written $|x_k - L| < \varepsilon$. This means that $\lim_{k\to\infty} x_k = L$, so that Statement (iii) implies Statement (i).

(b) Since ζ is bounded, there must exist a real number $M > 0$ such that $|z_k| \leq M$ for all k. Let $\varepsilon > 0$ be given. Then since ξ converges to 0, there must exist B in $\mathbb{R}$ such that if $k \geq B$ then $|x_k| < \varepsilon/M$. Then for $k \geq B$ one has

$$|z_k x_k| \leq M\,|x_k| \leq M\cdot\frac{\varepsilon}{M} = \varepsilon.$$

Thus, the sequence τ converges to 0, as claimed.

The following simple corollary is of unexpected importance in the theory.

III.2.10 Corollary

(a) Let r be a number in the interval $[0, 1)$, and let ρ denote the geometric sequence with initial term $A = 1$ and common ratio r; that is, $\rho = (1, r, r^2, \dots r^k, \dots)$ (see Example (I.9.2) (4)). Then ρ is convergent, and its limit is 0. In symbols:

$$\lim_{k\to\infty} r^{k-1} = 0 \text{ if } 0 < r < 1;$$

equivalently,

$$\lim_{k\to\infty} r^{k} = 0 \text{ if } 0 < r < 1;$$

(b) More generally, let $\xi = (x_1, x_2, \dots)$ be a of sequence of real numbers such that there exist a number r such that $0 \leq r < 1$, and a number $M \geq 0$, such that $|x_k| \leq M r^k$ for all sufficiently large k. Then the sequence ξ is convergent, and $\lim_{k\to\infty} x_k = 0$.

(c) The conclusion of Part (a) remains true if the hypothesis $0 < r < 1$ is replaced by the weaker hypothesis $|r| < 1$.

Proof

(a) The result is obviously true if $r = 0$, so assume now that $0 < r < 1$. It is clear, from the 'order' properties of real numbers, that ρ is a strictly decreasing sequence which is bounded below by 0. It follows from Part (a) of Theorem (III.2.5), i.e., the Monotonic-sequences Principle, that the sequence ρ converges to the number $L = \inf\{1, r, r^2, \dots r^k, \dots\}$. It is also clear that $0 \leq L \leq r^{k-1}$ for each index k; equivalently, $0 \leq L \leq r^k$ for each index k.

Claim One has $L = 0$.

Proof of Claim Since $0 < r < 1$, so it follows that $1/r > 1$. If the claim were incorrect, then it would follow that $L > 0$ as well, and thus $L/r > L$. Since L is the *greatest* lower bound of the set $\{r, r^2, \dots\}$, it follows that L/r cannot be a lower bound for this set, and thus there must exist k in $\mathbb{N}$ such that $r^k < L/r$. Multiply both of the terms of this inequality by the positive number r to conclude that $r^{k+1} < L$, which contradicts the fact that L is a lower bound for the set $\{r, r^2, \dots\}$. Thus, $L > 0$ is impossible, and the claim follows.

Note An alternative proof is given in Example (III.3.6) below.

(b) This statement follows easily from Part (a) together with Part (b) of the preceding theorem.

(c) This follows from Part (b) by replacing r in that statement by $|r|$, and letting $M = 1$.

III.2.11 Examples

(1) Let $\xi = (x_1, x_2, \dots)$ be the sequence given by the rule

$$x_k = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots + \frac{1}{k^2} \text{ for each } k \text{ in } \mathbb{N}.$$

Note that for each k in $\mathbb{N}$ one has $x_{k+1} = x_k + 1/(k+1)^2$. One computes the first few terms easily:

$$x_1 = 1;\, x_2 = x_1 + \frac{1}{2^2} = \frac{5}{4};\, x_3 = x_2 + \frac{1}{3^2} = \frac{49}{36};\, x_4 = \frac{49}{36} + \frac{1}{4^2} = \frac{820}{576};\, \dots$$

It is clear that the sequence ξ is strictly increasing. Less obvious, but true, is that this sequence is bounded above. Indeed, note that for each index $k \geq 2$ one easily computes that

$$\frac{1}{k^2} < \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k},$$

and thus

$$x_k < 1 + \left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \dots + \left(\frac{1}{k-1} - \frac{1}{k}\right)$$

In the sum of the terms inside the parentheses on the right, all the terms cancels out with the exception of the terms $1/1$ and $-1/k$, leaving one with

$$x_k \leq 1 + 1 - \frac{1}{k} = 2 - \frac{1}{k}.$$

In particular, $x_k < 2 - 1/k < 2$ for all $k \geq 2$. Of course the same holds for $k = 1$, so the sequence ξ is bounded above, by 2. It follows from the Monotonic-squences Principle (see Theorem (III.2.5) (a) above) that the sequence ξ is convergent.

Remark This is a good example of proving that a sequence is convergent, without having even a guess in advance the exact value of its limit L. The argument given above may appear to suggest that the desired limit is $L = 2$, but that is incorrect. Indeed, the correct value turns out to be $L = \pi^2/6$, where π denotes the famous constant from high-school geometry, with value $\pi = 3.14159\dots$, so that L is approximately 1.644, which is much less than 2. The proof that this is the correct value of L here is well beyond the techniques we have available at this point, but you can (informally) convince yourself by computing x_k for some large values of k directly from the fomula for x_k.

(2) Suppose that $\xi = (x_1, x_2, \dots)$ is given by the rule

$$x_k = 1 + \frac{1}{2} + \dots + \frac{1}{k} \text{ for all } k \text{ in } \mathbb{N}.$$

At first glance this sequence appears to be very similar to that considered in Example (1) above; for instance, it is also strictly increasing, and one goes from one term to the next by adding the reciprocal of some natural number. However, there is a major difference, because it turns out that the sequence in the current example is *not* bounded above.

To see that ξ is not bounded above, it suffices to find a subsequence which is unbounded. And in fact it is easy to show that the subsequence $\zeta = (z_1, z_2, \dots)$ given by $z_j = x_{2^j}$ satisfies

$$z_j > \frac{j}{2}.$$

To see why this is true, use Mathematical Induction, and let A be the set of j in $\mathbb{N}$ such that $z_j > j/2$. Clearly $1 \in A$, since $z_1 = x_2 = 1 + 1/2 > 1/2$. Next, suppose that $j \in A$, so that $z_j > j/2$. Note that

$$z_{j+1} = x_{2^{j+1}} = x_{2^j} + \frac{1}{2^j+1} + \frac{1}{2^j+2} + \dots \frac{1}{2^{j+1}}$$

Clearly the smallest of the fractions appearing on the right side of the last equation is the final fraction, $1/2^{j+1}$. Since there exactly 2^j such fractions, their sum is at least $2^j/2^{j+1} = 1/2$. Also, the term x_{2^j} appearing there is the same as z_j. Thus, one gets

$$z_{j+1} > z_j + \frac{1}{2} > \frac{j}{2} + \frac{1}{2} = \frac{j+1}{2}.$$

(The induction hypothesis, that $j \in A$, is used in the inequality $z_j > j/2$.) Thus $A = \mathbb{N}$, so $z_j > j/2$ for each j in $\mathbb{N}$, as claimed. It now follows from the Archimedean Principle that the subsequence ζ is unbounded, and thus that the original sequence ξ is also unbounded. Now Theorem (III.2.1) (c) implies that the sequence ξ is divergent; in fact, it has the limit $+\infty$.

(3) Let R be a positive number, and define a sequence $\xi = (x_1, x_2, \dots)$ by the rule

$$x_k = \frac{R^k}{k!} \text{ for } k = 1, 2, \dots$$

Then $\lim_{k \to \infty} x_k = 0$. To see this, let N be a positive integer such that $N \geq R$; then set $r = R/(N+1)$, so that $0 < r < 1$. (Such N exists by the Archimedean Property.) Note that

$$x_{N+1} = \frac{R^{N+1}}{(N+1)!} = \left(\frac{R^N}{N!}\right) \cdot \left(\frac{R}{N+1}\right) = r x_N.$$

By repeating this procedure, one gets $x_{N+2} < r x_{N+1} < r^2 x_N$, and so on. The general result is

$$x_{N+k} < r^k x_N \text{ for } k = 1, 2, \dots$$

It now follows from Part (b) of the preceding corollary that the sequence ξ converges to 0.

Note The sequence ξ in this example is not as strange as may first seem. Indeed, such ratios appear frequently in the so-called 'Power Series' one studied in elementary calculus.

III.2.12 Example – Heron's Method

Preliminary Remark As has already been mentioned, it is not hard to prove, using the Bisection Procedure, that $\sqrt{2}$ exists; in other words, that there exists a real number x such that $x^2 = 2$. As a computational tool, however, the 'Bisection' approach is rather inefficient.

There is a second procedure for computing square roots, however, that is generally much more efficient than the Bisection Procedure, and which uses the Nested Segments Theorem (see Theorem (II.4.31)). It is has long been called 'Heron's Method', in honor of the ancient Greek geometer Heron, from around the year 100. It is still widely used; see, for instance, Chapter Quote #2 at the start of this chapter.

Note In the twentieth century it was discovered that apparently the ancient Babylonians were also familiar with this method centuries before Heron. Because of this, it is sometimes called the 'Babylonian Method'.

The basic idea behind Heron's method is quite simple: Let C be a positive real number. It is clear that if u is a positive number such that $u^2 > C$, then (C/u) is a number such that $(C/u)^2 < C$. Indeed,

$$\left(\frac{C}{u}\right)^2 = \frac{C^2}{u^2} < \frac{C^2}{C} = C.$$

Likewise, if $u^2 < C$, then $(C/u)^2 > C$; and if $u^2 = C$, then $(C/u) = C$, and conversely.

In other words, *if* the positive number C has a positive square root $\sqrt{C}$, then for every number $u > 0$ the desired square root must lie in $\mathrm{Seg}\,[u, C/u]$. Heron's Method consists of using this basic idea repeatedly, together with one extra ingredient: since we know that the desired square root, if it exists, must lie in this segment, it makes sense to use the midpoint of this segment as an 'improved' guess of the actual value of the square root.

III.2.13 Theorem (Heron's 'Divide-and-Average' Method for Square Roots)

Let C and x_1 be positive real numbers, and let $\xi = (x_1, x_2, \dots)$ be the sequence of positive real numbers constructed recursively from C and x_1 as follows:

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{C}{x_k}\right), \qquad \text{(III.1)}$$

so that x_{k+1} is the midpoint of $\mathrm{Seg}\,[x_k, C/x_k]$. For each k in $\mathbb{R}$ let J_k denote the segment $\mathrm{Seg}\,[x_k, C/x_k]$. Then the segments J_1, $J_2, \dots$ form a nested sequence such that $\lim_{k\to\infty} |x_k - C/x_k| = 0$. In addition, the sequence $\xi = (x_1, x_2, \dots)$ converges to a number $L > 0$ such that $L^2 = C$.

Proof It is clear that at each stage of the construction the number x_k is positive. Indeed, $x_1 > 0$ by hypothesis. Then $C/x_1 > 0$ since the ratio of positive numbers is also positive. Then $x_2 > 0$ since x_2 is the average of two positive numbers. A similar argument shows that if x_k has been constructed and is positive, then x_{k+1}, as the average of positive numbers, is positive.

Since x_{k+1} is the midpoint of the segment $\mathrm{Seg}\,[x_k, C/x_k]$, it follows that x_{k+1} is an element of J_k. Likewise, it follows that C/x_{k+1} is an element of $\mathrm{Seg}\,[C/x_k, C/(C/x_k)]$. But $C/(C/x_k) = x_k$, and $\mathrm{Seg}\,[C/x_k, x_k] = \mathrm{Seg}\,[x_k, C/x_k]$ (see Part (c) of Theorem (II.3.4)); thus it follows that $C/x_{k+1} \in J_k$ as well. Since x_{k+1} and C/x_{k+1} are both in J_k, and segments are convex sets, it follows that $J_{k+1} \subseteq J_k$ for each k in $\mathbb{N}$; that is, the segments form a nested sequence, as claimed.

Next, notice that since x_{k+1} is the midpoint of the segment J_k, its distance from any point of J_k is at most half the length of that segment. In particular,

$$\left|x_{k+1} - \frac{C}{x_{k+1}}\right| \le \frac{1}{2}\left|x_k - \frac{C}{x_k}\right| \text{ for each } k \text{ in } \mathbb{N}.$$

By repeated use of this last fact, one gets

$$\left|x_{k+1} - \frac{C}{x_{k+1}}\right| \le \frac{1}{2^k}\left|x_1 - \frac{C}{x_1}\right| \text{ for each } k \text{ in } \mathbb{N}.$$

It follows easily that $\lim_{k\to\infty} |x_k - C/x_k| = 0$, as claimed.

One sees now that the Nested-Segment Theorem, as clarified by Remark (III.2.6), can be applied to conclude that the intersection $\bigcap_{k=1}^{\infty} J_k$ is a singleton set $\{L\}$ such that $L = \lim_{k\to\infty} x_k$.

Finally, note that since L is a member of each segment J_k, it follows as above that C/L is also in each J_k, and thus C/L is in the intersection of these segments. In particular, one gets $C/L = L$, hence $L^2 = C$, as required.

Remark Heron's 'Divide-and-Average' Method does appear in the typical textbook for elementary calculus, often as an optional topic; but usually it falls under the topic of 'Newton's Method for Calcuating Roots', and is treated as an application of the derivative, and not as part of the chapter on sequences.

Side Comment (on computing $\sqrt{2}$: Bisection vs Heron's Method):

Let us try to compute the irrational number $\sqrt{2}$ in two ways: using the Bisection Method, and using the Divide-and-Average Method. We have already carried out ten steps of the Bisection Method on $\sqrt{2}$ using the initial values $a_1 = 1$, $b_1 = 3$; see Example (II.4.18).

To carry out the Divide-and-Average Method one needs a choice of a single initial value x_1, compared to the need for two initial choices a_1 and b_1 in the Bisection Method. To make the current example comparable to Example (II.4.18), let us take $x_1 = (a_1 + b_1)/2$; that is, $x_1 = 2$. Here are the first five results:

k	1	2	3	4	5
x_k	2.0	1.5	1.41667	1.41421569	1.41421356

Compare these values with those obtained in Example (II.4.18):

$$a_{10} = \frac{362}{256} = 1.4140625; \quad b_{10} = \frac{363}{256} = 1.41796875$$

Note that the actual value of $\sqrt{2}$, to nine places, is known to be 1.414213562. Thus the error in approximating $\sqrt{2}$ by x_5 above is quite a bit smaller than the error in approximating $\sqrt{2}$ by either a_{10} or b_{10} in the Bisection Method. That is, Divide-and-Average gets much better accuracy than Bisection with much less computational work. This result is typical: Given comparable starting values a_1, b_1 and x_1 for the two methods, the Divide-and-Average Method converges to the desired square root much more rapidly than the Bisection method. However, the Bisection Method can be applied easily to many more situations, and the rapidity of convergence is normally not an issue in theoretical discussions.

It frequently happens that a sequence is constructed by using one formula for the terms with odd index and a different formula for terms with even index. The next result is often useful in such a situtation.

III.2.14 Theorem (The Odd/Even Limit Theorem)

Let $\xi = \{x_1, x_2, \ldots\}$ be a sequence of real numbers, and let ζ and τ be the subsequences of ξ formed from the odd and even terms, respectively, of ξ. That is,

$$\zeta = (z_1, z_2, z_3, \ldots z_j, \ldots) = (x_1, x_3, x_5, \ldots x_{2j-1}, \ldots);$$

likewise,

$$\tau = (t_1, t_2, t_3, \ldots t_j, \ldots) = (x_2, x_4, x_6, \ldots x_{2j}, \ldots).$$

Otherwise stated:

$$\xi = (z_1, t_1, z_2, t_2, \ldots).$$

Suppose that the subsequences ζ and τ both have the same limit L, where L is an extended real number. Then the original sequence ξ also has limit L. In symbols: if $\lim_{j\to\infty} z_j = L$ and $\lim_{j\to\infty} t_j = L$, then $\lim_{k\to\infty} x_k = L$.

Proof: Let y be any real number such that $y < L$. By Statement A of Theorem (III.1.9), applied to the sequence ζ and its limit L, one has $y < z_j$ for all but finitely many indices j; that is, $y < x_k$ for all but finitely many *odd* indices k. Likewise, the same statement, but now applied to the sequence τ and its limit L, implies that $y < x_k$ for all but finitely many *even* indices k. Since every index is either even or odd, it follows that Statement A of that theorem also applies to ξ. A similar argument shows that ξ satisfies Statement B of that theorem. It follows that $\lim_{k\to\infty} x_k = L$, as claimed.

III.2.15 Example

Define a sequence $\xi = (x_1, x_2, \ldots)$ by the rule

$$x_k = \begin{cases} \dfrac{1}{2} & \text{if } k \text{ is odd} \\[2ex] \dfrac{3k^2+k+5}{6k^2+1} & \text{if } k \text{ is even} \end{cases}$$

Let $\zeta = (z_1, z_2, \ldots)$ and $\tau = (t_1, t_2, \ldots)$ be the corresponding subsequences of ξ formed from the terms of odd index and even index, respectively. That is,

$$z_j = x_{2j-1} = \frac{1}{2} \text{ for each } j \text{ in } \mathbb{N}$$

and

$$t_j = x_{2j} = \frac{3(2j)^2+(2j)+5}{6(2j)^2+1} = \frac{12j^2+2j+5}{24j^2+1} \text{ for all } j \text{ in } \mathbb{N}.$$

It is clear by inspection that the subsequence ζ converges to 1/2. As for the subsequence τ, one could go through the usual calculations to determine its limit, but there is no need to do so. Indeed, the careful reader will notice that the formula for x_k when k is even agrees with the formula for x_j in Example (III.1.5) (3). Thus, ζ is a subsequence of the sequence studied in that earlier example. Since it was proved in that example that the sequence in question converges to 1/2, it follows from Part (b) of Theorem (III.2.1) that the subsequence τ in the current problem also converges to 1/2.

Since both of the subsequences ζ and τ converge to the same limit, namely $L = 1/2$, the 'Odd/Even Limit Theorem' can be used to conclude that the sequence ξ given here also converges to 1/2.

There Odd/Even Convergence Theorem works because the every natural number is either odd or even. The following more general result works in much the same way.

III.2.16 Theorem (Generalized Odd/Even Limit Theorem)

Let $\xi = (x_1, x_2, \ldots x_k, \ldots)$ be a sequence of real numbers. Suppose that there exists an extended real number L and a finite collection of infinite subsets $A_1, A_2, \ldots A_n$ of $\mathbb{N}$ and a finite subset S of $\mathbb{N}$, possibly empty, such that $A_1 \cup A_2 \cup \ldots \cup A_n = \mathbb{N} \setminus S$. Let $\xi_{A_1}, \xi_{A_2}, \ldots \xi_{A_n}$ be the subsequences of ξ that correspond to these infinite subsets of $\mathbb{N}$; see Definition (I.9.3). If each of these subsequences has limit L, then the original sequence ξ also has limit L.

The simple proof is left as an exercise.

III.2.17 Examples

(1) The Odd/Even Convergence Theorem corresponds to the case in which $n = 2$, A_1 is the set of odd natural numbers and A_2 is the set of even natural numbers.

(2) Let r_1, r_2 and r_3 be real numbers in the open interval $(0,1)$. Form a sequence ξ as follows:

$$\xi = (r_1, r_2^2, r_3^3, r_1^4, r_2^5, r_3^6, r_1^7, \ldots);$$

the pattern should be clear. Let $A_1 = \{1, 4, 7, 10, \ldots\}$, $A_2 = \{2, 5, 8, 11, \ldots\}$ and $A_3 = \{3, 6, 9, 12, \ldots\}$. It is easy to see that $A_1 \cup A_2 \cup A_3 = \mathbb{N}$. The corresponding subsequences are easy to figure out:

$$\xi_{A_1} = (r_1, r_1^4, r_1^7, \ldots); \quad \xi_{A_2} = (r_2^2, r_2^5, r_2^8, \ldots); \quad \xi_{A_3} = (r_3^3, r_3^6, r_3^9, \ldots)$$

Each of these sequences is a subsequence of a geometric sequence with common ratio in the open interval $(0,1)$. Since each such sequence converges to 0 (see Corollary (III.2.10)), it follows from the preceding theorem that the original sequence ξ also converges to 0.

(3) Suppose that $A_1, A_2, \ldots A_n, \ldots$ is an *infinite* collection of infinite subsets of $\mathbb{N}$ whose union is $\mathbb{N}$. One might conjecture that if ξ is a real sequence for which each subsequence ξ_{A_n} has the same limit L, then ξ itself must have limit L. Alas, this conjecture would be incorrect. For example, let r be a fixed number in the open interval $(0,1)$; if you prefer to be specific, let $r = 1/2$. Define a real sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ as follows:

> For each index k, express k in the form $k = 2^{n-1}\,(2\,m-1)$ for n, m in $\mathbb{N}$; see Example (I.8.10). Using this notation, set $x_k = r^{m-1}$.

For each n in $\mathbb{N}$ let $A_n = \{2^{n-1}{\cdot}1, 2^{n-1}{\cdot}3, 2^{n-1}{\cdot}5, \ldots 2^{n-1}{\cdot}(2\,m-1), \ldots\}$. It is easy to see that for each n one has $\xi_{A_n} = (1, r, r^2, \ldots r^{m-1}, \ldots)$. This is the geometric sequence with initial term 1 and common ratio r, which is known to converge to 0. However, the subsequence $\zeta = (z_1, z_2, \ldots z_n, \ldots)$ of ξ given by $z_n = x_{2^{n-1}}$ is the constant sequence $(1, 1, \ldots 1, \ldots)$, which does *not* have 0 as limit.

The Monotonic-Sequence Principle (see Theorem (III.2.5)) uses the concepts of 'supremum' and 'infimum' to compute limits of certain sequences. The next result reverses the roles and shows how to use limits of sequences to compute suprema and infima.

III.2.18 Theorem

Let X be a nonempty set of real numbers. Then there exist monotonic sequences $\alpha = (a_1, a_2, \ldots)$ and $\beta = (b_1, b_2, \ldots)$ of elements in X such that $\lim_{k\to\infty} a_k = \inf X$ and $\lim_{k\to\infty} b_k = \sup X$.

More precisely, if $\inf X$ is an element of X then one can take α to be a constant real sequence; but if $\inf X$ is *not* in X, then α can be chosen to be a strictly decreasing sequence.

Likewise, if $\sup X$ is an element of X then one can take β to be a constant real sequence; but if $\sup X$ is *not* in X, then β can be chosen to be a strictly increasing sequence.

Proof Let $A = \inf X$ and $B = \sup X$, First, suppose that $A \in X$. Then it is clear that all the terms of the constant sequence $(A, A, \ldots A, \ldots)$ are elements of X, and the limit of this sequence is $\inf X$, as required.

Next, suppose that $\inf X$ is *not* in X. There are two cases:

Case 1 Suppose that $\inf X \neq -\infty$, so that $\inf X$ is a real number. Define a sequence $\alpha = (a_1, a_2, \ldots)$ by the rule: a_1 is an element of X such that $\inf X < a_1 < (\inf X) + 1$. (The existence of such a_1 follows from the definition of 'infimum'.) Likewise, if $a_1, \ldots a_m$ have been defined then let a_{m+1} be any element of X such that $\inf X < a_{m+1} < a_m$ and $a_{m+1} < (\inf X) + \dfrac{1}{m}$. It is clear that $\alpha = (a_1, a_2, \ldots)$ is a strictly decreasing sequence of reals such that $\inf X < a_m < (\inf X) + 1/m$ for each m in $\mathbb{N}$. Now apply the Squeeze Property to conclude that $\inf X = \lim_{k\to\infty} a_k$.

Case 2 Suppose that $\inf X = -\infty$. Let a_1 be an element of X such that $a_1 < -1$. If $a_1, \ldots a_m$ have been defined, let a_{m+1} be any element of X such that $a_{m+1} < \min\{a_m, -m\}$. Then it is clear that $\alpha = (a_1, a_2, \ldots)$ is strictly decreasing and $\lim_{k\to\infty} a_k = -\infty = \inf X$, as required.

The analysis of $\sup X$ is similar, and is left to the reader.

III.3 Computational Rules for Limits of Real Sequences

Remark The preceding results are mainly of a theoretical nature. In contrast, the results in this section are more practical and computational. Some of them are already probably already familiar from elementary calculus; but several are almost certainly not.

Because the algebraic rules for real numbers differ in important ways from those for extended real numbers, in this section we separate the results for 'convergence' from those involving 'divergence to one of the infinities'. In particular, we often do the convergence proofs directly from Definition (III.1.1), even if a slicker proof is available: it is important that one have experience using the definitions.

The next two theorems involve convergent sequences. The first theorem involves properties of a single convergent sequence, while the second involves pairs of convergent sequences.

III.3.1 Theorem

Suppose that $\xi = (x_1, x_2, \ldots x_k \ldots)$ is a convergent sequence of real numbers. Let $L = \lim_{k\to\infty} x_k$. Then:

(a) (The Constant-Factor Rule for Convergent Real Sequences) One has $\lim_{k\to\infty}(c\,x_k) = c\,L$ for every real number c.

(b) (The Absolute-Value Rule for Convergent Real Sequences) One has $\lim_{k\to\infty} |x_k| = |L|$.

Proof

(a) This proof is left as an easy exercise.

(b) First note that by Inequality (II.4), the 'Reverse Triangle Inequality', one can write

$$0 \leq ||x_k| - |L|| \leq |x_k - L| \text{ for all } k \text{ in } \mathbb{N}. \tag{III.2}$$

Let $\varepsilon > 0$ be given. Then by the convergence hypothesis, there exists N so that if $k \geq N$ then $|x_k - L| < \varepsilon$. Combine this with Inequality (III.2) above to get that $||x_k| - |L|| < \varepsilon$ when $k \geq N$. The desired result follows.

III.3.2 Theorem

Suppose that $\alpha = (a_1, a_2, \dots)$ and $\beta = (b_1, b_2, \dots)$ are convergent sequences of real numbers. Let $A = \lim_{k\to\infty} a_k$ and $B = \lim_{k\to\infty} b_k$.

(a) (The Sum and Difference Rules for Convergent Real Sequences) Let $\sigma = (s_1, s_2, \dots)$, where $s_k = a_k + b_k$ for each k in $\mathbb{N}$; that is, σ is the sequence of sums of corresponding terms from α and β. Likewise, let $\delta = (d_1, d_2, \dots)$, where $d_k = a_k - b_k$ for every k in $\mathbb{N}$; that is, δ is the sequence of differences of corresponding terms from α and β. Then the sequences σ and δ are also convergent. Moreover, $\lim_{k\to\infty} s_k = A + B$ and $\lim_{k\to\infty} d_k = A - B$.

(b) (The Product Rule for Convergent Real Sequences) Let $\varphi = (p_1, p_2, \dots)$, where $p_k = a_k \cdot b_k$ for each k in $\mathbb{N}$; that is, φ is the sequence of products of corresponding terms from α and β. Then the sequence φ is also convergent, and $\lim_{k\to\infty} p_k = A \cdot B$.

(c) (The Quotient Rule for Convergent Real Sequences) Assume in addition that (i) all the terms b_k are nonzero and (ii) the limit B is also nonzero. Let $\rho = (r_1, r_2, \dots)$, where $r_k = a_k/b_k$ for each k in $\mathbb{N}$; that is, ρ is the sequence of ratios of corresponding terms from α and β. Then the sequence ρ is also convergent, and $\lim_{k\to\infty} r_k = A/B$.

Proof:

(a) First note that for all k in $\mathbb{N}$ one has

$$|(a_k + b_k) - (A + B)| = |(a_k - A) + (b_k - B)| \leq |a_k - A| + |b_k - B| \tag{III.3}$$

(The final inequality comes by using the Triangle Inequality.) Now let $\varepsilon > 0$ be given. Then, by the 'convergence' hypotheses, there exist numbers N_1 and N_2 in $\mathbb{N}$ such that

(i) if $k \geq N_1$, then $|a_k - A| < \varepsilon/2$;
(ii) if $k \geq N_2$, then $|b_k - B| < \varepsilon/2$.

Let $N = \max\{N_1, N_2\}$. If $k \geq N$ then $k \geq N_1$ and $k \geq N_2$, so by combining this with Inequality (III.3) one gets

$$|(a_k + b_k) - (A + B)| \leq |a_k - A| + |b_k - B| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \text{ if } k \geq N.$$

Thus $\lim_{k\to\infty}(a_k + b_k) = A + B$, as claimed.

The direct proof of the claim for the sequence δ is similar. Alternatively, one can note that $a_k - b_k = a_k + (-1)\, b_k$, and then combine the Sum Rule with the Constant-Factor Rule to get the desired result.

(b) Using the familiar 'Add-and-Subtract Trick', for each index k one can write

$$a_k\, b_k - AB = a_k\, b_k - a_k\, B + a_k\, B - A\, B = (a_k - A)B + a_k(b_k - B). \tag{III.4}$$

The hypotheses that $\lim_{k\to\infty} a_k = A$ and $\lim_{k\to\infty} b_k = B$ imply (by Part (d) of Theorem (III.2.5)) that $\lim_{k\to\infty}(a_k - A) = 0$ and $\lim_{k\to\infty}(b_k - B) = 0$. Note that the convergence hypothesis implies (by Part (b) of Theorem (III.2.1)) that the factors B and a_k in the expressions $(a_k - A)B$

and $a_k(b_k - B)$ in Equation (III.4) are bounded. Thus one can apply Part (f) of Theorem (III.2.5) to the terms $(a_k - A)B$ and $a_k(b_k - B)$ on the right side of Equation (III.4) to conclude that

$$\lim_{k\to\infty}(a_k - A)B \, = \, 0 \text{ and } \lim_{k\to\infty} a_k(b_k - B) \, = \, 0.$$

Finally, apply Part (a) of the current theorem to what has been shown to conclude that the left side of Equation (III.4) converges to $0 + 0 \, = \, 0$, and thus $\lim_{k\to\infty} a_k b_k \, = \, AB$, as claimed.

(c) Note that for all indices k one has

$$\left|\frac{1}{b_k} - \frac{1}{B}\right| \, = \, \left|\frac{B - b_k}{Bb_k}\right| \, = \, \frac{|B - b_k|}{|Bb_k|}. \tag{III.5}$$

Let $a_k \, = \, 1/(Bb_k)$. By Corollary (III.3.3) there exists $c \, > \, 0$ such that $|b_k| \geq c$ for all k. Thus $|a_k| \, = \, 1/|Bb_k| \leq c/|B|$, so the sequence $(a_1, a_2, \ldots)$ is bounded. Since $\lim_{k\to\infty} b_k \, = \, B$, it follows that $\lim_{k\to\infty} |B - b_k| \, = \, 0$. Thus by Part (f) of Theorem (III.2.5) one has $\lim_{k\to\infty} |B - b_k|/|Bb_k| \, = \, 0$. Compare this with Equation (III.5) to conclude that $\lim_{k\to\infty} \left(\frac{1}{B} - \frac{1}{b_k}\right) \, = \, 0$, which implies the desired result.

III.3.3 Corollary

Let $\xi \, = \, (x_1, x_2, \ldots)$ be a sequence of real numbers which converges to a nonzero number L. Assume, in addition, that none of the terms x_k of the sequence equal 0. Then there exists a number $\delta \, > \, 0$ such that $|x_k| \geq \delta$ for *all* the indices k.

Proof Let $\zeta \, = \, (z_1, z_2, \ldots z_k, \ldots)$ be the sequence given by the rule $z_k \, = \, |1/x_k|$ for each index k. It follows from the Quotient Rule for Convergent Real Sequences (see Part (c) of the preceding theorem), combined with the Absolute-Value Rule for Convergent Real Sequences (see Part (b) of Theorem (III.3.1)), that the sequence ζ conveges to $|1/L|$, and thus is bounded (see Part (b) of Theorem (III.2.1)). In particular, there exists a number $M \, > \, 0$ such that $|z_k| \leq M$, that is, $|1/x_k| \leq M$, for each k. It follows that $|x_k| \geq m \, = \, 1/M \, > \, 0$ for each k.

III.3.4 Corollary

The theory of convergence for bounded sequences of real numbers can be reduced to the corresponding theory for sequences all of whose values lie in the open unit interval $(0, 1)$. More precisely, suppose that $\xi \, = \, (x_1, x_2, \ldots x_k, \ldots)$ is a bounded sequence of reals. Let a and b be numbers such that $a \, < \, x_k \, < \, b$ for each index k. so that $x_k \in [a, b]$. Define a second sequence $\zeta \, = \, (z_1, z_2, \ldots z_k, \ldots)$ by the rule $z_k \, = \, \frac{x_k - a}{b - a}$ for each k for each k. Then $0 \, < \, z_k \, < \, 1$ for each k, and the original sequence ξ is convergent if, and only if, the new sequence ζ is convergent.

The simple proof is left as an exercise.

III.3.5 Remark

In the Quotient Rule (Part (c) of the preceding theorem), it is assumed that for all k one has $b_k \neq 0$. This is done so that all the ratios a_k/b_k are defined; in other words, so that ρ is a true sequence. In light of Definition (III.1.10), however, it is clear that one really need only require that for all sufficiently large k one has $b_k \neq 0$. And it is a simple exercise (left to the reader, of course) to show that the hypothesis $B \neq 0$, already included in the statement of the Quotient Rule, guarantees that $b_k \neq 0$ if k is large enough. Thus, the Quotient Rule for Convergent Sequences can be rephrased in the following more general form:

'Assume, in addition, that the limit B is nonzero. Then $\lim\limits_{k\to\infty} \dfrac{a_n}{b_n} = \dfrac{A}{B}$, where the limit is taken in the 'Extended Sense' of Definition (III.1.10)'

III.3.6 Examples

(1) In Corollary (III.2.10) it is shown that if $0 < r < 1$ then $\lim_{k\to\infty} r^k = 0$. Here is an alternate proof.

As in the earlier proof, note that the sequence $\xi = (r, r^2, r^3, \dots)$ is strictly decreasing and bounded below by 0, hence this sequence has a (finite) limit $L \geq 0$. Next, notice that $r^{k+1} = r^k \cdot r$, so that the sequence $\tau = (r^2, r^3, \dots)$ is obtained by multiplying each term of the sequence ξ by r. By the Constant Factor Rule for Convergent Sequences (i.e., Part (a) of Theorem (III.3.1)) it follows that τ converges to Lr. But clearly τ is a subsequence of ξ, so by Part (b) of Theorem (III.2.1) it follows that τ has the *same* limit as ξ. Thus $L = Lr$, hence $L(1-r) = 0$. Since $r \neq 1$, it follows that $L = 0$.

(2) Consider the sequence ξ whose k-th term x_k is given by

$$x_k = \frac{3k^2 + k + 5}{6k^2 + 1} \text{ for } k \text{ in } \mathbb{N}.$$

We determined the limit of this sequence in Part (3) of Example (III.1.5) by first guessing that this limit should be 1/2. Analyzing such a limit propblem becomes much easier using the limit laws obtained in this section.

Indeed, factor out the highest power of k in both the numerator and denominator of the fraction defining x_k, and simplify, to get:

$$x_k = \frac{k^2\,(3 + 1/k + 5/k^2)}{k^2\,(6 + 1/k^2)} = \frac{3 + 1/k + 5/k^2}{6 + 1/k^2}.$$

It is clear that

$$\lim_{k\to\infty}\left(3 + \frac{1}{5} + \frac{5}{k^2}\right) = 3, \text{ and } \lim_{k\to\infty}\left(6 + \frac{1}{k^2}\right) = 6.$$

It follows from the Quotient Rule for Limits that the original sequence ξ is convergent, and that $\lim_{k\to\infty} x_k = 3/6$, which agrees with the result obtained in the earlier treatment of this sequence.

There are a few 'algebraic rules' which apply to sequences which have *infinite* limits.

III.3.7 Theorem

Suppose that $\alpha = (a_1, a_2, \dots)$ and $\beta = (b_1, b_2, \dots)$ are sequences of real numbers.

(a) If $\lim_{k\to\infty} a_k = +\infty$ and β is bounded below, then $\lim_{k\to\infty}(a_k + b_k) = +\infty$.
Similarly, if $\lim_{k\to\infty} a_k = -\infty$ and β is bounded above, then $\lim_{k\to\infty}(a_k + b_k) = -\infty$.

(b) If $\lim_{k\to\infty} a_k = +\infty$ and there exists $m > 0$ such that $b_k \geq m$ for all k, then $\lim_{k\to\infty}(a_k\, b_k) = +\infty$.

Similarly, if $\lim_{k\to\infty} a_k = +\infty$ and there exists $m < 0$ such that $b_k \leq m$ for all k, then $\lim_{k\to\infty}(a_k\, b_k) = -\infty$.

(c) If $\lim_{k\to\infty} a_k = -\infty$ and there exists $m > 0$ such that $b_k \geq m$ for all k, then $\lim_{k\to\infty}(a_k\, b_k) = -\infty$.

Similarly, if $\lim_{k\to\infty} a_k = -\infty$ and there exists $m < 0$ such that $b_k \leq m$ for all k, then $\lim_{k\to\infty}(a_k\, b_k) = +\infty$.

(d) If there exists $m > 0$ such that $a_k \geq m$ for all k, and if $b_k > 0$ for all k and $\lim_{k\to\infty} b_k = 0$, then $\lim_{k\to\infty} a_k/b_k = +\infty$.

Similarly, if there exists $m > 0$ such that $a_k \geq m$ for all k, and if $b_k < 0$ for all k and $\lim_{k\to\infty} b_k = 0$, then $\lim_{k\to\infty} a_k/b_k = -\infty$.

(e) If there exists $m < 0$ such that $a_k \leq m$ for all k, and if $b_k > 0$ for all k and $\lim_{k\to\infty} b_k = 0$, then $\lim_{k\to\infty} a_k/b_k = -\infty$.

Similarly, if there exists $m < 0$ such that $a_k \leq m$ for all k, and if $b_k < 0$ for all k and $\lim_{k\to\infty} b_k = 0$, then $\lim_{k\to\infty} a_k/b_k = +\infty$.

The simple proofs are left as exercises.

The portions of the preceding theorems which involve $\lim_{k\to\infty} a_k/b_k$ omit two important cases:

The '0/0' Case In this case one has $\lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k = 0$.

The '∞/∞' Case In this case one has $\lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k = \pm\infty$.

The next result sometimes can used to handle these cases. It works by replacing the quotients a_k/b_k by new quotients c_k/d_k which may be able to avoid the 0/0 and ∞/∞ difficulties.

III.3.8 Theorem (Stoltz-Cesaro Theorem)

Let $\alpha = (a_1, a_2, \dots)$ and $\beta = (b_1, b_2, \dots)$ be sequences of real numbers. Assume that α and β satisfy the following conditions:

(i) β is strictly increasing, and

(ii) $\displaystyle\lim_{k\to\infty} \frac{a_{k+1} - a_k}{b_{k+1} - b_k} = L$ for some extended real number L.

Then:

(a) ('0/0 Case') Suppose that both the sequences α and β converge to 0. Then $\displaystyle\lim_{k\to\infty} \frac{a_k}{b_k} = L$.

(b) ('∞/∞' Case) Suppose that $b_k > 0$ for all k and that the sequence β has limit $+\infty$. Then $\displaystyle\lim_{k\to\infty} \frac{a_k}{b_k} = L$.

Notes to Parts (a) and (b) (i) Because the sequence β is assumed to be strictly increasing, in the '0/0' case the sequence β must approach 0 from below; in particular, $b_k < 0$ so the ratio a_k/b_k is defined.

(ii) The hypothesis in the '∞/∞' case, that $b_k > 0$ for every index k, is included to simplifiy the discussion. Indeed, the hypothesis that $\lim_{k\to\infty} b_k = +\infty$ already guarantees that $b_k > 0$, so that the fraction a_k/b_k makes sense, for all but a finite number of values of k. If one understands the limit here in the 'extended' sense of Definition (III.1.10), then the requirement that $b_k > 0$ for *all* k can be dropped and the result remains true.

(c) Similar results hold for both cases if, instead, one assumes that the sequence β is strictly decreasing, and in the '∞/∞' case one assumes that $\lim_{k\to\infty} b_k = -\infty$.

Proof Preliminary Discussion for the Proofs of Parts (a) and (*b*)

Let y be a real number such that $y < L$, and let y' be a real number such that $y < y' < L$. Hypothesis (ii) allows one to use Statement A of Theorem (III.1.9) to say that there exists a natural number N such that if $k \geq N$, then $y' < \dfrac{a_{k+1} - a_k}{b_{k+1} - b_k}$. Hypothesis (i) implies that $b_k - b_{k-1} > 0$ for each k, so it follows from the usual order properties of $\mathbb{R}$ that if $k \geq N$,

then $y'\,(b_{k+1} - b_k) < (a_{k+1} - a_k)$. If n is any natural number and $k \geq N$ then one gets the following string of inequalities:

$$y'\,(b_{k+1}-b_k) < (a_{k+1}-a_k),\ y'\,(b_{k+2}-b_{k+1}) < (a_{k+2}-a_{k+1}),\ \ldots y'\,(b_{k+n}-b_{k+n-1}) < (a_{k+n}-a_{k+n-1}),$$

Add both sides of these inequalities, using the fact that in this 'collapsing sum' most of the terms cancel, to get

$$y'\,(b_{k+n} - b_k) < (a_{k+n} - a_k)$$

Note that $b_k - b_{k+n} = (b_{k+1} - b_k) + (b_{k+2} - b_{k+1}) + \ldots + (b_{k+n} - b_{k+n-1})$, a sum of positive numbers, so $b_{k+n} - b_k > 0$. It follows that

$$y' < \frac{a_{k+n} - a_k}{b_{k+n} - b_k} \text{ if } k \geq N \text{ and } n \in \mathbb{N} \qquad (*)$$

Hypothesis (ii) likewise allows one to use Statement B of Theorem (III.1.9) to prove that if z and z' are real numbers such that $L < z' < z$, there exists N such that if $k \geq N$, then

$$z' > \frac{a_{k+n} - a_k}{b_{k+n} - b_k} \text{ if } k \geq N \text{ and } n \in \mathbb{N} \qquad (**)$$

Proof of (a) In this '0/0' case it is given that the sequences α and β converge to 0. Note that, because β is strictly increasing, it follows that $b_k < 0$ for each index k. Since every subsequence of a convergent sequence also converges to the same limit, it then follows that for each index k one has $\lim_{n\to\infty} a_{k+n} = 0$ and $\lim_{n\to\infty} b_{k+n} = 0$. In particular, if $k \geq N$ then by letting n approach ∞ in Inequality $(*)$, and using the Quotient Rule for Limits (which applies since $b_k > 0$), and recalling that y' was chosen so that $y < y'$, one gets

$$y < y' \leq \frac{(-a_k)}{(-b_k)} = \frac{a_k}{b_k} \text{ if } k \geq N.$$

That is, the sequence $(a_1/b_1, a_2/b_2, \ldots a_{k/b_k,\ldots})$ satisfies Statement A of Theorem (III.1.9). In a similar way, one can use Inequality $(**)$ to show that this sequence also satisfies Statement B of the same theorem. It now follows that $\lim\limits_{k\to\infty} \dfrac{a_k}{b_k} = L$, as claimed.

Proof of (b) Apply Statement A of Theorem (III.1.9) once again to show that there exists a natural number N such that if $k \geq N$, Inequality $(*)$ holds for all n in $\mathbb{N}$. Multiply both sides of the fraction which appears in Inequality $(*)$ by the positive quantity $b_{k+n} - b_k$ and do some simple algebra to get

$$y'\,(b_{k+n}) - y'\,b_k + a_k < a_{k+n}$$

This inequality is valid for every index $k \geq N$, but it is technically easier from here on to focus on the case $k = N$. One then gets

$$y'\,(b_{N+n}) - y'\,b_N + a_N < a_{N+n} \text{ for every } n \text{ in } \mathbb{N}$$

Divide by the positive quantity b_{N+n} to get

$$y' - y'\,\frac{b_N}{b_{N+n}} + \frac{a_N}{b_{N+n}} < \frac{a_{N+n}}{b_{N+n}} \text{ for every } n \text{ in } \mathbb{N}.$$

Since y' and thus N are held fixed – N is determined by the choice of z and z' – it is clear, from the hypothesis that $\lim_{k\to\infty} b_k = +\infty$, that

$$\lim_{n\to\infty}\left(-y'\,\frac{b_N}{b_{N+n}} + \frac{a_N}{b_{N+n}}\right) = 0.$$

In particular, since $y < y'$, there exists an index N' such that if $n \geq N'$ then

$$y < \left(-y'\,\frac{b_N}{b_{N+n}} + \frac{a_N}{b_{N+n}}\right) < \frac{a_{n+n}}{b_{N+n}} \text{ if } n \geq N'$$

This implies that

$$y < \frac{a_k}{b_k} \text{ if } k \geq N + N'$$

It follows that the sequence $(a_1/b_1, a_2/b_2, \ldots a_k/b_k)$ satisfies Statement A of Theorem (III.1.9). A similar proof shows that this sequence also satisfies Statement B, and thus this sequence has limit L, as claimed.

Proof of (c) It is a simple exercise to reduce this to the results obtained in Parts (a) and (b).

III.3.9 Examples

(1) Let $\alpha = (a_1, a_2, \ldots a_k, \ldots)$ and $\beta = (b_1, b_2, \ldots b_k, \ldots)$ be given by the formulas

$$a_k = 3\,k + \frac{5}{k^2} \text{ and } b_k = 7\,k + 2 - \frac{4}{k^3}.$$

Suppose one wishes to determine the convergence of the sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$, where $x_k = a_k/b_k$. It is easy to see that $\lim_{k\to\infty} a_k = +\infty$ and $\lim_{k\to\infty} b_k = +\infty$, so that the 'Standard Quotient Rules' described in Theorem (III.3.2) and in Theorem (III.3.7) do not apply. However, one sees easily that β is strictly increasing. Futhermore, one computes that

$$a_k - a_{k+1} = 3 + \frac{5}{k^2} - \frac{5}{(k+1)^2} \text{ and } b_k - b_{k+1} = 7 - \frac{4}{k^3} + \frac{4}{(k+1)^3}.$$

Note that these are both convergent sequences, with limits 3 and 7, respectively. It then follows directly from the standard Quotient Rule for Convergent Sequences, Part (c) of Theorem (III.3.2), that

$$\lim_{k\to\infty} \frac{a_k - a_{k+1}}{b_k - b_{k+1}} = \frac{3}{7}$$

Thus, the hypotheses of the Stoltz-Cesaro theorem are satisfied, so one can conclude that $\lim_{k\to\infty} x_k = 3/7$.

(2) Suppose that $x_k = a_k/b_k$, where $a_k = 5\,k^2 + 2\,k - 9$ and $b_k = 4 - k^2$. Once again, $\lim_{k\to\infty} a_k = +\infty$ and $\lim_{k\to\infty} b_k = -\infty$, so the standard Quotient Rule does not apply. However it is also clear that $b_{k+1} < b_k$ for each k, so it makes sense to try the Stoltz-Cesaro theorem. One computes:

$$a_k - a_{k+1} = 10\,k + 7 \text{ and } b_k - b_{k+1} = -2\,k - 1.$$

To use the Stoltz-Cesaro theorem, we next consider the limit of the sequence $\zeta = (z_1, z_2, \ldots z_k, \ldots)$, where $z_k = \dfrac{c_k}{d_k}$, with $c_k = a_k - a_{k+1} = 10\,k + 7$ and $d_k = b_k - b_{k+1} = -2\,k - 1$. Once again, one cannot use the standard Quotient Rule to compute $\lim_{k\to\infty}(c_k/d_k)$ because $\lim_{k\to\infty} c_k = +\infty$

and $\lim_{k\to\infty} d_k = -\infty$. However, if one tries the Stoltz-Cesaro Theorem one more time, one gets an answer. Indeed, $c_k - c_{k+1} = (10\,k + 7) - (10\,k + 17) = -7$; likewise, $d_k - d_{k+1} = (-2\,k - 1) - (-2\,k - 3) = 1$. It is clear that $\lim\limits_{k\to\infty} \dfrac{c_k - c_{k+1}}{d_k - d_{k+1}} = -7/1 = -7$. Apply the Stoltz-Cesaro theorem to get $\lim\limits_{k\to\infty} \dfrac{c_k}{d_k} = -7$; that is, $\lim\limits_{k\to\infty} \dfrac{a_k - a_{k+1}}{b_k - b_{k+1}} = -7$. This last fact then implies that $\lim\limits_{k\to\infty} \dfrac{a_k}{b_k} = -7$.

III.3.10 Remarks

(1) The reason for the name '0/0' case in Part (a) of the Stoltz-Cesaro theorem is fairly obvious: the '0' in the numerator corresponds to the hypothesis $\lim_{k\to\infty} a_k = 0$, while the '0' in the demominator corresponds to $\lim_{k\to\infty} b_k = 0$.

In Part (b), however, why the name '∞/∞' is not so clear, since there is no hypothesis of the form $\lim_{k\leftarrow\infty} a_k = \pm\infty$. Actually, such a hypothesis would be redundant 'most of the time'. Indeed, it is an easy exercise to show that the given hypotheses in Part (b) imply that $\lim_{k\to\infty} a_k = +\infty$ if $L > 0$, while they imply that $\lim_{k\to\infty} a_k = -\infty$ if $L < 0$.

(2) The Stoltz-Cesaro theorem is a specialized tool, to be used only when it is needed; mainly for limit problems of the form $0/0$ or ∞/∞. Using Stoltz-Cesaro on a limit problem which is *not* of these forms is likely to give no answer, or worse, the wrong anwser. For instance, suppose that for each k one has $a_k = 1$ and $b_k = 1/k$. One computes that $\lim\limits_{k\to\infty} (a_{k+1} - a_k)/(b_{k+1} - b_k) = 0$, since $a_{k+1} - a_k = 0$ for each k. However, one computes directly that $\lim\limits_{k\to\infty} a_k/b_k = +\infty$. Of course, the hypotheses of the Stoltz-Cesaro theorem are not satisfied in this example.

Moral of the Story Don't try to use a theorem if its hypotheses are not satisfied.

(3) The preceding discussion may bring back memories of elementary calculus; specifically, of *L'Hôpital's Rule*. (If you don't recall what that refers to, don't worry; it is covered in Chapter (V).) Indeed, the Stoltz-Cesaro theorem is often called **L'Hôpital's Rule for sequences of real numbers**. .

The following simple consequences of the Stoltz-Cesaro theorem are important in their own right. The first was stated explicitly, and proved, by Cauchy in his *Cours d'analyse* of 1821. In light of the simple relation between them, it seems likely that Cauchy was also aware of the truth of the second.

III.3.11 Corollary

(a) Let $\alpha = (a_1, a_2, \ldots a_k, \ldots)$ be a real sequence such that $\lim_{k\to\infty}(a_{k+1} - a_k) = L$ for some extended real number L. Then $\lim\limits_{k\to\infty} \dfrac{a_k}{k} = L$.

(b) Let $\xi = (x_1, x_2, \ldots x_k, \ldots)$ be a real sequence such that $\lim_{k\to\infty} x_k = L$ for some extended real number L. Let $\mu = (m_1, m_2, \ldots m_k, \ldots)$ be the related sequence given by

$$m_1 = x_1,\ m_2 = \frac{x_1 + m_2}{2}, \ldots m_k = \frac{m_1 + x_2 + \ldots + x_k}{k}, \ldots;$$

that is, m_k is the arithmetic mean (i.e., average) of the first k terms of the original sequence ξ. Then $\lim_{k\to\infty} m_k = L$.

Proof (a) This corresponds to the special case of the Stoltz-Cesaro theorem in which one sets $b_k = k$ for each index k.

(b) Define a sequence $\alpha = (a_1, a_2, \dots a_k, \dots)$ by the rule $a_k = x_1 + x_2 + \dots + x_k$ for each k. Note that $a_k/k = m_k$, and $a_{k+1} - a_k = x_{k+1}$. The hypothesis that ξ has limit L implies that $\lim_{k\to\infty} x_{k+1} = L$; that is, $\lim_{k\to\infty}(a_{k+1} - a_k) = L$. It follows from Part (a) of this Corollary, and the fact that $a_k/k = m_k$ for each k, that $\lim_{k\to\infty} m_k = L$.

Remarks (1) The numbers m_k defined above are called the **Cesaro means**, or, less commonly the **Cesaro averages**, associated with the sequence ζ, in honor of the Italian mathematician Ernesto Cesaro (c. 1888).

(2) The converses of the statements proved in the preceding corollary are not true. For example, it is an easy exercise to construct a sequence $\xi = (x_1, x_2, \dots x_k, \dots)$ which does not have a limit but whose corresponding sequence of Cesaro means does.

III.4 The Bolzano-Weierstrass Theorem

One of the most commonly used techniques in analysis the process of seeking a subsequence, of a given real sequence ξ, which has a limit even though ξ itself need not have a limit. The next result gives the theoretical underpinnings for that method. It is one of the most important theorems in *This Textbook*.

III.4.1 Theorem (The Bolzano-Weierstrass Theorem for Sequences of Real Numbers - Standard Form)

Every bounded sequence of real numbers has a convergent subsequence.

Standard Proof Since, by hypothesis, the sequence ξ is bounded, there exist real numbers a and b, with $a < b$, such that $a \leq x_k \leq b$ for every index k.

Now construct a bisection sequence $[a_1, b_1]$, $[a_2, b_2], \dots$ (see Definition (II.4.19)) as follows:

(i) Initial Step Let $[a_1, b_1]$ be any closed interval such that $x_k \in [a_1, b_1]$ for infinitely many values of the index k; for instance, $a_1 = a$ and $b_1 = b$ will do.

(ii) Inductive Step Suppose that for some m in $\mathbb{N}$ the interval $[a_m, b_m]$ has been constructed so that $x_k \in [a_m, b_m]$ for infinitely many indices k. It is clear that at least one of the two halves of the interval $[a_m, b_m]$ has the analogous property; namely, that there are infinitely many indices k such that x_k lies in that half of $[a_m, b_m]$. If the left-half interval has that property, define $[a_{m+1}, b_{m+1}]$ to be that left-half interval. If the left-half interval of $[a_m, b_m]$ does *not* have that property, then the right-half interval does; in that case, define $[a_{m+1}, b_{m+1}]$ to be that right-half interval. Continuing this way, one gets a bisection sequence $[a_1, b_1]$, $[a_2, b_2], \dots [a_m, b_m], \dots$ with the property that for each m there are infinitely many indices k such that $x_k \in [a_m, b_m]$.

For each m let A_m denote the set of indices k such that $x_k \in [a_m, b_m]$. It is clear that $\mathcal{A} = (A_1, A_2, \dots A_m, \dots)$ is a subsequence structure of infinite order; see Definition (I.9.8). Let $B = \{k_1 < k_2 < \dots < k_m < \dots\}$ be a cross section of $\mathcal{A}$; see Definition (I.9.12). Then the subsequence $\zeta = (x_{k_1}, x_{k_2}, \dots x_{k_m}, \dots)$ of ξ associated with B has the property that $k_m \in A_m$ and thus $x_{k_k} \in [a_m, b_m]$. The Bisection Principle implies that there is exactly one real number c which lies in each interval $[a_m, b_m]$. By Remark (III.2.6) (1), together with the Squeeze Property for sequences, it follows that the subsequence ζ converges to c.

There is a straight-forward extension of the Bolzano-Weierstrass Theorem which drops the requirement that the sequence ξ be bounded.

III.4.2 Theorem (Extended Bolzano-Weierstrass Theorem for Real Sequences – Extended Form)

Suppose that $\xi = (x_1, x_2, \dots)$ is a real sequence. Then there is a subsequence of ξ which has a limit (in the sense of Definition (III.1.6) above). More precisely:

(i) If ξ is unbounded above, then ξ has a strictly increasing subsequence whose limit is $+\infty$.
(ii) If ξ is unbounded below, then ξ has a strictly decreasing subsequence whose limit is $-\infty$.
(iii) If ξ is bounded above and below (i.e., bounded), then ξ has a convergent subsequence.

Proof

(i) Define an infinite-order subsequence structure $\mathcal{A} = (A_1, A_2, \dots A_m, \dots)$ as follows:

A_1 is the set of all indices k such that $x_k > 1$. Let n_1 be the smallest element of the set A_1. Since, by hypothesis, the sequence ξ is unbounded above, it is clear that A_1 is an infinite subset of $\mathbb{N}$.

A_2 is the set of all indices k such that $x_k > \max\{x_{n_1}, 2\}$. Let n_2 be the smallest element of A_2. It is clear that A_2 is an infinite subset of A_1, and that $n_2 > n_1$.

Continuing this way, suppose that A_j and n_j have are already defined for all $j = 1, \dots m$. Then A_{m+1} is the set of all indices k such that $x_k > \max\{x_m, m+1\}$, and n_{m+1} is the smallest element of A_{m+1}. It is clear that $n_1 < n_2 < \dots < n_m <$.

Let $B = \{n_1 < n_2 < \dots < n_m < \dots\}$; note that B is the minimal cross section of subsequence structure $\mathcal{A}$. It is clear that the subsequence $\xi_B = (x_{n_1}, x_{n_2}, \dots x_{n_m}, \dots)$ of ξ associated with the set B is strictly increasing and has limit $+\infty$, as required.

(ii) The proof in this case is similar to that in (i); the details are left as an exercise.

(iii) This case is simply a restatement of the standard Bolzano-Weierstrass Theorem (see Theorem (III.4.1)), which has already been proved.

III.4.3 Definition

Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be a sequence of real numbers. An extended real number which can be expressed as the limit of a subsequence of ξ is called a **subsequential limit of $\boldsymbol{\xi}$**. The set of all subsequential limits of a given sequence ξ is denoted by $\mathcal{L}[\xi]$.

Note that the extended Bolzano-Weierstrass Theorem can now be phrased to say that for every real sequence ξ the set $\mathcal{L}[\xi]$ is nonempty.

There is another version of the Bolzano-Weierstrass Theorem which is closer to Weierstrass' original formulation. It helps to first introduce some terminology.

III.4.4 Definition

Let X be a set of real numbers with infinitely many elements. A real number c is said to be an **accumulation point of $\boldsymbol{X}$**, and the set X is said to **accumulate at $\boldsymbol{c}$**, provided that for every $\varepsilon > 0$ there are infinitely many elements x in X such that $|x - c| < \varepsilon$.

Examples (1) Let $X = \{1, 1/2, 1/3, \dots 1/n, \dots\}$ be the set of reciprocals of natural numbers. It is easy to see that this set accumulates at exactly one real number, namely at $c = 0$.

(2) Let X' be the set from the preceding example, together with the number 0; thus, $X' = \{0, 1, 1/2, \dots 1/n \dots\}$. Once again, this set accumulates at exactly one real number, namely 0.

(3) Every rational number is an accumulation point of the set of all irrational numbers.

(4) Let $X = \mathbb{N}$. Then this set accumulates at no number. (Note that the definition does not allow $+\infty$ or $-\infty$ as accumulations points of a set.)

Remark Many texts use the terminology 'limit point of X' or 'cluster point of X' instead of 'accumulation point'.

(2) Examples (1) and (2) above illustrate the fact an accumulation point of a set is allowed, but not required, to be an element of that set.

(3) The preceding remark illustrates a subtle ambiguity of language. The phrase 'point of X' suggests – for obvious reasons – that one is speaking about an element of the set X. Adding the modifier 'accumulation' to that phrase suggests that one is speaking about an element of X which has a special property. The same type of ambiguity arises when, for example, states that a number M is the supremum of a set X: the phrasing suggests that this statement entails that M must be an element of X, when in fact it need not. Since it is unlikely that mathematicians will ever completely avoid such 'abuses of language', the only solution is to be aware of the problem and to read definitions carefully.

III.4.5 Theorem (Bolzano-Weierstrass Theorem for Sets of Real Numbers)

Let X be a bounded infinite subset of $\mathbb{R}$. Then X has at least one accumulation point.

Proof Let S be the set of real numbers y such that $y \leq x$ for all but a finite number of elements of the set X. Since, by hypothesis, the set X is bounded, hence bounded below, it is clear that $S \neq \emptyset$, since it obviously contains every lower bound of X. Also it is clear that S is a convex set which is unbounded below. Finally, the fact that X is bounded above and an infinite subset of $\mathbb{R}$ implies that $S \neq \mathbb{R}$, since every upper bound of X fails to satisfy the defining property of S. In summary, the set S satisfies the hypothses of the Bolzano's Right-Endpoint Principle. It follows that there exists a real number B such that either $S = (-\infty, B)$ or $S = (-\infty, B]$. It is easy to see that B is an accumulation point of the original set X.

Remarks (1) There is an obvious extension of the preceding theorem to unbounded infinite subsets of $\mathbb{R}$ if one allows $\pm\infty$ as possible accumulation points, but it hardly seems worth the effort.

(2) It would have been faster to prove this result as a simple corollary of the sequential version of the Bolzano-Weierstrass Theorem, but it seemed appropriate to use Bolzano's Principle on a theorem that includes his name. One can easily modify this argument to prove the 'sequential' version of the Bolzano-Weierstrass theorem using Bolzano's Principle.

(3) The statement of the Bolzano-Weierstrass Theorem for bounded infinite subsets makes sense in any ordered field. It is easy to show that in that context it is equivalent to all the other versions of 'Completeness' studied in Chapter (II), and thus could have been used as the 'Completeness Axiom'. One advantage in doing that would be that its statement probably requires the least preparation of all the candidates suggested for that axiom in Chapter (II). The biggest disadvantage to using it is

that it is even less 'obviously true', i.e., 'axiomatic', than, say, the Least-Upper-Bound Principle.

Parts (i) and (ii) of Theorem (III.4.2), the Extended Bolzano-Weierstrass Theorem for Sequences, suggest a natural question: Does every *bounded* sequence of real numbers have at least one *strictly* monotonic subsequence? A little thought makes it clear that the answer is 'No'; for example, a constant real sequence has no strictly monotonic subsequence. However, since such a sequence is monotonic, one is led to a modified question: does every bounded sequence have a *monotonic* subsequence? The answer to that question is provided by the next result.

III.4.6 Theorem (The Monotonic-Subsequences Theorem)

Every sequence of real numbers has a monotonic subsequence.

More precisely, if sequence does not have a constant subsequence, then it has a strictly monotonic subsequence.

Proof Parts (i) and (ii) of the preceding theorem verify the truth of the statement in the case of unbounded sequences. Indeed, the monotonic subsequences in that case can be chosen to be *strictly* monotonic.

Thus, let us now assume that $\xi = (x_1, x_2, \dots x_k, \dots)$ is a bounded real sequence. If ξ has a constant subsequence, that subsequence is automatically monotonic, so the result follows. Thus, assume for the rest of this proof that ξ does not have a constant subsequence. Equivalently, assume that there is no real number c such that $x_k = c$ for infinitely many indices k.

Special Case Assume that ξ is convergent, with limit L in $\mathbb{R}$. Let B be the set of indices k such that $x_k > L$, and let C be the set of indices k such that $x_k < L$. By the assumption that ξ does not have a constant subsequence, at least one of the sets B or C must be an infinite subset of $\mathbb{N}$. To be definite, assume that $B = \{k_1 < k_2 < \dots < k_m < \dots\}$ is infinite. Let $\zeta = \{z_1, z_2, \dots z_m, \dots\}$ be given by the formula $z_m = 1/(x_{k_m} - L)$, so each term z_m is positive. The hypothesis that $L = \lim_{k \to \infty} x_k = L$ implies that $\lim_{m \to \infty} z_m = +\infty$. In particular, Part (i) of the preceding theorem implies that ζ has a subsequence $\tau = (z_{k_{m_1}}, z_{k_{m_2}}, \dots z_{k_{m_n}} \dots)$ which is strictly increasing. It follows easily, from the usual order properties of $\mathbb{R}$, that the corrresponding subsequence $(x_{k_{m_1}}, x_{k_{m_2}}, \dots x_{k_{m_n}}, \dots)$ of ξ is strictly decreasing; in particular, it is strictly monotonic, as required.

Suppose, instead, that the set $C = \{j_1 < j_2 < \dots < j_n < \dots\}$ is infinite. Let $\sigma = (s_1, s_2, \dots s_k, \dots)$, be the sequence given by $s_k = -x_k$ for each index k; that is, $\sigma = -\xi$. Then σ converges to $L' = -L$, and C is the set of all indices for which $s_k > L'$. It follows from what was just proved, but now applied to the sequence σ and limit L', that σ has a strictly decreasing subsequence. Since $x_k = -s_k$, it follows that ξ has a subsequence which strictly increasing, hence strictly monotonic, as required.

General Case If ξ is bounded then, by the Bolzano-Weierstrass Theorem, it has a subsequence η which is convergent. By applying the results of the 'Special Case' above to η, one sees that η has a subsequence σ which is strictly monotonic. But σ is then a *sub*subsequence of ξ, hence a subsequence of ξ, which is strictly monotonic, as required.

Remark There are proofs of the Monotonic-Subsequences Theorem which are shorter and, in a sense, more elegant. An advantage of the proof given here is that it starts by proving a special case in which the claimed result is intuitively obvious, namely the case ξ is unbounded above. It then reduces all the remaining cases to this one, using the Bolzano-Weierstrass Theorem as the main tool. The technique of first proving the desired result in a simple special case, then reducing the other cases to this special case, is a powerful tool in mathematics.

The next results provide useful tools for showing whether given sequence has a limit.

III.4.7 Theorem

Let $\xi = (x_1, x_2, \dots)$ be a sequence of real numbers, and let L be an extended real number.

(a) A necessary and sufficient condition for ξ to not have L as a limit is that there exist an extended real number $L' \neq L$ and a subsequence $\tau = (t_1, t_2, \dots t_n, \dots)$ of ξ such that $\lim_{n\to\infty} t_n = L'$.

(b) A necessary and sufficient condition for ξ to have a limit is that all subsequences of ξ which has the same limit.

Proof

(a) The Condition is Necessary Suppose that ξ does not have limit L. Then either Statement A or Statement B of Theorem (III.1.9) must fail to hold. If Statement A fails to hold, then there exists a real number $y < L$ such that $x_k \leq y$ for infinitely many indices k. Let A be the set of such indices, so that A is an infinite subset of $\mathbb{N}$, and let $\zeta = (z_1, z_2, \dots z_k, \dots)$ be the corresponding subsequence ξ_A of ξ. By the Extended Bolzano-Weierstrass theorem, the sequence ζ has a subsequence $\tau = (t_1, t_2, \dots t_n, \dots)$ which has a limit; call this limit L'. Since $z_m \leq y$ for all m, it follows that $t_n \leq y$ for all n, and thus $L' \leq y < L$. Since τ is also a subsequence of the original sequence ξ, it follows from Part (c) of Theorem (III.2.1) that ξ does not have L as limit. A similar argument works if, instead, Statement B fails to hold.

The Condition is Sufficient Indeed, suppose that there is a subsequence τ of ξ which has limit $L' \neq L$. Then, by Part (a) of Theorem (III.2.1), the subsequence τ cannot have limit L, hence by Part (c) of the same theorem the original sequence ξ also cannot have limit L.

(b) The Condition is Necessary This is simply the statement of Part (c) of Theorem (III.2.1).

The Condition is Sufficient Suppose that every subsequence of ξ which has a limit has the same limit; call it L. Then, by the 'Necessary' portion of Part (a), it is *not* the case that ξ does *not* converge to L. In other words, ξ *does* converge to L.

Remark The converse of Part (c) of Theorem (III.2.1) can be expressed as follows:

If every subsequence of ξ has limit L, then ξ has limit L.

This coverse is, in fact, correct, but for trivial reasons: if *every* subsequence of ξ has limit L, then ξ, being itself a subsequence of ξ, has limit L.

In contrast, the hypotheses of Part (b) of Theorem (III.4.7) involve only a special type of subsequences of ξ. The proof of Part (b), far from being trivial, involves the results of Part (a), whose proof in turn uses the Extended Bolzano-Weierstrass theorem.

III.5 The Cauchy Criterion for Convergent Real Sequences

Introduction

The definition of 'limit of a real sequence' seems to require that one have a candidate L in mind for the purported limit. Indeed, most of the results and examples presented so far involve, in perhaps a hidden way, a plausible candidate for L. For example, in the proof of the Monotonic-sequences Principle, the candidate is the supremum of the term-set (when the sequence is monotonic up). Likewise, in the proof of the (standard) Bolzano-Weierstrass Theorem the candidate is the unique point c determined by the bisection sequence. (The main exceptions are those limits which are computed using algebra to reduce a complicated limit problem to a finite number of simpler problems whose answers are already known.)

The present section is devoted to a criterion for convergence which does not require a 'candidate' in advance. Note, however, that it is a criterion for *convergence*, not for 'existence of a limit', since it applies only to bounded sequences.

Preliminary Observation Anyone who deals with the convergence of sequences of real numbers is familiar with the following fact:

If a sequence $\xi = (x_1, x_2 \dots)$ of real numbers is convergent, then the distance between consecutive terms approaches 0; that is,

$$\lim_{k\to\infty} (x_{k+1} - x_k) = 0. \tag{III.6}$$

Actually, the following – apparently stronger – result holds.

III.5.1 Theorem

Suppose that $\xi = (x_1, x_2, \dots x_k \dots)$ is a sequence of real numbers. Let m be a natural number, and consider the sequence $\delta_{(\xi;m)} = (d_1, d_2, \dots d_k, \dots)$ defined by the rule

$$d_k = x_{m+k} - x_k \text{ for each } k \text{ in } \mathbb{N}.$$

That is, $\delta_{(\xi;m)} = (x_{m+1} - x_m,\ x_{m+2} - x_m,\ x_{m+3} - x_m,\ \dots)$.

Conclusion A necessary condition for the sequence ξ to be convergent is that for each m the associated sequence $\delta_{(\xi;m)}$ converges to 0; that is, for each m in $\mathbb{N}$ one has

$$\lim_{k\to\infty} (x_{m+k} - x_k) = 0 \tag{III.7}$$

Proof Suppose that ξ is convergent, and let $L = \lim_{k\to\infty} x_k$. Let $\zeta = (z_1, z_2, \dots z_k, \dots)$ be the sequence given by the rule $z_k = x_{m+k}$ for each k in $\mathbb{N}$. Note that ζ is a subsequence of ξ; thus, by Part (c) of Theorem (III.2.1), the sequence ζ also converges to L; that is, $\lim_{k\to\infty} x_{m+k} = L$. Note also that $d_k = z_k - x_k$ for each k in $\mathbb{N}$, so by Theorem (III.3.2) it follows that the sequence $\delta_{(\xi;m)}$ is also convergent, and

$$\lim_{k\to\infty} d_k = \left(\lim_{k\to\infty} x_{m+k}\right) - \left(\lim_{k\to\infty} x_k\right) = L - L = 0,$$

as required.

III.5.2 Remarks

(1) The phrase 'apparently stronger' is added before the word 'result' above. It is true that Theorem (III.5.1) does have a stronger conclusion than Equation (III.6) from the same hypothesis; indeed, that equation is the special case $m = 1$ of Condition (III.7). However, it is a simple exercise to show that if Equation (III.6) holds, then so does Condition (III.7) for *every m*.

(2) It is useful to reformulate Condition (III.7) above as follows, so the definition of 'convergence' is invoked directly:

$$\text{For each } \varepsilon > 0 \text{ and } m \text{ in } \mathbb{N}, \text{ there is } N \text{ in } \mathbb{N} \text{ such that if } k \geq N, \text{ then } |x_{k+m} - x_k| < \varepsilon \tag{III.8}$$

The preceding theorem provides a *necessary* condition for a sequence to be convergent. The next example, when combined with Remark (1) above, shows that, unfortunately, this condition is not *sufficient* for convergence. ('Unfortunately': many students lose points on calculus exams because they believe that the condition *is* sufficient to prove convergence.)

III.5.3 Example

Let $x_n = \sqrt{n}$. (The existence of square roots is proved in Theorem (III.2.13).) It is clear that this sequence is unbounded, and thus not convergent. Nevertheless, one has

$$x_{k+1} - x_k = \sqrt{k+1} - \sqrt{k} = \frac{(\sqrt{k+1} - \sqrt{k})\,(\sqrt{k+1} + \sqrt{k})}{\sqrt{k+1} + \sqrt{k}} = \frac{(\sqrt{k+1})^2 - (\sqrt{k})^2}{\sqrt{k+1} + \sqrt{k}} =$$

$$\frac{(k+1) - k}{\sqrt{k+1}\sqrt{k}} = \frac{1}{\sqrt{k+1} + \sqrt{k}}$$

It is then clear that $\lim_{k\to\infty}(x_{k+1} - x_k) = 0$.

There is another condition, quite similar to the alternate formulation given in Condition (III.8) above, which turns out to be not just necessary, but also sufficient, for a sequence of real numbers to be convergent. It is traditionally called the 'Cauchy Criterion for the convergence of a sequence of reals', although others stated versions of it earlier; for example, Bolzano published it in 1817.

III.5.4 Definition (The Cauchy Criterion and Cauchy Sequences in $\mathbb{R}$)

A sequence $\xi = (x_1, x_2, \dots)$ of real numbers is said to satisfy the **Cauchy criterion** provided

$$\text{For every } \varepsilon > 0, \text{ there exists a number } N \text{ in } \mathbb{N} \text{ such that if } k \geq N, \text{ then } |x_{k+m} - x_k| < \varepsilon \text{ for every } m \text{ in } \mathbb{N} \tag{III.9}$$

If this occurs then one says that ξ is a **Cauchy sequence (of real numbers)**.

Remarks (1) Compare Condition (III.9) given here with Condition (III.8) above. The only difference is the location of the phrase 'for every m in $\mathbb{N}$'. Nevertheless, this apparent 'small' change of phrasing has a major impact on the meaning of the conditions. Indeed, Condition (III.8) says that if both ε and m are chosen in advance, then one can chose a value of N, which depends on both ε and m, for which the given inequality holds. In contrast, Condition (III.9) requires that N can be chosen in terms of ε, but its choice should not depend on m.

(2) The Cauchy criterion is often phrased in the following alternate, but equivalent, form:

$$\text{For every } \varepsilon > 0 \text{ there exists a natural number } N \text{ such that if } j, l \geq N \text{ then } |x_j - x_l| < \varepsilon \tag{III.10}$$

The equivalence of this form with that given in the definition is easy to verify.

The importance of the concept of 'Cauchy sequence' in analysis is explained by the following result.

III.5.5 Theorem (The Cauchy Convergence Theorem for Real Sequences)

Suppose that $\xi = (x_1, x_2, \ldots)$ is a sequence of real numbers. Then a necesary and sufficient condition for ξ to be convergent is that ξ be a Cauchy sequence.

Proof:

(a) ('Necessity') Suppose that ξ converges to some number L. Let $\varepsilon > 0$ be given, and choose N in $\mathbb{N}$ large enough that if j satisfies $j \geq N$ then $|x_j - L| < \varepsilon/2$. In particular, suppose that $k \geq N$ and $m \in \mathbb{N}$, so that $m + k \geq N$ as well. Then one has

$$|x_{m+k} - L| < \frac{\varepsilon}{2} \text{ and } |x_k - L| < \frac{\varepsilon}{2}.$$

Combine this with the Triangle Inequality to get

$$|x_{m+k} - x_k| \leq |x_{m+k} - L| + |L - x_k| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon;$$

That is, ξ satisfies the Cauchy criterion, as required.

(b) ('Sufficiency') First note that ξ satisfies the Cauchy criterion, it is a bounded sequence. More precisely, let $\varepsilon_0 = 1$. (In reality, one can choose ε_0 to be any positive number; it is mere convention to choose it to be the best-known positive number, namely 1.) Choose N in $\mathbb{N}$ large enough that if $k \geq N$ then $|x_{m+k} - x_k| < 1$ for all m in $\mathbb{N}$. In particular, it follows that if $n \geq N$ then $|x_n - x_N| < 1$. Indeed, if $n = N$ the inequality is trivially true, while if $n > N$ it follows from the definition of N, using $k = N$ and $m = n - N$. Now use the Triangle Inequality to get

$$|x_n| = |(x_n - x_N) + x_N| \leq |x_n - x_n| + |x_N| < 1 + |x_N| \text{ for all } n \geq N$$

It follows from Part (c) of Theorem (II.3.16), 'A sequence which is eventually bounded is bounded', that the given sequence is bounded, as claimed.

It follows from the Bolzano-Weierstrass Theorem that the given sequence ξ has a convergent subsequence. Let $A = \{k_1 < k_2 < \ldots < k_n < \ldots\}$ be an infinite subset of $\mathbb{N}$ such that the subsequence $\zeta = (z_1, z_2, \ldots z_n, \ldots) = \xi_A$ is convergent, and let L the corresponding limit. In light of Part (c) of Theorem (III.2.1), it is clear that L is the only reasonable candidate to be the limit of the full sequence ξ. To see that the original sequence ξ does, in fact, converge to L, first note that, by the Triangle Inequality, for all k and m one has

$$|L - x_k| \leq |L - x_{m+k}| + |x_{m+k} - x_k| \qquad (*)$$

Now suppose that $\varepsilon > 0$ is given. Let N in $\mathbb{N}$ be large enough that if $k \geq N$, then $|x_{m+k} - x_k| < \varepsilon/2$. If $k \geq N$ choose m so that $m + k$ is of the form k_n for some index n, so that x_{m+k} is a term of the subsequence ζ, and so that m is large enough to guarantee that $|L - x_{m+k}| < \varepsilon/2$. (The existence of such m follows from the convergence of the subsequence ζ to L.) For such a choice of m, Inequality $(*)$ implies that for every $k \geq N$ one has

$$|L - x_k| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

The desired result follows.

Remark The theoretical importance of the Cauchy Criterion is that it applies to all sequences, and its hypotheses can be checked knowing only the terms of the sequence, without needing a candidate for the limit.

The next result illustrates a type of situation in which the Cauchy Criterion can be used.

III.5.6 Theorem

Suppose that $\xi = (x_1, x_2, \ldots)$ is a sequence of real numbers with the property that there exists a real number λ, with $0 \leq \lambda < 1$, such that

$$|x_{k+2} - x_{k+1}| \leq \lambda |x_{k+1} - x_k| \text{ for all sufficiently large } k \text{ in } \mathbb{N}. \tag{III.11}$$

Then the sequence ξ is a Cauchy sequence (and thus is convergent).

Proof: Case 1: Assume that Inequality (III.11) holds for *all* k in $\mathbb{N}$.

First note that the result is obvious if $\lambda = 0$ or if $x_2 = x_1$. Indeed, if $\lambda = 0$ then one clearly has $|x_{k+2} - x_{k+1}| = 0 \cdot |x_{k+1} - x_k|$ for all k in $\mathbb{N}$, so that $x_2 = x_3 = x_m = \ldots$ for all $m \geq 2$. Likewise, if $x_2 = x_1$ then it is equally clear that ξ is a constant sequence.

Thus, assume that $x_2 \neq x_1$ and that λ satisfies $0 < \lambda < 1$. It is easy to see, by repeatedly using the given hypothesis on ξ, that

$$|x_{k+1} - x_k| \leq \lambda^{k-1} |x_2 - x_1| \text{ for each } k \text{ in } \mathbb{N}$$

Indeed, let A be the set of all natural numbers k for which this inequality is valid. It is clear that $1 \in A$, since for $k = 1$ the condition to be verified becomes $|x_2 - x_1| \leq \lambda^0 |x_2 - x_1|$; and since $\lambda^0 = 1$ if $\lambda \neq 0$, this in turn reduces to $|x_2 - x_1| \leq 1 \cdot |x_2 - x_1|$, which is certainly true. Next, suppose that $k \in A$. Then

$$|x_{k+2} - x_{k+1}| \leq \cdot |x_{k+1} - x_k| \leq \lambda \cdot \left(\lambda^{k-1} |x_2 - x_1|\right) = \lambda^k |x_2 - x_1|.$$

Thus $(k+1)$ is also in A. Now the Principle of Mathematical Induction implies that $A = \mathbb{N}$, and the desired inequality follows.

To show that the sequence ξ is Cauchy, note that if k and m are in $\mathbb{N}$, then use the preceding inequality, together with the classic 'Add-and-Subtract Trick' (see Theorem (IX.1.8)) to get

$$|x_{k+m} - x_k| = |(x_{k+m} - x_{k+m-1}) + (x_{k+m-1} - x_{k+m-2}) + \ldots + (x_{k+1} - x_k)| \leq$$

$$|x_{k+m} - x_{k+m-1}| + \ldots + |x_{k+1} - x_k| \leq \left(\lambda^{k+m-2} + \lambda^{k+m-3} + \ldots + \lambda^{k-1}\right) |x_2 - x_1|.$$

By the results of Proposition (II.2.16) one can write

$$\lambda^{k+m-2} + \lambda^{k+m-3} + \ldots + \lambda^{k-1} = \lambda^{k-1}(\lambda^{m-1} + \lambda^{m-1} + \ldots + 1) = \lambda^{k-1}\left(\frac{\lambda^m}{1-\lambda}\right) = \frac{\lambda^{k+m-1}}{1-\lambda}.$$

Thus

$$|x_{k+m} - x_k| \leq \left(\frac{\lambda^{k+m-1}}{1-\lambda}\right) |x_2 - x_1| \leq \left(\frac{\lambda^k}{1-\lambda}\right) |x_2 - x_1|$$

It follows from Corollary (III.2.10) that for every $\varepsilon > 0$ there exists B such that if $k \geq B$ then $\lambda^k < \varepsilon \left(\frac{1-\lambda}{|x_2 - x_1|}\right)$. The fact that ξ is a Cauchy sequence now follows.

Case 2: Now consider the general case. That is, assume that there exists a natural number m such that Inequality (III.11) holds for all $k \geq m$. Let $\tau = (t_1, t_2, \ldots)$ be the sequence $(x_m, x_{m+1}, \ldots)$; that is $t_j = x_{m+j-1}$ for each j in $\mathbb{N}$. It is clear that the sequence τ satisfies the hypotheses for Case 1, and thus τ is a Cauchy sequence.

III.5.7 Example

Recall Heron's 'Divide-and-Average' method for computing square roots; see Theorem (III.2.13): If C and x_1 are positive real numbers, then Heron's method produces an infinite sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ such that for each index k one has $x_{k+1} = \frac{1}{2}\left(x_k + \frac{C}{x_k}\right)$. We have already seen that $x_k > 0$ for each k. One can use the preceding theorem to give an alternate proof that the sequence ξ is convergent and its limit is $\sqrt{C}$.

More precisely, note that

$$x_{k+2} - x_{k+1} = \frac{1}{2}\left(x_{k+1} + \frac{C}{x_{k+1}}\right) - \frac{1}{2}\left(x_k + \frac{C}{x_k}\right) =$$

$$\frac{1}{2}\left(x_{k+1} - x_k + C\left(\frac{1}{x_{k+1}} - \frac{1}{x_k}\right)\right) = \frac{1}{2}(x_{k+1} - x_k)\left(1 - \frac{C}{x_{k+1}\,x_k}\right)$$

From the definition of x_{k+1} in terms of x_k one computes that $2\,x_{k+1}\,x_k = x_k^2 + C > C$, so that $2 > C/(x_{k+1}\,x_k) > 0$. From this one gets $|1 - C/(x_{k+1}\,x_k)| < 1$ and thus

$$|x_{k+2} - x_{k+1}| < \frac{1}{2}\,|x_{k+1} - x_k|$$

Thus, the hypotheses of the previous theorem are satisfied, with $\lambda = 1/2$, so the sequence ξ is convergent. Let L be its limit. Since each $x_k > 0$, it follows that $L \geq 0$. If it were the case that $L = 0$, it would follow that $x_k + A/x_k$ would diverge to $+\infty$, which is impossible since ξ is convergent. Thus $L > 0$, so since $\lim_{k\to\infty} x_k = \lim_{k\to\infty} x_{k+1} = L$, it follows that $L = \frac{1}{2}\left(L + \frac{C}{L}\right)$, hence $2\,L^2 = L^2 + C$; that is, $L^2 = C$. In other words, L is the positive square root of C, as expected.

Side Comment (on what Cauchy and Bolzano knew about sequences)

Because of the length of this Side Comment, the reader should recall that nothing in any Side Comment is needed for the understanding of the main body or the appendices of *This Textbook*. In particular, this Side Comment is of a speculative nature; for example, it attempts to read the minds of Cauchy and Bolzano, both of whom are long dead. Thus, even more than is always the case, feel free to ignore it.

Most of the remarks below on what Cauchy and Bolzano may have been thinking are based on the following sources:

(1) Cauchy's *Cours d'analyse* of 1821.

(2) Bolzano's *Rein analytischer Beweis des Lehrsatzes . . .* of 1817.

On What Cauchy Knew With few exceptions, such as Part (a) of Corollary (III.3.11) in *This Textbook*, Cauchy treats the basic theory of limits of real sequences as if it was all 'well known' to his readers; so well known, in fact, that he does not provide explicit statements, much less proofs, of many of the important results such as the Sum and Product Rules for convergent sequences, as well as what we now call the (extended) 'Bolzano-Weierstrass Theorem for Real Sequences'. As for the theorem that a real sequence is convergent if, and only if, it is what we call a 'Cauchy sequence', Cauchy does at least clearly define that concept and state the theorem; but he gives no clue about how to prove it.

How could Cauchy think of such facts as being obvious, when modern treatments of analysis, such as *This Textbook*, work hard to state them explicitly and to prove them rigorously?

In the case of the Product Rule for convergent sequences, for instance, it can be argued that Cauchy is correct: it *is* 'obvious' that the rule is valid. Indeed, a common statement of that rule is that if a_k is close to A and b_k is close to B, then the product $a_k\,b_k$ is close to the the product $A\,B$. At the present time, as in Cauchy's time, the truth of that statement is accepted without thinking. For instance, if one uses a calculator to compute $\sqrt{2}$ times $\sqrt{3}$, one knows that the calculator does not use the exact values of $\sqrt{2}$ and $\sqrt{3}$; it must round off after a some number of decimal places. Nevertheless, one accepts the calculator's answer as being very close to the true value $\sqrt{6}$.

What about the Bolzano-Weierstrass Theorem? In Chapter VI of his Cours d'analyse, Cauchy considers sequences of the form $\xi = (x_1, x_2, \ldots x_k, \ldots)$, where $x_k = \sqrt[k]{|A_k|}$ for some quantities A_k. He then considers explicitly the set of all subsequential limits of such a sequence (see Definition (III.4.3)), and he makes important applications of them. (We consider these applications later in *This Textbook*.) Cauchy's applications make no sense unless every sequence of reals has a subsequence with a limit. But how could he think this fact is 'obvious'? What follows is a speculative attempt to read Cauchy's mind.

First of all, every mathematician of the time would have believed the truth of the extended Bolzano-Weierstrass Theorem for *unbounded* sequences, since all it says is that an unbounded sequence has a subsequence which diverges to one of the infinities, a fact that really *is* obvious. Thus, the issue reduces to the case of bounded sequences; that is, to the standard formulation of the Bolzano-Weierstrass theorem. One approach is to give the standard proof found earlier in *This Textbook*, using a Bisection-Sequence argument. Such an argument appears to be well within Cauchy's abilities, since he uses similar arguments elsewhere in his *Cours d'analyse*. However, what follows is an elementary proof which does make the (standard) Bolzano-Weierstrass theorem seem 'obviously true'. Indeed, it requires only a basic understanding of how the standard decimal representation of real numbers works, together with the following:

Fact Since there are only ten decimal digits, in any infinite list of decimal digits at least one such digit must appear infinitely often.

An Elementary Proof of the Standard Bolzano-Weierstrass Theorem Recall that we are given a bounded sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ of real numbers. There is no loss in generality in assuming that for each index k one has $0 < x_k < 1$; see Corollary (III.3.4).

Choose a decimal representation for each of the numbers x_k, and fix it for the discussion. Define a real number c whose decimal representations $c = 0.d_1\,d_2\,\ldots d_m\,\ldots$, where the digits $d_1,\,d_2, \ldots\,d_m\ldots$ are obtained as follows:

d_1 is the smallest decimal digit which appears as the first digit of infinitely many of the terms of the sequence ξ. (That there exists such d_1 follows from the 'Fact' above.) The infinitely many terms of the sequence ξ with first digit equal to d_1 then form a subsequence of ξ.

d_2 is the smallest decimal digit which appears as the *second* decimal digit of infinitely many of the terms of the subsequence formed in the preceding step. This leads to a subsubsequence of ξ whose terms each have first digit d_1 and second digit d_2.

Continuing this way by induction, suppose that decimal digits $d_1,\,d_2, \ldots\,d_m$ have already been constructed, together with a subsequence of ξ whose terms each have first m digits equal to $d_1,\,d_2, \ldots\,d_m$, respectively. Let d_{m+1} be the smallest decimal digit which appears as the $m+1$-st digit of infinitely many terms of the preceding subsequence. Clearly the resulting number $c = 0.d_1\,d_2\ldots d_m\,\ldots$ is a subsequential limit of ξ.

Remark The restriction to sequences with values in the interval $(0,1)$ is made to simplify the discussion; Cauchy would have had no problem dealing with the general decimal representation instead. Also, the proof given here is purposely formulated in a somewhat informal manner, much as a mathematician might have done in Cauchy's era, using minimal notation. It can easily be made to look more modern by introducing, say, a subsequence structure $\mathcal{A} = (A_1, A_2, \ldots A_m, \ldots)$, where A_1 is the infinite set of indices k for which the first digit of x_k is d_1, A_2 is the infinite set of k in A_1 for which the second digit of x_k is d_2, and so on.

If one grants that Cauchy had Bolzano-Weierstrass Theorem available as a tool – and that this theorem is 'obviously' true – then the fact that Cauchy sequences must be convergent does become an 'obvious' fact; see the proof of Theorem (III.4.1) above.

Perhaps the key issue is not that Cauchy could prove the sufficiency of his criterion, but that he thought of stating it in the first place.

On What Bolzano Knew It may appear silly to ask whether Bolzano knew the Bolzano-Weierstrass theorem, in either the set-theoretic or the sequential form, since his name appears in the title of the theorem. Nevertheless, the question bears asking.

The reason is that the connection of Weierstrass' name to the theorem was common by the 1880s, but it appears that Bolzano's name was attached to the result only about 1899. The justification given by some historians for adding Bolzano's name to that of Weierstrass for this result is that Weierstrass' result is a simple consequence of a theorem from Bolzano's 1817, the one called the Right-Endpoint Principle in *This Textbook*. Indeed, the proof given above for the set-theoretic version of this theorem shows that this is true. However, there is a big difference between having the tools to prove a given result and actually knowing to even state it in the first place. So the question becomes: did Bolzano know to even state the result?

It turns out that Bolzano did publish the statement of what we now call the Bolzano-Weierstrass theorem (for sets, at least) in 1830. Unfortunately, this paper was not discovered until a century later. Bolzano refers to a proof of it in one of his other papers, but apparently no copy of that paper has been discovered. In any event, it does seem clear that Bolzano knew the result in question, so it is fair to attach his name to it. It is also clear that Weierstrass knew and proved the result, independently of Bolzano, and proved it without using Bolzano's Right-Endpoint Principle. Thus attaching both names to the result makes good sense. Whether Bolzano knew the result as early as his 1817 seems to be still an open question.

The question of what Bolzano knew about what we call the Cauchy Criterion for convergence of a sequence is much clearer: He not only stated it, in essentially the same way that Cauchy did, but (unlike Cauchy) Bolzano actually tried to prove it.

Unfortunately, Bolzano's 'proof' seems totally confused, and the universal opinion is that it proves nothing.

However, there may be a core of truth in Bolzano's argument. Indeed, the fact that the Cauchy Criterion is sufficient to prove convergence is, indeed, 'obviously true', at least in a sense that could convince calculus students:

Plausibility Argument Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be a Cauchy sequence of real numbers. Since ξ is bounded, without loss of generality assume that $x_k \in (0,1)$ for every index k. (This is not a serious restriction on ξ; see Corollary (III.3.4). It what follows, 'digit' refers to 'decimal digit to the right of the decimal point'.

Let m be a natural number, and let N in $\mathbb{N}$ be large enough that for every n in $\mathbb{N}$ one has $|x_{N+n} - x_N| < 1/10^{m+1}$. Since the first m digits of $1/10^{m+1}$ are 0, it is clear that the first m digits of x_N and x_{N+n} are the same. Otherwise stated, all terms of the sequence from the N-th term on have the same first m digits; i.e., the first m digits of the sought limit are known. Since m is arbitrary, one can determine *all* the digits $d_1, d_2, \dots d_m, \dots$ of the purported limit. That is, the number $c = 0.d_1\,d_2 \dots d_m \dots$ is the only reasonable candidate to be the desired limit, and it is easy to see that this number works.

Before reading further, contemplate whether you find the preceding argument to be valid.

The argument *almost* works; unfortunately, there are exceptional cases in which it breaks down. For example, consider the numbers $x_1 = 0.1\,0\,0 \dots 0 \dots$ and $x_2 = 0.0\,9\,9 \dots 9 \dots$. Since $x_1 = x_2$, it follows that $|x_1 - x_2| < 1/10^{m+1}$ for *every* m; yet clearly these decimal representations have no digits in common.

In a sense, Bolzano's 'proof' seems to be similar to the plausibility argument just presented above. Indeed, he 'approximates', by x_n for some large n, the 'true value', of the limit being sought, and more-or-less states that the ability to do this to an arbitrarily small error somehow provides the desired limit. It could be he had the 'decimal digits' plausibility above argument in mind, but realized that it had to be modified.

Actually, there is a simple fix to the plausibility argument which is still based on the idea of determining the decimal digits of the desired limit. For example, Bolzano might have carried out the argument used in the 'Elementary Proof of the Standard Bolzano-Weierstrass Theorem' above, but only in the context of the Cauchy Criterion, to get the same c. (It really is obvious that a Cauchy sequence which has a convergent subsequence is convergent.)In any event, Bolzano does go on to make the observation that one might be tempted to think that the limit he obtains must be irrational. (He does not make clear *why* one might be so tempted.) He counters with the further observation that the sequence 0.1, 0.1 1,0.1 1 1, . . . does converge, and the limit, 1/9, is rational.

General Comments Most historians claim that, regardless of any arguments, given or not given, neither Cauchy not Bolzano could have proved that the Cauchy Criterion implies convergence, because neither understood the structure of the real numbers well enough – especially 'completeness'. How compelling is this argument?

It is clear that both men understood that the real numbers form what is called (in Chapter (II)) an 'ordered field'.

As for 'completeness', Cauchy explicitly stated that every bounded monotonic sequence of reals has a limit, which is equivalent to 'Completeness'. Indeed, he uses this, together with what we now call the Archimedes Property, which of course he also believed to hold, to prove what we call the Bisection Principle. That is, the facts about real numbers which Cauchy accepted as true, and thus which could be included in a list of axioms of the reals, included all the axioms given in Chapter (II) of *This Textbook*.

The story for Bolzano's understanding of 'Completeness' is more complicated. Indeed, Bolzano explicitly stated what in *This Textbook* is called the 'Right-Endpoint Principle', which is also equivalent to 'Completeness'. Had he simply stated it, we could say that Bolzano also knew that the real numbers satisfy 'Completeness', and thus had enough 'axioms' to do analysis. However, it appears that he did not view this Principle as being obvious enough to be 'axiomatic', so he tried to reduce it to something simpler, namely the fact that the Cauchy Criterion implies converegence; and since this fact in turn is not so 'axiomatic', he tried to prove it in terms of somethimg simpler yet. As was mentioned above, this attempt was basically unsuccessful.

In other words, both mathematicians understood the usual axioms of the real number systems, which is generally all that modern texts in analysis ask their readers to accept.

What Cauchy and Bolzano did *not* do is to prove that their axiom systems for the real numbers could be represented by some 'model' constructed from a simpler system such as the rational numbers. However, it can be argued that they did not seek such a 'model' for their axioms because such a model had existed for a long time, and was already familiar to everyone who had learned the basics of arithmetic: namely, the real numbers, described analytically in terms of the standard decimal representation, and not geometrically as points on a line. The idea of a sequence converging to a number being the same as being able to determine the decimal digits of that number in a step-by-step manner would have been familiar: that is how one computes, say, $\sqrt{2}$ using Bisection or Heron methods. The question for both mathematicians was not whether the real numbers exist; they may have thought of that question as a nonissue. Instead, the goal was defining concepts such as continuity and infinite sums more precisely, and proving theorems more rigorously, using the accepted properties of the real numbers – accepted, that is, as of 1817 and 1821.

Lest one accuse Cauchy and Bolzano as being naive in this matter, note that even those modern texts which include a construction of the reals from, say, the rationals, often carry out that development in an 'optional section': one can skip it yet still learn analysis. And even when such texts stress the importance of that construction, they often leave all the hard work to the reader as exercises.

III.6 Closed Subsets of R

III.6.1 Definition

A subset X of $\mathbb{R}$ is said to be **closed under sequential convergence** provided that the following condition is true:

If $\xi = (x_1, x_2, \dots)$ is a sequence of elements of X which converges to a number L, then the limit L is also in X.

It is convenient in this definition to allow X to be the empty set.
NOTE: The statement 'A subset X of $\mathbb{R}$ is closed under sequential convergence' is normally abbreviated to simply 'X is a closed subset' or, even more simply, 'X is closed' if no confusion is likely.

III.6.2 Examples

(1) Every finite subset of $\mathbb{R}$, including the empty set, is closed. Indeed, since there is no sequence ξ which lies in the empty set, so the fact that $\emptyset$ is closed is trivially true. And if X is a nonempty subset of $\mathbb{R}$, then any sequence in X which is convergent must eventually be constant, so its limit must be in X.

(2) Suppose that a and b are real numbers such that $a < b$.
Part (e) of Theorem (III.2.5) says, in effect, that $[a, b]$ is a closed subset of $\mathbb{R}$. Otherwise stated, a closed interval in $\mathbb{R}$ is a closed subset of $\mathbb{R}$.

In contrast, the interval $(a, b]$ is *not* a closed subset of $\mathbb{R}$. For example, let $x_k = a + (b - a)/k$ for each index k. Then it is clear that $x_k \in (a, b]$ for each k, and that $\lim_{k\to\infty} x_k = a$, but $a \notin (a, b]$.

In a similar manner one can prove that neither $[a, b)$ nor (a, b) is a closed subset of $\mathbb{R}$.

(3) In light of the preceding example, one might conjecture that an interval in $\mathbb{R}$ is a closed subset of $\mathbb{R}$ in the sense of Definition (III.6.1) above, if, and only if, it is a 'closed interval' in the sense of Definition (II.3.1). This conjecture is not true. Indeed, the interval $\mathbb{R}$ is obviously a closed subset of $\mathbb{R}$, but is not a closed interval. The same holds for intervals of the form $[a, +\infty)$ and $(-\infty, b]$, where a and b are real numbers.

(4) The set $\mathbb{Q}$ of all rational numbers is not closed in $\mathbb{R}$, since every irrational number is the limit of a sequence of rational numbers. Likewise, the set $\mathbb{R} \setminus \mathbb{Q}$ of all irrational numbers is not closed in $\mathbb{R}$.

(5) The Cantor Ternary Set is a closed subset of $\mathbb{R}$. Indeed, suppose that $\xi = (x_1, x_2, \dots x_k, \dots)$ is a Cauchy sequence of points in C. Define a number b in C as follows:

(i) Let B_1 be large enough that if $i, j \geq B_1$ then $|x_i - x_j| < 1/3$. It follows that all the numbers x_j with $j \geq B_1$ have the same first 1-free ternary digit; see Part (f) of Theorem (II.5.2). (Note that if a pair of 1-free ternary digits are not equal, then they differ by 2.)

(ii) In a similar mannner one shows that for each k in $\mathbb{R}$ there is B_k such that the first k 1-free ternary digits are the same for all x_j such that $j \geq B_k$.

From this one gets an infinite sequence of 1-free ternary digits d_1, $d_2, \dots$ $d_k, \dots$. Let $L = {}^{3}.\, d_1 d_2 \dots d_k \dots$. Then it is easy to see that $L = \lim_{k\to\infty} x_k$, and clearly L is an element of C. It follows that C is a closed subset of $\mathbb{R}$.

<u>Remark</u> It is an easy exercise to show that the other 'Cantor sets' described in Chapter (I) are also closed subsets of $\mathbb{R}$.

III.6.3 Remarks

(1) The ambiguity in the use of 'closed' described above in the context of intervals can be easily avoided. For example, many authors would carefully write something like 'Let I be an interval of the form $[a, b]$, where a and b are real numbers'. Others would write 'Let I be a closed bounded interval in $\mathbb{R}$'. In the latter case the meaning of 'closed' is still, technically speaking, ambiguous; however, under either interpretation the nature of the resulting set is not: it must be an interval which is a closed subset of $\mathbb{R}$ which is bounded in $\mathbb{R}$, which means it must be of the form $[a, b]$ with a, b in $\mathbb{R}$.

(2) The use of the word 'closed' in connection with limits is similar to the use of the same word in some other contexts; for instance:

'The set $\mathbb{N}$ of all natural numbers is closed under the operation of addition, but it is not closed under the operation of subtraction.'

'The set $\mathbb{Q}$ of all rational numbers is closed under the operations of addition, subtraction, multiplication, and division by nonzero numbers.'

That is, by simply adding natural numbers together, one cannot 'escape' outside the realm of natural numbers; but subtracting natural numbers can lead us outside the set $\mathbb{N}$ and into the larger set $\mathbb{Z}$ of all integers. In contrast, in $\mathbb{Q}$ it is impossible to 'escape' $\mathbb{Q}$ by doing the standard algebraic operations to numbers in $\mathbb{Q}$.

Thus $\mathbb{Q}$ is 'closed' in the algebraic sense, but *not* 'closed' in the limit sense.

III.7 EXERCISES FOR CHAPTER III

III - 1 In each part of this exercise, determine – directly from the definition of 'limit' – whether the given sequence is convergent or not.

(a) $\xi = (x_1, x_2, \dots x_k, \dots)$, where $x_k = \dfrac{3k+100}{k^2+3}$ for each k in $\mathbb{N}$.

(b) $\tau = (t_1, t_2, \dots)$ where $t_k = \dfrac{3k+17}{8k-17}$ for each k in $\mathbb{N}$.

(c) $\alpha = (a_1, a_2, \dots a_k, \dots)$, where $a_k = \dfrac{3k^2+(-1)^k k}{k^2}$.

(d) $\beta = (b_1, b_2, \dots b_k, \dots)$, where $b_k = \sqrt{k+1} - \sqrt{k}$ for each k in $\mathbb{N}$. (You may assume that every positive real number has a unique positive square root.)

III - 2 (a) Prove that if c is a real number such that $c > -1$, then $(1+c)^k \geq 1+kc$ for all natural numbers k.

(b) Use the conclusion of Part (a) to show that if $|r| < 1$ then the sequence $\xi = (1, r, r^2, \dots)$, whose k-th term is r^{k-1}, coverges to 0.

III - 3 (a) Prove or Disprove: Let $\tau = (t_1, t_2, \dots)$ be a convergent sequence of real numbers, and suppose that $\lim\limits_{k\to\infty} t_k \geq a$ for some number a. Then there exists a number N such that $k \geq N$ implies $t_k \geq a$.

(b) Prove or Disprove: Let $\xi = (x_1, x_2, \dots)$ be a convergent sequence of real numbers, and suppose that $\lim\limits_{k\to\infty} x_k > a$ for some number a. Then there exists a number N such that $k \geq N$ implies $x_k > a$.

III - 4 In both parts of this problem, $\alpha = (a_1, a_2, \dots)$ and $\beta = (b_1, b_2, \dots)$ are sequences of real numbers such that $a_k \leq b_k$ for all k in $\mathbb{N}$, and $\lim\limits_{k\to\infty} |b_k - a_k| = 0$.

(a) Prove or Disprove: Suppose that there exists a number c such that $a_k \leq c \leq b_k$ for all $k \in \mathbb{N}$. Then each of the sequences α and β is convergent.

(b) Prove or Disprove: Suppose that for each $k \in \mathbb{N}$ there exists a number c such that $a_k \leq c \leq b_k$. Then each of the sequences α and β is convergent.

III - 5 In each part of this problem, show that Statement (i) implies Statement (ii). Do this using only the axioms for an ordered field – that is, without using 'Completeness' of $\mathbb{R}$.

(a) (i) The Bisection Principle; (ii) The Monotonic Sequences Principle

(b) (i) The Monotonic Sequences Principle; (ii) The Bisection Principle.

(c) (i) The Monotonic-Up Sequences Principle; (ii) The Supremum Principle.

III - 6 Prove Parts (a) and (b) of Theorem (**??**)

III - 7 Prove Parts (c) and (d) of Theorem (**??**)

III - 8 Prove Parts (e) and (f) of Theorem (**??**)

III - 9 Prove Parts (g) and (h) of Theorem (**??**)

III - 10 Let $\xi = (x_1, x_2, \dots)$ be a sequence of *positive* real numbers, and let $\rho = (r_1, r_2, \dots)$ be the corresponding sequence of ratios: $r_k = x_{k+1}/x_k$ for each k in $\mathbb{N}$.

Prove or Disprove If the sequence ξ is convergent, then so is the sequence ρ.

III - 11 Definition Let X be a nonempty subset of $\mathbb{R}$. A point c in $\mathbb{R}$ is said to be an **accumulation point of** X provided that there exists a sequence $\xi = (x_1, x_2, \dots x_k, \dots)$ of points in X such that

(i) for all k in $\mathbb{N}$ one has $x_k \neq c$; (ii) $\lim_{k\to\infty} x_k = c$.

(a) Prove that if a subset X of $\mathbb{R}$ has an accumulation point then X is an infinite set.

(b) Give an example of a subset X of $\mathbb{R}$ which has exactly five accumulation points, three of which are elements of X, two which are not elements of X.

III - 12 (a) Prove the **Bolzano-Weierstrass Theorem for Sets in $\mathbb{R}$**: If X is a bounded infinite subset of $\mathbb{R}$, then X has an accumulation point. (See Exercise **III - 11** for the definition of 'accumulation point'.)

(b) Prove or Disprove If X is an uncountable subset of $\mathbb{R}$, then X has an accumulation point.

III - 13 Give an alternate proof of Theorem (**??**) along the lines of the proof of Theorem (III.2.14), the 'Odd/Even Convergence Theorem'.

III - 14 Suppose that ξ is a bounded infinite sequence of real numbers, and that $\mathcal{F} = (A_1, A_2, \dots A_n, \dots)$ is a countably infinite family of infinite subsets A_n of $\mathbb{N}$ such that $\mathbb{N} = \bigcup_{n=1}^{\infty} A_n$.

Prove or Disprove: If all of the subsequences of ξ corresponding to the subsets A_n in the family $\mathcal{F}$ converge to the same real number L, then the original sequence ξ also converges to L.

Note See Remark (**??**) (2) for the origin of this problem.

III - 15 Prove Parts (a), (b) and (c) of Theorem (**??**).

III - 16 Prove Parts (d) and (e) of Theorem (**??**).

III - 17 Prove Parts (f) and (g) of Theorem (**??**).

III - 18 Prove Part (h) of Theorem (**??**).

III - 19 Define a sequence $\xi = (x_1, x_2, \dots x_k, \dots)$ by the rule

$$x_1 = 1,\ x_{k+1} = \frac{1 + x_k}{2 + x_k} \text{ for each } k \text{ in } \mathbb{N}$$

(a) Show that the sequence ξ is monotonic down and bounded below by 0.

(b) Determine the limit of the sequence ξ.

III - 20 Suppose that $\alpha = (a_1, a_2, \dots)$ and $\beta = (b_1, b_2, \dots)$ are sequences such that

(i) $b_k = \dfrac{3 + a_k}{5 + a_k}$ for all k in $\mathbb{N}$. (ii) $\lim_{k\to\infty} b_k = 10$.

Determine whether the sequence α is convergent; if it is, determine its limit.

III - 21 In the 'Remark' immediately after the proof of Theorem (III.4.1) it is indicated that the Bolzano-Weierstrass Theorem is equivalent to the following result:

Every bounded sequence has a monotonic convergent subsequence.

Problem Prove this last result using the Infinimum/Supremum Principles.

Chapter IV

Continuity – The Basic Theory

Quotes for Chapter (IV):

(1) 'How fleeting are all human passions compared to the massive continuity of ducks.'
(Lord Peter Wimsey to Miss Harriet Vane, in the novel *Gaudy Nights*, by Dorothy L. Sayers)

Introduction

In the standard textbooks of elementary calculus one learns about the concepts of 'limits' and 'continuity' in the context of a real-valued function defined on a fairly simple set in $\mathbb{R}$ – usually an interval, or at worst the union of disjoint intervals. Generally speaking, the treatment of this concept in such texts is neither rigorous nor carried out in depth. For example, such texts do not normally consider 'complicated' subsets of $\mathbb{R}$, or carefully develop the structure of the real number system, or discuss in depth the convergence of real sequences. That is, calculus texts typically omit most of the material found in Chapters (I), (II) and (III) of *This Textbook*.

In texts on real analysis, such as *This Textbook*, one major goal is to revisit elementary calculus, but in a much more rigorous manner. Another major goal, however, is to prepare the reader for even more advanced modern analysis which involves 'spaces', for example, 'metric spaces' or 'topological spaces', that are much more general than $\mathbb{R}$ or even $\mathbb{R}^n$. The structure of the proofs given here is often the same as in the more general contexts, so studying the theory for real-valued functions defined on subsets of $\mathbb{R}$ is good preparation, in a familiar context, for the more advanced theories.

In *This Textbook* the basic theory of 'limits' and 'continuity' that is presented in Chapters (III) above and the present chapter provides ample background to cover the rigorous treatment of elementary calculus which is carried out in Chapters (V) and (VII) below.

In addition to the basic theory on limits and continuity referred to above, Chapter (VIII) expands on that theory. Any reader who prefers to finish the full theory of limits and continuity before starting the study of calculus can do so: simply procede directly from the present chapter to Chapter (VIII), and return to the calculus chapters later. The references to calculus in Chapter (VIII) usually involve facts which are familiar from elementary calculus and can be temporarily accepted 'on faith'.

IV.1 Continuity for Real-Valued Functions Defined in $\mathbb{R}$

Throughout this chapter all the functions under consideration are real-valued with domains being nonempty subsets of $\mathbb{R}$, unless stated explicitly otherwise.

There are two main approaches to the concept of 'continuity' that are commonly used for the theory of real-valued functions defined on a subset of $\mathbb{R}$. The next result says that it doesn't matter which approach one adopts as the 'official' definition.

IV.1.1 Theorem

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a (nonempty) subset X of $\mathbb{R}$. Let c be a point of the set X. Then the following statements are equivalent:

(i) For every sequence $\xi = (x_1, x_2, \dots)$ of numbers in X such that $\lim_{k\to\infty} x_k = c$, one has $\lim_{k\to\infty} f(x_k) = f(c)$.

(ii) For every number $\varepsilon > 0$ there exists a number $\delta > 0$ such that if d in X satisfies $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon$.

Proof Suppose that Statement (ii) holds. Let $\xi = (x_1, x_2, \dots)$ be a sequence of numbers in X such that $\lim_{k\to\infty} x_k = c$, and let $\varepsilon > 0$ be given. Statement (i) guarantees that there exists $\delta > 0$ so that if $d \in X$ and $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon$. For that δ the 'convergence' hypothesis for the sequence ξ impies that there exists a number N in $\mathbb{N}$ such that if $k \geq N$, then $|x_k - c| < \delta$. Combining these facts, one sees that for each $\varepsilon > 0$ there exists a number N in $\mathbb{N}$ such that if $k \geq N$, then $|f(x_k) - f(c)| < \varepsilon$. That is, $\lim_{k\to\infty} f(x_k) = f(c)$. It follows that Statement (i) holds.

To show that Statement (i) implies Statement (ii), it suffices to show the truth of the contrapositive assertion; that is, to prove that Statement (ii) being false implies that Statement (i) is also false.

Thus, suppose that Statement (ii) does not hold for the given f and c. This implies that there exists some $\varepsilon_0 > 0$ so that for every $\delta > 0$ there exists d in X such that $|d - c| < \delta$ but $|f(d) - f(c)| \geq \varepsilon_0$. Apply this successively for δ of the form $1/k$ with k in $\mathbb{N}$ to see that for each k in $\mathbb{N}$ there exists x_k in X such that $|x_k - c| < 1/k$ but $|f(x_k) - f(c)| \geq \varepsilon_0$. The first of these properties implies that $c = \lim_{k\to\infty} x_k$, while the second implies that the sequence $(f(x_1), f(x_2), \dots f(x_k), \dots)$ does *not* converge to $f(c)$. In particular, Statement (i) fails to hold.

IV.1.2 Definition

Let $f : X \to \mathbb{R}$ be a real-valued function whose domain is a (nonempty) subset X of $\mathbb{R}$. Let c be a point of X.

(1) One says that f **is continuous at c** if either Statement (i) or Statement (ii) of the preceding theorem holds; of course it then follows that *both* statements hold. One then refers to Statement (i) as the **sequential characterization of continuity**, and Statement (ii) as the **$\varepsilon\,\delta$ characterization of continuity**.

If $f : X \to \mathbb{R}$ is *not* continuous at c, then one says that **f is discontinuous at c**, and that **the point c is a discontinuity of** f.

(2) Let S be a nonempty subset of the set X. One says that ***f* is continuous on the subset *S*** provided $f : X \to \mathbb{R}$ is continuous at c (in the sense of Part (1) of this definition) for each number c in S.

(3) The function $f : X \to \mathbb{R}$ is said to be a **continuous function** if it is continuous on X (in the sense of Part (b)); that is, if f is continuous at each point of its domain X.

IV.1.3 Remarks

(1) Many texts in real analysis use the '$\varepsilon\,\delta$' characterization of continuity as their 'official' definition. For such texts the 'sequential' characterization becomes a theorem to be proved. Likewise, other such texts prefer to use the 'sequential' characterization as their 'official' definition of continuity. In those texts the '$\varepsilon\,\delta$' characterization then becomes a theorem. In any event, both types of texts must prove Theorem (IV.1.1). The approach taken in *This Textbook* avoids the problem of choosing which of these characterizations to take as the 'official' definition.

(2) Note also that to say that '$f : X \to \mathbb{R}$ has a discontinuity at a number c' requires that c be in X. This usage, which appears to be the norm in real analysis texts such as *This Textbook*, is *not* what is taught in elementary calculus. Indeed, a typical problem in a freshman calculus course might be to 'find the discontinuites of the function $f(x) = 1/\sqrt{x^2 - 1}$', and the 'correct' answer would be '$x = -1$ and $x = 1$'; correct, that is, in the context of elementary calculus. In real analysis, in contrast, the domain of this function consists of all x in $\mathbb{R}$ such that $x < -1$ or $x > 1$, so it is continuous at each point of its domain, and thus, in accordance with Definition (IV.1.2), has no discontinuities.

(3) Part (2) of the preceding definition appears to be innocuous, but it can lead to confusion. The problem is that the phrase 'f is continuous on S' is sometimes interpreted as meaning (using the 'sequential characterization'):

'*For every point c in S and for every sequence $\xi = (x_1, x_2, \dots)$ of points* in S *converging to c, the corresponding sequence of values $(f(x_1), f(x_2), \dots)$ converges to $f(c)$.*'

This is equivalent to saying that the restriction $g = f|_S$ of f to S, whose domain is the set S, is continuous at each point of S. This is *not* the intended meaning of Part (2) of the definition. For example, suppose that $f : X \to \mathbb{R}$ is a function with domain X such that f is not contiuous at a particular point c of X. Let S be the singleton set $\{c\}$. Then the function f is *not* continuous on the subset S, in the sense of Part (2) of the definition, since it fails to be continuous at a point of S, namely c. However, it is easy to see that the restriction $f|_S$ *is* continuous on S, since constant functions are continuous; see Example (IV.1.4) (1) below.

(4) It is useful, for future reference, to reformulate the condition for $f : X \to \mathbb{R}$ to be continuous on X, stated in Part (2) of the preceding definition with $S = X$, the full domain of f, in the following alternate form:

'*For every c in X and for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every d in X such that $|d - c| < \delta$, one has $|f(d) - f(c)| < \varepsilon$*'.

This alternate formulation does not refer to a previously defined concept of continuity at a single point of X. In that sense, this formulation is closer to the original definitions of continuity given by Bolzano and Cauchy, which focused on the continuity of a function on a full interval; see the following Side Comment.

Side Comment (on the approaches to continuity of Bolzano and Cauchy)

The modern concept of 'continuity' of a function $y = f(x)$, namely that small changes in the input x of the function cause small changes in the corresponding output y, took many years to evolve. Historians of mathematics tell us that the earliest treatments of continuity that agree in essence with the modern definitions were carried out by Bolzano and Cauchy in publications of 1817 and 1821, respectively. Both definitions focus on real-valued functions with domain some intervals in $\mathbb{R}$, instead of more general sets of numbers.

Bolzano's definition can be translated as follows:

'*According to a correct definition, the expression that a function* $\mathrm{f\,x}$ *varies according to the law of continuity for all values of x inside or outside certain limits means just that: if x is some such value, the difference $f(x+\omega) - f\,x$ can be made smaller than any given quantity provided ω can be taken as small as we please.*' (Translation given in '*Bolzano's Philosophy and the Emergence of Modern Mathematics'* by P. Rusnock.)

Remark Bolzano's definition is essentaily the same as the $\varepsilon\,\delta$ characterization described in Remark (IV.1.3) (4) above; in particular, the 'given quantity' corresponds to our ε, and the quantity ω is our δ; 'as small as we please' should then be interpreted to mean 'sufficiently small'. Of course when Bolzano says 'smaller than any given quantity' and 'as small as we please', he is referring to the absolute value of the quantities involved. The 'absolute value notation' $|\,|$ now taught in arithmetic did not become common in mathematics until after 1840.

The Cauchy definition expresses the same idea:

'*Let $f(x)$ be a function of the variable x, and suppose that for each value of x between two given limits, the function always takes a unique finite value. If, beginning with a value of x contained between these limits, we add to the variable x an infinitely small increment a, the function itself is incremented by the difference $f(x+a) - f(x)$, which depends both on the new variable a and on the value of x. Given this, the function $f(x)$ is a continuous function of x between the assigned limits if, for each value of x between these limits, the numerical value of the difference $f(x+a) - f(x)$ decreases indefinitely with the numerical value of a. In other words, the function $f(x)$ is continuous with respect to x between the given limits if, between these limits, an infinitely small increment in the variable always produces an infinitely small increment in the function itself.*' (Translation given in *Cauchy's Cours d'analyse: An Annotated Translation* by Bradley and Sandifer.)

Remark By 'infinitesimal quantity' Cauchy means 'a quantity which has 0 as its limit'. This evokes a sense of a quantity 'approaching' a value and thus a sense of 'time' and 'motion', physical concepts which lie outside the purely analytical treatment of analysis which Cauchy claimed to present. Indeed, Cauchy even uses the phrase 'successive values' of such a quantity in this context, which again suggests the idea of 'time'. However, Cauchy goes on to provide an explicit example of what he means by an 'infinitesimal quantity': $a = \frac{1}{4}, \frac{1}{3}, \frac{1}{5}, \frac{1}{6}, \ldots$. This example shows that, whatever Cauchy meant by 'infinitesimal quantity', he included all sequences of real numbers which converge to 0. The Cauchy approach to 'continuity' thus corresponds essentially to the standard 'sequence characterization' of continuity given above. Indeed, the sequence $(x_1 - c, x_2 - c, \ldots x_k - c, \ldots)$ corresponds to Cauchy's infinitesimal a: note that $c + (x_k - c) = x_k$.

Cauchy's 'numerical value' is the same as the modern 'absolute value'.

Note that both Cauchy and Bolzano use the word 'limit' in the sense of a quantity which bounds the variable x, not as a limit as in 'limit of a sequence'. For example, Bolzano allows his x to vary 'inside or outside certain limits'. This means restrictions such as $a < x < b$ or $-\infty < x < b$ and so on. Similarly, Cauchy 'limits' his variable x to lie in some sort of interval.

In addition, each author's definition could be translated in more modern language as 'f is continuous inside an interval provided it is continuous at each x in the interval'. However, at this time neither author assigns a name to what we now refer to as 'continuity at a single point c'; perhaps both would have found it silly to even consider a function if it were continuous at, say, only a single point. Indeed, later on Cauchy does define 'continuity at a single point x', but for him this means 'continuity on a full interval containing x'. In particular, at this early stage both Bolzano and Cauchy would describe 'continuity of $f : X \to \mathbb{R}$ on a subset S of X' along the lines of Remark (IV.1.3) (4) above. Of course neither author would have considered functions defined on arbitrary nonempty subsets of $\mathbb{R}$, as is done in *This Textbook*; that is a much more recent viewpoint.

IV.1.4 Examples

(1) Let X be a nonempty set of real numbers.

(i) If $f : X \to \mathbb{R}$ is a constant function with domain X, then f is continuous on X.

(ii) If $g : X \to X$ is the identity function with domain X, then g is continuous on X.

Indeed, suppose $\xi = (x_1, x_2, \dots x_k, \dots)$ is a sequence of points in X which converges to a point c in X. Let b be the constant value of the function f, so that $f(x_k) = b = f(c)$ for each index k. Then clearly $\lim_{k\to\infty} f(x_k) = b = f(c)$. Likewise, since $g(x) = x$ for each x in X, it follows that $g(x_k) = x_k$, and thus $\lim_{k\to\infty} g(x_k) = \lim_{k\to\infty} x_k = c = g(c)$. The claimed continuity follows.

(2) Let $f : \mathbb{R} \to \mathbb{R}$ be the Dirichlet function which was discussed in Example (I.6.8). Thus, if $x \in \mathbb{R}$ then $f(x) = 0$ if x is irrational, while $f(x) = 1$ if x is rational.

It is easy to see that f is discontinuous at every point of $\mathbb{R}$. Indeed, suppose that c is an irrational number, and let $\xi = (x_1, x_2, \dots)$ be a sequence of *rational* numbers such that $c = \lim_{k\to\infty} x_k$. Then $f(x_k) = 1$ for all k in $\mathbb{N}$, hence $\lim_{k\to\infty} f(x_k) = 1$. However, $f(c) = 0$, so it is *not* the case that $\lim_{k\to\infty} f(x_k) = f(c)$; thus, f is discontinuous at c.

Similarly, if c is a rational number, then let $\zeta = (z_1, z_2, \dots)$ be a sequence of *irrational* numbers which converges to c. One then notes that $f(z_k) = 0$ for all k, hence $\lim_{k\to\infty} f(z_k) = 0 \neq f(c)$ since $f(c) = 1$ when c is rational.

(3) Let $g : \mathbb{R} \to \mathbb{R}$ be given by the rule $g(x) = x\, f(x)$, where f is the Dirichlet function studied in the preceding example. It is a simple exercise to show that if $c = 0$ then g is continuous at c, but if $c \neq 0$ then g is discontinuous at c. That is, with the modern definition of 'continuity' it is possible for a function defined on an interval to be continuous at just a single point of its domain.

(4) Let $f : [0, +\infty) \to \mathbb{R}$ be given by the formula $f(x) = \sqrt{x}$ for all x in the domain $[0, +\infty)$ of f.

<u>Claim 1</u> If $c > 0$, then f is continuous at c.

<u>Proof of Claim 1</u> Note that, by the usual algebraic trick, one has

$$|f(d) - f(c)| = |\sqrt{d} - \sqrt{c}| = \left|(\sqrt{d} - \sqrt{c})\left(\frac{\sqrt{d} + \sqrt{c}}{\sqrt{d} + \sqrt{c}}\right)\right| = \frac{|d - c|}{\sqrt{d} + \sqrt{c}} \leq \frac{|d - c|}{\sqrt{c}} \text{ for all } d \geq 0$$

Now let $\varepsilon > 0$ be given, and let $\delta = \varepsilon\sqrt{c}$. Then it is clear that if $d \in X$ and $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon$, as required.

<u>Claim 2</u> The function f is also continuous at $c = 0$.

<u>Proof of Claim 2</u> Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be any sequence in X converging to 0, and consider the corresponding sequence of values, namely $\rho = (r_1, r_2, \dots r_k, \dots)$, where $r_k = f(x_k) = \sqrt{x_k}$ for each index k. Since the (convergent) sequence ξ is bounded, so is the sequence ρ. Let $\sigma = (s_1, s_2, \dots s_m, \dots)$ be any convergent subsequence of ρ, and let its limit be L. Then the sequence $(s_1^2, s_2^2, \dots s_m^2, \dots)$ is a subsequence of the original sequence ξ, hence by the Product Rule for Limits of Sequences, together with the fact that ξ converges to 0, one has $L^2 = 0^2 = 0$,

so $L = 0$. Now Part (b) of Theorem (III.4.7) applies to show that ρ converges to 0; that is, $\lim_{k\to\infty} f(x_k) = 0 = f(0)$. Thus f is also continuous at 0, as claimed.

(5) Let $P_0 = (x_0, y_0)$, $P_1 = (x_1, y_1), \ldots P_n = (x_n, y_n)$ be $n + 1$ points in $\mathbb{R}^2$ such that $x_0 < x_1 < \ldots < x_{n-1} < x_n$. For convenience set $a = x_0$ and $b = x_n$, and let I denote the closed interval $[a, b]$. Now let $h : I \to \mathbb{R}$ be the piecewise-linear interpolation through these $n + 1$ points; see Example (I.6.8). Then h is continuous at every point of I. More precisely, for each $\varepsilon > 0$ there exists $\delta > 0$ such that if $c, d \in I$ are such that $|d - c| < \delta$, then $|h(d) - h(c)| < \varepsilon$.

More precisely, for each $j = 1, 2, \ldots n$ let $s_j = (y_j - y_{j-1})/(x_j - x_{j-1})$; in geometric terms, s_j is the slope of the line segment in the plane $\mathbb{R}^2$ joining the points P_{j-1} and P_j. Let M be any positive number such that $M \geq \max\{|s_1|, |s_2|, \ldots |s_n|\}$. Then for each pair of numbers c and d in I one has $|h(d) - h(c)| \leq M\,|d - c|$.

To see this, note first that if x is a number such that $x_{j-1} \leq x \leq x_j$ for some index j, then, by the formula for 'linear interpolation', one has

$$h(x) = y_{j-1} + \left(\frac{y_j - y_{j-1}}{x_j - x_{j-1}}\right)(x - x_{j-1}) = y_{j-1} + s_j\,(x - x_{j-1}).$$

Without loss of generality, assume that $c < d$. Then there exist indices i and k, with $i \leq k$, such that $x_{i-1} \leq c \leq x_i$ and $x_{k-1} \leq d \leq x_k$. Assume that $i < k$; the case $i = k$ is even easier, and is left as an exercise. Using the formula for h obtained above, together with the old 'Add-and-Subtract' trick, one gets

$$h(d) - h(c) = (h(d) - h(x_{k-1})) + (h(x_{k-1}) - h(x_{k-2})) + \ldots + (h(x_i) - h(c)) =$$

$$s_k\,(d - x_{k-1}) + s_{k-1}\,(x_{k-1} - x_{k-2}) + \ldots + s_i\,(x_i - c).$$

Now use the Triangle Inequality, together with the fact that the coefficients $d - x_{k-1}$, $x_{k-1} - x_{k-2}, \ldots x_i - c$ of s_k, $s_{k-1}, \ldots s_i$ are nonnegative, to get

$$|h(d) - h(c)| \leq |s_k|\,(d - x_{k-1}) + |s_{k-1}|\,(x_{k-1} - x_{k-2}) + \ldots + |s_i|\,(x_i - c) \leq$$

$$M\,((d - x_{k-1}) + (x_{k-1} - x_{k-2}) + \ldots + (x_i - c)) = M\,|d - c| < M\left(\frac{\varepsilon}{M}\right) = \varepsilon.$$

Finally, suppose that c is any point of the interval I. Let $\varepsilon > 0$ be given, and let $\delta = \varepsilon/M$. Then, by the preceding result, for each d in I such that $|d - c| < \delta$ one has

$$|g(d) - g(c)| \leq M\,|d - c| < M\left(\frac{\varepsilon}{M}\right) = \varepsilon,$$

as required.

(6) Let h, $I = [a, b]$ and M be as in the preceding example. Let $h : \mathbb{R} \to \mathbb{R}$ be the extension of g to $\mathbb{R}$ (see Definition (**??**)) defined by the rule

$$g(x) = \begin{array}{ll} h(x) & \text{if } a \leq x \leq b \\ h(a) & \text{if } x < a \\ h(b) & \text{if x ¿ b} \end{array}$$

It is easy to see, from the preceding exercise, that g satisfies the analogous inequality

$$|g(d) - g(c)| \leq M\,|d - c| \text{ for every } c$$

Note The preceding examples illustrate the fact that sometimes it is easier to attack a 'continuity' problem by using the sequential characterization, while sometimes the $\varepsilon\delta$ characterization works more easily.

The conditions enjoyed by the piecewise-linear functions h in Example (5) and g in Example (6) above appear frequently in analysis and is given a name.

IV.1.5 Definition

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty subset X of $\mathbb{R}$. One says that ***f* satisfies a Lipschitz condition on *X***, or that ***f* is a Lipschitz function on *X***, provided there exists a real number $M > 0$ such that $|f(d) - f(c)| \leq M\,|d - c|$ for all c and d in the set X.

Remarks (1) The condition in question is named after the nineteenth-century German mathematician Rudolf Lipschitz, who recognized its importance in the theory of differential equations.

(2) The result in Example (5) above can be phrased to say that a continuous piecewise-linear function on an interval $[a, b]$ satisfies a Lipschitz condidtion on the set $[a, b]$. A similar remark applies to Example (6).

(3) It is obvious that if a function $f : X \to \mathbb{R}$ satisfies a Lipschitz condition on a set X, then f is continuous on X. However, the converse is not true. For example, it is an instructive exercise to show that the square-root function is *not* Lipschitz on, say, the closed interval $[0, 1]$.

IV.2 Standard Continuity Laws in R

The next several results allow one to construct functions which are continuous at a point, from other such functions, using the ordinary processes of algebra. If one uses the 'sequential characterization' of continuity, then proving these standard continuity laws becomes a trivial exercise in applying the standard limit laws for convergent sequences of real numbers from the preceding chapter; most such proofs are left to the reader. Some of these results are proved here using the '$\varepsilon\,\delta$' characterization, mainly to illustrate how such '$\varepsilon\,\delta$' proofs work.

IV.2.1 Theorem

Let $f : X \to \mathbb{R}$ be a real-valued function defineed on a (nonempty) subset X of $\mathbb{R}$.

(a) If f is continuous at a point c of X, then the function $|f|$ is also continuous at c.

(b) If f is continuous at a point c of X and A is any real number, then $A{\cdot}f$ is continuous at c.

(c) Suppose that S is a subset of X and that c is a point of S. Let $g = f|_S : S \to \mathbb{R}$. If f is continuous at c, then g is also continuous at c. However, the converse need not be true. That is, it can happen that g is continuous at c but f is not. (Compare with Remark (IV.1.3).)

(d) Suppose that X, S and c are as in Part (c), and suppose further that there is an open interval $I = (a, b)$ such that $c{\in}I$ and $I \subseteq S$. With this extra hypothesis, the converse of the conclusion in Part (c) *is* true. That is, if $g = f|_S : S \to \mathbb{R}$ is continuous at c, then $f : X \to \mathbb{R}$ is also continuous at c.

Similarly, suppose that X has a least element a, and suppose that there is an interval $I = [a, b)$ with $c \in I$ and $I \subseteq X$. Then the converse of Part (c) is true. Likewise, if X has a maximum element b and there is an interval $I = (a, b]$ such that $c \in I$ and $I \subseteq X$, then the converse of Part (c) is true.

Partial Proof

(a) By the definition of the function $|f|$, together with the Reverse Triangle Inequality, one has

$$|\,|f|(d) - |f|(c)\,| = |\,|f(d)| - |f(c)|\,| \leq |f(d) - f(c)|,$$

The rest of the proof of the desired result is now left as an easy exercise.

(b) The case in which $A = 0$ is trivial, and is left as an exercise.

Thus, assume that $A \neq 0$, and let $\varepsilon > 0$ be given. The hypothesis that f is continuous at c then implies that there is $\delta > 0$ such that if $d \in X$ and $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon/|A|$. Multiply both sides of this inequality by $|A|$ and do the obvious simplification to get

$$|(A{\cdot}f)(d) - (A{\cdot}f)(c)| = |A|{\cdot}|f(d) - f(c)| < \varepsilon \text{ if } d \in X \text{ and } |d - c| < \delta$$

The desired result follows.

(c) The proof that g is continuous at c is trivial, and is left as an exercise.

One simple example of the 'converse' not holding is this: Suppose that $X = \mathbb{R}$, $S = \mathbb{Q}$ and $c = 0$. Let $f : \mathbb{R} \to \mathbb{R}$ be the Dirichlet function, which is known to be discontinuous at every point; see Example (IV.1.4) (2). But by the definition of the Dirichlet function, $g = f|_{\mathbb{Q}} : \mathbb{Q} \to \mathbb{R}$ is a constant function, hence continuous at each point of its domain, including at 0.

(d) Write the open interval I in the form $I = (a, b)$, so that $a < c < b$. Let $\delta_1 > 0$ be small enough that $a < c - \delta_1 < c < c + \delta_1 < b$. Now let $\varepsilon > 0$ be given, and let $\delta_2 > 0$ be small enough that if $d \in S$ and $|d - c| < \delta_2$, then $|g(d) - g(c)| < \varepsilon$. Let $\delta = \min\{\delta_1, \delta_2\}$. If d in X satisfies $|d - c| < \delta$, then $|d - c| < \delta_1$ and thus $d \in I \subseteq S$; in particular, $d \in S$. It follows that $|f(d) - f(c)| = |g(d) - g(c)|$, by the definition of 'restriction'. Furthermore, one also has $|d - c| < \delta_2$, so $|g(d) - g(c)| < \varepsilon$. Combine these facts to conclude that if $d \in X$ and $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon$. It follows that $f : X \to \mathbb{R}$ is continuous at c, as required.

The other statements in this part of the theorem are proved in a similar manner.

IV.2.2 Theorem

Let $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ be real-valued functions defined on X and Y, respectively, where X and Y are nonempty subsets of $\mathbb{R}$. Suppose that $Z = X \cap Y \neq \emptyset$, and that c is a point of Z. Assume that $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ are both continuous at c, when viewed as function diagrams $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$, resepctively. Then:

(a) Then $(f + g) : Z \to \mathbb{R}$ and $(f - g) : Z \to \mathbb{R}$ are both continuous at c.

(b) Likewise, the function $(f{\cdot}g) : Z \to \mathbb{R}$ is continuous at c.

(c) Suppose, in addition, that $g(c) \neq 0$, and let $W = \{x \in Z : g(x) \neq 0\}$. Then $(f/g) : W \to \mathbb{R}$ is continuous at c.

Partial Proof (a) and (b): The reader is encouraged to prove these parts using the $\varepsilon\,\delta$ characterization of contiinuity.

(c) For simplicity, set $h = f{\cdot}g : Z \to \mathbb{R}$.

Special Case Suppose that $f(c) = 0$ and $g(c) = 0$. Then one has $h(c) = f(c)\,g(c) = 0$, hence $|h(d) - h(c)| = |f(d)\,g(d)|$. Let $\varepsilon > 0$ be given; without loss of generality one may assume that $\varepsilon < 1$. By hypothesis, $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ are continuous at c. It then follows from Part (c) of the preceding theorem that their restrictions to Z are continuous at c. Now let $\varepsilon > 0$; without loss of generality, assume that $0 < \varepsilon < 1$. It follows from this continuity that there exists $\delta > 0$ such that if $d {\in} Z$ and $|d - c| < \delta$, then $|f(d)| < \varepsilon$ and $|g(d)| < \varepsilon$. Since, by hypothesis, $0 < \varepsilon < 1$, it then follows that for such d one has $|h(d)| = |f(d){\cdot}g(d)| < \varepsilon^2 < \varepsilon$, as required.

General Case Define $F : Z \to \mathbb{R}$ and $G : Z \to \mathbb{R}$ given by the rules $F(x) = f(x) - f(c)$ and $G(x) = g(x) - g(c)$. Consider $H : Z \to \mathbb{R}$ given by $H(x) = F(x){\cdot}G(x)$, so that $F(c) = G(c) = 0$. It follows from Part (b), together with the fact that constant functions are continuous, that F and G satisfy the hypothesis of the special case just considered, and thus that H is continuous at c. However, one computes that for all x in Z one has

$$H(x) = f(x){\cdot}g(x) - f(c){\cdot}g(x) - f(x){\cdot}g(c) + f(c){\cdot}g(c);$$

that is,

$$f(x)\cdot g(x) = H(x) + f(c){\cdot}g(x) + f(x){\cdot}g(c) - f(c){\cdot}g(c).$$

It follows easily by repeated use of Parts (a) and (b) of the present theorem, together with Part (b) of the preceding theorem, that $f{\cdot}g : Z \to \mathbb{R}$ is continuous at c, as claimed.

IV.2.3 Examples

(1) Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a polynomial function on $\mathbb{R}$; that is, there exist finitely many real numbers c_0, $c_1, \ldots c_m$ such that

$$f(x) = c_0 + c_1\,x + c_2\,x^2 + \ldots + c_m\,x^m \text{ for all } x \text{ in } \mathbb{R}.$$

Then f is continuous on $\mathbb{R}$.

(2) Suppose that h is a rational function; that is, there exist polynomial functions f and g on $\mathbb{R}$, with g not equal to the zero polynomial, such that the domain D of h equals the set of all x such that $g(x) \neq 0$, and $h(x) = f(x)/g(x)$ for all x in D. Then h is continuous on D.

Both of these statements follow directly from the previous theorems, combined with Mathematical Induction. The details are left as an exercise.

The next result says, in effect, that 'continuity' behaves nicely under composition.

IV.2.4 Theorem

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty subset X of $\mathbb{R}$. Likewise, let $g : Y \to \mathbb{R}$ be such a function defined on a nonempty subset Y of $\mathbb{R}$. Assume that $f[X] \subseteq Y$. Let c and b be points of X and Y, respectively, such that $b = f(c)$. Suppose that f is continuous at c and that g is continuous at b. Then the composition $h = g{\circ}f$ is continuous at c.

Proof For sake of variety let us prove this result using the 'sequences characterization' of continuity.

Let $\xi = (x_1, x_2, \ldots)$ be a sequence in X such that $\lim_{k\to\infty} x_k = c$. It follows from the continuity hypothesis on f that $\lim_{k\to\infty} f(x_k) = f(c)$. Let $y_k = f(x_k)$ for each k in $\mathbb{N}$, so that the previous equation can be written $\lim_{k\to\infty} y_k = b$. (Recall that, by hypothesis, $f(c) = b$.) Then, since $f[X]\subseteq Y$, it also follows that $y_k{\in}Y$ for each k. Now the continuity hypothesis on g implies that $\lim_{k\to\infty} g(y_k) = g(b)$. Combine this with the fact that $h(x_k) = g(f(x_k)) = g(y_k)$ for each k in $\mathbb{N}$, and $h(c) = g(f(c)) = g(b)$ to get $\lim_{k\to\infty} h(x_k) = \lim_{k\to\infty} g(y_k) = g(b) = h(c)$, as required.

IV.2.5 Example

Let $f : [0,+\infty)\to\mathbb{R}$ be the standard Square-Root Function, given by $f(x) = \sqrt{x}$ for each $x \geq 0$. In Example (IV.1.4) (4) it is proved that f is continuous at each point of its domain $[0,+\infty)$. Since f maps $[0,+\infty)$ into itself, it follows that the composition $f{\circ}f : [0,+\infty)\to\mathbb{R}$ is defined, and by the preceding theorem this composition is continuous. This process can be repeated as often as one wishes.

Thus, for each n in $\mathbb{N}$ let $f_n : [0,+\infty)\to\mathbb{R}$ be the function obtained by composing f with itself $(n-1)$ times; thus, $f_1 = f$, $f_2 = f{\circ}f$, $f_2 = f{\circ}f{\circ}f$, and so on. Clearly each of these functions is also continuous on $[0,+\infty)$. It is easy to see that for each $x \geq 0$, $f_n(x)$ is the 2^n-th root of x; that is, $(f_n(x))^{2^n} = x$. Using the 'exponent' notation from high-school algebra, this can be written $f_n(x) = x^{1/2^m}$. Thus the function $g(x) = x^{1/2^m}$ is continuous on $[0.+\infty)$.

Now let m be any natural number, and $p_m : \mathbb{R}\to\mathbb{R}$ denote the 'm-th power function' on $\mathbb{R}$. This is a polynomial function, hence continuous on $\mathbb{R}$, so the composition $h(x) = (p_m{\circ}f_n)(x) = x^{m/2^n}$ is continuous on $[0,+\infty)$ as well. We shall see later in this chapter that the exponent $m/2^n$ can be replaced by any rational exponent, and even later by any real number, with the result still continuous on $[0,+\infty)$.

Definition (IV.1.2) seems to require that one verify that $\lim_{k\to\infty} f(x_k) = f(c)$ for *every* sequence $\xi = (x_1, x_2, \dots)$ in X which converges to c. The next result, however, shows that one can get by with checking it only for such sequences which are *monotonic*.

IV.2.6 Theorem

Let $f : X\to\mathbb{R}$ be a real-valued function whose domain is a (nonempty) subset X of $\mathbb{R}$. Let c be a point of X. If $\lim_{k\to\infty} f(x_k) = f(c)$ for every monotonic sequence $\xi = (x_1, x_2, \dots)$ in X converging to c, then f is continuous at c.

Proof We actually prove the contrapositive of the given statement. That is:

If f is *not* continuous at c then there exists a monotonic sequence $\zeta = (z_1, z_2, \dots)$ in X converging to c such that the corresponding sequence $(f(z_1), f(z_2), \dots)$ does *not* converge to $f(c)$.

Indeed, suppose that f is not continuous at c. Then there must exist a number $\varepsilon_0 > 0$ with the following property:

$$\text{For every } k \text{ in } \mathbb{N} \text{ there exists a number } x_k \text{ in } X \text{ such that } |x_k - c| < \frac{1}{k} \text{ and } |f(x_k) - f(c)| \geq \varepsilon_0 \qquad (*)$$

It is clear from $(*)$ that the sequence $\xi = (x_1, x_2, \dots)$ converges to c. It then follows from Theorem (III.4.6) that the sequence ξ has a subsequence $\zeta = (z_1, z_2, \dots)$ which is monotonic. The sequence ζ also converges to c, by Part (b) of Theorem (III.2.1). However, since $|f(x_k) - f(c)| \geq \varepsilon_0$ for all k, it follows that $|f(z_j) - f(c)| \geq \varepsilon_0$ for all j, and thus the sequence $(f(z_1), f(z_2), \dots)$ does not converge to $f(c)$. The desired result now follows.

The preceding result is especially useful in determining whether a function f is continuous at a point c in the case the formula for f to the left of c differs from that to the right of c. It leads naturally to the following generalization of the concept of 'continuity'.

IV.2.7 Definition

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty subset X of $\mathbb{R}$, and let c be a point of X. One says that f is **continuous at c from the left** provided that $\lim_{k\to\infty} f(x_k) = f(c)$ for every sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ in X which is monotonic up and converges to c. Likewise, one says that f is **continuous at c from the right** provided that $\lim_{k\to\infty} f(x_k) = f(c)$ for every sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ in X which is monotonic down and converges to c. If either of these conditions holds, then one says that **f has one-sided continuity at c**.

The preceding theorem can now be phrased more compactly.

IV.2.8 Corollary

A function $f : X \to \mathbb{R}$ is continuous at a point c of a nonempty subset X of $\mathbb{R}$ if, and only if, f has **two-sided continuity at c**; that is, f is continuous at c both from the left and from the right.

The simple proof is left as an exercise.

It is a useful exercise to use the preceding result to redo the analysis of the piecewise-linear function studied in Example (IV.1.4) above.

Remark If $c = \min X$ then f is automatically continuous from the left, since the only monotonic-up sequence in X with limit c is the constant sequence $(c, c, \ldots c, \ldots)$. In this situation one need only show that f is continuous at c from the right. Likewise, if $c = \max X$, then f is automatically continuous from the right at c, and one need only check continuity at c from the left.

IV.3 Continuity Theorems on Nonempty Subsets of R

The next theorem is always stated in elementary calculus, at least for functions defined on closed bounded intervals, and is used in a variety of important ways, both practical and theoretical; but it is almost never proved there. In contrast, the theorem which follows it is almost never even stated in elementary calculus.

IV.3.1 Theorem (The Extreme-Value Theorem in R)

Suppose that $f : X \to \mathbb{R}$ is a real-valued function which is continuous on a closed bounded set X in $\mathbb{R}$. Then the function f assumes minimum and maximum values on the set X. That is, there are numbers c and d in X such that $f(c) \leq f(x) \leq f(d)$ for all x in X.

Proof Let S be the image $f[X]$ of the set X under the function f; that is,

$$S = \{y : y = f(x) \text{ for at least one } x \text{ in the set } X\}.$$

It is clear that *if* such points c and d in X do exist, then one must have $f(c) = \inf S$ and $f(d) = \sup S$. Thus consider the quantities $m = \inf S$ and $M = \sup S$; we do not assume that either of these quantities is finite.

First note that, by Theorem (III.2.18), there exists a sequence $\zeta = (z_1, z_2, \ldots z_k, \ldots)$ in the set S such that $\lim_{k\to\infty} z_k = m$. (The fact that this sequence can be chosen to be monotonic down is not needed here.) The definition of the set S then implies that for each index k there exists at least one point x_k in X such that $f(x_k) = z_k$. Of course there is no reason to expect that the sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ has a limit. However, the hypothesis that X is bounded allows one to conclude, from the Standard Bolzano-Weierstrass Theorem for sequences, that the sequence ξ does have a subsequence $\sigma = (s_1, s_2, \ldots s_m, \ldots)$ which is convergent. Let $c = \lim_{m\to\infty} s_m$. The hypothesis that X is closed in $\mathbb{R}$ then implies that $c \in X$. Since, by hypothesis, f is continuous at every point of the set X, it is therefore continuous at c. Thus $\lim_{m\to\infty} f(s_m) = f(c)$. However, the sequence $(f(s_1), f(s_2), \ldots f(s_m), \ldots)$ is a subsequence of the sequence ζ, so $\lim_{m\to\infty} f(s_m) = \lim_{k\to\infty} z_k = \inf S$. Combine these facts to get that $\inf S = f(c)$ with c in X, and thus $\inf S$, i.e., $f(c)$, is an element of S, hence $f(c)$ is the minimum element of S. That is, $f(c)$ is the minimum value of f on the set X.

A similar proof works for the maximum value of f on X. Even simpler, just apply the preceding result for 'minimum' to the function $g = -f$.

IV.3.2 Corollary

Suppose that X and $f : X \to \mathbb{R}$ is as in the preceding theorem, and that for every x in X one has $f(x) \neq 0$. Then there exists a number $m > 0$ such that $|f(x)| \geq m$ for all x in X.

Proof Let $h : X = \mathbb{R}$ be defined by $h(x) = 1/|f(x)|$ for each x in X; note that $h(x) > 0$ for every x in I. Clearly h is continuous on I, so the preceding theorem implies that h is bounded. In particular, there exists $M > 0$ such that $0 < h(x) \leq M$ for all x in I; that is, $0 < 1/|f(x)| \leq M$. It follows that $|f(x)| \geq m > 0$, where $m = 1/M$.

Side Comment (on motivating the next important concept)

Because of the length of this Side Comment, it seems approriate to remind you that the content of a Side Comment is not needed for the logical development of the material in *This Textbook*, and no exercises are based on it. Nevertheless, you are encouraged to read the Side Comments for whatever extra insights they might provide.

The next big definition, namely Definition (IV.3.3) below, illustrates a major problem with the modern style of mathematical writing: in order to present topics in a 'logical' order, these topics must often be introduced without explaining historically how their importance became clear to mathematicians. For example, the definition in question 'logically' belong in the current chapter because it involves only continuity. In reality, however, it originally grew out of Cauchy's attempt to rigorously define the definite integral, a topic which appears – quite 'logically' – much later in *This Textbook*, and thus whose historical origins cannot readily be discussed at this point.

One common way out of this dilemma, certainly the easiest, and often the best, is for the textbook author to give the following advice:

'Trust me; you'll understand the true significance of the concept later on.'

However, sometimes it is possible to follow a 'pseudo-historical' approach to such a topic. This means considering a situation which makes sense at the current point of the logical treatment and which *could* have motivated the desired topic, but which historically did not play that role. This is the approach taken here to motivate Definition (IV.3.3). Of course, if this approach does not provide sufficient motivation for your needs, then follow the dictates of the advice given above.

Preliminary Remark

Suppose that $f : (a,b) \to \mathbb{R}$ is a continuous function whose domain is an open interval (a,b), where a and b are finite. Is it possible to 'extend' f to a continuous function on the closed interval $[a,b]$? That is, is there a function $g : [a,b] \to \mathbb{R}$ which is continuous on $[a,b]$ and whose restriction to (a,b) equals f?

The immediate answer is: it depends on the circumstances.

Example Consider the following functions, each with domain the open interval $(0,1)$:

$$f_1(x) = x\,(1-x); \qquad f_2(x) = \frac{1}{x\,(1-x)}.$$

Clearly f_1 extends to continuous $g_1 : [0,1] \to \mathbb{R}$, where $g_1(x) = x\,(1-x)$ for all x in $[0,1]$. However, f_2 cannot extend to a continuous function on $[0,1]$, since such an extension would be bounded on $[0,1]$ (by the Extreme-Value Theorem), hence its restriction to $(0,1)$ would also be bounded on $(0,1)$, while the given function f is clearly *unbounded* on $(0,1)$.

One way to analyse the issue is to predict what the values of such g – if it exists – must be. Of course the value $g(x)$ of the hoped-for extension g is clear when $0 < x < 1$: for such x one must have $g(x) = f(x)$. Thus the issue becomes this: what about $g(0)$ and $g(1)$? Let us consider $g(1)$.

Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be any sequence in $(0,1)$ such that $\lim_{k\to\infty} x_k = 1$. Then the only possible value for $g(1)$ is $\lim_{k\to\infty} f(x_k)$. This, in turn, would require that for every such sequence, the corresponding sequence $f{\circ}\xi = (f(x_1), f(x_2), \dots f(x_k), \dots)$ must be convergent to some number; that is, it must be a Cauchy sequence.

The issue, of extending a given continuous function $f : X \to \mathbb{R}$ from its original domain X to some superset of that domain, arises with domains that are much more complicated than simple intervals; and in those more general situations the connection with Cauchy sequences is still crucial. This connection leads to the following question: If $\xi = (x_1, x_2, \dots x_k, \dots)$ is a Cauchy sequence in the set X, under what circumstances must the corresponding sequence $f{\circ}\xi = (f(x_1), f(x_2), \dots f(x_k), \dots)$ also be a Cauchy sequence in $\mathbb{R}$?

To get a feel for the situation, it helps to consider what must hold for f to *not* not have this property. The answer is that there must exist a Cauchy sequence $\xi = (x_1, x_2, \dots x_k, \dots)$ in X for which the corresponding sequence $f{\circ}\xi = (f(x_1), f(x_2), \dots f(x_k) \dots)$ is not a Cauchy sequence. The latter fact would imply that there must exist $\varepsilon_0 > 0$ such that for every N in $\mathbb{N}$, no matter how large, there exists m in $\mathbb{N}$ for which $|f(x_{N+m}) - f(x_N)| \ge \varepsilon_0$. The hypothesis that ξ is a Cauchy sequence, however, implies that x_{N+m} and x_N can be made arbitrarily close to each other by simply choosing N sufficiently large. Thus one can conclude that in this situation, then the following condition must hold:

Condition A There exists $\varepsilon_0 > 0$ such that for every $\delta > 0$, no matter how small, there exist points c and d in X such that $|d-c| < \delta$ but $|f(d) - f(c)| \ge \varepsilon_0$.

It follows that the negation of Condition A is *sufficient* to guarantee that if ξ is Cauchy then so is $f{\circ}\xi$. This negation can be written as follows:

Condition B For every $\varepsilon > 0$ there exists $\delta > 0$ such that for every pair of points c and d in X which satisfy $|d-c| < \delta$ one has $|f(d) - f(c)| < \varepsilon$.

If one accepts that it is important to know about such issues, then it makes sense to give a name to Condition B. This name is given in Definition (IV.3.3).

IV.3.3 Definition

(1) Let $f : X \to \mathbb{R}$ be a real-valued function whose domain is a nonempty subset X of $\mathbb{R}$. Suppose that f enjoys the following property on the set X:

For every $\varepsilon > 0$ there exists $\delta > 0$ such that for each c in X if d in X satisfies $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon$.

Then one says that the function f is **uniformly continuous on the set** X.

(2) More generally, suppose as above that $f : X \to \mathbb{R}$ is a function with domain X, and that Y is a nonempty subset of X. One says that f is **uniformly continuous on the subset** Y provided the restriction $f|_Y : Y \to \mathbb{R}$, with domain Y, is uniformly continuous on Y in the sense of Part (a).

IV.3.4 Examples

Let $f : X \to \mathbb{R}$, with domain $X = \mathbb{R} \setminus \{0\}$, be given by the formula $f(x) = 1/x$. Note that f is continuous on the set X; that is, for each $c \neq 0$ the function f is continuous at c.

(1) The function f is uniformly continuous on the interval $[1, +\infty)$. Indeed, note that if $c, d \geq 1$ then

$$|f(d) - f(c)| = \left|\frac{1}{d} - \frac{1}{c}\right| = \left|\frac{c - d}{d\,c}\right| \leq |c - d| = |d - c| \text{ since } |d\,c| \geq 1.$$

Let $\varepsilon > 0$ be given, and let $\delta = \varepsilon$. Then the preceding inequality implies that if d and c are in the set $[1, +\infty)$ and satisfy $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon$.

(2) More generally, let $a > 0$ be any positive real number. Then f is uniformly continuous on the interval $[a, +\infty)$. The simple verification is left as an exercise.

(3) The continuous function f is *not* uniformly continuous on the subset $(0, 1]$. Indeed, one calculates as above that if $0 < c, d \leq 1$ then, as before,

$$|f(d) - f(c)| = \left|\frac{1}{d} - \frac{1}{c}\right| = \left|\frac{c - d}{d\,c}\right|.$$

But now one has $1/(d\,c) \geq 1/c$. Thus,

$$|f(d) - f(c)| \geq \frac{|d - c|}{c} \text{ for all } d \text{ in } (0, 1]$$

In particular, let $d = c/2$, so that $|d - c| = c/2$ and $|f(d) - f(c)| = |f(c/2) - f(c)| \geq (c/2)/c = 1/2$.

Now let $\varepsilon = 1/2$. There is no single $\delta > 0$ such that $|d - c| < \delta$ implies that $|f(d) - f(c)| < 1/2$ for *all* choices of d and c such that $0 < d, c \leq 1$. Indeed , suppose that such δ did exist. Choose $c = \min\{1, \delta\}$ and choose $d = c/2$. Then by the preceding analysis one would have $0 < c \leq 1$ and $0 < d < 1$, so that $|d - c| = c/2 < \delta$. By the (purported) construction of δ, this would imply that $|f(c/2) - f(c)| < 1/2$, contrary to what was obtained above.

(4) Let $g : \mathbb{R} \to \mathbb{R}$ be the 'squaring function'; that is, $g(x) = x^2$ for all x in $\mathbb{R}$. It is an instructive exercise to show that g is not uniformly continuous on $[1, +\infty)$.

IV.3.5 Remark

The statement of the preceding definition looks very much like the formulation, in Remark (IV.1.3) (2) above, of the condition for f to be continuous on X. Indeed, many students think the two formulations are equivalent ways of stating this condition.

In reality, they are *not* equivalent. It is true that being uniformly continuous on a set does imply being continuous on that set. However, Example (3) above illustrates the fact that the converse is not true.

The key to this difference in the meanings of these concepts is the location of the phrase 'there exists $\delta > 0$': in the formulation of Remark (IV.1.3), this phrase comes after the phrase 'for every $\varepsilon > 0$ *and* after the phrase 'for every c'. This means that the choice of δ depends on the previous choices of both ε and c. In the current definition, however, the phrase 'there exists $\delta > 0$' comes after the phrase 'for every $\varepsilon > 0$' but *before* the phrase 'for every c'. This means that, given a choice of ε one needs a choice δ, still depending on ε, but one which works *simultanously* for all c; that is, given the choice of ε, it requires that a 'uniform' choice of δ can be made which works for all c. This is why the word 'uniform' is included.

For future reference it is worth writing down more formally a fact which is alluded to in the preceding remark; its trivial proof is omitted.

IV.3.6 Theorem

If $f : X \to \mathbb{R}$ is uniformly continuous on X, then f is continuous on X. Otherwise stated: Uniform continuity on a set implies continuity on that set.

> Side Comment (on students' confusion over Theorem (IV.3.6)) Some students get confused over the fact that Theorem (IV.3.6) needs to be stated at all, much less proved. They reason, in effect, that the hypothesis 'f is uniformlly continuous' *means* that f is continuous – so it is pointless to even state this conclusion – but continuous in a special way, namely 'uniformly'. From this point of view it is not even necessary to know the meaning of 'continuous' to agree that the conclusion is correct. In the context of ordinary language, it is similar to saying 'The flower is bright pink'; one does not need to then assert 'Therefore the flower is pink', and one need not even know what 'bright' or 'pink' means, just their grammatical roles as adverb and adjective.
>
> The cause of the confusion, however, is that mathematical usage is often technical, and does not follow the usage of ordinary language. For example, the phrase 'uniformly continuous' has the same grammatical *format* as the phrase 'bright pink': an adverb modifying an adjective. However, the former phrase is defined technically, as a single unit, by Definition (IV.3.3), and not as the grammatical juxtaposition of an adverb (uniformly) and an adjective (continuous), each with their own previously defined meaning.
>
> One encounters similar confusion in elementary calculus; a good example appears there in the treatment of 'infinite series'. (If you are not familiar with this topic, simply skip this paragraph.) Indeed, one learns to distinguish between a series which is 'absolutely convergent' and a series which is 'conditionally convergent'. In the former case one proves a theorem which says that an absolutely convergent series is convergent; this leads calculus students to the same grammatical type of confusion as was just discussed concerning 'uniformly continuous'. In contrast, one does *not* prove that a conditionally convergent series is convergent, because the very *definition* of 'conditionally convergent' includes the hypothesis that the series in question is convergent. That is, the phrase 'conditionally convergent' *does* refer to a series which is

convergent, but convergent in a special way. (What actually confuses students here instead is why the adverb ‘conditionally’ is chosen to describe this special type of convergence.)

IV.3.7 Theorem (The Uniform-Continuity Theorem in $\mathbb{R}$)

Suppose that $f : X \to \infty$ is a continuous real-valued function on a nonempty closed and bounded subset X of $\mathbb{R}$. Then f is uniformly continuous on X.

Proof Suppose that the statement is *not* true for some f and X satisfying the hypotheses. Then there exists $\varepsilon_0 > 0$ such that for every $\delta > 0$ there exist c and d in X with the property that $|d - c| < \delta$ but $|f(d) - f(c)| \geq \varepsilon_0$; of course both c and d depend on δ. In particular, for each k in $\mathbb{N}$ let $\delta_k = 1/k$. Then there must exist c_k and d_k in X such that

$$|d_k - c_k| < 1/k \text{ but } |f(d_k) - f(c_k)| \geq \varepsilon_0 \qquad (*)$$

Since X is bounded, the standard Bolzano-Weierstrass Theorem implies that the sequence $\psi = (c_1, c_2, \dots c_k, \dots)$ must have a subsequence $(c_{k_1}, c_{k_2}, \dots c_{k_m}, \dots)$ which converges to some point p, which must be in X by the hypothesis that X is closed. Since $|d_{k_m} - c_{k_m}| < 1/k_m$ for each m, it follows easily from the Squeeze Theorem for real sequences that the sequence $(d_{k_1}, d_{k_2}, \dots d_{k_m}, \dots)$ also converges to p. It now follows from the continuity that $\lim_{m \to \infty} f(d_{k_m}) = \lim_{m \to \infty} f(c_{k_m}) = f(p)$. However, this contradicts Statement $(*)$ above, so the desired result follows.

Remark Many texts attribute the first treatment of ‘Uniform Continuity’ and the ‘Uniform Continuity Theorem in $\mathbb{R}$’ to the German mathematician E. Heine around 1871, and refer to Theorem (IV.3.7) as ‘Heine’s Theorem’. At various times, however, other names have been associated with these results, including Cantor, Borel, Schoenflies, Cousin and Lebesgue.

Historians now seem to agree that the concept and theorem were first clearly presented by Dirichlet in his lecture notes from around 1854. (These lecture notes were first published, however, only in 1904, so the confusion is understandable.) The name ‘Uniform Continuity Theorem’, given above to Theorem (IV.3.7), is not used uniformly in analysis (no pun intended), but it does have the advantage of being more descriptive than, say, the ‘Dirichlet-Heine-Cantor-Borel-Schoenflies-Cousin Theorem’.

Most of the applications of the Uniform Continuity Theorem appear later in *This Textbook*, but here is a simple one.

IV.3.8 Theorem

Suppose that $f : X \to \mathbb{R}$ is a uniformly continuous function on a nonempty subset X of $\mathbb{R}$.

(a) If $\xi = (x_1, x_2, \dots x_k, \dots)$ is a Cauchy sequence in X, then $f \circ \xi = (f(x_1), f(x_2), \dots f(x_k), \dots)$ is a Cauchy sequence in $\mathbb{R}$.

(b) Suppose that $\xi = (x_1, x_2, \dots x_k, \dots)$ and $\tau = (t_1, t_2, \dots t_k, \dots)$ are Cauchy sequences in X which both converge to the same number L in $\mathbb{R}$. Then the corresponding sequences $f \circ \xi$ and $f \circ \tau$ of values both converge to the same value.

Proof (a) Let $\varepsilon > 0$ be given, and let $\delta > 0$ be small enough that if c and d are in X such that $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon$. (Such δ exists by the hypothesis of uniform continuity.) Next, let N in $\mathbb{N}$ be large enough that $|x_{N+m} - x_N| < \delta$ for all m in $\mathbb{N}$. (Such N exists by the

'Cauchy' hypothesis.) Combining all this then yields $|f(x_{N+m}) - f(x_N)| < \varepsilon$ for all m in $\mathbb{N}$. It follows that $(f \circ \xi)$ is a Cauchy sequence, as required.

(b) It is given that ξ and τ both converge to the number L. Let $\sigma = (s_1, s_2, \dots s_m, \dots)$ be the sequence $(x_1, t_1, x_2, t_2, \dots)$. That is, $s_{2k-1} = x_k$ and $s_{2k} = t_k$ for each index k. It follows from the Odd/Even Limit Theorem (see Theorem (III.2.14)) that the sequence σ is also convergent to L, and thus, by the Cauchy Convergence Theorem, Theorem (III.5.5), σ is a Cauchy sequence. By Part (a) of the present theorem it then follows that $f \circ \sigma$ is a Cauchy sequence, and thus is convergent. By Theorem (III.2.1) it then follows that the subsequences $f \circ \xi$ and $f \circ \tau$ of $f \circ \sigma$ also converge to the same limit, as required.

IV.4 Continuity Theorems on Intervals in $\mathbb{R}$

Preliminary Remark In elementary calculus one considers, almost exclusively, real-valued functions defined on intervals in $\mathbb{R}$. This contrasts with the situations considered so far in this chapter, in which the domains can be arbitrary nonempty subsets of $\mathbb{R}$. The present section focuses on theorems which make sense only for functions defined on real intervals.

IV.4.1 Theorem (The Intermediate-Value Theorem – Standard Form)

Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function on a closed bounded interval $[a, b]$ in $\mathbb{R}$, and $f(a) \neq f(b)$. Then for every number y_0 strictly between $f(a)$ and $f(b)$ there exists at least one number x_0, with $a < x_0 < b$, such that $f(x_0) = y_0$.

Otherwise stated: If a number y_0 is intermediate between $f(a)$ and $f(b)$, then y_0 is a value of f on the interval $[a, b]$. (This explains the name of the theorem.)

Proof Without loss of generality assume that $f(a) < f(b)$, and define a bisection sequence $\mathcal{A} = ([a_1, b_1], [a_2, b_2], \dots [a_k, b_k], \dots)$, with initial interval $[a_1, b_1] = [a, b]$, as follows:

Step 1 Let c_1 be the midpoint of the interval $[a, b] = [a_1, b_1]$. If $f(c_1) \leq y_0$, then set $a_2 = c_1$ and $b_2 = b_1$; if, instead, $f(c_1) > y_0$, then set $a_2 = a_1$ and $b_2 = c_1$. In either case, one has $f(a_2) < f(b_2)$.

General Step Suppose that $[a_1, b_1], \dots \ [a_m, b_m]$, with $m \geq 2$, have been constructed so that for each $j = 1, \dots m-1$ the interval $[a_{j+1}, b_{j+1}]$ is one of the halves of the interval $[a_j, b_j]$ and one has $f(a_m) < f(b_m)$. Now let c_m be the midpoint of $[a_m, b_m]$, and define $a_{m+1} = c_m$ and $b_{m+1} = b_m$ if $f(c_m) \leq y_0$, while $a_{m+1} = a_m$ and $b_{m+1} = c_m$ if $f(c_m) > y_0$.

The resulting bisection sequence $[a_1, b_1] \supseteq [a_2, b_2] \supseteq \dots \supseteq [a_m, b_m] \supseteq$ then satisfies the condition $f(a_m) \leq y_0 < b_m$ for each index m. The Bisection Principle then implies that $\lim_{m \to \infty} a_m = \lim_{m \to \infty} b_m = c$, where c is the unique number which lies in each of the intervals of the bisection principle. It follows that $\lim_{m \to \infty} f(a_m) = f(c) = \lim_{m \to \infty} f(b_m)$. Since $f(x_m) \leq y_0 \leq f(b_m)$ for all m, it follows that $f(c) \leq y_0 \leq f(c)$; that is, $f(c) = y_0$, and the desired result follows.

Remark The proof given here is in essence the same as the proofs of Bolzano and Cauchy; see [BOLZANO 1817] and [CAUCHY 1821].

Example If $f(x) = x^2$ and $y_0 = 2$, then the procedure described above is essentially the same as that used to compute $\sqrt{2}$ in Example (II.4.18).

IV.4.2 Corollary (Intermediate-Value Theorem, Extended Form)

Suppose that $f : I \to \mathbb{R}$ is a continuous real-valued function defined on an interval I in $\mathbb{R}$; the interval I need not be bounded or closed. Let $S = f[I]$ be the image of I under f.

Let $A = \inf S$ and $B = \sup S$. If y_0 is any number such that $A < y_0 < B$, then there exists a number x_0 in I such that $f(x_0 = y_0)$.

Equivalently, the image $f[I]$ of f on I is either a singleton set (if f is constant on I), or an interval (if f is nonconstant on I).

Proof If $A = B$, so that f is constant on I, then I is the singleton $\{A\}$. In this case the hypothesis $A < y_0 < B$ is never satisfied, so the conclusion is automatically true. Thus, suppose, instead, that $A < B$, so that f is not constant on I. Let y_0 be such that $A < y_0 < B$. It follows from the approximation properties for infima and suprema that there exist numbers c and d in I such that $A < c < y_0 < d < B$. Since c and d are in I, it follows that there exist numbers a and b in I such that $c = f(a)$ and $d = f(b)$. Since $c \neq d$ it follows, from the basic definition of 'function' that $a \neq b$. It now follows from the preceding theorem that there exists x_0 between a and b, so that x_0 is an element of the interval I, such that $f(x_0) = y_0$, as required.

The fact that $f[I]$ is either a singleton set or an interval now follows easily.

The property described in the previous results is given a name.

IV.4.3 Definition

Let $f : I \to \mathbb{R}$ be a real-valued function whose domain is an interval I.

(1) One says that f has the **weak intermediate-value property on** I provided that if a and b are any numbers in I such that $f(a) \neq f(b)$, then for each number y_0 strictly between $f(a)$ and $f(b)$ there exists x_0 in I such that $f(x_0) = y_0$.

(2) One says that f has the **strong intermediate-value property on** I provided that if a and b are any numbers in I such that $f(a) \neq f(b)$, then for each number y_0 strictly between $f(a)$ and $f(b)$ there exists x_0, strictly between a and b, such that $f(x_0) = y_0$.

Remarks(1) To say that $f : I \to \mathbb{R}$ has the weak intermediate-value property on an interval I is equivalent to saying that the image set $f[I]$ is either a singleton set (if f is constant on I), or else $f[I]$ is is an interval in $\mathbb{R}$. Likewise, to say that $f : I \to \mathbb{R}$ has the strong intermediate-value property on I is equivalent to saying that for each pair of numbers a and b in I with $a < b$ the image of $[a, b]$ under f is either a singleton or an interval.

(2) It is clear that if a function has the strong intermediate-value property on an interval, then it certainly also has the weak intermediate-value property on that interval.

Examples (1) The Intermediate-Value Theorem implies that every function which is continuous on an interval I has the strong intermediate-value property on I.

(2) Consider the function $f : [0, 1] \to \mathbb{R}$ given by the rule

$$f(x) = x \text{ if } 0 < x < 1;\ f(0) = 1,\ f(1) = 0.$$

It is easy to verify that this function has the weak intermediate-value property on the interval $[0, 1]$, but not the strong.

As late as the mid-nineteenth century many mathematicians believed that if a function had the strong intermediate-value property on an interval, then it must be continuous on that interval. In 1875 the French mathematician G. Darboux gave an example of a function $f : [0, 1] \to \mathbb{R}$ which has the strong intermediate-value property on $[0, 1]$ but is discontinuous at 0. Much later the English mathematician J. Conway gave an example of a function $g : [0, 1] \to \mathbb{R}$ which also has the strong intermediate-value property but is discontinuous at every point of $[0, 1]$.

The next definition and theorem suggest that such examples must be fairly complicated.

IV.4.4 Definition

A function $f : [a, b] \to \mathbb{R}$ defined on a closed bounded interval $[a, b]$ is said to be **piecewise monotonic on** $[a, b]$ provided either

(i) f is monotonic on $[a, b]$; or

(ii) there exists a partition $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ of $[a, b]$ such that f is monotonic on each closed subinterval of the form $[x_{j-1}, x_j]$ for $j = 1, 2, \ldots n$.

Roughly speaking, f can be 'patched together' with finitely many monotonic functions. It is allowed that some of these functions be monotonic up, while the others are monotonic down.

Remark Note that Example (2) above is *not* piecewise monotonic on $[0, 1]$.

IV.4.5 Theorem

Let $I = [a, b]$ be a closed bounded interval in $\mathbb{R}$.

(a) Suppose that $f : I \to \mathbb{R}$ is a monotonic function on the interval I. Then f is continuous on I if, and only if, it has the weak intermediate-value property on I. The result remains true if 'weak' is replaced by 'strong'.

(b) Suppose that $g : I \to \mathbb{R}$ is a function such that, for each a and b in I with $a < b$, g is piecewise monotonic on $[a, b]$. Then g is continuous on I if, and only if, it has the strong intermediate-value property on I.

The simple proof is left as an exercise.

Example For each n in $\mathbb{N}$ let P_n be the ordered pair $\left(\frac{1}{n}, (-1)^{n-1}\right)$. Thus, $P_1 = (1, 1)$, $P_2 = (1/2, -1)$, and so on; speaking geometrically, if n is odd then P_n is the point in the Euclidean plane with abscissa $1/n$ and ordinate 1, while if n is even then P_n is the point with abscissa $1/n$ and ordinate -1.

Let $I = (0, 1]$, and define a function $g : I \to \mathbb{R}$ as follows: if $0 < x \leq 1$, let n be the largest natural number such that $\frac{1}{n+1} < x \leq \frac{1}{n}$. Then set $g(x)$ equal to the value of the linear function which interpolates the points P_{n+1} and P_n. It is easy to see that if $0 < a < b \leq 1$, then the function g is continuous, and thus has the strong intermediate-value property, on $[a, b]$. It follows from the preceding theorem that g is also continuous on I, and thus has the strong intermediate-value property on I. Now let $f : [0, 1] \to \mathbb{R}$ be given by $f(x) = g(x)$ if $0 < x \leq 1$, and $f(0) = 0$. It is easy to see that f also has the strong intermediate-value property on $[0, 1]$, but is *not* continuous at 0.

Remarks (1) The preceding example is a simplified version of Darboux's original example.

(2) For those whose French pronunciation is weak: the 'boux' in 'Darboux' is pronounced like the English word 'boo'.

There are several other important results which hold for continuous monotonic functions.

IV.4.6 Theorem

(a) Suppose that $f : I \to \mathbb{R}$ is a function which is continuous on an interval I which has at least one endpoint. Let J be the set of interior points of I. If the restriction of f to J is monotonic on J, then f is monotonic on I.

(b) If the word 'monotonic' in Part (a) is replaced throughout by the phrase 'strictly monotonic', the resulting statement remain true.

Proof The simple proof using the preceding theorem is left as an exercise.

If the function $f : I \to \mathbb{R}$ is known to be both continuous *and* strictly monotonic, then even more can be proved.

IV.4.7 Theorem

Suppose that $f : I \to \mathbb{R}$ is a continuous strictly monotonic function whose domain is an interval I. Let $J = f[I]$ be the image of f; note that the 'strictly monotonic' hypothesis implies that f is a bijection of I onto J, so by Part (b) of the preceding corollary the image set J is some type of interval in $\mathbb{R}$. Then:

(a) The intervals I and J contain the same number of endpoints as each other. That is, they both contain two endpoints, or both contain one endpoint, or both contain no endpoints.

(b) The function f is a bijection of I onto J, and the inverse function $f^{-1} : J \to I$ is continuous on J.

Proof

For brevity consider only the case in which I is an open interval (a, b) and f is strictly increasing. Only small changes of the argument, which are left as exercises, are need to handle the other cases.

(a) Let $A = \inf\{f(x) : a < x < b\}$, and let $B = \sup\{f(x) : a < x < b\}$. Then J is the open interval (A, B). Indeed, Part (b) of Corollary (IV.4.2) implies that the open interval (A, B) is a subset of J, and that the only way these intervals can not be equal is if J contains either A or B. However, it is impossible to have A be in the set J. Indeed, if this were to happen, then A would equal $f(x)$ for some x such that $a < x < b$. Let x' be a number such that $a < x' < x$, so that x' is also in I. Then the strict monotonicity would imply that $f(x') < f(x) = A$, which would contradict A being a lower bound for the values of f. Similarly, it is impossible for B to be in J, so $J = (A, B)$, as claimed.

(b) Since $f : (a, b) \to (A, B)$ is a bijection, it follows that the inverse function $f^{-1} : (A, B) \to (a, b)$ exists. It is clear that g is strictly increasing on (A, B), and its image is the interval (a, b). It then follows from Part (a) of Theorem (IV.4.5), appied to the function g, that g is continuous on (A, B), as claimed.

IV.4.8 Examples

(1) Let n be a natural number, and let $f : (0,+\infty) \to \mathbb{R}$ be the function, with domain $(0,\infty)$, given by the rule $f(x) = x^n$. Clearly f is continuous and strictly increasing on $(0,+\infty)$, and its inverse function $g : (0,+\infty) \to (0,+\infty)$ is the corresponding 'positive n-th root' function:

$$g(x) = \sqrt[n]{x} \text{ for } 0 < x < +\infty.$$

It follows from the preceding theorem that this function is continuous on $(0,+\infty)$.

(2) More generally, suppose that $r = m/n$ is a rational number, with m and n in $\mathbb{Z}$ and $n \geq 1$. Define the function $h : (0,+\infty) \to \mathbb{R}$ by the rule

$$h(x) = x^r = (\sqrt[n]{x})^m \text{ for } 0 < x < +\infty.$$

It is clear that is also continuous. Indeed, it is the composition of the 'm-th power' function with the function discussed in Example (1).

Remark It is a simple exercise to show that if the rational number r is expressed in a second way in the form $r = m'/n'$, with m' and n' in $\mathbb{Z}$ and $n' \geq 1$, then the resulting value of $h(x)$ is the same, at least for $0 < x < +\infty$.

The next result is almost never included in a course in elementary calculus; but we see in Chapter (V) that it is useful in that context.

IV.4.9 Theorem (The Piecewise-Linear Interpolation Theorem)

Suppose that $f : [a,b] \to \mathbb{R}$ is a real-valued function which is continuous on the closed bounded interval $[a,b]$. Let $\varepsilon > 0$ be given. Then there exists a continuous piecewise-linear function $g : \mathbb{R} \to \mathbb{R}$ on $\mathbb{R}$ such that $|f(x) - g(x)| < \varepsilon$ for every x in $[a,b]$.

Proof Let $a = x_0 < x_1 < \ldots < x_{k-1} < x_k = b$ be any partition of the interval $[a,b]$ into k subintervals; see Definition (II.3.1). Let $g : \mathbb{R} \to \mathbb{R}$ be the continuous piecewise-linear function such that $g(x_j) = f(x_j)$ for each $j = 0,1,2,\ldots k$; see Example (IV.1.4) (6). If x is any element of the interval $[a,b]$ we need to relate the quantity of interest, namely $|f(x) - g(x)|$, to the given information about f and g.

First note that there exists an index j such that $x_{j-1} \leq x \leq x_j$. In this interval one knows the formula for $g(x)$, given in Example (I.5.6) (6). Furthermore, the fact that f is continuous on the closed bounded set $[a,b]$, hence *uniformly* continuous there, implies that one can make $|f(u) - f(x_{j-1})|$ as small as one needs for each u in $[x_{j-1}, x_j]$ by having $|x_j - x_{j-1}|$ sufficently small. The connection between this data can be obtained, once again, by a clever use of the 'Add-and-Subtract Trick', combined with the Triangle Inequality. More precisely, if $x_{j-1} \leq x \leq x_j$, then one has

$$f(x)-g(x) = (f(x)-f(x_{j-1}))+(f(x_{j-1})-g(x)) = (f(x)-f(x_{j-1}))-\left(\frac{f(x_j)-f(x_{j-1})}{x_j-x_{j-1}}\right)(x-x_{j-1}).$$

(Note that this calculation uses the formula for linear interpolation between two points, together with the fact that $g(x_i) = f(x_i)$ for each $i = 0,1,\ldots k$.) Thus

$$|f(x)-g(x)| \leq |f(x)-f(x_{j-1})|+\left|\frac{f(x_j)-f(x_{j-1})}{x_j-x_{j-1}}\right||x-x_{j-1}| \leq |f(x)-f(x_{j-1})|+|f(x_j)-f(x_{j-1})| \qquad (*)$$

since $|x - x_{j-1}| \leq |x_j - x_{j-1}|$.

Now let $\varepsilon > 0$ be given. Let $\delta > 0$ be small enough that if c and d are any points of $[a, b]$ such that $|d - c| < \delta$, then $|f(d) - f(c)| < \varepsilon/2$; such δ exists because f is uniformly continuous on $[a, b]$. Choose the partition points x_j so that $|x_j - x_{j-1}| < \delta$ for each j; for example, choose k large enough that $|b - a|/k < \delta$, and for that k set $x_j = a + j\left(\dfrac{b-a}{k}\right)$. It then follows from Inequality $(*)$ above that if $x \in [a, b]$, then, as required,

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

IV.5 Limits and Continuity as in Elementary Calculus

As has been mentioned in the Introduction to this chapter, the way the topics of 'limits' and 'continuity' are treated in elementary calculus is somewhat different from the approach taken in more advanced analysis books. It is important to know both approaches, especially for the following chapter.

IV.5.1 Definition

Let c and L be extended real numbers.

(1) Suppose that f is a real-valued function which is defined on some open interval of the form (a, c), where $-\infty \leq a < c \leq +\infty$. One says that ***f* has left-hand limit *L* at *c*** provided that for every strictly increasing sequence $\xi = (x_1, x_2, \dots x_k, \dots)$ of reals such that $\lim_{k\to\infty} x_k = c$ one has $\lim_{k\to\infty} f(x_k) = L$. Then one writes $L = \lim_{x\to c-} f(x)$ or $\lim_{x\nearrow c} f(x) = L$; in spoken form, '$f(x)$ approaches L as x approaches c from below', or '$f(x)$ approaches L as x approaches c from the left'.

(2) Suppose, instead, that f is a real-valued function which is defined on some open interval of the form (c, b), where $-\infty \leq c < b \leq +\infty$. One says that ***f* has right-hand limit *L* at *c*** provided that for every strictly decreasing sequence $\xi = (x_1, x_2, \dots x_k, \dots)$ of reals such that $\lim_{k\to\infty} x_k = c$ one has $\lim_{k\to\infty} f(x_k) = L$. Then one writes $L = \lim_{x\to c+} f(x)$ or $\lim_{x\searrow c} f(x) = L$; in spoken form, '$f(x)$ approaches L as x approaches c from above', or '$f(x)$ approaches L as x approaches c from the right'.

(3) Finally, suppose that c is finite and that f is defined on open intervals of the form (a, c) and (c, b), where $-\infty \leq a < c < b \leq +\infty$. If both of the equations $\lim_{x\to c-} f(x) = L$ and $\lim_{x\to c+} = L$ are true, as described above, then one says that ***f* has two-sided limit *L* at *c***, or, more simply, that ***f* has limit *L* at *c***, and one writes $L = \lim_{x\to c} f(x)$.

Special Cases If c is the supremum of the domain of f, so that the expression $\lim_{x\to c+} f(x)$ makes no sense, then one often abbreviates the equation $\lim_{x\to c-} f(x) = L$ to $\lim_{x\to c} f(x) = L$. Likewise, if c is the infimum of the domain of f, then one often abbreviates the equation $\lim_{x\to c+} f(x) = L$ to $\lim_{x\to c} f(x) = L$. In particular, these conventions hold if $c = \pm\infty$. These 'abuses of notation' generally cause no problems.

IV.5.2 Remarks

(1) We often write $f(c-)$ for the quantity $\lim_{x \to c-} f(x)$ provided both c and this limit are finite. Likewise, we often write $f(c+)$ for the quantity $\lim_{x \to c+}$ provided both c and this second limit are finite. Of course, using either expression does not require that f b defined at c.

(2) In Part (1) of the preceding definition we do not require that the open interval (a, c) be the full domain of f or that x_k be in (a, c) for each index k; it suffices that this hold for all but a finite number of the indices. In effect, we are using the 'Extended Definition' of limits given in Definition (III.1.10). A similar comment holds for Part (2).

(3) In Part (1) of the preceding definition it is required that f be defined on *some* open interval of the form (a, c). Of course if that is true, then f is also defined on any open interval of the form (a', c), where $a \leq a' < c$. It may also happen that f is defined on an interval (a'', c) where $a'' < a$. It is clear from the preceding remark, however, combined with Part (b) of Theorem (IV.2.1), that neither the existence of the purported limit L, nor its value, depends on the particular choice of this open interval. A similar comment holds for the open interval (c, b) in Part (2) of the definition.

(4) If L is finite, then one often expresses the equation $\lim_{x \to c-} f(x) = L$ as something like 'the function f **converges** to L from the left'. Similar remarks hold for the equations $\lim_{x \to c+} f(x) = L$ and $\lim_{x \to c} f(x) = L$ when L is finite. This practice agrees with the usage with sequences, for which one uses the word 'converge' only when the limiting value is finite.

The restriction to strictly monotonic sequences in the preceding definition is mainly to aid the intuition dealing with one-sides limits. The following variation is often used in place of Parts (1) and (2).

IV.5.3 Theorem

Let c and L be extended real numbers.

(a) Suppose that f is a real-valued function defined on some open interval of the form (a, c). A necessary and sufficient condition that the equation $L = \lim_{x \to c-} f(x)$ hold, in the sense of the preceding definition, is that for every sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ of reals such that $x_k < c$ for each k and $\lim_{k \to \infty} x_k = c$ one has $\lim_{k \to \infty} f(x_k) = L$.

(b) Suppose, instead, that f is a real-valued function which is defined on some open interval of the form (c, b). A necessary and sufficient condition that the equation $L = \lim_{x \to c+} f(x)$ hold, in the sense of the preceding definition, is that for every sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ of reals such that $x_k > c$ for each k and $\lim_{k \to \infty} x_k = c$ one has $\lim_{k \to \infty} f(x_k) = L$.

(c) Finally, suppose that c is finite f is a real-valued function defined on some open interval containing c, except possibly at c itself. A necessary and sufficient condition that the equation $L = \lim_{x \to c} f(x)$ hold, in the sense of the preceding definition, is that for every sequence $\xi = (x_1, x_2, \ldots x_k, \ldots)$ of reals such that for each k $x_k \neq c$ and $\lim_{k \to \infty} x_k = c$ one has $\lim_{k \to \infty} f(x_k) = L$.

The proof is left as an exercise.

In elementary calculus the 'limit' concept is normally formulated without the uses of sequences. The next result summarizes that approach.

IV.5.4 Theorem

(a) The equation $\lim_{x\to c-} f(x) = L$ in Part (a) of the preceding definition is equivalent to the following pair of statements being true:

Statement A– For each real number y such that $y < L$, there exists a number a, with $a < c$, such that $y < f(x)$ for all x such that $a < x < c$.

Staement B– For each real number z such that $L < z$, there exists a number a, with $a < c$, such that $f(x) < z$ for all x such that $a < x < c$.

(b) The equation $\lim_{x\to c+} f(x) = L$ in Part (b) of the preceding definition is equivalent to the following pair of statements being true:

Statement A+ For each real number y such that $y < L$, there exists a number b, with $c < b$, such that $y < f(x)$ for all x such that $c < x < a$.

Statement B+ For each real number z such that $L < z$, there exists a number b, with $c < b$, such that $f(x) < z$ for all x such that $c < x < b$.

The simple proof is left as a simple application of Theorem (III.1.9).

IV.5.5 Remarks

(1) The phrasing of this theorem is slightly complex because it includes all the possibilities for the extended reals c and L: c can be finite or one of the infinities, as can L. Definition (IV.5.1) hides these complexities under the previously defined concept of a sequence of real numbers having a (possibly infinite) limit; see Theorem (III.1.9).

(2) If L is an upper bound for the image of f, then Statements B– and B+ are automatically true. Likewise, if L is a lower bound for the image of f, then Statements A– and A+ are automatically true. (Compare with Remark (**??**).)

(3) In elementary calculus the normal procedure is to formulate the meaning of the two-sided limit '$\lim_{x\to c} f(x) = L$' directly, without first defining one-sided limits.

(4) Let $\xi = (x_1, x_2, \ldots x_k, \ldots)$ be a sequence of real numbers. Recall that in the 'official' definition of 'sequence', ξ is actually a real-valued function with domain $\mathbb{N}$; with that viewpoint, x_k is simply an alternate notation for the function value $\xi(k)$. The limit of ξ, if it has one, then can be written as $\lim_{k\to\infty} \xi(k)$.

Suppose now that for some strange reason we decide to label the terms of this sequence using the letter x instead of the letter k, and to use the English letter f instead of the Greek letter ξ as the 'name' of this sequence. Then the notation for the sequential limit just written would change to $\lim_{x\to\infty} f(x)$.

Would this change of notation be legal? Absolutely yes. Would it be wise? Absolutely no, because it would be confusing to the reader: the latter notation looks like that introduced in Definition (IV.5.1) above. In *This Textbook* we shall try to make it clear which version of 'limit' is meant; in particular, we shall avoid such 'unwise' changes of notation.

IV.5.6 Corollary

Let c and L be real (i.e., finite) numbers.

(a) The equation $\lim_{x\to c-} f(x) = L$ in Part (a) of Definition (IV.5.1) is equivalent to the following condition: for every $\varepsilon > 0$ there is $\delta > 0$ such that if $c-\delta < x < c$, then $|f(x)-L| < \varepsilon$.

(b) The equation $\lim_{x \to c+} f(x) = L$ in Part (b) of Definition (IV.5.1) is equivalent to the following condition: for every $\varepsilon > 0$ there is $\delta > 0$ such that if $c < x < c+\delta$, then $|f(x)-L| < \varepsilon$.

The simple proof is left as an exercise.

Remark Note that in this formulation one still does not require that the function f be defined at the point c. Indeed, the requirements on x include $x < c$ in Part (a) and $x > c$ in Part (b).

The fact that we have already proved numerous results about limits of sequences makes it much easier to prove analogous results for limits of functions on intervals. The next three theorems illustrate this statement. In order to simplify the exposition, these results focus on 'two-sided' limits, which is the case that arises most frequently in practice; the reader is given the task to formulate the corresponding 'one-sided' results, as well as to prove most portions of these results. That is, we consider limits of the form $\lim_{x \to c} f(x)$, where c is finite, and the real-valued function f is defined at every point of some open interval (a,b) containing c except, possibly, at c itself. For convenience, write $X_c = (a,b) \setminus \{c\}$, so that f is defined on X_c.

IV.5.7 Theorem

Let $f : X_c \to \mathbb{R}$ be as above.

(a) If the function f has a limit at c, then this limit is unique. That is, if the statements $\lim\limits_{x \to c} f(x) = L_1$ and $\lim\limits_{x \to c} f(x) = L_2$ are both true, then $L_1 = L_2$.

(b) If there is a real number A such that $f(x) = A$ for all x in X_c, then $\lim\limits_{x \to c} f(x) = A$.

(c) Suppose that there exist real numbers m and M, with $m < M$, such that $f(x) \in [m, M]$ for all x in the set X_c. If f has a limit at c, then $\lim\limits_{x \to c} f(x)$ is also in the set $[m, M]$.

(d) (Squeeze Property for Limits on Intervals) Let f, g and h be real-valued functions defined on the set X_c, and let L be an extended real number. Suppose that the following conditions hold:
(i) $g(x) \in \mathrm{Seg}\,[f(x), h(x)]$ for all x in X_c.
(ii) $\lim\limits_{x \to c} f(x) = \lim\limits_{x \to c} h(x) = L$.
Then $\lim\limits_{x \to c} g(x) = L$.

(e) Suppose that the quantity L is a real number (i.e., L is finite). Then the following statements are equivalent:
(i) $\lim\limits_{x \to c} f(x) = L$;
(ii) $\lim\limits_{x \to c} |L - f(x)| = 0$;
(iii) $\lim\limits_{x \to c} (L - f(x)) = 0$.

(f) Suppose that $\lim\limits_{x \to c} f(x) = 0$. If g is real-valued function which is bounded on the set X_c, then $\lim\limits_{x \to c} f(x) \cdot g(x) = 0$.

IV.5.8 Theorem

Using the same notation as above, let $f : X_c \to \mathbb{R}$ be a real-valued function defined on the set $X_c = (a,b) \setminus \{c\}$.

(a) (Constant Factor Rule) Suppose that $\lim\limits_{x \to c} f(x) = L$ for some real number L. Then for every real number A one has $\lim\limits_{x \to c} (A \cdot f)(x) = A \cdot L$.

(b) (Absolute-value Rule) Suppose that $\lim\limits_{x\to c} f(x) = L$, where L is an extended real number. Then $\lim\limits_{x\to c} |f(x)| = |L|$.

(c) Suppose that $\lim\limits_{x\to c} f(x) = L$, where L is an extended real number. Also, assume that $L \neq 0$. Then there exists $\delta > 0$ such that if $0 < |c - x| \leq \delta$ then $f(x) \neq 0$.

More precisely:

Case (i) If $L > 0$ and m is any number such that $0 < m < L$, then there exists $\delta > 0$ so that $f(x) \geq m$ for all x such that $0 < |x - c| \leq \delta$.

Case (ii) If $L < 0$ and $m > 0$ is any number such that $L < -m < 0$, then there exists $\delta > 0$ so that $f(x) \leq -m$ for all x such that $0 < |x - c| \leq \delta$.

Proof of Case (i) of Part (c) Suppose that $L > 0$ and that the conclusion in (c) is *not* true for a given f. Then there would have to exist a number m, with $0 < m < L$, such that the following holds: for every k in $\mathbb{N}$ there exists a number x_k in the set X_c such that $0 < |c - x_k| < 1/k$ and $f(x_k) < m$. Let $\xi = (x_1, x_2, \dots)$ be the corresponding infinite sequence. Since, by construction, one has $|c - x_k| < 1/k$, it is clear that the sequence ξ converges to c. Since x_k is never equal to c, it follows that ξ does not have a constant subsequence. Theorem (III.4.6), the 'Monotonic-Subsequences Theorem', implies that ξ has a strictly monotonic subsequence $\zeta = (z_1, z_2, \dots z_n, \dots)$ which converges to c. It follows from Definition (IV.5.1) that $\lim_{n\to\infty} f(z_m) = L$. However, one also has, by construction, $f(x_k) < m$ for each k, hence $f(z_n) < m$ for each n. Thus by Part (c) of Theorem (III.2.5) one also has $L \leq m$, which contradicts the hypothesis that $0 < m < L$.

IV.5.9 Theorem

Using the same notation as above, let f and g be real-valued functions defined on the set $X_c = (a, b) \setminus \{c\}$. Assume that $\lim\limits_{x\to c} f(x) = A$ and $\lim\limits_{x\to c} g(x) = B$, where A and B are real numbers. Then:

(a) (Sum/Difference Rules) $\lim\limits_{x\to c} (f(x) + g(x)) = A + B$ and $\lim\limits_{x\to c} (f(x) - g(x)) = A - B$.

(c) (Product Rule) $\lim\limits_{x\to c} (f(x){\cdot}g(x)) = A{\cdot}B$.

(d) (Quotient Rule) Suppose that, in addition, $g(x) \neq 0$ when $x{\in}X_c$, and that $B \neq 0$. Then $\lim\limits_{x\to c} (f(x)/g(x)) = A/B$.

Note: The requirement that $g(x) \neq 0$ when $x \neq c$ is included just to simplify the statement of the result. In light of Part (c) of Theorem (IV.5.8), however, it is clear that this condition can be dropped since $B \neq 0$.

IV.5.10 Remark

The preceding result provides 'Sum, 'Difference', 'Product' and 'Quotient' rules for limits of functions on intervals. It is natural to ask about a correponding 'Composition' rule for limits of such functions. The obvious statement would be something like this:

'If $\lim\limits_{x\to c} f(x) = b$ and $\lim\limits_{y\to b} g(y) = L$, then $\lim\limits_{x\to c} g(f(x)) = L$.'

Intuitively, this equation seems quite plausible: As y gets close to b, $g(y)$ gets close to L, and as x gets close to a, the quantity $y = f(x)$ gets close to b; thus combining these facts seems to imply the 'law' stated above.

Unfortunately, this 'law' is not true in general. The underlying reason is the fact that in the expression $\lim_{y \to b} g(y)$ the value of g at b – or even whether g is defined at b – is irrelevent.

Example: Suppose that $f(x) = 1$ for all x in $\mathbb{R}$. Likewise, suppose that $g : \mathbb{R} \to \mathbb{R}$ is given by the rule that $g(y) = 4$ if $y \neq 1$, but $g(1) = 0$. It is clear that

$$\lim_{x \to 0} f(x) = 1 \text{ and } \lim_{y \to 1} g(y) = 4$$

However, $g(f(x)) = g(1) = 0$ for all x, so that $\lim_{x \to 0} g(f(x)) = 0$, not 4 as the 'law' would have predicted.

There are also analogs for 'limits on intervals' of the Monotonic-Sequences Principle (i.e., Part (a) of Theorem (III.2.5)). Of necessity, these analogs involve one-sided limits.

IV.5.11 Theorem (Limits for Monotonic Functions)

Let (a, b) be an open interval in $\mathbb{R}$, so that $-\infty \leq a < b \leq +\infty$.

(a) Suppose that a function $f : (a, b) \to \mathbb{R}$ is monotonic up on (a, b). Then the one-sided limits $\lim_{x \nearrow b} f(x)$ and $\lim_{x \searrow a} f(x)$ both exist. More precisely,

$$(i)\ \lim_{x \nearrow b} f(x) = \sup\{f(x) : a < x < b\} \text{ and } (ii)\ \lim_{x \searrow a} f(x) = \inf\{f(x) : a < x < b\} \qquad (*)$$

In particular,

$$\lim_{x \searrow a} f(x) \leq f(u) \leq \lim_{x \nearrow b} f(x) \text{ for all } u \text{ in } (a, b).$$

(b) Suppose, instead, that the function $f : (a, b) \to \mathbb{R}$ is monotonic down on (a, b). Then the one-sided limits $\lim_{x \nearrow b} f(x)$ and $\lim_{x \searrow a} f(x)$ both exist. More precisely,

$$(i)\ \lim_{x \nearrow b} f(x) = \inf\{f(x) : a < x < b\} \text{ and } (ii)\ \lim_{x \searrow a} f(x) = \sup\{f(x) : a < x < b\} \qquad (**)$$

In particular,

$$\lim_{x \searrow a} f(x) \geq f(u) \geq \lim_{x \nearrow b} f(x) \text{ for all } u \text{ in } (a, b).$$

Partial Proof We prove here only Equation (i) of $(*)$ in Part (a). The rest of the proof is quite similar and is left as an exercise.

Let $L = \sup\{f(x) : a < x < b\}$. It suffices to show that Statement A+ of Part (b) of Theorem (IV.5.4) holds; indeed, by Remark (IV.5.5) (2), Statement B+ of that theorem is true automatically since L is an upper bound of the image of f. Thus, let y be a number such that $y < L$. By the Approximation Property for Suprema, there exists u in (a, b) such that $y < f(u) \leq L$. Since f is monotonic up, it follows that if $u < x < b$ then $f(u) \leq f(x) \leq L$. That is, Statement A+ of Part (b) of Theorem (IV.5.4) holds, and the desired result follows.

Continuity in Elementary Calculus The definition of 'continuity' used in elementary calculus follows the approach to 'limits' studied here.

IV.5.12 Definition (Continuity as Defined in Elementary Calculus)

(a) A real-valued function f is said to **continuous from the left at a number c in the sense of calculus** provided that the following conditions hold:

(i) The function f is defined at c;
(ii) The limit $\lim_{x\to c-} f(x)$ exists;
(iii) $\lim_{x\to c-} f(x) = f(c)$.

(b) If, instead Conditions (i), (ii) and (iii) hold if one replaces $c-$ by $c+$ throughout, then one says that **continuous from the right at c in the sense of calculus**.

(c) If f is continuous both from the left and from the right at c, then one says that **f is continuous at c**.

Remarks (1) Recall that in *This Textbook* the symbols $f(c-)$ and $f(c+)$ are allowed only when the corresponding one-sided limits are finite; see Remark (IV.5.2) (1).

(2) In Part (a) of the current definition the use of the left-hand limit $f(c-)$ in (ii) tacitly assumes that f is defined on some open interval of the form (a, c). Also, our convention Likewise, in Part (b) there is the tacit assumption that f is defined on an open interval of the form (c, b). Finally, in Part (c) there is the tacit assumption that f is defined at every point of an open interval of the form (a, b) such that $a < c < b$, with the possible exception of c itself. The hypothesis that f is actually defined at c is treated separately by assuming Condition (i).

It is easy to see that the preceding definition agrees with Definition (IV.1.2) in the case X is an interval. Thus the general results about continuity obtained earlier in this chapter are valid in this context as well.

In Part (a) of the preceding definition, if Conditions (i) and (ii) hold, then the quantity $f(c) - f(c-)$ measures the extent to which Condition (iii) holds – or fails to hold: f is continuous from the left at c if, and only if, this difference is 0. A similar comment holds for right-hand limits: f is continuous from the right at c if, and only if, $f(c+) - f(c) = 0$. These quantities are given names.

IV.5.13 Definition (Jump Discontinuities)

(1) Suppose that a function f is defined at a real number c and has (finite) left-hand limit $f(c-)$ at c. Then the quantity $f(c) - f(c-)$ is called the **left-hand jump of f at** c.

If the left-hand jump of f is not zero, then one says that f has a **left-hand jump discontinuity** at c.

(2) Likewise, if f is defined at a number c and has (finite) right-hand limit $f(c+)$ at c, then the quantity $f(c+) - f(c)$ is called the **right-hand jump of f at** c.

If the right-hand jump of f is not zero, then one says that f has a **right-hand jump discontinuity** at c.

(3) If both (1) and (2) are applicable, then the sum of the left-hand jump and the right-hand jump of f at c is called the **total jump of f at c**. That is, it is the quantity

$$(f(c) - f(c-)) + (f(c+) - f(c)); \text{ that is, } f(c+) - f(c-)$$

If either the left-hand jump or the right-hand jump of f is not zero, then. one says that f has a **jump discontinuity** at c.

Note Example (1) below shows why the definition of 'jump discontinuity' here is *not* simply 'total jump equals zero'.

IV.5.14 Examples

(1) Let $f : \mathbb{R} \to \mathbb{R}$ be given by the rule

$$f(x) = 1 \text{ if } x \neq 0, f(0) = A \text{ for some real number } A.$$

It is clear that $f(0-) = f(0+) = 1$, so that the left-hand jump of f at 0 is $A - 1$ while the corresponding right-hand jump is $1 - A$. The total jump at 0 is then $(A - 1) + (1 - A) = 0$. It is clear that f is continuous at 0 if $A = 1$, discontinuous there if $A \neq 1$. In particular, if one redefines f at the one point $x = 0$ by resetting $f(0) = 0$, then the 'new' f is continuous there. More generally, if f satisfies $f(c-) = f(c+) = L$, but $f(c) \neq L$, then one says that f has a **removable discontinuity at c**. It is 'removable' in the sense that by merely redefining the value of $f(c)$, and at not other values of x, the resulting function is continuous at c.

(2) Let $f : \mathbb{R} \to \mathbb{R}$ be given by the rule

$$f(x) = \frac{|x|}{x} \text{ if } x \neq 0; \quad f(0) = A \text{ for some real number } A.$$

Note that $f(x) = -1$ if $x < 0$, $f(x) = +1$ if $x > 0$. It is clear that $\lim_{x \to 0-} f(x) = -1$ and $\lim_{x \to 0+} f(x) = +1$. Thus the left-hand jump of f at 0 is $A - (-1) = A + 1$, while the corresponding right-hand jump is $1 - A$. The total jump is then 2. Because the left-side and right-side limits of f at 0 are not equal, the function f is discontinuous at $x = 0$, regardless of the value A of $f(0)$. In other words, the discontinuity of this f at c cannot be removed by merely changing the value of $f(0)$.

(3) It is easy to see that if $f : \mathbb{R} \to \mathbb{R}$ is the Dirichlet function (see Example (I.5.6)), then for every c in $\mathbb{R}$ neither one-sided limit exists. In contrast, if $g : \mathbb{R} \to \mathbb{R}$ is the Thomae function (see the same definition), then g does have both left-sided and right-sided limits at every c in $\mathbb{R}$; indeed, both equal 0 at each c. The discontinuities of this function occur precisely at the rational numbers, and each one is removable.

There is an important class of functions for which the nature of any discontinuities is easy to analyse.

IV.5.15 Theorem

Suppose that $f : I \to \mathbb{R}$ is monotonic on an interval I, and let c be a point of I.

(a) If c is an interior point of I, then both of the one-sided limits $\lim_{x \to c-} f(x)$ and $\lim_{x \to c+} f(x)$ exist and are finite. If, instead, c is an endpoint of I, then the corresponding one-sided limit exists and is finite.

(b) If c is an endpoint of I, then f is continuous at c if, and only if, the corresponding one-sided jump of f at c equals 0. Likewise, if c is an interior point of I, then f is continuous at c if, and only if, the total jump $f(c+) - f(c-)$ of f at c equals 0.

Proof (a) Assume first that f is monotonic up on I and that c is an interior point of I; the other cases can be handled similarly. Let a and b be numbers in I such that $a < c < b$. By the monotonicity hypothesis one has $f(a) \leq f(x) \leq f(c)$ for all x in (a,c). Thus, by properties of 'supremum', one has

$$f(a) \leq \sup\{f(x) : a < x < c\} \leq f(c)$$

Part (a) of the preceding theorem then implies that $f(a) \leq f(c-) \leq f(c)$. Similarly, apply Part (a) of the preceding theorem to the interval (c,b) to get $f(c) \leq f(c+) \leq f(b)$. Combine these to get

$$f(a) \leq f(c-) \leq f(c) \leq f(c+) \leq f(b).$$

In particular, both $f(c-)$ and $f(c+)$ exist and are finite, as claimed.

If f is monotonic down on I and c is an interior point of I, then apply what was just obtained to the function $g = -f$. Finally, the modifications needed in the preceding argument to cover the case in which c is an endpoint of I are left as an easy exercise.

(b) The case in which c is an endpoint of I reduces to Definitions (IV.5.12) and (IV.5.13). Thus, suppose that c is an interior point of I. Note that the 'only if' portion of the claim also follows from the same definitions.

To prove the 'if' portion when c is an interior point of I, first assume that f is monotonic up on I. From the definition of 'total jump' one has

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

Since f is monotonic up on I, one has $f(c+) - f(c) \geq 0$ and $f(c) - f(c-) \geq 0$. It follows that $f(c+) - f(c-) = 0$ if, and only if, $f(c+) - f(c) = 0$ and $f(c) - f(c-) = 0$.

If f is montonic down on I, apply the result just obtained to $g = -f$.

The next theorem states that a function which is monotonic on an interval has, at worst, a countable number of discontinuities, all of which are jump discontinuities.

IV.5.16 Theorem

Suppose that $f : I \rightarrow \mathbb{R}$ is a function which is monotonic on an interval I. Let S be the set of discontinuities of f on the interval I. Then:

(a) Every discontinuity of f in I is a jump discontinuity. (At any endpoint of I one uses the appropriate concept of one-sided continuity.)

(b) The set S is countable (possibly empty).

Proof (a) This follows directly from Theorem (IV.5.15) and the definition of 'continuity as in calculus'.

(b) Let S_0 be the set of interior points of I at which f is discontinuous. Since I has at most two boundary points, the problem reduces to showing that the set S_0 is countable.

As usual, first assume that f is monotonic up, and suppose that $c \in S_0$. (The case of f being monotonic down then follows easily.) Associate with c the set $J_c = \{y : f(c-) \leq y \leq f(c+)\}$. Since $f(c-) < f(c+)$ (because f is monotonic up and discontinuous at c), it follows that J_c is a true interval in $\mathbb{R}$. In particular, J_c contains infinitely many rational numbers.

Now suppose that $S_0 \neq \emptyset$, and define a function $h : S_0 \rightarrow \mathbb{Q}$ by the rule that for each c in S_0, $h(c)$ is one of the rational numbers in the interval J_c.

Claim The function h is one-to-one on S_0.

Proof of Claim Suppose that c_1 and c_2 are points of S_0 such that $c_1 \neq c_2$; without loss of generality, assume that $c_1 < c_2$. It follows from Theorem (IV.5.11) that $f(c_1+) \leq f(c_2-)$. Since, by definition, $h(c_1) < f(c_1+)$ and $h(c_2) > f(c_2-)$, it then follows that $h(c_1) < h(c_2)$, and the claim follows.

It now follows that $h : S_0 \to \mathbb{Q}$ is a bijection of the set S_0 onto a subset of the countable set $\mathbb{Q}$, and thus is countable, as asserted.

Remark The approach followed here, which is based on the countability of the set of rational numbers, is standard. In the exercises an alternate approach to this result, also standard, is outlined. It relates the monotonicity to the countability in a more direct manner.

IV.6 EXERCISES FOR CHAPTER IV

IV - 1 Prove Part (a) of Theorem (**??**).

IV - 2 Let X be a nonempty subset of $\mathbb{R}$. Define $d_X : \mathbb{R} \to \mathbb{R}$ by the rule

$$d_X(y) = \inf \{t : t = |y - x| \text{ for some } x \text{ in } X\}$$

(a) Prove or Disprove: The function d_X is continuous at each point of its domain $\mathbb{R}$.

(b) Give an example of a nonempty set X in $\mathbb{R}$ such that d_X does not have a minimum value.

IV - 3 Let $f : (a,b) \to \mathbb{R}$ be a real-valued function defined on an open interval $I = (a,b)$.

Let c be a point of I. Show that f is continuous at c if, and only if, the following condition holds:

For every open subset U of $\mathbb{R}$ containing the point $f(c)$, the inverse image $f^{-1}[U]$ is an open subset of $\mathbb{R}$ containing c.

IV - 4 Let X be a nonempty subset of $\mathbb{R}$, and suppose that $f_1, f_2, \dots f_k$ are real-valued functions with domain X. Let c be a point of X.

(a) Prove that the function $\max\{f_1, \dots f_k\} : X \to \mathbb{R}$ is continuous at c.

(b) Prove that the function $\min\{f_1, \dots f_k\} : X \to \mathbb{R}$ is continuous at c.

IV - 5 Let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function whose domain is $\mathbb{R}$. Prove that f is continuous on $\mathbb{R}$ if, and only if, for every bounded open interval I in $\mathbb{R}$ the restriction of f to I is continuous on I. Be sure to make clear which part of your solution proves the 'if' portion of the statement, and which proves the 'only if' portion.

IV - 6 The proof of the Intermediate-Value Theorem for Continuous Functions (Theorem (**??**)) given in the *Notes* is based on the Bisection Principle.

Problem Give an alternate proof that is based on the Supremum Principle.

IV - 7 Prove or Disprove If $f : [a,b] \to \mathbb{R}$ is a continuous function, and if $g : [a,b] \to \mathbb{R}$ is defined by the rule

$$g(x) = \sup\{f(u) : a \leq u \leq x\} \text{ for } x \text{ in } [a,b],$$

then g is also continuous.

IV - 8 Let $f : \mathbb{R}\backslash\{1\} \to \mathbb{R}$ and $g : \mathbb{R}\backslash\{1\} \to \mathbb{R}$ be given by

$$f(x) = \frac{x^3 - 1}{x - 1} \text{ if } x \neq 1,$$

and

$$g(x) = \frac{x^2 + x - 2}{x - 1} \text{ if } x \neq 1$$

Problem Determine whether there is a choice of C in $\mathbb{R}$ such that the concatenation $f\&_{(1,C)}g$ is continuous on $\mathbb{R}$.

IV - 9 Prove or Disprove If $f : \mathbb{R} \to \mathbb{R}$ is a function such that f^2 is continuous on $\mathbb{R}$, then f is also continuous on $\mathbb{R}$.

IV - 10 Prove or Disprove If f and g are real-valued functions defined on $\mathbb{R}$, and f and $f\circ g$ are continuous, then g is continuous.

IV - 11 Definition (1) A real-valued function f defined on an interval I in $\mathbb{R}$ is said to be a **convex function on** I provided the following condition holds:

If x and y are in I, then $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ for all t such that $0 \leq t \leq 1$.

The function f is said to be **strictly convex on** I provided one has, in addition, when $x < y$ and $0 < t < 1$ one has $f(tx + (1-t)y) < tf(x) + (1-t)f(y)$.

(2) A real-valued function g on an interval I in $\mathbb{R}$ is said to be a **concave function on** I provided $-g$ is a convex function on I. Likewise, g is said to be **strictly concave on** I provided $-g$ is a strictly convex function on I.

Problem

(a) Give geometric interpretations, in terms of the relation between the graph of a function f on an interval I and the chords joining pairs of points on that graph, of the statements 'f is a convex function on I' and 'f is a strictly convex function on I'.

(b) Determine the convexity/concavity properties the 'absolute-value' function abs : $\mathbb{R} \to \mathbb{R}$, given by the rule abs$\,(x) = |x|$ for all x in $\mathbb{R}$. (By 'determine the convexity/concavity properties' is meant: determine the largest intervals on which the function is convex, strictly convex, concave or strictly concave.)

Warning The meanings of 'convex' and 'concave' given here are quite standard in advanced texts in analysis. However, most textbooks on elementary calculus use a different terminology. Specifically, they say 'concave up' instead of 'convex function' and 'concave down' instead of 'concave function'; a similar usage holds for 'strict' convexity/concavity. The use of the word 'concave' in the phrase 'concave up' to describe 'convexity' can sometimes cause confusion, so be careful.

IV - 12 (a) Show that if a real-valued function is convex on an open interval (a,b), then it is continuous on (a,b).

(b) Give an example of a function f, whose domain is a *closed* interval $[a,b]$, such that f is convex on $[a,b]$, but is not continuous on $[a,b]$.

IV - 13 Suppose that f is a continuous real-valued function defined on an interval I in $\mathbb{R}$ and that f satisfies the condition

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2} \text{ for all } x, y \text{ in } I.$$

Prove that f is convex on I.

IV - 14 Definition A function $f : \mathbb{R} \to \mathbb{R}$ is said to be an **additive function** provided $f(x+y) = f(x) + f(y)$ for all x and y in $\mathbb{R}$.

Prove that if f is an additive function which is continuous at 0 then f is a linear function; more precisely show that there exists a real number c such that $f(x) = cx$ for all x in $\mathbb{R}$.

IV - 15 Prove or Disprove If $f : \mathbb{R} \to \mathbb{R}$ is a continuous function such that $|f(y) - f(x)| \geq |y - x|$ for all x and y in $\mathbb{R}$, then f maps $\mathbb{R}$ onto $\mathbb{R}$.

IV - 16 Remark In the proof of the Extreme-Value Theorem found in the *Notes*, one obtains the result that if f is a continuous real-valued function defined on a closed bounded interval I, then

f is bounded above on I. However, this fact comes almost as an after-thought: one first shows that the supremum of f over I equals $f(c)$ for some c in I, and thus $\sup f[I]$ must be finite; in particular, f must be bounded above on I. In other words, one shows that f is bounded by *first* showing that f assumes a maximum value on I.

The goal of the present exercise is to show directly that such a function f must be bounded above on I; indeed, *two* approaches to this result are provided. The following execise then asks one to use this boundedness to prove that f assumes a maximum on I.

Suppose that $f : I \to \mathbb{R}$ is a real-valued function whose domain is a closed bounded interval $I = [a, b]$ in $\mathbb{R}$, and assume that f is continuous at each point of I.

(a) Give a direct proof of the fact that f is bounded on I by considering the set X of all $x > a$ in $[a, b]$ such that f is bounded on the subinterval $[a, x]$. In the course of events you may need to show that the set X is nonempty, and you may wish to consider the supremum of this set.

(b) Give a direct proof of the fact that f is bounded on I by noting that for each x in I there exists $\delta_x > 0$ such that f is bounded on the subset $U_x = (x - \delta_x, x + \delta) \cap I$. Then prove that there is a finite subset of the family $\mathcal{F} = \{U_x : x{\in}I\}$ whose union has I as a subset.

IV - 17 Give a proof, of the Extreme-Value Theorem for continuous real-valued functions defined on closed bounded intervals, which is based on the fact that continuous real-valued functions defined on closed bounded intervals are bounded functions.

IV - 18 Definition Let $f : X \to \mathbb{R}$ be a real-valued function whose domain is a nonempty subset X of $\mathbb{R}$, and let c be a point of X.

(i) One says that f is **upper-semicontinuous at** c provided that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if x in X satisfies $|x - c| < \delta$ then $f(x) \leq f(c) + \varepsilon$.

(ii) One says that f is **lower-semicontinuous at** c provided that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if x in X satisfies $|x - c| < \delta$ then $f(x) \geq f(c) - \varepsilon$.

(a) Give an example of a function which is upper-semicontinuous, but not continuous, at a point.

(b) Prove that the function f is continuous at the point c if, and only if, f is both upper-semicontinuous and lower-semicontinuous at c.

(c) Prove that if $f : [a, b] \to \mathbb{R}$ is upper-semicontinuous at each point of a closed interval $[a, b]$, then f assumes a maximum value for $[a, b]$ at some point of $[a, b]$.

IV - 19 Definition Let r be a rational number and let c be a positive real number. Then the number c^r, which is pronounced 'the r-th power of c', is defined as follows:

$$c^r = \begin{cases} 1 & \text{if } r = 0 \\ (\sqrt[n]{c})^m & \text{if } r > 0 \text{ and } r = m/n \text{ for natural numbers } m \text{ and } n \text{ in lowest terms} \\ c^{-r} & \text{if } r < 0 \end{cases}$$

As usual, the symbol $\sqrt[n]{c}$ denotes the unique positive real number u such that $u^n = c$.

(a) Suppose that $c > 0$ and that r is a rational number. Show that if one writes r in the form $r = p/q$, where p and q are integers and $q > 0$, then $c^r = (\sqrt[q]{c})^p$; that is, one does not need to express the rational number r in 'lowest terms'.

(b) Suppose that $c > 0$. Show that $c^{-r} = 1/c^r$ for every rational r.

(c) Suppose that $c > 0$. Show that $c^{r+s} = c^r c^s$ and $(c^r)^s = c^{(rs)}$ for all rationals r and s.

Chapter V

Derivatives and Antiderivatives in $\mathbb{R}$

Quotes for Chapter (V):

> (1) '*6accdæ13eff7i3l9n4o4qrr4s8t12vx*'
>
> (This is an anagram that appears in a document which Isaac Newton sent to Gottfried Leibniz, via Henry Oldenberg, the secretary of the Royal Society, in 1676. The numbers indicate how often the given letter appears in the anagram; for example, '6*a*' means that there are six occurrences of the letter *a*; likewise there is one occurrence of the ligature æ. The anagram summarizes, in a concealed way, the new techniques Newton used in obtaining the results he described in the letter. Can you figure out the message Newton hid in this anagram? Hint: The message was written in Latin.)
>
> (2) 'I turn away with fright and horror from this lamentable evil of functions which do not have derivatives.'
>
> (Translation of part of a letter from C. Hermite to L. Stieltjes.)
>
> (3) 'The proofs [of a certain theorem] in many text-books (and in the first three editions of this book) are inaccurate.'
>
> (From a footnote on Page 217 of the tenth edition of the celebrated text *A Course of Pure Mathematics* by G. H. Hardy. Can you guess which theorem was the cause of so many inaccurate proofs?)
>
> (4) 'Dans le calcul intégral il ma paru nécesaire de démontrer généralement l'existence des intégrales ou fonctions primatives avant de faire connaitre leur diverses propriétés.'
>
> Rough translation: 'In the integral calculus it seems necessary to me that one show in a general manner the existence of antiderivatives before making known their various properties.' (From the introduction to [CAUCHY 1823])

Introduction

The standard elementary (single-variable) calculus courses in the United States focus on two closely related concepts: the process of differentiation, and its inverse, the process of antidifferentiation. The purpose of the current chapter is to treat these topics, but more rigorously than in calculus. (These courses also include applications of these processes to subjects such as geometry, physics and economics, but such applications are not considered here.) Note that it is traditional in these courses to combine the treatment of antidifferentiation with the important topic of the definite integral, but in *This Textbook* we postpone the latter topic until Chapter (VII).

As was mentioned in the 'Preliminaries' for Chapter (I) (see Section (I.1)), readers of *This Textbook* should have already taken a course in elementary calculus. To avoid simply repeating the standard treatments which are already familiar from such a course, in this chapter we often present alternate treatments which can provide different insights to calculus. (Frequently the standard approaches are also included, but as exercises.) Many readers of analysis texts at the level of *This Textbook* eventually teach calculus themselves, perhaps as graduate teaching assistants; it is useful for such readera to know that there are alternate approaches.

V.1 The Derivative – Basic Definitions

The reader should be familiar, from elementary calculus, with the standard motivating examples which lead to the concept of 'derivative': slopes of curves, velocities, rates of change, and so on. Thus, we can omit those examples here and go straight to the formal definitions they suggest.

V.1.1 Definition (Differentiability; the Derivative)

Let $f : I \to \mathbb{R}$ be a real-valued function defined on an open interval I in $\mathbb{R}$.

(1) Suppose that c is a point of I. One says that **the function f is differentiable at c** provided that the expression $\lim\limits_{x \to c} \dfrac{f(x) - f(c)}{x - c}$ exists and is finite. The process of computing such limits is called **differentiation**. (The word 'differentiation' reflects the presence of the *differences* $f(x) - f(c)$ and $x - c$ in the preceding fraction.) If f is differentiable at each point c of a subset X of I, then one says that f is **differentiable on X**.

(2) Let X be the set of all c in I such that f is differentiable at c, in the sense of Part (1); note that X need not be an interval. If $X \neq \emptyset$ one can associate with f a second function $f' : X \to \mathbb{R}$ given by the rule

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} \text{ for all } c \text{ in } X. \tag{V.1}$$

The function f' is called **the derivative of f** (because it is *derived* from f; namely, by the process of differentiation).

V.1.2 Remarks

(1) The quantity $f(x)$ is defined for all x in the open interval I, but the expression $g(x) = \dfrac{f(x) - f(c)}{x - c}$ fails to be defined at $x = c$, since division by zero is not allowed. The fact that the derivative involves expressions of the form $\lim_{x \to c} g(x)$, where g is not defined at c, is the main reason that in calculus one follows the approach to limits found in Section (IV.5).

(2) Suppose that I is an open interval, J is an open subinterval of I, and c is a point of J (and thus a point of I). Let $f : I \to \mathbb{R}$ be a function defined on I, and let $h : J \to \mathbb{R}$ be the restriction of f to J. Then f is differentiable at c if, and only if, h is differentiable at c; and in this case one has $f'(c) = h'(c)$; see Remark (IV.5.2) (c).

The definition of 'derivative' used here agrees with that found in all elementary-calculus texts and in most texts on real analysis; in particular, we restrict this concept to functions defined on *open* intervals in $\mathbb{R}$. However, on occasion it is useful to allow the following simple extensions of the 'derivative' concept.

V.1.3 Definition (One-Sided Derivatives)

Suppose that f is a real-valued function defined on a half-open interval $(a, c]$. If $\lim_{x \nearrow c} \dfrac{f(x) - f(c)}{x - c}$ exists and is finite, then one says that that f is **differentiable from the left** at c. One denotes this limit by $f'_{-}(c)$ and calls it the **left-hand derivative of f at c**. Similarly, suppose that f is defined on an interval $[c, b)$. If $\lim\limits_{x \searrow c} \dfrac{f(x) - f(c)}{x - c}$ exists and is finite, one says that f is **differentiable from the right** at c. In this case one denotes that limit by $f'_{+}(c)$ and calls it the **right-hand derivative of f at c**. Each such quantity is called a **one-sided derivative**.

The next result states the obvious relation between one-sided derivatives and the usual (i.e., 'two-sided') derivative; its simple proof is left as an exercise.

V.1.4 Theorem

Let $f : I \rightarrow \mathbb{R}$ be a function defined on an open interval I, and let c be a point of I. Let $g : (a, c] \rightarrow \mathbb{R}$ and $h : [c, b) \rightarrow \mathbb{R}$ be the restrictions of f to the intervals $(a, c]$ and $[c, b)$, respectively. Then $f'(c)$ exists (in the sense of Definition (V.1.1) if, and only if, both of the one-sided derivatives $g'_{-}(c)$ and $h'_{+}(c)$ exist and are equal to each other. In this case one has $f'(c) = g_{-}(c) = h_{+}(c)$. ■

This result can be useful in dealing with the derivative of a function which is given by different formulas on adjacent intervals. Examples are given later in this chapter.

V.1.5 Theorem (Differentiability Implies Continuity)

Let $f : I \rightarrow \mathbb{R}$ be a real-valued function defined on an open interval I in $\mathbb{R}$.

(a) If f is differentiable at a point c of I, then f is continuous at c.

(b) The converse of (a) is not true. That is, it is *not* the case that the statement 'f is continuous at c' implies the statement 'f is differentiable at c'. Otherwise stated: A necessary condition for f to be differentiable at c is that it be continuous at c, but this 'continuity' condition is not sufficient to guarantee that f is differentiable at c.

<u>Proof</u> (a) Notice that if $x \in I$ and $x \neq c$, then

$$f(x) - f(c) = \left(\frac{f(x) - f(c)}{x - c}\right) \cdot (x - c)$$

Since, by hypothesis, f is differentiable at c, the first factor of the product on the right side of this equation approaches a finite limit. Let us denote this limiting value by L. (Of course we could denote it by $f'(c)$, but the issue here is that the limit exists and is finite.) Note also that the second factor on the right, namely $x - c$, also approaches the limiting value 0. It then follows from the Product Rule for Limits that limit of the right side, as x approaches c, exists and equals $L \cdot 0 = 0$. Thus, the same is true for the left side; that is, $\lim_{x \rightarrow c}(f(x) - f(c)) = 0$, so that f is continuous at c, as claimed.

(b) Examples (2) and (3) below involve functions which are continuous at a point but fail to be differentiable at that point. ■

V.1.6 Examples

(1) Let $f : \mathbb{R} \to \mathbb{R}$ be the function given by the rule

$$f(x) = \begin{cases} x^2 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

Let $c = 0$. One computes that if $x \neq 0$ then

$$\frac{f(x) - f(c)}{x - c} = \frac{f(x)}{x} = \begin{cases} x & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

It is clear that f is not continuous at c when $c \neq 0$, so by Theorem (V.1.5) the function f is *not* differentiable at any nonzero c.

In contrast, it is easy to see that f *is* continuous at $c = 0$. As for differentiability there, it is clear that $-|x| \leq \dfrac{f(x)}{x} \leq |x|$ for all $x \neq 0$, and thus, by Part (d) of Theorem (IV.5.7) (the Squeeze Property for Limits on Intervals), it follows that $\lim\limits_{x \to 0} \dfrac{f(x)}{x} = 0$. That is, f is differentiable at $c = 0$, and $f'(0) = 0$. In particular, this function is differentiable at exactly one point.

(2) Let $g : \mathbb{R} \to \mathbb{R}$ be the function given by the rule

$$g(x) = \begin{cases} x & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

By the argument used in Example (1) above, it is easy to show that g is continuous at exactly one number c, namely at $c = 0$; in particular, the only point at which it is even possible for g to be differentiable is $c = 0$. Note, however, that if $x \neq 0$ then one has

$$\frac{g(x) - g(0)}{x - 0} = \frac{g(x)}{x} = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

This expression does not approach a limit as x approaches 0, so g is not differentiable at the one point where there was a chance of this happening.

(3) Let $h : \mathbb{R} \to \mathbb{R}$ be the Absolute-Value Function; that is, $h(x) = |x|$ for every x in $\mathbb{R}$. Set $c = 0$ and note that if $x \neq 0$, then

$$\frac{h(x) - h(c)}{x - c} = \frac{|x - 0|}{x - 0} = \frac{|x|}{x} = \begin{cases} -1 & \text{if } x < 0 \\ +1 & \text{if } x > 0 \end{cases}$$

It is clear that the left-hand limit of the expression $|x|/x$ equals -1 while the right-hand limit of the same expression equals $+1$. That is, $h'_-(0)$ and $h'_+(0)$ exist, but are not equal. It follows from Theorem (V.1.4) that h is not differentiable at 0.

Remark There exist functions $h : \mathbb{R} \to \mathbb{R}$ with the property that h is continuous at *every* point of $\mathbb{R}$ yet differentiable at *no* point of $\mathbb{R}$; it is perhaps one of these functions which induced the 'horror' expressed by the great mathematician Hermite in Chapter Quote (2) at the start of this chapter.

The preceding examples emphasized the idea of a function being differentiable – or not differentiable – at a single point. Such examples play a small role in elementary calculus: the emphasis there is on functions which are differentiable at each point of an open interval. The following examples are of that type and should seem more familiar.

V.1.7 Some Standard Examples

(1) Let $f : \mathbb{R} \to \mathbb{R}$ be a linear function on $\mathbb{R}$; that is, there are real numbers A and B such that $f(x) = A\,x + B$ for all x in $\mathbb{R}$. Then f is differentiable on $\mathbb{R}$, and $f'(x) = A$ for each x in $\mathbb{R}$. Indeed, for each x and c in $\mathbb{R}$ with $x \neq c$ one has

$$\frac{f(x) - f(c)}{x - c} = \frac{(A\,x + B) - (A\,c + B)}{x - c} = \frac{A\,(x - c)}{x - c} = A, \text{ hence } \lim_{x \to c} \frac{f(x) - f(c)}{x - c} = A.$$

Two special cases have their own terminology:

(i) If $A = 0$, then f is the constant function of value B on $\mathbb{R}$.

(ii) If $A = 1$ and $B = 0$, then f is the identity function on $\mathbb{R}$.

Remark In light of Remark (V.1.2) (2) above, it follows that if $g : I \to \mathbb{R}$ is the restriction to an open interval I of a linear function $f : \mathbb{R} \to \mathbb{R}$, where $f(x) = A\,x + B$ for all x in $\mathbb{R}$ as above, then g is differentiable on I and $g'(c) = A$ for all x in I. For example, it follows easily from this that the absolute-value function h, given by

$$h(x) = |x| = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0 \end{cases},$$

is differentiable at all $c \neq 0$. Indeed, on the open interval $(-\infty, 0)$ h is the restriction to $(-\infty, 0)$ of the linear function with $A = -1$ and $B = 0$, so $h'(c) = -1$ if $c < 0$. A similar argument shows that $h'(c) = +1$ if $c > 0$.

Of course it has already been shown that the function h is not differentiable at $c = 0$; see Example (V.1.6) (3).

(2) Suppose the function $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = A\,x^2$ for all x in $\mathbb{R}$, where A is a fixed real number. Then for each x and c in $\mathbb{R}$ with $x \neq c$, one has

$$\frac{f(x) - f(c)}{x - c} = \frac{(A\,x^2 - A\,c^2)}{x - c} = A\left(\frac{x^2 - c^2}{x - c}\right) = A\left(\frac{(x - c)\,(x + c)}{x - c}\right) = A\,(x + c).$$

It follows from the usual limit laws that $\lim_{x \to c} A\,(x + c) = A\,(c + c) = 2\,A\,c$, and thus

$$\lim_{x \to c} \frac{f(x) - f(c)}{x - c} = 2\,A\,c.$$

That is, f is differentiable at c, and $f'(c) = 2\,A\,c$, for each c in $\mathbb{R}$.

(3) More generally, suppose that k is any natural number. By using the well-known factorization

$$x^k - c^k = (x - c)\,(x^{k-1} + x^{k-2}\,c + \ldots + x\,c^{k-2} + c^{k-1}),$$

it is easy to show that if $f(x) = x^k$ for all x, then $f'(c) = k\,c^{k-1}$ for every number c.

(4) Let $f : \mathbb{R}\backslash\{0\} \to \mathbb{R}$ be the 'reciprocal function'; that is, $f(x) = 1/x = x^{-1}$ for all $x \neq 0$. One then computes that

$$f(x) - f(c) = \frac{1}{x} - \frac{1}{c} = \frac{c - x}{x\,c} \text{ for all } x, c \neq 0.$$

Thus if one also has $x \neq c$, it follows that

$$\frac{f(x) - f(c)}{x - c} = \frac{c - x}{(x - c)\,x\,c} = = -\frac{(x - c)}{(x - c)\,x\,c} = -\frac{1}{x\,c}.$$

It follows from the basic laws for limits that $\lim_{x \to c} -\frac{1}{x\,c} = -\frac{1}{c^2}$, so that

$$\lim_{x \to c} \frac{f(x) - f(c)}{x - c} = \lim_{x \to c} -\frac{1}{x\,c} = -\frac{1}{c^2} \text{ for } c \neq 0.$$

That is, if $c \neq 0$ then f is differentiable at c and $f'(c) = -1/c^2$.

(5) Suppose that $f(x) = \sqrt{x} = x^{1/2}$ for all $x \geq 0$. Using a familiar trick from algebra, the so-called 'Multiply-and-Divide Trick', one then computes that

$$f(x) - f(c) = \sqrt{x} - \sqrt{c} = \left(\frac{(\sqrt{x} - \sqrt{c})(\sqrt{x} + \sqrt{c})}{\sqrt{x} + \sqrt{c}}\right) = \left(\frac{(\sqrt{x})^2 - \sqrt{x}\sqrt{c} + \sqrt{x}\sqrt{c} - (\sqrt{c})^2}{\sqrt{x} + \sqrt{c}}\right) = \left(\frac{x - c}{\sqrt{x} + \sqrt{c}}\right)$$

for all $x, c > 0$. (Note that the denominator $\sqrt{x} + \sqrt{c}$ is positive for all such x and c, so there is no 'division-by-zero' issue.) If, in addition $x \neq c$, then one has

$$\frac{f(x) - f(c)}{x - c} = \frac{x - c}{(x - c)(\sqrt{x} + \sqrt{c})} = \frac{1}{\sqrt{x} + \sqrt{c}}.$$

Since $\lim_{x \to c} \sqrt{x} = \sqrt{c}$ (see Example (IV.4.8)), it follows by the usual limit laws that

$$\lim_{x \to c} \frac{1}{\sqrt{x} + \sqrt{c}} = \frac{1}{2\sqrt{c}}.$$

Thus, one finally obtains

$$\lim_{x \to c} \frac{f(x) - f(c)}{x - c} = \lim_{x \to c} \frac{1}{\sqrt{x} + \sqrt{c}} = \frac{1}{2\sqrt{c}}.$$

In other words, if $c > 0$ then $f'(c)$ is defined, and

$$f'(c) = \frac{1}{2\sqrt{c}} = \frac{1}{2}\,c^{-1/2}.$$

In the case $c = 0$ it makes sense to consider the right-hand derivative. Since one must have $x > 0$ in that case, the corresponding difference quotion takes the form

$$\frac{f(x) - f(0)}{x - 0} = \frac{1}{\sqrt{x}}.$$

This expression approaches $+\infty$ as x decreases to 0, so the corresponding one-sided derivative fails to exist at $c = 0$, since by definition derivatives, whether one-sided or two-sided, are finite.

V.1.8 Remarks

There are several commonly used notations for the derivative that one should know.

(1) Many texts use the following formula to define the derivative:

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

One advantage of this formulation is that it works for those who insist that the input of both f and f' ought to be the same 'variable', often x. Note that h plays the role of $x - c$ which appears in

Definition (V.1.1) above. In particular, with this notation there is no need to introduce the extra letter c as was done in the preceding examples. Thus, instead of defining f by, say, the formula $f(x) = x^k$ and thus $f'(c) = k\,c^{k-1}$ as in Example (3) above, one could write the more familiar $f'(x) = k\,x^{k-1}$.

Similarly, many texts, especially the older ones, use the notation Δx in place of h. For those texts the definition becomes

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

The Greek letter Δ (i.e., upper-case delta) has the sound (at least in Classical Greek) of the English 'd'. It suggests the idea of 'difference'; indeed, $\Delta x = (x + \Delta x) - x$. That is, Δx is the change in the variable x in going from x to $x + \Delta x$.

Those authors who insist on the traditional 'variables' notation might write, say, $y = f(x)$ as the 'equation' of the function f, and refer to the expression $f(x + \Delta x) - f(x)$ as 'the change in y corresponding to the given change in x'. They would denote this change in the quantity y by Δy, and then write

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x},$$

and refer to the quantity $f'(x)$ as 'the derivative of y with respect to x' at the given value of the variable x. Such authors often write y' in place of $f'(x)$. The 'prime' notation here mentions the dependent variable y explicitly, but leaves it understood from the context that the corresponding independent variable is x.

(2) The 'prime' notation, f', for the derivative of the function f, is due to Lagrange. There are several older notations which one needs to know since they also are still in common use:

(a) **Leibniz' 'differential' notation** The most popular notation is the so-called 'differential' notation of Leibniz. Using the classical 'variables' description of functions, consider the function $y = f(x)$. Instead of comparing $f(x)$ with $f(x+\Delta x)$ for a small change Δx in the input variable x to obtain the corresponding small change $\Delta y = f(x+\Delta x) - f(x)$ in the output variable y, Leibniz introduced the symbol dx for an *infinitely small* – but nonzero – change in x, and dy for the corresponding *infinitely small* change in y. Note that in Leibniz' time an 'infinitely small' quantity is to be smaller in magnitude than any positive real number, yet still can be nonzero. Leibniz thinks of dx and dy as the horizontal and vertical sides, respectively of an infinitely small right triangle whose hypotenuse connects the 'neighboring' points (x, y) and $(x+dx, y+dy)$ of the graph $y = f(x)$. Leibniz then divides the (nonzero) dx into dy to get the slope of this hypotenuse; that is, the slope of $y = f(x)$ at the point (x, y). In terms of Lagrange's 'prime' notation, $f'(x) = dy/dx$; equivalently, $dy = f'(x)\,dx$.

NOTE: 'Infinitely small quantities', or 'infinitesimals', were used long before Leibniz and long after. One of the goals of the rigorization of analysis, as started in the nineteenth century, was to replace the use of 'infinitesimals' by the theory of limits; that is, by the approach followed, for example, in *This Textbook*. More recently, rigorous treatments of 'infinitesimals' have been developed under the heading of 'Non-standard Analysis'; see, for example, [ABRAHAM 1996].

(b) **Newton's 'dot' notation** As a physicist Newton had a dynamic view of calculus: quantities were viewed as functions of the time t, and calculus measured the time rate of change of such quantities. For example, if $u = f(t)$, then Newton would write $\dot{u}$ instead of du/dt or $f'(t)$. The 'dot' notation for differentiation with respect to the time is still in common use in physics and engineering.

(c) **Euler's 'D' notation** This notation is both obvious and simple. More precisely, if $y = f(x)$, then one writes Dy or $(Df)(x)$ or, more simply, $Df\,(x)$.

V.2 Derivatives of Higher Order

In most cases of interest to calculus, the derivative f' of a function f is itself a differentiable function, and thus its derivative $(f')'$ can also be studied. Repeating this idea leads to the following.

V.2.1 Definition (Derivatives of Higher Order)

Let $f : I \to \mathbb{R}$ be a real-valued function which is differentiable at each point of an open interval I.

(1) The **first derivative of** f, denoted by $f^{(1)}$, is the function f'. One also calls $f^{(1)}$ the **derivative of order 1** of f.

(2) If $f^{(1)}$ is also differentiable at each point of I, then its derivative, $(f^{(1)})'$, is called the **second derivative**, or **derivative of order** 2, of f; it is denoted by $f^{(2)}$.

(3) More generally, suppose that $n \in \mathbb{N}$ is such that the derivative of order n, $f^{(n)}$, has already been defined on I. If $f^{(n)}$ is also differentiable at each point of I, then its derivative, $(f^{(n)})' : I \to \mathbb{R}$, is called the **derivative of order** $n+1$, or the $(n+1)$**-st derivative** of f. It is denoted $f^{(n+1)}$.

(4) If, for some natural number n, $f^{(n)}$ is defined at each point of I, then one says that f is ***n*-times differentiable on *I***.

(5) For completeness, one sets $f^{(0)} = f$; that is, $f^{(0)}$ is the function which comes from f by not differentiating f at all. (Of course this notation makes sense even without the assumption that f is differentiable on I, but there is little need for it then.) Also, to agree with standard notation in elementary calculus, one lets f', f'', f''', ... be alternate notations for $f^{(1)}$, $f^{(2)}$, $f^{(3)}$, ... , although one rarely uses the 'prime' notation for, say, $n > 4$.

V.2.2 Example

Let $f : \mathbb{R} \to \mathbb{R}$ be given by the rule $f(x) = x^k$ for all x in $\mathbb{R}$, where k is a fixed natural number. Then by Example (V.1.7) above, one has $f'(x) = k\,x^{k-1}$ for all x. By repeated use of the same example, for all x one has $f''(x) = k\,(k-1)\,x^{k-2}$, $f'''(x) = k\,(k-1)\,(k-2)\,x^{k-3}$, and so on. Eventually one obtains $f^{(k)}(x) = k\,(k-1)\,(k-2), \ldots \cdot 2 \cdot 1 = k!$ for all x, so that $f^{(k)}$ is a constant function, and thus $f^{(n)}(x) = 0$ for all x if $n \in \mathbb{N}$ satisfies $n \geq k+1$.

V.2.3 Definition (C^k functions)

Let $f : I \to \mathbb{R}$ be defined at each point of an open interval I, and let n be a nonnegative integer.

(1) One says that f is **of class C^n** on I, or that f is **C^n** on I, provided that all the functions $f^{(0)}, f^{(1)}, \ldots f^{(n)}$ exist, and are continuous, at each point of I.

If f is of class C^k, but not of class C^{k+1}, on I, then one says that f is **strictly C^k** on I.

(2) If the function f is of class C^k on I for each k in $\mathbb{N}$, then one says that f is **of class C^∞**, or that f is **smooth**, on I.

Remark Most of the standard functions which one uses in elementary calculus are smooth. In particular, one typically does not encounter *strictly* C^k functions there with k finite. The next example fills that gap.

V.2.4 Example

Let $F : \mathbb{R} \to \mathbb{R}$ be the function given by the following rule:

$$F(x) = \begin{cases} -x^2 & \text{if } x \leq 0 \\ x^2 & \text{if } x > 0 \end{cases}$$

From Example (V.1.7) (2) above it is clear that F is C^∞ on the open intervals $(-\infty, 0)$ and $(0, +\infty)$. Indeed, one has $F'(x) = -2\,x$ if $x < 0$, while $F'(x) = 2\,x$ if $x > 0$; that is, $F'(x) = 2\,|x|$ for all $x \neq 0$.

The situation at $c = 0$ is more complicated. Indeed, if $x \neq 0$ one computes that

$$\frac{F(x) - F(0)}{x - 0} = \begin{cases} \dfrac{-x^2}{x} = -x & \text{if } x < 0 \\ \\ \dfrac{x^2}{x} = x & \text{if } x > 0 \end{cases}$$

As x approaches 0 from either side the corrresponding fractions $-x$ and x both approach 0. Thus, F is differentiable at $c = 0$, and $F'(0) = 0$, so that $F'(0) = 2\,|0|$. Combining these results, one sees that F is differentiable on $\mathbb{R}$, and F' is twice the absolute-value function. The latter function is continuous on $\mathbb{R}$, so F is C^1 on $\mathbb{R}$. However, F' is not differentiable at $c = 0$, so F is not C^2 on $\mathbb{R}$; indeed, it is not even twice-differentiable on $\mathbb{R}$.

V.2.5 Remarks

(1) In light of the preceding example, it is natural to ask whether there are C^k functions on an interval I which are $(k+1)$-times differentiable on I but not C^{k+1} on I. The answer is that such functions do exist. However, the construction of such examples is nontrivial, and is held off until later.

(2) The alternate notations for derivatives mentioned in Remark (V.1.8) above have extensions to derivatives of higher order. For example, second derivatives are written as follows: if $y = f(x)$, then

$$f^{(2)}(x) = f''(x) = y'' = \frac{d^2y}{dx^2} = \frac{d^2f(x)}{dx^2} = \ddot{y} = D^2y = D^2f(x).$$

Similarly, k-th order derivatives are written

$$f^{(k)}(x) = \frac{d^ky}{dx^k} = \frac{d^kf(x)}{dx^k} = D^ky = D^kf(x);$$

the corresponding 'prime' and 'dot' notations are omitted because when $k \geq 4$ they become difficult to read.

The notation 'd^ky/dx^k' sometimes confuses students. For example, if $k = 2$, then its genesis is the formula

$$f''(x) = (f')'(x) = \frac{d}{dx}\left(\frac{dy}{dx}\right).$$

The right side has two copies of d and one of y in the numerator, and two copies of dx in the denominator, which explains the abbreviation d^2y/dx^2. Note, in particular, that dx^2 is itself a shorthand for $(dx)^2$, not $d(x^2)$. Likewise, the notation D^2f is shorthand for $D(Df)$, not $(Df)^2$, so there are two copies of D but only one of f.

V.3 Computational Rules for Derivatives

Examples given above illustrate how to compute the derivative directly from its definition in some simple cases. If one needed the direct use of that definition in general, then calculus would be a much more difficult, and less useful, tool. The next several theorems summarize rules for differentiating functions which allow one to use such simple cases to compute many other derivatives without referring to that definition. Throughout these theorems, I is an open interval in $\mathbb{R}$, and c is a point of I.

V.3.1 Theorem

(a) (**Constant-Factor Rule for Differentiation**) Suppose that $f : I \to \mathbb{R}$ is differentiable at c. Let $h = k{\cdot}f$ for some number k; that is, $h(x) = k{\cdot}(f(x))$ for all x in I. Then h is also differentiable at c, and $h'(c) = k{\cdot}f'(c)$.

(b) (**Sum and Difference Rules for Differentiation**) Suppose that $f, g : I \to \mathbb{R}$ are differentiable at c. Let $h_1 = f + g$; that is, $h_1(x) = f(x) + g(x)$ for all x in I. Then h_1 is also differentiable at c, and $h_1'(c) = f'(c) + g'(c)$. Likewise, if $h_2 = f - g$, then h_2 is differentiable at c, and $h_2'(c) = f'(c) - g'(c)$.

(c) (**Linear-Combination Rule for Differentiation**) Suppose that $f_1, f_2, \ldots f_m : I \to \mathbb{R}$ are all differentiable at c. Let $h : I \to \mathbb{R}$ be a **linear combination** of these functions; that is, there exist constants k_1, $k_2, \ldots$ k_m such that

$$h(x) = k_1{\cdot}(f_1(x)) + k_2{\cdot}(f_2(x)) + \ldots + k_m{\cdot}(f_m(x)) \text{ for all } x \text{ in } I.$$

Then h is differentiable at c, and $h'(c) = k_1{\cdot}f_1'(c) + k_2{\cdot}f_2'(c) + \ldots + k_m{\cdot}f_m'(c)$.

The simple proofs are left as exercises. ■

Remark The 'Linear Combination' terminology comes from the mathematical subject of 'Linear Algebra'. In *This Textbook* one does not need to be familiar with that subject.

V.3.2 Theorem

Suppose that $f, g : I \to \mathbb{R}$ are functions, defined on an open interval I, which are differentiable at a point c of I.

(a) (**Product Rule for Differentiation**) Let $P = f{\cdot}g$ be the product of f and g; that is, $P(x) = f(x){\cdot}g(x)$ for all x in I. Then P is differentiable at c, and

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

(b) (**Quotient Rule for Differentiation**) Suppose, also, that for all x in I one has $g(x) \neq 0$. Let $Q = f/g$ be the quotient of f by g; that is, $Q(x) = f(x)/g(x)$ for all x in I. Then Q is differentiable at c, and

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

Special Case in which $f = 1$: The Reciprocal Rule for Differentiation If g is as in (b) above, then

$$\left(\frac{1}{g}\right)'(c) = -\frac{g'(c)}{(g(c))^2}.$$

Proof

(a) By a clever use of the 'Add-and-Subtract Trick' one can write

$$\frac{P(x) - P(c)}{x - c} = \frac{f(x)g(x) - f(c)g(c)}{x - c} = \frac{f(x)g(x) - f(c)g(x) + f(c)g(x) - f(c)g(c)}{x - c} =$$

$$\left(\frac{f(x) - f(c)}{x - c}\right) g(x) + f(c)\left(\frac{g(x) - g(c)}{x - c}\right) \qquad (*)$$

It follows, from the differentiability hypotheses on f and g at c, combined with the theorem that differentiability at c implies continuity there, that the first term on the right side of $(*)$ approaches $f'(c)g(c)$ as x approaches c. Likewise, the second term on the right side of $(*)$ approaches $f(c)g'(c)$. The desired result now follows easily.

(b) This is can be proved using a similar 'Add-and-Subtract Trick'; the details are left as an exercise. ■

Side Comment (on the statement and proof of Part (a)) The statement and proof of the Product Rule given above are both standard in calculus texts. In particular, the statement of the formula for the derivative of the product function P is given first, while the proof involves an unmotivated application of the 'Add-and-Subtract Trick'. This Side Comment tries to show how these steps might have arisen.

First of all, the statement of the formula for P' in reality came out of the proof. Indeed, one does not use that statement in carrying out the proof.

First, note that simply writing down the definition of $P'(c)$ leads one to the expression

$$P(x) - P(c) = f(x)g(x) - f(c)g(c) \qquad (**)$$

All that is known about f and g is that they are both differentiable at c, which facts involve the differences $f(x) - f(c)$ and $g(x) - g(c)$. The expression $f(x)g(x) - f(c)g(c)$ on the right side of Equation $(**)$ does not factor in a way that brings, say, $g(x) - g(c)$ into play. However, one may note that the difference $f(x)g(x) - f(x)g(c)$ *does* factor nicely as $f(x)(g(x) - g(c))$, by the Distributive Law from arithmetic. Then it is natural to consider how this 'nicer' expression compares with the right side of $(**)$, obtaining

$$f(x)g(x) - f(c)g(c) = f(x)g(x) - f(x)g(c) + f(x)g(c) - f(c)g(c),$$

as in the proof above. Fortunately, the expression $f(x)g(c) - f(c)g(c)$ can itself be factored as $(f(x) - f(c))g(c)$, so now the quantity of real interest, $f(x)g(x) - f(c)g(c)$, has been written as the sum of *two* terms that both can be factored in ways which allow the introduction of the differentiability hypotheses on f and g. The proof then contiues as above.

The preceding rules show how differentiation behaves when combined with the standard algebraic operations, such as addition and multiplication, and the standard names assigned to these rules relate directly to the names of these operations. The next rule describes how differentiation behaves when combined with the operation of composition, so it would seem reasonable to call it something like 'The Composition Rule for Differentiation'. In reality, however, for historical reasons it is called the 'Chain Rule'.

V.3.3 Theorem (Chain Rule for Differentiation)

Suppose that $h : I \to \mathbb{R}$ is a function, defined on an open interval I, such that h can be expressed on I as the composition $h = g{\circ}f$ of real-valued functions g and f. (This implies that f is defined on I and that g is defined on the image $f[I]$.) Assume that f is differentiable at a point c of I, and that g is differentiable at the corresponding point $d = f(c)$. (The latter fact implies that g is defined on some open interval containing d.) Then the composition $h = g{\circ}f$ is differentiable at c, and one has

$$h'(c) = g'(d){\cdot}f'(c); \text{ that is, } (g{\circ}f)'(c) = g'(f(c)){\cdot}f'(c).$$

Preliminary Discussion In elementary calculus the usual first step to convince students of the naturalness of this Rule is to note that

$$\frac{h(x)-h(c)}{x-c} = \frac{g(f(x))-g(f(c))}{x-c} = \left(\frac{g(f(x))-g(f(c))}{f(x)-f(c)}\right)\cdot\left(\frac{f(x)-f(c)}{x-c}\right) \qquad (*)$$

when $x{\in}I$ and $x \neq c$. Next, one notes that as x approaches c, the second factor on the right side of Equation $(*)$ approaches $f'(c)$. Likewise, as x approaches c the quantity $y = f(x)$ approaches $d = f(c)$ (by Theorem (V.1.5), 'Diffeerentiability implies Continuity'), hence the first factor on the right side of Equation $(*)$ approaches $g'(d)$. This argument can be found, for example, in Cauchy's famous *Leçons sur le calcul differential.*

As modern texts point out, however, this argument has a major gap: it assumes that if $x \neq c$, then one can divide by $f(x) - f(c)$; more precisely, it assumes that $f(x) \neq f(c)$ for all $x \neq c$, at least when x is sufficiently near c. Apparently this gap went unnoticed by many authors for a long time. For instance, as is alluded to in Chapter Quote (3) at the start of this chapter, the famous book 'A Course in Pure Mathematics', by G. H. Hardy, did not fix this gap until its fourth edition in 1925. (Indeed, the first correct proofs apparently appeared only in the 1870s; see the article by H. S. Carslaw in Volume 29 (1923) of the *Bulletin of the American Mathematical Society.*) Modern calculus texts usually provide a rigorous, but sophisticated, proof of a rather different style.

The proof given below is based mainly on Equation $(*)$, but it handles the 'major gap' issue by using the 'sequential' characterization of '$\lim\limits_{x\to c}$' given in Theorem (IV.5.3).

Proof Let $\xi = (x_1, x_2, \ldots x_k, \ldots)$ be a sequence of numbers in I, with $x_k \neq c$, such that $\lim\limits_{k\to\infty} x_k = c$. Let $y_k = f(x_k)$, so that $\lim_{k\to\infty} y_k = d$, by the continuity of f at c. To simplify the notation, for each index k set

$$r_k = \frac{h(x_k)-h(c)}{x_k-c} = \frac{g(y_k)-g(d)}{x_k-c}.$$

In light of Part (c) of Theorem (IV.5.3), it suffices to show that for every sequence ξ as above one has $\lim_{k\to\infty} r_k = g'(d){\cdot}f'(c)$.

Divide the indices k into two disjoint subsets A and B of $\mathbb{N}$:

$$A = \{k{\in}\mathbb{N} : y_k \neq d\} \text{ and } B = \{k{\in}\mathbb{N} : y_k = d\}$$

Note that $\mathbb{N} = A \cup B$, so that at least one of the subsets A or B must be infinite. There are three cases to consider:

Case (i) A is infinite and B is finite;
Case (ii) B is infinite and A is finite;
Case (iii) A and B are both infinite.

Suppose first that Case (i) holds. Then for all sufficiently large k one can write the following analog of Equation $(*)$ above:

$$r_k = \frac{g(f(x_k)) - g(f(c))}{x_k - c} = \left(\frac{g(y_k) - g(d)}{y_k - d}\right) \cdot \left(\frac{f(x_k) - f(c)}{x_k - c}\right) \qquad (**)$$

(Recall that $y_k = f(x_k)$ and that $d = f(c)$, so in this case for all sufficiently large k one has $y_k - d \neq 0$.) Since $\lim_{k \to \infty} y_k = d$ it follows that the first factor on the right side of Equation $(**)$ approaches $g'(d)$ as k approaches infinity. Similarly, the second factor on the right side of $(**)$ approaches $f'(c)$ as k approaches infinity. The Product Rule for Limits then implies that the left side of $(**)$ also approaches a finite limit, namely $g'(d) \cdot f'(c)$. That is, $\lim\limits_{k \to \infty} r_k = g'(d) \cdot f'(c)$, as required.

Suppose next that Case (ii) holds. In this case one has $f(x_k) - f(c) = 0$ for all sufficiently large k, so that one certainly has $\lim_{k \to \infty} r_k = 0$. Likewise one also has $\lim\limits_{k \to \infty} \dfrac{f(x_k) - f(c)}{x_k - c} = 0$. But since f is, by hypothesis, differentiable at c, it follows that $f'(c) = 0$ and thus $\lim_{k \to \infty} r_k = 0 = g'(d) \cdot f'(c)$, as required.

Finally, suppose that Case (iii) holds. Apply Case (i) to the subsequence of the sequence $\rho = (r_1, r_2, \ldots)$ corresponding to the set A of indices; likewise, apply Case (ii) to the subsequence of ρ corresponding to the set B. It then follows that both of these sequences of ρ converge to the same limit, namely $g'(d) \cdot f'(c)$. By the Generalized Odd/Even Limit Theorem, it follows that the sequence ρ itself converges to the same limit, $g'(d) \cdot h(c)$, as required.

Remark Some alternate proofs of the Chain Rule are outlined in the exercises.

The proof given above can be modified to show the following variation of the standard Chain Rule.

V.3.4 Theorem (Modified Chain Rule for Differentiation)

Suppose that $h : I \to \mathbb{R}$ is a function, defined on an open interval I, and that c is a point of I such that $h'(c)$ exists. Suppose further that h can be expressed on I as the composition $h = g \circ f$ of real-valued functions g and f.

(a) Assume that g is differentiable at the point $d = f(c)$ and that $g'(d) \neq 0$. (Recall that this implies that g is defined on some open interval containing d.) Also assume that f is continuous at c. Then f is also differentiable at c, and $f'(c) = h'(c)/g'(d)$.

(b) Similarly, assume instead that f is differentiable at c, with $f'(c) \neq 0$, and that g is continuous at d. Then g is differentiable at d and $g'(d) = h'(c)/f'(c)$.

Proof

(a) As in the proof of the preceding theorem, let $\xi = (x_1, x_2, \ldots x_k, \ldots)$ be a sequence of numbers in I, with $x_k \neq c$, such that $\lim\limits_{k \to \infty} x_k = c$, and let $y_k = f(x_k)$. As before, the hypothesis that f is continuous at c implies $\lim_{k\to\infty} y_k = d$. Set $z_k = \dfrac{f(x_k)-f(c)}{x_k-c}$. It suffices to show that for each such sequence ξ the corresponding sequence $\zeta = (z_1, z_2, \dots)$ converges to $h'(c)/g'(d)$.

As in the preceding proof, divide the indices k into two disjoint subsets A and B of $\mathbb{N}$:

$$A = \{k\in\mathbb{N} : y_k \neq d\} \text{ and } B = \{k\in\mathbb{N} : y_k = d\}$$

Once again, $\mathbb{N} = A\cup B$, so that at least one of the subsets A or B must be infinite, and there are three cases to consider:

Case (i) A is infinite and B is finite;
Case (ii) B is infinite and A is finite;
Case (iii) A and B are both infinite.

Suppose first that Case (i) holds, so that $y_k - d \neq 0$ if k is large enough. The hypothesis $g'(d) \neq 0$ then implies that

$$y_k - d \neq 0 \text{ and } \frac{g(y_k)-g(d)}{y_k-d} \neq 0 \text{ for all sufficiently large } k.$$

For all such k one then has

$$\frac{h(x_k)-h(c)}{x_k-c} = \frac{g(y_k)-g(d)}{x_k-c} = \left(\frac{g(y_k)-g(d)}{y_k-d}\right)\cdot\left(\frac{y_k-d}{x_k-c}\right),$$

and thus

$$\frac{f(x_k)-f(c)}{x_k-c} = \left(\frac{h(x_k)-h(c)}{x_k-c}\right)\Big/\left(\frac{g(y_k)-g(d)}{y_k-d}\right)$$

As k approaches infinity, the numerator on the right approaches $h'(c)$, while the demominator on the right approaches the nonzero number $g'(d)$. Thus the Quotient Rule for Limits implies that the fraction on the left approaches a finite limit, namely $h'(c)/g'(d)$.

Now suppose that Case (ii) holds. Then for all sufficiently large k one has $h(x_k) = h(c)$, so that $\lim\limits_{k\to\infty} \dfrac{h(x_k)-h(c)}{x_k-c} = 0$. By the hypothesis that h is differentiable at c it follows that $h'(c) = 0$. But likewise in this Case one has $f(x_k) = f(c)$ for sufficiently large k, so that $\lim\limits_{k\to\infty} \dfrac{f(x_k)-f(c)}{x_k-c} = 0$. Once again, this fraction approaches a finite limit, which again equals $0 = h'(c)/g'(d)$.

Finally, suppose that Case (iii) holds. Apply Case (i) to the subsequence of ζ which corresponds to the indices in A, and apply Case (ii) to the subsequence of ζ which corresponds to the indices in B. From what has just been shown, both of these subsequences converge to the same number, namely $h'(c)/g'(d)$ (which of course must equal 0). By the Generalized Odd/Even Limit Theorem, it follows that the sequence ζ converges to $h'(c)/g'(d)$, and the desired result follows.

(b) The proof of this part involves a simple modification of the preceding one, and is left as an exercise.

The next result applies the Modified Chain Rule to inverse functions.

V.3.5 Theorem (Inverse-Function Differentiation Rule)

Suppose that $f : I\to\mathbb{R}$ is a continuous strictly monotonic function on an open interval I. Let $J = f[I]$ be the image of I under f, so that by Theorem (IV.4.7) J is also an open interval, f is a bijection of I onto J, and f has an inverse function $g = f^{-1} : J\to I$ which is continuous on J.

Suppose in addition that f is differentiable at a point c of I, and that $f'(c) \neq 0$. Let $d = f(c)$. Then g is differentiable at d, and

$$g'(d) = \frac{1}{f'(c)} = \frac{1}{f'(f^{-1}(d))}. \tag{V.2}$$

Proof Let $h = g \circ f = f^{-1} \circ f$ on I, so that h is the identity function on I. In particular, $h'(c) = 1$. Part (b) of the Modified Chain Rule then implies that g is differentiable at d and that $g'(d) = h'(c)/f'(c) = 1/f'(c)$, as claimed. ■

What follows are some concrete applications of the preceding general differentiation rules.

V.3.6 Examples

Remark The first two examples below were obtained directly from the definition of the derivative in Example (V.1.7). The point here is that they can also be obtained from their simplest cases using the computational rules derived above.

(1) Suppose that k is a nonnegative integer, and let $f_k : \mathbb{R} \to \mathbb{R}$ be the 'k-th-Power Function'. That is,

$$f_k(x) = x^k \text{ for all } x \text{ in } \mathbb{R}.$$

Then f_k is differentiable at all points of $\mathbb{R}$, and one has

$$f_k'(x) = k\,x^{k-1} \text{ for all } x \text{ in } \mathbb{R}. \tag{V.3}$$

Note: It is understood that, in the current context, the expression x^0 is treated as the constant 1, even for $x = 0$. In elementary calculus one is also taught that the expression 0^0 is an 'indeterminant form' and has no definite meaning. These interpretations are in conflict, but in practice this normally causes no confusion.

In light of the preceding Note, the case $k = 0$ reduces to the fact that a constant on $\mathbb{R}$ function has derivative 0 at each point. If $k = 1$, then the formula reduces to $f_1'(x) = 1 \cdot x^0 = 1$, which was shown in Rxample (V.1.7) (1). Now suppose that the formula holds for a given natural number k, and note that $f_{k+1}(x) = f_k(x) \cdot f_1(x)$. Then the Product Rule for Derivatives implies that f_{k+1} is also differentiable on $\mathbb{R}$, and that

$$f_{k+1}(x) = f_k'(x) \cdot f_1'(x) + f_k(x) \cdot f_1'(x) = (k\,x^{k-1}) \cdot x + x^k \cdot (1) = (k+1)\,x^k,$$

as required. The desired result follows by the Principle of Mathematical Induction.

(2) Suppose that k is a *negative* integer. Let $f_k : \mathbb{R} \backslash \{0\} \to \mathbb{R}$ be given by

$$f_k(x) = x^k \text{ for all } x \neq 0.$$

Then f_k is differentiable at all points of $\mathbb{R} \backslash \{0\}$, and one has

$$f_k'(x) = k\,x^{k-1} \text{ for all } x \neq 0. \tag{V.4}$$

Once again the verification is obtained using the Principle of Mathematical Induction, starting with the case $k = -1$ carried out in Example (V.1.7) (4).

(3) Let n be a positive integer, and define the function $f_{1/n} : (0, +\infty) \to \mathbb{R}$ by the rule that for each $x > 0$ is $f_{1/n}(x)$ is $x^{1/n}$, the positive n-th root of x. Thus, $f_{1/n}$ is the inverse on $(0, +\infty)$ of the

function f_n described above, and by Theorem (IV.4.7) $f_{1/n}$ is continuous there. Since $f_n'(x) \neq 0$ if $x > 0$, it now follows from Theorem (V.3.5) that $f_{1/n}$ is differentiable on $(0, \infty)$, and that

$$f_{1/n}'(x) = \frac{1}{f_n'(x^{1/n})} = \frac{1}{n\, x^{(n-1)/n}} = \frac{1}{n}\, x^{-1+1/n} \text{ for } x > 0.$$

(4) Suppose that r is a rational number and $h_r : (0, \infty) \to \mathbb{R}$ is the function given by the rule $h_r(x) = x^r$; see Example (IV.4.8) (2). Express r in the form $r = m/n$, where m and n are integers and $n > 0$, so that $h_r = f_m \circ f_{1/n}$ on $(0, +\infty)$. It follows from the regular Chain Rule that h_r is differentiable on $(0, +\infty)$, and that for each $x > 0$ one has

$$h_r'(x) = f_m'(x^{1/n}) \cdot f_{1/n}'(x) = m\,(x^{1/n})^{m-1} \cdot \left(\frac{1}{n}\right)(x^{1/n-1}) = \frac{m}{n}\, x^{(m-1)/n} \cdot x^{-1+1/n} = \frac{m}{n}\, x^{-1+m/n} = r\, x^{r-1}.$$

It is convenient for ease of future reference to include the following generalization of the preceding examples.

V.3.7 Corollary (The Extended Power Rule for Differentiation – Rational Case)

Suppose that $f : I \to Y$ is differentiable on the open interval I and with values in a nonempty subset Y of $\mathbb{R}$. Let r be a rational number, and suppose that Y is such that $(f(x))^r$ is defined for each x in I. More precisely:

(1) If r is a nonnegative integer, then $Y \subseteq \mathbb{R}$.
(2) If r is a negative integer, then $Y \subseteq \mathbb{R} \setminus \{0\}$.
(3) If r is not an integer, then $Y \subseteq (0, \infty)$.

Define $g : I \to \mathbb{R}$ by the rule $g(x) = (f(x))^r$ for each x in I. Then $g'(x) = r\,(f(x))^{r-1} f'(x)$ for each x in I.

The simple proof is left as an exercise. ■

Remark There is an extension of this result which also holds for exponents r which are irrational. This is discussed later.

V.3.8 Examples

(1) Let $h : \mathbb{R} \setminus \{3\} \to \mathbb{R}$ be given by the rule

$$h(x) = \frac{5x - 17}{7x - 21} \text{ for all } x \neq 3.$$

This function is the ratio of linear functions f and g, where

$$f(x) = 5x - 17 \text{ and } g(x) = 7x - 21 \text{ for all } x.$$

Note that $f'(x) = 5$ and $g'(x) = 7$ for all x. It thus follows that h is differentiable at all points of its domain, and that the Quotient Rule for Derivatives can be used to get

$$h'(x) = \frac{5 \cdot (7x - 21) - (5x - 17) \cdot 7}{(7x - 21)^2} \qquad (*)$$

This answer is correct as it stands. However, by noting that

$$5{\cdot}(7x-21) - (5x-17){\cdot}7 \;=\; (35x-105)-(35x-119) \;=\; 14,$$

one sees that the answer can be simplified to

$$h'(x) \;=\; \frac{14}{(7x-21)^2} \qquad (**)$$

(2) Let h be the same function as in the preceding example, but now compute $h''(x)$. Of course this requires simply differentiating the function h' obtained above. By combining the Constant-Factor and Reciprocal Rules for Derivatives to the function h', one gets (using the form of h' given in Equation $(**)$)

$$h''(x) \;=\; -\frac{2{\cdot}14{\cdot}7}{(7x-21)^3} \;=\; -\frac{196}{(7x-21)^3}$$

Note that this calculation is quicker, and less prone to errors, than the corresponding one based on Equation $(*)$ above.

(3) (a) Suppose that $p:\mathbb{R}\to\mathbb{R}$ is a polynomial function on $\mathbb{R}$; that is, there are constants c_0, $c_1,\ldots\, c_k$ such that $p(x) \,=\, c_kx^k + c_{k-1}x^{k-1} + \,\ldots\, + c_1x + c_0$ for all x in $\mathbb{R}$. Then p is differentiable on $\mathbb{R}$; more precisely, p' is the polynomial given by

$$p'(x) \;=\; k\,a_kx^{k-1} + (k-1)\,a_{k-1}\,x^{k-2} + \;\ldots\; + 2\,a_2\,x + a_1.$$

By an obvious inductive argument one sees that p is a C^∞ function on $\mathbb{R}$, and that for each $m \,=\, 1,2,\,\ldots$ the function $p^{(m)}$ is a polynomial of degree at most $k-m$. In particular, $p^{(m)}(x) \,=\, 0$ for all x in $\mathbb{R}$ if $m \geq k+1$.

(b) Suppose that f is a rational function; that is, there exist polynomial functions p and q, with q not the zero polynomial, such that $f(x) \,=\, p(x)/q(x)$ for all x in $\mathbb{R}$ such that $q(x)\neq 0$. The domain of f is an open set U of the form $U \,=\, \mathbb{R}\backslash S$, where S is the set of all x in $\mathbb{R}$ such that $q(x) \,=\, 0$. (One knows from high-school algebra that S is a finite set; it may be empty.) Then f is differentiable on U, and one has

$$f'(x) \;=\; \frac{g'(x)h(x) - f(x)g'(x)}{\bigl(g(x)\bigr)^2} \text{ for all } x \text{ in } U.$$

It then follows from Part (a) of this example that f' is itself a rational function on the open set U. Once again, one can use induction to conclude that f is of class C^∞ on U, and that each derivative $f^{(m)}$ is a rational function on U.

Side Comments (on calculus pedagogy, Part 1) (1) Calculus teachers are familiar with the phenomenon of students taking an examination who ask ‘Do we need to simplify our answers?’. The usual answer from teachers is ‘Yes, you do’.

Of course, this raises the touchy issue of ‘How simple is simple enough?’ A Justice-Stewart type of response that ‘You’ll know it when you see it’ (see Chapter Quote (1) for Chapter (I)) is probably inadequate. Indeed, sometimes one form of the answer is better (i.e., ‘simpler’) for one purpose, while a different form of the answer is better for a different purpose. For instance, the expressions $\dfrac{2}{1-x^2}$ and $\dfrac{1}{1+x}+\dfrac{1}{1-x}$ both represent the same function for $x \neq \pm 1$. The former expression is ‘simpler’ if the issue is to determine where the function takes on the value

2, while the latter expression is 'simpler' if the the issue is to compute the fifth derivative of the function.

Likewise, in the case of Example (1) above, failure to simplify the expression $5{\cdot}(7x-21)-(5x-17){\cdot}7$ to 14 would not make the answer 'wrong'. However, it would make it harder to compute $h''(x)$, in that errors would be much more likely to enter.

(2) Many calculus students avoid using the 'Constant Factor Rule' for differentiation, preferring to use instead the Product Rule. For example, such a student might compute the derivative of the function $y = 5\,x^3$ as follows:

$$y' = 5'{\cdot}x^3 + 5{\cdot}(x^3)' = 0{\cdot}x^3 + 5{\cdot}3{\cdot}x^2 = 15\,x^2.$$

This is perfectly legal, albeit inefficient. But there is a deeper problem: it happens more often than one might expect that the student will write $5' = 1$, or maybe $5' = 5$, instead of $5' = 0$. If the calculation in question is just one step in, say, a max/min problem, the student may have turned a problem intended to have a quick and easy solution into one in which much valuable exam time is wasted getting the wrong answer or no answer at all. Similarly, students often avoid direct application of the Reciprocal Rule for differentiation to compute $(1/g)' = -g'/g^2$ and instead use the Quotient Rule $(1'{\cdot}g - 1{\cdot}g')/g^2$, opening the real possibility of introducing the error $1' = 1$.

V.4 Significance of the First Derivative

As one knows from elementary calculus, the derivative f' of a function f on an interval contains a great deal of information about the behavior of f on that interval. Much of that information is obtained through a study of the sign of the derivative; that is, by determining whether f' is positive or negative.

Note The approach followed in this section and the next differs substantially from that found in standard calculus texts. Indeed, the approach here is much closer to the original treatment of Lagrange in the eighteenth century, but with the gaps in his arguments filled in.

V.4.1 Theorem (First-Derivative Test for the Nonexistence of Extrema at a Point)

Let $f : [a,b] \to \mathbb{R}$ satisfy $f'(x_0) \neq 0$ at a point x_0 of a closed interval $[a,b]$. If x_0 is either of the endpoints a or b, then $f'(x_0)$ is the appropriate one-sided derivative.

(a) Suppose that $a \leq x_0 < b$. If $f'(x_0) > 0$, then f does not have a maximum for $[a,b]$ at x_0, while if $f'(x_0) < 0$, then f does not have a minimum for $[a,b]$ at x_0.

(b) Suppose that $a < x_0 \leq b$. If $f'(x_0) > 0$, then f does not have a minimum for $[a,b]$ at x_0, while if $f'(x_0) < 0$, then f does not have a maximum for $[a,b]$ at x_0.

(c) Suppose that $a < x_0 < b$. Then f has neither a maximum nor a minimum for $[a,b]$ at x_0.

Proof (a) Note that

$$f'(x_0) = f'_+((x_0)) = \lim_{x \searrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

In particular, the fraction $(f(x) - f(x_0))/(x - x_0)$ has the same sign as $f'(x_0)$ provided $x > x_0$ is sufficiently near x_0. In this case one has $x - x_0 > 0$, and it follows that $f(x) - f(x_0)$ has the same sign as $f'(x_0)$. For such x one then has $f(x) > f(x_0)$ if $f'(x_0) > 0$, while $f(x) < f(x_0)$ if $f'(x_0) < 0$. In the former case $f'(x_0)$ cannot be the maximum value of f on $[a, b]$, while in the latter case it cannot be the minimum value of f on $[a, b]$, as claimed.

(b) A similar proof applies.

(c) This follows immediately from Parts (a) and (b). ■

Remark Part (c) of the preceding result is often phrased in the following equivalent form:

V.4.2 Corollary (First-Derivative Test for Extrema)

Suppose that $f : I \to \mathbb{R}$ is defined on an open interval I and differentiable at a point c of I. If f assumes an extreme value (i.e., maximum or minimum) for I at c, then $f'(c) = 0$.

Alternate phrasing: A *necessary* condition for such f to assume an extreme value at c is that $f'(c) = 0$.

V.4.3 Remarks

(1) The name 'First-Derivative Test for Extrema' given to the preceding corollary is completely standard in calculus texts. Unfortunately, the name is also confusing, especially to students, since in fact one can *never* use this 'test' – by itself – to conclude that a function has a maximum or minimum value at a particular interior point: it is a *necessary*, but not a *sufficient*, condition for this to occur. Many students in elementary calculus have lost points on examinations because of this confusion.

In contrast, it makes good sense to call Theorem (V.4.1) the 'First-Derivative Test for the Nonexistence of Extrema at a Point'. Indeed, this result *does* provide sufficient conditions for a function to not have a extremum at a point.

(2) A point c in an open interval I at which $f'(c) = 0$ is called a **critical point of f**.

Note Some texts say that c is a critical point of f if either f is differentiable at c and $f'(c) = 0$, or f is defined, but not differentiable, at c. We avoid this usage in *This Textbook*.

In real-life mathematics the subject of 'Differential Calculus' is concerned primarily with functions that are differentiable at each point of an interval, and not just at isolated points. The remainder of this section is devoted primarily to such functions.

V.4.4 Theorem (Significance of the Sign of the First Derivative on an Interval)

Let $f : X \to \mathbb{R}$ be a real-valued function with domain $X \subseteq \mathbb{R}$, and suppose that $f'(x)$ is defined, and nonzero, at each point x of an interval $I \subseteq X$. Then:

(a) The derivative function f' is of constant sign on I. More precisely, either $f'(x) > 0$ for all x in I, or $f'(x) < 0$ for all x in I; equivalently, if a and b are numbers in the interval I, then the nonzero numbers $f'(a)$ and $f'(b)$ are of the same sign.

(b) If $f'(x) > 0$ for all x in I, then the original function f is strictly increasing on I. Likewise, if $f'(x) < 0$ for all x in I, then f is strictly decreasing on I. In either case, the image $J = f[I]$ of the interval I under f is an interval of the same type as I: either both are open intervals, both are closed intervals, or both are half-open intervals. Furthermore, the map $f : I \rightarrow J$ is a bijection.

Proof (a) Let a and b be points of I such that $a \neq b$; without lose of generality, suppose that they are labelled so that $a < b$. Since f is differentiable on the interval I, by Theorem (V.1.5) it is certainly continuous on the closed and bounded subinterval $[a, b]$. Thus, by the Extreme-Value Theorem, it must assume both a maximum and a minimum value for $[a, b]$ in $[a, b]$. The hypothesis that $f'(x)$ is never 0 for x in the interval I, combined with Part (c) of the preceding theorem, implies that these extreme values cannot be assumed at an interior point of $[a, b]$. If $f'(a) > 0$, then by Part (a) of that theorem it follows that that this maximum cannot occur at a, so it must occur at b, and nowhere else. Likewise, the minimum of f on $[a, b]$ must occur at a, and nowhere else. In particular, $f(a) < f(b)$. It then follows from Part (b) of the same theorem that $f'(b)$ cannot be negative, so one must have $f'(b) > 0$ as well.

A similar argument shows that if $f'(a) < 0$ then $f'(b) < 0$. In either case, $f'(a)$ and $f'(b)$ must have the same sign, as claimed.

(b) Assume, to be definite, that $f'(x) > 0$ for all x in I, and again let a and b be any numbers in I such that $a < b$. Then from the proof of Part (a) just given, it follows that $f(a) < f(b)$. Since a and b are arbitrary numbers in I with $a < b$, it follows that f is strictly increasing on I, as claimed.

If, instead, $f'(x) < 0$ for every x in I, apply what was just proved to the function $g = -f$.

The remaining claims, that the image set J is an interval of the same tpe as I, and that $f : I \rightarrow J$ is a bijection of I onto J, follow from Theorem (IV.4.7).

Remark The proofs just given are different from – and more direct than – the proofs of the same results found in nearly every calculus text. Indeed, those proofs are based on the so-called 'Mean-Value Theorem' (see below), a result that students find difficult to understand. The genesis of the proofs here is a paper by L. Bers; see [BERS 1967].

The next result is a simple application of Part (a) of the preceding result. It can be of considerable use in analysis, but is often treated as a mere curiosity. Standard texts in elementary calculus rarely even state it; indeed, apparently it went unobserved until Darboux proved it around 1875.

V.4.5 Theorem (The Intermediate-Value Theorem for Derivatives)

Suppose that f is differentiable at each point of an open interval I in $\mathbb{R}$. Let $m = \inf\{f'(x) : x \in I\}$ and $M = \sup\{f'(x) : x \in I\}$; we allow the possibilities $m = -\infty$ and $M = +\infty$. If p is a number such that $m < p < M$, then there exists a number c in I such that $f'(c) = p$.

Proof (by contradiction): Suppose that there is p such that $m < p < M$ but there is no c in I which satisfies $f'(c) = p$. By the Approximation Properties for infimum and supremum, there exist numbers q and r in the set $S = \{f'(x) : x \in I\}$ such that $q < p < r$. By definition of the set S, this implies that there exist numbers a and b in I such that $f'(a) = q$ and $f'(b) = r$, hence for which $f'(a) < p < f'(b)$. Define $g : I \rightarrow \mathbb{R}$ by the rule $g(x) = f(x) - p\,x$ for each x in I. Since $g'(x) = f'(x) - p$, it follows from the 'contradiction hypothesis' on p that for each x in I one has $g'(x) \neq 0$. It then follows from Part (a) of Theorem (V.4.4) that either $g'(x) > 0$ for all x in I, or $g'(x) < 0$ for all x in I. That is, either (i) $f'(x) > p$ for all x in I; or (ii) $f'(x) < p$ for all x in I. Statement (i) contradicts the fact that $q < p$: set $x = a$. Likewise, Statement (ii) contradicts the fact that $p < r$: set $x = b$. ■

V.4.6 Remarks

(1) If f' is constant on I, then $m = M$, so that the hypothesis, 'p is anumber such that $m < p < M$', of the claimed implication is never true, hence in this case the claim itself is automatically true. Of course if f' is *not* constant on I, then $m < M$ so the hypothesis $m < p < M$ holds for infinitely many values of p.

(2) In light of the Bolzano Endpoint-Principles, this result can be rephrased as follows: Suppose that f is differentiable at each point of an open interval I, and let J be the image of I under the derivative function f'. Then either J is a singleton set (if f' is constant on I), or J is also an interval (if f' is not constant on I), although not necessarily an open one.

(3) If the function f' were continuous on I, then the conclusion of the preceding result would follow directly from the standard Intermediate-Value Theorem for continuous functions. The fact that there exist functions f which are differentiable on I, but for which f' fails to be continuous on I – see Remark (V.2.5) (1) above – shows that the current result is not trivial.

The next result provides a useful application of Theorem (V.4.5).

V.4.7 Corollary (The Extended Intermediate-Value Theorem for Derivatives)

Suppose that f and g are differentiable on an open interval I. Assume, in addition, that $g'(x) > 0$ for each x in I. Let $Q(x) = f'(x)/g'(x)$ for each x in I. Suppose that c and d are numbers in I such that $c < d$ and $Q(c) \neq Q(d)$. Then for each number p between $Q(c)$ and $Q(d)$ there exists a number q between c and d such that $Q(q) = p$; that is, $\dfrac{f'(q)}{g'(q)} = p$.

Proof The hypothesis on g' implies that the image of I under g is an open interval J and that g has an inverse g^{-1} on J. Let $h(u) = (f \circ g^{-1})(u)$ for all u in J. By the Chain Rule and the rule for the derivative of an inverse, one has

$$h'(u) = f'(g^{-1}(u)) \cdot (g^{-1})'(u) = \frac{f'(g^{-1}(u))}{g'(g^{-1}(u))}.$$

Since every number x in I can be expressed in exactly one way in the form $x = g^{-1}(u)$ with u in J, it follows that for each x in I one has $f'(x)/g'(x) = h'(u)$ for some u in J. The desired result now follows by applying the usual IVT-D, Theorem 2, to the function h' on the open interval J.

With a little more work, one can also refine Theorem (V.4.4).

V.4.8 Theorem

(a) Suppose that $f'(x) \geq 0$ at each point x of some open interval I in $\mathbb{R}$. Then f is monotonic up on I.

(b) Likewise, suppose that $f'(x) \leq 0$ at each point x of some open interval I in $\mathbb{R}$. Then f is monotonic down on I

(c) In both Part (a) and in Part (b), suppose that there exist numbers x_1 and x_2 in I, with $x_1 < x_2$, such that $f(x_1) = f(x_2)$. Then f is constant on the closed interval $[x_1, x_2]$.

(d) Supppose, in contrast, that g is a function such that $g'(x)$ exists at each point x of the interval I, but g' is *not* of constant sign on I; that is, there exist points x_1 and x_2 in I such that $g'(x_1) < 0$ and $g'(x_2) > 0$. Then g is *not* monotonic, nor is g one-to-one, on I.

Proof

(a) Let k be any natural number, and let $g_k : I \to \mathbb{R}$ be given by the rule $g_k(x) = x/k + f(x)$ for each x in I. Then g_k is differentiable on I, and $g'(x) = 1/k + f'(x) \geq 1/k > 0$ since $f'(x) \geq 0$. It follows from Part (b) of Theorem (V.4.4) that g_k is strictly increasing on I. In particular, if x_1 and x_2 in I satisfy $x_1 < x_2$, then $g_k(x_2) - g(x_1) > 0$. This fact can be written as

$$\frac{x_2 - x_1}{k} + (f(x_2) - f(x_1)) > 0 \text{ for all } \varepsilon > 0.$$

Let the index k approach $+\infty$ and use standard limit laws for sequences to get $f(x_2) - f(x_1) \geq 0$; that is, $f(x_1) \leq f(x_2)$, as claimed.

(b) Apply the results of Part (a) to the function $-f$.

(c) This is an obvious property of functions which are monotonic on an interval.

(d) Without loss of generality assume that $x_1 < x_2$. By the Extreme-Value Theorem, the continuous function g assumes both its maximum and minimum values for the interval $[x_1, x_2]$ somewhere on $[x_1, x_2]$. Since, by hypothesis, one has $g'(x_1) < 0$ and $g'(x_2) > 0$, it follows from Parts (a) and (b) of Theorem (V.4.1) that the minimum value cannot occur at either x_1 or x_2, and thus must occur at some point c such that $x_1 < c < x_2$. In particular, one has $g(x_1) > g(c)$ and $g(c) < g(x_2)$. The former inequality implies that g is not monotonic up on the subinterval $[x_1, c]$, while the latter implies that g is not monotonic down on $[c, x_2]$. It follows that g is certainly not monotonic on the full interval I.

It is now an easy exercise to show that there must exist points u_1 and u_2, with $x_1 < u_1 < c < u_2 < x_1$, such that $g(u_1) = g(u_2)$, so that g is not one-to-one on I, as claimed. ∎

V.4.9 Corollary

Suppose that f satisfies the equation $f'(x) = 0$ for all points of an open interval I in $\mathbb{R}$. Then f is constant on I. (Of course, the converse is also true: if f is constant on I, then $f'(x) = 0$ for all x in I.)

Proof Note that the hypothesis implies that $f'(x) \geq 0$ for all x in I, and that $f'(x) \leq 0$ for all x in I. Thus by Part (a) of the preceding theorem it follows that f is monotonic up on I, while by Part (b) of that theorem it follows that f is monotonic down on I. The only way this can happen is if f is constant on I. ∎

V.4.10 Corollary

Let I be an open interval, and suppose that $f, g : I \to \mathbb{R}$ are differentiable at each point of I.

(a) Assume that $f'(x) \leq g'(x)$ for all x in the open interval I. Then

$$f(x_2) - f(x_1) \leq g(x_2) - g(x_1)$$

for all x_1, x_2 in I such that $x_1 \leq x_2$.

(b) Assume further that there exist x_1 and x_2 in the interval I, with $x_1 < x_2$, such that $f(x_2) - f(x_1) = g(x_2) - g(x_1)$. Then $g - f$ is constant on the closed interval $[x_1, x_2]$; equivalently, $f' = g'$ on $[x_1, x_2]$. In particular, if in addition there exists at least one number c in $[x_1, x_2]$ such that $f(c) = g(c)$, then $B = 0$ and $f = g$ on $[x_1, x_2]$.

(c) The corresponding results obtained by replacing the inequality '$\leq$' throughout (a) and (b) with '$\geq$' are also true.

Proof Apply Theorem (V.4.8) and Corollary (V.4.9) to the function $h = g - f$, then transpose terms in the obvious way. ■

The preceding results involve functions differentiable on an *open* interval. The next theorem provides information about what happens when endpoints are involved.

V.4.11 Theorem

Let $f : (a, b) \to \mathbb{R}$ and $g : (a, b) \to \mathbb{R}$ be differentiable on an open interval (a, b), where the endpoints a and b satisfy $-\infty \leq a < b \leq +\infty$; recall that the 'arrow' notation allows the possibility that f and g are defined at points outside the open interval (a, b) (at least if a or b is finite).

Case 1 Suppose that the endpoints a and b are finite and that f and g are also defined, and continuous, on the closed interval $[a, b]$.

(a) If $f'(x) < g'(x)$ for all x in the open interval (a, b), then for each x_1 and x_2 in $[a, b]$ such that $x_1 < x_2$ one has $f(x_2) - f(x_1) < g(x_2) - g(x_1)$.

(b) If, instead, $f'(x) \leq g'(x)$ for all x in (a, b), then $f(x_2) - f(x_1) \leq g(x_2) - g(x_1)$ for all x_1, x_2 in $[a, b]$ such that $x_1 < x_2$. Furthermore, if there exist x_1 and x_2 in $[a, b]$, with $x_1 < x_2$, such that $f(x_2) - f(x_1) = g(x_2) - g(x_1)$, then there exists a constant C such that $f(x) = g(x) + C$ for all x in the closed interval $[x_1, x_2]$. In particular, if in addition there exists at least one number c in $[x_1, x_2]$ such that $f(c) = g(c)$, then $C = 0$ and $f(x) = g(x)$ for *all* x in $[x_1, x_2]$.

Case 2 Suppose that the left endpoint a is finite and that h is continuous at a. Then the conclusions of Case (1) remain valid if the closed interval $[a, b]$ is replaced throughout that case by the half-open interval $[a, b)$. Similarly, suppose that the right endpoint b is finite and that h is continuous at b. Then the conclusions of Case (1) remain valid if the closed interval $[a, b]$ is replaced throughout that case by the half-open interval $(a, b]$.

Proof

Case 1 When $a < x_1 < x_2 < b$ in Parts (a) and (b), the desired results reduce to the case of open intervals obtained above. Now let x_1 approach a from the right and let x_2 approach b from the left, and use the continuity hypothesis for h. To obtain Part (c), apply Parts (a) and (b) to $h = g - f$.

Case 2 The proof is similar (and slightly easier), and is left as an exercise. ■

Remark Assuming that h is differentiable at the interior points of an interval and merely continuous at the endpoints may seem unduly fussy. However, there are many important functions for which these hypotheses are appropriate, for instance, the function f given by

$$f(x) = \sqrt{1 - x^2} \text{ for } -1 \leq x \leq 1.$$

This function, which arises in the description of the unit circle $x^2 + y^2 = 1$ in $\mathbb{R}^2$, is continuous on the closed interval $[-1,1]$ and differentiable on the open interval $(-1,1)$; but it is *not* differentiable at either endpoint.

V.5 Antiderivatives

In elementary calculus the first important procedure is that of 'differentiation': given a function F, determine its derivative $f = F'$. Students normally have little difficulty becoming proficient at this procedure because of the presence of the computation rules of differentiation discussed above.

Of equal importance in elementary calculus, but considerably more difficult for students, is the reverse procedure: given a function f, determine a function F of which f is the derivative. The present section studies this reverse procedure in some depth.

V.5.1 Definition (Antiderivatives)

Suppose that $f : I \rightarrow \mathbb{R}$ is function defined on an open interval I in $\mathbb{R}$.

(1) A function $F : I \rightarrow \mathbb{R}$ is said to be **an antiderivative of f on I** provided $F'(x) = f(x)$ for all x in I. If such F exists, then one also says that f is **antidifferentiable** on I.

(2) More generally, let k be a positive integer. A function $G : I \rightarrow \mathbb{R}$ is said to be a **k-th order antiderivative of f on I** provided $G^{(k)}(x) = f(x)$ for all x in I.

(3) The process of calculating an antiderivative of a given function f is called **antidifferentiation** of f.

V.5.2 Examples

(1) Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be given by $f(x) = 3\,x^2$ for all x. Then is is clear that F, given by $F(x) = x^3$, is an antiderivative of f. Why is it so clear? Because we have differentiated x^3 before (for instance as the special case $k = 3$ of Example (V.1.7)) and found the result to be $3\,x^2$. It is also clear from the simplest differentiation rules that if C is any constant, then the function G, given for all x by $G(x) = x^3 + C$, is also an antiderivative of f.

(2) More generally, l et F be any standard function from calculus. Differentiate F, and let f be the resulting function. Then F is an antiderivative of f, as is $F + C$ for any constant C.

(3) Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be given by the rule

$$f(x) = \frac{1}{(1+x^2)^{3/2}} \text{ for all } x \text{ in } \mathbb{R}$$

Claim The function $F : \mathbb{R} \rightarrow \mathbb{R}$ given by the rule

$$F(x) = \frac{x}{\sqrt{1+x^2}} \text{ for all } x \text{ in } \mathbb{R},$$

is an antiderivative of f.

To *prove* that this claim is correct is not hard: differentiate the function F, using the rules of differentiation, and simpify algebraically. The harder question to answer is this: Where did the formula for F come from?

(3) Let $f(x) = |x|$ for all x in $\mathbb{R}$. Then $f(x) = x$ if $x \geq 0$, and $f(x) = -x$ if $x < 0$. On the open interval $(0, +\infty)$ the function f has many antiderivatives; namely, any function on $(0, +\infty)$ of the form $G(x) = \dfrac{x^2}{2} + C_1$, where C_1 is constant. Likewise f has infinitely many antiderivatives on the interval $(-\infty, 0)$, namely functions of the form $H(x) = -\dfrac{x^2}{2} + C_2$. To get an antiderivative defined on all of $\mathbb{R}$, choose the constants C_1 and C_2 so that $\lim_{x \nearrow 0} H(x) = \lim_{x \searrow 0} G(x)$. This simply requires $C_1 = C_2$, so let us make the simplest choice, namely $C_1 = C_2 = 0$. Then define $F : \mathbb{R} \to \mathbb{R}$ by the rule

$$F(x) = \begin{cases} H(x) & = & -x^2/2 & \text{if } x < 0 \\ 0 & = & & \text{if } x = 0 \\ G(x) & = & x^2/2 & \text{if } x > 0 \end{cases}$$

This is the same function that appears in Example (V.2.4) above, where it is shown that F' is the absolute-value function.

As is indicated in Example (1) above, the use of the indefinite article 'an' in the definition of 'an antiderivative' is needed: If F is an antiderivative of f on an interval I, then for every constant function C the function $F + C$ is also an antiderivative of F on I. The next result shows that this is the only ambiguity possible for first-order antiderivatives.

V.5.3 Theorem

Suppose that F_1 and F_2 are antiderivatives of f on an open interval I. Then there exists a constant function C such that $F_2(x) = F_1(x) + C$ for all x in I.

Equivalently: The function $F_2 - F_1$ is constant on I.

Proof Note that, by Theorem (V.3.1), the function $F_2 - F_1$ is differentiable on I, and

$$(F_1 - F_2)'(x) = F_1'(x) - F_2'(x) = f(x) - f(x) = 0$$

for all x in I. It then follows from Corollary (V.4.9) that $F_2 - F_1$ is constant on I, as required. ■

Notes (1) Because of the preceding result, if F is a particular antiderivative of a given function f on an open interval I, then it is common to refer to the expression $F + C$ in which C is an 'arbitrary constant', as **the general antiderivative of f on I.**

(2) The corresponding ambiguity for higher-order antiderivatives is more complicated, and is considered later; see Theorem (**??**).

V.5.4 Corollary

(a) Suppose that $f : I \to \mathbb{R}$ is a function which has an antiderivative on an open interval I. Let c be any point in I, and let A be any real number. Then there exists a unique antiderivative F of f on I such that $F(c) = A$.

More precisely, if $G : I \to \mathbb{R}$ is any antiderivative of f on I, then the unique F with this property is given by

$$F(x) = A + G(x) - G(c) \qquad (*)$$

In particular, if $A = 0$ then the formula reduces to

$$F(x) = G(x) - G(c).$$

(b) Suppose that $f : I \rightarrow \mathbb{R}$ is a function defined on an open interval I and that a and b are elements of I with $a < b$. Suppose that there are numbers x_1, $x_2, \ldots x_k$ with $a < x_1 < x_2 < \ldots < x_k < b$ such that f has an antiderivative on each of the subintervals $[a, x_1]$, $[x_1, x_2], \ldots [x_{k-1}, x_k]$, $[x_k, b]$. Then for each A in $\mathbb{R}$ there exists a unique antiderivative $F : [a, b] \rightarrow \mathbb{R}$ of f on $[a, b]$ such that $F(a) = A$.

Proof

(a) Existence If G is any antiderivative of F on I, then clearly the function F given by Equation $(*)$ is also an antiderivative of f, since it differs from G by the constant $A - G(c)$. Furthermore, one computes that $F(c) = A + G(c) - G(c) = A$, as required.

Uniqueness Suppose that F_1 and F_2 are both antiderivatives of f on I such that $F_1(c) = A$ and $F_2(c) = A$. By the preceding theorem there exists a constant function C such that $F_2(x) - F_1(x) = C$ for all x in I. In particular, this condition must hold when $x = c$; that is,

$$0 = A - A = F_1(c) - F_1(c) = C,$$

so $C = 0$ and $F_2 = F_1$ on I, as claimed.

(b) For notational convenience, set $x_0 = a$ and $x_{k+1} = b$. By Part (a), f has a unique antiderivative $F_1 : [x_0, x_1] \rightarrow \mathbb{R}$ on $[x_0, x_1]$ such that $F_1(x_0) = A$. Then by Part (a) again f has a unique antiderivative $F_2 : [x_1, x_2] \rightarrow \mathbb{R}$ on $[x_1, x_2]$ such that $F_2(x_1) = F_1(x_1)$. Continuing on this way, one obtains functions F_1, $F_2, \ldots F_{k+1}$ such that

(i) for each $j = 1, 2, \ldots k + 1$, F_j is an antiderivative of f on the subinterval $[x_{j-1}, x_j]$.

(ii) $F_1(x_0) = A$; and for each $j = 1, 2, \ldots k$, $F_{j+1}(x_j) = F_j(x_j)$.

Now define $F : [a, b] \rightarrow \mathbb{R}$ by the rule

$$F(x) = F_j(x) \text{ if } x \in [x_{j-1}, x_j].$$

It is easy to show by Property (ii) above that $F(x)$ is well-defined at all x in $[a, b]$, even at $x = x_j$. It is also easy to show by using one-sided limits at the points x_j that $F'(x) = f(x)$ for all x in $[a, b]$; the details are left to the reader. The desired result now follows. ■

V.5.5 Examples

(1) Suppose that $f : \mathbb{R} \rightarrow \mathbb{R}$ is a monomial, in the sense of high-school algebra; that is, there is a nonzero constant a and a nonnegative integer k such that $f(x) = a\, x^k$ for all x in $\mathbb{R}$. (Note that if k is even, then $f(-x) = f(x)$ for all x, while if k is odd then $f(-x) = -f(x)$ for all x.) The function f has a unique antiderivative F on $\mathbb{R}$ which is also a monomial, namely the antiderivatives F such that $F(0) = 0$. Indeed, this antiderivative is given by $F(x) = b\, x^{k+1}$, where $b = a/(k + 1)$. Note that if f is a monomial of even degree, then F is a monomial of odd degree. Similarly, if f is of odd degree, then F is of even degree. In the former case one has $f(-x) = f(x)$ and $F(-x) = -F(x)$, while in the latter case one has $f(-x) = -f(x)$ and $F(-x) = F(x)$.

(2) More generally, suppose that I is an open interval of the form $(-b, b)$, where $b > 0$; the case $b = +\infty$ is allowed. Recall, from high-school algebra, that a function $g : I \rightarrow \mathbb{R}$ is said to be an **even function on I** provided $g(-x) = g(-x)$ for all x in I. Likewise, a function $h : I \rightarrow \mathbb{R}$ is said to be an **odd function on I** provided $h(-x) = -h(x)$ for all x in I.

Suppose that $f : I \rightarrow \mathbb{R}$ is a real-valued function with domain I, and that f has an antiderivative on I. Let $F : I \rightarrow \mathbb{R}$ be the unique antiderivative of f on I such that $F(0) = 0$.

(i) If f is an even function on the interval I, then F is an odd function on I.

(ii) If f is an odd function on I, then F is an even function on I.

Indeed, consider the functiom $G : I \to \mathbb{R}$ given by the rule $G(x) = F(-x)$ for all x in I. By the Chain Rule one sees that $G'(x) = -F'(-x) = -f(-x)$, so that $(-G)'(x) = f(-x)$.

In Case (i) one has $f(-x) = f(x)$, so that $(-G)'(x) = f(x)$. Since clearly $-G(0) = 0$, it follows that in this case one has $-G(x) = F(x)$; that is, $-F(-x) = F(x)$, which implies that F is an odd function, as claimed. Likewise, in Case (ii) one gets $(-G)'(x) = -f(x)$, which implies that $F(-x) = F(x)$, so that F is an even function, as claimed.

V.5.6 Remarks

(1) The use of the word 'antidifferentiation', to indicate the process opposite to the process of 'differentiation', seems reasonable; likewise for naming the result of that process an 'antiderivative'. Indeed, the 'antidifferentiation/antiderivative' terminology seems to be universal in the modern calculus texts.

However, for most of the centuries since calculus was first developed the words used instead of 'antidifferentiation' and 'antiderivative' were 'integration' and 'integral'; some authors used the word 'primitive' instead 'integral', and sometimes the word 'indefinite' is used before 'integral'.

Why is it important for you, the modern reader, to know this? Because the 'integration/integral' terminology for these concepts is still widely used; thus if you encounter it, you need to know what it means. Furthermore, although the use of the 'antiderivative' *terminology* , the *notation* used in connection with this concept remains stuck in the eighteenth century. More precisely, the way one writes the statement '$F + C$ is the general antiderivative of f' in mathematical symbols is this:

$$F(x) + C = \int f(x)\,dx$$

In the expression $\int f(x)\,dx$ the symbol $\int$ is called the **integral sign**, the function $f(x)$ is called the **integrand**, and the 'arbitrary constant' C is called the **constant of integration**. Even in those elementary-calculus textbooks which consistently use the 'antiderivative' terminology, the section in which one learns how to compute antiderivatives is normally titled something like 'Techniques of Integration', not 'Techniques of Antidifferentiation'.

(2) There is good reason to assert that the fundamental goal of elementary single-variable calculus is this:

Given a function f, to find its derivative; and given a function g, to find its (general) antiderivative.

Indeed, this is the significance of Chapter Quote 1 at the start of this chapter, namely

'*6accdæ13eff7i3l9n4o4qrr4s8t12vx*'.

Isaac Newton, one of the cofounders of calculus in the seventeenth century, used these letters as an anagram to disguise the following statement:

> *Data æquatione quotcunque fluentes quantitates involvente, fluxiones invenire: et vice versâ.*

That is,

> *Given an equation involving any number of fluent quantities, to find the fluxions: and vice versa.*

In Newton's terminology, a 'fluxion' is the rate of change (with respect to time) of a quantity; that is, a derivative; and a 'fluent' is the quantity whose change produces a give fluxion; that is, an antiderivative. Thus, according to Newton the purpose of calculus is to take derivatives and find antiderivatives.

The concepts of derivative and antiderivative are so intimately related that it should not be a surprise that for each fact realting to the one concept there is a corresponding fact relating to the other. The next result restates Corollary (V.4.10) in a format which handles the existence of more than one antiderivative in a simple manner.

V.5.7 Corollary (of Corollary (V.4.10))

Suppose that f and g are functions which have antiderivatives on an open interval I in $\mathbb{R}$, and assume that $f(u) \leq g(u)$ for all u in I. Let a be a point of I, and let F and G denote the antiderivatives of f and g on I such that $F(a) = G(a) = 0$.

(a) If $a \leq x$, then

$$F(x) \leq G(x),$$

with equality if, and only if, $F = G$ on the set $\mathrm{Seg}\,[a, x]$.

(b) If $x \leq a$, then

$$G(x) \leq F(x),$$

with equality if, and only if, $F = G$ on Seg [a,x].

Proof The result is simply a restatement of Corollary (V.4.10) ■

V.5.8 Examples

Throughout these examples I is an open interval in $\mathbb{R}$, $: I \to \mathbb{R}$ is a function with domain I, a and x are numbers in the interval I such that $a < x$. (The restriction to $a < x$ is to simplify the example; the case $x \leq a$ is considered later.)

(1) Suppose that f' is defined on I and that M_1 is a number such that $f'(u) \leq M_1$ for all u in $[a, x]$; let g_1 be the constant function such that $g_1(u) = M_1$ for all u in I. (The reason for the subscript will be clear soon.) Note that the antiderivative of f on I with value 0 at a is clearly $F_1 : I \to \mathbb{R}$ given by the formula $F_1(u) = f(u) - f(a)$. Likewise, the antiderivative of g_1 with value 0 at a is given by $G_1(u) = M_1\,(u - a)$. It follows from the preceding corollary that

$$f(x) - f(a) \leq M_1\,(x - a),$$

with equality if, and only if, $f(u) - f(a) = M_1\,(u - a)$ for all u in $[a, x]$. A similar argument shows that if m_1 is a number such that $m_1 \leq f'(u)$ for all u in $[a, x]$, then

$$m_1\,(x - a) \leq f(x) - f(a),$$

with equality if, and only if, $f(u) - f(a) = m_1\,(u - a)$ for all u in $[a, x]$.

(2) Now suppose that f'' is defined on I, and that $m_2 \leq f''(u) \leq M_2$ for all u in $[a, x]$. Apply the results of the preceding example to get

$$m_2\,(x - a) \leq f'(x) - f'(a) \leq M_2\,(x - a),$$

with equality on the left if, and only if, ???

The following question is natural:

> 'Under what circumstances does a given function have an antiderivative on a given interval?'.

One knows from elementary calculus many examples of functions whose antiderivatives can be computed. In contrast, it is easy to provide simple examples of functions $f : I \rightarrow \mathbb{R}$ which fail to have an antiderivative on an open interval I. For example, if $f : \mathbb{R} \rightarrow \mathbb{R}$ is the Dirichlet function, then there is no function $F : \mathbb{R} \rightarrow \mathbb{R}$ such that $F'(x) = f(x)$ for all x. This is obvious because the Dirichlet function clearly fails to possess the Intermediate-Value Property on any subinterval of $\mathbb{R}$. In other words, a complete answer to this question is likely to be complicated.

The next result generalizes the method used in the 'Absolute Value' example above.

V.5.9 Theorem

Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$, and let $\mathcal{D} = \{(x_0,y_0),(x_1,y_1),\ldots(x_k,y_k)\}$ be a set of 'data points' in $\mathbb{R}^2$ such that $a = x_0 < x_1 < \ldots < x_k = b$, and let $g : \mathbb{R} \rightarrow \mathbb{R}$ be the corresponding continuous piecewise-linear interpolating function through these points; see Example (IV.1.4) (6). Then for each real number c in $\mathbb{R}$ the function g has a unique antiderivative G on $\mathbb{R}$ such that $G(c) = 0$.

Proof It suffices to show that g has an antiderivative G which takes on the value 0 at $c = a = x_0$; the case for general choice of c follows easily.

Note that for each index $j = 1, 2, \ldots k$, if x satisfies the condition $x_{j-1} \leq x \leq x_j$, then by definition

$$g(x) = y_{j-1} + a_j\,(x - x_{j-1}) \text{ where } a_j = \frac{y_j - y_{j-1}}{x_j - x_{j-1}}.$$

It is clear that, on the inverval $[x_{j-1}, x_j]$, every function of the form $G_j(x) = a_j\,(x - x_{j-1})^2/2 + y_j\,(x - x_{j-1}) + c_j$, where c_j can be any real number, is an antiderivative of g on that interval such that $G_j(x_{j-1}) = c_j$. (At the endpoints of this interval G'_j refers to the appropriate one-sided derivatives.) The remainder of the construction of the desired function G is to choose the constants $c_1,\ c_2, \ldots\ c_k$ so that the antiderivatives $G_1,\ G_2, \ldots$ 'fit' together properly at the points x_j; more precisely:

The case $j = 1$ Since, as is observed above, one has $G_1(x_0) = c_1$, one must choose $c_1 = 0$.

The Case $j = 2$ Choose c_2 so that $G_1(x_1) = G_2(x_1)$. Since $G_2(x_1) = c_2$, this requires that one should choose $c_2 = G_1(x_1)$.

The case of general j is carried out similarly, so that $c_j = G_{j-1}(x_j)$ for each j.

Finally, define $G : [a,b] \rightarrow \mathbb{R}$ by the rule that if $x_{j-1} \leq x \leq x_j$, then $G(x) = G_j(x)$. One can apply Theorem (V.1.4) to conclude that $G'(x) = g(x)$ for all x in $[a,b]$. ■

Remark It is easy to show, using only simple Euclidean geometry, that if g is as above and G is any anitderivative of g on $[a,b]$, then $G(b) - G(a)$ equals the signed area between the graph $y = g(x)$ and the horizontal axis for $a \leq x \leq b$. ('Signed area' refers to the idea that area above the horizontal axis is positive, while area below that axis is negative.)

V.6 Significance of Derivatives of Higher Order

We have just seen that first derivative f' gives useful information about a given function f on an interval. Thus it is natural to ask what information derivatives of higher order can provide. The key is the following generalization of the construction in Part (c) of Corollary (**??**).

V.6.1 Example

Suppose that for some natural number n the function $f : I \to \mathbb{R}$ is n-times differentiable on an open interval I. Let c be any fixed number in I. Let $F_{n-1} : I \to \mathbb{R}$ denote the c-antiderivative of $f^{(n)}$ on I (see Remark (**??**)). Then F_{n-1} is given by the formula

$$F_{n-1}(x) \;=\; f^{(n-1)}(x) - f^{(n-1)}(c) \text{ for all } x \text{ in } I.$$

Similarly, the c-antiderivative F_{n-2} of F_{n-1} on I, which of course is the second-order c-antiderivative of $f^{(n)}$, is given by

$$F_{n-2}(x) \;=\; f^{(n-2)}(x) - f^{(n-2)}(c) - f^{(n-1)}(c)\,(x-c) \text{ for all } x \text{ in } I.$$

Continuing this way, one sees that the n-th order c-antiderivative of $f^{(n)}$ on I is the function $F_0 : I \to \mathbb{R}$ given by

$$F_0(x) \;=\; f(x) - f(c) - f'(c)\,(x-a) - \frac{f''(c)}{2}\,(x-c)^2 - \;\ldots\; - \frac{f^{(n-1)}}{(n-1)!} \text{ for all } x \text{ in } I.$$

Note that repeated use is made of the fact that if $g(x) \;=\; B\,(x-c)^m$ for some constant B, then the c-antiderivative of g is given by $\dfrac{B}{m+1}\,(x-c)^{m+1}$; see Example (**??**). In the case at hand, the constant B is of the form $f^{(k)}(c)/k!$.

The preceding calculation brings to light, in a natural way, certain polynomials which are associated with a function f having $n-1$ derivatives.

V.6.2 Definition

Let $f : I \to \mathbb{R}$ be a function which is n-times differentiable on an open interval I, where $n \in \mathbb{N}$. Let c be a point in I. The **Taylor polynomial of order $n-1$ of a function f centered at c**, denoted $\boldsymbol{p_{(f;n-1;c)}}$, is given by the rule

$$p_{(f;n-1;c)}(x) \;=\; f(c) + f'(c)\,(x-c) + \frac{f''(c)}{2}\,(x-c)^2 + \ldots + \frac{f^{(n-1)}(c)}{(n-1)!}\,(x-c)^{n-1}.$$

The expression $f(x) \approx p_{(f;n-1;c)}(x)$ is called the **Taylor approximation** of $f(x)$ of order $n-1$ at c; the difference $f(x) - p_{(f;n-1;c)}(x)$ is the **error** of this approximation.

The Taylor approximation is widely used in many applied areas, mainly because it is possible to estimate the size of the corresponding error.

V.6.3 Taylor's Theorem

Suppose that $f : I \to \mathbb{R}$ is n-times differentiable on an open interval I, where n is a natural number. Let c be a number in I.

(a) Let x be a number in I such that $x > c$, and suppose that for M_n is a quantity such that $f^{(n)}(u) \le M_n$ for all u in the closed interval $[c, x]$; the value $M_n = +\infty$ is allowed. Then one has

$$f(x) - p_{(f;n-1;c)}(x) \le \frac{M_n}{n!}(x-c)^n$$

with equality if, and only if, $f(u) - p_{(f;n-1;c)}(u) = \frac{M_n}{n!}(u-c)^n$ for every u in $[c, x]$; equivalently, if, and only if, $f^{(n)}(u) = M_n$ for all u in the interval $[c, x]$.

(b) Again let x be a number in I such that $x > c$, but now suppose, instead, that m_n is a quantity such that $m_n \le f^{(n)}(u)$ for all u in $[c, x]$; the value $m_n = -\infty$ is allowed. Then one has

$$\frac{m_n}{n!}(x-c)^n \le f(x) - p_{(f;n-1;c)}(x),$$

with equality if, and only if, $f(u) - p_{(f;n-1;c)}(u) = \frac{M_n}{n!}(u-c)^n$ for every u in $[c, x]$; equivalently, if, and only if, $f^{(n)}(u) = m_n$ for each u in $[c, x]$.

(c) Now let x be any number in I such that $x < c$, and suppose that m_n and M_n are quantities such that $m_n \le f^{(n)}(u) \le M_n$ for each u in the closed interval $[x, c]$; as usual, the values $m_n = -\infty$ and $M_n = +\infty$ are allowed. Then one has

$$(-1)^n \frac{m_n}{n!}(x-c)^n \le (-1)^n \Big(f(x) - p_{(f;n-1;c)}(x)\Big) \le (-1)^n fracM_n n!\,(x-c)^n.$$

Furthermore, one has equality on the left if, and only if, $f(u) - p_{(f;n-1;c)}(u) = \frac{M_n}{n!}(u-c)^n$ for each u in $[c, x]$; equivalently, if, and only if, $f^{(n)}(u) = m_n$ for each u in $[c, x]$. Similarly, one has equality on the right if, and only if, $f(u) - p_{(f;n-1;c)}(u) = \frac{M_n}{n!}(u-c)^n$ for every u in $[c, x]$; equivalently, if, and only if, $f^{(n)}(u) = M_n$ for every u in $[c, x]$.

(d) Let n be a positive integer, and suppose that $f : I \to \mathbb{R}$ is n-times differentiable on an open interval I. Let c be any number in I. Then for each x in I with $x \ne c$, there exists a number q strictly between x and c such that

$$f(x) - p_{(f;n-1;c)}(x) = \frac{f^{(n)}(q)}{n!}(x-c)^n.$$

This last equation often is written in the equivalent form

$$\frac{f(x) - p_{(f;n-1;c)}(x)}{(x-c)^n} = \frac{f^{(n)}(q)}{n!}.$$

Proof The simple proofs of Parts (a) and (b) are left as exercises.

(c) Let $J = \{y : -y \in I\}$, and define $g : J \to \mathbb{R}$ by the rule $g(y) = f(-y)$ for each y in J. Set $b = -c$. It follows directly from repeated use of the Chain Rule that g is n-times differentiable on J. More precisely, if y is in J and $x = -y$ is the coreesponding element of I, then one has

$$g^{(k)}(y) = (-1)^k f^{(k)}(-y) = (-1)^k f^{(k)}(x)$$

for each $j = 0, 1, \ldots n$. It follows easily for such k, x and y that $g^{(k)}(y)(y-b)^k = f^{(k)}(c)(x-c)^k$, hence

$$g(y) - p_{g;n-1}; b)(y) = f(x) - p_{f;n-1}; c)(x).$$

From the hypothesis $m_n \leq f^{(n)}(u) \leq M_n$ for u in $[x, c]$, and the fact that $(-1)^n\,(x-c)^n = (c-x)^n > 0$, it then follows that

$$\frac{m_n}{n!}\,(-1)^n(x) \leq (-1)^n\,g^{(n)}(v) \leq M_n \text{ for each } v \text{ in } [b, y].$$

Apply these results, together with the results of Parts (a) and (b), to the function $(-1)^n\,g$ to get the desired result.

(d) Case 1 Suppose first that $x > c$, and let $S = \{f^{(n)}(u) : u{\in}[c, x]\}$. Let $\hat{m}_n = \inf S$ and $\hat{M}_n = \sup S$. It is clear that $\hat{m}_n$ and $\hat{M}_n$ can be used for m_n and M_n, respectively, in Parts (a) and (b) above. More precisely, $\hat{m}_n$ is the largest value of m_n, and $\hat{M}_n$ is the smallest of M_n, that satisfy the hypotheses in Parts (a) and (b).

(i) Suppose that $\hat{m}_n = \hat{M}_n$, so that $f^{(n)}(u) = \hat{M}_n$ for all u in $[c, x]$. Then the 'if' portions of the 'equality' statements in (a) and (b) imply that

$$\frac{\hat{m}_n}{n!}\,(x-c)^n = f(u) - p_{(f;n-1;c)}(u) = \frac{\hat{M}_n}{n!}(u-c)^n \text{ for all } u \text{ in } [c, x].$$

In particular, in this case one can choose q to be any interior pointof the interval $[c, x]$; for example, $q = (c+x)/2$ works.

(ii),Suppose now that $m_n < M_n$ so that $f^{(n)}$ is not constant on the interval $[c, x]$. Then the 'only if' portions of the same statements imply that

$$\frac{\hat{m}_n}{n!}\,(x-c)^n < f(x) - p_{(f;n-1;c)}(x) < \frac{\hat{M}_n}{n!}(u-c)^n \text{ for all } u \text{ in } [c, x].$$

However, the definitions of $\hat{m}_n$ and $\hat{M}_n$ here, combined with with the Intermediate-Value Theorem for Derivatives, implies that every number between $\hat{m}_n$ and $\hat{M}_n$ is of the form $f^{(q)}$, and thus every number between $\dfrac{\hat{m}_n}{n!}\,(x-c)^n$ and $\dfrac{\hat{M}_n}{n!}\,(x-c)^n$ is of the form $\dfrac{f^{(n)}}{(}q)n!\,(x-c)^n$, for some q in (c, x). The desired result now follows.

Case 2 If, instead, one supposes that $x < c$, then the desired result follows in a similar manner from Part (c) above; a key observation is that $(-1)^n(x-c)^n = (c-x)^n > 0$.

V.6.4 Corollary (The Mean-Value Inequality)

Suppose that $f : I \to \mathbb{R}$ is differentiable on an open interval I. Let a and b be numbers in I such that $a < b$, and suppose that m and M are quantities such that $m \leq f'(u) \leq M$ for all u in the closed interval $[a, b]$; the values $m = -\infty$ and $M = +\infty$ are allowed. Then one has

$$m\,(b-a) \leq f(b) - f(a) \leq M\,(b-a),$$

with equality on either side being an equality if, and only if, f' is constant on the interval $[a, b]$.

The simple proof is left as an exercise. ■

V.6.5 Remarks

(1) The name 'Taylor' refers to the English mathematician Brook Taylor (1685-1731), although the preceding theorem is actually due to the Italian/French mathematician Joseph-Louis Lagrange (1736-1813).

(2) Many texts use the phrase 'Taylor polynomial of *degree* $n-1$' instead of 'of *order* $n-1$'. The problem with that terminology is that if $f^{(n-1)}(c) \,=\, 0$, then the degree of the polynomial $p_{(f;n-1;c)}$ is strictly less than $n-1$.

(3) The error $f(x) - p_{(f;n-1);c}$ is often called the **remainder** in the given Taylor approximation, and abbreviated as $R_{n-1}(x)$. That is, $R_{n-1}(x)$ is what 'remains' when one subtracts the approximate value $p_{(f;n-1;c)}(x)$ from the exact value $f(x)$ of f at x, so that one can write

$$f(x) \,=\, p_{(f;n-1);c}(x) + R_{n-1}(x).$$

In particular, the expression $f^{(n)}(q)\,(x-c)^n$ appearing in Part (d) above is called the **Lagrange form** of the remainder. There is a very different formulation of the same remainder, due to Cauchy, which must be postponed to a later chapter.

The following somewhat weaker result is sufficient for many applications, and covers both the cases $x \,>\, c$ and $x \,<\, c$.

V.6.6 Corollary

Suppose that for some number B_n one has $|f^{(n)}(x)| \le B$ for all x in some open interval I. Let c be any point of I. Then

$$|f(x) - p_{(f;n-1;c)}(x)| \le \frac{B}{n!}|x-a|^n \text{ for all } x \text{ in } I.$$

The simple proof is left as an exercise. ■

Special Case – The Mean-Value Theorems for Derivatives

The case $n \,=\, 1$ Theorem (V.6.3) corresponds to the simplest Taylor approximation, namely $f(x) \approx f(c)$. The corresponding inequalities from Parts (a) and (b) of that theorem is of the form

$$m_1\,(x-c) \le f(x) - f(c) \le M_1\,(x-c) \text{ for all } x \text{ in } I \text{ such that } x \,>\, c,$$

where m_1 is a lower bound for f' on $[c,x]$ and M_1 is the corresponding upper bound. Divide by the positive quantity $x-c$ to get the equivalent formulation

$$m_1 \le \frac{f(x)-f(c)}{x-c} \le M_1.$$

In texts for elementary calculus the letters c and x are usually replaced by the letters a and b, respectively, so that $a \,<\, b$, and the preceding result is written in the more familiar form

$$m_1 \le \frac{f(b)-f(a)}{b-a} \le M_1 \qquad (*)$$

Furthermore, one gets equality on either end of $(*)$ if, and only if, $m_1 = f'(u) = M_1$ for each u in $[a,b]$. Since $\dfrac{f(b)-f(a)}{b-a} = \dfrac{f(a)-f(b)}{a-b}$, it is clear that Inequality $(*)$ remains valid if, instead, $b < a$.

For historical reasons one refers to $(*)$ as the **Inequality Formulation of the Mean-Value Theorem for Derivatives**; or, more briefly, the **Mean-Value Inequality**.

Similarly, the case $n = 1$ in Part (d) of Theorem (V.6.3) usually is written

$$\frac{f(b)-f(a)}{b-a} = f'(q) \text{ for some } u \text{ such that } a < u < b.$$

This result is called the **Equation Formulation of the Mean-Value Theorem for Derivatives**; or, more briefly, the **Mean-value Equation**.

The next result is an obvious corollary of the Mean-Value Equation.

V.6.7 Corollary (Rolle's Theorem)

Suppose that $f : [a,b] \rightarrow \mathbb{R}$ is differentiable on an open interval (a,b) and continuous on the corresponding closed interval $[a,b]$. Assume further that $f(a) = f(b)$. Then there exists a number q such that $a < q < b$ such that $f'(c) = 0$.

Remark The order in which the topics are presented in the last two sections of *This Textbook* is essentially the reverse of the order followed in standard texts on elementary calculus. More precisely, such texts start the discussion with the statement and proof of Rolle's Theorem. This theorem is used to prove the Mean-Value Equation, which is then used to prove the results relating the sign of the derivative and monotonicity. In such texts the results concerning Taylor's Theorem normally occur much later, and are proved by a clever application of Rolle's Theorem.

One way to view this is that often in mathematics there is more than one way to develop a subject, and that it is useful for learners to see alternate approaches. In the next Side Comment other reasons for preferring the approach used in *This Textbook* are given.

Side Comment (on reducing the roll of Rolle's theorem) NEED TO WRITE

V.6.8 Example

In physics and engineering the Taylor approximation is used frequently to obtain useful values for processes that are subjet to known physical laws. For instance, the analysis of vibrating membranes leads to the following differential equation: $f'' + f = 0$, where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a C^2 function to be determined. Note that this equation implies that f is actually a C^∞ function. Indeed, if written as $f'' = -f$, it implies that f'' is also a C^2 function, hence that f is actually a C^4 function and that $f^{(4)} = -f^{(2)} = f$. Repeat this argument to get that f is C^∞, and that

$$f''' = -f'; \quad f^{(4)} = f; \quad f^{(5)} = f'' = -f; \quad f^{(6)} = -f'; \text{ and so on.}$$

In particular, all the derivatives $f^{(n)}$ with $n \geq 2$ can be expressed simply in terms of f and f'. It follows that for each c in $\mathbb{R}$ the coefficients of the Taylor polynomials of f at c are determined by the numbers $f(c)$ and $f'(c)$:

$$\frac{f''(c)}{2!} = -\frac{f(c)}{2!}; \quad \frac{f^{(3)}(c)}{3!} = -\frac{f'(c)}{3!}; \quad \frac{f^{(4)}(c)}{4!} = \frac{f(c)}{4!}; \text{ and so on.}$$

To simplify this example, consider the case $c = 0$ and $f(0) = 1$, $f'(0) = 0$. One then gets

$$p_{(f;0;0)}(x) = p_{(f;1;0)}(x) = 1; \quad p_{(f;2;0)}(x) = p_{(f;3;0)}(x) = 1-\frac{x^2}{2}; \quad p_{(f;4;0)}(x) = p_{(f;5;0)}(x) = 1-\frac{x^2}{2}+\frac{x^4}{4!};$$

and so on. To be definite, consider the case $n = 5$, so that the corresponding Taylor approximation of order 4 takes the form

$$f(x) \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!}.$$

This approximation is of little value since it gives no indication of how accurate it is. However, if one can obtain upper and lower bounds on $f^{(5)}$, then one can obtain such information.

In this case there is a trick: multiply both sides of the original equation $f'' + f = 0$ by f' to get $f''{\cdot}f + f{\cdot}f' = 0$. The left side of this last equation is seen to be $\frac{1}{2}\left((f')^2 + f^2\right)'$, so that $(f')^2 + f^2$ must be some constant B. Substitute $c = 0$ into f' and f to get $0^2 + 1^2 = B$, so that $(f'(x))^2 + (f(x))^2 = 1$ for all x. In particular, one has $-1 \leq f(x) \leq 1$ and $-1 \leq f'(x) \leq 1$, so that $-1 \leq f^{(n)}(x) \leq 1$, for all x. In particular, one has

$$-\frac{x^5}{5!} \leq f(x) - \left(1 - \frac{x^2}{2} + \frac{x^4}{4!}\right) \leq \frac{x^5}{5!} \text{ for all } x > 0.$$

If, say, $x = 1$, then this says the corresponding Taylor approximation of $f(1)$ has error less than $1/5! = 1/120 < 0/01$. By choosing larger values of n one can approximate the value $f(1)$ with the value of an appropriate Taylor polynomial to any desired accuracy.

V.6.9 Standard Form for the Taylor Remainder

The 'inequality' form of the remainder/error in the Taylor approximation, as given in Theorem (V.6.3), is not the version that one finds in most calculus texts. Instead, the following formulation is standard.

V.6.10 Theorem

Let n be a positive integer, and suppose that $f : I \to \mathbb{R}$ is n-times differentiable on an open interval I. Let c be any number in I. Then for each x in I with $x \neq c$ there exists a number u strictly between x and c such that

$$f(x) - p_{(f;n-1;c)}(x) = f_{(n)}(u)\,(x-c)^n.$$

Proof Consider first the situation in which $x > c$. Let $U_n = \{f^{(n)}(u) : u{\in}(c,x)\}$, and set $m_n = \inf U$ and $M_n = \sup U_n$.

Case 1 Suppose that m_n and M_n are both finite. Then by Theorem (V.6.3) one has

$$\frac{m_n}{n!}(x-c)^c \leq f(x) - p_{(f;n-1;c)}(x) \leq \frac{M_n}{n!}(x-c)^n$$

By Part (a) of Theorem (V.6.3)

Side Comment (on the mean-value theorem for derivatives) The 'Mean-Value Theorem' described in Theorem (V.6.11) below has an interesting history. It was known to Lagrange in the late 18th century, and to Cauchy in the 1820's; indeed, Cauchy also presented a stronger version; see Theorem (V.6.14). Unfortunately, none of their proofs would be considered rigorous today. In *This Textbook* we follow the usual custom of naming the original version, Theorem (V.6.11), after Lagrange, and the stronger version, Theorem (V.6.14), after Cauchy.

The Lagrange/Cauchy formulation of the original theorem can be paraphrased as follows:

Suppose that $y = f(x)$ has continuous derivative $f'(x)$ for all x in some interval $a \leq x \leq b$. Let m be the minimum value of of f' on the interval $[a, b]$, and let M be the corresponding maximum value of f'. Then

$m(b{-}a) \leq f(b){-}f(a) \leq M(b{-}a)$; *equivalently, the ratio* $\dfrac{f(b) - f(a)}{b - a}$ *lies in the interval* $[m, M]$.

Stated this way, the theorem says that the amount the function f changes over the interval $[a, b]$, i.e., the quantity $f(b) - f(a)$, depends not just on the size of the interval, i.e., on $b - a$, but also on the size of the derivative, i.e. on m and M.

However, there is an alternate phrasing which is shorter, and which has become the standard one. To arrive at it, note that since (by hypothesis) f' is continuous on $[a, b]$, it follows that $[m, M]$ is the image of $[a, b]$ under f'; see Part (b) of Corollary (IV.4.2). Thus, the conclusion of the theorem, namely that the ratio $\dfrac{f(b) - f(a)}{b - a}$ lies in $[m, M]$, is equivalent to saying that this ratio is a value of f' on $[a, b]$; that is, there exists c in $[a, b]$ such that

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

This alternate formulation, which both Lagrange and Cauchy stated, has some pleasant features. For example, it involves a single equation, not a pair of inequalities. It also avoids the need to mentioning the extreme values m and M of f'.

Unfortunately, these 'pleasant features' are also the worst features of the alternate formulation. In particular, this formulation shifts the focus away from the size of f' (i.e., m and M), and towards the number c. Indeed, many students in elementary calculus think that this theorem is about determining the value of c; but of course, it is not, since one almost never needs to do that. Compare this situation with two other major theorems of elementary calculus which also involve the existence of a number c in $[a, b]$; namely, the Extreme-Value Theorem and Intermediate-Value Theorem. In those theorems, finding effective methods for computing c is a central issue. Thus, it is not surprising that students believe (albeit incorrectly) that computing c explicitly is what the Mean-Value Theorem asks of them.

Here is a summary of the proof along the line of Lagrange and Cauchy, with editorial references, to some relevant theorems in *This Textbook*, marked off by brackets:

Note first that the maximum and minimum values M and m of f' referred to in the statement of the theorem do exist, by the Extreme-Value Theorem, because of the hypothesis that f' is continuous on $[a, b]$. Then $f'(x) - m \geq 0$ for all x. That is, the function $g(x) = f(x) - m\,x$ satisfies the inequality $g'(x) \geq 0$, and thus $g(b) - g(a) \geq 0$ [see Theorem (V.4.8)]. That is, $f(b) - f(a) \geq m\,(b - a)$. A similar argument shows that $f(b) - f(a) \leq M\,(b - a)$.

In contrast, the approach to the Mean-Value Theorem most widely used now in elementary calculus texts is based on the proof attributed to Ossian Bonnet in the mid 1800s. (This is the familiar proof that starts with Rolle's Theorem, and then introduces a cleverly-chosen auxiliary function to get the general result.) Indeed, in calculus one carries out this proof first and then later uses the conclusion to prove Theorem (V.4.8) and Corollary (V.4.9).

The Bonnet proof is a model of mathematical elegance: it is simple, it avoids the continuity hypothesis on f', and it gets c to be an interior point. Its one weakness – and in the context of teaching elementary calculus, this is a major weakness – is that it hides the fact that the size of the derivative has a bearing on how much the function can change.

The proof of the Lagrange Mean-Value Theorem given below harks back to the Lagrange/Cauchy viewpoint, in that it directly relates the ratio $\frac{f(b)-f(a)}{b-a}$ to the size of the derivative. However, this proof also maintains the theoretical advantages of the Bonnet approach, in that it does not require continuity of f', and it does guarantee that the number c can be chosen to be an interior point of the interval.

V.6.11 Theorem (The Mean-Value Theorem for Derivatives)

Suppose that $f : [a,b] \to \mathbb{R}$ is differentiable at each point of the open interval (a,b) and continuous on the closed interval $[a,b]$.

(a) (Lagrange Formulation) Let S be the set of numbers of the form $f'(x)$ for x in the open interval (a,b). Suppose that m is a lower bound of the set S and that M is an upper bound of S. Then

$$m \leq \frac{f(b)-f(a)}{b-a} \leq M;\ \text{equivalently, } m \leq \frac{f(a)-f(b)}{a-b} \leq M. \qquad \text{(V.5)}$$

Furthermore, if f' is not constant on the interval (a,b), then both of the inequalities in (V.5) are strict; that is, in this case the fraction $(f(b)-f(a))/(b-a)$ equals neither m nor M.

(b) (Standard Formulation) There exists a number c, with $a < c < b$, such that

$$\frac{f(b)-f(a)}{b-a} = f'(c);\ \text{equivalently, } \frac{f(a)-f(b)}{a-b} = f'(c) \qquad \text{(V.6)}$$

Equivalently, there exists c in the open interval (a,b) such that

$$f(b)-f(a) = f'(c)\,(b-a);\ \text{equivalently, } f(a)-f(b) = f'(c)\,(a-b) \qquad \text{(V.7)}$$

Proof

(a) The inequality $m \leq \frac{f(b)-f(a)}{b-a}$ is obviously satisfied if $m = -\infty$, and is strict, since the fraction is finite. The analogous statement is true if $M = +\infty$.

Thus, it suffices to consider the case in which m and M are finite. Under this assumption, let $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ be the functions given by the formulas $g(x) = m\,x$ and $h(x) = M\,x$. Then $g'(x) = m$ and $h'(x) = M$for all x in $\mathbb{R}$, so

$$g'(x) \leq f'(x) \leq h'(x) \text{ for all } x \text{ in } (a,b).$$

Let $\hat{g}$, $\hat{f}$ and $\hat{h}$ be the unique antiderivatives of g', f' and h' on (a,b) which have value 0 at a; that is,

$$\hat{g}(x) = g(x) - g(a) = m, (x-a)\,\hat{f}(x) = f(x) - f(a),\ \text{and } \hat{h}(x) = h(x) - h(a) = M\,(x-c).$$

It then follows from Corollary (V.4.10) that if $x \in [a,b]$ then

$$m\,(x-a) \leq f(x) - f(a) \leq M\,(x-a) \text{ for all } x \text{ in } (a,b).$$

In particular, if one sets $x = b$ in the preceding set of inequalities and then divide by the positive number $b-a$, one gets the desired inequality.

(b) Part (a) is valid for every choice of lower bound m and upper bound M of the set S, so chose $m = \inf S$ and $M = \sup S$. In Cases (1) and (2) above, one can choose c to be any number in (a,b) since, as is shown above, $f'(x) = \dfrac{f(b)-f(a)}{b-a}$ for *all* x in (a,b). In Case (3), with this choice of m and M, the desired result follows from the Intermediate-Value Theorem for Derivatives (Theorem (V.4.5)). ■

Remarks (1) Many texts refer to what we call the 'Standard Version' here as the 'Lagrange Version'. In fact, Lagrange stated both versions, as did Cauchy some years later. The name 'Cauchy' is traditionally attached to certain generalizations; see below.

(2) In principle the most precise phrasing of Part (a) of the preceding theorem comes from choosing $m = \inf S$ and $M = \sup S$. However, computing infima and suprema can be difficult, whereas finding less precise values for m and M, which are still precise enough for the problem at hand, is often much simpler.

In the preceding theorem we assume that $a \neq b$. The next slight generalization, whose proof is omitted, drops this requirement. Of course the formulation no longer involves division by the number $b - a$, since this quantity might equal zero. This formulation is easier to generalize to multivariable calculus.

V.6.12 Corollary (The Mean-Value Theorem – Segment Form)

Suppose that f is differentiable at each point of a segment $\mathrm{Seg}\,[a,b]$ in $\mathbb{R}$; we do not assume that $a < b$. Then there exists a number c in $\mathrm{Seg}\,[a,b]$) such that

$$f(b) - f(a) = f'(c)\,(b-a) \text{ for some } c \text{ in } \mathrm{Seg}\,[a,b]. \qquad \text{(V.8)}$$

Remark If $b = a$, then $\mathrm{Seg}\,[a,b] = \{a\}$, so one must choose $c = a = b$. In this case, however, $b - a = 0$ and $f(b) - f(a) = 0$, so the actual value of the factor $f'(c)$ is irrelevant.

It is traditional to single out the following special case of the Standard Mean-Value Theorem and give it a name.

V.6.13 Theorem (Rolle's Theorem)

Suppose that $f : [a,b] \to \mathbb{R}$ is differentiable at each point of the open interval (a,b) and continuous on the closed interval $[a,b]$. Suppose, in addition, that $f(a) = f(b)$. Then there exists c, with $a < c < b$, such that $f'(c) = 0$.

Proof

The hypothesis $f(a) = f(b)$ implies that $\dfrac{f(b)-f(a)}{b-a} = 0$. Now apply Part (b) of the Standard Mean-Value Theorem.

Remark Michel Rolle stated the preceding result in 1691, at least for polynomial functions; apparently his name was not attached to it until 1835. The proof of his theorem given above is not the same as one finds in elementary calculus; for that proof, see the exercises.

Cauchy also provided the following major generalization of the Mean-Value Theorem.

V.6.14 Theorem (The Cauchy Mean-Value Theorem)

Suppose that f and g are continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Suppose, in addition, that for all x in (a, b) one has $g'(x) \neq 0$.

(a) There exists a number p in (a, b) such that

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(p)}{g'(p)}. \tag{V.9}$$

(b) Let $M = \sup \left\{ \frac{f'(x)}{g'(x)} : a < x < b \right\}$, and let $m = \inf \left\{ \frac{f'(x)}{g'(x)} : a < x < b \right\}$. Then

$$m \leq \frac{f(b) - f(a)}{g(b) - g(a)} \leq M$$

Furthermore, if the ratio f'/g' is not constant on the interval (a, b), then the preceding inequalities are strict.

Remark Cauchy's proof of this result is actually along the lines of the proof given above for the Lagrange Mean-Value Theorem. Indeed, we could have saved space by simply stating and proving the Cauchy version, and then pointing out that the standard version corresponds to the special case $g(x) = x$. However, separating the statements of these results is customary; and doing so allows us to give an alternate proof which shows that the Cauchy version, which is obviously more general than the Lagrange formulation of the Mean-Value Theorem, can actually be viewed as a special case of the Lagrange version.

Proof

(a) By Theorem (V.4.4), the hypothesis that g' is never 0 implies that g' is either always positive or always negative. Assume first that $g'(x) > 0$ for all x in (a, b).

By Theorem (IV.4.7) it follows that the image $g[a, b]$ of $[a, b]$ under g is also a closed interval $[c, d]$, with $c = g(a)$ and $d = g(b)$. Also the inverse $g^{-1} : [c, d] \rightarrow [a, b]$ is continuous on $[c, d]$, and satisfies the conditions $g^{-1}(c) = a$ and $g^{-1}(d) = b$. It also follows that g maps the open interval (c, d) onto the open interval (a, b).

Now consider the function $h : [c, d] \rightarrow \mathbb{R}$ defined by the rule

$$h(y) = f(g^{-1}(y)) \text{ for all } y \text{ in } [c, d];$$

that is, $h = f \circ g^{-1}$. It is clear from what was just stated that h is continuous on $[c, d]$. Likewise, it is clear from the Chain Rule that h is differentiable on the open interval (c, d); more precisely, if $x \in (a, b)$ and $y = g(x)$, so that $x = g^{-1}(y)$, then

$$h'(y) = f'(g^{-1}(y)) \cdot (g^{-1})'(y) = \frac{f'(x)}{g'(x)} \tag{V.10}$$

Apply Part (b) of the Lagrange Mean-Value Theorem to the function h to conclude that there exists a number q in (c, d) such that

$$\frac{h(d) - h(c)}{d - c} = h'(q) \quad (*)$$

Let $p = g^{-1}(q)$. In light of Equation (V.10) and the definition of h, Equation $(*)$ can be written

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(p)}{g'(p)},$$

as required.

If $g'(x) < 0$ for all x in (a, b), then apply what was just proved to the functions f and $-g$.

(b) It is clear from Equation $EqnE.110$ above that the quantities M and m defined above also satisfy the conditions $M = \sup \{h'(y) : c < y < d\}$ and $m = \inf \{h'(y) : c < y < d\}$. The desired result now follows from Part (a) of Theorem (V.6.11). ■

There is a slight generalization of the Cauchy Mean-Value Theorem which, aptly enough, many calculus texts call the 'Generalized Mean-Value Theorem'; we include it here for sake of completeness. Its main advantage is that it treats the functions f and g more equally; in particular, there is no hypothesis concerning the nonvanishing of either function or their derivatives.

V.6.15 Theorem (The Generalized Mean-Value Theorem)

Suppose that f and g are continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Then there exists a number p in (a, b) such that

$$(f(b) - f(a))g'(p) = f'(p)(g(b) - g(a)). \qquad \text{(V.11)}$$

The simple proof is left as an exercise. ■

V.6.16 Remarks

(1) In most elementary calculus texts the Lagrange Mean-Value Theorem is used to prove those results in Section (V.4) which relate the sign of the derivative f' to the changes in the values of the original function f.

(2) The Lagrange Mean-Value Theorem is sometimes called the 'Law of the Mean', to distinguish it from various 'mean-value theorems for definite integrals'. In contrast, French mathematics texts often use the name 'Theorem of Finite Changes' (or, more precisely, the French equivalent of that phrase), reflecting the fact that the theorem is concerned with how the derivative affects the changes in the values of the function.

(3) Many authors drop the explicit reference to 'Lagrange' and use the phrase 'Mean-Value Theorem' to refer to the Lagrange version, Theorem (V.6.11), studied above. We often follow that usage in *This Textbook.*

(4) What we call here the Cauchy Version of the Mean-Value Theorem is called the Generalized Mean-Value Theorem by some texts.

Side Comment (on calculus pedagogy, Part 2) (a) Most students find the name 'Mean-Value Theorem' to be obscure. To understand its roots, one must use some ideas from elementary calculus that have not yet been covered in *This Textbook.* However, since the purpose is to explain terminology, and not to prove anything, it seems safe enough to pause the purely logical flow of material for a moment.

Recall from elementary integral calculus: if a variable quantity u is a continuous function of x for $a \leq x \leq b$, then one defines the *average value* $\overline{u}$ *of* u *over the interval* $[a, b]$ by the equation

$$\overline{u} = \frac{1}{b-a} \int_a^b u \, dx \qquad (*)$$

Note that in subjects such as probability and statistics one frequently uses the phrase 'mean value' instead of 'average value'. In particular, suppose that $u = f'(x)$, where $y = f(x)$ is a function whose derivative f' is continuous on the closed interval $[a,b]$. Then combining the preceding definition of 'mean value' with the Fundamental Theorem of Calculus, one gets

$$\overline{u} = \frac{1}{b-a}\int_a^b f'(x)\,dx = \frac{f(b)-f(a)}{b-a}.$$

This equation says that the fraction $\dfrac{f(b)-f(a)}{b-a}$ which appears in the Lagrange Mean-Value Theorem is, in fact, a 'mean value': it equals the average (i.e., mean) value of the function f' over the interval $[a,b]$.

The Cauchy Mean-Value Theorem has a similar interpretation in terms of average values. It is based on the 'Change-of-Variables' formula from elementary integral calculus.

More precisely, let us introduce a new variable t by the equation $t = g(x)$, $a \leq x \leq b$. Assume, for simplicity, that g' is continuous and positive on $[a,b]$, so that t ranges over the interval $[c,d]$, where $c = g(a)$ and $d = g(b)$. Then g^{-1} is defined and continuously differentiable on $[c,d]$, with $a = g^{-1}(c)$ and $b = g^{-1}(d)$. One can then write $x = g^{-1}(t)$, so that the variable $y = f(x)$ can be expressed as a function of t:

$$y = f(g^{-1}(t)), \quad c \leq t \leq d.$$

Then by the Chain Rule one has

$$\int \frac{dy}{dx}\,dx = \int \frac{dy}{dx}\left(\frac{dx}{dt}\,dt\right) = \int \frac{dy}{dt}\,dt$$

Thus, the Change-of-Variables Formula says

$$f(b) - f(a) = \int_a^b \frac{dy}{dx}\,dx = \int_c^d \frac{dy}{dt}\,dt$$

Divide both sides by $g(b) - g(a) = d - c$ to get

$$\frac{f(b)-f(a)}{g(b)-g(a)} = \frac{1}{d-c}\int_c^d \frac{dy}{dt}\,dt.$$

Thus, the ratio $\dfrac{f(b)-f(a)}{g(b)-g(a)}$ which appears in the Cauchy Mean-Value Theorem is the average value over the interval $[c,d]$ of the quantity dy/dt, where $y = f(x)$ for x in $[a,b]$.

(b) The careful reader may have noticed the consistent use above of a hyphen in the phrase 'mean-value theorem'. The reason for this may become clearer if one substitutes the equivalent phrase 'average-value theorem'. Without the hyphen, one could interpret the phrase 'average value theorem' as referring to a 'value theorem' which is neither too complicated nor too easy, but just 'average' in difficulty; of course, that is *not* the intended meaning.

The next well-known result is one of the standard applications of the Cauchy Mean-Value Theorem.

V.6.17 Theorem (L'Hôpital's Rule for Functions)

(a) Suppose that f and g are functions defined on a half-open interval $(c,b]$; the case $c = -\infty$ is allowed. Assume that f and g satisfy the following hypotheses:

(i) $\lim_{x\to c+} f(x)$ and $\lim_{x\to c+} g(x)$ both exist, and one of the following cases holds:
('0/0 case') each of these limits equals 0;
('∞/∞ case') each limit equals one of the infinities $+\infty$ or $-\infty$.
(ii) $f'(x)$ and $g'(x)$ exist for all x in $(c,b]$;
(iii) $g'(x) \neq 0$ if $x\in(c,b]$;
(iv) $\displaystyle\lim_{x\to c+} \frac{f'(x)}{g'(x)} = L$. (The quantity L can be either a real number or one of the infinities.)

Then $\displaystyle\lim_{x\to c+} \frac{f(x)}{g(x)} = L$ as well.
Note: If $c = -\infty$ then the notation $\lim_{x\to c+}$ means the same as the more usual $\lim_{x\to -\infty}$.

(b) If, instead, f and g are defined on an interval of the form $[a,c)$, then the corresponding results for left-hand limits at c also hold. Likewise, the corresponding results for two-sided limits at c hold in the case for which f and g are defined on the intervals $[a,c)$ and $(c,b]$ with $a < c < b$.

Proof

(a) In light of Definition (IV.5.1), it suffices to prove the following:

Claim For each strictly monotonic sequence $\xi = (x_1, x_2, \ldots)$ in $(c,b]$ such that $\lim_{k\to\infty} x_k = c$, one has $\lim_{k\to\infty} f(x_k)/g(x_k) = L$.

To see that 'Claim' is true, define sequences $\alpha = (a_1, a_2, \ldots)$ and $\beta = (b_1, b_2, \ldots)$ by the rules $a_k = f(x_k)$ and $b_k = g(x_k)$ for each k in $\mathbb{N}$. Note that from Hypothesis (iii) above it is clear that the function g is strictly monotonic on $(c,b]$ and thus the sequence β is strictly monotonic. Furthermore, by the Cauchy Mean-Value Theorem, for each k in $\mathbb{N}$ one has

$$\frac{a_k - a_{k+1}}{b_k - b_{k+1}} = \frac{f(x_k) - f(x_{k+1})}{g(x_k) - g(x_{k+1})} = \frac{f'(z_k)}{g'(z_k)}$$

for some z_k between x_k and x_{k+1}. It follows, from the hypothesis that $\lim_{k\to\infty} x_k = c$ and the Squeeze Property for sequences, that $\displaystyle\lim_{k\to\infty} z_k = c$, and thus, by Hypothesis (iv), one has $\displaystyle\lim_{k\to\infty} \frac{a_k - a_{k+1}}{b_k - b_{k+1}} = L$. The desired result now follows by applying Theorem (III.3.8), the Stoltz-Cesaro theorem.

(b) The proof here is similar, and is left as an exercise. ■

V.6.18 Remarks

(1) The spelling 'L'Hôpital' used here is common, but not universal. Many texts write 'L'Hospital' instead; some of these also drop the apostrophe. No matter how one decides to spell it, however, the pronunciation is the same: *Loh pea tahl.*
In other words, the 's' in the second spelling is silent. (Of course the 'H' and the apostrophe are also silent; it is French, after all.) The advantage of using the 'L'Hôpital' spelling is that the letter 's' is not visible, so there is no temptation for a student to pronounce it 'Luh Hahs Pee Tal', as many do.

(2) It appears that M. L'Hôpital is honored with the name of this result because it appeared in his calculus text (the first such text); but the result itself is due to Johann Bernoulli.

V.6.19 Corollary

Suppose that $f : I \to \mathbb{R}$ is differentiable at every point of an open interval I. Then a necessary and sufficient condition for f' to be continuous at a point c of I is that $\lim\limits_{x \to c} f'(x)$ exist. (It is *not* assumed here that this limit is finite.)

The simple proof is left as an exercise. ■

V.7 The First Derivative and Linear Approximations

The description of 'derivative' given in Definition (V.1.1) is standard. The next result provides an alternative formulation.

V.7.1 Theorem

Let $f : I \to \mathbb{R}$ be a real-valued function whose domain is an open interval I in $\mathbb{R}$, and let c be a point of I. Then the following statements are equivalent:

(i) The function f is differentiable at c. That is, the quantity $\lim\limits_{x \to c} \dfrac{f(x) - f(c)}{x - c}$ exists and is finite.

(ii) There exists a number A such that for every $\varepsilon > 0$ there exists a number $\delta > 0$ so that if $|x - c| < \delta$ then

$$|f(x) - f(c) - A(x - c)| \leq \varepsilon |x - c| \quad (*)$$

Proof Suppose that Statement (i) is true, and let $A = f'(c)$. Then one has $A = \lim_{x \to c} \dfrac{f(x) - f(c)}{x - c}$, which means (by definition of 'limit') that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $0 < |x - c| < \delta$ then

$$\left| \frac{f(x) - f(c)}{x - c} - A \right| < \varepsilon.$$

Multiply both sides of this inequality by the positive quantity $|x - c|$, and use the fact that if u and v are real numbers, then $|u \cdot v| = |u| \cdot |v|$, to get

$$|f(x) - f(c) - A(x - c)| < \varepsilon |x - c| \quad (**)$$

if $0 < |x - c| < \delta$. Inequality $(**)$ is strict, and it clearly fails to hold if one allows $x = c$. However, $(**)$ certainly implies the weaker inequality $(*)$ when $0 < |x - c| < \delta$; and Inequality $(*)$ is trivially true when $x = c$, since in that case it reduces to the statement $0 \leq 0$. In other words, Statement (i) implies Statement (ii).

Conversely, suppose that Statement (ii) is true. Let $\varepsilon > 0$ be given, and let $\delta > 0$ be such that

$$|f(x) - f(c) - A(x - c)| \leq \frac{\varepsilon}{2} |x - c| \quad (***)$$

for all x such that $0 \leq |x - c| < \delta$. Then Inequality $(***)$ remains true if one restricts x even further to satisfy $0 < |x - c| < \delta$. In that situation, however, one has $0 < \dfrac{\varepsilon}{2} |x - c| < \varepsilon |x - c|$, and thus

$$|f(x) - f(c) - A(x - c)| < \varepsilon |x - c| \text{ if } 0 < |x - c| < \delta.$$

Divide both sides of this last (strict) inequality by the positive quantity $|x-c|$, and carry out the obvious algebraic simplification, to get

$$\left|\frac{f(x)-f(c)}{x-c} - A\right| < \varepsilon \text{ if } 0 < |x-c| < \delta.$$

It follows from this argument that

$$\lim_{x\to c} \frac{f(x)-f(c)}{x-c} = A.$$

In particular, f is differentiable at c, i.e., Statement (i) is true. Indeed, this argument shows that $f'(c) = A$. ■

V.7.2 Remarks

(1) The preceding result can be interpreted in terms of the accuracy near c of 'linear approximations' of the function f. More precisely, suppose that Statement (ii) of the preceding theorem is satisfied with a certain choice of A (which of course we know equals $f'(c)$). Let $g : \mathbb{R}\to\mathbb{R}$ be another linear function such that $g(c) = f(c)$; thus $g(x) = f(c) + m(x-c)$ for some constant $m \neq A$. We wish to compare the accuracy of the two approximations,

$$f(x) \approx f(c) + A(x-c) \text{ and } f(x) \approx f(c) + m(x-c).$$

Note that, by Inequality (II.4), the 'Modified Triangle Inequality', one has

$$|f(x)-f(c)-m(x-c)| = |(A-m)(x-c) + (f(x)-f(c)-A(x-c))| \geq$$

$$||A-m||x-c| - |f(x)-f(c)-A(x-c)|| \geq |A-m||x-c| - |f(x)-f(c)-A(x-c)|.$$

Let $\varepsilon > 0$ satisfy $0 < \varepsilon < |A-m|/2$, and let $\delta > 0$ be small enough that

$$|f(x)-f(c)-A(x-c)| \leq \varepsilon|x-c| \text{ if } 0 < |x-c| < \delta.$$

For such x one then has

$$|f(x)-f(c)-m(x-c)| \geq |A-m||x-c| - |f(x)-f(c)-A(x-c)| \geq$$

$$|A-m||x-c| - \frac{1}{2}|A-m||x-c| = \frac{1}{2}|A-m||x-c| > 0.$$

This implies that

$$\left|\frac{f(x)-f(c)-A(x-c)}{f(x)-f(c)-m(x-c)}\right| \leq 2\frac{|f(x)-f(c)-A(x-c)|}{|A-m||x-c|} \leq \frac{2\varepsilon}{|A-m|}$$

Since $A = f'(c)$, one can interpret this result as follows: as x approaches c, the error in approximating $f(x)$ by the function $f(c) + f'(c)(x-c)$ becomes 'infinitely small' in comparison with the error in approximating $f(x)$ by $f(c) + m(x-c)$ when $m \neq f'(c)$. This fact is reflected by the following standard terminology.

(2) In elementary calculus, the straight line given by the equation $y = f(c) + f'(c)(x-c)$ is called the **tangent line** to the graph $y = f(x)$ at the point $(c, f(c))$. The preceding analysis then shows that near the point $(c, f(c))$ this tangent line approximates the graph of f infinitely better than any other straight line that passes through that point.

The preceding results give rise to the following standard terminology.

V.7.3 Definition

Suppose that $f : I \to \mathbb{R}$ is a function defined on an open interval I in $\mathbb{R}$, and that c is a point of I. Suppose that f is differentiable at c, and define the linear function $L : \mathbb{R} \to \mathbb{R}$ by the rule $L(x) = f(c)+f'(c)(x-c)$ for all x in $\mathbb{R}$. Then the approximation $f(x) \approx L(x)$ is called the **tangent-line approximation of f at c**. The function L itself is called the **best linear approximation of f at c**.

V.7.4 Remarks

(1) The content of Theorem (V.7.1) can be summarized to say that one can define the derivative of a real-valued function of a single real variable either in terms of limits (Statement (i)) or in terms of tangent lines (Statement (ii)). The 'limit' viewpoint is usually preferred in single-variable calculus because the proofs of basic facts, such as the Product Rule and the Chain Rule, tend to be more straight forward using the 'limit' approach.

The situation is more complicated in multivariable calculus. Indeed, in that context the two viewpoints correspond to different concepts: the obvious generalization of the 'limit' approach leads to the concept of 'Partial Derivatives'; while the obvious generalization of the 'tangent line' approach leads to a stronger concept, the 'Total Derivative', sometimes called the 'Frechet derivative'. It turns out, for example, that for functions of two or more variables the analog of Theorem (V.1.5), the 'Differentiability-Implies-Continuity' Theorem, remains valid for 'total derivatives', but does not hold for partial derivatives.

(2) In the preceding we follow the standard usage in high-school algebra and refer to a function of the form $L(x) = ax + b$, with a and b constant, as a 'linear function of x'. This usage is quite standard in single-variable analysis, and needs no apologies. However, in the subject of Linear Algebra the word 'linear' has a somewhat more restricted meaning; for example, for a function $f : \mathbb{R} \to \mathbb{R}$ to be called 'linear' in Linear Algebra, it would be required (among other things) that $f(0) = 0$. Thus, some careful authors use the word 'affine' to describe a function of the form $ax + b$, and restrict 'linear' to the case $b = 0$. Such an author would refer to instead of linear approximations. Since this issue is of little consequence in single-variable analysis, we do not use the 'affine' terminology any further in *This Textbook*.

The tangent-line approximation is among the most elementary of the many schemes used in applied mathematics to simplify problems. The basic principle of such approximations is easy to grasp: Suppose that you need to perform a certain type of operation on a complicated function f. Replace f by a simpler function g which provides a good approximation of f, and for which the desired operation is easy to carry out; then carry out the operation on g. The expectation is that the result of doing the operation on g should provide a decent approximation of the result of actually doing the operation on f.

One of the most familiar of these 'approximation schemes' is for solving equations of the form $f(x) = 0$, where f is a real-valued function which is differentiable at each point of an open interval I. (This method is also referred to as the **Newton-Raphson method**.) The idea is simple: choose a point x_0 in I, preferably one which is close to a root r of the equation $f(x) = 0$. Let $g_0 : \mathbb{R} \to \mathbb{R}$ be the corresponding tangent-line approximation of f at x_0, so that $g_0(x) = f(x_0)+f'(x_0)(x-x_0)$. If x_0 really is close to r, then one could reasonably expect the line $y = f(x_0) + f'(x_0)(x - x_0)$ to remain close to the graph $y = f(x)$, at least throughout the segment $\mathrm{Seg}\,[x_0, r]$. And if that is the case, then it is reasonable to presume that solving the equation $f(x) = 0$ is almost the same as solving the *linear* equation $g_0(x) = 0$. The latter equation is easy to solve:

$$f(x_0) + f'(x_0)(x - x_0) = 0 \text{ implies } x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

(Of course in this expression one tacitly assumes that $f'(x_0) \neq 0$.) Moreover, if the *equation* $g_0(x) = 0$ is almost the same as the equation $f(x) = 0$, then it seems plausible that the *solution* $x_1 = x_0 - f(x_0)/f'(x_0)$ of the approximate equation $g(x) = 0$ is even closer to the desired root r than x_0 is. And if that is the case, then applying the same idea to the tangent-line approximation of f at x_1 should produce an even better approximation of the desired root r. The complete procedure thus produces the familiar infinite sequence x_0, $x_1, \ldots x_k, \ldots$ given recursively by the rule

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \text{ for } k = 1, 2, \ldots.$$

It is assumed here that $f'(x_k) \neq 0$ and that x_k being in the domain of f implies that x_{k+1} is also. Note that in the special case $f(x) = x^2 - C$, where C is a positive constant, Newton's Method with $x_0 > 0$ is the same as Heron's Method for computing square roots.

Another important linear approximation scheme is given by linear interpolation (see Example (I.5.6)). For this method, one needs *two* points on the graph of the given function f. More precisely, suppose that $(a, f(a))$ and $(b, f(b))$ are points such that $a \neq b$. Then the linear function $g : \mathbb{R} \to \mathbb{R}$ which interpolates these points is given by the formula

$$g(x) = f(a) + \left(\frac{f(b) - f(a)}{b - a}\right)(x - a) \text{ for all } x \text{ in } \mathbb{R}.$$

Note that the graph $y = g(x)$ of the function g is what one calls in elementary calculus the 'secant line' to the graph $y = f(x)$ through the points $(a, f(a))$ and $(b, f(b))$.

As with the tangent line approximation, one can use the 'secant line approximation' $y = g(x)$ to approximate solutions of the equation $f(x) = 0$. More precisely, suppose that $(x_0, f(x_0))$ and $(x_1, f(x_1))$ are points of the graph of f with $x_0 \neq x_1$, and let g_0 be the corresponding linear interpolation function. Now consider the equation $g_0(x) = 0$; one gets the solution

$$x = x_0 - f(x_0)\left(\frac{x_1 - x_0}{f(x_1) - f(x_0)}\right).$$

As with Newton's Method, this process can be repeated to produce of an infinite sequence of approximations of the solution of $f(x) = 0$. More precisely, if x_0, $x_1, \ldots$ x_k have been defined, then set

$$x_{k+1} = x_{k-1} - f(x_{k-1})\left(\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}\right) \text{ for each } k = 1, 2, \ldots$$

In numerical analysis this procedure is called the .

In numerical analysis it is customary to estimate the errors in the tangent line and secant line approximations in terms of the second derivative of the original function f. The following estimates, which involve only the first derivative of f, are cruder, but are sufficiently good for our current purposes. To simplify the discussion, we assume that the function f to be approximated has continuous first derivative.

V.7.5 Theorem

Suppose that $f : I \to \mathbb{R}$ is a C^1 function on an open interval I in $\mathbb{R}$. Let a and b be numbers in I such that $a < b$, and let M be an upper bound for $|f'|$ on the closed bounded interval $[a,b]$. (The existence of the finite upper bound M follows from the Extreme-Value Theorem applied to the continuous function f' on $[a,b]$.)

(a) Suppose that $c \in [a,b]$, and let $g : \mathbb{R} \to \mathbb{R}$ be the tangent line approximation of f at c; that is, $g(x) = f(c) + f'(c)(x-c)$ for all x. Then

$$|f(x) - g(x)| \leq 2M|x-c| \leq 2M|b-a| \text{ for all } x \text{ in } [a,b]. \tag{V.12}$$

(b) Let $g : \mathbb{R} \to \mathbb{R}$ be the linear interpolation of f on the interval $[a,b]$; that is,

$$g(x) = f(a) + \left(\frac{f(b)-f(a)}{b-a}\right)(x-a).$$

Then

$$|f(x) - g(x)| \leq M|b-a| \text{ for all } x \text{ in } [a,b]. \tag{V.13}$$

The key to the proof is the following result:

V.7.6 Lemma

Suppose that f, I, a, b and M are as in the statement of the preceding theorem. Let $g : \mathbb{R} \to \mathbb{R}$ be a linear function, so that $g'(x) = m$ for some constant m such that $|m| \leq M$. Assume further that there is a number c in $[a,b]$ such that $f(c) = g(c)$; speaking geometrically, assume that the graph of f and the graph of g intersect somewhere in $[a,b]$. Then for all x in $[a,b]$ one has

$$|f(x) - g(x)| \leq 2M \max\{|b-c|, |c-a|\}.$$

Proof of Lemma Note that if $x \in [a,b]$ then, since $f(c) = g(c)$, there exists u in Seg $[x,c]$ such that

$$|f(x) - g(x)| = |(f(x)-f(c)) - (f(c)-g(x))| = |(f(x)-f(c)) - (g(c)-g(x))| =$$

$$|f'(u)(x-c) - m(x-c)| = |f'(u) - m| \cdot |x-c| \leq 2M|x-c|.$$

Since $c \in [a,b]$, one must have either $a \leq x \leq c$ or $c \leq x \leq b$. In the former case one gets $|x-c| \leq |c-a|$, while in the lattr one gets $|x-c| \leq |b-c|$. Thus in both cases one has $|x-c| \leq \max\{|b-c|, |c-a|\}$, and the lemma follows.

Proof of Theorem

(a) Since $m = g'(a) = f'(a)$ by construction of the tangdent line at $(a, f(a))$, it follows that m is a value of f' on $[a,b]$ and thus $|m| \leq M$. In addition, one has $g(a) = f(a)$. The desired result now follows from the lemma with $c = a$.

(b) Let m be the (constant) value of the interpolating linear function g. Then by the definition of 'interpolatoion', combined with the standard Mean-Value Theorem one has

$$m = \frac{g(b)-g(a)}{b-a} = \frac{f(b)-f(a)}{b-a} = f'(\xi) \text{ for some } \xi \text{ in } (a,b).$$

In partcular, m is a value of f' and thus $|m| \leq M$. Let $\mu = (a+b)/2$ be the midpoint of the interval $[a,b]$, and note that M is an upper bound for $|f'|$ on each of the subintervals $[a,\mu]$ and $[\mu,b]$. Suppose that $x \in [a,b]$. Then x is in (at least one) of the subintervals $[a,\mu]$ or $[\mu,b]$. In the former case the lemma, with $c = a$, implies that $|f(x) - g(x)| \leq 2M|\mu - a| = M|b-a|$. Likewise, in the latter case, the same lemma, but now with $c = b$, implies $|f(x) - g(x)| \leq M|b-a|$. The desired result now follows. ■

Remark A generalization of Part (b) is given in the exercises.

V.7.7 Remark

The careful reader may be grumbling that the main hypothesis in the preceding theorem, namely that f be of class C^1 on I, is stronger than necessary. Indeed, this hypothesis could be replaced by the following weaker one:

'*The derivative f' is defined on the interval I, and for each a and b in I with $a < b$, f' is bounded on the subinterval* $[a,b]$.'

In light of Example (**??**), this alternate hypothesis is certainly weaker than the C^1 hypothesis actually used, and the proof using the weaker hypothesis is essentially the same as the one given above. So why not use the weaker hypothesis?

Answer The issue of balancing generality with ease-of-use arises frequently in mathematics (as in other subjects). In the current situation, expressing the result in terms of f being C^1 makes the statement marginally easier to remember, while being sufficiently general for our purposes.

V.8 Cauchy's Theorem on the Existence of Antiderivatives

The next theorem is one of the most important in analysis.

V.8.1 Theorem (Cauchy's Antiderivative Theorem)

Let $f : I \to \mathbb{R}$ be a function which is continuous on an interval I in $\mathbb{R}$. Then the function f has an antiderivative on I. More precisely, if c is any point of I, then f has a unique antiderivative F on I whose value at c is 0. (As usual, if the interval I includes an endpoint p, then $F'(p)$ refers to the appropriate one-sided derivative.)

Remark The proof given here differs quite a bit from that of Cauchy. It is due to H. Lebesgue; see [LEBESGUE 1905]. Cauchy's original proof is outlined in Chapter (VII).

Side Comment (on Lebesgues' proof of Cauchy's Antiderivative Theorem) The purpose of this Side Comment is outline the basic structure of the detailed proof given below. As usual, one can procede directly to the proof without reading this Side Comment.

The concept of 'continuity' is not obviously related to that of 'antiderivative', so the truth of Cauchy's Antiderivative Theorem comes somewhat as a surprise. Lebesgue's proof makes this relation believable. Indeed, consider any pair of numbers in I such that $a < c < b$, so that, by Theorem (IV.4.9), there exists a sequence of continuous piecewise linear functions $g_k : \mathbb{R} \to \mathbb{R}$ such that $|f(x) - g_k(x)| < 1/k$ for each $x \in [a,b]$; in particular, $\lim_{k\to\infty} g_k(x) = f(x)$ for each

$x \in [a,b]$. By Theorem (V.5.9), each g_k has a unique antiderivative G_k over $\mathbb{R}$ with value 0 at c, so that $G'_k(x) = g_k(x)$ for each x in $[a,b]$. It is easy to show that for each x in $[a,b]$ the sequence $(G_1(x), \ldots G_k, \ldots)$ is Cauchy, and thus is convergent. Let $F : [a,b] \to \mathbb{R}$ be defined by the rule $F(x) = \lim_{k\to\infty} G_k(x)$. Lebesgue shows that F is the desired antiderivative of f on $[a,b]$.

In summary: f has an antiderivative over $[a,b]$ because f can be 'nicely' approximated over $[a,b]$ by functions which themselves have antiderivatives over $[a,b]$. The details of the proof below use the precise meaning of 'nicely'. The final step, extending this solution to the full interval I, is then straight-forward.

Proof Let a and b be points of I such that $a \leq c \leq b$. Since f is continuous on I, one sees that for each positive integer k there is a continuous function $g_k : \mathbb{R} \to \mathbb{R}$ such that

Condition (i) $|g_k(x) - f(x)| < 1/k$ for all x in the interval $[a,b]$;

Condition (ii) g_k has a unique antiderivative G_k on $\mathbb{R}$ such that $G_k(c) = 0$.

Indeed, Theorem (IV.4.9) implies that there is a continuous piecewise-linear function on $\mathbb{R}$ which satisfies Condition (i), while Theorem (V.5.9) implies that every such function satisfies Condition (ii). Note also that Condition (i) implies that for each x in $[a,b]$ one has $\lim\limits_{k\to\infty} g_k(x) = f(x)$.

Claim 1 For each x in $[a,b]$ the sequence $\Gamma_x = \{G_1(x), G_2(x), \ldots\}$ is a Cauchy sequence.

Proof of Claim 1 Let k and m be positive integers. Note that if $x \in [a,b]$ then, by Corollary (V.6.12), the Segment Form of the Mean-Value Theorem applied to the function $G_{k+m} - G_k$, there exists a number $\hat{x}$ in $\mathrm{Seg}\,[x,c]$ such that the following holds:

$$G_{k+m}(x) - G_k(x) = (G_{k+m} - G_k)(x) - (G_{k+m} - G_k)(c) = (G_{k+m} - G_k)'(\hat{x})(x-c) =$$

$$(G'_{k+m}(\hat{x}) - G'_k(\hat{x}))(x-c) = (g_{k+m}(\hat{x}) - g_k(\hat{x}))(x-c).$$

It follows, after doing a judicious 'add-and-substract' trick and using the Triangle Inequality, that

$$|G_{k+m}(x) - G_k(x)| \leq (|g_{k+m}(x) - f(x)| + |f(x) - g_k(x)|)\,|x-c| \leq \left(\frac{1}{k+m} + \frac{1}{k}\right)|b-a| \leq \frac{|b-a|}{2k}.$$

In particular, if $\varepsilon > 0$ is given, then one has

$$|G_{k+m}(x) - G_k(x)| \leq \frac{|b-a|}{2k} < \varepsilon \qquad \text{(V.14)}$$

for all k in $\mathbb{N}$ such that $k > |b-a|/(2\varepsilon)$ and all m in $\mathbb{N}$. In particular, the sequence $\Gamma(x)$ is Cauchy, as claimed.

The preceding result suggests a natural candidate for the desired antiderivative of the given function f; namely, define $F : [a,b] \to \mathbb{R}$ by the rule $F(x) = \lim_{k\to\infty} G_k(x)$ for all x in the interval $[a,b]$. (The fact that for each x in $[a,b]$ this limit exists and is finite follows from Claim 1.)

Claim 2 The function F just described is the desired antiderivative of f. More precisely, for each x_0 in the open interval (a,b) and for each $\varepsilon > 0$, there exists $\delta > 0$ such that if h satisfies $0 < |h| < \delta$ then

$$\left|\frac{F(x_0+h) - F(x_0)}{h} - f(x_0)\right| \leq \varepsilon \quad (*)$$

Proof of Claim 2 Let x_0 be any point of (a,b), and let h be any nonzero real number small enough that $x_0 + h$ is also in (a,b); more precisely, $|h| < \min\{|x_0 - a|, |b - x_0|\}$. Then for each k in $\mathbb{N}$ one has, for some $\hat{x}$ in $\mathrm{Seg}\,[x_0, x_0+h]$,

$$G_k(x_0+h) - G_k(x_0) - g_k(x_0)h = G'_k(\hat{x})h - g_k(x_0)h = (g_k(\hat{x}) - g_k(x_0))h.$$

It then follows from the Extended Triangle Inequality that

$$|G_k(x_0+h) - G_k(x_0) - g_k(x_0)h| \leq (|g_k(\hat{x}) - f(\hat{x})| + |f(\hat{x}) - f(x_0)| + |f(x_0) - g_k(x_0)|) \cdot |h| \qquad (**)$$

Now let $\varepsilon > 0$ be given. Using the uniform continuity of f on the interval $[a,b]$, choose $\delta > 0$ small enough that if x and y are any points of $[a,b]$ such that $|y-x| < \delta$, then one has $|f(y)-f(x)| < \varepsilon/3$. It then follows from Condition (ii) above that if $0 < |h| < \delta$ then $(**)$ implies

$$|G_k(x_0+h) - G_k(x_0) - g_k(x_0)h| < \left(\frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3}\right)|h| = \varepsilon|h|$$

Divide by $|h|$ and do some simple algebraic simplification to get

$$\left|\frac{G_k(x_0+h) - G_k(x_0)}{h} - g_k(x_0)\right| < \varepsilon$$

This last inequality is true for *all* sufficiently large k. In particular, by letting k approach ∞ one gets the desired inequality $(*)$ for all h such that $0 < |h| < \delta$.

Since this is true for all $\varepsilon > 0$, it follows that $F'(x_0)$ exists and equals $f(x_0)$. Since x_0 can be any point of (a,b), it follows that F is an antiderivative of f on (a,b). The fact that $F(c) = 0$ follows from the fact that, by definition, $G_k(c) = 0$ for each k.

Summary For each a and b in I such that $a < c < b$, there exists a unique antiderivative F of f on (a,b) such that $F(c) = 0$.

The desired result, namely that f has a unique antiderivative, whose value at c is 0, on the *full* open interval I, now follows by letting a approach $\inf I$ and b approach $\sup I$; the details are left to the reader. ■

V.9 Numerical Methods for Antiderivatives

If $f : I \to \mathbb{R}$ is a continuous function on the open interval I, then the existence of an antiderivative $F : I \to \mathbb{R}$ of f on I is guaranteed by the preceding results. The numbers of greatest interest for such an antiderivative are differences of the form $F(b) - F(a)$, with a and b in I and $a < b$. In elementary calculus one normally computes such differences using a method that is independent of the specific choice of a and b: from the formula for $f(x)$ on I, obtain an explicit formula for $F(x)$; then substitute $x = a$ and $x = b$ into the latter formula, and subtract.

Unfortunately, in many cases determining an explicit formula for F from the formula for f can be difficult or even impossible. In such cases one may need to use methods which depend on the specific choice of a and b and which provide only estimates of $F(b) - F(a)$. Most of these methods fall under the heading of so-called 'numerical methods'; there is a vast literature of such methods. We consider several of the most familar here.

Perhaps the most straight-forward way to estimate the difference $F(b) - F(a)$, where $F' = f$, is to use the Mean-Value Inequality, Corollary (V.6.4). In the present context this Inequality takes the more precise form

$$m\,(b-a) \leq F(b) - F(a) \leq M\,(b-a),$$

where m is the minimum value of f on $[a,b]$ and M is the corrsponding maximum value; these extreme values exist because f is continuous on $[a,b]$. Of course the quantities m and M need not be close to each other, so this estimate may be of little value. To improve it, divide the interval $[a, b]$ into short subtintervals and apply the analogous inequalities on each of them. More precisely, let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be a partition of $[a, b]$ into n subintervals $[x_{j-1}, x_j]$, $1 \leq j \leq n$. For each such index j let m_j be the minimum value of f on the subinterval $[x_{j-1}, x_j]$, and let M_j be the corresponding maximum. Then from the equation

$$F(b) - F(a) = (F(x_n) - F(x_{n-1})) + \ldots + (F(x_1) - F(0)), \text{ that is, } F(b) - F(a) = \sum_{j=1}^{n} \Delta F_j(x),$$

where $\Delta F_j(x) = F(x_j) - F(x_{j-1})$, one obtains the estimate

$$m_1\,\Delta x_1 + m_2\,\Delta x_2 + \ldots + m_n\,\Delta x_n \leq F(b) - F(a) \leq M_1\,\Delta x_1 + M_2\,\Delta x_2 + \ldots + M_n\,\Delta x_n;$$

as usual, $\Delta x_j = x_j - x_{j-1}$. For brevity, denote the quantities $m_1\,\Delta x_1 + m_2\,\Delta x_2 + \ldots + m_n\,\Delta x_n$ and $M_1\,\Delta x_1 + M_2\,\Delta x_2 + \ldots + M_n\,\Delta x_n$ by $L(f; \mathcal{P})$ and $U(f; \mathcal{P})$, respectively. In particular, the numbers $L(f; \mathcal{P})$ and $U(f; \mathcal{P})$ determine a full range of reasonable estimates $F(b) - F(a) \approx A$: namely, let A be any number such that $L(f; \mathcal{P}) \leq A \leq U(f; \mathcal{P})$. Here are a few of the commonly used ways to choose the estimating number A:

(1) Let $\zeta = (z_1, z_2, \ldots z_n)$ be an ordered list of 'sample points' such that for each index j one has $z_j \in [x_{j-1}, x_j]$. It is clear that $m \leq f(z_j) \leq M_j$ for each j, so that $F(b) - F(a) \approx A_{\mathcal{P};\zeta}$, where $A_{\mathcal{P};\zeta} = f(z_1)\,\Delta x_1 + \ldots + f(z_n)\,\Delta\, x_n$. Here are several popular 'Rules' for choosing the sample points z_j:

(Left-hand Rule) $z_j = x_{j-1}$; that is,

$$A_{\mathcal{P};\lambda} = \sum_{j=1}^{n} f(x_{j-1})\,\Delta x_j.$$

(Right-hand Rule) $z_j = x_j$; that is,

$$A_{\mathcal{P};\rho} = \sum_{j=1}^{n} f(x_j)\,\Delta x_j.$$

(Midpoint Rule) $z_j = (x_{j-1} + x_j)/2$; that is,

$$A_{\mathcal{P};\mu} = \sum_{j=1}^{n} f\left(\frac{x_{j-1} + x_j}{2}\right)\,\Delta x_j.$$

(Minimum-Value Rule) z_j is a point in $[x_{j-1}, x_j]$ at which $f(z_j) = m_j$; that is,

$$A_{\mathcal{P};\min} = L(f; \mathcal{P})$$

(Maximum-Value Rule) z_j is a point in $[x_{j-1}, x_j]$ at which $f(z_j) = M_j$; that is,

$$A_{\mathcal{P};\max} = U(f; \mathcal{P})$$

(Mean-Value-Theorem Rule) z_j is a point in the interval $[x_{j-1}, x_j]$ at which $f(z_j) = \left(\dfrac{F(x_j) - F(x_{j-1})}{x_j - x_{j-1}}\right)$; that is,

$$A_{\mathcal{P};\mathrm{MVT}} = \sum_{j=1}^{n} f(z_j)\,\Delta x_j = \sum_{j=1}^{n} (F(x_j) - F(x_{j-1})) = F(b) - F(a).$$

Remark The final rule provides an approximation, $A_{\mathcal{P};\mathrm{MVT}} \approx F(b) - F(a)$, which in fact is exactly correct. Unfortunately, it is almost always impossible to determine the corresponding values of the points z_j without already knowing the exact value of $F(b) - F(a)$ in advance, making this rule of little practical value. Similarly, the difficulty of finding extrema for the function f on the n subintervals makes the Minimum/Maximum-Value Rules of little value.

(2) Since each of the estimating values $A_{\mathcal{P};\lambda}$, $A_{\mathcal{P};\rho}$ and $A_{\mathcal{P};\mu}$ satisfies the condition $L(f;\mathcal{P}) \leq A \leq U(f;\mathcal{P})$, it follows that any linear combination $p\,A_{\mathcal{P};\lambda} + q\,A_{\mathcal{P};\rho} + r\,A_{\mathcal{P};\mu}$, with $0 \leq p, q, r \leq 1$ and $p+q+r = 1$, also satisfies this condition. The two most familiar examples are these:

(Trapezoid Rule: $p = q = 1/2, r = 0$) $A_{\mathcal{P};\tau} = \dfrac{A_\lambda + A_\rho}{2}$.

(Simpson's Rule: $p = q = 1/6, r = 1/3$) $A_{\mathcal{P};\sigma} = \dfrac{A_{\mathcal{P};\tau} + 2\,A_{\mathcal{P};\mu}}{3}$.

Remarks (1) For simplicity, many calculus texts formulate these rules only in the special case the partition $\mathcal{P}$ has constant spacing; that is, all n of the subintervals $[x_{j-1}, x_j]$ are of equal length; equivalently, $\Delta x_j = (b-a)/n$ for each j.

(2) Lebesgue's proof of Cauchy's Antiderivative Theorem, given above, involves repeated use of the Trapezoid Rule.

The next result shows that these estimates can be guaranteed to be as accurate as one wishes by choosing the partition $\mathcal{P}$ appropriately. The key to this result is the observation that from the inequalities

$$L(f;\mathcal{P}) \leq F(b) - F(a) \leq U(f;\mathcal{P}) \text{ and } L(f;\mathcal{P}) \leq A_{\mathcal{P};\zeta} \leq U(f;\mathcal{P})$$

obtained above, it follows that

$$0 \leq |(F(b) - F(a)) - A_{\mathcal{P};\zeta}| \leq U(f;\mathcal{P}) - L(f;\mathcal{P}) \qquad (*)$$

V.9.1 Theorem

Suppose that $f : I \to \mathbb{R}$ is continuous on an open interval I, and that $F : I \to \mathbb{R}$ is an antiderivative of f on I. Let a and b be points of I such that $a < b$. Then:

(a) For every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ is any partition of $[a,b]$ whose mesh $||\mathcal{P}||$ satisfies $||\mathcal{P}|| < \delta$, then for every list $\zeta = (z_1, z_2, \ldots z_n)$ of sample points associated with $\mathcal{P}$ one has

$$|(F(b) - F(a)) - A_{\mathcal{P};\zeta}| < \varepsilon.$$

(As usual, $||\mathcal{P}||$ is defined to be the the largest of the numbers Δx_j, $1 \leq j \leq n$.)

(b) The quantity $F(b) - F(a)$ is the unique number B such that $L(f;\mathcal{P}) \leq B \leq U(f;\mathcal{P})$ for every partition $\mathcal{P}$ of the interval $\subseteq a, b]$.

Proof (a) By Theorem (IV.3.7), the function f is uniformly continuous on the closed bounded interval $[a,b]$. Thus, given $\varepsilon > 0$ let $\delta > 0$ be small enough that if c and d are any points in $[a,b]$ such that $|d-c| < \delta$, then $|f(d) - f(c)| < \varepsilon/(b-a)$. Let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be a partition as above such that $||\mathcal{P}|| < \delta$. Then for each index $j = 1, 2, \ldots n$ one has $|d-c| \leq (x_j - x_{j-1}) < \delta$ for each c and d in the subinterval $[x_{j-1}, x_j]$. In particular, choose c so that $f(c) = m_j$ and choose d so that $f(d) = M_j$. Then $0 \leq M_j - m_j < \varepsilon/(b-a)$. Multiply by the positive quantity Δx_j and sum over j to get

$$0 \leq \sum_{j=1}^{n} (M_j - m_j)\,\Delta x_j < \frac{\varepsilon}{b-a} \sum_{j=1}^{n} \Delta x_j = \varepsilon.$$

It follows from Inequality $(*)$ above that $|(F(b) - F(a)) - A_{\mathcal{P};\zeta}| < \varepsilon$, as required.

(b) This follows easily from the preceding. ■

V.9.2 Remark

The preceding discussion provides a natural motivation for one of the most characteristic notations found in classical analysis. Indeed, recall the equation $F(b) - F(a) = \sum_{j=1}^{n} \Delta F_j(x)$ mentioned above, which expresses the 'whole difference' $F(b) - F(a)$ as the sum of 'partial differences' $\Delta F_j(x)$ as j runs from $j = 1$ to $j = n$. Leibniz' treatment of calculus often replaces such discrete sums of ordinary quantities by 'continuous' sums of infinitely small quantities. Thus, he expresses $F(b) - F(a)$ as the 'continuous sum' of infinitely small differences $dF(x)$ as x runs from $x = a$ to $x = b$. Instead of our use of the upper case Greek letter Σ as a reminder of the first letter of the word 'sum', he uses the symbol $\int$, an elongated version of the letter 'S', for the same purpose when summing infinitely small quantities. Thus, when combined with the formula $dF(x) = F'(x)\,dx$ obtained above, one gets, using the hypothesis $F'(x) = f(x)$, that

$$F(b) - F(a) = \int dF(x) = \int f(x)\,dx.$$

It is assumed here that one knows that x runs from a to b either from the context or by stating it in words; the standard calculus notation $\int_a^b f(x)\,dx$ arose only at the beginning of the nineteenth century.

The process in calculus of obtaining 'whole' quantities by summing infinitely small 'partial quantities', as above, is called 'integration', from the Latin for 'whole' or 'complete'. Likewise, the symbol $\int$ is called the 'integral sign', and the quantity $\int f(x)\,dx$ is the 'integral of the function f'. From the time of Leibniz until the nineteenth century, the concept of 'integrals' was identified with what we now call 'antiderivatives'. Even now, most calculus texts refer to the methods of computing antiderivatives as 'Techniques of Integration'. The reason for the modern practice of separating the concept of 'integral' from that of 'antiderivative' grew from profound studies of Fourier, published in 1822, on the theory of heat flow. This transition is discussed at the begining of the next chapter.

V.10 The Standard Transcendental Functions

All the explicit functions considered so far in this chapter can be constructed using only the basic operations of algebra a finite number of times. One knows from elementary calculus, of course, that there other functions which also play vital roles in calculus; namely the standard exponential, logarithmic, trigonometric and inverse trigonometric functions. These special functions are sometimes called the **Standard Transcendental Functions** (because their construction 'transcends' the use of finite algebra).

The goal of the present section is to develop these functions rigorously by using the properties of antiderivatives obtained in the preceding section. This treatment is quite instructive, but somewhat tedious. Some instructors may elect to simply assume as 'known' the standard facts about these functions which one learns in elementary calculus and thus save time by skipping portions of this section.

The key facts needed to construct the standard transcendental functions are given in the next result.

V.10.1 Theorem

(a) There is a unique function $F : (0,+\infty) \rightarrow \mathbb{R}$ such that $F'(x) = \dfrac{1}{x}$ for all $x > 0$, and $F(1) = 0$.

(b) There is a unique function $G : (-1,1) \rightarrow \mathbb{R}$ such that $G'(x) = \dfrac{1}{\sqrt{1-x^2}}$ for all x in $(-1,1)$ and $G(0) = 0$.

Proof Both statements follow from the Cauchy Antiderivative Theorem because the functions $f : (0,+\infty) \rightarrow \mathbb{R}$ and $g : (-1,1) \rightarrow \mathbb{R}$, given by $f(x) = 1/x$ and $g(x) = \sqrt{1-x^2}$, respectively, are continuous on their given domains.

V.10.2 Definition

(1) The function F described in Part (a) of the preceding theorem is called the **natural logarithm function**; it is denoted ln, so that $F(x) = \ln x$ for all $x > 0$.

(2) The function G described in Part (b) of the preceding theorem is called the **(principle) arcsine function**, and is denoted Arcsin. (The reason for using the adjective 'principle' and upper-case letter 'A' will be explained later.)

Properties of the Natural Logarithm Function and Related Functions

The next result shows that the function ln defined above has the usual properties familiar from elementary calculus.

V.10.3 Theorem

The natural logarithm function ln has the following properties:

(a) $\ln'(x) = \dfrac{1}{x}$ for all $x > 0$, and $\ln(1) = 0$.

(b) $\ln(x{\cdot}y) = \ln(x) + \ln(y)$ for all $x, y > 0$.

(c) $\ln(x^{-1}) = -\ln(x)$ for all $x > 0$.

(d) $\ln(x^k) = k\ln x$ for all $x > 0$ and all k in $\mathbb{Z}$.

(e) $\lim_{x\rightarrow+\infty} \ln x = +\infty$, and $\lim_{x\searrow 0} \ln x = -\infty$.

(f) The function ln is a bijection of the interval $(0,+\infty)$ onto $\mathbb{R}$.

(g) $\lim_{x\rightarrow 0+} x^k \ln x = 0$ for every k in $\mathbb{N}$.

Partial Proof

(a) These are just the defining properties of the function ln; they are included in the list for ease of reference.

(b) For fixed $y > 0$ define functions g and h on $(0, +\infty)$ by the rules

$$g(x) = \ln(x{\cdot}y), \qquad h(x) = \ln(x) + \ln(y).$$

By the Chain Rule one has $g'(x) = \dfrac{y}{xy} = \dfrac{1}{x}$, while from the Addition Rule for Derivatives (and the fact that y is 'constant') one has $h'(x) = \dfrac{1}{x}$. That is, $g' = h'$ for each point of the domain $(0, +\infty)$. Thus, by Corollary (V.4.9), one has $h = g + C$ for some constant function C. However, note that $g(1) = \ln(1{\cdot}y) = \ln(y)$, and $h(1) = \ln(1) + \ln(y) = \ln(y)$. It follows that one must have $C = 0$ and thus $g = h$. In particular, for each $x, y > 0$ one has $\ln(x{\cdot}y) = \ln(x) + \ln(y)$, as claimed.

(c) Note that, by Part (b), one has $\ln(x) + \ln(x^{-1}) = \ln(x{\cdot}x^{-1}) = \ln(1) = 0$. The desired result follows easily.

(d), (e) and (f): The simple proofs of these parts of the theorem are left as exercises.

(g) Notice that one can express the quantity $x^k \ln x$ in the form $\dfrac{f(x)}{g(x)}$, where $\lim_{x\to 0+} f(x) = -\infty$ and $\lim_{x\to 0+} g(x) = +\infty$. Indeed, let $f(x) = \ln(x)$ and $g(x) = x^{-k}$. Note also that $f'(x) = 1/x$ and $g'(x) = -kx^{-k-1}$. Thus

$$\frac{f'(x)}{g'(x)} = \frac{1/x}{(-k)x^{-k-1}} = -\frac{x^k}{k}.$$

Since $\lim_{x\to 0+} x^k = 0$, L'Hôpital's Rule can be used to conclude that $\lim_{x\to 0+} x^k \ln x = 0$. ■

It follows from Part (f) of the preceding theorem that the function ln has an inverse $\ln^{-1} : \mathbb{R} \to (0, +\infty)$ which maps $\mathbb{R}$ onto the set of positive real numbers.

V.10.4 Definition

The function $\ln^{-1} : \mathbb{R} \to (0, +\infty)$ described above is called the (standard) **exponential function**, and is denoted by exp.

The next result summarizes the main properties of the function exp.

V.10.5 Theorem

(a) The function exp is differentiable at each point of $\mathbb{R}$, and $\exp'(x) = \exp(x)$ for all x in $\mathbb{R}$; also, $\exp(0) = 1$.

(b) $\exp(x + y) = (\exp(x)){\cdot}(\exp(y))$ for all x and y in $\mathbb{R}$.
Special case: $\exp(0) = 1$.

(c) If $x \in \mathbb{R}$ and $m \in \mathbb{Z}$ then $\exp(mx) = (\exp x)^m$.
Special case: $\exp(-x) = 1/\exp(x)$ for all x in $\mathbb{R}$.

(d) More generally, if $x \in \mathbb{R}$ and $r \in \mathbb{Q}$, then $\exp(rx) = (\exp x)^r$. ('Rational exponents' are defined in Example (IV.4.8).)

(e) Define e to be the number $\exp(1)$. Then $\ln(e) = 1$. In addition, $\exp(r) = e^r$ for all rational numbers r.

(f) $\lim_{x \to +\infty} x^k e^{-x} = 0$ for each k in $\mathbb{N}$.

(g) Let I be an open interval in $\mathbb{R}$, and suppose that $f : I \to \mathbb{R}$ is a function such that $f'(x) = f(x)$ for all x in I. Then there exists a constant A such that $f(x) = A\exp(x)$ for all x in I.

The proofs of Parts (a) through (f) follow easily from corresponding properties of the logarithm function. The proof of (g) can be obtained by differentiating the quotient $f/\exp$. The details are left as exercises. ■

In Definition (II.1.4) we introduced the standard 'power' notion c^p, where c is a real number and p is a natural number, as a shorthand for the repeated multiplication $c{\cdot}c{\cdot}\ldots{\cdot}c$, with the factor c appearing p times. Later we extended this 'exponent' notation to include expressions of the form c^p in which p can be any integer, at least if $c \neq 0$. We then further extended the notation c^p to make sense when p is any rational number, although now c needs to be a *positive* real number. In light of Part (e) of the preceding theorem, it is now possible to extend the exponent notation one more time to allow for exponents which can be any *real* number.

V.10.6 Definition

(1) The number $e = \exp(1)$ described in Part (e) of Theorem (V.10.5) is called the .

(2) If x is a real number then the x **power of the number** e is the number $\exp x$. It is usually denoted by the expression e^x, which is pronounced 'e to the power x'.

(3) More generally, if $b > 0$ and x is any real number, then the x **power of** b is the number $\exp(x{\cdot}\ln(b))$; that is, $e^{x\cdot\ln b}$. It is usually denoted by the expression b^x, pronounced 'b to the power x'.

The next result summarizes some of the main facts associated with the 'exponent notation'.

V.10.7 Theorem

Let b be a positive real number.

(a) If x is a real number then $b^x > 0$ and $b^{-x} = 1/b^x$; in particular, $b^0 = 1$.

(b) If x and y are real numbers then $b^{x+y} = b^x{\cdot}b^y$ and $(b^x)^y = b^{(xy)}$.

(c) If x a rational number, or an integer, or a natural number, then the value of the number b^x given in the preceding definition agrees with the value assigned to such expressions in our earlier definitions.

(d) Let $f : \mathbb{R} \to \mathbb{R}$ be defined by the formula $f(x) = b^x$ for each x in $\mathbb{R}$. Then f is differentiable on $\mathbb{R}$, and one has $f'(x) = (\ln(b))\,{\cdot}b^x$.

The simple proof is left as an exercise. ■

In Section (**??**) we construct certain C^k 'bump functions'; see Definition (**??**). Because the underlying functions for this construction are polynomials, these bump functions are not of class C^∞; they are *strictly* C^k. With the introduction of the transcendental function exp, however, one can produce C^∞ bump functions. The key is the following result.

V.10.8 Lemma

(a) For every k in $\mathbb{N}$ one has

$$\lim_{x \to 0+} \frac{e^{-1/x}}{x^k} = 0 \tag{V.15}$$

(b) For every k in $\mathbb{N}$ one has

$$\lim_{x \to 0} \frac{e^{-1/x^2}}{x^k} = 0 \tag{V.16}$$

Proof

(a) Let $u = 1/x$. Then one has

$$\frac{e^{-1/x}}{x^k} = u^k e^{-u}.$$

Moreover, one sees that $u \to +\infty$ as $x \to +0$. The desired result now follows from Part (g) of Theorem (V.10.5).

(b) This follows easily from Part (a); the details are left as an exercise. ■

V.10.9 Example

Let $f_0 : \mathbb{R} \to \mathbb{R}$ be given by the formula

$$f_0(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ e^{-1/x} & \text{if } x > 0 \end{cases}$$

Then f_0 is of class C^∞ at each point x of $\mathbb{R}$, even at $x = 0$; indeed, at $x = 0$ one has $f_0^{(k)}(0) = 0$ for all k in $\mathbb{N}$.

To see why this is the case, let $g : (0, +\infty) \to \mathbb{R}$ be given by the formula $g(x) = e^{-1/x}$. It is easy to see (by Mathematical Induction) that g is of class C^∞ at each point of the interval $(0, +\infty)$. In fact, each of the derivatives $g^{(k)}(x)$ is the sum of finitely many terms of the form $\frac{c_j}{x^j} e^{-1/x}$ for certain constants c_j. In particular, it follows from repeated use of Part (a) of the preceding lemma that the one-sided derivatives of f_0 at $x = 0$ from the right all exist and equal 0. And since the same is trivially true for the one-sided derivatives at 0 from the left, the desired result follows from the Concatenation Theorem for Derivatives.

Following the pattern used in Example (**??**) and Definition (**??**), one can use the function f_0 described above to construct C^∞ 'bump functions'. We summarize the construction as follows.

V.10.10 Definition

(1) Define $B^{[\infty]}_{[0,1]} : \mathbb{R} \to \mathbb{R}$ by the rule

$$B^{[\infty]}_{[0,1]}(x) = \begin{cases} 0 & \text{if } x \leq 0 \text{ or } x \geq 1 \\ f_0(x) f_0(1-x) & \text{if } 0 < x < 1. \end{cases}$$

Let M denote the maximum value of this function on the interval $[0,1]$, and set $\hat{B}^{[\infty]}_{[0,1]}(x) = B^{[\infty]}_{[0,1]}(x)/M$ for each x in $\mathbb{R}$. The function $\hat{B}^{[\infty]}_{[0,1]}$ is called the **normalized** C^∞ **bump function on** $[0,1]$.

(2) More generally, if $[a,b]$ is any closed interval in $\mathbb{R}$, then the function $B^{[\infty]}_{[a,b]} : \mathbb{R} \to \mathbb{R}$ given by the rule

$$B^{[\infty]}_{[a,b]}(x) \,=\, B^{[\infty]}_{[0,1]}\left(\frac{x-a}{b-a}\right) \text{ for all } x \text{ in } \mathbb{R}$$

is called the **normalized C^{∞} bump function on** $[a,b]$.

Remark These C^{∞} bump functions may appear to be mere curiosities, but they actually play an important role in advanced analysis as 'test functions' for distributions. The topic of 'distributions' is outside the scope of *This Textbook*.

Properties of the Arcsine Function and Related Functions

As in the case of the function ln, the function 'Arcsine' has been introduced into our theory in a purely analytical manner, as an antiderivative of an algebraically defined function. And as in the case of the logarithm, this function has an inverse function which is of great importance in analysis. The basis for all this is the following result.

V.10.11 Theorem

(a) The function Arcsin : $(-1,1) \to \mathbb{R}$ is an odd function, and is strictly increasing, on $(-1,1)$. Furthermore, its image is a bounded open interval of the form $(-c,c)$, where c is given by $c \,=\, \lim_{x \nearrow 1} \text{Arcsin}\, x$.

(b) The limit c described in Part (a) is finite. More precisely, it is a real number such that $0 \,<\, c \leq 2$.

(c) Let $g : (-c,c) \to (-1,1)$ be the inverse function Arcsin^{-1}, where the quantity c is described in Part (a). Then g satisfies the following conditions:

$$(i) \quad g''(x) \,=\, -g(x) \text{ for all } x \text{ in } (-c,c); \qquad (ii) \quad g(0) \,=\, 0 \text{ and } g'(0) \,=\, 1. \qquad \text{(V.17)}$$

In addition, g is also an odd function on $(-c,c)$.

Proof (a) To save writing, let us set $G(x) \,=\, \text{Arcsin}\,(x)$ for all x in the interval $(-1,1)$.

The fact that G is an odd function on $(-1,1)$ follows easily from the results obtained in Example (V.5.5) above, combined with the obvious fact that the function $f : (-1,1) \to \mathbb{R}$ given by the rule $f(x) \,=\, 1/\sqrt{1-x^2}$ is even on the interval $(-1,1)$.

Next, note that $G'(x) \,>\, 0$ for all x in the interval $(-1,1)$, so by Theorem (V.4.4), G is strictly increasing on $(-1,1)$. It now follows from Theorem (IV.4.7) that G is a bijection of $(-1,1)$ onto some open interval I in $\mathbb{R}$. Because G is an odd function, it follows that I is an open interval of the form $(-c,c)$, where $c \,=\, \lim_{x \nearrow 1} G(x)$ is either a positive real number or $+\infty$, as claimed.

(b) To estimate the size of the quantity c, observe that if $0 \leq x \,<\, 1$, then

$$\frac{1}{\sqrt{1-x^2}} \,=\, \frac{1}{\sqrt{(1+x)\,(1-x)}} \leq \frac{1}{\sqrt{1-x}} \qquad (*)$$

It is easy to see that if $H(x) \,=\, -2\sqrt{1-x}$, then $H'(x) \,=\, 1/\sqrt{1-x}$ for $0 \leq x \,<\, 1$. Combine this with Inequality $(*)$ above and Part (a) of Corollary (V.4.10), with the roles of f and g in that corollary being played here by G and H, respectively, to get

$$G(x) \,=\, G(x) - G(0) \leq H(x) - H(0) \,=\, 2 - 2\sqrt{1-x} \,<\, 2 \text{ for all } x \text{ such that } 0 \leq x \,<\, 1.$$

The desired inequality $0 < c \leq 2$ now follows by letting x approach 1 from below.

(c) The fact that g is differentiable on $(-c, c)$ follows from Theorem (**??**). One then computes $g'(y)$ using Equation (V.2): if $g(y) = x$, so that $y = \text{Arcsin}x$, then

$$g'(y) = \frac{1}{\text{arcsin}'(x)} = \frac{1}{(1/\sqrt{1-x^2})} = \sqrt{1-x^2} = \sqrt{1-g^2(y)} \text{ for each } y \text{ in } (-c, c).$$

From this one gets

$$g''(y) = \frac{-g(y)g'(y)}{\sqrt{1-g^2(y)}} = \frac{-g(y)\sqrt{1-g^2(y)}}{\sqrt{1-g^2(y)}} = -g(y) \text{ for each } y \text{ in } (-c, c),$$

as required. The fact that $g(0) = 0$ and $g'(0) = 1$ follows easily. ■

V.10.12 Remark

In elementary calculus we define the trigonometric functions and their inverses geometrically, in terms of angles (using radian measure). With that geometric interpretation, the Arcsine function has domain $(-\pi/2, \pi/2)$, which allows us to identify the number c described above with the geometrically defined number $\pi = 3.14159\ldots$. However, the reason for introducing this function first is because it can be done rigorously, using antiderivatives, without any reliance on geometry. The same holds for the number π: it can now be defined – without using geometry – in terms of the number c.

V.10.13 Definition (The Number π)

The number π is given by the formula

$$\pi = 2 \lim_{x \nearrow 1} \arcsin x. \tag{V.18}$$

Remark With this new notation one can replace the number c in our previous discussion with the expression $\pi/2$. Thus, one can write

$$\text{Arcsin} : (-1, 1) \rightarrow \left(\frac{-\pi}{2}, \frac{\pi}{2}\right),$$

with the

It would natural to define the function $g = \text{Arcsin}^{-1}$ described above to be the sine function, but there is a problem: the standard sine function one uses in calculus is defined on *all* of $\mathbb{R}$, while g is defined only on the bounded interval $(-\pi/2, \pi/2)$. This can be fixed by recalling that the usual sine function satisfies certain identities which allow one to extend it from the interval $(-\pi/2, \pi/2)$ to all other numbers. For example, in calculus one uses the fact that $\sin(\pm\pi/2) = \pm 1$. Likewise, one has the identity $\sin(x - k\pi) = (-1)^k \sin x$ for each k in $\mathbb{Z}$. These facts lead one to the following definition; basically, it describes the only function which satisfies these required identities and which agrees with g on the interval $(-\pi/2, \pi/2)$. Likewise, it defines the cosine function to be the derivative of the sine function.

V.10.14 Definition

(1) The function $g : (-\pi/2, \pi/2) \to \mathbb{R}$ described above is called the **restricted sine function**, and is denoted by $Sin : (-\pi/2, \pi/2) \to \mathbb{R}$; note the upper-case letter 'S'.

(2) The **standard sine function**, denoted $\sin : \mathbb{R} \to \mathbb{R}$, is given by the following rules:
(i) $\sin x = g(x)$ for $-\pi/2 < x < \pi/2$;
(ii) if $-\pi/2 + k\pi < x < \pi/2 + k\pi$, for some k in $\mathbb{Z}$, then $\sin x = (-1)^k g(x - k\pi)$.
(iii) $\sin(\pi/2 - k\pi) = (-1)^k$ for each k in $\mathbb{Z}$.

(3) The **cosine function** , denoted $\cos : \mathbb{R} \to \mathbb{R}$, is defined to be $\sin'$.

V.10.15 Theorem

The functions sin and cos defined above satisfy the following conditions:

(a) The function sin is C^∞ on $\mathbb{R}$, and $\sin'' = -\sin$. Moreover, $\sin(0) = 0$ and $\sin'(0) = 1$.

(b) The function cos is C^∞ on $\mathbb{R}$ and satisfies the equations $\cos' x = -\sin x$ and $\cos'' x = -\cos x$ for all x in $\mathbb{R}$. Moreover, $\cos(0) = 1$ and $\cos'(0) = 0$.

(c) One has $\sin^2(x) + \cos^2(x) = 1$ for all x in $\mathbb{R}$.

(d) The solutions of the equation $\sin x = 0$ are precisely the numbers of the form $k\pi$ with k in $\mathbb{Z}$. Likewise the solutions of the equation $\cos x = 0$ are precisely the numbers $(2k-1)\pi/2$ for k in $\mathbb{Z}$.

(e) Suppose $f : (a,b) \to \mathbb{R}$ is a function, with domain an open interval (a,b), such that $f'' = -f$ on (a,b). Then there exist unique constants A and B such that $f(x) = A\cos x + B\sin x$ for all x in (a,b).

(f) The sine and cosine functions satisfy the following 'Addition Formulas':

$$\sin(x_1 + x_2) = \sin x_1 \cos x_2 + \cos x_1 \sin x_2$$

and

$$\cos(x_1 + x_2) = \cos x_1 \cos x_2 - \sin x_1 \sin x_2$$

for all x_1 and x_2 in $\mathbb{R}$.

(g) The sine and cosine functions are periodic with period 2π. That is,

$$\sin(x + 2\pi) = \sin x \text{ and } \cos(x + 2\pi) = \cos x \text{ for all } x \text{ in } \mathbb{R}.$$

The proofs of these well-known properties are left as exercises. ■

The remaining basic trigonometric functions, namely tan, cot, sec and csc can be defined in terms of sin and cos as in elementary calculus, and their standard properties can then be derived in the usual way from the properties of sine and cosine obtained above. Similarly, the various inverse trig functions can be defined in terms of the standard trig functions in the standard manner. The details for all this are left to the exercises. Henceforth we shall feel free to use all these standard facts.

V.11 EXERCISES FOR CHAPTER V

V - 1 Prove the claim in Example (E.1.5) (2) that the function F in that example is differentiable precisely at the points r_1, $r_2, \ldots r_m$.

V - 2 Let $f : [0, +\infty) \to \mathbb{R}$ be the square-root function, so that $f(x) = \sqrt{x}$ for all $x > 0$. In Example (E.1.9) (2) on Page 240 it is shown that $f'(x)$ exists and equals $1/(2\sqrt{x})$ for every $x > 0$. However, the case $x = 0$ is not discussed there; of course in that case one would have to consider a one-sided derivative.

Consider the following argument:

'*From the formula* $f'(x) = 1/(2\sqrt{x})$ *for* $x > 0$ *one sees that* $\lim_{x\searrow 0} f'(x) = +\infty$. *This implies that* $f'(0)$ *cannot exists, since we require the values of* f' *to be finite.*

(a) Explain why this argument, although intuitively convincing, is not valid.

(b) Give a correct proof of the fact that the square root function is not differentiable (from the right) at 0.

V - 3 The proof of the Product Rule for Differentiation given in the *Notes* (see Page 247) is a little different from the 'standard proof' that one finds in most texts.

Problem Carry out the standard proof.

(Hint: This proof starts by using the ever-popular 'Add-and-Subtract' trick:

$$f(x)g(x) - f(c)g(c) = f(x)g(x) - f(x)g(c) + f(x)g(c) - f(c)g(c).$$

That is, one adds and subtracts the (cleverly chosen) quantity $f(x)g(c)$.)

V - 4 Prove the Quotient Rule for Differentiation (Part (b) of Theorem E.2.2 on Page 247).

V - 5 In the 'Remark' on Page 252, right after the proof of the Chain Rule, it is pointed out that if one assumes *continuous* first derivatives then one can get a much simpler proof of this result. Here is a more precise statement:

The Weakish Chain Rule Suppose that $f : I \to \mathbb{R}$ and $g : J \to \mathbb{R}$ are functions defined on open intervals I and J in $\mathbb{R}$. Assume that $f(x) \in J$ for all x in I, so that the composition $h = g \circ f : I \to \mathbb{R}$ is defined. Let c be a number in I and let $d = f(c)$, so $d \in J$. If f is differentiable at c and g is *continuously* differentiable at d, then h is differentiable at c, and $h'(c) = g'(d) \cdot f'(c)$. (Note that the hypothesis of g being differentiable at d tacitly requires that g' be defined on some open subinterval of J containing d.)

Problem Give a short – but rigorous – direct proof of the Weakish Chain Rule. ('Direct Proof': You are not allowed to simply say 'it's a special case of the regular Chain Rule'. However, you are free to use results, such as the Lagrange Mean-Value Theorem, which are not themselves based on the regular Chain Rule.)

V - 6 Let $f : I \to \mathbb{R}$ be a function defined on an open interval I.

(a) If $f''(c) > 0$ at some point c in I, then $f(x) > f(c) + f'(c)(x - c)$ for all x sufficiently near c.

(b) If $f''(x) > 0$ for all x in I, then for every c in I one has $f(x) > f(c) + f'(c)(x - c)$ for all x in I with $x \neq c$.

(c) Give a geometric interpretation of these results in terms of 'tangent lines'.

(d) Suppose that $f'(x)$ exists for each x in I, and assume that f has exactly one critical point in I; that is, there is exactly one value of x in I such that $f'(x) = 0$. Let c be that unique critical point.

Problem: Show that if $f''(c)$ exists and $f''(c) > 0$, then f has a strict minimum for I at c. That is, if $x \in I$ and $x \neq c$, then $f(x) > f(c)$.

V - 7 Let $f : I \to \mathbb{R}$ be a function defined on an open interval I. Assume that $f''(x) > 0$ for all x in I.

(a) Show that if a and b are in I, with $a < b$, then

$$f((1-t)a + tb) < (1-t)f(a) + tf(b) \text{ for all } t \text{ such that } 0 < t < 1.$$

(b) Give a geometric interpretation of this result in terms of the graph of f.

V - 8 The standard proof of Rolle's Theorem (Corollary E.4.2 on Page 266), as taught in courses on elementary calculus, does *not* assume that one has already proved the Lagrange Mean-Value Theorem (Theorem E.4.1 on Page 264). Indeed, the standard proof of the Mean-Value Theorem taught in such courses uses Rolle's Theorem.

(a) Give the standard proof of Rolle's Theorem. (Hint: What can you say about the location of the maximum and minimum values of f on $[a, b]$ in light of the hypothesis that $f(b) = f(a)$?)

(b) Use Rolle's Theorem to give the standard proof of the Lagrange Mean-Value Theorem. (Hint: Let $g : \mathbb{R} \to \mathbb{R}$ be the linear function whose graph passes through the endpoints, $(a, f(a))$ and $(b, f(b))$, of the graph of f on $[a, b]$. What does Rolle's Theorem say about $h = f - g$?)

V - 9 Suppose that $f : [0, 1] \to \mathbb{R}$ is continuous on $[0, 1]$ and differentiable on $(0, 1)$. Assume that $f(0) = 0$. Prove that if the derivative f' is monotonic up on the open interval $(0, 1)$, then so is the function $g : (0, 1) \to \mathbb{R}$ given by $g(x) = f(x)/x$.

V - 10 Suppose that f is differentiable on an open interval $(c - \delta, c + \delta)$, where $\delta > 0$. Let h satisfy $0 < h < \delta$.

(a) Show that there exists t with $0 < t < h$ such that

$$\frac{f(c+h) - f(c-h)}{h} = f'(c+t) + f'(c-t)$$

(b) Show that there exists τ with $0 < \tau < h$ such that

$$\frac{f(c+h) - 2f(c) + f(c-h)}{h} = f'(c+\tau) - f'(c-\tau).$$

(c) Show that if $f''(c)$ exists then

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

(d) Give an example of differentiable f such that the limit on the right side of Equation $(*)$ exists and is finite, but where $f''(c)$ does not exist.

V - 11 (a) Prove Parts (d) and (e) of Theorem E.6.14 (see Pages 287-288).

(b) Prove Part (g) of Theorem E.6.16 (see Pages 288-289.

V - 12 Prove that, for each x in $\mathbb{R}$, $e^x = \lim_{k\to\infty}\left(1+\dfrac{x}{k}\right)^k$.

Hint: Consider the quantity $\ln\left(\left(1+\dfrac{x}{k}\right)^k\right)$.

Note: Part of the solution of this exercise should include a proof of the fact that the limit in question actually exists; you cannot simply assume it.

V - 13 Let $g : I \to \mathbb{R}$ be a function which has an antiderivative on the open interval I.

Problem Determine the functions $f : I \to \mathbb{R}$, if any, such that $f'(x) = g(x)f(x)$ for all x in I.

V - 14 Suppose that f and g are C^2 functions on an open interval I. Let c be a point of I.

(a) Prove that the product functions $f{\cdot}g'$ and $g{\cdot}f'$ both have antiderivatives on I, and that

$$D_c^{-1}(f{\cdot}g') = f{\cdot}g - f(c){\cdot}g(c) - D_c^{-1}(g{\cdot}f').$$

(b) Suppose that f is a real-valued function such that f' is defined at each point of an open interval I in $\mathbb{R}$. Likewise, suppose that g is a function which has an antiderivative on an open interval J in $\mathbb{R}$; let $G : J \to \mathbb{R}$ be such an antiderivative. Assume that $f[I] \subseteq J$, so the composition $h = g{\circ}f : I \to \mathbb{R}$ is defined.

Problem: Show that the function $h{\cdot}f'$ has an antiderivative on I, and that if c is a point of I then $D_c^{-1}(h{\cdot}f') = G{\circ}f - G(f(c))$.

Remark In elementary calculus the result stated in Part (a) is called the **Law of Integration-by-Parts**, while the result in Part (b) is called the **Substitution Law**.

V - 15 Prove or Disprove: If $f : \mathbb{R} \to \mathbb{R}$ satisfies $|f(y) - f(x)| \leq (y - x)^2$ for all x and y in $\mathbb{R}$, then f is constant.

V - 16 Let $f : [a, b] \to \mathbb{R}$ be a function which is differentiable at each point of a closed bounded interval $[a, b]$. (At each endpoint one uses the appropriate one-sided derivative.) Consider the following statements about f:

(i) The function f is C^1 on $[a, b]$.

(ii) For every $\varepsilon > 0$ there exists $\delta > 0$ so that if t and x are in $[a, b]$ and $0 < |t - x| < \delta$, then $\left|\dfrac{f(t) - f(x)}{t - x} - f'(x)\right| < \varepsilon$.

Problem:

(a) Prove or disprove that Statement (i) implies Statement (ii).

(b) Prove or disprove that Statement (ii) implies Statement (i).

V - 17 Let f and g be functions which are continuous on a closed bounded interval $[a, b]$ and differentiable on the open interval (a, b). Prove that there exists c in (a, b) such that

$$f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a)).$$

V - 18 (a) Show that the sine function satisfies the equation $\sin''(x) = -\sin(x)$ except *possibly* at points of the form $\pi/2 - k\pi$ with k in $\mathbf{Z}$.

(b) Show that the sine function is continuous at each point of the form $\pi/2 - k\pi$ with k in $\mathbf{Z}$.

(c) Prove Part (a) of Theorem E.6.24 (see Page 293). (Hint: At points of the form $x = \pi/2 - k\pi$ consider using L'Hôpital's Rule.)

(d) Prove Part (b) of Theorem E.6.24.

V - 19 Note: In this problem you may assume Parts (a) and (b) of Theorem E.6.24 (see page 293).

(a) Prove Parts (c) and (d) of Theorem E.6.24.

(b) Prove Parts (e) and (f) of Theorem E.6.24. (Hint: You may wish to prove (f) first.)

V - 20 (a) Define the four remaining basic trigonometric functions (tan, cot, sec and csc) in terms of the sine and cosine functions. Make it clear at which points of $\mathbb{R}$ these functions fail to be defined.

(b) Derive the standard formulas for the derivatives of the functions tan, cot, sec and csc. ('Standard formula': the derivatives of tan and sec should be expressed in terms of tan and sec; likewise, the derivatives of cot and csc should be expressed in terms of cot and csc.)

(c) Prove that the tangent function maps the open interval $(-\pi/2, \pi/2)$ bijectively onto $\mathbb{R}$.

(d) Let $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$ be the inverse (relative to the interval $(-\pi/2, \pi/2)$) of the tangent function. Derive the formula for the derivative of the function arctan.

V - 21 Suppose that $f : I \to \mathbb{R}$ is a C^k function on the open interval I, and that c and x are in I. Throughout this exercise let p_{k-1} denotes the Taylor polynomial of order $k-1$ for f about the center c.

(a) Let $H_k : I \to \mathbb{R}$ be given by the rule $H_k(t) = \frac{1}{(k-1)!} f^{(k)}(t)(x-t)^{k-1}$ for all t in I. Show that

$$f(x) = p_{k-1}(x) + \left(D_c^{-1} H_k\right)(x) \qquad (*)$$

Hint: Define $h : I \to \mathbb{R}$ by the rule (REVISED)

$$h(t) = f(t) + f'(t)(x-t) + \frac{1}{2} f''(t)(x-t)^2 + \ldots + \frac{f^{(k-1)}(t)}{(k-1)!}(x-t)^{k-1} \text{ for all } t \text{ in } I.$$

Compute h' and notice the considerable simplification that occurs.

Remark Equation $(*)$ is often called the **Integral Form of the Taylor Formula with Remainder**; compare it with the 'Derivative Form' presented in Theorem E.7.9. (The word 'integral' here is used in the classical sense of 'indefinite integral; that is, 'antiderivative'.)

(b) Show that if m in $\mathbb{N}$ satisfies $1 \leq m \leq k$, then there exists τ in Seg$[c, x]$ such that

$$f(x) = p_{k-1}(x) + \frac{f^{(k)}(\tau)}{m(k-1)!}(x-\tau)^{k-m}(x-c)^m \qquad (**)$$

Note: The case $m = k$ in Equation $(**)$ is the 'Derivative Form' from Theorem E.7.9. The case $m = 1$ is called the **Cauchy Form** of the remainder.

(c) Let $f(x) = \ln(1+x)$ for $x > -1$, with $c = 0$.

Problem Use the Cauchy Form described above to show that $\lim_{k \to \infty} p_{k-1}(x) = f(x)$ for all x such that $-1 < x < 1$.

Chapter VI

Existence of Antiderivatives in R

In elementary calculus one learns that many important applied problems can be solved by determining antiderivatives of appropriate functions. For example, if $f : [a, b] rightarrow \mathbb{R}$ is a continuous real-value function defined on the closed interval $[a, b]$ such that $f(x){\geq}0$ for all x in $[a, b]$, and if $F : [a, b]{\to}[\mathbb{R}]$ IS an antiderivative of f on $[a, b]$, then the area under the graph of f over $[a, b]$ equals $F(b) - F(a)$.

Likewise, in the sciences often what can say about an important unknown quantity $Q(t)$ which depends on the elapsed time in some situation involves relations between its derivatives $Q'(t)$, $Q''(t)$ and so on, and knowable physical quantities. For example, suppose that an object of mass m is thrown vertically down from a high tower with initial height H_0 above the Earth and initial velocity V_0 at time $t = 0$. Assume that air resistance can be ignored and that H_0 is not too large. Then Newton's Law of Gravity says that the gravitational force $F(t)$ of the Earth on this object is a constant F_0 throughout its fall (i.e., until the object hits the Earth). That is, $F(t) = m\,A(t)$, where $A(t)$ is the acceleration of the object, so that $A(t)$ is a constant $A_0 = F_0/m$ for all such t. But, by definition, $A(t) = f''(t)$, where $f(t)$ denotes the height of the object at time t. That is, one knows the second derivative of the function of the function f. Because the force of gravity pulls objects *down*, so $f(t)$ is *deacreasing*it is easy to see that the acceleration A_0 is a *negative* constant $-g$, where $g > 0$.

Chapter VII

The Riemann Integral

Quotes for Chapter (VII):

> (1) Leibniz's 'Whole difference equal to sum of the partial differences.....'
>
> (2) 'It is frequently claimed that Lebesgue integration is as easy to teach as Riemann integration. This is probably true, but I have yet to be convinced that it is as easy to learn.'
> T. Korner, 'A Companion to Analysis'

VII.1 The Riemann Integral

Introduction One of the most important concepts in elementary calculus is that of the definite integral $\int_a^b f$ of a bounded function f over an interval $[a, b]$. Cauchy gave an analytical treatment of this quantity in [CAUCHY 1823] as the basis for his proof of Theorem (V.8.1), the Antiderivative Theorem; indeed, he stressed the need for such a theorem in the opening of that book. In contrast with Lebesgue's proof of the Antiderivative Theorem given in Chapter (V), however, Cauchy's approach does not involve approximating f by functions known to possess antiderivatives on $[a, b]$.

Later developments in analysis made it important to extend Cauchy's concept of the definite integral to one which can apply to discontinuous functions which may not have antiderivatives. To be useful, such extensions needed to maintain the major properties enjoyed by the concept in Cauchy's context of continuous functions. It turns out that for such an extension the expression $\int_a^b f$ does not make sense for every bounded function f on an interval $[a, b]$. Thus before computing the value of such an integral, one must first determine whether the expression even makes sense under the given extension; that is, whether f is 'integrable' over $[a, b]$ under this extension. The first major extension of Cauchy's formulation of the definite integral is due to Riemann; see [RIEMANN 1854]. Several other extensions of the definite integral concept appeared over the following decades, and it has become customary to refer to Riemann's extension as the 'Riemann integral'.

In *This Textbook* we follow an approach to the Riemann integral, due to Darboux, which simplifies Riemann's original treatment; see [DARBOUX 1875]. This approach can be based on the following question:

'What features should any reasonable concept of $\int_a^b f$ include?'

Certainly one would want any such concept to include the basic rules of ordinary integral calculus. These include:

Rule 1 Suppose f is integrable on $[a,b]$. If m and M satisfy $m \leq f(x) \leq M$ for all x in $[a,b]$, then $m\,(b-a) \leq \int_a^b f \leq M\,(b-a)$. In particular, these should hold if $m = \inf\{f(x) : x\in[a,b]\}$ and $M = \sup\{f(x) : x\in[a,b]\}$. Of course since $b-a > 0$, the left-hand inequality holds automatically if $m = -\infty$; likewise, the right-hand inequality holds automatically if $M = +\infty$ since $b-a > 0$. Hence this rule is of significance primarily when m and M can be chosen to be finite; that is, when f is bounded on $[a,b]$. We assume such boundedness throughout this discussion; unbounded functions are considered later.

Rule 2 Suppose again that (bounded) f is integrable on $[a,b]$. If $a < c < b$, then f should be integrable over each of the subintervals $[a,c]$ and $[c,b]$. Furthermore, one should have $\int_a^b f = \int_a^c f + \int_c^b f$.

VII.1.1 Remark

Repeated use of these two rules forms the basis for Darboux's approach to the definite integral. For example, repeated use of the first portion of Rule 2 would imply that if f is integrable on $[a,b]$, and if $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ is any partition of $[a,b]$, then f is integrable on each subinterval $[x_{j-1},x_j]$ of this partition, $1 \leq j \leq n$. Furthermore, Rule 1 would then imply that if for each $j = 1,2,\ldots n$ one sets $m_j = \inf\{f(x) : x\in[x_{j-1},x_j]\}$ and $M_j = \sup\{f(x) : x\in[x_{j-1},x_j]\}$, then one has

$$m_j\,(x_j - x_{j-1}) \leq \int_{x_{j-1}}^{x_j} f \leq M_j\,(x_j - x_{j-1}) \text{ for each } j = 1,2,\ldots n.$$

Summing over the preceding inequalities and using the second portion of Rule 2 then implies that

$$\sum_{j=1}^{n} m_j\,(x_j - x_{j-1}) \leq \int_a^b f \leq \sum_{j=1}^{n} M_j\,(x_j - x_{j-1}).$$

The simple proof is left as an exercise.

Side Comment (on Fourier series and the Riemann integral) STILL TO BE WRITTEN

The next definition assigns names to the quantities which appear in the preceding Remark.

VII.1.2 Definition

Let $f : [a,b] \to \mathbb{R}$ be a function which is bounded on a closed interval $[a,b]$; continuity of f is not assumed here, nor is it assumed that $[a,b]$ is the full domain of f.

For a partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ of $[a,b]$, let $m_j = \inf\{f(x) : x\in[x_{j-1},x_j]\}$ and let $M_j = \sup\{f(x) : x\in[x_{j-1},x_j]\}$. For convenience write $\Delta x_j = x_j - x_{j-1}$ for each index j. The number $L(f;\mathcal{P}) = \sum_{j=1}^{n} m_j\,\Delta x_j$ is called the **lower Darboux sum** associated with the function f and the partition $\mathcal{P}$; similarly, the number $U(f;\mathcal{P}) = \sum_{j=1}^{n} M_j\,\Delta x_j$ is the corresponding **upper Darboux sum**. The quantity $\Delta(f;\mathcal{P}) = U(f;\mathcal{P}) - L(f;\mathcal{P})$ is the corresponding **Darboux difference** associated with the function f and the partition $\mathcal{P}$.

Remarks (1) The notations m_j, M_j and Δx_j are ambiguous, in that they assume that the context makes clear which function f and partition $\mathcal{P}$ is under consideration. Normally this ambiguity causes no problems.

(2) If f is continuous on $[a,b]$, then m_j and M_j are the minimum and maximum values of f on the subinterval $[x_{j-1}, x_j]$, so the notation $L(f;\mathcal{P})$ and $U(f;\mathcal{P})$ agrees with that used in Section (V.5).

The conclusions of Remark (VII.1.1) above can now be written to say that any reasonable definition of $\displaystyle\int_a^b f$ should satisfy the inequalities

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

for every partition $\mathcal{P}$ of $[a,b]$. In particular, $\int_a^b f$ should be an upper bound for the set of all numbers of the form $L(f;\mathcal{P})$ and a lower bound for the set of numbers of the form $U(f;\mathcal{P})$. Thus one ought to have

$$\sup\nolimits_{\Pi[a,b]} \{L(f;\mathcal{P})\} \leq \int_a^b f \leq \inf\nolimits_{\Pi[a,b]} \{U(f;\mathcal{P})\},$$

where the sup and inf are both taken over the set $\Pi([a,b])$ of all partitions $\mathcal{P}$ of $[a,b]$. The next result outline properties of these quantities.

VII.1.3 Lemma

Suppose that $f : [a,b] \to \mathbb{R}$ is bounded on the interval $[a,b]$.

(a) If $\mathcal{P}$ is any partition of $[a,b]$, then $L(f;\mathcal{P}) \leq U(f;\mathcal{P})$. Furthermore one has $\Delta(f;\mathcal{P}) \geq 0$, with equality if, and only if, f is constant on $[a,b]$.

(b) Suppose that $\mathcal{P}$ and $\mathcal{Q}$ are partitions of $[a,b]$ such that $\mathcal{Q}$ is a refinement of $\mathcal{P}$; that is, $\mathcal{P} \subseteq \mathcal{Q}$. Then

$$L(f;\mathcal{P}) \leq L(f;\mathcal{Q}) \leq U(f;\mathcal{Q}) \leq U(f;\mathcal{P}).$$

(c) If $\mathcal{Q}$ and $\mathcal{R}$ are partitions of $[a,b]$, then $L(f;\mathcal{Q}) \leq U(f;\mathcal{R})$. In particular, one has

$$L(f;\mathcal{Q}) \leq \sup\nolimits_{\Pi[a,b]} \{L(f;\mathcal{P})\} \leq \inf\nolimits_{\Pi[a,b]} \{U(f;\mathcal{P})\} \leq U(f;\mathcal{R})$$

for every pair of partitions $\mathcal{Q}$ and $\mathcal{R}$ of $[a,b]$.

(d) Let c be a number such that $a < c < b$. Then a partition $\mathcal{P}$ of $[a,b]$ contains the point c if, and only if, there are partitions $\mathcal{Q}$ and $\mathcal{R}$ of the intervals $[a,c]$ and $[c,b]$, respectively, such that $\mathcal{P} = \mathcal{Q} \cup \mathcal{R}$. Furthermore, when this situation holds one has $L(f;\mathcal{P}) = L(f;\mathcal{Q}) + L(f;\mathcal{R})$ and $U(f;\mathcal{P}) = U(f;\mathcal{Q}) + U(f;\mathcal{R})$.

Proof (a) This statement follows directly from the observation that if S is any nonempty set of real numbers, then $\inf S \leq \sup S$, with equality if, and only if, S is a singleton set.

(b) Let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be any partition of $[a,b]$. Suppose first that $\mathcal{Q}$ is obtained from $\mathcal{P}$ by adjoining a single new point q to the finite set $\mathcal{P}$. To be precise, suppose that $x_{k-1} < q < x_k$ for some index k. Then one has

$$L(f;\mathcal{P}) = m_1\,\Delta x_1 + \ldots + m_{k-1}\,\Delta x_{k-1} + m_k\,\Delta x_k + m_{k+1}\,\Delta x_{k+1} + \ldots + m_n\,\Delta x_n.$$

In contrast, the sum expressing $L(f;\mathcal{Q})$ consists of exactly the same terms, except that the single term $m_k\,\Delta x_k = m_k\,(x_k - x_{k-1})$ is replaced by two terms $m'_k\,(q - x_{k-1}) + m''_k\,(x_k - q)$, where $m'_k = \inf\{f(x) : x_{k-1} \leq x \leq q\}$ and $m''_k = \inf\{f(x) : q \leq x \leq x_k\}$. It is clear from properties of the infimum that $m_k \leq m'_k$ and $m_k \leq m''_k$, and thus

$$m_k\,\Delta x_k = m_k\,(x_k - x_{k-1}) = m_k\,((q - x_{k-1}) + (x_k - q)) \leq m'_k\,(q - x_{k-1}) + m''_k\,(x_k - q).$$

In other words, to obtain $L(f;\mathcal{Q})$ one replaces the term $m_k\,\Delta x_k$ in the sum for $L(f;\mathcal{P})$ by the sum $m'_k\,(p - x_{k-1}) + m''_k\,(x_k - p)$, which is either bigger than or at worst equal to the first term. The equation $L(f;\mathcal{P}) \leq L(f;\mathcal{Q})$ follows in this special case. To get the general case, in which $\mathcal{Q}$ is obtained from $\mathcal{P}$ by adjoining any finite number of new elements, simply repeat the preceding argument and use mathematical induction. A similar argument shows that $U(f;\mathcal{Q}) \leq U(f;\mathcal{P})$. Finally, by Part (a) the inequality $L(f;\mathcal{Q}) \leq U(f;\mathcal{Q})$ is true.

(c) Let $\mathcal{Q} = \mathcal{P} \cup \mathcal{R}$, so that $\mathcal{Q}$ is a refinement of both $\mathcal{P}$ and $\mathcal{R}$. Then by Part (b) one has

$$L(f;\mathcal{P}) \leq L(f;\mathcal{Q}) \leq U(f;\mathcal{Q}) \leq U(f;\mathcal{R})$$

The desired results then follow from the transitivity property of order and the definitions of 'sup' and 'inf'.

(d) The simple proof is left as an exercise.

The preceding results suggest the following.

VII.1.4 Definition

Let f be a function bounded on an interval $[a,b]$.

A number B is said to be a **Darboux number for f on $[a,b]$** provided $L(f;\mathcal{P}) \leq B \leq U(f;\mathcal{P})$ for every partition $\mathcal{P}$ of $[a,b]$.

Remark The preceding discussion can now be reformulated to say that in any reasonable definition of $\int_a^b f$, this number must be a Darboux number of f on $[a,b]$. Furthermore, Part (c) of the preceding result implies that such Darboux numbers do exist; namely $\sup_{\Pi[a,b]}\{L(f;\mathcal{P})\}$ and $\inf_{\Pi[a,b]}\{U(f;\mathcal{P})\}$, together with any number between these values. Darboux now focusses on the case in which there is only one such number, hence only one reasonable choice for $\int_a^b f$.

VII.1.5 Definition

Let f be a function bounded on an interval $[a,b]$.

One says that f is **integrable on $[a,b]$ in the sense of Darboux** provided that there exists precisely one Darboux number B for f on $[a,b]$. If this is the case, then one writes $B = \displaystyle\int_a^b f$, and one calls the expression $\displaystyle\int_a^b f$ the **Riemann integral of f over $[a,b]$**. In this context the function f is called the **integrand** and the numbers a and b are called the **limits of integration**.

Remarks (1) It is common to use the classical 'variables' notation for functions and write something like $\displaystyle\int_a^b f(x)\,dx$ in place of the simpler expression $\displaystyle\int_a^b f$. The role of the quantity x here is the as the **variable of integration**. The letter x can be replaced by any letter that is not already in mathematical use here.

(2) The 'Darboux lower sum' and 'Darboux upper sum' terminology is standard in analysis; in contrast, the 'Darboux difference' and 'Darboux number' terminology, while convenient, is not.

(3) In calculus texts the quantity $\int_a^b f(x)\,dx$ described above is normally introduced following an earlier approach of Riemann; see below. Riemann's approach to this quantity is technically more complicated than that of Darboux, but it avoids the use of the concepts of suprema and infima, concepts which are considered too difficult for elementary calculus. Since the two approaches lead to precisely the same set of integrable functions and the same values for their integrals, and Riemann's approach is earlier than that of Darboux, it is customary to assign the name 'Riemann integral' to this quantity, regardless of which approach is taken, but to distinguish between 'Darboux integrable' and 'Riemann integrable' to indicate which approach one follows in obtaining the Riemann integral. However, there are exceptions: some authors refer to the 'Darboux integral' instead.

For some purposes one needs a somewhat more general formulation of 'Darboux number'.

VII.1.6 Lemma

Let $\mathcal{R}$ be a particular (fixed) partition of $[a,b]$. Then a necessary and sufficient condition for a number B to be a Darboux number for a function f bounded on $[a,b]$ is that

$$L(f;\mathcal{Q}) \leq B \leq U(f;\mathcal{Q}) \text{ for every refinement } \mathcal{Q} \text{ of } \mathcal{R}.$$

Proof

The necessity of the given condition is obvious. To show its sufficiency, suppose that B satisfies the given condition, and let $\mathcal{P}$ be any partition of $[a,b]$. As in the proof of Part (c) above, let $\mathcal{Q} = \mathcal{P}\cup\mathcal{R}$, so that $\mathcal{Q}$ is a refinement of $\mathcal{R}$. Then the hypothesis here says that $L(f;\mathcal{Q}) \leq B \leq U(f;\mathcal{Q})$, so that by Part (c) again one has

$$L(f;\mathcal{P}) \leq L(f;\mathcal{Q}) \leq B \leq U(f;\mathcal{Q}) \leq U(f;\mathcal{R})$$

for every partition $\mathcal{P}$ of $[a,b]$. The claimed sufficiency now follows.

VII.1.7 Examples

In the following examples, the word 'integrable' means 'integrable in the sense of Darboux'.

(1) If $f : I \to \mathbb{R}$ is continuous on an open interval I, then, by Part (b) of Theorem (V.9.1), for every pair of numbers a and b in I such that $a < b$, the function f is integrable on $[a,b]$. Furthermore, $\int_a^b f = F(b) - F(a)$, where F is any antiderivative of f on I.

(2) If $f : [a,b] \to \mathbb{R}$ is monotonic up on $[a,b]$, then f is integrable on $[a,b]$. This is trivially true by Part (1) if $f(a) = f(b)$, since in this case f is constant, hence continuous, on $[a,b]$.

Now assume that $f(a) < f(b)$. Suppose, in contradiction, that f is *not* integrable on $[a,b]$. Then there exist Darboux numbers B_1 and B_2 for f on $[a,b]$ with $B_1 < B_2$, such that

$$L(f;\mathcal{P}) \leq B_1 < B_2 \leq U(f;\mathcal{P}) \text{ for every partition } \mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\} \text{ of } [a,b].$$

Using the notation from Part (1) of Definition (VII.1.2), together with the hypothesis that f is monotonic up, one gets

$$L(f;\mathcal{P}) \,=\, \sum_{j=1}^{n} f(x_{j-1})\,\Delta x_j \,\leq\, B_1 \,<\, B_2 \leq \sum_{j=1}^{n} f(x_j)\,\Delta x_j \,=\, U(f;\mathcal{P}).$$

It follows that

$$0 \,<\, B_2 - B_1 \leq \sum_{j=1}^{n} \left(f(x_j) - f(x_{j-1})\right)\,\Delta x_j \,\leq\, \left(\sum_{j=1}^{n}(f(x_j) - f(x_{j-1}))\right) ||\mathcal{P}|| \,=\, (f(b) - f(a))\,||\mathcal{P}||$$

for *every* partition $\mathcal{P}$ of $[a,b]$. Now chose $\mathcal{P}$ so that $||\mathcal{P}|| \,<\, (B_2 - B_1)/(f(b) - f(a))$ to get the contradiction.

Clearly the conclusion, that f is integrable, remains true if one assumes instead that f is monotonic *down* on $[a,b]$.

(3) Let $f : [0,1] \rightarrow \mathbb{R}$ be the restriction to $[0,1]$ of the Dirichlet function, so that for x in $[0,1]$ one has $f(x) \,=\, 0$ if x is irrational while $f(x) \,=\, 1$ if x is rational. Then clearly $L(f;\mathcal{P}) \,=\, 0$ and $U(f;\mathcal{P}) \,=\, 1$ for every partition $\mathcal{P}$ of $[0,1]$. In particular, f is *not* integrable on $[0,1]$. Indeed, if B is any of the infinitely many numbers in the closed interval $[0,1]$, then B is a Darboux number for f on [0,1].

(4) Let $f : [0,1] \rightarrow \mathbb{R}$ be the restriction to $[0,1]$ of the Thomae function; that is, for x in $[0,1]$ one has $f(x) \,=\, 0$ if x is irrational, or if $x \,=\, 0$; while if $x \,>\, 0$ is rational and of the form $x \,=\, p/q$, with p, q being natural numbers having no common factors bigger than 1, then $f(x) \,=\, 1/q$. It is clear that if $\mathcal{P}$ is any partition of $[0,1]$, then $L(f;\mathcal{P}) \,=\, 0$ (since every subinterval of $[0,1]$ includes irrational points), and that $U(f;\mathcal{P}) \leq 1$ (since $1/q \leq 1$ for each q in $\mathbb{N}$). In particular, 0 is a Darboux number for f on $[0,1]$. To see that 0 is the only such Darboux number, suppose that B satisfies $0 \,<\, B \leq 1$, and let n be any natural number such that $1/n \,<\, B/3$. If x in $[0,1]$ satisfies $f(x) \geq B$, so that $f(x) \geq B/3 \,>\, 1/n$, then x can be written in the form p/q, as described above, with $1 \leq p \leq q \,<\, n$. Clearly there are at most n^2 such values of x. Now let $\mathcal{P}$ be the partition of $[0,1]$ into n^3 subintervals of equal length $1/n^3$. At most $2\,n^2$ of these subintervals contain x such that $f(x) \,>\, 1/n$. (The factor 2 reflects the fact that such x might be an endpoint of two contiguous subintervals.) The sum of the terms in the expression for $U(f;\mathcal{P})$ from those subintervals then is strictly bounded above by $2\,n^2/n^3 \,=\, 2/n$. The sum of the remaining terms appearing in $U(f;\mathcal{P})$ is clearly bounded above by $1/n$, so that $U(f;\mathcal{P}) \,<\, 3/n \,<\, B$, so that B is not a Darboux number for f on $[0,1]$.

Remark The last example shows that f being Riemann integrable on $[a,b]$ implies nothing about f having an antiderivative on $[a,b]$. Indeed, the Intermediate-Value Theorem for Derivatives implies that the Thomae function has an antiderivative on *no* subinterval of $[0,1]$.

VII.1.8 Corollary

Suppose that $f : [a,b] \rightarrow \mathbb{R}$ is bounded on the interval $[a,b]$ and that $\mathcal{R}$ is a fixed partition of $[a,b]$. Let $L_{(f;\mathcal{R})} \,=\, \{L(f;\mathcal{P}) : \mathcal{P} \text{ is a refinement of } \mathcal{R}\}$, and let $U_{(f;\mathcal{R})} \,=\, \{U(f;\mathcal{Q}) :\; \mathcal{Q} \text{ is a refinement of } \mathcal{R}\}$. Then:

(a) The sets $L_{(f;\mathcal{R})}$ and $U_{(f;\mathcal{R})}$ are nonempty bounded subsets of $\mathbb{R}$, and one has $\sup L_{(f;\mathcal{R})} \leq \inf U_{(f;\mathcal{R})}$. Furthermore, a number B is a Darboux number for f on $[a,b]$ if, and only if, $\sup L_{(f;\mathcal{R})} \leq B \leq \inf U_{(f;\mathcal{R})}$.

(b) The function f is integrable on $[a,b]$ in the sense of Darboux if, and only if, $\sup L_{(f;\mathcal{R})} = \inf U_{(f;\mathcal{R})}$. If this equality holds, then the common value of these two expressions equals $\displaystyle\int_a^b f$.

Proof (a) The fact that $L_{(f;\mathcal{R})}$ and $U_{(f;\mathcal{R})}$ are nonempty sets of numbers follows from the fact that there exists at least one partition of the given interval $[a,b]$ which is a refinement of $\mathcal{R}$; for example, $\mathcal{R}$ itself. It then follows from Part (b) of the preceding theorem that the set $L_{(f;\mathcal{R})}$ is bounded below and the set $U_{(f;\mathcal{R})}$ is bounded above, namely by the numbers $L(f;\mathcal{R})$ and $U(f;\mathcal{R})$, respectively. Furthermore, it follows from Part (d) of the preceding theorem that if $x \in L_{(f;\mathcal{R})}$ and $y \in U_{(f;\mathcal{R})}$, then $x \leq y$; in particular, X is bounded and Y is bounded below. It now follows from Part (d) of Theorem (II.4.14) that $\sup L_{(f;\mathcal{R})} \leq \inf U_{(f;\mathcal{R})}$, as claimed. It also follows that if B is any number such that $\sup L_{(f;\mathcal{R})} \leq B \leq \inf U_{(f;\mathcal{R})}$, then B has the desired property.

(b) From the proof of the preceding part of this theorem it is clear that if $\sup L_{(f;\mathcal{R})} < \inf U_{(f;\mathcal{R})}$, then there are infinitely many values of B such that $L(f;\mathcal{P}) \leq B \leq U(f;\mathcal{P})$ for every refinement $\mathcal{P}$ of $\mathcal{R}$, in which case f is *not* integrable on $[a,b]$ in the sense of Darboux. In contrast, it likewise follows that if $\sup L_{(f;\mathcal{R})} = \inf U_{(f;\mathcal{R})}$, then there is precisely one value of B for which $L(f;\mathcal{P}) \leq B \leq U(f;\mathcal{P})$ for every such partition; namely the common value of $\sup L_{(f;\mathcal{R})}$ and $\inf U_{(f;\mathcal{R})}$. The desired result now follows. ■

VII.1.9 Remarks

(1) It follows from Part (a) of the preceding corollary that the numbers $\sup L_{(f;\mathcal{R})}$ and $\inf U_{(f;\mathcal{R})}$ described there do not depend on the choice of the fixed partition $\mathcal{R}$ of $[a,b]$. In particular, suppose that one chooses $\mathcal{R}$ to be the 'trivial' partition $\{a,b\}$ of $[a,b]$, so that $L_{(f;\mathcal{R})} = L_{(f;\{a,b\})}$ and $U_{(f;\mathcal{R})} = U_{(f;\{a,b\})}$. Then the set of partitions $\mathcal{P}$ referred to in the corollary is simply set of *all* partitions of $[a,b]$.

(2) There are several popular notations for the quantities $\sup L_{(f;\mathcal{R})}$ and $\inf U_{(f;\mathcal{R})}$ considered above, for example:

(i) $\sup L_{(f;\mathcal{R})} = \underline{I}(f;[a,b])$; $\inf U_{(f;\mathcal{R})} = \overline{I}(f;[a,b])$.

(ii) $\sup L_{(f;\mathcal{R})} = \underline{\int}_a^b f$; $\inf U_{(f;\mathcal{R})} = \overline{\int}_a^b f$.

Because of the notations in (ii), many texts refer to $\sup L_{(f;\mathcal{R})}$ as the **lower Darboux integral of f on** $[a,b]$, and to $\inf L_{(f;\mathcal{R})}$ as the **upper Darboux integral of f on** $[a,b]$. In *This Textbook* we avoid the use of such specialized notations and terminology, but the reader needs to know that it does appear elsewhere.

There are several alternate ways to characterize the property of a function being integrable on an interval. In the following theorem and its proof, the word 'integrable' means 'integrable in the sense of Darboux'.

VII.1.10 Theorem

Let $f:[a,b] \to \mathbb{R}$ be a function which is bounded on the interval $[a,b]$. Then the following statements are equivalent:

Statement (i) The function f is integrable in the sense of Darboux on $[a,b]$; that is, f has exactly one Darboux number for $[a,b]$.

Statement (ii) For every $\varepsilon > 0$ there exists a partition $\mathcal{R}$ of $[a,b]$ such that $\Delta(f;\mathcal{R}) < \varepsilon$, where, as usual, $\Delta(f;\mathcal{R})$ is the (nonnegative) Darboux difference $U(f;\mathcal{R}) - L(f;\mathcal{R})$ associated with the partition $\mathcal{R}$.

Statement (iii) For every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathcal{R}$ is any partition of $[a,b]$ for which $||\mathcal{R}|| < \delta$, then

$$\Delta(f;\mathcal{R}) < \varepsilon.$$

Statement (iv) There exists a number B with the following property: for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ is any partition of $[a,b]$ for which $||\mathcal{P}|| < \delta$, then for any ordered list $\zeta = (z_1, z_2, \ldots z_n)$ of numbers with $x_{j-1} \leq z_j \leq x_j$ for each j one has

$$\left| B - \sum_{j=1}^{n} f(z_j)\Delta x_j \right| < \varepsilon.$$

Proof The fact that Statement (i) implies Statement (ii) is a simple consequence of Part (b) of Corollary (VII.1.8). Indeed, suppose that f is integrable on $[a,b]$, so that by that corollary one has $\sup L_{(f;\{a,b\})} = \inf U_{(f;\{a,b\})}$. Let B be the common value of these quantities. It then follows easily from the Approximation Property for infima and suprema that for every $\varepsilon > 0$ there exist partitions $\mathcal{P}$ and $\mathcal{Q}$ of $[a,b]$ such that

$$L(f;\mathcal{P}) > B - \frac{\varepsilon}{2} \text{ and } U(f;\mathcal{Q}) < B + \frac{\varepsilon}{2}.$$

Now let $\mathcal{R} = \mathcal{P} \cup \mathcal{Q}$. The desired inequality follows from Part (b) of Lemma (VII.1.3).

As for the converse, namely that Statement (ii) implies Statement (i), recall that for every partition $\mathcal{R}$ of $[a,b]$ one has

$$L(f;\mathcal{R}) \leq \sup L_{(f;\{a,b\})} \leq \inf U_{(f;\{a,b\})} \leq U(f;\mathcal{R})$$

If Statement (ii) holds, then for every $\varepsilon > 0$ there exists $\mathcal{R}$ such that

$$0 \leq U_{(f;\{a,b\})} - L_{(f;\{a,b\})} \leq U(f;\mathcal{R}) - L(f;\mathcal{R}) < \varepsilon,$$

which implies that $U_{(f;\{a,b\})} = L_{(f;\{a,b\})}$, and thus, by Part (b) of Corollary (VII.1.8), f is integrable on $[a,b]$.

To see that Statement (ii) implies Statement (iii), let $\mathcal{R} = \{a = y_0 < y_1 < \ldots < y_{N-1} < y_N = b\}$ be a partition of $[a,b]$ such that $\Delta(f;\mathcal{R}) < \varepsilon/2$; such $\mathcal{R}$ exists by the hypothesis that Statement (ii) holds. Let δ_1 be the minimum of the lengths of the N subintervals of $[a,b]$ determined by the partition $\mathcal{R}$; this is (probably) *not* the δ referred to in Statement (iii). To set up the search for the true δ, let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be any partition of $[a,b]$ such that $||\mathcal{P}|| < \delta_1$, and let $\mathcal{Q} = \mathcal{P} \cup \mathcal{R}$; note that $\mathcal{Q}$ is a refinement of $\mathcal{R}$, so that $\Delta(f;\mathcal{Q}) < \varepsilon/2$ as well. Because $0 < x_j - x_{j-1} < \delta_0$ for each j, it is clear that for each j either the subinterval $[x_{j-1}, x_j]$ of the partition $\mathcal{P}$ is also a subinterval of $\mathcal{Q}$, or there exists exactly one index k such that $x_{j-1} < y_k < x_j$; in that case $[x_{j-1}, x_j]$ is the union of exactly two subintervals for $\mathcal{Q}$, namely $[x_{j-1}, y_k]$ and $[y_k, x_j]$. It follows that one can write $\Delta(f;\mathcal{P}) = \Delta(f;\mathcal{Q}) + S$, where S is the sum of at most $N-1$ terms, each of the form $(M_j - m_j)(x_j - x_{j-1}) - ((M'_k - m'_k)(y_k - x_{j-1}) + (M''_k - m''_k)(x_j - y_k))$. In this last expression, as usual $m_j = \inf\{f(x) : x_{j-1} \leq x \leq x_j\}$, and $M_j = \sup\{f(x) : x_{j-1} \leq x \leq x_j\}$; likewise, m'_k, M'_k and m''_k, M''_k are the analogous quantities on $[x_{j-1}, y_k]$ and $[y_k, x_j]$, respectively. Let $L = \sup\{f(x) : a \leq x \leq b\} - \inf\{f(x) : a \leq x \leq b\}$, Then each term forming the sum S

above has magnitude no bigger than $3\,L\,||\mathcal{P}||$. Since S has at most $N-1$ such terms, it follows that certainly $|S| \leq 3\,(N-1)\,L\,||\mathcal{P}|| \leq 3\,N\,(L+1)\,||\mathcal{P}||$. Now let $d_2 = \varepsilon/(6\,N\,(L+1))$, and let $\delta = \min\{\delta_1, \delta_2\}$. Then $||\mathcal{P}|| < \delta$ implies that $|S| < \varepsilon/3$, and thus

$$\Delta(f;\mathcal{P}) = \Delta(f;\mathcal{Q}) + S \leq \Delta(f;\mathcal{Q}) + |S| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

as required. (The use of N and $L+1$ the weaker estimate for $|S|$, instead of the more accurate L and $N-1$, is to avoid the possible division-by-zero problem in defining δ_2.)

The proof that Statement (iii) implies Statement (ii) is obvious.

Next, suppose that Statement (iii) holds, and thus, as has been shown, Statements (i) and (ii) also hold, so that f is integrable on $[a,b]$. Let $B = \int_a^b f$. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that $0 \leq U(f;\mathcal{P}) - L(f;\mathcal{P}) < \varepsilon$ for any partition $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ for which $||\mathcal{P}|| < \delta$. Note that, by definition of 'integral', one has $L(;f\mathcal{P}) \leq B \leq U(f;\mathcal{P})$. Furthermore, if $\zeta = (z_1, z_2, \ldots z_n)$ is as in Statement (iv), then it is clear that

$$L(f;\mathcal{P}) \leq \sum_{j=1}^{n} f(z_j)\Delta x_j \leq U(f;\mathcal{P})$$

It follows that

$$\left| B - \left(\sum_{j=1}^{n} f(z_j)\Delta x_j \right) \right| \leq U(f;\mathcal{P}) - L(f;\mathcal{P}) < \varepsilon,$$

as required.

As for the converse, suppose that f is *not* integrable on $[a,b]$. Then $\sup L_{(f;\{a,b\})} < \inf U_{(f;\{a,b\})}$, so there must exist Darboux numbers B_1 and B_2 for f on $[a,b]$ such that $\sup L_{(f;\{a,b\})} < B_1 < B_2 < \inf U_{(f;\{a,b\})}$. Let $\varepsilon_0 = B_2 - B_1 > 0$. It is a simple exercise to show, using the Approximation Property for infima and suprema, that if $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ is any partition of $[a,b]$, then there exist ordered lists $\zeta = (z_1, z_2, \ldots z_n)$ and $\tau = (t_1, t_2, \ldots t_n)$, satisfying $x_{j-1} \leq z_j, t_j \leq x_j$ for each $j = 1, 2, \ldots n$, such that

$$\sum_{j=1}^{n} f(z_j)\,\Delta x_j \leq B_1 < B_2 \leq \sum_{j=1}^{n} f(t_j)\,\Delta x_j.$$

This implies that $\left| \left(\sum_{j=1}^{n} f(t_j)\,\Delta x_j \right) - \left(\sum_{j=1}^{n} f(z_j)\,\Delta x_j \right) \right| \geq B_2 - B_1 = \varepsilon_0$. Since this holds for *every* partition of $[a,b]$, it follows that Statement (iv) cannot hold.

Remarks (1) The description of integrability given by Statement (iv) in the preceding theorem is Riemann's original formulation of the concept; for that reason in *This Textbook* we say that a function which satisfies this statement is **integrable in the sense of Riemann**; or, more briefly, it is **Riemann integrable**.

(2) It is easy to see that if Statement (iv) holds for f on $[a,b]$, then the number B mentioned there equals $\int_a^b f$; in particular, B is unique. In particular, Darboux's approach yields exactly the same class of integrable functions as does Riemann's, and exactly the same value for the integral. The usual custom in analysis texts – including *This Textbook* – is to refer to the final product by the name 'Riemann integral', even when the tools used in its analysis are based on the approach of Darboux.

(3) Statements (ii) and (iii) in the preceding theorem are analogs of the Cauchy Criterion for the convergence of sequences of numbers. More precisely, they can be used to prove the existence of a certain number B without requiring in advance a particular candidate for this number.

(4) There are several other approaches to integration which are more general than that of Riemann/Darboux and lead to different classes of 'integrable' funnctions. These approaches are associated with names such as Lebesgue, Stieltjes and Henstock, and are not studied here.

VII.2 Basic Properties of the Riemann Integral

This section considers some of the basic properties of the Riemann integral. The following lemma forms the connecting link with the preceding section.

VII.2.1 Lemma

Suppose that $f : [a, b] \to \mathbb{R}$ is a bounded function on $[a, b]$, and that $a < c < b$. Then a number B is a Darboux number for f on $[a, b]$ if, and only if, there exist Darboux numbers C and D for f on $[a, c]$ and $[c, b]$, repectively, such that $B = C + D$.

Proof Suppose first that C and D are Darboux numbers for f on $[a, c]$ and $[c, b]$, respectively. Then by Part (b) of Corollary (VII.1.8) one has

$$\sup L_{(f;\mathcal{Q})} \leq C \leq \inf U_{(f;\mathcal{Q})}$$

and

$$\sup L_{(f;\mathcal{R})} \leq C \leq \inf U_{(f;\mathcal{R})},$$

where $\mathcal{Q}$ and $\mathcal{R}$ are partitions of $[a, c]$ and $[c, b]$, respectively. By Part (e) of Lemma (VII.1.3), to determine the Darboux numbers for f on $[a, b]$ it suffices to restrict attention to partitions $\mathcal{P}$ of $[a, b]$ such that $c \in \mathcal{P}$; that is, to refinements $\mathcal{P}$ of the special partition $\mathcal{P}_0 = \{a, c, b\}$. For such $\mathcal{P}$ one can write $\mathcal{P} = \mathcal{Q} \cup \mathcal{R}$, where $\mathcal{Q} = \mathcal{P} \cap [a, c]$ is a partition of $[a, c]$ and $\mathcal{R} = \mathcal{P} \cap [c, b]$ is a partition of $[c, b]$. By Part (c) of Lemma (VII.1.3) one has

$$L(f; \mathcal{P}) = L(f; \mathcal{Q}) + L(f; \mathcal{R}) \leq C + D \leq U(f; \mathcal{Q}) + U(f; \mathcal{R}) = U(f; \mathcal{P}).$$

It follows that $B = C + D$ is a Darboux number for f on $[a, b]$, as claimed.

As for the converse, let $\mathcal{Q}$ and $\mathcal{R}$ be partitions of $[a, c]$ and $[c, b]$, respectively, and let $\mathcal{P} = \mathcal{Q} \cup \mathcal{R}$ be the corresponding partition of $[a, b]$ containing c. It is easy to see that if $\mathcal{P}_0 = \{a, c, b\}$ as above, while $\mathcal{Q}_0 = \{a, c\}$ and $\mathcal{R}_0 = \{c, b\}$ are the corresponding trivial partitions of $[a, c]$ and $[c, b]$, respectively, then

$$\sup L_{(f;\mathcal{P}_0)} = \sup L_{(f;\mathcal{Q}_0)} + \sup L_{(f;\mathcal{R}_0)}.$$

Likewise, one has

$$\inf U_{(f;\mathcal{P}_0)} = \inf U_{(f;\mathcal{Q}_0)} + \inf U_{(f;\mathcal{R}_0)}.$$

In other words, if B_1 is the lowest Darboux number for f on $[a, b]$, then one has $B_1 = C_1 + D_1$, where C_1 and D_1 are the lowest Darboux numbers for f on $[a, c]$ and $[c, b]$, respectively. The analogous equation $B_2 = C_2 + D_2$ holds for the corresponding highest Darboux numbers. Finally, note that if B is any Darboux number for f on $[a, b]$, then one can write

$$B = t\,B_1 + (1 - t)\,B_2 \text{ for some number } t \text{ in } [0, 1].$$

One then gets $B = C + D$ where

$$C = t\,C_1 + (1-t)\,C_2 \text{ and } D = t\,D_1 + (1-t)\,D_2$$

are the desired Darboux numbers of f on $[a,c]$ and $[c,b]$, respectively.

VII.2.2 Theorem

(a) Suppose that $f : [a,b] \to \mathbb{R}$ can be expressed in the form $f = A{\cdot}g$ for some constant A, where $g : [a,b] \to \mathbb{R}$ is integrable on $[a,b]$. Then f is also integrable on $[a,b]$, and one has

$$\int_a^b f = A{\cdot}\int_a^b g.$$

(b) Suppose that $f : [a,b] \to \mathbb{R}$ can be expressed in the form $f = f_1 + f_2$, where the functions $f_1 : [a,b] \to \mathbb{R}$ and $f_2 : [a,b] \to \mathbb{R}$ are both integrable on $[a,b]$. Then the function f is also integrable on $[a,b]$, and one has

$$\int_a^b f = \left(\int_a^b f_1\right) + \left(\int_a^b f_2\right).$$

More generally, suppose that f can be expressed on $[a,b]$ in the form

$$f = A_1{\cdot}g_1 + \ldots + A_n{\cdot}g_n,$$

where $A_1, \ldots A_n$ are real constants and $g_1, \ldots g_n$ are integrable on $[a,b]$. Then f is also integrable on $[a,b]$, and one has

$$\int_a^b f = A_1 \int_a^b g_1 + \ldots + \int_a^b g_n.$$

(c) Let c be any number such that $a < c < b$. Then $f : [a,b] \to \mathbb{R}$ is integrable on $[a,b]$ if, and only if, it is integrable on each of the intervals $[a,c]$ and $[c,b]$. Furthermore, if this happens, then one has

$$\int_a^b f = \left(\int_a^c f\right) + \left(\int_c^b f\right).$$

(d) Suppose that $f : [a,b] \to \mathbb{R}$ satisfies the condition that $f(x) = h(x)$ for all but finitely many values of x in $[a,b]$, where $h : [a,b] \to \mathbb{R}$ is integrable on $[a,b]$. Then f is also integrable on $[a,b]$, and one has $\displaystyle\int_a^b f = \int_a^b h$.

Proof (a) The simple proof breaks into the separate cases $A > 0$, $A < 0$ and $A = 0$, and is left as an exercise.

(b) For convenience, set $B_1 = \displaystyle\int_a^b f_1$ and $B_2 = \displaystyle\int_a^b f_2$. Then by the definition of 'integral', the numbers B_1 and B_2 are the only ones which satisfy

$$L(f_1;\mathcal{P}) \le B_1 \le U(f_1;\mathcal{P}) \text{ and } L(f_2;\mathcal{Q}) \le B_2 \le U(f_2;\mathcal{Q}) \text{ for all partitions } \mathcal{P} \text{ and } \mathcal{Q} \text{ of } [a,b].$$

By summing these inequalities one gets, for $\mathcal{R} = \mathcal{P} \cup \mathcal{Q}$, that if B is any Darboux number for f, then

$$L(f_1;\mathcal{P}) + L(f_2;\mathcal{Q}) \leq L(f_1;\mathcal{R}) + L(f_2;\mathcal{R}) \ = \ L(f_3;\mathcal{R}) \leq$$
$$B \leq U(f;\mathcal{R}) \leq U(f_1;\mathcal{R}) + U(f_2;\mathcal{R}) \leq U(f_1;\mathcal{P}) + U(f_2;\mathcal{Q}).$$

In particular, one has

$$L(f_1;\mathcal{P}) + L(f_2;\mathcal{Q}) \leq B \text{ for all partitions } \mathcal{P} \text{ and } \mathcal{Q} \text{ of } [a,b].$$

Keeping $\mathcal{P}$ fixed for the moment, take the supremum over $\mathcal{Q}$ and use the fact that B_2 is the supremum of the numbers of the form $L(f_2;\mathcal{Q})$ to get

$$L(f_1;\mathcal{P}) + B_2 \leq B \text{ for every partition } \mathcal{P} \text{ of } [a,b];$$

that is,

$$L(f_1;\mathcal{P}) \leq B - B_2 \text{ for every partition } \mathcal{P} \text{ of } [a,b].$$

In a similar manner one gets

$$B - B_2 \leq U(f_1;\mathcal{P}) \text{ for every partition } \mathcal{P} \text{ of } [a,b].$$

These facts, when combined with the fact that B_1 is the unique Darboux number for f_1 on $[a,b]$, implies that $B - B_2 \ = \ B_1$. In other words, there is precisely one Darboux number for $f \ = \ f_1 + f_2$ on $[a,b]$, namely $B \ = \ B_1 + B_2$. The desired result follows.

The corresponding statement about the function $f \ = \ A_1{\cdot}g_1 + \ \ldots A_n{\cdot}g_n$ follows easily using mathematical induction on n.

(c) This follows easily from the preceding lemma.

(d) The simple proof is left as an exercise.

The next result also uses the integrability of given functions to prove the integrability of a function $f : [a,b] \to \mathbb{R}$ formed from them. In contrast with the preceding theorem, however, it does not provide the precise value of $\displaystyle\int_a^b f$ from the values of the other integrals.

VII.2.3 Theorem

(a) Suppose that $f : [a,b] \to \mathbb{R}$ can be expressed as a product $f \ = \ f_1{\cdot}f_2$, where $f_1 : [a,b] \to \mathbb{R}$ and $f_2; [a,] \to \mathbb{R}$ are integrable on $[a,b]$. Then f is also integrable on $[a,b]$.

More generally, if f can be expressed as the product $f \ = \ f_1{\cdot}f_2{\cdot}\ldots{\cdot}f_k$ of finitely many functions which are integrable on $[a,b]$, then f is also integrable on $[a,b]$.

(b) Suppose that $f : [a,b] \to \mathbb{R}$ can be expressed as a quotient $f \ = \ f_1/f_2$, where $f_1 : [a,b] \to \mathbb{R}$ and $f_2 : [a,b] \to \mathbb{R}$ are integrable on $[a,b]$. Suppose further that there exists a constant $c > 0$ such that $|f_2(x)| \geq c$ for all x in $[a,b]$. Then f is also integrable on $[a,b]$.

(c) Suppose that $f : [a,b] \to \mathbb{R}$ can be expressed in the form $f \ = \ |g|$ for some function $g : [a,b] \to \mathbb{R}$ such that g is integrable on $[a,b]$. Then f is also integrable on $[a,b]$.

Proof We shall give direct proofs of these results later, based on an important theoretical characterization of integrability to be developed in the next section. Direct simple proofs are also outlined in the exercises at the end of this chapter. ■

Remarks (1) The usual phrasing of Part (a) of the preceding theorem is this:

‘The product of finitely many functions that are integrable on $[a, b]$ is also integrable on $[a,]$.’ Similar comments about the usual phrasings of Part (b) of this theorem and Parts (a) and (b) of Theorem (VII.2.2).

In real-life mathematics, however, one does not normally start with a pair of functions f_1 and f_2 and then form their product $f = f_1{\cdot}f_2$ (or their sum or quotient). Instead, one starts with a function f of interest and seeks out a way to express it in terms of simpler functions using operations such as addition, multiplication, division and so on. The phrasings of such results in *This Textbook* attempt to reflect that more realistic usage.

(2) The converses of various parts of Theorems (VII.2.2) and (**??**) are not true. For example, if $f : [a, b] \rightarrow \mathbb{R}$ is already known to be integrable on $[a, b]$, then knowing that $f = f_1 + f_2$ or $f = f_1{\cdot}f_2$ or $f = f_1/f_2$ on $[a, b]$ does *not* imply that f_1 and f_2 are also integrable on $[a, b]$. Likewise, knowing that $|f|$ is integrable on $[a, b]$ does not imply that f is integrable on $[a, b]$. It is a simple exercise to find appropriate counterexamples.

There are simple inequalities associated with the Riemann integral which are important for both practical and theoretical discussions.

VII.2.4 Theorem

Suppose that f and g are Riemann integrable on the interval $[a, b]$, and that $f(x) \leq g(x)$ for all x in $[a, b]$. Then

$$\int_a^b f \leq \int_a^b g.$$

Furthermore, if in addition both f and g are continuous on $[a, b]$, then one gets the case of equality in the preceding inequality if, and only if, $f(x) = g(x)$ for all x in $[a, b]$.

ProofDefine $h : [a, b] \rightarrow \mathbb{R}$ by the rule $h(x) = g(x) - f(x)$ for all x in $[a, b]$. By Parts (a) and (b) of Theorem (VII.2.2) it follows that h is also integrable on $[a, b]$ and that $\displaystyle\int_a^b h = \int_a^b g - \int_a^b f$. It is clear from the ‘inequality’ hypothesis on f and g that $h(x) \geq 0$ for every x in $[a, b]$, which implies that $L(h; \mathcal{P}) \geq 0$ for every partition $\mathcal{P}$ of $[a, b]$. Since $\displaystyle\int_a^b h$ is the supremum of Darboux sums of the form $L(h'\mathcal{P})$, it follows that $\int_a^b h \geq 0$, which in turn implies that $\displaystyle\int_a^b g > \int_a^b f$, as required. The statement about ‘equality’ when f and g are continuous on $[a, b]$ is left as an exercise. ■

VII.3 More on Antiderivatives and the Riemann Integral

There are important relations between the Riemann integral and the concept of antiderivatives. This was illustrated in Example (VII.1.7), where it was it was pointed out that if $f : [a, b] \rightarrow \mathbb{R}$ is continuous on $[a, b]$, and F is any antiderivative of f on $[a,]$, then f is Riemann integrable on $[a, b]$ and $\displaystyle\int_a^b f = F(b) - F(a)$. In this section we explore these relations more completely.

The first step is to introduce a slight extension of the integral notation. More precisely, up to now the expression $\int_a^b f$ assumes that $[a,b]$ is an interval in $\mathbb{R}$; in particular, it requires that the number at the bottom of the integral sign be strictly smaller than the one at the top. However, even in elementary calculus one finds it useful to allow $a = b$ or $a > b$ in such expressions. The motivation for the standard choice for this extension is the observation that the right side of the preceding equation *does* make sense independent of the relation between a and b. More precisely, one has

$$F(a) - F(b) \,=\, -(F(b) - F(a)) \text{ and } F(c) - F(c) \,=\, 0 \text{ for every number } c \text{ in } [a,b].$$

Thus it makes sense to extend the meaning of the integral as follows: If f is integrable on an interval $[a,b]$, then

$$\int_b^a f \,=\, -\left(\int_a^b f\right) \text{ and } \int c^c f \,=\, 0 \text{ for every } c \text{ in } [a,b].$$

In *This Textbook* we refer to this as the **extended integral notation**IndBintegralsextended integral notation.

Using this extended notation, one can reformulate Part (c) of Theorem (VII.2.2) as follows:

VII.3.1 Theorem

Suppose that $f : I \to \mathbb{R}$ is Riemann integrable on a closed bounded interval I. Then for each triple of numbers a, b and c in the interval I, regardless of the relative order of these numbers, one has

$$\int_a^b f \,=\, \int_a^c f + \int_c^b; \text{ equivalently, } \int_a^c f + + \int_b^a f \,=\, 0.$$

The simple proof is left as an exercise.

Example Suppose that $f : [a,b] \to \mathbb{R}$ is continuous, and that $u, v : J \to [a,b]$ are differentiable functions on an open interval J with values in $[a,b]$. Define $G : J \to \mathbb{R}$ by the rule

$$G(t) \,=\, \int_{u(t)}^{v(t)} f \text{ for all } t \text{ in } J;$$

in light of the extended integral notation, this makes sense no matter the relative order of the numbers $u(t)$ and $v(t)$. It is easy to see that G is differentiable on J, and

$$G'(t) \,=\, f(v(t)){\cdot}v'(t) - f(u(t)){\cdot}u'(t) \text{ for all } t \text{ in } J.$$

Indeed, the continuity hypothesis implies that f has an antiderivative F on $[a,b]$, and the Cauchy Fundamental Theorem of Calculus implies that

$$G(t) \,=\, F(v(t)) - F(u(t)).$$

The desired result follows from the Chain Rule for Derivatives.

The proof of Cauchy's Antiderivative Theorem given in the preceding chapter differs from Cauchy's original proof. The next result carries out his approach in a slightly more general manner.

VII.3.2 Theorem

Suppose that $f : [a, b] \to \mathbb{R}$ is Riemann integrable on an interval $[a, b]$. Let c be any number in $[a, b]$, and define a new function $F : [a, b] \to \mathbb{R}$ by the rule

$$F(x) = \int_c^x f \text{ for each } x \text{ in } [a, b];$$

note that once again the extended integral notation is in use. Then:

(a) The function F is continuous on $[a, b]$.

(b) If x_0 is a point of $[a, b]$ at which f is continuous, then F is differentiable at x_0. and $F'(x_0) = f(x_0)$. (If x_0 is one of the endpoints a or b, then the appropriate one-sided notions of continuity and differentiability are understood to apply.)

Proof

Note that if x_1 and x_2 are in $[a, b]$ with $x_1 \leq x_2$, then one has

$$F(x_2) - F(x_1) = \int_a^{x_2} f - \int_a^{x_1} f = \int_a^{x_2} + \int_{x_1}^{a} f = \int_{x_1}^{x_2} f;$$

see Remark (**??**). If M is an upper bound for $|f|$ on $[a, b]$, it then follows that

$$|F(x_2) - F(x_1)| \leq M\left(\alpha(x_2) - \alpha(x_1)\right).$$

Parts (a) and (b) now follow easily.

To get a more precise estimate, suppose that $a < x_1 < x_2 < b$, and let $m(x_2) = \inf\{f(x) : x_1 \leq x \leq x_2\}$ and $M(x_2) = \sup\{f(x) : x_1 \leq x \leq x_2\}$. Then one has

$$m(x_2)(\alpha(x_2) - \alpha(x_1)) \leq F(x_2) - F(x_1) \leq M(x_2)(\alpha(x_2) - \alpha(x_1)).$$

Divide all the terms in this inequality by the positive number $x_2 - x_1$ to get

$$m(x_2)\left(\frac{\alpha(x_2) - \alpha(x_1)}{x_2 - x_1}\right) \leq \frac{F(x_2) - F(x_1)}{x_2 - x_1} \leq M(x_2)\left(\frac{\alpha(x_2) - \alpha(x_1)}{x_2 - x_1}\right)$$

The hypothesis that f is continuous at x_1, which implies that $\lim_{x_2 \to x_1} m(x_2) = \lim_{x_2 \to x_1} M(x_2) = f(x_1)$, combined with the hypothesis that α is differentiable at x_1, implies (by the 'Squeeze Theorem') that

$$\lim_{x_2 \searrow x_1} \frac{F(x_2) - F(x_1)}{x_2 - x_1} = f(x_1)\alpha'(x_1).$$

A similar argument shows that the left-hand derivative of F at x_1 exists and equals $f(x_1)\alpha'(x_1)$. The desired result follows.

VII.3.3 Corollary

Suppose that $f : [a, b] \to \mathbb{R}$ is Riemann integrable on $[a, b]$. Let $F : [a, b] \to \mathbb{R}$ be defined by $F(x) = \int_a^x f(t)\, dt$ for each x in $[a, b]$. Then F is continuous on $[a, b]$, and if f is continuous at x, then $F'(x)$ exists and one has $F'(x) = f(x)$.

Proof This follows immediately from the previous theorem in the case $\alpha(x) = x$ for all x.

Remark When f is continuous on $[a,b]$, the preceding corollary reduces essentially to Theorem (V.8.1), 'Cauchy's Antiderivative Theorem'.

In the preceding one uses the definite integral to construct an antiderivative of a continuous function. The next result reverses the relationship between antiderivatives and integrals.

VII.3.4 Theorem

Suppose that $f:[a,b]\rightarrow \mathbb{R}$ is Riemann integrable on $[a,b]$, and suppose that there exists continuous $F:[a,b]\rightarrow \mathbb{R}$ such that $F'(x) = f(x)$ for all x in $[a,b]$. Then $\displaystyle\int_a^b f(x)\,dx = F(b)-F(a)$.

Proof Let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_k = b\}$ be a partition of $[a,b]$, and consider the telescoping sum

$$F(b)-F(a) = \sum_{j=1}^{k}\left(F(x_j)-F(x_{j-1})\right).$$

By the Mean-Value Theorem, for each j there exists t_j in (x_{j-1},x_j) such that

$$F(x_j)-F(x_{j-1}) = F'(t_j)\Delta x_j = f(t_j)\Delta x_j.$$

It follows that $F(b)-F(a)$ equals the Riemann sum $\sum_{j=1}^{k} f(t_j)\Delta x_j$, and thus one has

$$L(\mathcal{P},f)\leq F(b)-F(a)\leq U(\mathcal{P},f).$$

However, by hypothesis f is Riemann integrable on $[a,b]$, so there exists only one number A such that $L(\mathcal{P},f)\leq A\leq U(\mathcal{P},f)$, namely $A = \int_a^b f(x)\,dx$. The desired result now follows.

Remark Some authors refer to Theorem (VII.3.2) (or Corollary (VII.3.3)) as the 'First Fundamental Theorem of Calculus', and to Theorem (VII.3.4) as the 'Second Fundamental Theorem of Calculus'. Other authors number these results in the opposite way. And yet other authors combine the two as individual parts of a single 'Fundamental Theorem'. In *This Textbook* we shall refer to each as 'The Fundamental Theorem of Calculus', and let the context make it clear to the reader which version is being used, or simply give the number of the theorem.

VII.3.5 Corollary

Suppose that $f:[a,b]\rightarrow \mathbb{R}$ is continuous and $\alpha:[a,b]\rightarrow \mathbb{R}$ is monotonic up. Assume further that α' is defined and Riemann integrable on $[a,b]$. Then

$$\int_a^b f\,d\alpha = \int_a^b f(x)\alpha'(x)\,dx. \qquad \text{(VII.1)}$$

Proof Since, by hypothesis, f is continuous on $[a,b]$ and $\alpha'\in\mathcal{R}_{[a,b]}$, it follows from Theorem (**??**) that $f\in\mathcal{R}_{[a,b]}(\alpha)$, and from Theorem (**??**) that $f\cdot\alpha'\in\mathcal{R}_{[a,b]}$. Furthermore, it follows from Theorem (VII.3.2) that the function $F:[a,b]\rightarrow \mathbb{R}$ given by the rule $F(x) = \displaystyle\int_a^x f\,d\alpha$ is continuous on $[a,b]$ and differentiable on (a,b), and that $F'(x) = f(x)\alpha'(x)$. Then Theorem (VII.3.4) allows one to conclude that $F(b)-F(a) = \displaystyle\int_a^b f(x)\alpha'(x)\,dx$. The desired result now follows by noting that $F(b)-F(a) = \displaystyle\int_a^b f\,d\alpha$.

Remark With a little more work one can prove the following more general result.

VII.3.6 Theorem

Suppose that $f : [a,b] \rightarrow \mathbb{R}$ is in $\mathcal{R}_{[a,b]}(\alpha)$ and $\alpha : [a,b] \rightarrow \mathbb{R}$ is monotonic up. Assume further that α' is defined and Riemann integrable on $[a,b]$. Then $f\alpha'$ is Riemann integrable on $[a,b]$, and Equation (VII.1) holds.

See Page 131 of Rudin's 'Principles of Mathematical Analysis' (3rd edition) for a proof.

VII.3.7 Remark

The left side of Equation (VII.1) involves the Riemann-Stieltjes integral $\int_a^b f\,d\alpha$, in which the integrator α is assumed to be monotonic up on $[a,b]$. This restriction on α was imposed by the initial motivation of the integral in terms of weighted averages; but it also facilitated the technical development of the theory. For instance, the Darboux approach makes repeated use of the fact that $\Delta\alpha_j \geq 0$.

The restriction of α to be monotonic up implies that the factor α', which appears in the Riemann integral $\int_a^b f(x)\alpha'(x)\,dx$ on the right side of Equation (VII.1), must be nonnegative. However, the integral itself makes perfectly good sense even if α' changes sign in the interval $[a,b]$. This suggests that it might be useful to extend the concept of the Riemann-Stieltjes integral to allow integrators α which are not of fixed monotonicity throughout the interval $[a,b]$. Such extensions have been developed – indeed, it appears that even Stieltjes allowed such extentions. An excellent source for such an extended treatement of the Riemann-Stieltjes integral can be found in Apostol's 'Mathematical Analysis (2nd edition)'. Instead of defining integrabilty in terms of Riemann's Condition, which uses the hypothesis that $\Delta\alpha_j \geq 0$, Apostol uses the Riemann sum approach; see Theorem (**??**). This approach allows quite general integrators. However, one soon restricts to the case in which α is of bounded variation on $[a,b]$, since in this case the basic theorem, that every continuous function on $[a,b]$ is in $\mathcal{R}_{[a,b]}(\alpha)$, remains valid.

VII.4 Miscellaneous Results on the Riemann Integral

There are numerous results for the Riemann integral which are worth singling out.

VII.4.1 Theorem (Integration-by-Parts)

Suppose that f and g are differentiable on $[a,b]$ and that their derivatives are Riemann integrable on $[a,b]$. Then fg' and $f'g$ are both integrable on $[a,b]$, and one has

$$\int_a^b f(x)g'(x)\,dx \;=\; f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x)\,dx. \tag{VII.2}$$

Proof Since f and g are differentiable on $[a,b]$, they are certainly continuous, hence Riemann integrable, on $[a,b]$; and since, by hypothesis, f' and g' are Riemann integrable on $[a,b]$, it follows that the products $f'g$ and fg' are also in $\mathcal{R}_{[a,b]}$, and thus so is their sum $f'g + fg'$. However, by the

Product Rule for Derivatives, this last function is the derivative on $[a,b]$ of the function $H = fg$. Now Theorem (VII.3.4) implies that

$$H(b) - H(a) = \int_a^b f'(x)g(x)\,dx + \int_a^b f(x)g'(x)\,dx$$

Transpose the term $\int_a^b f'(x)g(x)\,dx$ in this last equation, and note that $H(b) - H(a) = f(b)g(b) - f(a)g(a)$, to get the desired Equation (VII.2).

VII.4.2 Theorem (First Mean-Value Theorem for Riemann Integrals)

(a) Suppose that $f{\in}\mathcal{R}_{[a,b]}$, and let $m^* = \inf\{f(x) : x{\in}[a,b]\}$ and $M^* = \sup\{f(x) : x{\in}[a,b]\}$. Suppose that $g : [a,b] \to \mathbb{R}$ is a nonnegative Riemann integrable function on $[a,b]$. Then there exists a number μ, with $m^* \leq \mu \leq M^*$, such that

$$\int_a^b f(x)g(x)\,dx = \mu \int_a^b g(x)\,dx$$

If, in addition, f is continuous on $[a,b]$, then μ can be chosen to be of the form $f(c)$ for some c in $[a,b]$.

(b) If the function f in Part (a) is continuous on $[a,b]$, and if $g(x) = 1$ for all x in $[a,b]$, then one has $\frac{1}{b-a}\int_a^b f(x)\,dx = f(c)$ for some number c in $[a,b]$.

Proof (a) Because g is nonnegative, it is clear that $m^*g(x) \leq f(x)g(x) \leq M^*g(x)$ for all x in $[a,b]$. Then from Part (c) of Theorem (**??**) one sees that

$$m^* \int_a^b g(x)\,dx \leq \int_a^b f(x)g(x)\,dx \leq M^* \int_a^b g(x)\,dx.$$

If the integral $\int_a^b g(x)\,dx$, which appears on either end of this string of inequalities, equals 0, then clearly one also has $\int_a^b f(x)g(x)\,dx = 0$, so μ can be any number such that $m^* \leq \mu \leq M^*$. If, instead, that integral is *not* zero, then choose μ by the rule

$$\mu = \frac{\int_a^b f(x)g(x)\,dx}{\int_a^b g(x)\,dx}.$$

It is clear that $m^* \leq \mu \leq M^*$.

If f is also continuous on $[a,b]$, then the Intermediate-Value Theorem for Continuous Functions implies that $\mu = f(c)$ for some c in $[a,b]$.

(b) This follows easily from Part (a).

VII.4.3 Theorem (Second Mean-Value Theorem for Riemann Integrals)

Suppose that $f : [a,b] \rightarrow \mathbb{R}$ is a monotonic-up function that is differentiable on $[a,b]$ and for which $f' \in \mathcal{R}_{[a,b]}$, and that g is continuous on $[a,b]$. Then there exists c in $[a,b]$ such that

$$\int_a^b f(x)g(x)\,dx \;=\; f(a)\int_a^c g(x)\,dx + f(b)\int_c^b g(x)\,dx$$

Proof Define $G : [a,b] \rightarrow \mathbb{R}$ by the rule $G(x) \;=\; \displaystyle\int_a^x g(t)\,dt$. Note that G is differentiable because of the hypothesis that g is continuous, and of course $G' \;=\; g$ is Riemann integrable on $[a,b]$. Then Theorem (VII.4.1), 'Integration-by-Parts', can be apllied to the functions f and G to yield

$$\int_a^b f(x)g(x)\,dx \;=\; \int_a^b f(x)G'(x)\,dx \;=\; (f(b)G(b) - f(a)G(a)) - \int_a^b G(x)f'(x)\,dx \qquad (*)$$

Now apply Part (a) of the preceding theorem, with G and f' here playing the roles of f and g, respectively, in that earlier theorem. Then, since G is certainly continuous on $[a,b]$, one gets

$$\int_a^b G(x)f'(x)\,dx \;=\; G(c)\int_a^b f'(x)\,dx \;=\; G(c)(f(b)-f(a)) \;=\; \left(\int_a^c g(x)\,dx\right)(f(b)-f(a))$$

for some c in $[a,b]$. Combine this last result with Equation $(*)$ to get

$$\int_a^b f(x)g(x)\,dx \;=\; (f(b)G(b) - f(a)G(a)) - G(c)(f(b)-f(a)) \qquad (**)$$

By the definition of G one has $G(a) \;=\; 0$, $G(b) \;=\; \displaystyle\int_a^b g(x)\,dx$, and $G(c) \;=\; \displaystyle\int_a^c g(x)\,dx$. After doing the obvious simplification, including noting that

$$f(b)G(b) - f(b)G(b) \;=\; f(b)\left(\int_a^b g(x)\,dx - \int_a^c g(x)\,dx\right) \;=\; f(b)\int_c^b g(x)\,dx,$$

the desired result follows.

VII.4.4 Theorem (Change-of-Variables Theorem for Riemann Integrals)

Suppose that $f : [a,b] \rightarrow \mathbb{R}$ is continuous on an interval $[a,b]$, and that $g : [c,d] \rightarrow \mathbb{R}$ has continuous first derivative on an interval $[c,d]$. Suppose further that $a \leq g(t) \leq b$ for all t in $[c,d]$. Then

$$\int_{g(c)}^{g(d)} f(x)\,dx \;=\; \int_c^d f((g(t)))g'(t)\,dt \qquad \text{(VII.3)}$$

Proof Let F be an antiderivative of f on $[a,b]$; such F exists because f is continuous. Likewise, let $H \;=\; F{\circ}g : [c,d] \rightarrow \mathbb{R}$; the composition makes sense because g maps $[c,d]$ to a subset of $[a,b]$. Then H is the composition of differentiable functions, so the Chain Rule can be used to say that H is differentiable on $[c,d]$, and that $H'(t) \;=\; F'(g(t)){\cdot}g'(t) \;=\; f(g(t)){\cdot}g'(t)$ for all t in $[c,d]$. Since f, g and g' are all continuous, the function $H' \;=\; (f{\circ}g){\cdot}g'$ is certainly continuous on $[c,d]$. It now follows from Theorem (VII.3.4) – applied to $(f{\circ}g){\cdot}g'$ that

$$\int_c^d (f{\circ}g){\cdot}g' \;=\; H(d) - H(c) \;=\; F(g(d)) - F(g(c)).$$

However, since f is continuous on $[a,b]$, one can use Theorem (VII.3.4) again, but this time on f, to say that

$$\int_u^v f(x)\,dx \;=\; F(u) - F(v) \text{ for all } u,\, v \text{ in } [a,b].$$

In particular, one has $F(g(d)) - F(g(c)) \;=\; \displaystyle\int_{g(c)}^{g(d)} f(x)\,dx$. The desired result now follows.

Remarks

(1) Some analysis texts impose the requirement that the function g be strictly increasing on $[a,d]$, so that g is a bijection of $[c,d]$ onto $[a,b]$. Clearly this restriction is not needed; however, it does make the name 'change of variables' seem more appropriate.

(2) Note that there is no restriction that $g(c) \;<\; g(d)$. Indeed, we even allow the possibility that $g(c) \;=\; g(d)$. If this occurs, Equation (VII.3) takes the particularly simple form

$$\int_{g(c)}^{g(c)} f(x)\,dx \;=\; \int_c^d f((g(t)))g'(t)\,dt;$$

that is,

$$\int_c^d f((g(t)))g'(t)\,dt \;=\; 0.$$

Note that this holds independently of the choice of function f. For instance, suppose that $f(x) \;=\; e^{-x^2}$ and $g(t) \;=\; \sin t$ for $0 \leq t \leq 2\pi$. Then, when read backwards, Equation (VII.3) takes the form

$$\int_0^{2\pi} e^{-\sin^2 t} \cos t\,dt \;=\; \int_0^0 e^{-x^2}\,dx.$$

One cannot write down an antiderivative for the integrand $f(x) \;=\; e^{-x^2}$ which appears on the right side of this last equation, but one can still see that the value of the integral on the right – and thus the value of the integral on the left – must equal 0.

VII.5 Existence Theorems for the Riemann Integral

In his original treatment of the integral, Riemann carries out a more detailed analysis of the quantity $\Delta(f;\mathcal{P})$ in Statement (iii); see [RIEMANN 1854]. Several decades later Henri Lebesgue greatly improved this analysis; see [LEBESGUE 1901]. The approach followed here, however, focuses on a somewhat simpler approach arising from the work of Camille Jordan; see [JORDAN ????]. The next definition clarifies the ideas involved.

VII.5.1 Definition

Let $f : X \to \mathbb{R}$ be a function such that f bounded on the (nonempty) set X. Then the **oscillation of *f* over the set *X*** is the number $\Omega(f;X)$ given by the formula

$$\Omega\,(f;X) \;=\; \sup\{|f(x_2) - f(x_1)| : x_1, x_2 {\in} X\}.$$

As usual, the 'function diagram' notation $f : X \to \mathbb{R}$ allows the possibility that the full domain of f includes points outside the set X.

VII.5.2 Remarks

(1) Some authors define $\Omega(f;X)$ by the formula

$$\Omega\,(f;X) \;=\; \sup\{f(x_2)-f(x_1) : x_1,x_2{\in}X\};$$

that is, they omit absolute-value signs found in the definition above. Likewise, some authors define this expression by

$$\Omega\,(f;X) \;=\; \sup\{f(x) : x{\in}X\} - \inf\{f(x) : x{\in}X\}.$$

It is a simple exercise to show that these definitions all yield the same value for $\Omega(f;X)$. In *This Textbook* we freely use all these variations. It is also easy to show that if Y is a nonempty subset of X, then $0 \leq \Omega(f;Y) \leq \Omega(f;X)$.

(2) It is clear that the quantities m_j and M_j that appear in Definition (VII.1.2) are related by the rule

$$M_j - m_j \;=\; \Omega\,(f;[x_{j-1},x_j]).$$

(3) The symbol Ω is the upper-case version of the Greek letter 'omega'; the lower-case version of this letter is ω. This Greek letter corresponds closely to the English letter 'O'; in the present context it is used to remind one of the first letter of the word 'oscillation'.

There is a close relationship between the concept of 'oscillation over a set' and 'continuity'. For simplicity, we consider here only sets which are closed intervals.

Thus, suppose, as usual, that $f:[a,b]\to\mathbb{R}$ is bounded on $[a,b]$, and let c be a point of $[a,b]$. It follows easily from Theorem (III.4.7) that a necessary and sufficient condition for f to <u>not</u> be continuous at c is this:

'There exists a sequence $(x_1,x_2,\ldots x_n,\ldots)$ in $[a,b]$, converging to c, such that the corresponding sequence $(f(x_1),f(x_2),\ldots f(x_n),\ldots)$ of values converges to some number $L \neq f(c)$.' (The possibility $L \;=\; \pm\infty$ is excluded by the boundedness hypothesis on f.)

Of course if c equals one of the endpoints a or b, then 'continuity' in the present context means the appropriate one-sided continuity.

It follows that a reasonable indicator of the degree to which f fails to be continuous at c is the width of the set $S_{(f;c)}$ consisting of all numbers L which can be expressed as such a limit. Note that this set is bounded, because the function f is, and it is nonempty, because it contains the number $f(c)$. A commonly used measure of the size of $S_{(f;c)}$ is the number $\omega_f(c) \;=\; \sup S_{(f;c)} - \inf S_{(f;c)}$, called the **oscillation of f at** c; this is often abbreviated to $\omega(c)$ if the function $f:[a,b]\to\mathbb{R}$ is understood from the context.

It is clear that for each c in $[a,b]$ one has $\omega(c) \geq 0$, and that $\omega(c) \;=\; 0$ if, and only if, f is continuous at c. Otherwise stated, the set $D_{(f;[a,b])}$ of all the discontinuities of f in $[a,b]$ is precisely the set of c such that $\omega(c) \;>\; 0$. For each k in $\mathbb{N}$ let D_k denote the set of points c in $[a,b]$ such that $\omega(c) \geq 1/k$.

VII.5.3 Definition

Let S be a subset of a closed interval $[a,b]$ in $\mathbb{R}$. One says that S is a **null set in the sense of Jordan**, or, briefly, the set S is **Jordan null**, provided that the following condition holds:

For every $\varepsilon \;>\; 0$ there exists a partition $\mathcal{P} \;=\; \{a \;=\; x_0 \;<\; x_1 \;<\; \ldots \;<\; x_{n-1} \;<\; x_n \;=\; b\}$ of $[a,b]$ such that the sum $\Sigma(S;\mathcal{P}) \;<\; \varepsilon$, where $\Sigma(S;\mathcal{P})$ denotes the sum of the lengths of those subintervals $[x_{j-1}, x_j]$ of $\mathcal{P}$ having nonempty intersection with S. If $S = \emptyset$, this sum is defined to equal 0.

VII.5.4 Remark

The specific choice of interval $[a, b]$ containing a nonempty set S as a subset does not affect whether S is Jordan null. In particular, there is no loss in generality by assuming that $a < \inf S$ and $b > \sup S$. In what follows it is often convenient to make that assumption.

VII.5.5 Lemma

In this Lemma $[a, b]$ is an interval in $\mathbb{R}$.

(a) Suppose that S is a subset of $[a, b]$.

(i) If $\mathcal{P}$ and $\mathcal{Q}$ are partitions of $[a, b]$ such that $\mathcal{Q}$ is a refinement of $\mathcal{P}$, then $\Sigma(S; \mathcal{Q}) \leq \Sigma(S; \mathcal{P})$.

(ii) If T is also a subset of $[a, b]$, then $\Sigma(S \cup T; \mathcal{R}) \leq \Sigma(S; \mathcal{R}) + \Sigma(T; \mathcal{R})$ for every partition $\mathcal{R}$ of $[a, b]$.

(iii) If T is a subset of $[a, b]$ such that $S \subseteq T$, then $\Sigma(S; \mathcal{R}) \leq \Sigma(T; \mathcal{R})$ for every partition $\mathcal{R}$ of $[a, b]$.

(b) Assume now that S is a nonempty subset of $[a, b]$ such that $a < \inf S$ and $b > \sup S$, as in Remark (VII.5.4) above.

(i) Let $\mathcal{Q}$ be any partition of $[a, b]$. Then there exists a partition $\mathcal{P} = \{a = x_0 < \ldots < x_n = b\}$, satisfying $\Sigma(S; \mathcal{P}) = \Sigma(S; \mathcal{Q})$, such that if a subinterval $[x_{k-1}, x_k]$ of $\mathcal{P}$ has nonempty intersection with S, then each subinterval of $\mathcal{P}$ adjacent to $[x_{k-1}, x_k]$ is disjoint from S. In particular, for such $\mathcal{P}$ none of the partition points x_j, $j = 0, 1, \ldots n$ is in S.

(ii) Let $U = (J_1, J_2, \ldots J_N)$ be any finite ordered list of open subintervals of $[a, b]$ such that $S \subseteq \bigcup_{i=1}^{N} J_i$; that is, the intervals in the list U form an open cover of S. Write $J_i = (c_i, d_i)$ for each $i = 1, 2, \ldots N$. Let $\mathcal{P} = \{a, c_1, d_1, \ldots c_N, d_N, b\}$ be the partition of $[a, b]$ formed from the endpoints of these open intervals, together with the numbers a and b. (There is no assumption that these endpoints are in any particular order or even that they are distinct; recall that a partition of $\mathcal{P}$ is simply a finite subset of $\mathcal{P}$ containing both a and b.) Then

$$\Sigma(S; \mathcal{P}) \leq \sum_{i=1}^{N} (d_i - c_i).$$

Remark The conclusions are obviously true if $S = \emptyset$.

Proof (a) The simple proof is left as an exercise.

(b) (i) Among all the partitions $\mathcal{R}$ for which $\Sigma(S; \mathcal{R}) = \Sigma(S; \mathcal{Q})$, there is at least one for which the number of subintervals of $[a, b]$ determined by $\mathcal{R}$ is a minimum. Choose $\mathcal{P}$ to be such a partition. Suppose that there exists an index k such that the adjacent subintervals $[x_{k-1}, x_k]$ and $[x_k, x_{k+1}]$ both contain points of S; note that this implies that $1 \leq k \leq n - 1$. Then the sum forming the quantity $\Sigma(S; \mathcal{P})$ includes the summands $(x_k - x_{k-1})$ and $(x_{k+1} - x_k)$, which together add up to $x_{k+1} - x_{k-1}$. Now let $\mathcal{T}$ be the partition of $[a, b]$ obtained by removing the number x_k from the set $\mathcal{P}$. Then $\Sigma(S; \mathcal{T}) = \Sigma(S; \mathcal{P}) = \Sigma(S; \mathcal{Q})$, but the partition $\mathcal{T}$ has one fewer subintervals than the partition $\mathcal{P}$, contradicting the definition of $\mathcal{P}$. That is, $\mathcal{P}$ has the desired 'non-adjacency' property. In particular, if $1 \leq k \leq n - 1$, then x_k, which belongs to adjacent subintervals, cannot be an element of S. The fact that $a = x_0$ and $b = x_n$ also cannot be in S follows from the hypothesis that $a < \inf S$ and $b > \sup S$.

(ii) The proof here is by mathematical induction on the number N.

Initial Step Suppose that $N = 1$, so that $S \subseteq (c_1, d_1)$. Let $\mathcal{P} = \{a \leq c_1 < d_1 \leq b\}$ be the corresponding partition of $[a, b]$. Then $[c_1, d_1]$ is a subinterval of the partition $\mathcal{P}$ such that $S \subseteq (c_1, d_1) \subseteq [c_1, d_1]$, so that clearly $\Sigma(S; \mathcal{P}) = (d_1 - c_1)$, and thus $\Sigma(S; \mathcal{P}) \leq (d_1 - c_1)$, as required.

Induction Step Suppose that the claim is true for $N = k$. To see that it remains true when $N = k + 1$, let S be a subset of $[a, b]$, let $U = (J_1, J_2, \ldots J_{k+1})$ be an ordered list of intervals $J_i = [c_i, d_i]$ in $[a, b]$ which form an open cover of S. Let $V = (J_1, \ldots J_k)$ be the sublist of U formed from the first k terms in U, and let $T_1 = \bigcup_{i=1}^{k}(S \cap J_i)$; clearly the intervals in the list V form an open cover of T_1. Let $\mathcal{Q}$ be the partition of $[a, b]$ formed, as above, from the endpoints of the intervals J_i, $1 \leq i \leq k$. It follows from the induction hypothesis that

$$\Sigma(T_1; \mathcal{Q}) \leq \sum_{i=1}^{k}(d_i - c_i).$$

Similarly, let $T_2 = S \setminus T_1$. It is clear that for each $i = 1, 2, \ldots k$ one has $T_2 \cap J_i = \emptyset$, and thus $T_2 \subseteq (c_k, d_k)$. Let $\mathcal{R} = \{a, c_{k+1}, d_{k+1}, b\}$. It then follows from the truth of the given result in the case $N = 1$ that

$$\Sigma(T_2, \mathcal{R}) \leq (d_{k+1}, c_{k+1}).$$

Finally, let $\mathcal{P} = \mathcal{Q} \cup \mathcal{R}$. Then it follows from Part (a) of Lemma (VII.5.5), together with the equation $S = T_1 \cup T_2$, that

$$\Sigma(S; \mathcal{P}) \leq \Sigma(T_1; \mathcal{P}) + \Sigma(T_2; \mathcal{P}) \leq \Sigma(T_1; \mathcal{Q}) + \Sigma(T_2; \mathcal{R}) \leq \left(\sum_{i=1}^{k}(d_i - c_i)\right) + (d_{k+1} - c_{k+1}).$$

The desired result now follows. ■

VII.5.6 Corollary

(a) Every subset of a Jordan-null set is also Jordan null.

(b) A necessary and sufficient condition for a bounded subset S of $\mathbb{R}$ to be Jordan null is that its closure $\overline{S}$ in $\mathbb{R}$ be Jordan null.

(c) Let S be a nonempty subset of an interval $[a, b]$. The condition for S to be Jordan null can be weakened slightly to the following: For every $\varepsilon > 0$ there exists a partition $\mathcal{P}$ of $[a, b]$ and a Jordan-null subset T of S, which may depend on $\mathcal{P}$, such that $\Sigma(S \setminus T; \mathcal{P}) < \varepsilon$.

(d) Let S be a nonempty subset of an interval $[a, b]$. Then a necessary and sufficient condition for S to be Jordan null is that for every $\varepsilon > 0$ there exists a finite open cover $U = \{J_1, J_2, \ldots J_N\}$, formed from open intervals $J_i = (c_i, d_i)$, $i = 1, 2, \ldots N$, such that $\sum_{i=1}^{N}(d_i - c_i) < \varepsilon$.

Proof (a) The simple proof is left as an exercise.

(b) It is clear from Part (a) that if the closure $\overline{S}$ is a Jordan null set, then so is S itself. Now suppose, conversely, that S is Jordan null. Without loss of generality assume that S is a subset of an interval $[a.b]$ such that $a < \inf S$ and $b > \sup S$. Let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be a partition of $[a, b]$, and let K be the set of indices k for which $S \cap [x_{k-1}, x_k] \neq \emptyset$. Assume that $\mathcal{P}$ is chosen, as in Part (b) of the preceding lemma, so that if $k{\in}K$, then each subinterval of $\mathcal{P}$ adjacent to the subinterval $[x_{k-1}, x_k]$ is disjoint from S. One certainly has

$$S \;=\; \bigcup\nolimits_{k\in K} (S \cap [x_{k-1}, x_k])\,,$$

which, because each interval $[x_{k-1}, x_k]$ is closed in $\mathbb{R}$, implies that

$$\overline{S} \;=\; \bigcup\nolimits_{k\in K} \left(\overline{S} \cap [x_{k-1}, x_k]\right).$$

Unfortunately, the latter equation does *not* imply that K is also the set of all indices k such that $\overline{S} \cap [x_{k-1}, x_k] \neq \emptyset$. Indeed, it is possible that $\overline{S} \cap [x_{k-1}, x_k]$ includes one or both of the endpoints x_{k-1} or x_k, and thus that an adjacent interval *does* include points of $\overline{S}$. The simple solution is to modify the partition $\mathcal{P}$ by widening, slightly, the intervals of the form $[x_{k-1}, x_k]$ with $k{\in}K$, thus simultaneously narrowing the other subintervals. For example, let $\delta \,>\, 0$ be small enough that if $k{\in}K$ and $1 \,<\, k \,<\, n$, then

$$\frac{x_{k-2} + x_{k-1}}{2} \;<\; x_{k-1} - \delta \text{ and } x_k + \delta \;<\; \frac{x_{k+1} + x_k}{2}.$$

Similarly, let δ also be small enough so that if $k \,=\, 1$ is in K, then $x_1 + \delta \,<\, (x_1 + x_2)/2$, while if $k \,=\, n$ is in K, then $x_{n-1} - \delta \,>\, (x_{n-2} + x_{n-1})/2$. For each $k{\in}K$ replace x_k by $x_k + \delta$ and x_{k-1} by $x_{k-1} - \delta$, as appropriate, in $\mathcal{P}$, and leave the other partition points of $\mathcal{P}$ unchanged, to obtain a new partition $\mathcal{Q} \,=\, \{a \,=\, y_0 \,<\, y_1 \,<\, \ldots y_{n-1} \,<\, y_n \,=\, b\}$ of $[a, b]$. With this partition, the original set K is now the set of indices for which $\overline{S} \cap [y_{k-1}, y_k] \neq \emptyset$, and the adjacent subintervals for $\mathcal{Q}$ do not intersect $\overline{S}$. Since K has at most n elements, and for each k in K there are at most two extensions of length δ, it follows that

$$\Sigma(\overline{S}; \mathcal{Q}) \;=\; \sum_{k\in K} (y_k - y_{k-1}) \leq 2\,n\,\delta + \Sigma(S; \mathcal{P}).$$

Finally, let $\varepsilon \,>\, 0$ be given, let $\mathcal{P}$ be chosen so that in addition one has $\Sigma(S; \mathcal{P}) \,<\, \varepsilon/2$, and choose $\delta \,>\, 0$ small enough as above and also small enough that $\delta \,<\, \varepsilon/(4\,n)$. It follows that $\Sigma(\overline{S}; \mathcal{P}) \,<\, \varepsilon$, and thus $\overline{S}$ is a Jordan null set, as claimed.

(c) Let $\varepsilon \,>\, 0$ be given, and let $\mathcal{P}$ and T be chosen so that $\Sigma(S \setminus T; \mathcal{P}) \,<\, \varepsilon/2$ and T is Jordan null. Let $\mathcal{Q}$ be a partition of $[a, b]$ such that $\Sigma(T; \mathcal{Q}) \,<\, \varepsilon/2$, and let $\mathcal{R} \,=\, \mathcal{P} \cup \mathcal{Q}$. Then, since $S \,=\, (S \setminus T) \cup T$, and $\mathcal{R}$ is a refinement of both $\mathcal{P}$ and $\mathcal{Q}$, it follows from Part (a) of Lemma (VII.5.5) that

$$\Sigma(S; \mathcal{R}) \leq \Sigma(S \setminus T; \mathcal{R}) + \Sigma(T; \mathcal{R}) \leq \Sigma(S \setminus T; \mathcal{P}) + \Sigma(T; \mathcal{Q}) \;<\; \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \;=\; \varepsilon.$$

(d) This follows directly from Part (b) of Lemma (VII.5.5) above.

VII.5.7 Examples

(1) It is easy to show that every finite subset of $\mathbb{R}$ is Jordan null.

(2) Suppose that S and T are Jordan null sets in $\mathbb{R}$. Then $W \,=\, SS \cup T$ is also a Jordan null set. Indeed, since (by definition) S and T are bounded subsets of $\mathbb{R}$, it follows that so is W. Let $[a, b]$ be any interval such that $W \subseteq [a, b]$, so that S and T are also subsets of $[a, b]$. (Recall that, by Remark (VII.5.4) (2), the choice of the particular interval $[a, b]$does not matter here.) Let $\varepsilon \,>\, 0$

be given, and let $\mathcal{P}$ and $\mathcal{Q}$ be partitions of $[a,b]$ such that $\Sigma(S;\mathcal{P}) < \varepsilon/2$ and $\Sigma(T;\mathcal{Q}) < \varepsilon/2$. Let $\mathcal{R} = \mathcal{P}\cup\mathcal{Q}$, so that $\mathcal{R}$ is a refinement of both $\mathcal{P}$ and $\mathcal{Q}$. Then, by Remark (VII.5.4) (1), one has

$$\Sigma(W;\mathcal{R}) \leq \Sigma(S;\mathcal{R}) + \Sigma(T;\mathcal{R}) \leq \Sigma(S;\mathcal{P}) + \Sigma(T;\mathcal{Q}) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

The desired result follows.

By repeatedly applying this result, it is easy to see that the union of finitely many Jordan null sets is also Jordan null.

(3) Let S be the set of rational numbers in the unit interval $[0,1]$. It is clear that if $\mathcal{P}$ is any partition of $[0,1]$, then *every* subinterval of $\mathcal{P}$ has nonempty intersection with S, so that $\Sigma(S;\mathcal{P}) = 1$. In particular, S is *not* Jordan null. Of course, this set S is countable, and thus is the union of countably many Jordan null sets, namely, singleton sets.

(4) Consider, instead, the set $S = \{1, 1/2, 1/3, \dots 1/n, \dots\}$ consisting of all the reciprocals of natural numbers. This set is also the union of countably many Jordan-null sets, but here it is an easy exercise to show that S *is* Jordan null.

(5) It is an interesting exercise to show that the Cantor Middle-Thirds set, although uncountable, *is* Jordan null.

The preceding examples illustrate the complexity of the relation between 'Jordan null' and 'countable unions'. The next result clarifies this relation somewhat.

VII.5.8 Theorem

Suppose that S is a bounded subset of $\mathbb{R}$ which can be expressed as the union of a countable family of Jordan-null sets. If, in addition, S is a closed subset of $\mathbb{R}$, then S is also Jordan null.

Proof Without loss of generality, assume that $S \neq \emptyset$ and that a and b are numbers such that $a < \inf S$ and $b > \sup S$. Then, by hypothesis, there is a sequence $X_1, X_2, \dots X_n, \dots$ of Jordan-null subsets of $[a,b]$ such that $S = \bigcup_{j=1}^{\infty} X_j$, and it follows from Part (b) of Lemma (VII.5.5) that for each index j there exists a partition $\mathcal{P}_j$ of $[a,b]$ such that for which no partition point of $\mathcal{P}_j$ is in the set X_j; denote the k-th subinterval of $\mathcal{P}_j$ by $[c_{jk}, d_{jk}]$. For each index j let K_j denote the set of numbers k such that the k-th subinterval of the partition $\mathcal{P}_j$ has nonempty intersection with X_j, and let $V_j = \bigcup_{k\in K_j}(c_{jk}, d_{jk})$ be the union of the interiors of such subintervals of $\mathcal{P}_j$. Then $X_j \subseteq V_j$, since X_j has no partition points of $\mathcal{P}_j$, and $\Sigma(X_j;\mathcal{P}_j) = \sum_{k\in K_j}(d_{jk} - c_{jk}) < \varepsilon/2^j$. Since $S = \bigcup_{j=1}^{\infty} X_j \subseteq \bigcup_{j=1}^{\infty} V_j$, it follows that the open sets V_j, $j = 1, 2, \dots$ form an open cover of S. Since, by hypothesis, S is closed and bounded in $\mathbb{R}$, it follows from the Heine-Borel Theorem that there is a finite subcollection $V_{j_1}, V_{j_2}, \dots V_{j_m}$, with $j_1 < j_2 < \dots < j_m$, of these open sets which also covers S. To simplify the notation, let $N = j_m$. Then the collection of sets $V_j, 1 \leq j \leq N$, is also a finite open cover of S, as is the collection of open intervals of the form (c_{jk}, d_{jk}), with $k\in K_j$ and $1 \leq j \leq N$. The sum of the lengths of these open intervals satisfies the conditions

$$\sum_{j=1}^{N}\sum_{k\in K_j}(d_{jk} - c_{jk}) = \sum_{j=1}^{N}\Sigma(X_j;\mathcal{P}_j) < \sum_{j=1}^{N}\frac{\varepsilon}{2^j} < \varepsilon.$$

Since ε can be any positive number, it follows from Part (b) of Lemma (VII.5.5) that S is Jordan null, as claimed. ∎

VII.5.9 Example

The subset $S = [a,b]$ cannot be expressed as the countable union of Jordan-null subsets of $[a,b]$. Indeed, since S is a closed subset of $[a,b]$, if it could be so expressed, then it would itself be Jordan null, contraary to what has already been proved.

The preceding theorem can be used to provide an important characterization of Riemann integrabilty.

VII.5.10 Theorem

Suppose that $g : [c,d] \to \mathbb{R}$ is a bounded function on the interval $[c,d]$, and let D denote the set of discontinuities of g on $[c,d]$; at the endpoints c and d use the appropriate 'one-sided' notion of continuity. Then a necessary and sufficient condition for g to be Riemann integrable on $[c,d]$ is that D can be expressed as the union of a countable family of Jordan-null sets.

ProofTo simplify the discussion, let a and b be numbers so that $a < c$ and $b > d$, so that $a < \inf D$ and $b > \sup D$. Define $f : [a,b] \to \mathbb{R}$ by the rule

$$f(x) = \begin{cases} 0 & \text{if } a \le x < c \\ g(x) & \text{if } c \le x \le d \\ 0 & \text{if } d < x \le b \end{cases}$$

Denote the set of discontinuities of f on $[a,b]$ by S, again using the appropriate notion of one-sided continuity at the endpoints a and b. Note that S consists of the points of D, possibly augmented by one or both of the endpoints c and d. It is clear from Examples (VII.5.7) (2) and (3) that D can be expressed as the union of a countable family of Jordan-null sets if, and only if, the same property holds for S. Likewise, it is clear from Parts (b) and (d) of Theorem (**??**) that g is integrable on $[c,d]$ if, and only if, f is integrable on $[a,b]$. Thus, it suffices to consider the integrability of f on $[a,b]$, for which the set S of discontinuities satisfies the simplifying condition $a < \inf S$ and $b > \sup S$.

Suppose first that f is integrable on $[a,b]$. For each natural number q let S_q denote the set of points x in $[a,b]$ such that $\omega_f(x) \ge 1/q$. For such q let $\varepsilon > 0$ be given, and let $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b\}$ be any partition of $[a,b]$ such that $\Delta(f;\mathcal{P}) < \varepsilon q$. Let $T = \mathcal{P}$. Clearly if $x \in S_q \setminus T$, then $x_{k-1} < x < x_k$ for exactly one index k. It follows that for such x one has $1/q \le \omega_f(x) \le M_k - m_k$, and thus $(x_k - x_{k-1}) \le (M_k - m_k)\,(x_k - x_{k-1})\,q$. Let K be the set of indices k such that $(S_q \setminus T) \cap [x_{k-1}, x_k] \ne \emptyset$. Then one has

$$\Sigma(S_q \setminus T;\mathcal{P}) = \sum_{k\in K}(x_k - x_{k-1}) \le \sum_{k\in K}(M_k - m_k)\,(x_k - x_{k-1})\,q \le \sum_{j=1}^{n}(M_j - m_j)\,(x_j - x_{j-1})\,q < \frac{\varepsilon}{q}q = \varepsilon.$$

It now follows from Part (c) of Corollary (VII.5.6) that S_q is Jordan null. Clearly S is the union of the countable family of Jordan null sets of the form S_q with q in $\mathbb{N}$.

Conversely, suppose that the stated condition holds, and let $Y_1, Y_2, \ldots Y_n, \ldots$ be a sequence of Jordan-null sets whose union is S. For each q in $\mathbb{N}$ let S_q be as above. It is easy to see that S_q is a closed and bounded subset of $\mathbb{R}$. Since S_q is the union of the Jordan null sets of the form $X_j = Y_j \cap S_q$ for j in $\mathbb{N}$, it follows from Theorem (**??**) that, for each q, S_q is also Jordan null. Theorem??? (REFERENCE???) then implies that f is integrable on $[a,b]$, hence g is integrable on $[c,d]$, as claimed. ■

The preceding result illustrates the importance of a set being expressible as the countable union of Jordan-null subsets of $\mathbb{R}$. This property is equivalent to a related concept due to Lebesgue.

VII.5.11 Definition

A subset S of $\mathbb{R}$ is said to be **of Lebesgue measure zero**, or, more briefly, **Lebesgue null**,, provided it can be expressed as the union of countably many Jordan null subsets of $\mathbb{R}$.

Remarks (1) In contrast with the definition of 'Jordan null', the definition 'Lebesgue null' does not restrict S to be a *bounded* subset of $\mathbb{R}$. This makes sense because every subset S of $\mathbb{R}$ can be expressed as a countable union of bounded sets; for example, the sets of the form $S \cap [k, k+1]$ with k in $\mathbb{Z}$.

(2) Lebesgue's original formulation of a set being of (Lebesgue) measure zero is slightly different. The equivalence of his definition and the one given here is outlined in an exercise.

With this new terminology, one can now phrase Theorem (VII.5.10) in the following more standard form:

VII.5.12 Theorem

Suppose that $g : [c, d] \to \mathbb{R}$ is bounded on the interval $[c, d]$, and let D denote the set of discontinuities of g on $[c, d]$; at the endpoints c and d use the appropriate 'one-sided' notion of continuity. Then a necessary and sufficient condition for g to be Riemann integrable on $[c, d]$ is that D be a set of Lebesgue measure zero.

The significance of the preceding result is that it shows that the (Riemann) integrability of a bounded function on an interval depends only on the set D of discontinuitiesof that function, and not the detailed nature of those discontinuities at individual points of D. The next result illustrates the power of this formulation.

It is convenient for future reference to point out the following simple facts about sets which are Lebesgue null. The simple proofs are left as exercises.

VII.5.13 Theorem

(a) Every Jordan-null set is Lebesgue null. In particular, every finite sybset of $\mathbb{R}$ is Lebesgue null.

(b) Every subset of a Lebesgue-null set is Lebesgue null.

(c) The union of a countable family of Lebesgue-null sets is Lebesgue null. (Recall that the analogous statement with 'Lebesgue' replaced by 'Jordan' is not true.) In particular, every countable subset of $\mathbb{R}$ is Lebesgue null.

(d) If S contains an interval as a subset, then S is *not* Lebesgue null.

In Examples (VII.1.7), Theorem (VII.2.2) and Theorem (VII.2.3) we determine the integrability, or nonintegrability, in several particular cases. Each case requires a separate argument based on the specific nature of the function at hand. The following result shows how Theorem (VII.5.12) allows a unified treatment of such cases.

VII.5.14 Examples

(1) If $f : [a,b] \to \mathbb{R}$ is continuous on $[a,b]$, then it is Riemann integrable on $[a,b]$. Indeed, in this case the set D of discontinuities of f is the empty set, which is a finite set.

(2) The Dirichlet function is *not* Riemann integrable on any interval $[a,b]$, since the corresponding set of discontinuities is $[a,b]$ itself, which is not Lebesgue null.

(3) If $f : [a,b] \to \mathbb{R}$ is bounded on $[a,b]$ and the set of discontinuities of f in $[a,b]$ is countable, then f is Riemann integrable on $[a,b]$.
Special Cases The set of discontinuities of the Thomae function in the interval $[0,1]$ is a subset of the rationals, hence is countable. Likewise, the set of discontinuities of a function $f : [a,b] \to \mathbb{R}$ that is monotonic on $[a,b]$ is countable. Since both functions are clearly bounded, it follows that both are Riemann intergable.

(4) Suppoose that $f : [a,b] \to \mathbb{R}$ is bounded on $[a,b]$ and that c is a number such that $a < c < b$. Then f is Riemann integrable on $[a,b]$ if, and only if, it is Riemann integrable on each of the subintervals $[a,c]$ and $[c,b]$. Indeed, let D be the set of discontinuities of f on $[a,b]$. Likewise, let D_1 and D_2 be the corresponding sets for $[a,c]$ and $[c,b]$, respectively, but using the appropriate one-sided limits at c. It is clear that either $D = D_1 \cup D_2$ or $D = D_1 \cup D_2 \cup \{c\}$. In either case, it follows from Theorem (VII.5.13) that D is Lebesgue null if, and only if, both of the sets D_1 and D_2 are Lebesgue null.

(5) Suppose that f_1 and f_2 are both Riemann integrable on $[a,b]$. Let D_1 be the set of discontinuities of f_1 in $[a,b]$, and let D_2 be the corresponding set for f_2. Then the set D of discontinuities in $[a,b]$ of the function $f = f_1 + f_2$ is a subset of the finite union $D_1 \cup D_2$ of Lebesgue-null sets, and thus is itself Lebesgue null. Since f_1 and f_2 must be bounded on $[a,b]$, so is their sum f. Thus f is also Riemann integrable on $[a,b]$.
A similar argument shows that the product $g = f_1 \cdot f_2$ of the Riemann integrable functions f_1 and f_2 is Riemann integrable on $[a,b]$, then g is Riemann integrable on $[a,b]$. Likewise, if there is a constant $c > 0$ such that $|f_2(x)| \geq c$ for all x in $[a,b]$, then the quotient $h = f_1/f_2$ is bounded on $[a,b]$, and its set of discontinuities is a subset of the Lebesgue-null set $D_1 \cup D_2$, so that h is also Riemann integrable on $[a,b]$.

Remark In Part (b) of Theorem (VII.2.2) we prove not only that the sum $f = f_1 + f_2$ of two Riemann integrable functions is integrable, but also the formula $\int_a^b f = \int_a^b f_1 + \int_a^b f_2$. The proof of that formula can be simplified if one already has proved, as above, that f is integrable on $[a,b]$. Indeed, from the integrability of f_1 and f_2 one has

$$L(f_1;\mathcal{P}) \leq \int_a^b f_1 \leq U(f_1;\mathcal{P}) \text{ and } L(f_2;\mathcal{P}) \leq \int_a^b f_2 \leq U(f_2;\mathcal{P})$$

for every partition $\mathcal{P}$ of $[a,b]$. It then follows from the standard properties of Darboux sums that for each such partition $\mathcal{P}$ one has

$$L(f_1+f_2;\mathcal{P}) \leq L(f_1;\mathcal{P})+L(f_2;\mathcal{P}) \leq \left(\int_a^b f_1\right)+\left(\int_a^b f_2\right) \leq U(f_1;\mathcal{P})+U(f_2;\mathcal{P}) \leq U(f_1+f_2;\mathcal{P}).$$

In particular, one has

$$L(f;\mathcal{P}) \leq \left(\int_a^b f_1\right)+\left(\int_a^b f_2\right) \leq U(f;\mathcal{P}).$$

for every such partition $\mathcal{P}$. However, since f is already proved to be integrable on $[a,b]$, it follows that there is only one number which lies between $L(f;\mathcal{P})$ and $U(f;\mathcal{P})$ for each such partition $\mathcal{P}$, namely $\int_a^b f$. The desired equation now follows.

The following consequnce of Theorem (VII.5.12) makes it easy to prove the Riemann integrability of a wide class of functions.

VII.5.15 Corollary

Suppose that $f : [a,b] \to \mathbb{R}$ is Riemann integrable on $[a,b$. Suppose further that $H : X \to \mathbb{R}$ is a function which is defined on a subset X of $\mathbb{R}$ containing all numbers of the form $h(x)$ with x in $[a,b]$. If H is continuous and bounded on X, then the composition $g = H \circ f : [a,b] \to \mathbb{R}$ is also Riemann integrable on $[a,b]$.

Proof The hypotheses certainly imply that the expression $(F \circ g)(x)$ is defined for every x in $[a,b]$. In addition, Theorem (IV.2.4) implies that g is continuous at each x in $[a,b]$ at which f is continous, hence the set S of discontinuities of g in $[a,b]$ is a subset of the set D of discontinuities of f in $[a,b]$. It follows easily that S is a set of Lebesgue measure zero, and thus h is Riemann integrable on $[c,d]$, as claimed.

Example It was proved in Theorem (VII.2.3) that if $f : [a,b] \to \mathbb{R}$ is Riemann integrable on $[a,b]$, then so is the function $|f|$. This fact also follows from the preceding corollary by choosing H in that result to be the absolute-value function.

VII.6 EXERCISES FOR CHAPTER VII

VII - 1 Preliminary Remarks The definition of $\int_a^b f\,d\alpha$ given in the *Notes* is the same as the one found in, say, Rudin's 'Principles of Mathematical Analysis'. Some authors use the following definition, which yields a somewhat different theory.

Alternate Definition Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$, and as usual let $\mathcal{P}_{[a,b]}$ denote the set of partitions of the interval $[a,b]$.

(i) If $\mathcal{P} = \{a = x_0 < x_1 < \ldots < x_k = b\}$ is a partition of $[a,b]$ then the **mesh of the partition** $\mathcal{P}$ is the number $||\mathcal{P}||$ given by

$$||\mathcal{P}|| = \max\{|x_1 - x_0|, |x_2 - x_1|, \ldots |x_k - x_{k-1}|\}.$$

Speaking geometrically, $||\mathcal{P}||$ is the length of the longest subinterval of $[a,b]$ determined by the partition $\mathcal{P}$.

(ii) Suppose that $f : [a,b] \to \mathbb{R}$ is a bounded function and $\alpha : [a,b] \to \mathbb{R}$ is monotonic up. Using the notation and terminology of Lemma H.1.7 and Definition H.1.8, one says that f is 'Riemann-Stieltjes Integrable' in this alternate sense provided that the following condition holds:

There exists a number A such that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathcal{P}$ is any partition of $[a,b]$ such that $||\mathcal{P}|| < \delta$, then for every choice list $\tau = (t_1, t_2, \ldots t_k)$ associated with $\mathcal{P}$ one has

$$|S(\mathcal{P}, f, \alpha, \tau) - A| < \varepsilon.$$

(Recall that $S(\mathcal{P}, f, \alpha, \tau) = \sum_{j=1}^{k} f(t_j)\Delta\alpha_j$.) The number A is then the value of the integral.

Problem

(a) Show that if f is Riemann-Stieltjes integrable on $[a,b]$ with respect to α in this alternate sense, then it is also Riemann-Stieltjes integrable on $[a,b]$ with respect to α in the sense of Definition H.1.4, and the integrals are equal.

(b) Give an example of $[a,b]$, f and α for which the integral exists in the sense of Definition H.1.4, but not in this alternate sense.

(c) Show that in the case of the ordinary Riemann integral, i.e., when $\alpha(x) = x$ for all x, the two definitions are equivalent; that is, they produce the same class of integrable functions, and the same values for the integrals.

VII - 2 Prove Part (a) of Theorem H.2.4 (Hint: Use Theorem H.1.9)

VII - 3 Prove Part (b) of Theorem H.2.4 (Hint: Use Theorem H.1.9)

VII - 4 For each *even* natural number m, let $\alpha^{(m)}$ be the integrator described on Page 415, in Example H.1.10 (6), the 'Simpson's Rule' example.

(a) Determine the value $\alpha^{(m)}(x)$ for each x in $[a,b]$.

(b) Show that if $f : [a,b] \to \mathbb{R}$ is continuous, then $\int_a^b f = \lim_{m\to\infty} \int_a^b f\,d\alpha^{(m)}$.

VII - 5 Prove or Disprove If $f \in \mathcal{R}_{[0,1]}$, then $\lim_{k\to\infty} \int_0^1 x^k f(x)\,dx = 0$.

Chapter VIII

Advanced Topics in Limits and Continuity

Introduction

In this chapter we return to the study of 'Limits' and of 'Continuous Functions', concepts which were introduced in Chapters (III) and (IV). Most of the topics discussed in the present chapter are *not* normally covered in elementary calculus courses.

VIII.1 Limits and Suprema/Infima; $\limsup$ and $\liminf$

The relations between suprema/infima and limits of monotonic sequences suggests that one might be able to describe 'Completeness' in terms of monotonic sequences. Indeed, some texts use one of the following pair of equivalent statements as their choice of the Completeness Axiom for $\mathbb{R}$:

'A sequence which is monotonic up and bounded above must be convergent'

or

'A sequence which is monotonic down and bounded below must be convergent'

Of course we have seen these statements as the 'Monotonic-Sequence Principle' in Part (b) of Theorem (III.2.5); thus they are results proved using our version of the Completeness Axiom, the Bisection Principle. In those other texts, however, one treats the Bisection Principle as a theorem to be proved using the Monotonic Sequence Principle. As usual, it is merely a matter of mathematical taste as to which statement one takes as the Completeness Axiom.

There is a simple construction which associates a pair of monotonic sequences, one 'up', the other 'down', with any given bounded sequence of real numbers.

VIII.1.1 Construction

Let $\xi = (x_1, x_2, \dots x_n, \dots)$ be a bounded sequence of real numbers. Thus, there exist real numbers m and M such that $m \leq x_k \leq M$ for all k in $\mathbb{N}$. From this one can easily construct examples of sequences $\alpha = (a_1, a_2, \dots)$ and $\beta = (b_1, b_2, \dots)$ with the following properties:

(i) α is monotonic up, and $a_k \leq x_k$ for all k;

(ii) β is monotonic down, and $b_k \geq x_k$ for all k.

For instance, simply let $a_k = m$ and $b_k = M$ for all k in $\mathbb{N}$.

One can do much better; indeed, one can construct a unique 'best possible' pair of sequences associated with ξ that satisfy (i) and (ii).

First note that if a sequence α exists which satisfies (i), then clearly a_1 must satisfy $a_1 \leq \inf\{x_1, x_2, \ldots\}$. For if a_1 did not satisfy this inequality, then one would have $a_1 > \inf\{x_1, x_2, \ldots\}$. By the defining properties of the infimum of a set, this in turn would imply that there must exist an index k such that $a_1 > x_k$. From the requirement that the sequence α should be monotonic up one then sees that $a_k \geq a_1$ and thus $a_k > x_k$, contradicting (i). This argument also shows that if such α exists, then one must have

$$a_k \leq \inf\{x_k, x_{k+1}, \ldots\} \text{ for each } k \text{ in } \mathbb{N}.$$

A similar argument shows that if a sequence β which satisfies (ii) exists, then β must satisfy

$$b_k \geq \sup\{x_k, x_{k+1}, \ldots\} \text{ for each } k \text{ in } \mathbb{N}.$$

The preceding discussion motivates the following.

VIII.1.2 Definition

(a) Let $\xi = (x_1, \ldots x_k, \ldots)$ be a sequence of real numbers which is bounded above. The **upper envelope associated with** ξ, denoted ξ^+, is the sequence whose k-th term $M_k(\xi)$ is given by the rule

$$M_k(\xi) = \sup\{x_k, x_{k+1}, \ldots\} \text{ for all } k \text{ in } \mathbb{N}.$$

(The fact that the suprema $M_k(\xi)$ exist, and are numbers, follows from the Supremum Principle.)

(b) Similarly, if the sequence ξ is bounded below then the **lower envelope associated with** ξ, denoted ξ^-, has k-th term $m_k(\xi)$ given by

$$m_k(\xi) = \inf\{x_k, x_{k+1}, \ldots\} \text{ for all } k \text{ in } \mathbb{N}.$$

(The fact that these infima exist, and are real numbers, follows from the Infimum Principle.)

Note If the context makes clear which sequence ξ is under consideration, one may abbreviate the notations $M_k(\xi)$ and $m_k(\xi)$ to M_k and m_k, respectively.

VIII.1.3 Remark

There is an obvious extension of the concept of 'upper envelope' to the case in which ξ is unbounded above. Indeed, in this case it is clear that $\sup\{x_k, x_{k+1}, \ldots\} = +\infty$ for *all* indices k, so the natural definition would be $\xi^+ = (+\infty, +\infty, \ldots)$. Likewise, there is an obvious way to extend the notion of 'lower envelope' to allow sequences which are unbounded below. However, it does not appear to be worth the effort to introduce these extensions, so we don't.

The next result shows how questions about convergence of arbitrary sequences can be reduced to the theory for monotonic sequences. To simplify the statement of the theorem, the hypothesis of 'boundedness' is included. This is a reasonable restriction, however, since unbounded sequences cannot be convergent.

VIII.1.4 Theorem (The Upper/Lower-Envelopes Theorem)

Let $\xi = (x_1, x_2, \dots)$ be a bounded sequence of real numbers, and let $\xi^+ = (M_1, M_2, \dots)$ and $\xi^- = (m_1, m_2, \dots)$ be the corresponding upper and lower envelopes associated with ξ, as described in Definition (VIII.1.2). Then:

(a) The upper envelope $\xi^+ = (M_1, M_2, \dots)$ is monotonic down, and the lower envelope $\xi^- = (m_1, m_2, \dots)$ is monotonic up. Furthermore, one has

$$m_k \leq x_k \leq M_k \text{ for each } k \text{ in } \mathbb{N}.$$

(b) The monotonic sequences ξ^+ and ξ^- are bounded, and thus are convergent.

(c) The original sequence ξ is convergent if, and only if, $\lim_{k\to\infty}(M_k - m_k) = 0$. When this occurs, one has $\lim_{k\to\infty} x_k = \lim_{k\to\infty} m_k = \lim_{k\to\infty} M_k$.

Proof:

(a) For convenience let A_k denote the set $\{x_k, x_{k+1}, \dots\}$, so that $m_k = \inf A_k$, and $M_k = \sup A_k$. It is clear that $A_{k+1} \subseteq A_k$ for each k in $\mathbb{N}$, so that by Part (b) of Theorem (**??**) one has, for each index k,

$$m_{k+1} = \inf A_{k+1} \geq \inf A_k = m_k, \text{ and } M_{k+1} = \sup A_{k+1} \leq \sup A_k = M_k.$$

That is, the claimed monotonicity holds.

(b) The monotonicity properties of the sequences ξ^+ and ξ^- imply that $m_k \geq m_1$ and $M_k \leq M_1$ for all indices k. Combining this with the hypothesis $m_k \leq x_k \leq M_k$ (and, of course, using Transitivity of Order in $\mathbb{R}$), one then obtains

$$m_1 \leq m_k \leq x_k \leq M_k \leq M_1 \text{ for all indices } k.$$

In particular, both of the sequences ξ^+ and ξ^- are bounded below by m_1 and bounded above by M_1. The fact that these sequences are convergent then follows from the Monotonic-Sequence Principle (Part (b) of Theorem (III.2.5))

(c) Let $A = \lim_{k\to\infty} m_k$ and $B = \lim_{k\to\infty} M_k$ be the (real) limits whose existence is proved in Part (b).

(i) Assume that $\lim_{k\to\infty}(M_k - m_k) = 0$, so that $A = B$. Since, by Part (a), one also has $m_k \leq x_k \leq M_k$, the Squeeze Property for Sequences (Part (c) of Theorem (III.2.5)) can be used to conclude that $\lim_{k\to\infty} x_k$ exists and equals the common limit of ξ^- and ξ^+.

Conversely, suppose that the sequence ξ is convergent, and let $L = \lim_{k\to\infty} x_k$. Let y and z be real numbers such that $y < L < z$. Then there exists a number B such that if $k \geq B$ then $y < x_k < z$. From this it is clear that for $k \geq B$ the number y is a lower bound for the set A_k and z is an upper bound for A_k. Thus, by the basic properties of 'supremum' and 'infimum', if $k \geq B$ then $y \leq m_k \leq M_k \leq z$. Thus, by Theorem (**??**) it follows that the sequences ξ^- and ξ^+ converge to L.

To illustrate the preceding result, let us use it to give a second proof of Part (b) of Theorem (III.2.1). That is, suppose $\xi = (x_1, x_2, \dots)$ is a convergent sequence of real numbers, with $\lim_{k\to\infty} x_k = L$, and that $\zeta = (z_1, z_2, \dots)$ is a subsequence of ξ. We want to show that ζ also converges to L by using the preceding theorem.

Proof Using Theorem (VIII.1.4): First, recall from Theorem (**??**) that to each infinite subset A of $\mathbb{N}$ there is a (unique) strictly increasing bijection $\Psi_A : \mathbb{N} \to A$ of $\mathbb{N}$ with A, given as follows:

$$\Psi_A(1) = \min A; \Psi_A(j+1) = \min A \backslash \{\Psi_A(1), \dots \Psi_A(j)\} \text{ for each } j \text{ in } \mathbb{N}.$$

Also recall (from the same theorem) that the subsequence ζ of the given sequence ξ can be expressed in the form $\zeta = \xi\circ\Psi_A$ for at least one infinite subset A of $\mathbb{N}$. (In terms of the notation used in Proof (A) above, $A = \{k_1, k_2, \ldots\}$, and $\Psi_A(j) = k_j$ for each j.) Since $k_j = \Psi_A(j) \geq j$ for each j in $\mathbb{N}$, and Ψ_A is strictly increasing, it follows that the set $\{z_j, z_{j+1}, \ldots\}$ is a subset of the set $\{x_j, x_{j+1}, \ldots\}$. Now apply Theorem (**??**) to conclude that

$$\inf\{x_j, x_{j+1}, \ldots\} \geq \inf\{z_j, z_{j+1}, \ldots\} \text{ and } \sup\{z_j, z_{j+1}, \ldots\} \leq \sup\{x_j, x_{j+1}, \ldots\}$$

Using the notation of Definition (VIII.1.2), one can then say

$$m_j(\xi) \leq m_j(\zeta) \leq M_j(\zeta) \leq M_j(\xi) \text{ for each index } j. \qquad \text{(VIII.1)}$$

By the Upper/Lower-Envelopes Theorem (Theorem (VIII.1.4)), combined with the hypothesis that $\lim_{k\to\infty} x_k = L$, one knows that the sequences $\xi^+ = (M_1(\xi), M_2(\xi), \ldots)$ and $\xi^- = (m_1(\xi), m_2(\xi), \ldots)$ both converge to L. Now apply Inequality (VIII.1) and the Squeeze Property for Sequences (Part (c) of Theorem (III.2.5)) to conclude that the sequences $\zeta^+ = (M_1(\zeta), M_2(\zeta), \ldots)$ and $\zeta^- = (m_1(\zeta), m_2(\zeta), \ldots)$ also both converge to L. Finally, apply the Upper/Lower-Envelopes Theorem again to conclude that the subsequence ζ converges to L, as claimed.

We have seen repeatedly that it is important to analyse the convergence properties of subsequences of a given sequence. The next definition provides terminology to aid in that analysis.

VIII.1.5 Definition

Let $\xi = (x_1, x_2, \ldots)$ be a real sequence. A quantity L is said to be a **subsequential limit of** ξ if there exists a subsequence $(x_{k_1}, x_{k_2}, \ldots)$ of ξ such that $L = \lim_{j\to\infty} x_{k_j} = L$. Note that the quantity L can be a real number or one of the infinities $+\infty$, $-\infty$.

The set of all subsequential limits of ξ is denoted by $\mathcal{L}[\xi]$.

VIII.1.6 Remarks

(1) It follows from Theorem (III.4.2), (i.e., the Extended Bolzano-Weierstrass Theorem), that the set $\mathcal{L}[\xi]$ is nonempty.

(2) It follows from Part (g) of Theorem (**??**) that ξ has a limit if, and only if the set $\mathcal{L}[\xi]$ has precisely one element. Also, when this condition occurs, that element equals $\lim_{k\to\infty} x_k$.

There is a simple characterization of the set of subsequential limits of a sequence.

VIII.1.7 Theorem

Let $\xi = (x_1, x_2, \ldots)$ be a sequence of real numbers.

(a) Let L be a real number. Then L is an element of $\mathcal{L}[\xi]$ if, and only if, for every pair of numbers y and z such that $y < L < z$, there are infinitely many indices k such that $y < x_k < z$. Alternate Phrasing: L is an element of $\mathcal{L}[\xi]$ if, and only if, for every $\varepsilon > 0$ there are infinitely many indices k such that $|L - x_k| < \varepsilon$.

(b) The quantity $+\infty$ is an element of $\mathcal{L}[\xi]$ if, and only if, the sequence ξ is unbounded above.

(c) Likewise, the quantity $-\infty isin \mathcal{L}[\xi]$ if, and only if, ξ is unbounded below.

Proof (a) Suppose that $L \in \mathcal{L}[\xi]$, and let $(x_{k_1}, x_{k_2}, \ldots)$ be a subsequence of ξ which converges to L. Let y and z be numbers such that $y < L < z$. Then by Part (a) of Theorem (III.2.1) there is a number B such that if $j \geq B$ then $y < x_{k_j} < z$. Since the indices k_1, $k_2, \ldots$ form a strictly increasing sequence of natural numbers, it follows that there are infinitely many different indices k_j with $j \geq B$, Thus there are infinitely many indices k such that $y < x_k < z$; for instance, the k's of the form k_j with $j \geq B$.

Conversely, suppose that for each y and z in $\mathbb{R}$ such that $y < L < z$ there are infinitely many indices k such that $y < x_k < z$. Choose an infinite strictly increasing sequence of indices $k_1 < k_2 < \ldots$ as follows:

(i) x_{k_1} satisfies $L - 1 < x_{k_1} < L + 1$.

(ii) Suppose that indices $k_1 < k_2 < \ldots k_m$ have been chosen. By hypothesis, there are infinitely many indices k such that

$$L - \frac{1}{m+1} < x_k < L + \frac{1}{m+1}. \quad (*)$$

From among these, choose k_{m+1} so that $k_{m+1} > k_m$. Then it is clear that the subsequence $(x_{k_1}, x_{k_2}, \ldots)$ has the property that

$$|L - x_{k_{j_m}}| < \frac{1}{m} \text{ for each } m \text{ in } \mathbb{N}.$$

Since $\lim_{m \to \infty} 1/m = 0$, the Squeeze Property implies that this subsequence converges to L.

The proof that the alternate phrasing also works is left to the reader; see the proof of Part (a) of Theorem (III.2.1).

(b) Suppose that $+\infty \in \mathcal{L}[\xi]$. Then (by the definition of the set $\mathcal{L}[\xi]$) there is a subsequence of ξ which has $+\infty$ as limit. Clearly that subsequence is unbounded above, and thus ξ itself is unbounded above.

Conversely, suppose that ξ is unbounded above. Then, by Case (ii) of Theorem (III.4.2), there exists a subsequence of ξ which has $+\infty$ as its limit; thus, $+\infty \in \mathcal{L}[\xi]$.

(c) Apply the conclusions of Part (b) to the sequence $(-x_1, -x_2, \ldots)$. The details are left to the reader.

The preceding result allows one to easily show that the limit properties of a sequence of real numbers do not depend on the order in which one writes down the terms of the sequence.

VIII.1.8 Theorem

Let $\xi = (x_1, x_2, \ldots)$ be a sequence of real numbers, and suppose that $\sigma = (s_1, s_2, \ldots)$ is a sequence obtained by permuting the terms of the sequence ξ. That is, suppose that σ can be expressed in the form $\sigma = \xi \circ F$, where $F : \mathbb{N} \to \mathbb{N}$ is a bijection of $\mathbb{N}$ onto $\mathbb{N}$. Then $\mathcal{L}[\sigma] = \mathcal{L}[\xi]$.

Proof First, suppose that L is a real number in the set $\mathcal{L}[\sigma]$, and let y and z be numbers such that $y < L < z$. By Theorem (VIII.1.7) there are infinitely many indices j such that $y < s_j < z$. Let A be the set of such indices, and let $B = F[A]$. Since F is one-to-one, the set B is also an infinite subset of $\mathbb{N}$. Suppose that $k \in B$, so that $k = F(j)$ for a unique j in A. Then, since $s_j = x_{F(j)} = x_k$, it follows that x_k also satisfies $y < x_k < z$ for all k in the infinite set B. Thus, by Theorem (VIII.1.7) again, one sees that $L \in \mathcal{L}[\xi]$ as well. A similar argument shows that if one of the infinities is in $\mathcal{L}[\sigma]$ then it is also in $\mathcal{L}[\xi]$. Combining these results leads to the conclusion that $\mathcal{L}[\sigma] \subseteq \mathcal{L}[\xi]$.

By reversing the roles of σ and ξ in the preceding argument, which one can do because F is invertible and thus one can write $\xi = \sigma \circ F^{-1}$, one sees that $\mathcal{L}[\xi] \subseteq \mathcal{L}[\sigma]$ as well. Thus, $\mathcal{L}[\sigma] = \mathcal{L}[\xi]$, as claimed.

VIII.1.9 Corollary

Suppose that ξ and σ are real sequences which differ only by a permutation of their indices; that is, there exists a bijection $F : \mathbb{N} \to \mathbb{N}$ of $\mathbb{N}$ onto itself such that $\sigma = \xi \circ F$. Then ξ has a limit if, and only if, σ has a limit; and when this occurs, their limits are equal.

Proof Combine the results of the preceding theorem with the result stated in Remark (VIII.1.6) (2) above.

VIII.1.10 Examples

(1) Let $\xi = (x_1, x_2, \dots)$ be the sequence given by $x_k = (-1)^{k-1}$ for each k in $\mathbb{N}$; that is, $\xi = (1, -1, 1, -1, \dots)$. This sequence is bounded (note that $|x_k| = 1$ for all k), so neither $+\infty$ nor $-\infty$ is in $\mathcal{L}[\xi]$. Likewise, if L is a real number such that $L \neq 1$ and $L \neq -1$, then there exists $\varepsilon > 0$ such that $|L - x_k| \geq \varepsilon$ for all indices k. Indeed, let $\varepsilon = \min\{|L-1|, |L+1|\}$.

In contrast, it is clear that 1 and -1 are both in $\mathcal{L}[\xi]$. For instance, no matter which $\varepsilon > 0$ is chosen, there are infinitely many indices k for which $|1 - x_k| < \varepsilon$; indeed, one has $1 - x_k = 0$ whenever k is odd. Likewise, $|-1 - x_k| = 0 < \varepsilon$ if k is even.

Thus one has $\mathcal{L}[\xi] = \{-1, 1\}$.

(2) Recall that the set $\mathbb{Q}$ of all rational numbers is countable (see Corollary (I.8.12)). Thus, there exists a bijection $\alpha : \mathbb{N} \to \mathbb{Q}$ which maps $\mathbb{N}$ one-to-one onto $\mathbb{Q}$. As such, the map α is an infinite sequence $(a_1, a_2, \dots)$ of rational numbers in which each rational number appears exactly once.

Claim: $\mathcal{L}[\alpha] = \mathbb{R} \cup \{-\infty, +\infty\}$.

Proof of Claim First, note that the set $\mathbb{Q}$ is unbounded above and below, so the same is true for the sequence α. Thus $-\infty$ and $+\infty$ are elements of the set $\mathcal{L}[\alpha]$.

Next, notice that if L is a real number and if y and z are numbers such that $y < L < z$, then there are infinitely many rational numbers in the open interval (y, z). Since each rational number corresponds to exactly one index k, there exist infinitely many indices k such that $y < a_k < z$.

The claim now follows by applying Part (a) of Theorem (VIII.1.7).

VIII.1.11 Theorem

Let $\xi = (x_1, x_2, \dots)$ be a sequence of real numbers, and let $\mathcal{L}[\xi]$ be the corresponding set of subsequential limits of ξ. Then the set $\mathcal{L}[\xi]$ has both a maximum element and a minimum element. That is, there are quantities L_1 and L_2 such that

(i) L_1 and L_2 are elements of $\mathcal{L}[\xi]$, and

(ii) $L_1 \geq L \geq L_2$ for all L in $\mathcal{L}[\xi]$.

(Of course, we allow the possibility that L_1 or L_2 could be an infinity.)

Proof: Let us first show that $\mathcal{L}[\xi]$ has a maximum element. There are several cases to consider.

Case 1 Suppose that $\mathcal{L}[\xi]$ is a singleton set $\{L\}$. (By Part (g) of Theorem (**??**) this corresponds to the situation in which $\lim_{k\to\infty} x_k = L$.) In this case one sees that the choice $L_1 = L$ is the desired maximum element. (Of course it is also the desired minimum element, but we are not yet ready to discuss the minimum element in general.)

Case 2 Suppose that $+\infty \in \mathcal{L}[\xi]$. Then clearly $L_1 = +\infty$ is the desired maximum.

Case 3 Suppose that $\mathcal{L}[\xi]$ is not a singleton set, and assume also that $+\infty$ is not in $\mathcal{L}[\xi]$. Then, by Part (a) of Theorem (VIII.1.7), ξ is bounded above. Let M in $\mathbb{R}$ be an upper bound for ξ, so that $M \geq x_k$ for all k. Let $\zeta = (z_1, z_2, \dots)$ be any subsequence of ξ which has a limit. Since ζ is a subsequence of ξ, it follows that $M \geq z_j$ for all j in $\mathbb{N}$, and thus $M \geq \lim_{j\to\infty} z_j$. Now let $X = \mathcal{L}[\xi] \backslash \{-\infty\}$. Then X is a nonempty set of real numbers which is bounded above by the number M. Let $L_1 = \sup X$, so that L_1 is a real number and $M \geq L_1 \geq L$ for all L in X. Of course $L_1 > -\infty$, so certainly $L_1 \geq L$ for all L in $\mathcal{L}[\xi]$; that is, L_1 satisfies Condition (ii) above.

Next, note that because of Theorem (III.2.18) one knows that there exists a monotonic-up sequence of numbers $b_1, b_2, \dots$ in X (hence in $\mathcal{L}[\xi]$) such that $L_1 = \lim_{j\to\infty} b_j$. Since $b_j \in \mathcal{L}[\xi]$, it follows that there exists a sequence of subsequences $\tau_1 = (t_{11}, t_{12}, \dots)$, $\tau_2 = (t_{21}, t_{22}, \dots)$, ... of ξ such that $b_j = \lim_{m\to\infty} t_{jm}$ for each $j = 1, 2, \dots$.

Now let $\varepsilon > 0$ be given. By the Alternate Phrasing of Part (a) of Theorem (VIII.1.7) there exist infinitely many k such that $|L_1 - b_k| < \varepsilon/2$. Let p be one such index, and consider the corresponding subsequence $\tau_p = (t_{p1}, t_{p2}, \dots)$, so that $b_p = \lim_{j\to\infty} t_{pj}$. By the definition of convergent sequence, there exist infinitely many indices j such that $|b_p - t_{pj}| < \varepsilon/2$. By the Triangle Inequality one then has

$$|L_1 - t_{pj}| = |L_1 - b_p| + |b_p - t_{pj}| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for infinitely many indices j. Since each number t_{pj} is of the form x_{k_j} for some index k_j, and since the indices k_j, $j = 1, 2, \dots$ form a strictly increasing sequence, it follows that $|L_1 - x_k| \leq \varepsilon$ for infinitely many indices j. Thus, by Theorem (VIII.1.7) again, it follows that $L_1 \in \mathcal{L}[\xi]$, as claimed.

The proof that $\mathcal{L}[\xi]$ has a minimum element can be carried out in a similar manner. Or, better yet, one can simply apply to the sequence $-\xi$ the result just obtained about the maximium element; the details are left to the reader.

VIII.1.12 Definition

Let $\xi = (x_1, x_2, \dots)$ be a sequence of real numbers, and (as usual) let $\mathcal{L}[\xi]$ denote the corresponding set of subsequential limits of ξ. The maximum element of $\mathcal{L}[\xi]$ is called the **limit superior of** ξ. It is denoted by an expression such as $\limsup\limits_{k\to\infty} x_k$, or, on occasion,by $\limsup \xi$. (The symbol $\limsup$ is pronounced 'lim soup'.) This quantity is also called the **upper limit of the sequence** ξ. As such, it is sometimes denoted by an expression such as $\overline{\lim}_{k\to\infty} x_k$; but some authors combine the phrase 'upper limit' with the '$\limsup$' notation.

Likewise, the minimum element of the set $\mathcal{L}[\xi]$ is called the **limit inferior**, or the **lower limit**, of the sequence ξ, and it is denoted by expressions such as $\liminf\limits_{k\to\infty} x_k$ or $\underline{\lim}_{k\to\infty} x_k$.

VIII.1.13 Examples

(1) Let $\xi = (x_1, x_2, \dots)$ be the sequence given by $x_k = (-1)^{k-1}$ for each k in $\mathbb{N}$. From the results obtained in Example (VIII.1.10) (1) one sees that $\limsup \xi = +1$, $\limsup \xi = -1$.

(2) Let $\alpha = (a_1, a_2, \dots)$ be a sequence of real numbers. Let X be the set of all real numbers $R \geq 0$ such that the sequence $\rho = (a_1 R, a_2 R^2, a_3 R^3, \dots)$ is a bounded sequence. It is clear

that the set X is nonempty; for example, $R = 0$ is obviously in X. Then one can show that $\sup X = 1/\limsup \xi$, where $\xi = (|a_1|, \sqrt[2]{|a_2|}, \sqrt[3]{|a_3|}, \ldots \sqrt[k]{|a_k|} \ldots)$.

Note It appears that the concept – although not the terminology – of 'limit superior' is due Cauchy in his proof of a version of the preceding result.

VIII.1.14 Theorem

Let $\xi = (x_1, x_2, \ldots)$ be a sequence of real numbers.

(a) A necessary and sufficient condition for $\limsup \xi = +\infty$ is that the sequence ξ be unbounded above.

Likewise, a necessary and sufficient condition for $\liminf \xi = -\infty$ is that the sequence ξ be unbounded below.

(b) Suppose that ξ is bounded above, and let L be a real number. Then a necessary and sufficient condition for L to equal $\limsup_{\,k \to \infty} x_k$ is that for every number $\varepsilon > 0$ one has
(i) $x_k > L + \varepsilon$ for only finitely many indices k; and
(ii) $x_k > L - \varepsilon$ for infinitely many indices k.

(c) Suppose that ξ is bounded below, and let L be a real number. Then a necessary and sufficient condition for L to equal $\liminf_{\,k \to \infty} x_k$ is that for every number $\varepsilon > 0$ one has
(i) $x_k < L - \varepsilon$ for only finitely many indices k; and
(ii) $x_k < L + \varepsilon$ for infinitely many indices k.

(d) Suppose that ξ is bounded above, and let $\xi^+ = (M_1(\xi), M_2(\xi), \ldots)$ denote the upper envelope associated with ξ (see Definition (VIII.1.2)). Then

$$\limsup_{k \to \infty} x_k = \lim_{k \to \infty} M_k(\xi). \tag{VIII.2}$$

Likewise, suppose that ξ is bounded below, and let $\xi^- = (m_1(\xi), m_2(\xi), \ldots)$ denote the lower envelope associated with ξ. Then

$$\liminf_{k \to \infty} x_k = \lim_{k \to \infty} m_k(\xi). \tag{VIII.3}$$

The simple proof is left as an exercise.

VIII.1.15 Corollary

Let $\xi = (x_1, x_2, \ldots)$ be a sequence of real numbers. A necessary and sufficient condition for the sequence ξ to have a limit L is that

$$\limsup_{k \to \infty} x_k = \liminf_{k \to \infty} x_k = L.$$

Proof This follows from Parts (a) and (d) of the preceding theorem when combined with the results of Theorem (VIII.1.4).

NOTE: Some mathematics texts use Equation (VIII.2) as the *definition* of the limit superior of a real sequence which is bounded above. Likewise, they use Equation (VIII.3) as the *definition* of the limit inferior of a real sequence which is bounded below. Normally, however, such texts would write these equations in the form

$$\limsup_{k \to \infty} x_k = \lim_{k \to \infty} (\sup \{x_k, x_{k+1}, \ldots\})$$

and

$$\liminf_{k\to\infty} x_k = \lim_{k\to\infty} \left(\inf\{x_k, x_{k+1}, \dots\}\right).$$

That is, they probably would not introduce the auxiliary notion of 'enveloping sequence'.

In contrast, some texts define the concepts of $\limsup \xi$ and $\liminf \xi$ in terms of the conditions stated in Parts (b) and (c) of Theorem (VIII.1.14).

Remark on the 'sup' and 'inf' Terminology

Many students get confused when trying to sort out the differences between the word 'supremum' and the phrase 'limit superior'. One obvious source of this confusion is that mathematicians have elected to use the same abbreviation, namely 'sup', for both 'supremum' and 'superior'. A similar confusion holds between 'infimum' and 'limit inferior'. Thus, it may be useful to briefly consider the linguistic backgrounds of these words and phrases.

It has already been stated that the words 'supremum' and 'infimum' are of Latin origin; indeed, one often uses the Latin version of their plurals, ('suprema' and 'infima' respectively). In any event, these words are nouns which mean (roughly) 'highest one' and 'lowest one', respectively.

In contrast, the phrases 'limit superior' and 'limit inferior' are word-for-word translations into English of the Latin phrases 'limes superior' and 'limes inferior'. Unfortunately, word-for-word translations between languages with different structures often produce awkward phrasings. Indeed, in Latin the words 'superior' and 'inferior' are adjectives; and as such they are placed after the noun they modify, 'limes', because that's proper Latin grammar. English, in contrast, is a language in which an attributive adjective is normally placed *before* the noun it modifies; thus better translations would have been 'superior limit' and 'inferior limit'. Of course, these last phrases have virtually the same meanings in English as the corresponding 'upper limit' and 'lower limit' mentioned in Definition (VIII.1.12).

The situation gets even less clear if a textbook introduces the notations $\limsup$ and $\liminf$, but fails to tell its readers what phrases they correspond to. Lacking any guidance, the reader of such a textbook is then likely to conclude – incorrectly – that these symbols should be pronounced 'limit supremum' and 'limit infimum'.

Special cases of the first two of these topics, 'Uniform Convergence' and 'Uniform Continuity' have already made appearances in Chapter (V), but without being mentioned explicitly by name. This reflects the normal situation in the development of mathematical concepts: formal definitions normally arise *after* one knows that the idea is useful and thus worth naming, not at the beginning of the theory.

A good way to prepare for the introduction of the new concepts of 'Uniform Convergence' and 'Uniform Continuity' would be to carefully reread the proof of Theorem (V.8.1). (That theorem states that if f is a C^1 function on an open interval I, then f has an antiderivative on I.) Indeed, we use some of the details of that proof to motivate the definitions of the new concepts below. While rereading that earlier proof, ask yourself 'What is the real heart of the proof? What features makes it work?'

VIII.2 Uniform Convergence of a Sequence of Functions

RE-DO LIGHT OF CHANGES IN CHAPTER E

In the proof of Theorem (V.8.1) we use the fact that, for every index n, one has

$$|f(x) - g_n(x)| \leq \frac{M(b-a)}{n} \text{ for each } x \text{ in } [a,b].$$

This fact implies that for each x in $[a,b]$ the sequence $\gamma(x) = (g_1(x), g_2(x), \dots)$ converges to $f(x)$. However, what that inequality tells us is, in a subtle way, much stronger than that; this extra strength is needed to complete of the proof of Theorem (V.8.1). The next definition clarifies the subtlety involved.

VIII.2.1 Definition (Pointwise Convergence; Uniform Convergence)

Let $\gamma = (g_1, g_2, \dots)$ be a sequence of real-valued functions which are all defined on a nonempty set $X \subseteq \mathbb{R}$.

(1) One says that the sequence γ **converges pointwise on** X **to a function** $f : X \to \mathbf{R}$ provided that for each x in X the numerical sequence $\gamma(x) = (g_1(x), g_2(x), \dots)$ converges to the number $f(x)$. That is, provided the following holds:

$$\text{for every } x \text{ in } X \text{ if } \varepsilon > 0 \text{ then there exists } B \text{ such that } k \geq B \text{ implies that } |f(x) - g_k(x)| < \varepsilon \qquad (I)$$

(2) One says that the sequence γ **converges uniformly on** X **to a function** $f : X \to \mathbf{R}$ provided the following holds:

$$\text{if } \varepsilon > 0 \text{ then there exists } B \text{ such that } k \geq B \text{ implies that } |f(x) - g_k(x)| < \varepsilon \text{ for every } x \text{ in } X \qquad (II)$$

On the surface, Statements (I) and (II) appear to be nearly the same; indeed, the only real difference is the placement of the phrase 'for every x in X'. But the location of that phrase affects the meaning of the statement. Indeed, in Statement (I) by placing the phrase 'for every x in X' first, one is saying that if you first chose x in X and $\varepsilon > 0$, then one can find a number B for which $|f(x) - g_k(x)| < \varepsilon$ when $k \geq B$; in particular, the value of B may depend in an essential way on the choice of both x and ε. In Statement (II), however, the choice of B can be made given only ε; in particular, B can be chosen so that $|f(x) - g_n(x)| < \varepsilon$ when $k \geq B$, and that this same B works *simultaneously* for all x in X.

VIII.2.2 Examples

(1) The sequence $\gamma = (g_1, g_2, \dots)$ which appears in the proof of Theorem (V.8.1) converges uniformly on $[a,b]$ to the function f in that theorem. Indeed, suppose that $\varepsilon > 0$ is given, and let $B = M(b-a)/\varepsilon$. Then it follows from the inequality $|f(x) - g_n(x)| \leq \frac{M(b-a)}{n}$ for each x in $[a,b]$ which appears in the proof of that theorem that if $n > B$ then $|f(x) - g_n(x)| < \varepsilon$ for all x in $[a,b]$.

(2) Let $\Gamma = (G_1, G_2, \dots)$ and F be as in the proof of the same theorem. It is easy to check that the sequence Γ converges uniformly on $[a,b]$ to F.

(3) It is clear that if a sequence $\gamma = (g_1, g_2, \dots)$ converges uniformly on X to a function f, then it certainly converges pointwise on X to f.

(4) One does not have to consider exotic situations to find a sequence of functions which converges pointwise, but not uniformly, to a function on a set. For instance, let $X = \mathbb{R}$ and let

$g_k(x) = \left(x + \dfrac{1}{k}\right)^2$ for all x in $\mathbb{R}$. It is clear that for each x in $\mathbb{R}$ one has $\lim_{k\to\infty} g_k(x) = x^2$; that is, the sequence γ converges pointwise on $\mathbb{R}$ to the function $f : \mathbb{R}\to\mathbb{R}$ given by $f(x) = x^2$. However, the sequence γ does not converge uniformly on $\mathbb{R}$ to f. Indeed, suppose that $\varepsilon > 0$ is given. Note that

$$|f(x) - g_k(x)| = \left|x^2 - \left(x + \frac{1}{k}\right)^2\right| = \left|\frac{2x}{k} + \frac{1}{k^2}\right| \geq \left|\frac{2|x|k - 1}{k^2}\right|,$$

where the final inequality comes from using the Modified Triangle Inequality. It is easy to see that if $x > (1 + \varepsilon k^2)/(2k)$ then $|f(x) - g_k(x)| > \varepsilon$.

(5) In the preceding example the fact that the domain $\mathbb{R}$ is unbounded is crucial to the proof that the sequence γ fails to be uniformly convergent. Indeed, it is a simple exercise to show that the same sequence γ does converge uniformly to the same function f on any closed bounded interval $[a, b]$.

(6) For each k in $\mathbb{N}$ let $g_k = \hat{B}^{[1]}_{[\frac{1}{k+1},\frac{1}{k}]}$ be the normalized C^1 bump function on the interval $\left[\dfrac{1}{k+1}, \dfrac{1}{k}\right]$; see Definition (**??**). It is an easy exercise to show that the sequence $\gamma = (g_1, g_2, \ldots)$ converges pointwise to the zero function on $\mathbb{R}$, but that it fails to converge uniformly on any interval of the form $[0, b]$ with $b > 0$. In contrast, the sequence *does* converge uniformly on each interval $[a, b]$ with $a > 0$.

(7) Let $g_k(x) = x^k$ for all x in $[0, 1]$. It is easy to see that the sequence $\gamma = (g_1, g_2, \ldots)$ converges pointwise on $[0, 1]$ to the function $f : [0, 1]\to\mathbb{R}$ given by the rule

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

It is an easy exercise to show that the sequence γ fails to be uniformly convergent on $[0, 1]$.

(8) Let p_k denote the Taylor polynomial of order k for the exponential function exp about the center point $c = 0$; that is,

$$p_k(x) = 1 + x + \frac{x^2}{2} + \ldots + \frac{x^k}{k!}$$

In Theorem (**??**) it is proved that

$$|\exp x - p_k(x)| = \frac{\exp(r_k)}{(k+1)!}|x|^{k+1} \leq \exp(|x|)\frac{|x|^{k+1}}{(k+1)!}$$

for every x in $\mathbb{R}$. In particular, if $R > 0$ is given, then for each x in the interval $[-R, R]$ one has

$$\exp(|x|)\frac{|x|^k}{k!} \leq e^R\frac{R^k}{k!}.$$

It follows that the sequence of these Taylor polynomials converges uniformly to exp on $[-R, R]$.

VIII.2.3 Theorem

Let $\gamma = (g_1, g_2, \ldots)$ be a sequence of real-valued functions defined on a nonempty set $X \subseteq \mathbb{R}$.

(a) If γ converges uniformly on X to a function $f : X \rightarrow \mathbb{R}$, then it converges uniformly to f on every nonempty subset of X.

(b) If $X = A_1 \cup A_2 \cup \ldots \cup A_m$, where for each $j = 1, 2, \ldots m$ A_j is a nonempty subset of X, then γ converges uniformly on X to a function $f : X \rightarrow \mathbb{R}$ if, and only if, it converges uniformly to f on each subset A_j.

(c) Suppose that f is a function which is defined on X. For each k let $M_k = \sup\{|f(x) - g_k(x)| : x \in X\}$. Then a necessary and sufficient condition for the sequence γ to converge uniformly on X to f is that $\lim_{k \to \infty} M_k = 0$.

Note: It is possible that some of the quantities M_k might equal $+\infty$. The notation $\lim_{k \to \infty} M_k$ here then tacitly assumes that this can occur only for finitely many indices k; compare with Remark (**??**) and Definition (III.1.10).

The simple proof is left as an exercise.

Cauchy claimed to prove the following result (or at least a result that is easily seen to be equivalent)):

'*If a sequence of continuous functions converges pointwise to a function f on an interval I, then the limit function f is also continuous on I.*'

Unfortunately, this statement in not correct, as Example (7) above illustrates. The next result, which appears to be due to Weierstrasse in its current form, shows that Cauchy's conclusion can be obtained if one replaces the hypothesis of 'pointwise convergence' with 'uniform convergence'.

VIII.2.4 Theorem (The 'Uniform-Convergence-Preserves-Continuity' Theorem)

Let X be a nonempty subset of $\mathbb{R}$.

(a) Let $\gamma = (g_1, g_2, \ldots)$ be a sequence of real-valued functions with domain X. Assume that the sequence γ converges uniformly on X to a function $f : X \rightarrow \mathbb{R}$. If each function g_k, $k = 1, 2, \ldots$, is continuous at a certain point c of X, then the function f is also continuous at c. Likewise, if each function g_k is continuous on X, then f is continuous on X.

(b) Let $\gamma = (g_1, g_2, \ldots)$ be a sequence of real-valued functions which are defined and continuous on X. Suppose that γ converges pointwise on X to a function $f : X \rightarrow \mathbb{R}$, and that for each interval $[a, b]$ for which $X \cap [a, b] \neq \emptyset$ the convergence to f is uniform. Then f is continuous on X.

Proof (based on the '$\varepsilon\delta$' characterization of continuity)

(a) Note that for each x in X and each k in $\mathbb{N}$ one has

$$|f(x) - f(c)| = |(f(x) - g_k(x)) + (g_k(x) - g_k(c)) + (g_k(c) - f(c))| \leq |f(x) - g_k(x)| + |g_k(x) - g_k(c)| + |g_k(c) - f(c)| \qquad (*)$$

Now let $\varepsilon > 0$ be given, and let B be large enough so that if $k \geq B$ then $|f(z) - g_k(z)| \leq \varepsilon/3$ for all z in X. Let k be such an index, and let $\delta > 0$ be small enough that if $|x - c| < \delta$ then $|g_k(x) - g_k(c)| < \varepsilon/3$. (Such δ exists because of the hypothesis that g_k is continuous at c.) With this choice of k, Inequality $(*)$ then implies

$$|f(x) - f(c)| \leq |f(x) - g_k(x)| + |g_k(x) - g_k(c)| + |g_k(c) - f(c)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$$

for all x in X such that $|x - c| < \delta$. The continuity of f at c now follows. Of course if each g_k is continuous at each point of X, then c can be any element of X, hence f is continuous on X.

(b) This result follows easily by applying Part (a) to the sets of the form $X \cap [a, b]$; the details are left as an exercise.

VIII.2.5 Examples

(1) The preceding theorem provides an explanation for the fact that the sequence in Example (VIII.2.2) (7) fails to be uniformly convergent on $[0, 1]$. Indeed, the functions $g_k(x) = x^k$ in that example are continuous on $[0, 1]$, but the limit function f fails to be continuous at 1.

(2) The hypothesis, in the preceding theorem, that the continuous functions g_k converge *uniformly* to f on X, is sufficient to guarantee the continuity of the limit function f, but is certainly not necessary; see Example (VIII.2.2) (6). Indeed, even the hypothesis that the functions g_k be continuous is not necessary to get f continuous; the reader is invited to find a suitable example.

The '$\varepsilon/3$' proof given above is the standard one found in most texts. The classic text *Principles of Mathematical Analysis* by W. Rudin bases the proof on the 'sequential' characterization of continuity, using the following result.

VIII.2.6 Lemma

Suppose that $\gamma = (g_1, g_2, \dots)$ is a sequence of functions defined on a nonempty subset X of $\mathbb{R}$, and assume that γ converges uniformly on X to a function $f : X \to \mathbb{R}$. Let $\xi = (x_1, x_2, \dots) : \mathbb{N} \to \mathbb{R}$ be a Cauchy sequence in X. Suppose that for each n in $\mathbb{N}$ the corresponding sequence $g_n \circ \xi = (g_n(x_1), g_2(x_2), \dots)$ of values is also Cauchy. Then the sequence $f \circ \xi = (f(x_1, f(x_2), \dots))$ is Cauchy. Furthermore, for each n one let $A_n = \lim g_n \circ \xi$, and likewise let $A = \lim f \circ \xi$. Then the sequence $(A_1, A_2, \dots)$ is convergent, and one has $\lim_{n \to \infty} A_n = A$.

Note: This last equation is sometimes written

$$\lim_{n \to \infty} \lim_{k \to \infty} g_n(x_k) = \lim_{k \to \infty} \lim_{n \to \infty} g_n(x_k).$$

Proof Note that for each m, k and n in $\mathbb{N}$ one has

$$|f(x_{m+k}) - f(x_m)| \leq |f(x_{m+k}) - g_n(x_{m+k})| + |g_n(x_{m+k}) - g_n(x_m)| + |g_n(x_m) - f(x_m)| \qquad (*)$$

Let $\varepsilon > 0$ be given, and let B_1 in $\mathbb{N}$ be large enough that if $n \geq B_1$ then $|f(y) - g_n(y)|$ for all y in X. (Such B_1 exists because of the hypothesis that the sequence γ converges uniformly to f on X.) In particular, for such n one has $|f(x_{m+k}) - g_n(x_{m+k})| < \varepsilon/3$ and $|g_n(x_m) - f(x_m)| < \varepsilon/3$ for each m and k in $\mathbb{N}$, since the points of the sequence ξ are in X. Now fix $n \geq B_1$, and for that n let B_2 be large enough that if $m \geq B_2$ then $|g_n(x_{m+k}) - g_n(x_m)| < \varepsilon/3$. (Such B_2 exists because of the hypothesis that the sequence $(g_n(x_1), g_n(x_2), \dots)$ is Cauchy.) Then from $(*)$ one gets

$$|f(x_{m+k}) - f(x_m)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \text{ for all } m \geq B_2 \text{ and all } k.$$

It follows that the sequence $(f(x_1), f(x_2), \dots)$ is Cauchy, as claimed. Note that the limits A and A_n which appear in the statement of the result exist and are finite because Cauchy sequences are convergent.

By doing a similar analysis on the inequality

$$|A - A_n| \leq |A - f(x_k)| + |f(x_k) - g_n(x_k)| + |g_n(x_k) - A_n|,$$

one also sees that $A = \lim_{n\to\infty} A_n$, as required.

Remarks

(1) It is easy to give an alternate proof of Theorem (VIII.2.4) based on the results of the precedeing lemma; the details are left as an exercise.

(2) Notice that the preceding lemma does not assume any continuity of the functions g_n or f; nor does it assume that the limit of the Cauchy sequence ξ lies in X.

As in the theory of limits of sequences of numbers, there are useful concepts of 'Cauchy sequences' in the context of pointwise and uniform convergence.

VIII.2.7 Definition

Let X be a nonempty subset of $\mathbb{R}$, and let $\gamma = (g_1, g_2, \dots)$ be a sequence of real-valued functions with domain X.

(1) The sequence γ is said to be **pointwise Cauchy on** X provided for each x in X the sequence $\gamma(x) = (g_1(x), g_2(x), \dots)$ is a Cauchy sequence. That is: for every x in X and every $\varepsilon > 0$, there exists B such that if n in $\mathbb{N}$ satisfies $n \geq B$ then $|g_{n+k}(x) - g_n(x)| < \varepsilon$ for all k in $\mathbb{N}$.

(2) The sequence γ is said to be **uniformly Cauchy on** X provided that for every $\varepsilon > 0$, there exists B such that if n in $\mathbb{N}$ satisfies $n \geq B$ then $|g_{n+k}(x) - g_n(x)| < \varepsilon$ for all k in $\mathbb{N}$ and all x in X.

Example In the proof of Part (a) of Theorem (V.8.1), the sequence $\Gamma = (G_1, G_2, \dots)$ introduced there is shown to satisfy the inequality

$$|G_{n+k}(x) - G_n(x)| \leq \frac{2M(b-a)^2}{n}$$

for all x in $[a, b]$. Thus the sequence Γ is uniformly Cauchy on $[a, b]$.

VIII.2.8 Theorem

Let X be a nonempty subset of $\mathbb{R}$, and let $\gamma = (g_1, g_2, \dots)$ be a sequence of real-valued functions with domain X.

(a) A necessary and sufficient condition for γ to converge pointwise on X to some function is that γ be pointwise Cauchy on X.

(b) A necessary and sufficient condition for γ to converge uniformly on X to some function is that γ be uniformly Cauchy on X.

The simple proof is left as an exercise.

Uniform convergence works well at 'preserving continuity', but is not nearly so useful in dealing with differentiability.

VIII.2.9 Examples

(1) For each k in $\mathbb{N}$ let $g_k : \mathbb{R} \to \mathbb{R}$ be given by the formula

$$g_k(x) = \sqrt{x^2 + \frac{1}{k}}.$$

Each of these functions is differentiable on $\mathbb{R}$. In addition, it is easy to show that the sequence $\gamma = (g_1, g_2, \dots)$ converges uniformly on $\mathbb{R}$ to the absolute-value function abs. The latter function fails to be differentiable at 0. Thus, it is *not* the case that the uniform limit of functions which are differentiable on an interval needs to be differentiable on that interval.

(2) For each k in $\mathbb{N}$ let $g_k : \mathbb{R} \to \mathbb{R}$ be given by $g_k(x) = (\sin kx)/\sqrt{k}$. Then it is obvious that the sequence $\gamma = (g_1, g_2, \dots)$ converges uniformly on $\mathbb{R}$ to the zero function. However the corresponding sequence of derivatives $g_k'(x) = \sqrt{k}\cos kx$ does not converge to a differentiable function; for example, it fails to even remain bounded at $x = 0$.

In contrast, uniform convergence works quite well with *anti*derivatives.

VIII.2.10 Theorem (The 'Uniform-Converence-Preserves-Antidifferentiability' Theorem)

Let $\gamma = (g_1, g_2, \dots)$ be a sequence of real-valued functions defined on an open interval I in $\mathbb{R}$. Assume that γ converges pointwise on I to a function $f : I \to \mathbb{R}$, and that on each closed bounded subinterval $[a, b]$ of I the convergence is uniform. If each of the functions g_k has an antiderivative on I, then f has an antiderivative on I. More precisely, fix a point c in I, and set $G_k = D_c^{-1} g_k$. Then the sequence $\Gamma = (G_1, G_2, \dots)$ converges pointwise on I to a function $F : I \to \mathbb{R}$ such that $F'(x) = f(x)$ for all x in I, and $F(c) = 0$. The sequence Γ converges uniformly to F on each closed bounded subinterval $[a, b]$ of I.

Proof Let a and b be numbers in I such that $a < c < b$.
Claim 1 The sequence Γ is uniformly Cauchy on $[a, b]$.

Proof of Claim 1 For each n and k in $\mathbb{N}$ and each x in $[a, b]$ one has

$$|(G_{n+k} - G_n)(x)| = |(G_{n+k} - G_n)(x) - (G_{n+k} - G_n)(c)| = |(G_{n+k} - G_n)'(\hat{x})(x - c)| = |(g_{n+k}(\hat{x}) - g_n(\hat{x}))(x - c)|$$

for some number $\hat{x}$ in $\mathrm{Seg}\,[x, c] \subseteq [a, b]$. Since, by hypothesis, the sequence γ is uniformly convergent on $[a, b]$, it follows that the sequence Γ is also uniformly Cauchy on $[a, b]$. Indeed, let $\varepsilon > 0$ be given, and let B in $\mathbb{N}$ be large enough that if $n \geq B$ then $|g_{n+k}(z) - g_n(z)| < \varepsilon/|b - a|$ for all z in $[a, b]$. In particular this inequality holds for $z = \hat{x}$, so one gets

$$|(G_{n+k} - G_n)(x)| \leq |(g_{n+k}(\hat{x}) - g_n(\hat{x}))(x - c)| < \left(\frac{\varepsilon}{|b - a|}\right)|x - c| \leq \varepsilon$$

for all n and k in $\mathbb{N}$ such that $n \geq B$ and all x in $[a, b]$. Claim 1 follows.

Now define $F : \mathbb{R} \to \mathbb{R}$ by the rule $F(x) = \lim_{n \to \infty} G_n(x)$ for each x in I. Thus, the sequence Γ converges uniformly to F on each closed bounded subinterval $[a, b]$ of I. In particular, one also has $F(c) = \lim_{n \to \infty} G_k(c) = 0$, since by construction one has $G_n(c) = 0$ for each n.

Claim 2 For each x in I one has $F'(x) = f(x)$.

Proof of Claim 2 Fix x in I, and let a and b in I be chosen so that $x, c \in (a, b)$. For each t in $[a, b]$ and each n and k in $\mathbb{N}$ one has

$$|(G_{n+k} - G_n)(t) - (G_{n+k} - G_n)(x)| = \left|(G_{n+k} - G_n)'(\hat{x})(t - x)\right| = |g_{n+k}(\hat{x}) - g_n(\hat{x})|\,|t - x|$$

for some $\hat{x}$ in $\mathrm{Seg}\,[x, t]$. Now let $Y = [a, b] \setminus \{x\}$. Then for t in Y the preceding equation can be written

$$\left|\left(\frac{G_{n+k}(t) - G_{n+k}(x)}{t - x}\right) - \left(\frac{G_n(t) - G_n(x)}{t - x}\right)\right| = |g_{n+k}(\hat{x}) - g_n(\hat{x})| \qquad (*)$$

For convenience, define $\varphi_m : Y \to \mathbb{R}$ by the rule

$$\varphi_m(t) = \frac{G_m(t) - G_m(x)}{t - x}$$

Then $(*)$ can be written

$$|\varphi_{n+k}(t) - \varphi_n(t)| \leq |g_{n+k}(\hat{x}) - g_n(\hat{x})|.$$

Let $\varepsilon > 0$ be given, and let B in $\mathbb{N}$ be large enough that if $n \geq B$ then $|g_{n+k}(z) - g_n(z)| < \varepsilon$ for all z in $[a, b]$. Then one has

$$|\varphi_{n+k}(t) - \varphi_n(t)| \leq \varepsilon \text{ for all } n \geq B \text{ and all } k \text{ in } \mathbb{N} \text{ and for all } t \text{ in } Y.$$

That is, the sequence $\Phi = (\varphi_1, \varphi_2, \dots)$ is uniformly Cauchy on Y. Define $\psi : Y \to \mathbb{R}$ by

$$\psi(t) = \frac{F(t) - F(x)}{t - x}$$

Then it follows from the fact that $\lim_{k \to \infty} G_k(t) = F(t)$ that the sequence Φ converges uniformly on Y to ψ.

We are now ready to apply Lemma (VIII.2.6), with the roles of X, g_n and f in the lemma being played here by Y, φ_n and ψ, respectively. Indeed, let $\tau = (t_1, t_2, \dots)$ be a sequence of points in Y converging to x; note that τ is a Cauchy sequence in Y, but its limit, x, is not in Y. Since, by hypothesis, G_n is differentiable at x, one sees that for each n in $\mathbb{N}$, one has $\lim_{k \to \infty} \varphi_n(t_k) = A_n$, where $A_n = G'_n(x) = g_n(x)$. Lemma (VIII.2.6) then implies the sequence $(A_1, A_2, \dots\dots)$ converges to some number A. However, one has $A_n = g_n(x)$, and by hypothesis $\lim_{n \to \infty} g_n(x) = f(x)$. Thus, $A = f(x)$, and this holds for every choice of the sequence τ. In addition, it also follows from that same lemma that the sequence $(\psi(t_1), \psi(t_2), \dots)$ also converges to A, that is, to $f(x)$, independently of the choice of τ. That is, one has

$$\lim_{k \to \infty} \frac{F(t_k) - F(x)}{t_k - x} = f(x).$$

Since this is independent of the choice of sequence τ, it follows that $\lim\limits_{t \to x} \dfrac{F(t) - F(x)}{t - x} = f(x)$. This implies that $F'(x)$ exists and equals $f(x)$. Since this argument works for every x in I, it follows that $F'(x) = f(x)$ for all x in I, as claimed.

The preceding result is often phrased as follows, in terms of derivatives instead of antiderivatives.

VIII.2.11 Corollary

Suppose that $\Gamma = (G_1, G_2, \dots)$ is a sequence of real-valued functions which are differentiable on an open interval I, and assume that there is a point c in I such that the numerical sequence $\Gamma(c)$ is convergent. If the corresponding sequence $\gamma = (g_1, g_2, \dots)$ of derivatives $g_k = G'_k$ converges uniformly on I to some function f, then the sequence Γ converges uniformly on I to a function F such that $F' = f$ on I.

The reader is encouraged to see why this corollary is equivalent to the preceding theorem. (The only feature that requires any thought is to see where the hypothesis, that the sequence $\Gamma(c) = (G_1(c), G_2(c), \dots)$ is convergent, fits in.)

One of the iconic results of classical analysis is the Weierstrass Approximation Theorem. Weierstrass was 70 years of age when he published it.

VIII.2.12 Theorem (The Weierstrass Approximation Theorem)

Suppose that $f : [a,b] \to \mathbb{R}$ is continuous on the closed bounded interval $[a,b]$. Then there is an infinite sequence $\varphi = (p_1, p_2, \dots)$ of polynomials which converges uniformly to f on $[a,b]$.

Proof The proof given here is essentially that of Serge Bernstein. It is ultimately based on ideas from probability theory. We follow the treatment in the text 'Advanced Calculus' by Patrick Fitzgerald.

Background – Coin Tosses Suppose that one has a two-sided coin which is 'weighted' so that the probability of 'Heads' is p and thus the probability of 'Tails' is $q = 1-p$. Now do an experiment in which the coin is tossed n times. Then it is known that the probability of getting exactly j heads out of n tosses is $C(n,j)p^j(1-p)^j$, where $C(n,j)$ denotes the **binary coefficient** $\dfrac{n!}{j!(n-j)!}$. In Bernstein's proof one can think of what is going on in probabilistic terms, but that is not required here.

Note that, for all x in $[0,1]$, we have the following identities:

$$\sum_{j=0}^{n} C(n,j)x^j(1-x)^{n-j} = 1 \qquad (I)$$

This follows from the Binomial Theorem: $(a+b)^n = \sum_{j=0}^{n} C(n,j)a^j b^{n-j}$, with $a = x$, $b = 1-x$.

$$\sum_{j=0}^{n} \left(\frac{j}{n}\right) C(n,j)x^j(1-x)^{n-j} = x \qquad (II)$$

This follows from the Binomial Theorem applied to $x(x+(1-x))^{n-1}$.

In Equations (I) and (II) n can be in any natural number. In the next equation one needs $n \geq 2$:

$$\sum_{j=0}^{n} \left(\frac{j(j-1)}{n(n-1)}\right) C(n,j)x^j(1-x)^{n-j} = x^2 \qquad (III)$$

This follows from the Binomial Theorem applied to $x^2(x+(1-x))^{n-2}$.

Claim For each n in $\mathbb{N}$ and each x in $[0,1]$ one has

$$\sum_{j=0}^{n} \left(x - \frac{j}{n}\right)^2 C(n,j)x^j(1-x)^{n-j} = \frac{x(1-x)}{n} \qquad (IV)$$

Proof of Claim The result is trivially true if $n = 1$, so assume that $n \geq 2$. Let L_n denote the left side of Equation (IV). Then one can write

$$L_n = \sum_{j=0}^{n} \left(x^2 - \frac{2xj}{n} + \frac{j^2}{n^2}\right) C(n,j)x^j(1-x)^{n-j} = A_n + B_n + C_n,$$

where

$$A_n = x^2 \sum_{j=0}^{n} C(n,j)x^j(1-x)^{n-j}, \quad B_n = (-2x)\sum_{j=0}^{n} \frac{j}{n} C(n,j)x^j(1-x)^{n-j}, \quad C_n = \sum_{j=0}^{n} \frac{j^2}{n^2} C(n,j)x^j(1-x)^{n-j}.$$

From (I) and (II) above one sees that $A_n = x^2$ and $B_n = -2x^2$. Also, one easily computes that

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

Now Equations (II) and (III) can be applied to get

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

Thus,

$$A_n+B_n+C_n \;=\; x^2-2x^2+\left(\frac{n-1}{n}\right)\left(x^2+\frac{x}{n-1}\right) \;=\; \frac{x(1-x)}{n},$$

as required.

It is easy to see that if one can prove the Approximation Theorem for continuous functions on the interval $[0,1]$, then the theorem is true for the general interval $[a,b]$, so we restrict our attention to the case $[a,b] \;=\; [0,1]$.

Thus, consider a function $f:[0,1]\to\mathbb{R}$ which is continuous on $[0,1]$. For each n in $\mathbb{N}$, Bernstein uses as the approximating n-th degree polynomial $p_n(x) \;=\; \sum_{j=0}^{n} f\left(\frac{j}{n}\right)B_{n,j}(x)$. Notice that, by Equation (I) above, one has

$$|f(x)-p_n(x)| \le \sum_{j=0}^{n}\left|f(x)-\left(\frac{j}{n}\right)\right| C(n,j)x^j(1-x)^{n-j} \qquad (*)$$

Now let $\varepsilon \;>\; 0$ be given. Since f is uniformly continuous on $[0,1]$, there exists $\delta \;>\; 0$ so that $|f(y)-f(x)| \;<\; \varepsilon/2$ for all x,y in $[0,1]$ such that $|y-x| \le \delta$. Let M be the maximum value of f on $[0,1]$. Then for every x in $[0,1]$ and every $j \;=\; 0,1,2,\ldots n$ one of the following must hold:

$$\left|x-\frac{j}{n}\right| \;<\; \delta, \text{ hence } \left|f(x)-f\left(\frac{j}{n}\right)\right| \;<\; \frac{\varepsilon}{2};$$

or

$$\left|x-\frac{j}{n}\right| \;\ge \delta, \text{ hence } \left|f(x)-f\left(\frac{j}{n}\right)\right| \;\le 2M \le \frac{2M}{\delta^2}\left(x-\frac{j}{n}\right)^2$$

Thus one has

$$\left|f(x)-f\left(\frac{j}{n}\right)\right| \;<\; \frac{\varepsilon}{2}+\frac{2M}{\delta^2}\left(x-\frac{j}{n}\right)^2.$$

for all x in $[0,1]$ and all $j \;=\; 0,1,2,\ldots n$. Multiply both sides of this last inequality by $C(n,j)x^j(1-x)^{n-j}$ and sum over j to get

$$|f(x)-p_n(x)| \le \frac{\varepsilon}{2}\sum_{j=0}^{n}C(n,j)x^j(1-x)^{n-j}+\frac{2M}{\delta^2}\sum_{j=0}^{n}\left(x-\left(\frac{j}{n}\right)\right)^2 C(n,j)x^j(1-x)^{n-j}.$$

In light of Equations (I) and (IV), one then has

$$|f(x)-p_n(x)| \le \frac{\varepsilon}{2}+\frac{2M}{\delta^2}\frac{x(1-x)}{n} \le \frac{\varepsilon}{2}+\frac{2M}{n\delta^2}.$$

Clearly if $n \;>\; \frac{4M}{\varepsilon\delta^2}$, then one gets

$$|f(x)-p_n(x)| \;<\; \frac{\varepsilon}{2}+\frac{\varepsilon}{2} \;=\; \varepsilon.$$

It follows that the sequence $(p_1,p_2,\ldots)$ converges uniformly to f on $[0,1]$, sa required.

VIII.3 Extending Continuous Functions

In Example (IV.2.5) (1) it is pointed out that if $f : X \to \mathbb{R}$ is a function which is continuous on a nonempty set $X \subseteq \mathbb{R}$, and if S is a nonempty subset of X, then the restriction $f|_S$ of f to S is also continuous on S.

It is natural to ask to what extent the converse statement holds; that is, to ask whether a real-valued function which is continuous on a set has a continuous extension on a larger set. (See Definition (I.5.1) to review the meanings of 'restriction' and 'extension'.) The next result gives an important partial answer.

VIII.3.1 Theorem (The Tietze Extension Theorem in R)

Suppose that $g : S \to \mathbb{R}$ is a function whose domain is a nonempty subset S of $\mathbb{R}$. Assume that S is a closed subset of $\mathbb{R}$, and that g is continuous at each point of its domain S. Let T be a subset of $\mathbb{R}$ such that $S \subseteq T$. Then there is an extension $f : T \to \mathbb{R}$ of g to T which is continuous on T.

Proof Assume first that $T = \mathbb{R}$. If $S = \mathbb{R}$ there is nothing to prove – simply let $f = g$ – so assume that S is a proper subset of $\mathbb{R}$. Let $U = \mathbb{R}\backslash S$. Then, by Theorem (**??**), U is the union of a countable family $\mathcal{F}$ of mutually disjoint open intervals.

To simplify the exposition, assume for the moment that S is a *bounded* closed subset of $\mathbb{R}$; that is, a compact subset. Then the open intervals in $\mathcal{F}$ are of three types: $I_1 = (-\infty, m)$, where m is the minimum element of S; $I_2 = (M, +\infty)$, where M is the maximum element of S; and (a, b), where a and b are certain elements of S with $a < b$. (The case $(-\infty, +\infty)$ cannot occur, since $S \neq \emptyset$ implies $U \neq \mathbb{R}$.)

In reality, there are infinitely many ways to extend g to a function which is continuous on all of $\mathbb{R}$, but perhaps the simplest one is constructed as follows:

(i) On the interval $I_1 = (-\infty, m)$ the function f must be chosen to be continuous and to approach the value $g(m)$ as $x \nearrow m$. The simplest choice is to make $f(x) = g(m)$ for all $x < m$, so f is constant on the open interval I_1.

(ii) Likewise, the simplest choice that can do the job on $I_2 = (M, +\infty)$ is $f(x) = g(M)$ for all $x > M$.

(iii) On intervals of the form (a, b), one must choose $f(x)$ so that $f(x) \to g(a)$ as $x \searrow a$, while $f(x) \to g(b)$ as $x \nearrow b$. The obvious simplest way to do that in a continuous manner is to use the process of 'linear interpolation'; see Example (**??**).

That is, define $f : \mathbb{R} \to \mathbb{R}$ by the rule

$$f(x) = \begin{cases} g(x) & \text{if } x \in S \\ g(b) & \text{if } x \text{ is in an interval in } \mathcal{F} \text{ of the form } (-\infty, b) \\ g(a) & \text{if } x \text{ is in an interval in } \mathcal{F} \text{ of the form } (a, +\infty) \\ (1-t)g(a) + tg(b) & \text{if } x \text{ is in an interval in } \mathcal{F} \text{ of the form } (a, b) \\ & \text{and } t \text{ in } (0,1) \text{ is such that } x = (1-t)a + tb \end{cases} \tag{VIII.4}$$

It is obvious from the first line of Equation (VIII.4) that f is an extension of g. Now let c be any real number. In light of Theorem (IV.2.6), in order to show that f is continuous at c it suffices to verify that $\lim_{k\to\infty} f(x_k) = f(c)$ for every *monotonic* sequence $\xi = (x_1, x_2, \dots)$ in $\mathbb{R}$ which converges to c.

Case 1 Suppose that c is *not* in S. Then c must lie in exactly one of the intervals that form the family $\mathcal{F}$; call that interval I_c. By Theorem (**??**) one must have x_k in I_c for k sufficiently large.

If I_c is of the form $(-\infty, m)$ then $f(c) = f(x_k) = g(m)$ for all sufficiently large k; in particular, $\lim_{k\to\infty} f(x_k) = f(c)$. A similar argument shows that if $I_c = (M, +\infty)$ then $\lim_{k\to\infty} f(x_k) = f(c)$.

If I_c is of the form (a, b) for certain elements a and b in S, the analysis is slightly harder. In this situation one can write $c = (1-t)a + tb$ for a unique t in $(0,1)$. Likewise, when k is large enough that $x_k{\in}I_c$ one has $x_k = (1-t_k)a + t_k b$ for certain t_k in $(0,1)$. It is clear that the condition $\lim_{k\to\infty} x_k = c$ implies $\lim_{k\to\infty} t_k = t$. Thus from the equations

$$f(x_k) = f((1-t_k)a + t_k b) = (1-t_k)g(a) + t_k g(b),$$

valid for sufficiently large k, and

$$f(c) = f((1-t)a + tb) = (1-t)g(a) + tg(b),$$

it follows that

$$\lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} ((1-t_k)g(a) + t_k g(b)) = (1-t)g(a) + tg(b) = f(c).$$

Thus, if c is not in S then f is certainly continuous at c.

Case 2 Suppose that $c{\in}S$, and suppose, to be definite, that the sequence ξ is monotonic *up*, so that $x_k \le c$ for each k. Let $\varepsilon > 0$ be given, and let $\delta > 0$ be small enough that if $x{\in}S$ and $|x - c| < \delta$ then $|g(x) - g(c)| < \varepsilon$. (That such $\delta > 0$ exists follows from the hypothesis that g is continuous at each point of S.)

Suppose that there exists a point u in S such that $c-\delta < u < c$. Then the assumption, that the sequence ξ is monotonic up and converges to c, implies that $u < x_k \le c$ for all sufficiently large k. For such k either $x_k{\in}S$ or $x_k \notin S$. In the former situation one has $|f(x_k) - f(c)| = |g(x_k) - g(c)| < \varepsilon$, by definition of δ and the fact that f is an extension of g. In the latter situation there must exist a, b in S, with $u \le a < b \le c$, such that $(a, b){\in}\mathcal{F}$ and $x_k{\in}(a, b)$. Then $f(x_k) = (1-t)g(a) + tg(b)$ for some t in $(0,1)$. Also, $|c - a| < \delta$ and $|c - b| < \delta$, so one gets

$$|f(c) - f(x_k)| = |g(c) - ((1-t)g(a) + tg(b))| = |(1-t)(g(c) - g(a)) + t(g(c) - g(b))| \le$$

$$(1-t)|g(c) - g(a)| + t|g(c) - g(b)| < (1-t)\varepsilon + t\varepsilon = \varepsilon.$$

In any event, if such u exists, then $|f(c) - f(x_k)| < \varepsilon$ for all sufficiently large k.

Suppose, instead, that no such u exists. Then either $c = m$ or there exists a in S, with $a \le c-\delta$, such that the open interval (a, c) is in the family $\mathcal{F}$. In the former situation one has $x_k \le m$ for all k, hence $f(x_k) = g(c)$ by definition. In the latter case one has x_k in (a, c) for all sufficiently large k and $f(x_k) = (1-t_k)g(a) + t_k g(c)$ for some t_k in $(0,1)$, by definition of f on points outside S. As in Case (1) above, it is clear that $\lim_{k\to\infty} t_k = 1$ since $\lim_{k\to\infty} x_k = c$. This implies

$$\lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} (1-t_k)g(a) + tg(c) = 0{\cdot}g(a) + 1{\cdot}g(c) = g(c).$$

In particular, if k is sufficiently large then $|f(x_k) - f(c)| < \varepsilon$.

It follows that, in all circumstances, $\lim_{k\to\infty} f(x_k) = f(c)$. A similar argument works for sequences $\zeta = (z_1, z_2, \dots)$ which are monotonic down and converge to c.

The conclusion then is that f is continuous at all c in $\mathbb{R}$, so that f is an extension of g to $\mathbb{R}$ which is continuous on $\mathbb{R}$, as required.

Suppose now that the set T referred to in the statement of the theorem is a proper subset of $\mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ be any continuous extension of g to $\mathbb{R}$. Then clearly $f|_T : T \to \mathbb{R}$ is a continuous extension of g to T.

Remark The 'Tietze' referred to in the title of the preceding theorem is the Austrian mathematician Heinrich Tietze (1880-1964), and it is part of his work in the field of General Topology. His general extension theorem, which is well outside the scope of these *Notes*, applies to spaces much more complicated than the set $\mathbb{R}$.

VIII.3.2 Definition

Suppose that $g : S \to \mathbb{R}$ is a continuous function whose domain is a nonempty closed subset S of $\mathbb{R}$. The continuous extension $f : \mathbb{R} \to \mathbb{R}$ constructed from g by the 'linear interpolation' technique given above is called the **piecewise-linear extension of g to $\mathbb{R}$**. Likewise, if T is a subset of $\mathbb{R}$ such that $S \subseteq T$, then the restriction to T of the function f just described is called the **piecewise-linear extension of g to T**.

VIII.3.3 Examples

(1) Let $C \subseteq [0,1]$ denote the Cantor Ternary Set; see Definition (**??**). The set C is closed; see Example (**??**) (3)). There is a natural function $g : C \to \mathbb{R}$ given by the following rule:

If $x \in C$, express x as a one-free ternary decimal $0 \overset{(3)}{\cdot} d_1 d_2, \dots d_k \dots$ with all the ternary digits d_k either 0 or 2. For each k let $c_k = d_k/2$, so that for each k, c_k is either 0 or 1. Then $g(x)$ is the number with *binary* representation $0 \overset{(2)}{\cdot} c_1 c_2 \dots c_k \dots$.

The function g is continuous on the closed set C. Indeed, let $\varepsilon > 0$ be given, and let δ be a number of the form $1/3^m$, where m in $\mathbb{N}$ is large enough that $1/2^m < \varepsilon$. Let x and x' be numbers in C whose 1-free ternary representations are $x = \overset{(3)}{\cdot} d_1 d_2 \dots$ and $x' = \overset{(3)}{\cdot} d'_1 d'_2 \dots$. If $|x - x'| < \delta$ then one has $d_j = d'_j$ for $j = 1, 2, \dots m$; see Theorem (II.5.2). Next, for each j let $c_j = d_j/2$ and $c'_j = d_j/2$ as above. Then $c_j = c'_j$ for each $j = 1, 2, \dots m$, hence the numbers $g(x)$ and $g(x')$ have binary representations which agree for the first m binary digits. It is clear from this that $|g(x) - g(x')| \leq 1/2^m < \varepsilon$. The continuity of g on C follows.

Definition The piecewise-linear extension of g to the closed interval $[0,1]$ is called the **Cantor Function**. It is a simple exercise to prove that the Cantor function is monotonic up on $[0,1]$, and is constant on each of the 'middle-thirds' open intervals which are removed from $[0,1]$ to form the set C; see Remark (**??**).

(2) Let $S = \{x_0, x_1, x_2, \dots x_{m-1}, x_m\}$ be a finite (nonempty) set of points in $\mathbb{R}$, labelled in 'strictly increasing order'; that is, so that

$$x_0 < x_1 < x_2 < \dots < x_{m-1} < x_m.$$

Let $g : S \to \mathbb{R}$ be a function defined on S. Since S is finite, it is automatically closed in $\mathbb{R}$, and the function g is automatically continuous on S.

Let $h : [x_0, x_m] \to \mathbb{R}$ denote the piecewise-linear extension of g to the closed interval $T = [x_0, x_m]$. It is easy to characterize the function h geometrically. Indeed, the graph of h is the m-sided polygonal line in $\mathbb{R}^2$ whose sides are the line segments connecting the points $(x_{j-1}, g(x_{j-1}))$ to $(x_j, g(x_j))$ for each $j = 1, 2, \dots m$.

Note In light of the terminology introduced in Definition (**??**), one refers to the function h constructed here as the **piecewise linear function associated with** g.

Now consider the analogous 'extension problem' for a continuous function $g : S \to \mathbb{R}$ whose domain S is an *arbitrary* nonempty subset S of $\mathbb{R}$. In light of the previous theorem, it is clear that the real issue is whether g can be extended continuously to the closure $\overline{S}$ of S. Indeed, if it can be so extended, then the preceding theorem guarantees that g can be extended continuously to *every* set T such that $S \subseteq T \subseteq \mathbb{R}$. The next examples show that analysis when S is *not* closed is far from obvious.

VIII.3.4 Examples

(1) Let S be the half-open interval $(0, 1]$, and define $g : S \to \mathbb{R}$ by the rule $g(x) = 1/x$ for all x such that $0 < x \leq 1$. It is clear that g is continuous at each point of S. Nevertheless, g does *not* have a continuous extension to the closure $\overline{S} = [0, 1]$ of S. Indeed, if such a continuous extension $f : [0, 1] \to \mathbb{R}$ were to exist, it would have to be bounded (by the Extreme-Value Theorem), and thus any restriction of f – including g – would also have to be bounded. However, g is clearly unbounded, since $\lim_{x \searrow 0} 1/x = +\infty$.

(2) Consider the function $g : (0, 1] \to \mathbb{R}$ defined as follows:

(i) If $x = 1/(2m)$ for some m in $\mathbb{N}$ then $g(x) = 0$; and if $x = 1/(2m-1)$ for some m in $\mathbb{N}$ then $g(x) = 1$.

(ii) Suppose that x satisfies the inequalities $\dfrac{1}{2m} < x < \dfrac{1}{2m-1}$ for some m in $\mathbb{N}$. Express x in the form $(1-t)/(2m) + t/(2m-1)$ for a unique t in $(0, 1)$, and then set $g(x) = t$.

It is easy to check that g is continuous on $(0, 1]$. Even more, the function g is bounded on $(0, 1]$. Nevertheless, g does not have a continuous extension to $[0, 1]$. To see this, note that if f is such an extension, then one must have, in particular,

$$f(0) = \lim_{k \to \infty} f\left(\frac{1}{k}\right).$$

However $f(1/k) = g(1/k)$ for each k in $\mathbb{N}$, and half of the terms $g(1/k)$ equal 1 (namely, when k is odd), while half equal 0 (namely, when k is even). Thus, any such extension must have the impossible property that $f(0) = 0$ and $f(0) = 1$. In other words, no such extension exists.

(3) Suppose that $h : \to \mathbb{R}$ is a function which is continuous at each point of an open interval (a, b) in $\mathbb{R}$, and let c be a point of (a, b). Let $S = (a, b)\backslash\{c\} = (a, c) \cup (c, b)$, and define $g : S \to \mathbb{R}$ by the rule

$$g(x) = \frac{h(x) - h(c)}{x - c} \text{ for all } x \text{ in } S.$$

Notice that the fraction on the right side of this equation does not make sense when $x = c$, so it is natural to ask whether g has a continuous extension from S to (a, b). The analysis of this question is left as an exercise; but the main conclusion is that such an extension $f : (a, b) \to \mathbb{R}$ exists if, and only if, h is differentiable at c. Moreover, in this case one has

$$f(x) = \begin{cases} \dfrac{h(x) - h(c)}{x - c} & \text{if } x \in S; \\ \\ h'(c) & \text{if } x = c. \end{cases}$$

Remark Some authors reverse the thinking used in this example to give an alternate approach to the derivative. This approach is called the **Caratheodory Definition of the Derivative** in honor of the mathematician who introduced it.

It is easy to state a general criterion for when a continuous function on a set extends continuously to the closure of that set. In order to simplify the discussion, it helps to introduce some terminology.

VIII.3.5 Definition

Let $g : S \rightarrow \mathbb{R}$ be a real-valued function whose domain is a nonempty subset S of $\mathbb{R}$. One says that **the function g preserves the Cauchy-Sequence Property on S** provided that if $\xi = (x_1, x_2, \dots)$ is a Cauchy sequence of points in S, then $g \circ \xi = (g(x_1), g(x_2), \dots)$ is also a Cauchy sequence in $\mathbb{R}$.

Note: One often paraphrases the statement 'g preserves the Cauchy-Sequence Property on S' as '**g preserves Cauchy sequences on S**' or '**g maps Cauchy sequences in S to Cauchy sequences in $\mathbb{R}$**'.

VIII.3.6 Examples

(1) In Examples (VIII.3.4) (1) and (2) it is easy to show that the sequence $g(1), g(1/2), \dots g(1/k), \dots$ is not Cauchy. Thus it is not the case that g preserves Cauchy sequences on S.

(2) Suppose that $g : S \rightarrow \mathbb{R}$ is continuous on a closed nonempty subset S of $\mathbb{R}$. Then g preserves Cauchy sequences on S.

Indeed, suppose that $\xi = (x_1, x_2, \dots)$ is a Cauchy sequence in S. By Theorem (III.5.5) ('The Cauchy Criterion'), the sequence ξ converges to some number c in $\mathbb{R}$. By the hypothesis that the set S is closed, the limit of the sequence ξ must be in the set S. By the hypothesis that g is continuous at each point of S it follows that $\lim_{k \to \infty} g(x_k) = g(c)$. By the Cauchy Criterion (again) it follows that the sequence $(g(x_1), g(x_2), \dots)$ is also a Cauchy sequence, as required.

(3) Suppose that $g : I \rightarrow \mathbb{R}$ is differentiable at each point of an open interval I. Assume further that g' is bounded on I; more precisely, assume there exists $M > 0$ in $\mathbb{R}$ such that $|g'(x)| \leq M$ for all x in I. Then g preserves Cauchy sequences on I.

To see this, let $\xi = (x_1, x_2, \dots)$ be a Cauchy sequence in I. If $\varepsilon > 0$ is given, let B be large enough that if $j, k \geq B$ then $|x_k - x_j| < \varepsilon/M$. It follows from the Lagrange Mean-Value Theorem that if $j, k \geq B$ then there exists z_{jk} in I such that

$$|g(x_k) - g(x_j)| = |g'(x_{jk})| \cdot |x_k - x_j| \leq M \cdot \left(\frac{\varepsilon}{M}\right) = \varepsilon.$$

Thus, the sequence $(g(x_1), g(x_2), \dots)$ is also Cauchy, as required.

The following is the analog, for the preceding definition, of Theorem (IV.2.6).

VIII.3.7 Theorem

Let $g : S \rightarrow \mathbb{R}$ be a function whose domain is a nonempty subset S of $\mathbb{R}$. A necessary and sufficient condition for g to preserve the Cauchy-Sequence Property on S is that if $\xi = (x_1, x_2, \dots)$ is a bounded monotonic sequence in S then $g \circ \xi = (g(x_1), g(x_2), \dots)$ is a Cauchy sequence in $\mathbb{R}$.

One can prove this result using an argument similar to that used in the proof of Theorem (IV.2.6) above. The details are left as an exercise.

Notice that in the previous theorem, as in Definition (VIII.3.5), we do *not* assume that the function g is continuous on S. The next result shows that this continuity occurs automatically.

VIII.3.8 Theorem

Let S be a nonempty suubset of $\mathbb{R}$, and suppose that $g : S \to \mathbb{R}$ preserves the Cauchy-Sequence Property on S, in the sense of Definition (VIII.3.5) above. Then g is continuous on S.

Proof Let c be a point of S, and let $\xi = (x_1, x_2, \dots)$ be any sequence in S converging to c. Let $\tau = (x_1, c, x_2, c, \dots)$ be the sequence formed by inserting a c between each term of the sequence ξ. It follows from Theorem (III.2.14), the 'Odd/Even Convergence Theorem', that the sequence τ also converges to c, and thus it must be a Cauchy sequence in S. By the hypothesis that g preserves Cauchy sequences on S it then follows that the sequence $g \circ \tau = (g(x_1), g(c), g(x_2), g(c), \dots)$ is a Cauchy sequence in $\mathbb{R}$, and thus must be convergent to some L in $\mathbb{R}$. It then follows that the subsequence of even-order terms, namely $(g(c), g(c), \dots)$ converges to L, from which one gets $L = g(c)$. But it also follows from the convergence of $g \circ \tau$ to L that the subsequence of odd-order terms also converges to $g(c)$. That is,

$$\lim_{j \to \infty} g(x_j) = L = g(c).$$

Thus g is continuous at c, as claimed.

VIII.3.9 Theorem (Continuus Extensions and Cauchy Sequences)

Suppose that $g : S \to \mathbb{R}$ is a function whose domain is a nonempty set $S \subseteq \mathbb{R}$. Note that we do *not* assume that g is continuous on S.

(a) If the function g has a continuous extension to the closure $\overline{S}$ of S, then this extension is unique.

(b) A necessary and sufficient condition for g to have a continuous extension to the closure $\overline{S}$ of S is that g preserve the Cauchy-Sequence Property on S, in the sense of Definition (VIII.3.5) above.

(c) More generally, a necessary and sufficient condition for g to have a continuous extension to the closure $\overline{S}$ of S is that g preserve the Cauchy-Sequence Property on every *bounded* nonempty subset of S.

Proof

(a) Suppose that $f : \overline{S} \to \mathbb{R}$ is a continuous extension of g to $\overline{S}$. Let c be a point of $\overline{S}$, and let $\xi = (x_1, x_2, \dots)$ be a sequence of points in S such that $c = \lim_{k \to \infty} x_k$. (Such a sequence must exist because of Theorem (**??**).) Then one must have

$$f(c) = \lim_{k \to \infty} f(x_k) = \lim_{k \to \infty} g(x_k). \tag{VIII.5}$$

The first equation reflects the fact that f is continuous at c; the second equation reflects the fact that f is an extension of g. The fact that $f(c) = \lim_{k \to \infty} g(x_k)$ then implies that the value of f

at c is completely determined by the original function g. In particular, there can be only one such function f, as claimed.

(b) Suppose first that g has a continuous extension f to $\overline{S}$. Let $\xi = (x_1, x_2, \dots)$ be a Cauchy sequence in S. Then ξ converges to some number c in $\mathbb{R}$, and this point c is an element of $\overline{S}$. Thus, by the continuity of f, one has $\lim_{k\to\infty} f(x_k) = f(c)$. In particular, the sequence $f\circ\xi = (f(x_1), f(x_2), \dots)$ is convergent, hence it is a Cauchy sequence. But $f(x_k) = g(x_k)$ for each k, since f is an extension of g from S to $\overline{S}$ and each $x_k \in S$. Thus g preserves the Cauchy-Sequence Property on S.

Conversely, suppose g preserves the Cauchy-Sequence Property on S. Note that, by Theorem (VIII.3.8), it follows that g is continuous on S. Let c be a point of $\overline{S}$, and let $\xi = (x_1, x_2, \dots)$ and $\zeta = (z_1, z_2, \dots)$ be sequences in S converging to c. Then, by Theorem (III.2.14) (the 'Odd/Even Convergence Theorem'), the sequence $\tau = (x_1, z_1, x_2, z_2, \dots)$ also converges to c. In particular, τ is a Cauchy sequence in S, hence (by the hypothesis that g preserves Cauchy sequences) $g\circ\tau = (g(x_1), g(z_1), g(x_2), g(z_2), \dots)$ is a Cauchy sequence in $\mathbb{R}$. Thus the sequence $g\circ\tau$ converges to some number L. By Theorem (III.2.1) (b) it follows that the sequences $g\circ\xi$ and $g\circ\zeta$ also converge to L.

Now define $f(c)$ to be $\lim_{k\to\infty} g(x_k)$, where $\xi = (x_1, x_2, \dots)$ is any sequence in S converging to c. The preceding discussion shows that the value $f(c)$ depends only on c, and not on the choice of sequence ξ. In other words, this process determines a function $f : \overline{S} \to \mathbb{R}$.

Finally, suppose that the function f just constructed is *not* continuous at some point c of $\overline{S}$. Then there exists $\varepsilon_0 > 0$ so that for each k in $\mathbb{N}$ there exists z_k in $\overline{S}$ such that $|c - z_k| < 1/(2k)$ but $|f(c) - f(z_k)| \geq \varepsilon_0$. Since $z_k \in \overline{S}$, it follows from the definition of f above that there must exist x_k in S such that $|z_k - x_k| < 1/(2k)$ and $|f(z_k) - g(x_k)| < \varepsilon_0/2$. Now use the Triangle Inequality to obtain

$$|c - x_k| \leq |c - z_k| + |z_k - x_k| < \frac{1}{2k} + \frac{1}{2k} = \frac{1}{k}$$

In particular, one has $\lim_{k\to\infty} x_k = c$; thus, by the construction of f, one also has $\lim_{k\to\infty} g(x_k) = f(c)$. Thus, there exists a number B such that if $k \geq B$ then $|f(c) - g(x_k)| < \varepsilon_0/2$. Now use the Triangle Inequality to get, for $k \geq B$,

$$|f(c) - f(z_k)| \leq |f(c) - g(x_k)| + |g(x_k) - f(z_k)| < \frac{\varepsilon_0}{2} + \frac{\varepsilon_0}{2} = \varepsilon_0.$$

This contradicts the defining condition on the numbers z_k, namely that $|f(c) - f(z_k)| \geq \varepsilon_0$ for *all* indices k.

Since assuming that f fails to be continuous at some point of $\overline{S}$ leads to a contradiction, it follows that f is continuous at *each* point of $\overline{S}$.

(c) Suppose that g has a continuous extension to $\overline{S}$, and let W be any nonempty bounded subset of S. If ξ is a Cauchy sequence in W then it is clearly a Cauchy sequence in S, so by Part (b) $g\circ\xi$ is a Cauchy sequence in $\mathbb{R}$. Thus g preserves the Cauchy-Sequence Property on W.

Conversely, suppose that g preserves the Cauchy-Sequence Property on W for every nonempty bounded subset W of S. Let $\xi = (x_1, x_2, \dots)$ be a Cauchy sequence in S. Since ξ is convergent it must be bounded; see Theorem (III.2.1) (c). That is, there exists a real number $M > 0$ such that $-M \leq |x_k| \leq M$ for all k in $\mathbb{N}$. Let $W = S \cap [-M, M]$. Then W is a bounded subset of S which contains the terms x_k of the sequence ξ. By the hypothesis that g preserves the Cauchy-Sequence Property on bounded nonempty subsets of S, it follows that $g\circ\xi$ is a Cauchy sequence. Since ξ can be any Cauchy sequence in S, it follows that g preserves the Cauchy-Sequence Property on S. It then follows from Part (b) that g has a continuous extension to $\overline{S}$, as claimed.

VIII.3.10 Important Example – Exponential Functions

Let b be a real number such that $b > 1$. Let $S = \mathbb{Q}$, the set of all rational numbers; note that $\overline{S}$, the closure of the set of rational numbers, is the set $\mathbb{R}$.

It is known that for every rational number r there is a well-defined number b^r; see Example (IV.4.8). Thus, for this fixed b, let $g_b : \mathbb{Q} \to \mathbb{R}$ be given by the rule

$$g_b(x) = b^x \text{ for all } x \text{ in } \mathbb{Q}.$$

Let $\rho = (r_1, r_2, \dots)$ be a bounded monotonic sequence of rational numbers; to be definition, assume ρ is monotonic-up and bounded above by B. Then, by the usual 'order properties' of powers the corresponding sequence of values, $g_b \circ \rho = (b^{r_1}, b^{r_2}, \dots)$ is also monotonic up and bounded above by b^B. In particular, by the Monotonic-Sequences Principle, the sequence $g_b \circ \rho$ is convergent, hence Cauchy. A similar argument shows that if ρ is monotonic down and bounded below, then $g_b \circ \rho$ is Cauchy. It then follows from Theorem (VIII.3.7) that g_b preserves the Cauchy-Sequence Property on $S = \mathbb{Q}$. Thus by Theorem (VIII.3.9) the function g_b has a unique continuous extension $f_b : \mathbb{R} \to \mathbb{R}$.

VIII.3.11 Definition

Let b be a number such that $b > 1$, and let $g_b : \mathbb{Q} \to \mathbb{R}$ and $f_b : \mathbb{R} \to \mathbb{R}$ be as in the preceding example. Then one calls the function f_b the **exponential function with base** b. The 'modern' notation for this function is $\exp_b$; the old-fashioned 'variables' notation for this function is $y = b^x$.

If $0 < b < 1$ then one defines b^x to mean $\left(\dfrac{1}{b}\right)^x$; and one sets $1^x = 1$.

Note: It is an easy exercise to show that the exponential functions satisfy the usual 'laws for exponents':

$$b^x \cdot b^y = b^{x+y}; \quad (b^x)^y = b^{xy}; \quad b^{-x} = \frac{1}{b^x};$$

and so on. Using the 'exp' notation, this becomes

$$\exp_b(x+y) = \exp_b(x) \cdot \exp_b(y); \quad \exp_{\exp_b x} y = \exp_b(xy); \quad \exp_b(-x) = \frac{1}{\exp_b(x)}.$$

Theorem (VIII.3.9) reduces the problem of the existence of continuous extensions to the question of whether a function on a bounded set preserves the Cauchy-Sequence Property. The next result provides an alternate approach that does not mention Cauchy sequences.

VIII.3.12 Theorem (Continuous Extensions and Uniform Continuity)

Suppose that $g : S \to \mathbb{R}$ is a function defined on a *bounded* nonempty subset S of $\mathbb{R}$. The following statements are equivalent:

Statement (1) The function g has a continuous extension to the closure $\overline{S}$ of S.

Statement (2) The function g is uniformly continuous on S.

Proof

Suppose that Statement (1) is true. If Statement (2) were *not* true, there would exist $\varepsilon_0 > 0$ with the following property: for each k in $\mathbb{N}$ there exist points t_k and z_k in S such that $|t_k - z_k| < 1/k$

but $|g(t_k)-g(z_k)| \geq \varepsilon_0$. Let $\tau = (t_1, t_2, \dots)$ and $\zeta = (z_1, z_2, \dots)$ be the corresponding sequences. Since, by hypothesis, S is a bounded set, it would then follow from the Bolzano-Weierstrass Theorem (i.e., Theorem (III.4.1)) that the sequence τ) has a convergent subsequence $\rho = (r_1, r_2, \dots)$. Let c be the limit of the subsequence ρ; clearly c would have to be an element of $\overline{S}$. Let A be an infinite subset of $\mathbb{N}$ such that ρ is the subsequence of τ corresponding to A, and let $\sigma = (s_1, s_2, \dots)$ be the subsequence of ζ corresponding to the same set A. The condition $|t_k - z_k| < 1/k$ would imply that $|r_j - s_j| < 1/j$ for each j in $\mathbb{N}$, and thus that σ must also converge to c. It then follows from the results of Theorem (VIII.3.9) that

$$f(c) = \lim_{j \to \infty} g(r_j) = \lim_{j \to \infty} g(s_j) \qquad (*)$$

In contrast, the inequality $|g(t_k) - g(z_k)| \geq \varepsilon_0$ for all k would imply $|g(r_j) - g(s_j)| \geq \varepsilon_0$ for all j, contradicting Equation $(*)$ above.

Since assuming that Statement (1) is true but Statement (2) is false leads to a contradiction, then Statement (1) implies Statement (2), as claimed.

Conversely, suppose that Statement (2) is true, and let $\xi = (x_1, x_2, \dots)$ be a Cauchy sequence in S. Let $\varepsilon > 0$ be given. Then, by Statement (2), there exists $\delta > 0$ so that if y, z in S satisfy $|y - z| < \delta$ then $|g(y) - g(z)| < \varepsilon$. But by the definition of 'Cauchy sequence', there exists B so that if $j, k \geq B$ then $|x_j - x_k| < \delta$. Combine these facts to conclude that if $j, k \geq B$ then $|g(x_j) - g(x_k)| < \varepsilon$. Thus, the sequence $g \circ \xi = (g(x_1), g(x_2), \dots)$ is a Cauchy sequence in $\mathbb{R}$. Thus, Statement (2) implies Statement (1), as claimed.

VIII.3.13 Remarks

(1) The proof above that Statement (2) implies Statement (1) does not use the hypothesis that S is a bounded set.

In contrast, the proof that Statement (1) implies Statement (2) definitely needs that hypothesis. For example, let $g : \mathbb{Q} \to \mathbb{R}$ be given by the formula $g(r) = r^2$ for every rational number r. This function certainly has a continuous extension to $\overline{\mathbb{Q}} = \mathbb{R}$, namely the squaring function $f(x) = x^2$ for all x in $\mathbb{R}$; that is, Statement (1) holds in this case. However, Statement (2) does not hold; see Example (**??**) (2).

VIII.4 Open Sets and Continuity in R

There is a beautiful characterization of the concept of 'continuity' which does not mention 'limit' or 'convergence'; neither does it mention 'ε' or 'δ'. In fact, it formulates the concept entirely in terms of 'open sets'.

VIII.4.1 Theorem

Let $f : X \to \mathbb{R}$ be a real-valued function whose domain is a nonempty subset X of $\mathbb{R}$. A necessary and sufficient condition for f to be continuous on X is this:

Condition O For every open set V in $\mathbb{R}$ the corresponding inverse image $f^{-1}[V]$ is of the form $X \cap U$ for some open set U in $\mathbb{R}$.

Proof Suppose that Condition O holds. Let x be a point of X and let $\varepsilon > 0$ be given. If one sets $V = (f(x)-\varepsilon, f(x)+\varepsilon)$, then V is an open set in $\mathbb{R}$ – indeed, it is an open interval – and thus (by Condition O) there exists an open set U in $\mathbb{R}$ such that $f^{-1}[V] = U \cap X$. In particular, since $f(x) \in V$ by construction, it follows that $x \in U \cap X$ and thus $x \in U$. By definition of U being 'open', it then follows that there exists $\delta > 0$ such that the open interval $(x-\delta, x+\delta)$ is a subset of U, and thus

$$(x-\delta, x+\delta) \cap X \subseteq U \cap X = f^{-1}[V].$$

That is, if $y \in X$ and $|y-x| < \delta$, then $f(y) \in V$ and thus $|f(y)-f(x)| < \varepsilon$. Thus, f is continuous at x; and since this works for every x in X, it follows that f is continuous on X.

Conversely, suppose that f is continuous on X. Let V be an open subset of $\mathbb{R}$, and let x be a point of $f^{-1}[V]$, so that $f(x) \in V$. By the definition of V being open, it follows that there exists $\varepsilon_x > 0$, depending on x, such that $(f(x)-\varepsilon_x, f(x)+\varepsilon_x) \subseteq V$. It then follows, by the continuity of f, that there is a corresponding $\delta_x > 0$ such that if $y \in X$ and $|y-x| < \delta_x$, then $|f(y)-f(x)| < \varepsilon_x$; in particular, $f(y) \in V$, hence $y \in f^{-1}[V]$. Let $U_x = (x-\delta_x, x+\delta_x)$. Then U_x is an open set in $\mathbb{R}$ and $U_x \cap X \subseteq f^{-1}[V]$. Let $U = \bigcup_{x \in f^{-1}[V]} U_x$. Then U is an open subset of $\mathbb{R}$, since each set U_x is an open interval; and $U \subseteq f^{-1}[V]$, since the same is true for each set $U_x \cap X$. Finally, $f^{-1}[V] \subseteq U \cap X$, since if $x \in f^{-1}[V]$ then $x \in U_x$ and (of course) $x \in X$, hence $x \in U \cap X$. Thus, $f^{-1}[V] = U \cap X$, with U open in $\mathbb{R}$, and the desired result follows.

VIII.4.2 Remark

The argument given above can be modified easily to characterize the continuity of a function at a single point in terms of inverse images of open sets; the details are left as an exercise.

The preceding theorem may, at first, appear to be just a curiosity. In fact, it may seem almost an annoyance: 'Do we really need yet another characterization of continuity?' If anything, this characterization seems even *less* useful than the sequential and $\varepsilon\delta$ characterizations of continuity, in that it obscures the fundamental idea that a small change in the input of the function causes a small change in the corresponding output.

However, it often happens in the evolution of mathematics that looking at a subject in a new way leads to far-reaching generalizations. That is the situation here. In fact, the branch of mathematics called 'General Topology' (or 'Point-set Topology') is based on the concept of open sets. Then the preceding theorem can be – and is – used as the basis for the concept of 'continuity' in this subject. Likewise, the concept of 'closed subset' can be defined in terms of open sets; see Corollary (**??**).

Although a treatment of 'point-set topology' is well outside the scope of these *Notes*, it does make sense to seek other concepts in $\mathbb{R}$ which can be characterized completely in terms of the open subsets of $\mathbb{R}$. The next couple of results do just that. To simplify the phrasing, however, it is useful to introduce some terminology.

VIII.4.3 Definition

(1) Let $\mathcal{F}$ be a family of sets. One says that the family $\mathcal{F}$ **covers a set** X, or that $\mathcal{F}$ **is a covering of** X, provided that each element x in X is an element of at least one of the sets in the family $\mathcal{F}$. Equivalent formulation: $X \subseteq \bigcup \mathcal{F}$.

(b) If, in addition, each set in the family $\mathcal{F}$ is an open subset of $\mathbb{R}$, then one says that $\mathcal{F}$ is an **open cover of** X.

Remark The sets in the family need not be mutually disjoint; indeed, in most cases of interest these sets can have considerable overlap. Also, the union of the sets in the family $\mathcal{F}$ can be – and often is – strictly bigger than X.

VIII.4.4 Theorem (The Heine-Borel Theorem)

Let X be a nonempty subset of $\mathbb{R}$. Then the following conditions are equivalent:

(1) The set X is a compact subset of $\mathbb{R}$.

(2) If $\mathcal{F}$ is a family of open subsets of $\mathbb{R}$ which covers X, then there is a finite subfamily $\mathcal{F}'$ of $\mathcal{F}$ which also covers X.

Proof Suppose that X is compact. Then X is bounded, so $a = \inf X$ and $b = \sup X$ are finite; and X is closed, so a and b are elements of X. Clearly X is a subset of the closed interval $[a, b]$. Let $U_0 = [a, b]\backslash X$. Then it is easy to see that U_0 is an open subset of $\mathbb{R}$. Indeed, by Corollary (**??**) one knows that $\mathbb{R}\backslash X$ is an open set in $\mathbb{R}$, and by Theorem (**??**) one knows that this open set can be expressed as the disjoint union of open intervals in $\mathbb{R}$. Since a and b are in X, and thus not in $\mathbb{R}\backslash X$, two of the disjoint open intervals in question are $(-\infty, a)$ and $(b, +\infty)$. Since $\mathbb{R}\backslash X$ is the disjoint union of the sets $(-\infty, a)$, $(b, +\infty)$ and $[a, b]\backslash X = U_0$, it follows that U_0 is the union of the remaining disjoint intervals forming $\mathbb{R}\backslash X$. Thus, U_0 is open.

Now let $\mathcal{F}$ be an open cover of X, and let $\mathcal{F}_0 = \mathcal{F} \cup \{U_0\}$. Then $\mathcal{F}_0$ is an open cover of the interval $[a, b]$. Suppose that no finite subfamily of $\mathcal{F}_0$ is an open cover of $[a, b]$. Construct a bisection sequence I_1, $I_2, \ldots$ (see Definition (**??**)) as follows:

(i) $I_1 = [a, b]$.

(ii) If I_k has the property that no finite subfamily of $\mathcal{F}_0$ covers I_k, then the same must be true for at least one of the two halves of I_k; let I_{k+1} be one of those halves.

It follows from the Bisection Principle that the intersection of the intervals I_k is a singleton set $\{c\}$. Let U_c be an element of the family $\mathcal{F}_0$ such that $c \in U_c$. Let $\delta > 0$ be small enough that $(c - \delta, c + \delta) \subseteq U_c$. If k is large enough that $(b - a)/2^{k-1} < \delta$, then $I_k \subseteq (c - \delta, c + \delta) \subseteq U_c$, contrary to the hypothesis that no finite subfamily of $\mathcal{F}_0$ covers I_k.

Thus, there must be a finite subfamily $\mathcal{F}_0'$ of $\mathcal{F}_0$ which covers $[a, b]$. Without loss of generality, assume that $U_0 \in \mathcal{F}_0'$; indeed, adding one more set to a finite family of sets still gives a finite family, and its union still covers $[a, b]$. However, $X = [a, b]\backslash U_0$, so the union of the sets in the smaller subfamily $\mathcal{F}' = \mathcal{F}_0'\backslash\{U_0\}$ also covers X, since U_0 covers no points of X. That is, Condition (1) implies Condition (2).

Conversely, suppose that Condition (2) holds. For each x in X let $U_x = (x - 1, x + 1)$. Then clearly the family $\mathcal{F} = \{U_x : x \in X\}$ is an open cover of X. Since Condition (2) holds, there must be a finite subfamily $\{U_{x_1}, U_{x_2}, \ldots U_{x_m}\}$ which covers X. It follows that if $x \in X$ then $x \in U_{x_j}$ for some $j = 1, 2, \ldots m$, and thus $|x| \leq \max\{|x_1| + 1, |x_2| + 1, \ldots |x_m| + 1\}$; in particular, X is bounded.

Next, suppose that X is not closed, and let $\xi = (x_1, x_2, \ldots)$ be a sequence in X which converges to some number c not in X. Then it is clear that the set $Y = \{c, x_1, x_2, \ldots x_k, \ldots\}$ is closed in $\mathbb{R}$. Let $U_0 = \mathbb{R}\backslash Y$, so that U_0 is an open subset of $\mathbb{R}$. Also, for each k in $\mathbb{N}$ let $\delta_k = |c - x_k|/2$. Note that $\delta_k > 0$ (since c is not in X but x_k is), and that $\lim_{k\to\infty} \delta_k = 0$. Now let $U_k = (x_k - \delta_k, x_k + \delta_k)$ for each k in $\mathbb{N}$, and let $\mathcal{F} = \{U_0, U_1, \ldots U_k, \ldots\}$. Clearly the family $\mathcal{F}$ covers X; thus by Condition (2) there is a finite subfamily $\mathcal{F}'$ with the same property. Without lose of generality one may assume that there is an index m such that $\mathcal{F}' = \{U_0, U_1, \ldots U_m\}$. (Indeed, if the sets U_{j_1}, U_{j_2}, $\ldots$ U_{j_n} together form a cover of X, simply let m be the largest of the indices

$j_1, \dots j_n$.) If k is large enough that $\delta_k < \delta_j/2$ for each $j = 1, 2, \dots m$, then

$$|c - x_k| = 2\delta_k \leq \delta_j \text{ for each } j = 1, 2, \dots m.$$

Thus, by the Modified Triangle Inequality one has

$$|x_j - x_k| = |(x_j - c) + (c - x_k)| \geq ||x_j - c| - |x_k - c|| \geq 2\delta_j - \delta_j = \delta_j.$$

In particular, x_k is not in U_j for $j = 1, 2, \dots m$. Since x_k is also not in U_0, it follows that X is not the union of the sets U_0, $U_1, \dots U_m$, contrary to Condition (2).

One can also characterize the convex sets in $\mathbb{R}$ in terms of open sets.

VIII.4.5 Theorem

Let X be a nonempty subset of $\mathbb{R}$. Then the following statements are equivalent:

(1) The set X is a convex subset of $\mathbb{R}$.

(2) For every pair of open sets U and V in $\mathbb{R}$ such that $U \cup V \supseteq X$, $U \cap X \neq \emptyset$ and $V \cap X \neq \emptyset$, one has $U \cap V \neq \emptyset$.

The simple proof is left to the reader.

VIII.5 Miscellaneous Results

Example (VIII.2.2) (6) shows that it is possible for a sequence $(f_1, f_2, \dots)$ of continuous functions to converge pointwise to a continuous function f on a compact set, but without the convergence being uniform on that set. A key feature of that example is the sudden appearance – and disappearance – of 'bumps' at which the the sequence stops moving steadily towards the limit function, at least for a while. The next result shows that something like those 'bumps' would be needed in any other example of this phenomenon.

VIII.5.1 Theorem (Dini's Uniform-Convergence Theorem)

Suppose that $\varphi = (f_1, f_2, \dots f_k, \dots)$ is a sequence of continuous functions which converges pointwise on a nonempty compact set X to a continuous function f.

(a) Suppose, in addition, that this convergence is 'monotonic up' on X, in the sense that for each x in X the numerical sequence $\varphi(x) = (f_1(x), f_2(x), \dots)$ is monotonic up. Then the convergence of the sequence φ to f is uniform on X.

(b) Likeswise, if the convergence is monotonic down, then φ converges uniformly to f on X.

The straight-forward proof is left as an exercise.

VIII.5.2 Theorem

Suppose that X is a compact nonempty subset of $\mathbb{R}$ and that $f : X \to \mathbb{R}$ is a continuous function with domain X. Let $Y = f[X]$ denote the image of the function f. Then:

(a) The set Y is compact.

(b) If, in addition, the function f is one-to-one on X, then the inverse function $f^{-1} : Y \to X$ is continuous on Y.

The proof is left as an exercise.

Many of the most fruitful methods for solving equations of various sorts, in both pure and applied mathematics, are based on the concept of a 'fixed point'.

VIII.5.3 Definition

Suppose that $f : X \to X$ is a function defined on a set X such that all the values of f are also in X. A point c in X is said to be a **fixed point of** f provided $f(c) = c$.

VIII.5.4 Example

Let C be a fixed positive number, and define $f : (0, +\infty) = (0, +\infty)$ by the rule

$$f(x) = \frac{1}{2}\left(x + \frac{C}{x}\right).$$

A fixed point for this function would be a number $x > 0$ such that $f(x) = x$; that is,

$$\frac{1}{2}\left(x + \frac{C}{x}\right) = x$$

Multiply both sides by $2x$ to get the equivalent condition $x^2 + C = 2x^2$, i.e., $x^2 = C$. That is, seeking a fixed point for this function is the same as seeking the positive square root of the given number C.

In Theorem (III.2.13) Heron's Method instructs one to choose an initial approximation $x_1 > 0$ of $\sqrt{C}$, and use it to form the sequence of approximations $(x_1, x_2, \ldots)$, where $x_{k+1} = \frac{1}{2}\left(x_k + \frac{C}{x_k}\right)$, $k = 1, 2, \ldots$. In terms of the function f in the preceding example, this recursive formula can be written $x_{k+1} = f(x_k)$; Theorem (III.2.13) then guarantees that this sequence converges to a fixed point of f. The existence of examples such as this have led to an extensive theory of fixed points. The next result illustrates that theory.

VIII.5.5 Theorem (The Banach Fixed-Point Theorem in $\mathbb{R}$)

Suppose that $f : X \to X$ is a continuous function defined on a nonempty closed set in $\mathbb{R}$, and whose values are also in X. Assume in addition that the following condition is satisfied:

There is a constant λ, with $0 \le \lambda < 1$, such that $|f(y) - f(x)| \le \lambda|y - x|$ for all x, y in X $(*)$

Then f has a unique fixed point c in X. More precisely, let x_0 be any number in X, and let $\xi = (x_0, x_1, x_2, \ldots)$ be the sequence defined recursively by the rule $x_k = f(x_{k-1})$. Then $c = \lim_{k \to \infty} x_k$.

Proof Note that, by Inequality $(*)$, for each $n \ge 1$ one has

$$|x_{n+1} - x_n| = |f(x_n) - f(x_{n-1})| \le \lambda|x_n - x_{n-1}|.$$

By repeatedly using this fact one finally gets

$$|x_{n+1} - x_n| \leq \lambda^n |x_1 - x_0|.$$

Then for each k in $\mathbb{N}$ one has

$$|x_{n+k+1}-x_n| \leq |x_{n+k+1}-x_{n+k}|+|x_{n+k}-x_{n+k-1}|+\ldots+|x_{n+1}-x_n| \leq \left(\lambda^{n+k} + \lambda^{n+k-1} + \ldots + \lambda^n\right)|x_1-x_0|.$$

That is, by Part (a) of Theorem (II.2.16),

$$|x_{n+k+1} - x_n| \leq \frac{\lambda^{n+k+1}}{1-\lambda}|x_1 - x_0|.$$

Since $|\lambda| < 1$ it follows that $\lim_{n\to\infty} \lambda^{n+k+1} = 0$, it follows easily that the sequence ξ is a Cauchy sequence, and thus converges to some number c. Since $x_n \in X$ for each n, and X is closed by hypothesis, it follows that c is also in X. In addition, since f is continuous, one has $\lim_{n\to\infty} f(x_n) = f(c)$. And since ξ converges to c, it follows that $\lim_{n\to\infty} x_{n+1} = c$. Combining all this with the recursive formula $x_{n+1} = f(x_n)$. one gets $c = f(c)$; that is, c is a fixed point of f.

Now suppose that c_1 and c_2 are both fixed points of f in X. Then

$$|c_2 - c_1| \leq |f(c_2) - f(c_1)| \leq \lambda|c_2 - c_1|.$$

Since $|\lambda| < 1$, it follows that $|c_2 - c_1| = 0$; that is, $c_2 = c_1$. In other words, the fixed point is unique, as claimed.

VIII.5.6 Example

Let $f(x) = \dfrac{1}{2}\left(x + \dfrac{2}{x}\right)$, and let $X = [1, 2]$. It is easy to show that f maps X into X. Also, note that $f'(x) = \dfrac{1}{2}\left(1 - \dfrac{2}{x^2}\right)$. It is clear that if $x \in [1, 2]$ then $|f'(x)| \leq 1/2$. Then it follows from the Mean-Value Theorem that $|f(y) - f(x)| \leq \frac{1}{2}|y - x|$ for all x, y in $[1, 2]$. Then the Banach Fixed-Point Theorem can be used to show that if $x_0 \in [1, 2]$ and if $x_{n+1} = x_n$ for each n in $\mathbb{N}$, then $\lim_{n\to\infty} x_n = \sqrt{2}$. (This is a special case of Heron's Method.)

VIII.5.7 Theorem (The Brouwer Fixed-Point Theorem in $\mathbb{R}$)

Suppose that $f : [a, b] \to [a, b]$ is a continuous function defined on a closed bounded interval $[a, b]$ with values in the same interval. Then f has at least one fixed point in $[a, b]$.

Proof Since f maps $[a, b]$ into $[a, b]$, it follows that $a \leq f(x) \leq b$ for all x in $[a, b]$. Thus one also has $a - x \leq f(x) - x \leq b - x$ for all such x. In particular, $0 \leq f(a) - a \leq b - a$ and $a - b \leq f(b) - b \leq 0$. In particular, let M be the maximum value of $f(x) - x$ for x in $[a, b]$ and let m be the corresponding minimum value. Then $m \leq 0$ and $M \geq 0$, so by the Intermediate-Value Theorem for Continuous Functions, there must exist a number c in $[a, b]$ such that $f(c) - c = 0$. This c is a fixed point of f in $[a, b]$.

VIII.5.8 Remarks

(1) Both the Banach Theorem and the Brouwer Theorem described here are very special cases of the full theorems that bear the same names.

(2) A function which satisfies Condition ($*$) in Theorem (VIII.5.5) is called a **contraction mapping on** X. For that reason, the Banach theorem is often called the **Contraction-Mapping Theorem** (or the **Contraction-Mapping Principle**).

VIII.6 Discontinuities of Functions

In this section we complement our previous study of continuous functions with the consideration of functions which *fail* to be continuous at some points of their domains. The discussion begins by providing a measure of 'how badly' a function is discontinuous at a point.

VIII.6.1 Definition (Oscillation over a Set)

Let $f : S \to \mathbb{R}$ be a *bounded* function defined on a nonempty set S of $\mathbb{R}$. The **oscillation of** f **over** S is the number $\Omega_S(f)$ given by the formula

$$\Omega_S(f) \,=\, \sup\{|f(y) - f(x)| : x, y{\in}S\}. \tag{VIII.6}$$

In the terminology of Definition (**??**), $\Omega_S(f)$ is the **diameter of the image set** $f[S]$.

Remark Let f and S be as above, and x be a point of the set S. Define $\varphi_{S,f,x} : (0, +\infty) \to \mathbb{R}$ by the rule

$$\varphi_{S,f,x}(r) \,=\, \Omega_{S\,\cap\,(x-r,x+r)}(f)$$

It follows from Theorem (**??**) that $\varphi_{S,f,x}(r) \geq 0$ for all $r > 0$, and that as r decreases so does $\varphi_{S,f,x}(r)$. In particular, $\lim_{r\searrow 0} \varphi_{S,f,x}(r)$ exists and is a nonnegative real number.

Note: Usually the set S, the function f and the point x under discussion are clear from the context. In such a case we normally write $\varphi(r)$ instead of the more precise $\varphi_{S,f,x}(r)$.

VIII.6.2 Definition (Oscillation at a Point)

Let f, S, x and φ be as in the preceding 'Remark'. Then the **oscillation of** f **at** x **with respect to** S is the number $\omega_S(f; x)$ given by

$$\omega_S(f; x) \,=\, \lim_{r\searrow 0} \varphi(r).$$

That is,

$$\omega_S(f; x) \,=\, \lim_{r\searrow 0} \Omega_{S\,\cap\,(x-r,x+r)}(f). \tag{VIII.7}$$

If the set S is clear from the context – for example, if S is the full domain of f – one sometimes abbreviates the notation to $\omega(f; x)$; and to emphasize that this quantity can be viewed as a 'function of x on S', it is often written $\omega_f(x)$.

VIII.6.3 Theorem

Suppose that $f : S \to \mathbb{R}$ is a bounded function whose domain is the nonempty set S. Let c be a point of S. Then a necessary and sufficient condition for f to be continuous at c is that $\omega_S(f;c) = 0$.

Proof Suppose that $\omega_S(f;c) = 0$. Let $r > 0$ be given. By definition of φ, if x in S satisfies $|x - c| < r$, then one has

$$|f(x) - f(c)| \leq \sup\{|f(y) - f(z)| : y, z \in S \cap (c - r, c + r)\} = \varphi(r) \qquad (*)$$

Now let $\varepsilon > 0$ be given. By the hypothesis that $\omega_S(f;c) = 0$, it follows that there exists $\delta > 0$ such that if $0 < r < \delta$ then $0 \leq \varphi(r) < \varepsilon$. For such δ one then gets, using Inequality $(*)$ above, $|f(x) - f(c)| < \varepsilon$ when $x \in S$ and $|x - c| < \delta$. Thus f is continuous at c, as claimed.

Conversely, suppose that f is continuous at c. Note that if x and y are in S, then

$$|f(y) - f(x)| = |f(y) - f(c) + f(c) - f(x)| \leq |f(y) - f(c)| + |f(c) - f(x)| \qquad (**)$$

Now let $\varepsilon > 0$ be given, and let $\delta > 0$ be small enough that if $z \in S$ and $|z - c| < \delta$, then $|f(z) - f(c)| < \varepsilon/4$. By Inequality $(**)$ above it follows that if $0 < r < \delta$ and $x, y \in S \cap (c - r, c + r)$, then

$$|f(y) - f(x)| \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

Since this inequality holds for *all* x and y in $S \cap (c - r, c + r)$, it follows that the supremum of the set of all such numbers $|f(x) - f(y)|$ is also no bigger than $\varepsilon/2$. That is, if $0 < r < \delta$ then

$$\varphi(r) \leq \frac{\varepsilon}{2} < \varepsilon.$$

It follows that $\lim_{r \searrow 0} \varphi(r) = 0$; that is, $\omega_S(f;c) = 0$, as claimed.

In light of the preceding theorem, it is natural to think of the quantity $\omega_S(f;c)$ as providing an informal 'measure' of the 'degree of discontinuity of f at c': the bigger this quantity, the 'more discontinuous at c' the function f is.

Because of Theorem (IV.5.11), the discontinuities of a monotonic function defined on an interval I are easy to classify; but first some more terminology.

VIII.6.4 Definition

Suppose that $f : I \to \mathbb{R}$ is defined on an interval I in $\mathbb{R}$, and that c is an interior point of I. Assume that the one-sided limits $\lim_{x \nearrow c} f(x)$ and $\lim_{x \searrow c} f(x)$ both exist and are finite. In light of Remark (**??**) (2), this can be stated, 'suppose that $f(c-)$ and $f(c+)$ exist'.

(1) If $f(c-) \neq f(c+)$ then one says that f has a **jump discontinuity at** c, or that f has a **discontinuity of Type 1 at** c. One refers to the quantity $f(c+) - f(c-)$ as the **jump of** f **at** c.

(2) If $f(c-) = f(c+)$ but this common value does not equal $f(c)$, then one says that f has a **removable discontinuity at** c.

(3) Similar terminology is used if $c \in I$ but c is an endpoint of I. Thus, if c is the left endpoint of I and $f(c+)$ exists, one says that f has a jump discontinuity at c provided $f(c+) \neq f(c)$; one then calls the quantity $f(c+) - f(c)$ the jump of f at c. Likewise, if c is the right endpoint of I and $f(c-)$ exists, one says that f has a jump discontinuity at c provided $f(c-) \neq f(c)$, and one then calls the difference $f(c) - f(c-)$ the jump of f at c.

(4) If $f(c-)$ and $f(c)$ both exist and are finite, then one also calls the difference $f(c) - f(c-)$ the **left-hand jump of f at c**. One sometimes denotes this jump by $\Delta_{-}f(c)$. Likewise, if $f(c+)$ and $f(c)$ both exist and are finite, then one calls $f(c+) - f(c)$ the **right-hand jump of f at c**, denoted $\Delta_{+}f(c)$.

Remarks

(1) Some texts use a slightly different definition of 'jump discontinuity', in that they allow the possibility that $f(c-) = f(c+)$ as long as $f(c)$ takes on a different value.

(2) The reason for the name 'removable discontinuity' is because by merely redefining the value of f at c to be the common value $f(c-) = f(c+)$, one can make f continuous at c.

VIII.6.5 Theorem

Suppose that $f : [a,b] \to \mathbb{R}$ is monotonic up on the closed bounded interval $[a,b]$. Then:

(a) Every discontinuity of f on $[a,b]$ is a jump discontinuity, and the jump of f at each such point is a positive number.

(b) Let A denote the set of discontinuities of f on $[a,b]$. If $x_1, x_2, \ldots x_k$ is any finite (nonempty) collection of points in A, with $a \leq x_1 < x_2 < \ldots < x_k \leq b$, then the sum of the jumps of f at these points is at most $f(b) - f(a)$.

(c) The set of discontinuities of f on $[a,b]$ is countable.

(d) A similar result holds for any function $g : [a,b] \to \mathbb{R}$ which is monotonic down on $[a,b]$, except that in this case the jumps are all negative numbers, and the sum of the jumps of g at any finite set of discontinuities is bounded below by $g(b) - g(a)$.

Proof

(a) This is essentially the content of Theorem (IV.5.11).

(b) Suppose, to be definite, that $a < x_1$ and $x_k < b$. Note that the sum S of the jumps at the points $x_1, \ldots x_k$ is given by

$$S = (f(x_1+) - f(x_1-)) + (f(x_2+) - f(x_2-)) + \ldots + (f(x_{k-1}+) - f(x_{k-1}-)) + (f(x_k+) - f(x_k-))$$

Form the following 'collapsing sum' $f(b) - f(a) =$

$$f(b) + (f(x_k+) - f(x_k+)) + (f(x_k-) - f(x_k-)) + (f(x_{k-1}+) - f(x_{k-1}+)) + (f(x_{k-1}-) - f(x_{k-1}-)) + \ldots +$$

$$+ \ldots + (f(x_1+) - f(x_1+)) + (f(x_1-) - f(x_1-)) - f(a).$$

After using the Generalized Associative and Commutative Laws for Addition to rearrange terms, one can rewite the preceding equation as

$$f(b) - f(a) = (f(b) - f(x_k+)) + (f(x_k+) - f(x_k-)) + (f(x_k-) - f(x_{k-1}+)) + + \ldots$$

$$+ \ldots + (f(x_2+) - f(x_2-)) + (f(x_2-) - f(x_1+)) + (f(x_1+) - f(x_1-)) + (f(x_1-) - f(a)) \qquad (*)$$

Note that the right side of this last equation consists of two types of terms:

(i) The jumps of f; that is, the terms of the form $f(x_j+) - f(x_j-)$ for $j = 1, 2, \ldots k$.

(ii) The 'nonjump' terms; that is, the terms of the form $f(x_j-) - f(x_{j-1}+)$ for $j = 2, 3, \ldots k$, as well as the two 'endpoint' terms $f(b) - f(x_k+)$ and $f(x_1-) - f(a)$.

It is clear from the hypothesis that f is monotonic up, combined with the fact that $x_{j-1} < x_j$ for each j, that all the terms of Type (ii) are nonnegative. Thus removing them from the right side of ($*$) produces a smaller quantity on the right. More precisely, one gets

$$f(b) - f(a) \geq f(x_k+) - f(x_k-) + (f(x_{k-1}+) - f(x_{k-1}-)) + \ldots + (f(x_1+) - f(x_1-)).$$

That is, $f(b) - f(a)$ is at least as large as the sum of the jumps of f at x_1, $x_2, \ldots x_k$, as claimed.

The reader is encouraged to made the minor changes needed in the preceding argument to handle the case in which f has a jump at either of the endpoints a or b.

(c) For each positive integer m let A_m denote the set of all discontinuities x of f on $[a, b]$ such that the jump of f at x is at least $1/m$. It follows by Part (b) that the set A_m cannot have N elements if N is a positive integer such that $N > m(f(b) - f(a))$. Indeed, the sum of N jumps, each at least $1/m$ in size, must be at least N/m, and if $N > m(f(b) - f(a))$, the sum of these jumps would have to exceed $f(b) - f(a)$, contrary to the conclusions of Part (b).

Finally, note that by Part (a), if x is a discontinuity of f on $[a, b]$ then the jump of f at x is positive. In particular, there must exist a positive integer m such that this jump is at least as large as $1/m$. In other words, the set D of all discontinuities of f on $[a, b]$ is the union of the countable family of sets A_1, $A_2, \ldots$, and each set in this family is countable (in fact, finite). Now Theorem (I.8.11) implies that the union D is itself a countable set, as claimed.

(d) The case of a function g which is monotonic down on $[a, b]$ can be reduced to the monotonic-up case just discussed: look at the function $f = -g$, which is monotonic up on $[a, b]$. As is usual in such reductions, the details are left as an exercise.

The preceding theorem tells us, in effect, that if $f : [a, b] \to \mathbb{R}$ is monotonic, then its discontinuities cannot be overly 'wild'. Indeed, the set of points at which the function f *is* continuous is an uncountable subset of $[a, b]$, since one obtains this set by removing the (at most countably many) discontinuities of f from $[a, b]$; see Theorem (**??**). Moreover, the discontinuities themselves cannot involve 'wild' fluctuations of the values of f, since they are all simple jump discontinuities.

It is also clear that any function formed, by any reasonable algebraic method, from a finite number of functions that are monotonic on $[a, b]$ need not be monotonic, yet also cannot be too wild. For instance, if f_1, $f_2, \ldots f_k$ are monotonic on $[a, b]$ and c_1, $c_2, \ldots c_k$ are constants, then the linear combination $g = c_1 f_1 + \ldots + c_k f_k$ has at most a countable number of discontinuities, and each of these discontinuities is, at worst, either a jump discontinuity or a removable discontinuity. Similarly, the product of finitely many monotonics, while not necessarily monotonic itself, has at worst a countable number of either jump or removable discontinuities.

With all this in mind, it makes sense to consider the following class of functions.

VIII.6.6 Definition

Let I be an interval in $\mathbb{R}$. A function $f : I \to \mathbb{R}$ is said to be of **hybrid type on the interval I** provided there exist functions $g : I \geq \mathbb{R}$ and $h : I \to \mathbb{R}$, both being monotonic *up* on I, such that $f(x) = g(x) - h(x)$ for all x in I.

The set of all functions f, with domain I, which are of hybrid type on I is denoted $\mathcal{H}_I$.

Remarks

(1) Since $-h$ is monotonic down when h is monotonic up, it would have been possible to phrase the definition as follows:

'The function f is of hybrid type if it can be expressed as the *sum* of a monotonic up function with a monotonic down function on I.'
Experience shows, however, that having g and h be of the same type of monotonicity, and using the minus sign, ultimately causes less confusion than omitting the minus sign but having g and h of opposite monotonicity. More importantly, expressing a function as the difference of monotonic-up functions is standard in the literature; while expressing them as the sum of a monotonic-up and a monotonic-down is not standard.

(2) The terminology of a 'hybrid type' of function, although quite reasonable, is not standard. We switch to the standard terminology later, when it makes more sense.

VIII.6.7 Examples

(1) Let $g(x) = 3x^2$ and $h(x) = 2x^3$. Then both functions are monotonic on the interval $I = [0, +\infty)$. Now let $f : [0, +\infty) \rightarrow \mathbb{R}$ be the corresponding hybrid function, given by $f(x) = g(x) - h(x) = 3x^2 - 2x^3$ for all $x \geq 0$. By using elementary calculus one easily sees that f is increasing on the interval $[0, 1]$ and decreasing on the interval $[1, +\infty)$; in other words, f is a true 'hybrid', in the sense of being neither monotonic up throughout the entire interval $[0, +\infty)$ nor monotonic down throughout.

(2) Suppose that $f : [0, +\infty) \rightarrow \mathbb{R}$ is a function with the following property: there exists a number $c > 0$ such that f is monotonic up on the subinterval $[0, c]$ and monotonic down on $[c, +\infty)$. Then f is of hybrid type on $[0, +\infty)$. Indeed, one can choose g and h so that h is constant on $[0, c]$ and g is constant on $[c, +\infty)$. More precisely, define g and h as follows:

$$g(t) = \begin{cases} f(t) & \text{for } 0 \leq t \leq c \\ f(c) & \text{for } c < t < +\infty \end{cases}$$

Likewise,

$$h(t) = \begin{cases} 0 & \text{for } 0 \leq t \leq c \\ f(c) - f(t) & \text{for } c < t < +\infty \end{cases}$$

It is clear that $f(t) = g(t) - h(t)$ for *all* t in $[0, +\infty)$, and that g and h are both monotonic up on $[0, +\infty)$. (Don't be fooled by the minus sign in the expression $f(c) - f(t)$ which is used for h on $[1, +\infty)$. On that portion of the domain the function f is monotonic *down*, hence $f(c) - f(t)$ is monotonic *up* on the same portion.) Thus, the function f is of hybrid type.

Remarks on This Example:

(i) The method illustrated here generalizes to any function which alternates between intervals on which it is monotonic up and intervals on which it is monotonic down; think 'sine' and 'cosine'. The idea in the general case is quite similar to what appears here, but the detailed calculations can be a bit messy. The general case is left as an exercise.

(ii) There is a simple 'physical' interpretation of the function f in this example. Namely, think of the input t for the quantity $f(t)$ as the time, and the corresponding output $x = f(t)$ as the location at time t of an object moving along the x-axis. Then the function g provides a record of of the *forward* motion of the object; that is, motion to the right along the x-axis. Indeed, for any time interval $[t_1, t_2]$, with $0 \leq t_1 < t_2$, the difference $g(t_2) - g(t_1)$ is the total distance the object moved to the right during that time interval. Likewise, $h(t_2) - h(t_1)$ is the total distance the object moved to the left during that time interval. Thus $f(t_2) - f(t_1) = (g(t_2) - g(t_1)) - (h(t_2) - h(t_1))$ is the *net* distance traveled by the object during the given time interval.

(3) Let $f : [0,+\infty) \to \mathbb{R}$ be the function discussed above in Example (1). From the results of that example, it is clear that the method used in Example (2), with $c = 1$, applies here as well. That method then yields the expression $f(x) = g(x) - h(x)$, where

$$g(x) = \begin{cases} 3x^2 - 2x^3 & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } 1 < x < +\infty \end{cases}$$

and

$$h(x) = \begin{cases} 0 & \text{for } 0 \leq x \leq 1 \\ 1 - (3x^2 - 2x^3) & \text{for } 1 < x < +\infty \end{cases}$$

(4) The Dirichlet function (see Example (IV.1.4) (1)) is *not* of hybrid type. Indeed, it has uncountably many discontinuities, whereas a function of hybrid type can have at most countably many such points.

(5) Suppose that $f : I \to \mathbb{R}$ is a function with bounded derivative on an interval I. Then f is a hybrid function in I. Indeed, let M be an upper bound for $|f'|$ on I, and let c be a point of I. Define $g, h : I \to \mathbb{R}$ by

$$g(x) = f(c) + M(x-c) \text{ and } h(x) = \left(D_c^{-1}(M - f')\right)(x) \text{ for all } x \text{ in } I.$$

Clearly $g'(x) = M = 0$ for all x in I, and $h'(x) = M - f'(x) = 0$ for all x in I. Thus g and h are certainly monotonic up on I. It is also easy to verify that $f(x) = g(x) - h(x)$ for all x in I. It now follows that f is of hybrid type on I.

> Historical Note The theory of hybrid functions was developed by Camille Jordan in an 1881 paper. It grew out of a theorem of Dirichlet which plays an important role in the development of the theory of Fourier series.
>
> More specifically, Dirichlet showed that if $g : [0,\delta] \to \mathbb{R}$ is continuous on an interval $[0,\delta]$, and if g is piecewise monotonic on $[0,\delta]$ (see Definition (**??**) (4)), then
>
> $$\lim_{p \to +\infty} \frac{2}{\pi} \int_0^{\delta} g(t) \frac{\sin(pt)}{t}\, dt = g(0+)$$
>
> Dirichlet remarked that the 'piecewise monotonicity' condition could be weakened and still yield this equation. Jordan showed, in fact, that g being a hybrid function on $[0,\delta]$ would suffice.

VIII.6.8 Remarks

(1) If $f : I \to \mathbb{R}$ can be expressed on I as the difference of monotonic functions, this expression is not unique. Indeed, note that if $f = g - h$, where g and h are monotonic on I, then $f = (\psi + g) - (\psi + h)$, where $\psi : I \to \mathbb{R}$ is any function that is monotonic up on I; clearly $\psi + g$ and $\psi + h$ are also monotonic up on I.

(2) The ambiguity in the choice of g and h in the decomposition $f = g - h$ can be reduced somewhat. Indeed, choose a point a in I, and note that if $f : I \to \mathbb{R}$ is a hybrid function then it is possible to find monotonic functions $\tilde{g} : I \to \mathbb{R}$ and $\tilde{h} : I \to \mathbb{R}$ such that $f(x) = f(a) + \hat{g}(x) - \tilde{h}(x)$

for all x in I, and $\tilde{g}(a) = \tilde{h}(a) = 0$. Indeed, express f as $f = g - h$, where g and h are monotonic up on I, and then set $\tilde{g}(x) = g(x) - g(a)$ and $\tilde{h}(x) = h(x) - h(a)$ for all x in I. Then

$$f(x) = f(a) + \tilde{g}(x) - \tilde{h}(x).$$

We shall say that such a decomposition of f is **based at the number** a.

Note: To simplify the notation later on, we shall normally transpose the term $f(a)$ to the left side and express the desired condition as

$$f(x) - f(a) = \tilde{g}(x) - \tilde{h}(x).$$

(3) In light of the preceding remarks, it is natural to ask whether every decomposition $f(x) - f(a) = \tilde{g}(x) - \tilde{h}(x)$ of the type considered in Remark (2) arises, from some fixed decomposition $f - f(a) = g_a - h_a$, by adding a suitable monotonic-up function ψ as in Remark (1). More precisely, do there exist monotonic-up functions $g_a, h_a : I \to \mathbb{R}$, satisfying $g_a(a) = h_a(a) = 0$ and $f - f(a) = g_a - h_a$, such that if $\tilde{g}$ and $\tilde{h}$ are as in Remark (2), then there exists monotonic-up $\psi : I \to \mathbb{R}$ such that $\tilde{g} = \psi + g_a$ and $\tilde{h} = \psi + h_a$? If such g_a and h_a exist, it is easy to see that $\psi(a) = 0$. Since ψ is monotonic up on I, one must then have, for each x in I,

$$\psi(x) \geq 0 \text{ if } x \geq a \text{ and } \psi(x) \leq 0 \text{ if } x \leq a.$$

Since $\tilde{g}(x) = \psi(x) + g_a(x)$ for all x in I, it follows that

$$\tilde{g}(x) \geq g_a(x) \text{ if } x \geq a \text{ and } \tilde{g}(x) \leq g_a(x) \text{ if } x \leq a.$$

These inequalities provide some information about the nature of the desired functions g_a and h_a, assuming that they actually exist. Namely, it is clear that, when $x \geq a$, $g_a(x)$ must be a lower bound of the set of numbers of the form $\tilde{g}(x)$; while when $x \leq a$, $g_a(x)$ must be an upper bound for the set of numbers of the form $g(\tilde{x})$. It then is reasonable to guess that perhaps $g_a(x)$ should be the *greatest* of the lower bounds of this set when $x \geq a$, and that $g_a(x)$ should be the *least* of the upper bounds of this set of numbers when $x \geq a$ but equal the *greatest* of the lower bounds when $x \leq a$. Similar remarks hold concerning $h_a(x)$ in relation to numbers of the form $\tilde{h}(x)$.

(4) There is an alternate viewpoint which leads to essentially the same conclusions. Namely, suppose that $\tilde{g}$ and $\tilde{h}$ could be expressed as $\tilde{g} = \psi + g_a$ and $\tilde{h} = \psi + h_a$ as in the preceding remark. Then for each x_1 and x_2 in I with $x_1 < x_2$ one would have

$$\tilde{g}(x_2) - \tilde{g}(x_1) = (\psi(x_2) - \psi(x_1)) + (g_a(x_2) - g_a(x_1)) \geq g_a(x_2) - g_a(x_1)$$

with the final inequality reflecting the fact that ψ is montonic up. Moreover, this final inequality would actually reduce to an equation if, and only if, ψ is constant on the interval $[x_1, x_2]$. Likewise, one would have

$$\tilde{h}(x_2) - \tilde{h}(x_1) = (\psi(x_2) - \psi(x_1)) + (h_a(x_2) - h_a(x_1)) \geq h_a(x_2) - h_a(x_1).$$

That is, the 'preferred' functions g_a and h_a would represent f as the difference of monotonic-up functions in an 'optimal' way, with a minimum of 'jumping'.

The next results shows that the ideas suggested in Remarks (3) and (4) above are valid.

VIII.6.9 Theorem (Jordan's Theorem for Functions of Hybrid Type)

Let I be an interval in $\mathbb{R}$. Suppose that $f : I \to \mathbb{R}$ is a function of hybrid type on the interval I. Then for each point a in I there is a unique pair of functions $g_a, h_a : I \to \mathbb{R}$ with the following properties:

(i) The functions g_a and h_a are monotonic up on I.

(ii) $f(x) - f(a) = g_a(x) - h_a(x)$ for all x in I, and $g_a(a) = h_a(a) = 0$.

(iii) Suppose that $g, h : I \to \mathbb{R}$ are monotonic-up functions on I such that $f(x) - f(a) = g(x) - h(x)$ for all x in I, and $g(a) = h(a) = 0$. Then $g(x) \geq g_a(x)$ and $h(x) \geq h_a(x)$ for all x in I such that $x \geq a$; likewise, $g(x) \leq g_a(x)$ and $h(x) \leq h_a(x)$ for all x in I such that $x < a$.

Proof Let $\mathcal{G}_{I;a}$ denote the set of all functions $g : I \to \mathbb{R}$ such that g and $g - f$ are monotonic up on I, and $g(a) = 0$.

Claim The set $\mathcal{G}_{I;a}$ is nonempty.

Proof of Claim The hypothesis that f is of hybrid type on I means that there exists a pair of functions $\hat{g}$ and $\hat{h}$ which are monotonic up on I such that $f = \hat{g} - \hat{h}$. Let $g = \hat{g} - \hat{g}(a)$ Clearly g is monotonic up on I and $g(a) = 0$. Also,

$$g - f = \hat{g} - \hat{g}(a) - f = (\hat{g} - \hat{g}(a)) - \left(\hat{g} - \hat{h}\right) = \hat{h} - \hat{g}(a).$$

Since $\hat{h}$ is monotonic up, it follows that $\hat{h} - \hat{g}(a)$, i.e., $g - f$, is also monotonic up. Thus, the function g constructed this way is an element of $\mathcal{G}_{I;a}$; in particular, the set $\mathcal{G}_{I;a}$ is nonempty, as claimed.

To construct the desired functions g_a and h_a, first note that if $g \in \mathcal{G}_{I;a}$ then, since g is monotonic up on I, one has $g(x) \geq g(a) = 0$ for all x in I such that $x \geq a$; likewise, $g(x) \leq 0$ for all x in I such that $x \leq a$.

Now define $g_a : I \to \mathbb{R}$ by the rule

$$g_a(x) = \begin{cases} \inf\{g(x) : g \in \mathcal{G}_{I;a} \text{ if } x \in I \text{ and } x \geq a\} \\ \sup\{g(x) : g \in \mathcal{G}_{I;a} \text{ if } x \in I \text{ and } x \leq a\}. \end{cases} \qquad \text{(VIII.8)}$$

Define $h_a : I \to \mathbb{R}$ to be the function given by

$$h_a(x) = g_a(x) - (f(x) - f(a)) \text{ for each } x \text{ in } I.$$

Remarks (1) The indicated 'sup' and 'inf' exist since $\mathcal{G}_{I;a} \neq \emptyset$.

(2) Equation (VIII.8) provides two different formulas for the number $g_a(a)$. However, the two formulas yield the same value. Indeed, by definition, if $g \in \mathcal{G}_{I;a}$ then $g(a) = 0$, and the supremum and infimum of the singleton set $\{0\}$both equal 0.

(3) Since each g in $\mathcal{G}_{I;a}$ is monotonic up on I, it then becomes clear that the infimum and supremum in Equation (VIII.8) are both finite.

Proof of Property (i) Suppose g_a is *not* monotonic up on I. Then there must exist numbers x_1 and x_2 in I, with $x_1 < x_2$, such that $g_a(x_1) > g_a(x_2)$.

Case 1 Suppose that $x_1 \geq a$, so that $a \leq x_1 < x_2$. By Equation (VIII.8) one then has $g_a(x_2) = \inf\{g(x_2) : g \in \mathcal{G}_{I;a}\}$. It then follows from the defining properties of 'infimum' (see Definition (**??**)) that there exists an element $\hat{g}$ in $\mathcal{G}_{I;a}$ such that $g_a(x_1) > \hat{g}(x_2) \geq g_a(x_2)$. Since $\hat{g} \in \mathcal{G}_{I;a}$, one knows that $\hat{g}$ is monotonic up on I and thus $\hat{g}(x_1) \leq \hat{g}(x_2)$. However, by the definition of $g_a(x_1)$ as the infimum of numbers of the form $g(x_1)$ with g in $\mathcal{G}_{I;a}$, it also follows that $g_a(x_1) \leq \hat{g}(x_1)$. Combining all the above then yields

$$g_a(x_1) \leq \hat{g}(x_1) \leq \hat{g}(x_2) < g_a(x_1),$$

which is clearly impossible. Thus, if x_1 and x_2 are elements of I such that $a \leq x_1 < x_2$, it follows that $g_a(x_1) \leq g_a(x_2)$.

Case 2 Suppose that $x_2 \leq a$, so that $x_1 < x_2 < a$. Assuming that $g_a(x_1) > g_a(x_2)$ then implies, because $g_a(x_1)$ is defined as a certain supremum, that there is $\hat{g}$ in $\mathcal{G}_{I;a}$ such that $g_a(x_1) \geq \hat{g}(x_1) > g_a(x_2)$. As before, one knows that $\hat{g}$ is monotonic up on I, so that $\hat{g}(x_2) \geq \hat{g}(x_1) > g_a(x_2)$. However, the definition of $g_a(x_2)$ as a certain supremum implies $g_a(x_2) \geq \hat{g}(x_2)$. Combining the preceding then implies

$$g_a(x_2) \geq \hat{g}(x_2) \geq \hat{g}(x_1) > g_a(x_2),$$

which is impossible. Thus if x_1 and x_2 are elements of I such that $x_1 < x_2 \leq a$, then $g_a(x_1) \leq g_a(x_2)$.

Case 3 The only situation yet to be handled is when $x_1 < a \leq x_2$. By Case (1) it is clear that $g_a(a) \leq g_a(x_2)$. Likewise, by Case (2) it is clear that $g_a(x_1) \leq g_a(a)$. Now apply the Transitivity Law for Order to conclude that $g_a(x_1) \leq g_a(x_2)$, as required.

Next, define h_a by the rule

$$h_a(x) = g_a(x) - (f(x) - f(a)) \text{ for all } x \text{ in } I.$$

An argument similar to that given for g_a can be used to show that h_a is monotonic up on I. Indeed, suppose that h_a is *not* monotonic up on I. Then, as before, there exist numbers x_1 and x_2 in I, with $x_1 < x_2$, such that $h_a(x_1) > h_a(x_2)$. That is, one has

$$g_a(x_1) - (f(x_1) - f(a)) > g_a(x_2) - (f(x_2) - f(a));$$

equivalently,

$$g_a(x_1) - f(x_1) > g_a(x_2) - f(x_2).$$

As above, there are three cases to consider.

Case 1′ Suppose that $a \leq x_1 < x_2$. For convenience let $C = g_a(x_1) - f(x_1)$, so that $C > g_a(x_2) - f(x_2)$. Since

$$g_a(x_2) - f(x_2) = \inf\{g(x_2) : g \in \mathcal{G}_{I;a}\} - f(x_2) = \inf\{g(x_2) - f(x_2) : g \in \mathcal{G}_{I;a}\},$$

it follows from properties of 'infimum' that there exists a number of the form $\hat{g}(x_2) - f(x_2)$, with $\hat{g}$ in $\mathcal{G}_{I;a}$, such that $C > \hat{g}(x_2) - f(x_2) \geq g_a(x_2) - f(x_2)$. That is,

$$g_a(x_1) - f(x_1) > \hat{g}(x_2) - f(x_2) \geq g_a(x_2) - f(x_2).$$

However, since $\hat{g} \in \mathcal{G}_{I;a}$, one knows that $\hat{g} - f$ is monotonic up on I. In particular, one then has

$$\hat{g}(x_2) - f(x_2) \geq \hat{g}(x_1) - f(x_1) \geq g_a(x_1) - f(x_1);$$

the final inequality follows from the definition of $g_a(x_1)$ as an infimum. Combining all these results then yields $g_a(x_1) - f(x_1) > \hat{g}(x_2) - f(x_2) \geq g_a(x_1) - f(x_1)$, which is impossible.

Case 2′ A similar proof shows that the inequality $h_a(x_1) > h_a(x_2)$, with $x_1 < x_2$, is impossible if $x_1 < x_2 < a$.

Case 3′ If $x_1 < a \leq x_2$, then an argument using the results of Case (1′) and Case (2′) shows that once again it is impossible to have $h_a(x_1) > h_a(x_2)$.

Proof of Property (ii) The formula $f(x) - f(a) = g_a(x) - h_a(x)$ follows easily from the definition of of $h_a(x)$ as the quantity $g_a(x) - (f(x) - f(a))$. Substituting $x = a$ into the formula above then yields $0 = f(a) - f(a) = g_a(a) - h_a(a)$, so that in any event one has $h_a(a) = g_a(a)$. The fact that $g_a(a) = 0$ follows from one of the defining properties of the set $\mathcal{G}_{I;a}$; namely, if $g \in \mathcal{G}_{I;a}$, then $g(a) = 0$. In particular, $g_a(a)$ is the infimum of the singleton set $\{0\}$, hence $g_a(a) = 0$.

Proof of Property (iii) If g and h are as described in the statement of (iii) then clearly $g \in \mathcal{G}_{I;a}$.
Case 1 Suppose that x in I satisfies $x \geq a$. Then the definition of $g_a(x)$ as an infimum implies $g(x) \geq g_a(x)$. Likewise, the fact that $f(x) - f(a) = g(x) - h(x) = g_a(x) - h_a(x)$ implies that $h(x) = g(x) - (f(x) - f(a)) \geq g_a(x) - (f(x) - f(a)) = h_a(x)$.
Case 2 Suppose that x in I satisfies $x < a$. Then the definition of $g_a(x)$ as a supremum implies $g(x) \leq g_a(x)$. Likewise, the formula $f(x) - f(a) = g(x) - h(x) = g_a(x) - h_a(x)$ implies $h(x) \leq h_a(x)$.

Finally, the fact that the functions g_a and h_a constructed above are the only ones which satisfy (i), (ii) and (iii) now follows easily. Indeed, suppose that $\overline{g_a}$ and $\overline{h}_a$ are also functions which have Properties (i), (ii) and (iii). Since $\overline{g}_a$ and $\overline{h}_b$ satisfy (i) and (ii), it follows from the fact that g_a and g_b satisfy (iii) that $\overline{g}_a(x) \geq g_a(x)$ and $\overline{h}_a(x) \geq h_a(x)$ for all x in I such that $x \geq a$. By reversing the roles of $\overline{g}_a$ and $\overline{h}_a$ with those of g_a and h_a, one likewise gets $g_a(x) \geq \overline{g}_a(x)$ and $h_a(x) \geq \overline{h}_a(x)$ for all x in I such that $x \geq a$. Thus, $\overline{g}_a(x) = g_a(x)$ and $\overline{h}_a(x) = h_a(x)$ for all x in I such that $x \geq a$. A similar argument shows that $\overline{g}_a(x) = g_a(x)$ and $\overline{h}_a(x) = h_a(x)$ for all x in I such that $x < a$.

VIII.6.10 Definition

Let $f : I \to \mathbb{R}$ be a function of hybrid type on an interval I, and let a be a point of I. The expression $f - f(a) = g_a - h_a$ obtained above is called the **Jordan splitting of f at a**; the functions g_a and h_a are then called the corresponding **components of the Jordan splitting of f at a**.

Remark Some authors would use the phrase 'Jordan *decomposition*' instead of 'Jordan *splitting*'. However, the 'Jordan decomposition' terminology is also used in other parts of mathematics, so we prefer the 'splitting' terminology to lessen the chance for confusion.

The next pair of results tell us, in effect, that the Jordan splitting provides an *optimal* representation of a hybrid function as the difference of two monotonic-up functions, in the sense indicated in Remark (VIII.6.8) (4) above.

VIII.6.11 Theorem

Suppose that $f : I \to \mathbb{R}$ is a function of hybrid type on the interval I. Let a be an element of I and let $f - f(a) = g_a - h_a$ be a Jordan splitting of f; note that, in particular, one has $g_a(a) = h_a(a) = 0$. Let g and h be monotonic-up functions on I such that $f = g - h$. Suppose that x_1 and x_2 are elements of I such that $x_1 < x_2$. Then

$$g(x_2) - g(x_1) \geq g_a(x_2) - g_a(x_1) \tag{VIII.9}$$

and

$$h(x_2) - h(x_1) \geq h_a(x_2) - h_a(x_1). \tag{VIII.10}$$

Moreover, if one gets equality in either of these inequalities, then g and h differ from g_a and h_a by constants on the interval $[x_1, x_2]$. More precisely, $g(x) = g_a(x) + g(x_1) - g_a(x_1)$ and $h(x) = h_a(x) + h(x_1) - h_a(x_1)$ for all x in $[x_1, x_2]$.

Proof Let $\hat{g} = g - g(a)$ and $\hat{h} = h - h(a)$. Then it is clear that $\hat{g}$ and $\hat{h}$ are in $\mathcal{G}_{I;a}$.

Now let x_1 and x_2 be in I and satisfy $x_1 < x_2$. Then one has

$$(i) \quad g(x_2) - g(x_1) = \hat{g}(x_2) - \hat{g}(x_1) \text{ and } (ii) \quad h(x_2) - h(x_1) = \hat{h}(x_2) - \hat{h}(x_1) \quad (*)$$

There are several cases to consider.

Case 1 Suppose that $x_1 \leq a \leq x_2$. (Note that this includes the cases $a = x_1 < x_2$ and $x_1 < a = x_2$.) Then by the very constructions of the functions g_a and h_a one has

$$\hat{g}(x_1) \leq g_a(x_1) \text{ and } \hat{h}(x_1) \leq h_a(x_1).$$

Likewise,

$$\hat{g}(x_2) \geq g_a(x_2) \text{ and } \hat{h}(x_2) \geq h_a(x_2).$$

It follows easily that

$$\hat{g}(x_2) - \hat{g}(x_1) \geq g_a(x_2) - g_a(x_1) \text{ and } \hat{h}(x_2) - \hat{h}(x_1) \geq h_a(x_2) - h_a(x_1).$$

Combine this with Part (i) of Equation $(*)$ above to get

$$g(x_2) - g(x_1) \geq g_a(x_2) - g_a(x_1);$$

likewise, combine with Part (ii) of $(*)$ to get

$$h(x_2) - h(x_1) \geq h_a(x_2) - h_a(x_1);$$

Case 2 Now assume that $a < x_1 < x_2$. Suppose that it is *not* the case that $g(x_2) - g(x_1) \geq g_a(x_2) - g_a(x_1)$; or, in light of Equation $(*)$, suppose that $\hat{g}(x_2) - \hat{g}(x_1) < g_a(x_2) - g_a(x_1)$. Define a new function $G : I \to \mathbb{R}$ by the rule

$$G(x) = \begin{cases} g_a(x) & \text{if } x \leq x_1 \\ g(x) - g(x_1) + g_a(x_1) & \text{if } x_1 < x < x_2 \\ g_a(x) - ((g_a(x_2) - g_a(x_1)) - (g(x_2) - g(x_1))) & \text{if } x \geq x_2 \end{cases}$$

Then it is easy to verify that $G \in \mathcal{G}_{I;a}$; the details are left as an exercise. Furthermore, if $g(x_2) - g(x_1) < g_a(x_2) - g_a(x_1)$, then it follows that $G(x_2) < g_a(x_2)$; but this would contradict the construction of $g_a(x_2)$ as an infimum.

A similar argument shows that one must have $h(x_2) - h(x_1) \geq h_a(x_2) - h_a(x_1)$.

Case 3 Assume now that $x_1 < x_2 < a$. Then an argument similar to that in Case (2) shows that $g(x_2) - g(x_1) \geq g_a(x_2) - g_a(x_1)$ and that $h(x_2) - h(x_1) \geq h_a(x_2) - h_a(x_1)$.

Inequalities (VIII.9) and (VIII.10) now follow.

Finally, suppose that $x \in [x_1, x_2]$. Then by what has just been proved one has

$$g(x_2) - g(x) \geq g_a(x_2) - g_a(x) \text{ and } g(x) - g(x_1) \geq g_a(x) - g_a(x_1)$$

Add the terms of these inequalities to get

$$g(x_2) - g(x_1) = (g(x_2) - g(x)) + (g(x) - g(x_1)) \geq (g_a(x_2) - g_a(x)) + (g_a(x) - g_a(x_1)) = g_a(x_2) - g_a(x_1).$$

Now suppose that $g(x_2) - g(x_1) = g_a(x_2) - g_a(x_1)$. Then all the inequalities above reduce to equations. In particular, one must have

$$g(x_2) - g(x) = g_a(x_2) - g_a(x) \text{ for all } x \text{ in } [x_1, x_2]$$

It follows that $g_a(x) - g(x) = g_a(x_2) - g(x_2)$ for all x in $[x_1, x_2]$; that is, the functions g and g_a differ by a constant on the interval $[x_1, x_2]$, as claimed.

A similar argument shows that if $h(x_2) - h(x_1) = h_a(x_2) - h_a(x_1)$, then h and h_a differ by a constant on $[x_1, x_2]$.

VIII.6.12 Corollary

Suppose that f is a function of hybrid type on the interval I; let a be a point of I, and let $g, h : I \to \mathbb{R}$ be functions on I. Then the following conditions are equivalent:

(i) The functions g and h are monotonic up on I and $f = g - h$.

(ii) There exists a monotonic-up function $\psi : I \to \mathbb{R}$ such that $g = \psi + g_a + f(a)$ and $h = \psi + h_a$.

Proof The fact that Statement (ii) implies Statement (i) is obviously true; see Remark (VIII.6.8) (1).

Now suppose that Statement (i) is true. Set $\psi = g - g_a - f(a)$.

Claim 1 The function ψ is monotonic up on I.

Proof of Claim 1 Note that if x_1 and x_2 are in I and $x_1 < x_2$, then by Inequality (VIII.9) one has

$$\psi(x_2) - \psi(x_1) = (g(x_2) - g(x_1)) - (((g_a(x_2) - f(a)) - (g_a(x_1) - f(a))) = (g(x_2) - g(x_1)) - (g_a(x_2) - g_a(x_1)) \geq 0$$

Thus ψ is monotonic up on I, as claimed.

Claim 2 One also has $h = \psi + h_a$.

Proof of Claim 2 Note that $f = g - h = g_a - h_a + f(a)$, hence $g - g_a - f(a) = h - h_a$. The claim now follows from the definition of ψ.

The next result shows that the dependence on the base point a of the Jordan splitting $f - f(a) = g_a - h_a$ is simple.

VIII.6.13 Corollary

Suppose that $f : I \to \mathbb{R}$ is a function of hybrid type on the interval I. Let a and b be elements of I, and let g_a, h_a and g_b, h_b be the corresponding Jordan splittings of f. Then

$$g_b(x) = g_a(x) - g_a(b) \text{ and } h_b(x) = h_a(x) - h_a(b) \text{ for all } x \text{ in } I.$$

In particular, for each pair of numbers x_1, x_2 in I one has

$$g_b(x_2) - g_b(x_1) = g_a(x_2) - g_a(x_1) \text{ and } h_b(x_2) - h_b(x_1) = h_a(x_2) - h_a(x_1).$$

The simple proof is left as an exercise.

The preceding results allows one to determine Jordan splittings in some important cases.

VIII.6.14 Theorem

Suppose that $f : I \to \mathbb{R}$ is a function of hybrid type on the interval I. Let a be a point of I, and let g_a and h_a denote the components of the corresponding Jordan splitting.

(a) If f is monotonic up on I then h_a is constant zero on I; likewise, if f is monotonic down on I, then g_a is constant zero on I.

(b) Suppose that c is a point of I which is not the left-hand endpoint of I, so that the left-hand jumps $\Delta_- f(c)$, $\Delta_- g_a(c)$ and $\Delta_- h_a(c)$ at c of the functions f, g_a and h_a exist and are finite. (See Part (4) of Definition (VIII.6.4) for the descriptions of the left-hand and right-hand jumps.) Then:

(i) If $\Delta_- f(c) \geq 0$ then $\Delta_- h_a(c) = 0$ and $\Delta_- g_a(c) = \Delta_- f(c)$.

(ii) If, instead, $\Delta_- f(c) \leq 0$ then $\Delta_- g_a(c) = 0$ and $\Delta_- h_a(c) = -\Delta_- f(c)$.

Likewise, suppose that c is not a right-hand endpoint of I. Then:

(iii) If $\Delta_{+}f(c) \geq 0$ then $\Delta_{+}g_a(c) = \Delta_{+}f(c)$ and $\Delta_{+}h_a(c) = 0$.

(iv) If $\Delta_{+}f(c) \leq 0$ then $\Delta_{+}g_a(c) = 0$ and $\Delta_{+}h_a(c) = \Delta_{+}f(c)$.

Proof

(a) If f is monotonic up on I, define $g : I \to \mathbb{R}$ and $h : I \to \mathbb{R}$ by the rules

$$g(x) = f(x) \text{ and } h(x) = 0 \text{ for all } x \text{ in } I.$$

Then g is monotonic up on I because of the hypothesis on f, and h is monotonic up on I because it is constant on I. Also, one has $g - h = f - 0 = f$. Thus, Theorem (VIII.6.11), and in particular Inequality (VIII.10), can be used to imply that for each x_1 and x_2 in I with $x_1 < x_2$ one has

$$0 = h(x_2) - h(x_1) \geq h_a(x_2) - h_a(x_1) \geq 0.$$

(The equation $0 = h(x_2) - h(x_1)$ uses the fact that h is a constant function on I; the inequality $h_a(x_2) - h_a(x_1) \geq 0$ reflects the fact that the function h_a is monotonic up on I.) It follows that $0 \geq h_a(x_2) - h_a(x_1) \geq 0$, which implies that $h_a(x_2) = h_a(x_1)$. Since this is true for every such pair of numbers x_1 and x_2 in I, it follows that h_a is constant on I, as claimed; and since $h_a(a) = 0$, that constant value is zero.

To prove that g_a is constant zero if f is monotonic down, apply the preceding result to the monotonic-up function $-f$.

(b) (i) Suppose that $\Delta_{-}f(c) \geq 0$. Since $f = g_a - h_a + f(a)$, it follows from basic limit laws that

$$\Delta_{-}f(c) = \Delta_{-}g_a(c) - \Delta_{-}h_a(c) \qquad (*)$$

Note that, by definition, $\Delta_{-}g_a(c) = g_a(c) - \lim_{x \nearrow c} g_a(c) \geq 0$, where the inequality on the right uses the fact that g_a is monotonic up on I and in the given limit x approaches c from the left. Likewise, one has

$$\Delta_{-}h_a(c) = h_a(c) - \lim_{x \nearrow c} h_a(x) \geq 0 \qquad (**)$$

For convenience, set $\Delta = \Delta_{-}h_a(c)$, and suppose, in contradiction to the conclusion of Statement (i), that $\Delta \neq 0$. In light of Inequality $(**)$ it follows that $\Delta > 0$. Now define $g, h : I \to \mathbb{R}$ by the rules

$$g(x) = \begin{cases} g_a(x) + f(a) & \text{if } x \in I \text{ and } x < c \\ g_a(x) + f(a) - \Delta & \text{if } x \in I \text{ and } x \geq c, \end{cases}$$

and

$$h(x) = \begin{cases} h_a(x) & \text{if } x \in I \text{ and } x < c \\ h_a(x) - \Delta & \text{if } x \in I \text{ and } x \geq c, \end{cases}$$

It is obvious that $g(x) - h(x) = g_a(x) - h_a(x) + f(a) = f(x)$ for all x in I. It is also clear that g and h are monotonic up on I. Indeed, suppose that x_1 and x_2 are in I and $x_1 < x_2$. If $x_2 < c$ then $x_1 < c$ hence

$$g(x_2) - g(x_1) = (g_a(x_2) + f(a)) - (g_a(x_1) + f(a)) = g_a(x_2) - g_a(x_1) \geq 0$$

because g_a is monotonic up on I. Likewise, if $x_1 \geq c$, so that $x_2 > c$, then

$$g(x_2) - g(x_1) = (g_a(x_2) + f(a) - \Delta) - (g_a(x_1) + f(a) - \Delta) = g_a(x_2) - g_a(x_1) \geq 0.$$

Finally, if $x_1 < c$ and $x_2 \geq c$, then $g(x_2) = g_a(x_2) + f(a) - \Delta \geq g_a(c) + f(a) - \Delta$, and $g(x_1) = g_a(x_1) + f(a)$. It then follows that

$$g(x_2) - g(x_1) \geq (g_a(c) + f(a) - \Delta) - (g_a(x_1) + f(a)) = g_a(c) - g_a(x_1) - \Delta$$

However, it follows, from the fact that g_a is monotonic up on I, together with Equation $(*)$, that

$$g_a(c) - g_a(x_1) \geq \Delta_- g_a(c) = \Delta_- f(c) + \Delta \geq \Delta,$$

where the inequality on the right uses the hypothesis that $\Delta_- f(c) \geq 0$. It follows from these results that $g(x_2) - g(x_1) \geq g_a(c) - g_a(x_1) - \Delta \geq 0$. It is clear now that g is monotonic up on I. An even simpler argument shows that h is also monotonic up on I. It now follows that g and h satisfy the hypotheses for Theorem (VIII.6.11). In particular, Inequality (VIII.10) then implies that if x_1 is in I and $x_1 < c$, then $h(c) - h(x_1) \geq h_a(c) - h_a(x_1)$; that is, $(h_a(c) - \Delta) - h_a(x_1) \geq h_a(c) - h_a(x_1)$. This inequality cannot be correct, since $\Delta > 0$. Thus, assuming $\Delta_- h_a(c) \neq 0$ leads to a contradiction, so one must have $\Delta_- h_a(c) = 0$, as claimed. Combining this with Equation $(*)$ then implies $\Delta_- g_a(c) = \Delta_- f(c)$, also as claimed, so Statement (i) is true.

The proofs of Statements (ii), (iii) and (iv) are similar, and are left as exercises.

VIII.6.15 Corollary

Suppose that $f : I \to \mathbb{R}$ is a function of hybrid type on the interval I. Let a be a point of I, and let g_a and h_a denote the components of the corresponding Jordan splitting. If f is continuous at a point c of I, then g_a and h_a are also continuous at c.

Proof The hypothesis that f is continuous at c implies that each one-sided jump of f at c is 0. It then follows from the preceding theorem that each one-sided jump at c of g_a and of h_a is also 0, and thus g_a and h_a are both continuous at c, as claimed.

VIII.6.16 Example

Let $f(x) = \cos x$ on the interval $I = [0, +\infty)$. One knows from elementary calculus that the function f is continuous on I. Also, f is monotonic down on intervals of the form $[(2k-2)\pi, (2k-1)\pi]$, and monotonic up on intervals of the form $[(2k-1)\pi, 2k\pi]$, with k in $\mathbb{N}$. It then follows from Theorem (VIII.6.14) that for each a in I, one has g_a constant on every interval of the form $[(2k-2)\pi, (2k-1)\pi]$, while h_a is constant on every interval of the form $[(2k-1)\pi, 2k\pi]$, with k in $\mathbb{N}$.

In particular, let $a = 0$. Then, since $f(0) = \cos 0 = 1$, one sees that

$$g_0(x) = 0,\ h_0(x) = 1 - (\cos x),\ 0 \leq x \leq \pi.$$

Note that on the interval $[0, \pi]$ we do have

$$f(x) = g_0(x) - h_0(x) + f(0), \quad g_0(0) = h_0(0) = 0,$$

aa required. Similarly, one hsa

$$g_0(x) = (\cos x) + 1,\ h_0(x) = 2,\ \pi \leq x \leq 2\pi.$$

Note: If you are confused by the numbers 1 and 2 appearing in the expressions for g_0 and h_0, recall that in this interval h_0 must be constant. However, since f is continuous on I it follows from

Corollary (VIII.6.15)) that g_a and h_a must also be continuous on I. In particular, the constant value of h_0 on the second subinterval $[\pi, 2\pi]$ must equal the value of h_0 at the right endpoint of the first subinterval $[0, \pi]$. That value is $h_0(\pi) = 1-\cos(\pi) = 1-(-1) = 2$. Then the number 1 appearing in the formula for g_0 comes from the fact that $g_0(x) = f(x)-f(0)+h_0(x) = \cos x-1+2 = \cos x+1$.

The reader is encouraged to determine the formulas for g_0 and h_0 on the intervals $[2\pi, 3\pi]$, $[3\pi, 4\pi]$, and so on.

In the preceding discussion, there are no restrictions on the type of interval I being used. In most treatments of these topics, however, one considers only closed bounded intervals; that is, intervals of the form $[\alpha, \beta]$, with α and β in $\mathbb{R}$ and $\alpha < \beta$. The next result tells us, in effect, that restricting attention to such intervals does not materially affect the theory.

VIII.6.17 Theorem

Suppose that $f : I \to \mathbb{R}$ is a function defined on an interval I. Let a be a point of I, and suppose that there exists a nonempty family $\mathcal{F}$ of subintervals of I such that

(i) the union of the intervals in the family $\mathcal{F}$ is the original interval I;

(ii) the number a is an element of each interval J in the family $\mathcal{F}$;

(iii) for each interval J in $\mathcal{F}$, the restriction $f|_J : J \to \mathbb{R}$ of f to J is of hybrid type on J.

Then f is of hybrid type on I.

Outline of Proof Let J be an interval in the family $\mathcal{F}$. By Statement (iii), together with Theorem (VIII.6.9), there exist unique functions $g_{a;J} : J \to \mathbb{R}$ and $h_{a;J} : \to \mathbb{R}$, monotonic up on J, such that $f|_J = g_{a;J} - h_{a;J} + f(a)$ and $g_{a;J}(a) = h_{a;J}(a) = 0$. Now define functions $g, h : I \to \mathbb{R}$ by the following rule:

If $x \in I$, let J be an interval in the family $\mathcal{F}$ such that $x \in J$; such J exists because Statement (i) holds. Now define $g(x) = g_{a;J}(x)$ and $h(x) = h_{a;J}(x)$. It is easy to see, by the 'uniqueness' of the functions $g_{a;J}$ and $h_{a;J}$ mentioned above, that if x is an element of more than one interval in the family $\mathcal{F}$, it does not matter which such interval is used as J. It is then easy to verify that g and h are monotonic up on I, that $g(a) = h(a) = 0$, and that $f(x) = g(x) - h(x) + f(a)$ for all x in I, and that the expresion $f = g - h + f(a)$ is the Jordan splitting at a of f.

VIII.6.18 Remark (Jordan's Observation)

In his original paper on these topics, Jordan gave a simple criterion for a function $f : [a, b] \to \mathbb{R}$, defined on a closed bounded interval $[a, b]$, to be of hybrid type. The criterion is based on the following observation:

Suppose that $f : [a, b] \to \mathbb{R}$ can be expressed in the form $f = g - h$, for monotonic-up functions g and h. Note that if c and d are any points of $[a, b]$ such that $c < d$, then

$$f(d) - f(c) = (g(d) - g(c)) - (h(d) - h(c)) \leq g(d) - g(c),$$

and

$$-(f(d) - f(c)) = (h(d) - h(c)) - (g(d) - g(c)) \leq h(d) - h(c);$$

in both cases one uses the fact that g and h are monotonic up, and thus $g(d) - g(c) \geq 0$ and $h(d) - h(c) \geq 0$. Since $|f(d) - f(c)|$ equals one of the numbers $f(d) - f(c)$ or $-(f(d) - f(c))$, it follows that $|f(d) - f(c)|$ is no larger than one of the nonnegative numbers $g(d) - g(c)$ and $h(d) - h(c)$. That is, $|f(d) - f(c)| \leq \max\{(g(d) - g(c)), (h(d) - h(c))\}$. More generally, let x_0, x_1,

$x_2, \ldots x_k$ be points of I such that $a = x_0 < x_1 < \ldots < x_{k-1} < x_k = b$; that is, $\{x_0, \ldots x_k\}$ is a partition of the interval $[a, b]$. One then has

$$|f(x_j) - f(x_{j-1})| \leq \max\{(g(x_j) - g(x_{j-1})), (h(x_j) - h(x_{j-1}))\} \text{ for each } j = 1, 2, \ldots k.$$

By adding these quantities, one then gets

$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| \leq \sum_{j=1}^{k} \max\{(g(x_j) - g(x_{j-1})), (h(x_j) - h(x_{j-1}))\}.$$

The sum on the right side of this last inequality is not easy to analyse directly. However, note that, since g and h are monotonic up, one has

$$\max\{(g(x_j) - g(x_{j-1})), (h(x_j) - h(x_{j-1}))\} \leq (g(x_j) - g(x_{j-1})) + (h(x_j) - h(x_{j-1})).$$

Thus

$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| \leq \sum_{j=1}^{k} \max\{(g(x_j) - g(x_{j-1})), (h(x_j) - h(x_{j-1}))\} \leq$$

$$\sum_{j=1}^{k}(g(x_j) - g(x_{j-1})) + \sum_{j=1}^{k}(h(x_j) - h(x_{j-1})) = (g(b) - g(a)) + (h(b) - h(a)). \tag{VIII.11}$$

In other words, a necessary condition for $f : [a, b] \to \mathbb{R}$ to be of hybrid type on $[a, b]$ is that the set of numbers of the form $\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})|$, where $a = x_0 < x_1 < \ldots < x_{k-1} < x_k = b$ is a partition of $[a, b]$, is bounded above. What Jordan proved, in effect, is that this condition is also sufficient for f to be of hybrid type on $[a, b]$.

VIII.6.19 Definition

Let $f : [a, b] \to \mathbb{R}$ be a real-valued function defined on a closed bounded interval $[a, b]$. Let $\mathcal{P}_{[a,b]}$ denote the set of all partitions of the interval $[a, b]$.

(1) Suppose that $P = \{a = x_0 < x_1 < \ldots < x_k = b\}$ is an element of $\mathcal{P}_{[a,b]}$, and associate with f and P the number $V(f, P)$ be given by the formula

$$V(f, P) = \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})|, \tag{VIII.12}$$

Then the quantity $V_{[a,b]}(f) = \sup\{V(f, P) : P \in \mathcal{P}_{[a,b]}\}$ is called the **total variation of f over the interval $[a, b]$**.

(2) One says that f **is of bounded variation on** $[a, b]$ provided the quantity $V_{[a,b]}$ is a real number; that is, provided $V_{[a,b]}(f) < +\infty$.

Note: In light of the definition of 'supremum', one can say instead that f is of bounded variation on $[a, b]$ provided that there exists a number M such that $V(f, P) \leq M$ for all partitions P of $[a, b]$. This follows more closely the approach used by Jordan in his treatment of the concept.

In the original treatment of the concept of 'function of bounded variation', Jordan first discusses the (slightly) more primitive ideas of 'positive variation' and 'negative variation'.

VIII.6.20 Definition

Let $f : [a,b] \to \mathbb{R}$ be a real-valued function defined on a closed bounded interval $[a,b]$. Suppose that $P = \{a = x_0 < x_1 < \ldots < x_k = b\}$ is a partition of $[a,b]$, and associate with f and P the numbers $V^+(f,P)$ and $V^-(f,P)$ given by the formulas

$$V^+(f,P) \,=\, \sum_{j=1}^{k} (f(x_j) - f(x_{j-1}))^+\,, \text{ and } V^-(f,P) \,=\, \sum_{j=1}^{k} (f(x_j) - f(x_{j-1}))^-\,. \qquad \text{(VIII.13)}$$

(As usual, if c is a real number, then c^+ and c^- denote, respectively, the positive and negative part of c; see Definition (**??**).) Then the quantities $V^+_{[a,b]}(f)$ and $V^-_{[a,b]}(f)$, given by

$$V^+_{[a,b]}(f) \,=\, \sup\{V^+(f,P) : P{\in}\mathcal{P}_{[a,b]}\} \text{ and } V^-_{[a,b]}(f) \,=\, \sup\{V^-(f,P) : P{\in}\mathcal{P}_{[a,b]}\}$$

are called, respectively, the **total positive variation** and the **total negative variation** of f over the interval $[a,b]$. If $V^+_{[a,b]}(f)$ is a real number (equivalently: if there is a number M^+ such that $V^+(f,P) \le M^+$ for all partitions P of $[a,b]$), then one says that f is **of bounded positive variation on** $[a,b]$. The concept of f being of **bounded negative variation on** $[a,b]$ is defined similarly.

VIII.6.21 Theorem

Suppose that $f : [a,b] \to \mathbb{R}$ is defined on the closed bounded interval $[a,b]$. Then:

(a) For each partition P of the interval $[a,b]$, if P' is a refinement of P, then $0 \le V^+(f,P) \le V^+(f,P')$, $0 \le V^-(f,P) \le V^-(f,P')$ and $0 \le V(f,P) \le V(f,P')$.

In particular, if P_0 is any partition of $[a,b]$, then one has

$$V^+_{[a,b]}(f) \,=\, \sup\{V^+(f,P') : P'{\in}\mathcal{P}_{[a,b]} \text{ and } P' \supseteq P_0\}.$$

Likewise, one has

$$V^-_{[a,b]}(f) \,=\, \sup\{V^-(f,P') : P'{\in}\mathcal{P}_{[a,b]} \text{ and } P' \supseteq P_0\},$$

and

$$V_{[a,b]}(f) \,=\, \sup\{V(f,P') : P'{\in}\mathcal{P}_{[a,b]} \text{ and } P' \supseteq P_0\}.$$

In other words, there is no loss in generality in computing these total variations by restricting one's attention to partitions of $[a,b]$ which contain a given finite collection of points in $[a,b]$.

(b) For each partition P of $[a,b]$ one has

$$V(f,P) \,=\, V^+(f,P) + V^-(f,P) \qquad \text{(VIII.14)}$$

and

$$f(b) - f(a) \,=\, V^+(f,P) - V^-(f,P) \qquad \text{(VIII.15)}$$

(c) The function f is of bounded variation on $[a,b]$ if, and only if, $V^+_{[a,b]}(f)$ and $V^-_{[a,b]}(f)$ are both finite. When this occurs, one has

$$V_{[a,b]}(f) \,=\, V^+_{[a,b]}(f) + V^-_{[a,b]}(f) \qquad \text{(VIII.16)}$$

and

$$f(b) - f(a) \,=\, V^+_{[a,b]}(f) - V^-_{[a,b]}(f). \qquad \text{(VIII.17)}$$

(d) If f is of bounded variation on $[a, b]$ then f is of bounded variation on every subinterval of $[a, b]$. Moreover, if u, v and w are numbers such that $a \leq u < v < w \leq b$, then

$$V^{+}_{[u,w]}(f) = V^{+}_{[u,v]}(f) + V^{+}_{[v,w]}(f). \tag{VIII.18}$$

Likewise, one has

$$V^{-}_{[u,w]}(f) = V^{-}_{[u,v]}(f) + V^{-}_{[v,w]}(f) \tag{VIII.19}$$

and

$$V_{[u,w]}(f) = V_{[u,v]}(f) + V_{[v,w]}(f). \tag{VIII.20}$$

Proof

(a) Suppose that $P = \{x_0 = a < x_1 < \ldots < x_{k-1} < x_k = b\}$.

Special Case Suppose that partition P' has exactly one more point than P. Then $P' \backslash P = \{c\}$ for some number c in $[a, b]$. Clearly there is an index j_0, with $1 \leq j_0 \leq k$, such that $x_{j_0 - 1} < c < x_{j_0}$. Assume, for the moment, that $1 < j_0 < k$. Then it is clear that

$$V^{+}(f, P') = (f(x_1) - f(x_0))^{+} + \ldots + (f(c) - f(x_{j_0-1}))^{+} + (f(x_{j_0}) - f(c))^{+} + \ldots + (f(x_k) - f(x_{k-1}))^{+}.$$

In comparison, on has

$$V(f, P)^{+} = (f(x_1) - f(x_0))^{+} + \ldots + (f(x_{j_0}) - f(x_{j_0-1}))^{+} + \ldots + (f(x_k) - f^{+}_{x_{k-1}}.$$

In other words, the sums forming the quantities $V(f, P')$ and $V(f, P)$ have exactly the same terms, with the following exception: the single term $(f(x_{j_0}) - f(x_{j_0-1}))^{+}$ in the sum for $V^{+}(f, P)$ corresponds to the two terms $(f(c) - f(x_{j_0-1}))^{+} + (f(x_{j_0}) - f(c))^{+}$ in the sum for $V^{+}(f, P')$. However, notice that

$$(f(x_{j_0}) - f(x_{j_0-1}))^{+} = (f(x_{j_0}) - f(c) + f(c) - f(x_{j_0-1}))^{+} =$$
$$((f(x_{j_0}) - f(c)) + (f(c) - f(x_{j_0-1})))^{+} \leq (f(c) - f(x_{j_0-1}))^{+} + (f(x_{j_0}) - f(c))^{+},$$

where the final inequality comes from Part (d) of Theorem (**??**). It follows easily that $V^{+}(f, P) \leq V^{+}(f, P')$, in this special case that P' has exactly one point more than P.

If P' is an arbitrary refinement of P, then P' can be obtained by adjoining finitely many points to P, one after another, and repeatedly using the special case above, together with the transitivity property of order.

Now suppose that P_0 is a partition of $[a, b]$.

Case 1 If $V^{+}_{[a,b]}(f) = +\infty$, then for each M in $\mathbb{R}$ there exists a partition P of $[a, b]$ such that $V^{+}(f, P) \geq M$. Let $P' = P \cup P_0$, so that P' is a refinement of P which contains P_0. By what was just proved, one then has $V^{+}(f, P') \geq V^{+}(f, P) \geq M$. It follows that $\sup\{V^{+}(f, P') : P' \in \mathcal{P}_{[a,b]} \text{ and } P' \supseteq P_0\} \geq M$. Since this holds for each M, it follows that the supremum on the left equals $+\infty = V^{+}_{[a,b]}(f)$, as required.

Case 2 If $V^{+}_{[a,b]}(f)$ is finite, then for every $\varepsilon > 0$ there exists a partition P of $[a, b]$ such that

$$V^{+}_{[a,b]}(f) - \varepsilon < V^{+}(f, P) \leq V^{+}_{[a,b]}(f).$$

Let $P' = P \cup P_0$. Then by reasoning similar to that used above it follows that

$$V^{+}_{[a,b]}(f) - \varepsilon < \sup\{V^{+}(f, P') : P' \in \mathcal{P}_{[a,b]} \text{ and } P' \supseteq P_0\} \leq V^{+}_{[a,b]}(f).$$

Since this is true for each $\varepsilon > 0$, the equation $V^+_{[a,b]}(f) = \sup\{V^+(f,P') : P'\in\mathcal{P}_{[a,b]}$ and $P' \supseteq P_0\}$ follows.

A similar proof works in for $V^-(f,P)$ and $V(f,P)$; the details are left as an exercise.

(b) Equation (VIII.14) follows easily from the observation that

$$|f(x_j) - f(x_{j-1})| = (f(x_j) - f(x_{j-1}))^+ + (f(x_j) - f(x_{j-1}))^- \,;$$

see Part (b) of Theorem (**??**). Equation (VIII.15) follows in a similar manner from the observation that $f(x_j) - f(x_{j-1}) = (f(x_j) - f(x_{j-1}))^+ - (f(x_j) - f(x_{j-1}))^-$, together with the fact that

$$\sum_{j=1}^{n}(f(x_j) - f(x_{j-1})) = f(b) - f(a).$$

(c) First, suppose that f is of bounded variation on $[a,b]$. Then, by definition, there exists M in $\mathbb{R}$ such $V(f,P) \leq M$ for all partitions P of $[a,b]$. It then, by Equation (VIII.14), one sees that

$$V^+(f,P) \leq V^+(f,P)+V^-(f,P) = V(f,P) \leq M \text{ and } V^-(f,P) \leq V^+(f,P)+V^-(f,P) = V(f,P) \leq M$$

for all partitions P of $[a,b]$. Conversely, suppose that there exist numbers M^+ and M^- such that $V^+(f,P) \leq M^+$ and $V^-_{[a,b]}(f) \leq M^-$ for all such partitions P. Then, by Equation (VIII.14) and the definitions of $V^+_{[a,b]}(f)$ and $V^-_{[a,b]}(f)$, for every partition P of $[a,b]$ one has

$$V(f,P) = V^+(f,P) + V^-(f,P) \leq M^+ + M^-.$$

Thus, f is of bounded variation on $[a,b]$.

Now suppose that f is of bounded variation on $[a,b]$. Let $\varepsilon > 0$ be given, and chose partitions P^+ and P^- of $[a,b]$ such that

$$V^+_{[a,b]}(f) - \varepsilon/2 < V^+(f,P^+) \leq V^+_{[a,b]}(f), \text{ and } V^-_{[a,b]}(f) - \varepsilon/2 < V^-(f,P^-) \leq V^-_{[a,b]}(f).$$

(That such partitions P^+ and P^- exist follows from the definition of $V^+_{[a,b]}(f)$ and $V^-_{[a,b]}(f)$ as suprema, and the fact, just proved, that when f is of bounded variation on $[a,b]$ then all the preceding quantities are real numbers, so the indicated arithmetic makes sense.) Let P' be any partition of $[a,b]$ which is simultaneously a refinement of P^+ and P^-; in other words, let P' be a partition such that $P' \supseteq P^+ \cup P^-$. Then, by Part (a) (and the meaning of 'supremum') one obtains

$$V^+_{[a,b]}(f) - \varepsilon/2 < V^+(f,P') \leq V^+_{[a,b]}(f), \text{ and } V^-_{[a,b]}(f) - \varepsilon/2 < V^-(f,P') \leq V^-_{[a,b]}(f).$$

By the orders properties of $\mathbb{R}$ it then follows that

$$V^+_{[a,b]}(f) + V^-_{[a,b]}(f) - \varepsilon < V^+(f,P') + V^-(f,P') \leq V^+_{[a,b]}(f) + V^-_{[a,b]}(f)$$

By Part (b) one can then write

$$V^+_{[a,b]}(f) + V^-_{[a,b]}(f) - \varepsilon < V(f,P') \leq V^+_{[a,b]}(f) + V^-_{[a,b]}(f)$$

Since this holds for *every* partition P' of $[a,b]$ such that $P' \supseteq P^+ \cup P^-$, it follows from the last portion of Part (a) (and properties of suprema) that

$$V^+_{[a,b]}(f) + V^-_{[a,b]}(f) - \varepsilon < V_{[a,b]}(f) \leq V^+_{[a,b]}(f) + V^-_{[a,b]}(f).$$

Since this is true for all $\varepsilon > 0$, Equation (VIII.16) follows.

Equation (VIII.17) follows in a like manner by using Equation (VIII.15).

(d) The proof of this part is similar to what has come before, and is left as an exercise.

VIII.6.22 Theorem (Jordan's Bounded-Variation Theorem)

A necessary and sufficient condition for a function $f : [a,b] \to \mathbb{R}$ to be of hybrid type on a closed bounded interval $[a,b]$ is that f be of bounded variation on $[a,b]$.

Proof The fact that this condition is necessary is obvious; see 'Jordan's Observation' above.

To see that it is also sufficient, note that if f is of bounded variation on $[a,b]$, then (by Part (d) of Theorem (VIII.6.21)) for each x such that $a < x \leq b$ the function f is of bounded variation on the subinterval $[a,x]$. Thus by Part (c) of that theorem it follows that $V^{+}_{[a,x]}(f)$ and $V^{-}_{[a,x]}(f)$ are finite. Define functions $g, h : [a,b] \to \mathbb{R}$ be the rules

$$g(a) = 0, \quad g(x) = V^{+}_{[a,x]}(f) \text{ if } a < x \leq b \text{ and } h(a) = 0, \quad h(x) = V^{-}_{[a,x]}(f) \text{ if } a < x \leq b.$$

It follows easily from Part (d) of Theorem (VIII.6.21) that g and h are monotonic up on $[a,b]$, and it follows from Equation (VIII.17) that $f(x) = f(a) + g(x) - h(x)$ for all x in $[a,b]$. Thus, f is a function of hybrid type on $[a,b]$, as claimed.

VIII.6.23 Remarks

(1) It is a simple exercise to show that the functions g and h obtained in the proof of the preceding theorem are the functions g_a and h_a used to form the Jordan splitting of f at the left-end point a; see Definition (VIII.6.10).

(2) Many analysis texts use the alternate formula

$$f(x) - f(a) = V_{[a,x]}(f) + (V_{[a,x]} - f(x)) \tag{VIII.21}$$

to express $f(x) - f(a)$ as the difference of the monotonic-up functions $V_{[a,x]}(f)$ and $V_{[a,x]}(f)$. In light of Remark (1) above and Equation (VIII.16), this is the same as writing

$$f(x) - f(a) = (g_a(x) + h_a(x)) + ((g_a(x) + h_a(x)) - (g_a(x) - h_a(x)) = (g_a(x) + h_a(x)) - 2h_a(x).$$

Since g_a and h_a are monotonic up on $[a,b]$, it follows that Equation (VIII.21) does represent $f(x) - f(a)$ as the difference of two monotonic-up functions.

(3) It is easy to show that a function f is of hybrid type on an interval I if, and only if, f is of bounded variation on each bounded subinterval $[a,b]$ of I. It is standard in analysis to use only the 'bounded variation' terminology, so from here on we drop the 'hybrid-type' terminology.

We end with two examples which illustrate the fact that a continuous function can fail to be of bounded variation, while a function of bounded variation can 'wiggle' rather badly.

VIII.6.24 Examples

(1) Let $y_1, \; y_2, \ldots y_k, \ldots$ be a sequence of positive numbers such that $\lim_{k\to\infty} y_k = 0$. Let $X = \left\{1, \frac{1}{2}, \frac{1}{2^2}, \ldots \frac{1}{2^{k-1}}, \ldots\right\} \cup \{0\}$, and define $\varphi : X \to \mathbb{R}$ by the rule

$$\varphi(x) = \begin{cases} 0 & \text{if } x = 1/2^{2m-2} \text{ for some } m \text{ in } \mathbb{N} \text{ or if } x = 0 \\ y_m & \text{if } x = 1/2^{2m-1} \text{ for some } m \text{ in } \mathbb{N} \end{cases}$$

It is clear that X is a closed subset of $\mathbb{R}$ and that φ is continuous on X. Let $f : [0,1] \to \mathbb{R}$ be the piecewise-linear continuous extension of φ to $[0,1]$ guaranteed by Theorem (VIII.3.1), the Tietze Extension Theorem. It is easy to see that $V_{[1/2^{2m},1/2^{2m-2}]}(f) = 2y_m$, Thus, by repeated use of Equation (VIII.20), one gets $V_{[1/2^{2m},1]}(f) = 2y_1 + 2y_2 + \ldots + 2y_m$ It is easy to see that a necessary and sufficient condition for f to be of bounded variation on the full domain $[0,1]$ is that there exist a number M such that $y_1 + y_2 + \ldots + y_m \leq M$ for all m in $\mathbb{N}$.

Special Case Let $y_m = \ln\left(\frac{m+1}{m}\right)$ for each m in $\mathbb{N}$. It is easy to see that $y_m > 0$ and $\lim_{m\to\infty} y_m = 0$. However, by the usual properties of logarithms, one also has

$$y_1+y_2+\ldots+y_m = \ln\left(\frac{2}{1}\right)+\ln\left(\frac{3}{2}\right)+\ldots+\ln\left(\frac{m+1}{m}\right) = \ln\left(\frac{2}{1}\cdot\frac{3}{2}\cdot\ldots\frac{m}{m-1}\cdot\frac{m+1}{m}\right) = \ln(m+1).$$

Since $\lim_{m\to\infty} \ln(m+1) = +\infty$, it is clear that with this choice of y_m the resulting function f is *not* of bounded variation on $[0,1]$.

(2) Let $Z = \{z_1, z_2, \ldots z_k, \ldots\}$ be a countably infinite subset of the open interval $(0,1)$, with the labeling set up so that if $i \neq j$ then $z_i \neq z_j$. Let $(y_1, y_2, \ldots)$ be a strictly decreasing infinite sequence of positive numbers such that $\lim_{k\to\infty} y_k = 0$. Define $f : [0,1] \to \mathbb{R}$ by the rule

$$f(x) = \left\{ \begin{array}{ll} y_k & \text{if } x = z_k \text{ for some } k \text{ in } \mathbb{N} \\ 0 & \text{if } x \text{ is not in } Z \end{array} \right\}$$

Let $\mathcal{P} = \{0 = x_0 < x_1 < \ldots < x_{2n-1} < x_{2n} = 1\}$ be a partition of the interval $[0,1]$ into an even number $2n$ of subintervals, and consider the sum $S = \sum_{j=1}^{2n} |f(x_j) - f(x_{j-1})|$ The only way that one of the terms $|f(x_j) - f(x_{j-1})|$ can be nonzero is if at least one of the endpoints x_{j-1} or x_j is of the form z_k, and the largest that term could be is y_k. Moreover, a point z_k can be an endpoint of no more than two of the intervals determined by the partition $\mathcal{P}$. Since the sequence $(y_1, y_2, \ldots)$ is strictly decreasing, the most the sum S could be is $2y_1 + 2y_2 + \ldots + 2y_n$. It now follows that if there is a number M such that $y_1 + y_2 + \ldots + y_m \leq M$ for all m in $\mathbb{N}$, then f is of bounded variation.

Special Case Let $Z = (z_1, z_2, \ldots)$ be the set of all rational numbers in the interval $(0,1)$, and let $y_k = a^k$, where a is a fixed number such that $0 < a < 1$. Then it is clear that $y_1 + \ldots + y_m < 1/(1-a)$. The corresponding function f then is of bounded variation on $[0,1]$; however, it is monotonic on *no* subinterval of $[0,1]$, because the rationals are dense in $[0,1]$.

VIII.7 EXERCISES FOR CHAPTER VIII

VIII - 1 In each part of this exercise, either construct an example of a sequence ξ for which $\mathcal{L}[\xi]$ is the indicated set, or show that no such sequence exists.

(a) $\mathcal{L}[\xi] = \{-1, 3, \pi\}$ (b) $\mathcal{L}[\xi] = \mathbb{N}$

VIII - 2 Note: In doing Part (b) of this exercise you may use the result of Part (a), whether you proved it or not. Likewise, in doing Part (c) you may assume the results of Parts (a) and (b).

(a) Prove Part (b) of Theorem C.6.15. (Hint: Show that Condition (i) is a necessary and sufficient condition for the number L to satisfy the inequality $L \geq \limsup \xi$, while Condition (ii) is a necessary and sufficient condition for L to satisfy the inequality $L \leq \limsup \xi$.)

(b) Prove Part (c) of Theorem C.6.15.

(c) Prove Part (d) of Theorem C.6.15.

VIII - 3 Let $\xi = (x_1, x_2, \ldots)$ be a sequence of real numbers. Show that

$$\liminf_{k \to \infty} x_k = -\limsup_{k \to \infty} -x_k \text{ and } \limsup_{k \to \infty} x_k = -\liminf_{k \to \infty} -x_k.$$

(As usual, in the case of the quantities $+\infty$ and $-\infty$ one has $-(+\infty) = -\infty$ and $-(-\infty) = +\infty$.)

VIII - 4 Suppose that $\alpha = (a_1, a_2, \ldots)$ and $\beta = (b_1, b_2, \ldots)$ are bounded sequences of real numbers.

(a) Show that

$$\limsup_{n \to \infty}(a_n + b_n) \leq \left(\limsup_{n \to \infty} a_n\right) + \left(\limsup_{n \to \infty} b_n\right)$$

(b) Find an example of bounded α and β for which the inequality in Part (a) is strict.

VIII - 5 Suppose that $\xi = (x_1, x_2, \ldots)$ is a sequence of positive numbers.

(a) Prove that $\limsup_{k \to \infty} \sqrt[k]{x_k} \leq \limsup_{k \to \infty} \frac{x_{k+1}}{x_k}$. (Hint: Look at the 'Hint' for Part (a) of Exercise **VIII - 2** .)

(b) Prove that $\liminf_{k \to \infty} \frac{x_{k+1}}{x_k} \leq \liminf_{k \to \infty} \sqrt[k]{x_k}$.

VIII - 6 Suppose that $X_1, X_2, \ldots X_k, \ldots$ is a nested sequence of nonempty compact subsets of $\mathbb{R}$. ('Nested': For each k one has $X_{k+1} \subseteq X_k$.) Prove that the intersection of the family $\{X_k : k \in \mathbb{N}\}$ is nonempty. Note: This result is often called the **Cantor Intersection Theorem**.

VIII - 7 Prove or Disprove If X is a nonempty subset of $\mathbb{R}$ such that every continuous function $f : X \to \mathbb{R}$ is bounded on X, then X is compact.

VIII - 8 Let $f : X \to \mathbb{R}$ be a continuous function whose domain X is a nonempty compact subset of $\mathbb{R}$. Let Y be the image $f[X]$ of X under f.

(a) Prove that Y is also a compact subset of Y.

(b) Suppose that, in addition, f maps X bijectively onto Y. Prove that the inverse map $f^{-1} : Y \to X$ is continuous on Y.

(c) Determine whether conclusion of Part (b) remains true if the hypothesis 'X is compact' is omitted.

VIII - 9 Let $\xi = (x_1, x_2, \dots)$ be a sequence of real numbers, and let $M = \sup\{x_1, x_2, \dots\}$.

Prove or Disprove: The quantity M is an element of $\mathcal{L}[\xi]$ if for each index j one has $x_j \neq M$.

Prove or Disprove: The quantity M is an element of $\mathcal{L}[\xi]$ only if for each index j one has $x_j \neq M$.

VIII - 10 Definition Let $\mathcal{F}$ be a nonempty family of real-valued functions defined on a nonempty subset X of $\mathbb{R}$. One says that the family $\mathcal{F}$ is **uniformly bounded on** X provided that there exists a number M such that $|f(x)| \leq M$ for all points x in X and for all functions f in the family $\mathcal{F}$.

(a) Give an example of a nonempty family $\mathcal{F}$ of functions $f : \mathbb{R} \to \mathbb{R}$ such that each function f in $\mathcal{F}$ is bounded on $\mathbb{R}$ but $\mathcal{F}$ is not uniformly bounded on $\mathbb{R}$.

(b) Prove or Disprove If a sequence of functions $f_k : X \to \mathbb{R}$ converges uniformly on X to a function $f : X \to \mathbb{R}$, and if for each k the function f_k is bounded on X, then the family $\mathcal{F} = \{f_1, f_2, \dots\}$ is uniformly bounded on X.

VIII - 11 Suppose that $\varphi = (f_1, f_2, \dots)$ and $\gamma = (g_1, g_2, \dots)$ are sequences of real-valued functions defined on a nonempty set X in $\mathbb{R}$. Assume that φ converges uniformly on X to f and γ converges uniformly on X to g.

(a) Prove that the sequence $\sigma = ((f_1 + g_1), (f_2 + g_2), \dots)$ converges uniformly on X to $f + g$.

(b) Give an example of such sequences φ and γ for which the corresponding sequence $\mu = (f_1 \cdot g_1, f_2 \cdot g_2, \dots)$ of products fails to be uniformly convergent on X.

(c) Prove that if one assumes, in addition to the uniform convergence on X of φ and γ, that each f_k and each g_k is bounded on X, then the sequence μ does converge uniformly to $f \cdot g$ on X.

VIII - 12 Let $[a, b]$ be a fixed compact interval in $\mathbb{R}$, with $a < b$; and for each n in $\mathbb{N}$ let $f_n : [a, b] \to \mathbb{R}$ be a continuous function.

(a) Prove or Disprove: A *necessary* condition for the sequence $\varphi = (f_1, f_2, \dots)$ to converge uniformly on $[a, b]$ is that for every Cauchy sequence $\xi = (x_1, x_2, \dots)$ in $[a, b]$ the corresponding sequence of values $(f_1(x_1), f_2(x_2), \dots f_k(x_k), \dots)$ is Cauchy in $\mathbb{R}$.

(b) Prove or Disprove: A *sufficient* condition for the sequence $\varphi = (f_1, f_2, \dots)$ to converge uniformly on $[a, b]$ is that for every Cauchy sequence $\xi = (x_1, x_2, \dots)$ in $[a, b]$ the corresponding sequence of values $(f_1(x_1), f_2(x_2), \dots f_k(x_k), \dots)$ is Cauchy in $\mathbb{R}$.

VIII - 13 Prove or Disprove: If $f : \mathbb{R} \to \mathbb{R}$ is a bounded monotonic continuous function, then f is uniformly continuous on $\mathbb{R}$.

VIII - 14 Suppose that $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ are functions defined on nonempty subsets X and Y, respectively, and assume that $f[X] \subseteq Y$.

Prove or Disprove: If f is uniformly continuous on X and g is uniformly continuous on Y, then their composition $h = g \circ f$ is uniformly continuous on X.

VIII - 15 Let X be a nonempty subset of $\mathbb{R}$.

(a) Prove or Disprove: If $f : X \to \mathbb{R}$ is uniformly continuous on X then $|f| : X \to \mathbb{R}$ is uniformly continuous on X.

(b) Prove or Disprove: If $|f| : X \to \mathbb{R}$ is uniformly continuous on X then $f : X \to \mathbb{R}$ is uniformly continuous on X.

VIII - 16 Let $f : \to \mathbb{R}$ be a real-valued function whose domain is a nonempty subset X of $\mathbb{R}$. Prove that the following statements are equivalent:

(i) The function f is continuous on X.

(ii) For every closed subset Y of $\mathbb{R}$ the set $f^{-1}[Y]$ is of the form $X \cap Z$ for some closed subset Z of $\mathbb{R}$.

VIII - 17 Prove Theorem F.4.5

VIII - 18 Prove Dini's Theorem (see Theorem F.5.1) using the Cantor Intersection Theorem (see Exercise **VIII - 6**) and the results of Exercise **VIII - 16** above. (Hint: First reduce to the case in which the functions f_n are decreasing pointwise on X to the zero function. Then for each $\varepsilon > 0$ consider the sets $Y_n(\varepsilon)$ of the form $Y_n(\varepsilon) = \{x \in X : f_n(x) \geq \varepsilon\}$.)

VIII - 19 Definition (1) A function $f : \mathbb{R} \to \mathbb{R}$ is said to be an **open function** provided $f[U]$ is open in $\mathbb{R}$ for each open set $U \subseteq \mathbb{R}$.

(2) A function $f : \mathbb{R} \to \mathbb{R}$ is said to be a **closed function** provided $f[X]$ is closed in $\mathbb{R}$ for each closed set $X \subseteq \mathbb{R}$.

Problem (a) Give an example of a continuous function $f : \mathbb{R} \to \mathbb{R}$ which is *not* an open function.

(b) Give an example of a continuous function $f : \mathbb{R} \to \mathbb{R}$ which is *not* a closed function.

VIII - 20 (a) Prove or Disprove If f_1 and f_2 are hybrid functions defined on a closed bounded interval $[a, b]$, then their sum, $f_1 + f_2$, is hybrid on $[a, b]$.

(b) Prove or Disprove If f_1 and f_2 are hybrid functions defined on a closed bounded interval $[a, b]$, then their product, $f_1 \cdot f_2$, is hybrid on $[a, b]$.

VIII - 21 Suppose that $f : [a, b] \to \mathbb{R}$ is of bounded variation on $[a, b]$.

(a) Prove or Disprove The function $|f|$ is of bounded variation on $[a, b]$.

(b) Prove that if there is a constant $c > 0$ such that $f(x) \geq c$ for all x in $[a, b]$, then $1/f$ is of bounded variation on $[a, b]$.

VIII - 22 Suppose that $\varphi = (f_1, f_2, \dots)$ is a sequence of functions which converges pointwise on I to a function f.

(a) Prove or Disprove If each function f_k is monotonic up on I then so is f.

(b) Prove or Disprove If each function f_k is of bounded variation on I then so is f.

VIII - 23 Suppose that $g,\ f_1,\ f_2, \dots\ f_k, \dots$ are all monotonic up on an open interval (a, b). Suppose further that $\lim_{k \to \infty} f_k(x) = g(x)$ for every *rational* number in (a, b). Prove that if c is an irrational number in (a, b) such that g is continuous at c, then $\lim_{k \to \infty} f_k(c) = g(c)$.

VIII - 24 Let $f : [a, b] \to \mathbb{R}$ be a function of bounded variation on the closed bounded interval $[a, b]$. Assume also that f also has the 'Intermediate-Value Property' on $[a, b]$; that is, for each x_1 and x_2 in $[a, b]$ with $x_1 < x_2$, if y is between $f(x_1)$ and $f(x_2)$ then there exists c in (x_1, x_2) such that $f(c) = y$.

Problem Show that f is continuous on $[a, b]$.

VIII - 25 Prove the claim made in Remark (1) of F.6.23 on Page 351

Chapter IX

Infinite Sums (and Products) of Numbers

Quotes for Chapter (IX):

(1) $1+2+4+8+\ldots+2^k+\ldots\ =-1$

(Formula often attributed to the great 18th-century Swiss mathematician Leonhard Euler)

(2) $1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots+\frac{1}{2^k}+\ldots\ =+\infty$

(The so-called 'Zeno's Formula'; associated with the famous 'Achilles and the Tortoise' paradox of Zeno of Elea c. 450BC)

(3) $\frac{\pi}{2}\ =\ \frac{1{\cdot}2{\cdot}2{\cdot}4{\cdot}4{\cdot}\ldots{\cdot}(2k)(2k){\cdot}\ldots}{3{\cdot}3{\cdot}5{\cdot}5{\cdot}\ldots{\cdot}(2k-1){\cdot}(2k-1)\ldots}$

(Formula attributed to the 16-th century English mathematician John Wallis)

(4) 'The Method of Fluxions and Infinite Series, with its Applications to the Geometry of Curve-Lines'

(Title of the 1736 translation – from Latin into English – by John Colson, of Newton's 1671 (unpublished) book on calculus. 'Fluxion' was Newton's terminology for what we call 'derivative'.)

Introduction

Almost everyone who has a reason to read *This Textbook* has encountered, in a course in elementary calculus, the concept of 'an infinite sum of numbers', but probably under the title of 'an infinite *series* of numbers'. In most such calculus courses the theory of infinite series is based on a previous treatment of infinite *sequences* of numbers. The level of rigor one encounters in the treatment of sequences in such a course tends to be much lower than that found in *This Textbook*; but if one accepts those results on sequences, then the proofs of the theorems for infinite series based on them are usually correct. It follows that we *could* base the treatment of 'infinite series' in *This Textbook* directly on the theory of sequences which is developed in Chapter (III). Such an approach would have both advantages and disadvantages:

(i) The main advantage of such an approach here would be that everything would look familiar: the statements and proofs of theorems here would be identical with the analogous treatment in elementary calculus.

(ii) The main *dis*advantage of such an approach here would be that everything would look familiar: the reader who actually paid attention to the statements and proofs of theorems in that earlier course would see essentially nothing new here and thus would learn nothing new.

Instead, in *This Textbook* we approach these ideas from a more advanced viewpoint which normally is not stressed in elementary calculus courses.

Regardless of what approach is taken, the first issue to be handled is to respond to an obvious question:

'Why would anyone wish to form the sum of infinitely many numbers?'

The appropriate response is that such sums arise fairly naturally, even in 'applied' areas.

IX.0.1 Examples

(1) Every grade-school student is taught about the decimal representation of numbers. For instance,

$$\frac{1}{3} = 0.333333\ldots,$$

where the string of 'dots' indicates that the expression on the right side of the equation 'goes on forever'.

What does the expression $0.333333\ldots$ mean? For example, the *finite* expression 0.333 is a shorthand for a certain *finite* sum:

$$0.333 = \frac{3}{10} + \frac{3}{100} + \frac{3}{1000}$$

In the same spirit, the 'infinite decimal' $0.333333\ldots$ should indicate the corresponding*infinite* sum:

$$0.333333\ldots = \frac{3}{10} + \frac{3}{10^2} + \frac{3}{10^3} + \ldots + \frac{3}{10^k} + \ldots$$

That is, one can express the fraction $1/3$ as an infinite sum:

$$\frac{1}{3} = \frac{3}{10} + \frac{3}{10^2} + \frac{3}{10^3} + \ldots + \frac{3}{10^k} + \ldots$$

(2) In Section (**??**) we saw that if $f : \mathbb{R} \to \mathbb{R}$ is a C^∞ function and c is a number in $\mathbb{R}$, then one can write the Taylor Approximation

$$f(x) = f(c) + f'(c)(x-c) + \ldots + \frac{f^{(k)}(c)}{k!} + E_k(x),$$

where the remainder $E_k(x)$ is given by Taylor's Formula with Remainder. We also saw in a few cases (the exponential, sine and cosine functions with $c = 0$) that $\lim_{k\to\infty} E_k(x) = 0$. Such examples lead one to consider the possibility of getting an exact formula for $f(x)$ as an 'infinite sum' by letting $k \to \infty$, thereby eliminating the error E_k entirely by pushing it 'off to infinity':

$$f(x) = f(c) + f'(c)(x-c) + \frac{f''(c)}{2}(x-c)^2 + \ldots + \frac{f^{(k)}(c)}{k!} + \ldots$$

For instance, the analysis carried out in Theorem (**??**) suggests the following formulas:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \frac{x^k}{k!} + \ldots$$

$$\sin x \,=\, x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + (-1)^{k-1}\frac{x^{2k-1}}{(2k-1)!} + \ldots$$

$$\cos x \,=\, 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + (-1)^{k-1}\frac{x^{2k-2}}{(2k-2)!} + \ldots$$

The use of such 'infinite sums' has been one of the primary tools of calculus from the earliest days. Indeed, as the book title referred to in Chapter Quote (4) hints, Newton's concept of we now call a 'function' seemed to include that it can be expressed by an infinite sum of simple functions.

Needless to say, we could consider many more such examples from applied mathematics; but to save time let us simply assume that we all agree it is important to consider infinite sums in mathematics, and then move on to describe what such a sum might mean.

There are two main approaches to the theory of infinite sums.

(A) The 'Ordered Sum' Approach One thinks of an 'infinite sum' as the result of successively adding numbers, one after the other, – hence the name 'Ordered Sum' – until all the numbers to be added are exhausted. This is the natural extension of the idea of 'finite ordered sum' described in Definition (II.1.4), except that the additions need not stop. Of course with this approach one needs to impose on the numbers being added the order in which the addition is to take place.

The example $\frac{1}{3} = \frac{3}{10} + \frac{3}{10^2} + \frac{3}{10^3} + \ldots + \frac{3}{10^k} + \ldots$ given above illustrates the analog, for infinite sums, of the 'ordered sum' approach: one orders the numbers being added in terms of the exponents k in the denumerators of the fractions $3/10^k$.

(B) The 'Unordered Sum' Approach In this approach one need not impose an order on the numbers being added.

To make the distinction between these approaches clearer to the intuitiion, consider the following finite situation:

One has a nonempty box B containing n distinct physical objects, with each object x in the box having positive physical mass $m(x)$.

Problem Determine the total mass M of the objects in the box B.

Solution 1 Remove the n objects in the box B in some order: x_1 is the first object removed, x_2 is the second ,... x_n is the final one removed. Then form the (finite) ordered sum $m(x_1) + m(x_2) + \ldots + m(x_n)$, as in Definition (II.1.4). The result is the desired mass M.

Solution 2 Place the entire box B on, say, the left side of a balance scale; adjust for the weight of the box when empty. Then place enough standard-sized weights on the right side to overcome the weight of the box on the left. Replace standard weights on the right side by less heavy standard weights, and keep doing this until the two sides come into balance. Note that this 'weighing' procedure does not require that one list out the objects in the box B, or even to know the number n or the individual weights of these objects.

Note that Solution 1 uses an ordered sum, but in fact the choice of the order in which the the objects are removed from the box is irrelevant.

In elementary calculus one normally studies infinite 'Ordered Sums' in some depth, but 'Unordered Sums' not at all. However, the theory for 'Unordered Sum' is actually simpler, so in *This Textbook* we begin with it.

Final Note It turns out that the theory of 'infinite sums of real numbers' provides us, more or less for free, a corresponding theory for 'infinite products of real numbers'. This is why the phrase 'and Products' is included in the title of this chapter. That phrase is in parentheses, however, because the ‘infinite products’ portion plays a much less prominent role in *This Textbook* than does the ‘infinite sums’ part. The ‘products’ treatment is largely relegated to the exercises.

IX.1 Finite Unordered Sums

The theory of unordered infinite sums to be presented here is based on properties of *finite* sums. Thus it is useful to reformulate the finite theory in a way that makes the generalization to infinite sums simpler.

Recall that in Definition (II.1.4) the ordered sum of k real numbers begins with an ordered k-tuple $(x_1, x_2, \dots x_k)$ of real numbers. By Definition (**??**), such a k-tuple is a real-valued function $f : \mathbb{N}_k \to \mathbb{R}$ from the set $\mathbb{N}_k$ into $\mathbb{R}$; the ‘orderedness’ of the corresponding ordered sum $\sum_{j=1}^{k} x_j$ derives from the standard ordering on the domain $\mathbb{N}_k$ of the function f. The simplest way to formally introduce the idea of an ‘unordered’ finite sum of k real numbers is to replace the ordered set $\mathbb{N}_k$ by a general set X, not necessarily ordered in any standard way, with k elements.

IX.1.1 Definition

Let X be a finite nonempty set with exactly $k \geq 2$ elements, and let $f : X \to \mathbb{R}$ be a real-valued function whose domain contains X as a subset.

(1) The **(unordered) sum of the function values of f over the set X** is the number $\sum_X f$ given by

$$\sum_X f \;=\; f(g(1)) + f(g(2)) + \dots + f(g(k)),$$

where g is any bijection of $\{1, 2, \dots k\}$ onto X. It is sometimes convenient to use an ‘index’ notation such as $\sum_{x \in X} f(x)$ instead of $\sum_X f$.

(2) The **product of the function values of f over X** is the number $\prod_X f$ given by

$$\prod\nolimits_X f \;=\; f(g(1)) \cdot f(g(2)) \cdot \dots \cdot f(g(k)).$$

As with unordered sums, it is sometimes convenient to use an ‘index’ notation, such as $\prod_{x \in X} f(x)$, instead of $\prod_X f$.

(3) If X is a singleton set, $X = \{c\}$, then it is convenient to set $\sum_X f = \prod_X f = f(c)$. Likewise, it is convenient to write $\sum_\emptyset f = 0$ and $\prod_\emptyset f = 1$ for every function f.

IX.1.2 Remarks

(1) The conventions that $\sum_\emptyset f = 0$ and $\prod_\emptyset f = 1$ do not conflict with the requirement that the domain of a function must be a nonempty set. Indeed, in the preceding definition it is required only that the set X over which the sum and product are to be taken be a *subset* of the domain of the function f. The empty set fulfills that requirement automatically.

(2) Theorem (II.1.5) shows that the expressions $\sum_X f$ and $\prod_X f$ depend only on f and X, but not on the specific choice of bijection g; that is, they do not depend on the order in which one adds or multiplies the values of f. Note that is the *expressions* – that is, the manner in which the sums and products are written down – that are being distinguished as either ‘ordered’ or ‘unordered’ here, not the *values* assigned to these expressions; indeed, the main conclusion of Theorem (II.1.5)

is that the values are the same, at least in the case of finite sums. We shall see, however, that in the generalization to infinite sums the commutative law is not true in general.

(4) In the special case $X = \mathbb{N}_k$ the notations for the two types of sums take the form $\sum_{j\in\mathbb{N}_k} f(j)$ (unordered sum) and $\sum_{j=1}^{k} f(j)$ (ordered sum). When using the former notation we ignore the standard order on $\mathbb{N}_k$; when using the latter notation we use the standard order.

Needless to say, the commutative law does not have much significance in the context of unordered sums. In contrast, the associative laws for addition and multiplication also have versions which apply to the unordered situation.

IX.1.3 Theorem (Generalized Associative Laws for Unordered Finite Sums and Products)

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty finite set X with exactly k elements. Let $C_1, C_2, \ldots C_r$ be nonempty subsets of X that are mutually disjoint and whose union is X. Then

$$\sum\nolimits_X f = (\sum\nolimits_{C_1} f) + (\sum\nolimits_{C_2})f + \ldots + (\sum\nolimits_{C_r} f) \qquad \text{(IX.1)}$$

and

$$\prod\nolimits_X f = (\prod\nolimits_{C_1} f)\cdot(\prod\nolimits_{C_2} f)\cdot\ldots\cdot(\prod\nolimits_{C_r} f). \qquad \text{(IX.2)}$$

Proof We shall verify Equation (IX.1); the proof for Equation (IX.2) is similar, and is left as an exercise.

Let A be the set of all r in $\mathbb{N}$ for which Equation (IX.1) is true.

Clearly $1\in A$; indeed, if $r = 1$ then $C_r = C_1 = X$, and the condition to be proved reduces to $\sum_X f = \sum_{C_1} f$, which is true because $C_1 = X$.

Next, suppose that $r = 2$, and suppose that C_1 and C_2 are nonempty mutually disjoint subsets of X such that $X = C_1 \cup C_2$. For convenience let $Y = C_1$ and $Z = C_2$. Let p be the number of points in Y, and let $y_1, y_2, \ldots y_p$ be these points. Likewise, let q be the number of points of Z, and let $z_1, z_2, \ldots z_q$ be these points. By Part (c) of Theorem (I.7.7), one then has $p + q = k$. Define $g : \mathbb{N}_k \to X$ by the rule

$$g(j) = \begin{cases} y_j & \text{if } 1 \leq j \leq p \\ z_{j-p} & \text{if } p+1 \leq j \leq k = p+q. \end{cases}$$

Then it is clear that g is a bijection of $\mathbb{N}_k$ onto X. To simplify the notation, let us write $x_j = g(j)$ for all j in $\mathbb{N}_k$. Thus one has

$$x_1 = y_1,\, x_2 = y_2,\, \ldots x_p = y_p,\, x_{p+1} = z_1,\, x_{p+1} = z_2,\, \ldots x_k = x_{p+q} = z_q.$$

By definition one has

$$\sum\nolimits_X f = x_1 + x_2 + \ldots + x_k.$$

Now from Part (b) of Theorem (II.1.5) it follows that

$$\sum\nolimits_X f = x_1 + x_2 + \ldots + x_p = (x_1 + x_2 + \ldots + x_p) + (x_{p+1} + x_{p+2} + \ldots + x_{p+q}) = \sum\nolimits_Y f + \sum\nolimits_Y f,$$

as required. In other words, $r = 2$ is in the set A.

Next, suppose that $r \in A$, with $r \geq 2$, and suppose that $C_1, C_2, \ldots C_r, C_{r+1}$ are mutually disjoint nonempty subsets of X whose union is X. Let $Y = C_1 \cup C_2 \cup \ldots \cup C_r$ and $Z = C_{r+1}$. Then the fact that 2 is in A implies that

$$\sum\nolimits_X f = \sum\nolimits_Y f + \sum\nolimits_Z f.$$

And the induction hypothesis that r is in A implies that

$$\sum\nolimits_Y f = \sum\nolimits_{C_1} f + \sum\nolimits_{C_2} f + \ldots + \sum\nolimits_{C_r} f.$$

Combining these facts with the definition of ordered finite sums, one finally obtains

$$\sum\nolimits_X f = \Big(\sum\nolimits_{C_1} f + \sum\nolimits_{C_2} f + \ldots + \sum\nolimits_{C_r} f\Big) + \sum\nolimits_{C_{r+1}} f = \sum\nolimits_{C_1} f + \sum\nolimits_{C_2} f + \ldots + \sum\nolimits_{C_r} f + \sum\nolimits_{C_{r+1}} f.$$

In other words, $r+1$ is also in A. Now the induction is complete, and $A = \mathbb{N}$ and the desired result follows.

Example Suppose that $X = \{a, b, c\}$ is a set with exactly three elements. Let $C_1 = \{a\}$ and $C_2 = \{b, c\}$. Then Equation (IX.1) takes the form

$$\sum\nolimits_X f = \sum\nolimits_{C_1} f + \sum\nolimits_{C_2} f;$$

that is,

$$\sum\nolimits_X f = f(a) + (f(b) + f(c)).$$

Likewise, if one applies Equation (IX.1) to the sets $C_1' = \{a, b\}$ and $C_2' = \{c\}$, one gets

$$\sum\nolimits_X f = \sum\nolimits_{C_1'} f + \sum\nolimits_{C_2'} f = (f(a) + f(b)) + f(c).$$

In particular,

$$f(a) + (f(b) + f(c)) = (f(a) + f(b)) + f(c).$$

If one sets $x = f(a)$, $y = f(b)$ and $z = f(c)$, then this result takes the more familiar form

$$x + (y + z) = (x + y) + z.$$

Similar applications of Equation (IX.2) imply

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

In other words, the simplest nontrivial case of this theorem reduces to the usual Associative Laws for Addition and Multiplication. It is for this reason the the results of Theorem (IX.1.3) are called '*Generalized* Associative Laws for Unordered Finite Sums'.

IX.1.4 Example

Consider the equation

$$1 + (4 + 1) + (3 + 2) = 1 + 1 + 2 + 3 + 4 \qquad (*)$$

Proving this directly from the axioms could be time consuming. Here is how to get it from the preceding theorems.

First, note that there are 5 terms in the expression on the left, so let X be a set with exactly 5 elements. To be definite, let $X = \{a, b, c, d, e\}$. Define $f : X \to \mathbb{R}$ by the rule $f(a) = 1$, $f(b) = 4$, $f(c) = 1$, $f(d) = 3$, $f(e) = 2$. Then note that

$$X = C_1 \cup C_2 \cup C_3,$$

where $C_1 = \{a\}$, $C_2 = \{b, c\}$, $C_3 = \{d, e\}$. Note that the left side of Equation $(*)$ equals

$$(\sum_{C_1} f) + (\sum_{C_2} f) + (\sum_{C_3} f).$$

Likeswise,

$$\sum_X f = f(a) + f(b) + f(c) + f(d) + f(e) = 1 + 4 + 1 + 3 + 2 = 1 + 1 + 2 + 3 + 4,$$

where the last equation follows from the General Commutative Law for Addition. The desired equation now follows by applying the Generalized Associative Law.

IX.1.5 Remark

The careful reader may be bothered by the fact that the right side of Equation (IX.1) is written as an 'ordered sum', while the left side is an 'unordered sum'; see Remark (IX.1.2) (1) for the meanings of these phrases. This (minor) aesthetic flaw can easily be fixed by observing that the sets C_1, $C_2, \ldots C_r$ discussed in Theorem (IX.1.3) form a *partition* of the original set X; see Definition (**??**) (1). More precisely, let $\mathcal{F} = \{C_1, \ldots C_r\}$; then $\mathcal{F}$ is the partition in question. This suggests the following reformulation of Theorem (IX.1.3); the simple proof is left as an exercise for the reader.

IX.1.6 Theorem (Alternate Formulation of the Generalized Associative Laws for Unordered Finite Sums)

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a nonempty finite set X with exactly k elements. Let $\mathcal{F}$ be a partition of X, and define a function $\hat{f} : \mathcal{F} \to \mathbb{R}$ by the rule

$$\hat{f}(C) = \sum_C f \text{ for each } C \text{ in the family } \mathcal{F}.$$

Then

$$\sum_X f = \sum_{\mathcal{F}} \hat{f}. \tag{IX.3}$$

Likewise, let $\tilde{f} : \mathcal{F} \to \mathbb{R}$ be given by the rule

$$\tilde{f}(C) = \prod_C f \text{ for each } C \text{ in } \mathcal{F}.$$

Then

$$\prod_X f = \prod_{\mathcal{F}} \tilde{f}. \tag{IX.4}$$

IX.1.7 'Real-life' Example

The Story Line A certain person, whom we call 'Pat', owns a chain of three stores in Chicago. The following table shows the net profit for each store during the six-month period January 1 through June 30 of a certain year:

Month	Jan	Feb	Mar	Apr	May	Jun
Store A	1213	1502	1101	987	322	−206
Store B	−433	−106	224	783	1200	1305
Store C	213	714	949	851	770	312

Problem Determine the net profit that Pat received from the chain of stores during this period.

Solution First note that the entire situation can be formulated along the lines of the previous discussion. Indeed, let X be the set of all ordered pairs (u, v) where u is one of the three store letters A, B, C, and v is one of the six months Jan, Feb, Mar, Apr, May, Jun. Next, define $f : X \to \mathbb{R}$ by the rule $f(u, v)$ is the net profit of Store u during Month v. Then it is clear that the number Pat wishes to know is $\sum_X f$. This quantity is, by its nature, an 'unordered sum', since there is no uniquely natural way to list out the 18 numbers of the table to form a single ordered sum. However there are two obvious ways to 'partition' the data in the table:

Partition by 'Store' Let S_A be the subset of X corresponding to the six data points associated with Store A. That is

$$S_A = \{(A, \mathrm{Jan}), (A, \mathrm{Feb}), \ldots (A, \mathrm{Jun})).$$

Likewise, let S_B and S_C denote the subsets of X corresponding to Stores B and C, respectively. Clearly the sets S_A, S_B and S_C form a partition $\mathcal{F} = \{S_A, S_B, S_C\}$ of X. Note that the corresponding function $\hat{f} : \mathcal{F} \to \mathbb{R}$ is then given by

$$\hat{f}(S_A) = \sum\nolimits_{S_A} f = \text{ the net profit of Store A,}$$

and likewise for $f(\hat{S_B})$ and $\hat{f}(S_C)$. Theorem (IX.1.6) then takes the form

$$\sum\nolimits_X f = \sum\nolimits_{\mathcal{F}} \hat{f}.$$

This is then interpreted as saying that the net profit over the period for the entire chain is the sum of the profits of the individual stores over that period.

In terms of the 'table' notation used here: $\sum_{\mathcal{F}} \hat{f}$ calculates the sum of the entries of the table 'row-by-row'.

Partition by 'Month' Let T_{Jan} be the subset of X corresponding to the three data points associated with the month of January. That is,

$$T_{\mathrm{Jan}} = \{(A, \mathrm{Jan}), (B, \mathrm{Jan}), (C, \mathrm{Jan})\}.$$

Likewise, let $T_{\mathrm{Feb}}, \ldots T_{\mathrm{Jun}}$ denote the subsets of X corresponding to February, March etc. Clearly the family $\mathcal{G}$, given by

$$\mathcal{G} = \{T_{\mathrm{Jan}}, T_{\mathrm{Feb}}, T_{\mathrm{Mar}}, T_{\mathrm{Apr}}, T_{\mathrm{May}}, T_{\mathrm{Jun}}\},$$

forms a second partition of X. Now define $f^{\#} : \mathcal{G} \to \mathbb{R}$ by the rule

$$f^{\#}(T_{\mathrm{Jan}})f = \sum\nolimits_{T_{\mathrm{Jan}}} = \text{ the sum of the net profits from each of the three stores for January,}$$

and likewise for the other five months. In this case Theorem (IX.1.6) takes the form

$$\sum_X f = \sum_{\mathcal{G}} f^{\#}.$$

This is then interpreted as saying that the net profit of the chain over the given six-month period is equal to the sum of the net profits for each of these months of the chain.

In terms of the 'table' notation used here: $\sum_{\mathcal{G}} f^{\#}$ calculates the sum of the entries of the table 'column-by-column'.

Practical Note It is clear that the quantities $\sum_{\mathcal{F}} \hat{f}$ and $\sigma_{\mathcal{G}} f^{\#}$ must equal each other, since they are both equal to $\sum_X f$. Acountants use the fact that adding row-by-row ought to produce the same results as adding column-by-column as a way to (partially) verify the accuracy of their calculations.

Remark The formulation of the 'General Associative Law' in Theorem (IX.1.6), i.e., in terms of partitions, may seem artificial. However, we need to use an analog of this formulation in our treatment of 'infinite sums' in Chapter G. Also, such 'partition' formulations appear frequently in other branches of advanced mathematics.

The Generalized Commutative and Associative Laws can be used to prove the following facts which, although simple, are surprisingly useful.

IX.1.8 Theorem

(a) Let a and b be real numbers. If x_1, $x_2, \ldots$ x_k, are real numbers, then

$$b - a = (b - x_k) + (x_k - x_{k-1}) + (x_{k-1} - x_{k-2}) + \ldots + (x_2 - x_1) + (x_1 - a).$$

(b) Let c and d be nonzero real numbers. If y_1, $y_2, \ldots$ y_k are nonzero real numbers, then

$$\frac{d}{c} = \left(\frac{d}{x_k}\right) \cdot \left(\frac{x_k}{x_{k-1}}\right) \cdot \left(\frac{x_{k-1}}{x_{k-2}}\right) \cdot \ldots \left(\frac{x_2}{x_1}\right) \cdot \left(\frac{x_1}{a}\right)$$

Proof

(a) For convenience set $x_0 = a$ and $x_{k+1} = b$, so the equation to be proved takes the form

$$x_{k+1} - x_0 = (x_{k+1} - x_k) + (x_k - x_{k-1}) + (x_{k-1} - x_{k-2}) + \ldots + (x_2 - x_1) + (x_1 - x_0) \qquad (*)$$

Note that the expression on the right side of Equation $(*)$ involves $2k + 2$ indexed quantities:

$$x_0, x_1, \ldots x_k, x_{k+1}, -x_1, \ldots - x_k$$

(Recall that a term such as $x_2 - x_1$ really is an addition, namely $x_2 + (-x_1)$.) There does not appear to be a completely obvious choice of labeling set X and labeling function f for this situation. Based on the order in which the terms are written on the right side of Equation $(*)$, we choose $X = \{0, 1, 2, \ldots 2k + 1\}$ as the labeling set, and define the labeling function $f : X \to \mathbb{R}$ by the rule

$$f(j) = \begin{cases} x_{k+1-j} & \text{if } 0 \leq j \leq k \\ -x_{2k+1-j} & \text{if } k + 1 \leq j \leq 2k + 1 \end{cases}$$

This labeling corresponds to writing the numbers that appear in the desired sum in the following order:

$$x_{k+1}, x_k, x_{k-1}, \ldots x_2, x_1, -x_k, -x_{k-1}, \ldots - x_2, -x_1, -x_0.$$

Now let C_0, $C_2, \ldots C_k$ be the subsets of X that correspond under f to indices which appear in the differences in the parentheses on the right side of $(*)$:

$C_0 = \{0, k{+}1\}, \quad C_1 = \{1, k{+}2\}, \quad C_2 = \{2, k{+}3\}, \quad \cdots C_{k-1} = \{k{-}1, 2k\}, \quad C_k = \{k, 2k{+}1\}$

The general rule is $C_j = \{j, k+1+j\}$ for $0 \leq j \leq k$. Note that $f(j) + f(k+1+j) = x_{k+1-j} + (-x_{k-j}) = x_{k+1-j} - x_{k-j}$. Thus the right side of Equation $(*)$ equals

$$\sum\nolimits_{C_0} f + \sum\nolimits_{C_1} f + \ldots + \sum\nolimits_{C_k} f.$$

Since the sets C_0, $C_1, \ldots C_k$ are disjoint nonempty subsets of X whose union is X, it then follows from the Generalized Associative Law for Addition that the right side of Equation $(*)$ equals $\sum_X f$.

Next, define new subsets C'_0, $C'_1, \ldots C'_k$ of X as follows:

$$C'_0 = \{0, 2k+1\}, \quad C'_1 = \{1, k+1\}, \quad C'_2 = \{2, k+2\}, \ldots C'_k = \{k, 2k\}.$$

Much as above, one sees that $f(0)+f(2k+1) = x_{k+1} - x_0 = b - a$, $f(1)+f(k+1) = x_k - x_k = 0$, $f(2) + f(k+2) = x_{k-1} - x_{k-1} = 0, \ldots f(k) + f(2k) = x_1 - x_1 = 0$. Thus the sum

$$\sum\nolimits_{C'_0} f + \sum\nolimits_{C'_1} f + \ldots + \sum\nolimits_{C'_k} f$$

equals the sum $(b - a) + 0 + 0 + \ldots + 0$; that is, the value is $b - a$, which of course is the *left* side of Equation $(*)$. But it is clear that the sets C'_0, $C'_1, \ldots$ C'_k are also nonempty mutually disjoint subsets of X whose union is X. Thus the Generalized Associative Law for Addition applies once again to imply that the left side of Equation $(*)$ also equals $\sum_X f$.

Since, as just been shown, both sides of Equation $(*)$ equal the same quantity, namely $\sum_X f$, it follows that the two sides of this equation equal each other; which is a fancy way of saying that the equation is true.

(b) The proof of this part is similar, and is left as an exercise.

Remark These equations should bring back happy memories from high-school algebra: they form the basis for the classic 'Add-and-Subtract' Trick (Part (a)) and the 'Multiply-and-Divide Trick' (Part (b)).

The preceding results keep addition and multiplication separated. The next result combines both operations.

IX.1.9 Theorem (Generalized Distributive Law)

Suppose that $f_1, f_2, \ldots f_m : X \to \mathbb{R}$ are real-valued functions defined on a nonempty finite set X, and that c_1, $c_2, \ldots$ c_m are real numbers. Then:

$$\sum\nolimits_X (c_1 f_1 + c_2 f_2 + \ldots + c_m f_m) = c_1 \cdot \sum\nolimits_X f_1 + c_2 \cdot \sum\nolimits_X f_2 + \ldots + c_m \cdot \sum\nolimits_X f_m \qquad \text{(IX.5)}$$

In particular, one has as the following: $\sum_X (-f) = -\sum_X f$.

The simple proof is left as an exercise.

To see how this result generalizes the usual Distributive Law (see Axiom A3), consider the special case in which $m = 1$ and the set X has exactly two elements: $X = \{a, b\}$. Then on the left side of the equation one has

$$\sum\nolimits_X c_1 f_1 = c_1 f_1(a) + c_1 f_1(b),$$

while on the right side one has

$$c_1 \sum\nolimits_X f_1 \;=\; c_1(f_1(a) + f_1(b)).$$

The resulting equality $c_1 f_1(a) + c_1 f_1(b) \;=\; c_1(f_1(a) + f_1(b))$ is the Distributive Law.

The preceding calculations illustrate an important feature of proving well-known facts directly from the axioms: the process, although certainly instructive, can be quite tedious. It is also quite annoying, since we are proving results that we have been using, sometimes unconsciously, almost all our lives. From this point on, however, we shall follow the usual custom and normally leave such low-level proofs to the reader. Sometimes these proofs will be omitted without comment, sometimes with a blithe statement such as 'Now simplify to get ... '. In any event, the reader is left the task of filling in the gaps in such calculations, often by invoking the Generalized Commutative, Associative and Distributive Laws.

Pedagogical Comment The axioms listed above all involve standard algebraic facts about real numbers that everyone has known – and used, often without realizing it – since grade school. Because of that familiarity, there is a tendency to think of these axioms as 'obvious' or even 'trivial'; Chapter Quote 1 suggests a reason for this tendency.

Nevertheless, the rather bland presentation of these axioms does hide some nonobvious facts. For example, consider the following equation:

$$25350{\times}47 \;=\; 325{\times}3666 \qquad (*)$$

The odds are high that most people would not find this equation as being 'obviously true'; indeed, most would have to verify it by simply performing the indicated multiplications and noting that each ends up equaling 1191450. However, one easily checks that $25350 \;=\; 325{\cdot}78$ and $3666 \;=\; 78{\cdot}47$, so that Equation $(*)$ is just a rewriting of the following special case of the Associative Law for Multiplication:

$$(325{\cdot}78){\cdot}47 \;=\; 325{\cdot}(78{\cdot}47)$$

If the special case $(*)$ of this Associative Law is not 'obvious', then it would appear that the general Associative Law should be even less 'obvious'.

So why do we accept Axiom A2 as 'obvious'? Almost certainly *not* because we have tried out this axiom on dozens of triples x, y and z and verified that it works, much as we just did with Equation $(*)$. The likely answer to this question is that we accept this axiom because in grade school 'Teacher said it is so', and we alwys believe our teachers (at least in grade school).

IX.2 Unordered Infinite Sums of Real Numbers

Preliminary Comments

We wish to extend the concept of 'sum of the values of a function $f : X \to \mathbb{R}$ over the set X', described for finite sets X in Definition (IX.1.1), to the case in which X is infinite. In symbolic terms: we wish to make sense of the notation $\sum_X f$ even when X is infinite.

Unfortunately, it seems to be impossible to formulate a 'reasonable' definition of such a sum for *arbitrary* real-valued functions defined on an infinite nonempty set, even if one allows the possibility of the 'values' $+\infty$ and $-\infty$. As will become evident, the main difficulty involves infinite sums in which infinitely many of the terms are positive and infinitely many are negative. Because of this, we start with the special case in which $f(x) \geq 0$ for all x in X. In symbols: Our first goal is to make sense, as best possible, of the expression $\sum_X f$ when X is an arbitrary nonempty set and $f : X \to \mathbb{R}$ is a *nonnegative* function on X. Once that is done, the next goal would then be to build on this to extend the ideas to as wide a class of functions as possible.

Remark The approach to be followed below can be thought of as the obvious extension of the practice long used in doing finite sums ‘by hand. Indeed, when adding a *finite* list of numbers ‘by hand’, the simplest way is to first compute the sum of the positive numbers in the list, then the sum the negative numbers, but with the minus signs ignored; finally, subtract the latter sum from the former to obtain the desired result. This reflects the reality that, in hand calculations at least, subtractions are more difficult to perform than are additions: this method requires only one subtraction.

Note that even in the area of modern numerical analysis, in which the additions and subtractions are carried out by high-speed computers, so the difference in effort between subtraction and addition becomes negligible, the recommendation is still to add the positive and negative terms separately; but now the reason is to minimize the loss of significant digits.

Temporary Notation In order to minimize the possibilities for confusion between the ‘finite sum’ and ‘infinite sum’ cases, for the remainder of these ‘Preliminary Comments’ we use the symbolism $\overline{\sum}_X f$ to denote the quantity already described in Definition (IX.1.1). That is, if X is a finite nonempty set with exactly k distinct elements, then in what follows we set

$$\overline{\sum}_X f \;=\; f(g(1)) + f(g(2)) + \,\ldots\, + f(g(k)),$$

where g is any bijection of the set $\mathbb{N}_k \;=\; \{1, 2, \ldots k\}$ onto X.

Let us list some ‘guidelines’ that will help us choose an appropriate definition.

Situation $f : X \to \mathbb{R}$ is a real-valued function, defined on a nonempty set X, such that $f(x) \geq 0$ for all x in X. Then any reasonable rule for assigning a definite ‘value’ to the expression $\sum_X f$ ought to satisfy the following conditions:

Guideline 1 In the special case in which X is a finite set, then the rule for $\sum_X f$ should assign same value $\overline{\sum}_X f$ given by Definition (IX.1.1). In terms of the ‘Temporary Notation’ given above:

$$\sum_X f \;=\; \overline{\sum}_X f \text{when } X \text{ is a finite nonempty set.}$$

Guideline 2 The value assigned by the rule to the expression $\sum_X f$ should satisfy the inequality

$$0 \leq \sum_X f \;\leq\; +\infty.$$

Guideline 3 If Y is a nonempty subset of X, then the rule should imply that

$$\sum_Y f \leq \sum_X f.$$

In particular, if the subset Y satisfies $\sum_Y f \;=\; +\infty$, then one ought to have $\sum_X f \;=\; +\infty$.

Guideline 4 Let S_1, $S_2, \ldots\, S_r$ be nonempty subsets of X which are mutually disjoint and whose union is X. Assume also that for each $j \;=\; 1, 2, \ldots k$ the rule assigns a *finite* value to each expression $\sum_{S_j} f$, $1 \leq j \leq r$, are all finite. Then the rule ought to assign to the expression $\sum_X f$ the value

$$\sum_X f \;=\; \left(\sum_{S_1} f\right) + \left(\sum_{S_2}\right) f + \,\ldots\, + \left(\sum_{S_r} f\right)$$

In particular, in this case $\sum_X f$ should also be finite.

Remarks on the Guidelines

(1) Guideline 1 simply reaffirms that we are *extending* the definition of $\sum_X f$ from the case in which it was already defined to a wider class of situations.

(2) The reason for including the possibility $\sum_X f = +\infty$ in Guideline 2 comes from examples such as $f : \mathbb{N} \to \mathbb{R}$ in which $f(k) = 1$ for all k in $\mathbb{N}$. Any reasonable definition of $\sum_{\mathbb{N}} f$ for this f ought to correspond intuitively to the equation $1 + 1 + \ldots + 1 + \ldots = +\infty$. And of course the reason for *not* allowing $\sum_X f$ to be less than zero is because, in the situation under consideration we assume that $f(x) \geq 0$ for all x. (However, the fact that the great Euler was willing to write the equation appearing in Chapter Quote (1) might give one cause for worry about this guideline. More on that later.)

(3) Guidelines 3 and 4 simply assert that some properties which are obvious for finite sums ought to extend to infinite sums, at least in the case of nonnegative summands. In particular, Guideline 4 corresponds to the Generalized Associative Law for Addition (see Theorem (IX.1.3)).

(4) Note that in Guideline 4, only the *sums* $\sum_{S_j} f$, $1 \leq j \leq r$, are asssumed to be finite. The *sets* S_j are still allowed to have infinitely many elements. Indeed, that is the only case of real interest, since if all the sets S_j, $1 \leq j \leq r$, are finite sets then so is X; then by Guideline 1, the result would revert to the finite case, which was already studied in Section (II.1).

Note One could of course come up with other Guidelines which any 'reasonable' definition of $\sum_X f$ ought to satisfy. The four guidelines given here, however, are enough for our purposes. In particular, they suffice to point us towards the definition given below.

The connecting link between the guidelines just given and the official definition given below is the following:

Observation Suppose that we have defined a notion of $\sum_X f$, at least for nonnegative functions f, which satisfies the preceding guidelines. Let $U_{X;f}$ denote the set of all real numbers of the form $\overline{\sum}_W f$, where W is a nonempty *finite* subset of X. Then

$$\sum\nolimits_X f \geq \sup U_{X;f}.$$

Proof of Observation First note that the set $U_{X;f}$ is nonempty. Indeed, let c be any element of X; such c exists because X itself is nonempty. Let $W = \{c\}$, and note that (by Definition (IX.1.1)) one has $\overline{\sum}_W f = f(c)$, so the number $f(c)$ is an element of $U_{X;f}$.

Next, it follows from Guidelines 1 and 3 that

$$\sum\nolimits_X f \geq \sum\nolimits_W f = \overline{\sum}_W f$$

for every finite nonempty subset W of X. Thus, $\sum_X f$ must be an upper bound for the set $U_{X;f}$. But $\sup U_{X;f}$ is the *least* of such upper bounds, so it follows that $\sum_X f \geq \sup U_{X;f}$, as claimed.

It follows from the preceding discussion that any 'reasonable' definition of '$\sum_X f$' for nonnegative functions f defined on an nonempty set X must satisfy

$$\sup U_{X;f} \leq \sum\nolimits_X f \leq +\infty.$$

These last inequalities describe the extremes of the possible 'reasonable' definitions of $\sum_X f$. It also suggests two obvious candidates for a 'reasonable' definition of $\sum_X$.

The Maximal Candidate Suppose that $f : X \to \mathbb{R}$ is a nonnegative function defined on an nonempty set X. Then define

$$\sum\nolimits_X^{\max} f = \begin{cases} 0 & \text{if } f(x) = 0 \text{ for all } x \text{ in } X \\ \overline{\sum}_W f & \text{if the set } W = \{x : f(x) > 0\} \text{ is finite and nonempty} \\ +\infty & \text{if } f(x) > 0 \text{ for infinitely many } x. \end{cases}$$

The Minimal Candidate Suppose that $f : X \to \mathbb{R}$ is a nonnegative function defined on a nonempty set X. Let $U_{X;f}$ be as above, and define

$$\sum\nolimits_X^{\min} f \;=\; \sup U_{X;f}.$$

It can be shown that both the 'Maximal Candidate' $\sum_X^{\max}$ and the 'Minimal Candidate' $\sum_X^{\min}$ satisfy the four guidelines. However, the 'Maximal' case is of little interest; indeed, it assigns the *same* value, $+\infty$, to every sum of infinitely many positive numbers.

In contrast, the 'Minimal Candidate' has $\sum_X f$ being finite in as many situations as is possible. Thus, we use it in Section (IX.2) below as our 'official' definition of $\sum_X f$ for positive functions f.

It has been mentioned above that the concept of the 'Unordered Sum' of the values of a positive function $f : X \to \mathbb{R}$ corresponds intuitively to a process of 'weighing', at least in the case of finite sums. More precisely, think of X as being a collection of physical objects, and for each x in X let $f(x)$ denote the weight of the object x. Recall how the process of 'weighing' works: place the collection X on the left-hand tray of the scale, and then place on the right-hand tray a stack of standard weights that is at least as heavy X. Remove standard weights from the right tray until the two trays come into balance; at that point, the total weight in the right-hand tray equals the total weight of X. More precisely, let $V_{X;f}$ denote the set of numbers v such that v is larger than the total mass of X; in order for X to have finite total mass, the set $V_{X;f}$ must be nonempty. The process of removing weights from the right-hand tray then corresponds to the mathematical process of determining the infimum of the set $V_{X;f}$.

The key to this weighing process is to have a way of knowing that a given stack of standard weights is larger than the total mass of X. The basic assumption behind the 'minimal candidate' is that any number which is at least as large as all the numbers in $U_{X;f}$ must be in $V_{X;f}$. In other words, $V_{X;f}$ is the set of all upper bounds of $U_{X;f}$. Since $\sup U_{X;f}$ is the *least* of the upper bounds of $U_{X;f}$, it follows that $\sum_X^{\min} f \;=\; \inf V_{X;f}$; that is, $\sum_X^{\min}$ is the result of removing the excess weights from the right-hand tray until balance is reached.

The obvious way to extend the concept of $\sum_X^{\min} f$ from the case in which f is nonnegative on X to the case of arbitrary f is to use the results of Corollary (**??**) That is, set

$$\sum\nolimits_X^{\min} f \;=\; \left(\sum\nolimits_X^{\min} f^+\right) - \left(\sum\nolimits_X^{\min} f^-\right) \qquad (*)$$

where the values $\sum_X^{\min} f^+$ and $\sum_X^{\min} f^-$ are defined as above for nonnegative functions.

Unfortunately, there is a problem which can arise when X is an infinite set: the quantities $\sum_X^{\min} f^+$ and $\sum_X^{\min} f^-$ might both equal $+\infty$. That is, Equation $(*)$ might take the form $\sum_X^{\min} f \;=\; \infty - \infty$. However, as is well known from elementary calculus, the expression $\infty - \infty$ is an 'indeterminate form' which cannot be assigned a definite value as written.

This difficulty is a real one, and cannot be completely resolved. However, if at least one of the quantities $\sum_X^{\min} f^+$ or $\sum_X^{\min} f^-$ is finite, then a definite answer can be given. For example, if $\sum_X^{\min} f^+ \;=\; +\infty$ but $\sum_X^{\min} f^-$ is finite, then Equation $(*)$ takes the form

$$\infty - \text{ a finite number.}$$

It is natural to assign the value $+\infty$ to that expression.

The fruit of these considerations is summarized in the 'official' definition given below. In it, we drop the provisional 'min' from the notation $\sum_X^{\min}$, since its only purpose was to distinguish itself from the other candidate, $\sum_X^{\max}$.

IX.2.1 Definition (Unordered Infinite Sums)

Let X be a nonempty set of real numbers.

(1) Suppose that $g : X \to \mathbb{R}$ is a nonnegative function on X; that is, $g(x) \geq 0$ for all x in X. Let $U_{X;g}$ denote the set of all numbers of the form $\sum_W g$, where W is a finite nonempty subset of X, and $\sum_W g$ is the corresponding unordered finite sum, as described in Definition (IX.1.1). Then

$$\sum\nolimits_X g \;=\; \sup U_{X;g}.$$

(2) Suppose that $f : X \to \mathbb{R}$ is a real-valued function defined on X, and let f^+ and f^- be the corresponding positive and negative parts of f, as in Definition (**??**). Let $A \,=\, \sum_X f^+$ and $B \,=\, \sum_X f^-$, as defined in Part (1) above, with g replaced by f^+ and f^{-1}, respectively.

(i) If A and B are both finite, then $\sum_X f \,=\, A - B$.

(ii) If $A \,=\, +\infty$ and B is finite, then $\sum_X f \,=\, +\infty$. Likewise, if A is finite and $B \,=\, +\infty$, then $\sum_X f \,=\, -\infty$.

(3) Suppose that $f : X \to \mathbb{R}$ is a real-valued function, defined on a nonempty set X. Let $A \,=\, \sum_X f^+$, $B \,=\, \sum_X f^-$, as above.

(i) If at least one of the quantities A or B is finite then one says that **the unordered sum $\sum_X f$ is defined**, or that **the sum exists**. In contrast, if $A \,=\, B \,=\, +\infty$, then one says that **the unordered sum $\sum_X f$ is not defined**, or that it is the **indeterminate form** $\infty - \infty$.

(ii) Suppose, in addition, X is an infinite set, and assume that the infinite sum $\sum_X f$ is defined, as described in (i). If $\sum_X f \,=\, S$ for some finite number S, then one says that the unordered sum $\sum_X f$ **is convergent** and that $\sum_X f$ **converges to** S. In contrast, if $\sum_X f \,=\, +\infty$ or $\sum_X f \,=\, -\infty$, then one says that the unordered sum $\sum_X f$ **diverges to $+\infty$ or to $-\infty$**, respectively.

Note: In ordinary mathematical usage, statements such as '$\sum_X f$ is a finite sum', or '$\sum_X f$ is an infinite sum', can be ambiguous. For example, the first statement might mean that the sum involves only a finite number of terms (i.e., X is a finite set), while the second statement might mean that the sum involves infinitely many terms (i.e., X is an infinite set). However, these statements could well be interpreted as referring to the *value* assigned to the sum,; that is, the first statement might mean that the value is finite (even though X could be infinite), while the second could mean that the value is $+\infty$ or $-\infty$. In *This Textbook* we try avoid such ambiguity by using the phrases 'finite sum' and 'infinite sum' to indicate that the set X is finite or infinite, respectively; and to use 'convergent' and 'divergent to $\pm\infty$' for the alternate meanings.

IX.2.2 Remarks

(1) Note that we allow ourselves the freedom to write down expressions such as $\sum_X f$ *before* we know whether 'the sum exists'. This is similar to the freedom we allowed ourselves with limits; see Remark (III.1.7).

(2) The preceding definition provides a fairly conservative approach to infinite sums because it refuses to assign a value to the expression $\sum_X f$ when both of the quantities A and B in Part (2) are infinite. There are other, less conservative, approaches which sometimes assign a definite value in that case as well. Indeed, the 'Ordered Sum' approach to infinite sums, which is discussed in the next section, is an example of such a less conservative approach. There is an entire theory of such 'summation methods'.

(3) The preceding definition is said to be only 'fairly' conservative, because one could have chosen to use an approach that is drastically more so. For example, in his famous 'Achilles and the Tortoise' paradox, the ancient Greek philosopher Zeno of Elea appears to take the viewpoint that every infinite sum of positive terms should be treated as having value $+\infty$; that is, he chose the 'Maximal Candidate'.

(3) Definition (IX.2.1) allows us the flexibility of saying that certain unordered infinite sums are defined, but diverge (whether to $+\infty$ or $-\infty$). In practice, however, most of the interesting results involve unordered sums which are convergent to some (finite) value.

IX.2.3 Examples

(1) Let X be an infinite set and let c be a real number. Define $f : X \to \mathbb{R}$ by the rule $f(x) = c$ for all x in X. It is easy to show that $\sum_X f$ is defined, and that

$$\sum\nolimits_X f = \begin{cases} +\infty & \text{if } c > 0 \\ 0 & \text{if } c = 0 \\ -\infty & \text{if } c < 0. \end{cases}$$

Note The main point of this example is that the quantity $\sum_X f$ can be defined even if the set X is quite large; in particular, X is allowed to be uncountable.

(2) Let $X = \mathbb{N}$ and let $f : \mathbb{N} \to \mathbb{R}$ be given by the rule

$$f(k) = \frac{1}{k^2} \text{ for each } k \text{ in } \mathbb{N}.$$

Since $f(k) > 0$ for all k in $\mathbb{N}$, it follows that the infinite sum $\sum_{\mathbb{N}} f$ is certainly defined. The real issue is whether this sum is finite or infinite.

Let W be a finite nonempty subset of $\mathbb{N}$, and let k be the largest element of W. Then certainly one has $W \subseteq \mathbb{N}_k$, where $\mathbb{N}_k = \{m \in \mathbb{N} : 1 \leq m \leq k\}$ (see Definition (I.2.1)). Let $x_k = \sum_{\mathbb{N}_k} f$, so by Definition (IX.1.1) one can write

$$x_k = f(1) + f(2) + \ldots + f(k) = 1 + \frac{1}{4} + \frac{1}{9} + \ldots + \frac{1}{k^2}.$$

Clearly the sequence $\xi = (x_1, x_2, \ldots)$ satisfies $x_1 = 1$ and $x_{k+1} = x_k + 1/(k+1)^2$ for each k in $\mathbb{N}$. In other words, ξ satisfies the conditions of the sequence studied in Example (III.2.11) (1). It follows from the results of that example that $x_k < 2$ for all k. Thus one has

$$\sum\nolimits_W f \leq \sum\nolimits_{\mathbb{N}_k} f < 2,$$

so that 2 is an upper bound for the set $U_{\mathbb{N};f}$. Thus, by basic properties of 'supremum', it follows that

$$\sum\nolimits_{\mathbb{N}} f = \sup U_{\mathbb{N};f} \leq 2.$$

In particular, $\sum_{\mathbb{N}} f$ is finite.

The logical next question would be this: since we now know that $\sum_{\mathbb{N}} f$ is finite, what is its exact value? It turns out that this question can be answered, but only with considerable difficulty; we do not address this question here.

(3) Let $X = \mathbb{N}$ and let $f : \mathbb{N} \to \mathbb{R}$ be given by the rule

$$f(k) = \frac{1}{k} \text{ for each } k \text{ in } \mathbb{N}.$$

An analysis similar to that used in Example (2) above, but this time based on Example (III.2.11) (2), shows that $\sum_X f = +\infty$. Indeed, let k be any element of $\mathbb{N}$, and let $W_k = \mathbb{N}_{2^k} = \{1, 2, \ldots 2^k\}$. The results of Example (III.2.11) (2) tell us that

$$\sum\nolimits_{W_k} f = \sum\nolimits_{\mathbb{N}_{2^k}} f \geq \frac{k}{2}.$$

By the Archimedean Property, the fraction $k/2$ can become arbitrarily large; and it is clear that $U_{\mathbb{N};f}$ contains all the numbers of the form $\sum_{W_k} f$. Thus, it follows that $\sup U_{\mathbb{N};f} = +\infty$. That is, $\sum_{\mathbb{N}} f = +\infty$, as claimed.

(4) Let $X = \mathbb{N}$ and let r be a number such that $-1 < r < 1$. Define $g : \mathbb{N} \to \mathbb{R}$ by the rule $g(i) = r^{i-1}$. (Note that if $r = 0$ and $i = 1$, then the expression r^{i-1} becomes the indeterminate form 0^0; see the discussion after Lemma (**??**). As is stated in that discussion, we set this expression equal to 1.)

Case 1 Suppose that $r = 0$. Then $g(1) = 1$ while $g(i) = 0$ if $i \geq 2$; in particular, $g(i) \geq 0$ for all i in $\mathbb{N}$. It is clear that if W is any finite nonempty subset of $\mathbb{N}$, then either $\sum_W g = 0$ (namely, when $1 \notin W$), or $\sum_W g = 1$ (when $1 \in W$). It is clear from this that $\sum_{\mathbb{N}} g = 1$.

Case 2 Suppose that $0 < r < 1$, so that $g(i) > 0$ for all i in X. Let W be a finite nonempty subset of $\mathbb{N}$, and let k be the largest of the elements of W, so that $W \subseteq \mathbb{N}_k$. One can then use Part (a) of Theorem (II.2.16) to conclude that

$$\sum\nolimits_W g \leq \sum\nolimits_{\mathbb{N}_k} g = 1 + r + \ldots + r^{k-1} = \frac{1 - r^k}{1 - r} < \frac{1}{1 - r}.$$

In particular, $1/(1 - r)$ is an upper bound for the set $U_{\mathbb{N};g}$, so that $\sup U_{\mathbb{N};g} \leq \dfrac{1}{1-r}$. Likewise, one knows that for every $\varepsilon > 0$ there exists k such that $0 < r^k < \varepsilon(1 - r)$. For such k the set $W = \{1, 2, \ldots k\}$ satisfies

$$\sum\nolimits_W g = \frac{1 - r^k}{1 - r} > \frac{1 - \varepsilon(1 - r)}{1 - r} = \frac{1}{1 - r} - \varepsilon$$

and thus $\sup U_{\mathbb{N};g} > \dfrac{1}{1 - r} - \varepsilon$ for every $\varepsilon > 0$. It now follows that $\displaystyle\sum_{\mathbb{N}} g = \frac{1}{1 - r}$.

Case 3 Suppose that that $-1 < r < 0$, so that $r^{i-1} > 0$ if i is odd, and $r^{i-1} < 0$ if i is even. It follows that $g^+(2j + 1) = |r|^j$, and $g^+(2j) = 0$, for each j in $\mathbb{N}$. It then follows from the results of Case 2 above – but with r in that case replaced by r^2 – that the infinite unordered sum $\sum_{\mathbb{N}} g^+$ is convergent and has value $\dfrac{1}{1 - r^2}$. In a similar manner one can easily show that $\sum_{\mathbb{N}} g^- = \dfrac{(-r)}{1 - r^2}$. Thus, by Definition (IX.2.1) the unordered sum $\sum_{\mathbb{N}} g$ is convergent, and one has

$$\sum\nolimits_{\mathbb{N}} g = \sum\nolimits_{\mathbb{N}} g^+ - \sum\nolimits_{\mathbb{N}} g^- = \left(\frac{1}{1 - r^2}\right) - \left(-\frac{r}{1 - r^2}\right) = \frac{1 + r}{1 - r^2} = \frac{1}{1 - r}.$$

Summary: If $-1 < r < 1$, and $g(i) = r^{i-1}$ for each i in $\mathbb{N}$, then the unordered sum $\sum_{\mathbb{N}} g$ is convergent, and its value is $1/(1 - r)$.

The next result summarizes some basic properties of unordered sums.

IX.2.4 Theorem

Let $g : X \to \mathbb{R}$ be a real-valued function defined on a nonempty set X such that $g(x) \geq 0$ for all x in X. Then:

(a) The quantity $\sum_X g$ is defined (see Part (2) of Definition (IX.2.1)), and $\sum_X g \geq 0$. Furthermore, one gets $\sum_X g = 0$ if, and only if, $g(x) = 0$ for all x in X.

(b) Suppose that Y is a nonempty subset of X. Then $\sum_Y g \leq \sum_X g$. In particular, if $\sum_Y g = +\infty$ then $\sum_X g = +\infty$.

If, instead, the quantity $\sum_Y g$ is finite, and $\sum_Y g = \sum_X g$, then either $Y = X$ or, when $Y \neq X$, one has $g(x) = 0$ for all x in $X \backslash Y$. Conversely, if either $Y = X$ or $g(x) = 0$ for all x in $X \backslash Y$, then $\sum_Y g = \sum_X f$.

(c) A similar result holds if $g(x) \leq 0$ for all x in X, except that now one has $\sum_Y g \geq \sum_X g$. In particular, if $\sum_Y g = -\infty$ then $\sum_X g = -\infty$.

The simple proof is left as an exercise.

The next result allows one to determine directly, in terms of f, whether an unordered infinite sum $\sum_X f$ is convergent, without needing to first reduce everything to f^+ and f^-.

IX.2.5 Theorem

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a infinite set X.

Let L be a real number. Then the following statements are equivalent:

(i) The unordered sum $\sum_X f$ is convergent and has the value L in $\mathbb{R}$.

(ii) For every $\varepsilon > 0$ there exists a finite nonempty subset W_ε of X such that if W is any finite subset of X with $W \supseteq W_\varepsilon$, then

$$\left|L - \left(\sum\nolimits_W f\right)\right| < \varepsilon.$$

Alternate Formulation of (ii): For every choice of numbers y and z such that $y < L < z$, there exists a finite subset $W_{y;z}$ of X such that if W is any finite subset of X with $W \supseteq W_{y;z}$, then $y < \sum_W f < z$.

Proof First note that it is obvious that the two formulations of Statement (ii) are equivalent.

Suppose that Statement (i) is true. Then, by definition, the quantities $A = \sum_X f^+$ and $B = \sum_X f^-$ are finite. It follows, from the definitions of these quantities as $\sup U_{X;f^+}$ and $\sup U_{X;f^-}$ and the properties of 'sup', that there exist finite nonempty subsets W_1 and W_2 of X such that

$$A - \frac{\varepsilon}{2} < \sum\nolimits_{W_1} f^+ \leq A < A + \frac{\varepsilon}{2}$$

and

$$B - \frac{\varepsilon}{2} < \sum\nolimits_{W_2} f^- \leq A < A + \frac{\varepsilon}{2}.$$

Let $W_\varepsilon = W_1 \cup W_2$, and suppose that W is a finite subset of X such that $W \supseteq W_\varepsilon$. Then one has $W \supseteq W_1$, hence

$$A - \frac{\varepsilon}{2} < \sum\nolimits_{W_1} f^+ \leq \sum\nolimits_W f^+ \leq A < A + \frac{\varepsilon}{2};$$

the inequality $\sum_{W_1} f^+ \leq \sum_W f^+$ holding because $f^+(x) \geq 0$ for all x and $W \supseteq W_1$. Likewise, one has

$$B - \frac{\varepsilon}{2} < \sum\nolimits_{W_2} f^- \leq \sum\nolimits_W f^- \leq B < B + \frac{\varepsilon}{2}.$$

Multiply each term in this string of inequalities by (-1) to get

$$-B - \frac{\varepsilon}{2} < -B \leq -\left(\sum\nolimits_W f^-\right) < -B + \frac{\varepsilon}{2}$$

Combining these inequalities then yields

$$A - B - \varepsilon < \sum\nolimits_W f^+ - \sum\nolimits_W f^- < A - B + \varepsilon.$$

In light of Theorem (IX.1.8), combined with the fact that $f = f^+ - f^-$, one can rewrite the preceding as

$$A - B - \varepsilon < \sum\nolimits_W f < A - B + \varepsilon.$$

Since $A - B = L = \sum_X f$, this last result can be written

$$\left|\sum\nolimits_X f - \sum\nolimits_W f\right| < \varepsilon,$$

as required. Thus, Statement (i) implies Statement (ii).

Suppose, instead, that Statement (ii) is true.

Claim 1 The unordered sum $\sum_X f$ is convergent.

Proof of Claim 1 Choose $\varepsilon = 1$ in Statement (ii), and let W_1 be a finite nonempty subset of X such that if W is a finite subset of X satisfying $W \supseteq W_1$, then $|L - \sum_W f| < 1$.

Suppose $\sum_X f^+ = +\infty$. Let $D = \sum_{W_1} f^-$, so that D is a fixed nonegative real number. Then, by the hypothesis that $\sum_X f^+ = \sup U_{X;f^+}$, there must exist a finite nonempty subset W' of X such that

$$\sum\nolimits_{W'} f^+ \geq 1 + |L| + D$$

By properties of finite sums, we may ignore any terms in the sum on the left side of this inequality equal to 0, since such terms do not change the value of that sum. That is, we may assume, without loss of generality, that W' is chosen so that $f^+(x) > 0$ for all x in W'. It then follows that $f^-(x) = 0$ for all x in W'. Now let $W = W' \cup W_1$. Since $W \supseteq W'$, it is clear that

$$\sum\nolimits_W f^+ \geq \sum\nolimits_{W'} f^+ \geq 1 + |L| + D$$

Also, since $W \supseteq W_1$, it follows by Statement (ii) that

$$\left|L - \sum\nolimits_W f\right| < 1$$

Finally, because $f^-(x) = 0$ for all x in W', it follows from basic properties of finite sums that

$$\sum\nolimits_W f^- = \sum\nolimits_{W_1} f^- = D.$$

Next, use the fact that $f = f^+ - f^-$ and Theorem (IX.1.8) to see that

$$\sum\nolimits_W f^+ = \sum\nolimits_W f + \sum\nolimits_W f^- = \left(\sum\nolimits_W f - L\right) + L + \sum\nolimits_W f^- = \left(\sum\nolimits_W f - L\right) + L + D.$$

Apply the Triangle Inequality to this last equation, together with the inequalities obtained above, to get

$$1 + |L| + D \leq \sum\nolimits_W f^+ \leq \left|L - \sum\nolimits_W f\right| + |L| + D < 1 + |L| + D.$$

This implies the impossible inequality $1 + |L| + D < 1 + |L| + D$. That is, assuming that $\sum_X f^+ = +\infty$ leads to a contradiction. Likewise, assuming that $\sum_X f^- = +\infty$ would also lead to a similar contradiction. Thus, both of these unordered sums must be finite, hence $\sum_X f$ is defined and is finite, as claimed.

Claim 2 Let $A = \sum_X f^+$ and $B = \sum_X f^-$. Then $L = A - B$.

Proof of Claim 2 Let $\varepsilon > 0$ be given, and let W_ε be a finite nonempty subset of X such that if W is any subset of X satisfying $W \supseteq W_\varepsilon$ then $|L - \sum_W f| < \varepsilon/3$. Now let W_1 and W_2 be a finite subsets of X such that $0 \leq A - \sum_{W_1} f^+ < \varepsilon/3$, and $0 \leq B - \sum_{W_2} f^- < \varepsilon/3$. By an argument similar to that used in Part (a) of this proof, together with the hypothesis that Statement (ii) holds, one sees that

$$|A - \sum_W f^+| < \frac{\varepsilon}{3},\ |B - \sum_W f^-| < \frac{\varepsilon}{3},\ \text{and}\ |L - \sum_W f| < \frac{\varepsilon}{3}.$$

Next, notice that from the equation $f = f^+ - f^-$ one gets $\sum\limits_W f = \left(\sum\limits_W f^+\right) - \left(\sum\limits_W f^-\right)$, hence

$$(A - B) - L = \left(A - \sum_W f^+\right) - \left(B - \sum_W f^-\right) - \left(L - \sum_W f\right).$$

Now use the Triangle Inequality, together with the properties obtained above for the set W, to get

$$|(A - B) - L| \leq \left|A - \sum_W f^+\right| + \left|B - \sum_W f^-\right| + \left|L - \sum_W f\right| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3}.$$

In summary: For every $\varepsilon > 0$ one has $|(A - B) - L| < \varepsilon$. The only way this can happen is if $L = A - B$. In light of the definition of A and B, this can be written

$$L = \sum_X f^+ - \sum_X f^- = \sum_X f.$$

That is, Statement (ii) implies Statement (i), as claimed.

IX.2.6 Remarks

(1) Many authors use the property described in Statement (ii) above as the *definition* of what it means for the unordered sum $\sum_X f$ to converge to L; our definition would, for them, be a theorem to be proved. On the whole, the approach to 'unordered sums' based on the property in Staement (ii) seems fussier than the '$f = f^+ - f^-$' approach used in *This Textbook.* The 'Statement (ii)' approach does have one major advantage: it can be generalized to situations involving functions whose values are *not* real numbers, and thus for which the decomposition $f = f^+ - f^-$ may not make sense. (For those in the know: Think of functions $f : X \to B$, where B is a Banach space.)

(2) The equations '$\sum_X f = +\infty$' and '$\sum_X f = -\infty$' have similar characterizations in terms of finite sums of the form $\sum_W f$. The precise statements of these characterizations, and their proofs, are left as an exercise. (Hint: Use the 'Alternate Formulation of (ii)' as a guide.)

IX.2.7 Corollary

Let X be a nonempty set.

(a) Suppose that f and g are real-valued functions on X such that the unordered sums $\sum_X f$ and $\sum_X g$ are convergent. Then the sum $\sum_X (f + g)$ is also convergent, and one has

$$\sum_X (f + g) = \sum_X f + \sum_X g.$$

(b) Suppose that $f : X \to \mathbb{R}$ is a function such that $\sum_X f$ is convergent. Then for every real number c the sum $\sum_X c{\cdot}f$ is also convergent, and one has

$$\sum\nolimits_X c{\cdot}f \;=\; c\sum\nolimits_X f.$$

(c) More generally, suppose that $f_1,\, f_2, \ldots\, f_k$ are functions such that each of the ordered sums $\sum_X f_j$, $1 \le j \le k$, is convergent. Also, let $c_1, \;\ldots\, c_k$ be real numbers. Then the unordered sum $\sum_X (c_1{\cdot}f_1 + \,\ldots\, + c_k{\cdot}f_k)$ is convergent, and one has

$$\sum\nolimits_X (c_1{\cdot}f_1 + \,\ldots\, + c_k{\cdot}f_k) \;=\; c_1\sum\nolimits_X f_1 + \,\ldots\, + c_k\sum\nolimits_X f_k.$$

Proof

(a) Set $A \;=\; \sum_X f$, $B \;=\; \sum_X g$, and $C \;=\; A + B$. Let $\varepsilon \;>\; 0$ be a positive number. By Theorem (IX.2.4), there exist finite nonempty subsets Y_ε and Z_ε of X such that if Y and Z are finite subsets of X such that $Y \supseteq Y_\varepsilon$ and $Z \supseteq Z_\varepsilon$, then

$$\left|A - \sum\nolimits_Y f\right| \;<\; \frac{\varepsilon}{2} \text{ and } \left|B - \sum\nolimits_Z g\right| \;<\; \frac{\varepsilon}{2} \qquad (*)$$

Now set $W_\varepsilon \;=\; Y_\varepsilon \cup Z_\varepsilon$, so that W_ε is a finite nonempty subset of X. Suppose that W is a finite subset of X such that $W \supseteq W_\varepsilon$. Since the set W is finite, it follows from Theorem (IX.1.3) that $\sum_W (f + g) \;=\; \sum_W f + \sum_X g$. Using this, and the Triangle Inequality, one gets

$$|C{-}\sum\nolimits_W (f{+}g)| \;=\; |(A{+}B){-}\sum\nolimits_W (f{+}g)| \;=\; \left|\left(A - \sum\nolimits_W f\right) + \left(B - \sum\nolimits_W g\right)\right| \;\le\; \left|A - \sum\nolimits_W f\right| {+} \left|B - \sum\nolimits_W g\right|$$

Now apply Inequality $(*)$ to the right-most terms above to get

$$|C - \sum\nolimits_W (f + g)| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \;=\; \varepsilon.$$

Theorem (IX.2.5) now implies that $C \;=\; \sum_X (f + g)$.

(b) and (c): The simple proofs are left to the reader.

The unordered infinite sums of greatest interest in analysis are the *convergent* sums. The next result provides the most common tool for determining that a sum is convergent without necessarily determining the actual value of that sum.

IX.2.8 Theorem

Let $f : X \to \mathbb{R}$ be a real-valued function defined on an infinite set X. Then the following statements are equivalent:

(i) The unordered infinite sum $\sum_X f$ is convergent.
(ii) The unordered infinite sum $\sum_X |f|$ is convergent.
(iii) There exists a number $M \ge 0$ such that

$$\sum\nolimits_W |f| \le M \text{ for every finite nonempty subset } W \text{ of } X.$$

Proof Let $A \;=\; \sum_X f^+$ and $B \;=\; \sum_X f^-$.

Suppose that Statement (i) is true. Then, by definition, the quantities A and B are numbers, not $+\infty$. Since $|f| = f^{+} + f^{-}$, Part (a) of Corollary (IX.2.7) implies that $\sum_X |f|$ is a convergent and that

$$\sum_X |f| = \sum_X f^{+} + \sum_X f^{-} = A + B.$$

Thus, $\sum_X |f|$ is convergent; that is, Statement (ii) is true.

Suppose that Statement (ii) is true, and let $M = \sum_X |f|$. Then $M = \sup U_{X;|f|}$. In particular, M is an upper bound for the set of all numbers of the form $\sum_W |f|$, where W is any finite nonempty subset of X. That is, Statement (iii) is true.

Finally, suppose that Statement (iii) is true. Then the number M is an upper bound of the set $U_{X;|f|}$. Since $\sup U_{X;|f|}$ is the least of such upper bounds, it follows that $\sum_X |f| \leq M$. Also $|f| = f^{+} + f^{-}$ and $f^{+}(x) \geq 0$ and $f^{-}(x) \geq 0$ for all x in x. Thus it follows that $f^{+}(x) \leq |f|(x)$ and $f^{-}(x) \leq |f|(x)$ for all x in X. In particular, for every finite nonempty subset W of X one has

$$\sum_W f^{+} \leq \sum_W |f| \leq M \text{ and } \sum_W f^{-} \leq \sum_W |f| \leq M.$$

Thus, M is an upper bound for both of the sets $U_{X;f^{+}}$ and $U_{X;f^{-}}$, which implies that $\sum_X f^{+} \leq M$ and $\sum_X f^{-} \leq M$. In particular, $\sum_X f$ is convergent; that is, Statement (i) is true.

IX.2.9 Corollary

Let $f : X \to \mathbb{R}$ be a real-valued function on an infinite set X, and assume that the unordered sum $\sum_X f$ is defined (see Part (3) of Definition (IX.2.1)). If Y is a nonempty subset of X, then the unordered sum $\sum_Y f$ is also defined. Furthermore:

(i) If $\sum_Y f = +\infty$ then $\sum_X f = +\infty$.

(ii) If $\sum_Y f = -\infty$ then $\sum_X f = -\infty$.

(iii) If the sum $\sum_X f$ is convergent, then so is $\sum_Y f$.

The simple proof is left as an exercise.

The next result is the generalization, to infinite unordered sums, of Theorem (IX.1.6), the 'Alternate Formulation of the Generalized Associative Law (for finite sums)'.

IX.2.10 Theorem (Generalized Associative Law for Infinite Unordered Sums)

generalized associative law for infinite unordered sums

Let $f : X \to \mathbb{R}$ be a real-valued function on an infinite set X, and assume that the unordered sum $\sum_X f$ is defined. Let $\mathcal{F}$ be a partition of X into a family of disjoint nonempty subsets (see Definition (**??**)). Suppose that for every S in $\mathcal{F}$ the unordered sum $\sum_S f$ is convergent. (The fact that $\sum_S f$ is defined for each S in $\mathcal{F}$ follows from Corollary (IX.2.9) above.) Define a function $\hat{f} : \mathcal{F} \to \mathbb{R}$ by the rule

$$\hat{f}(S) = \sum_S f \text{ for each } S \text{ in } \mathcal{F}.$$

Then the unordered sum $\sum_{\mathcal{F}} \hat{f}$ is defined. Furthermore, one has

$$\sum_X f = \sum_{\mathcal{F}} \hat{f}; \text{ that is, } \sum_X f = \sum_{S \in \mathcal{F}} \sum_S f. \qquad \text{(IX.6)}$$

Proof

The proof breaks naturally into two cases.

Case 1 Suppose that $\sum_{\mathcal{F}}\hat{f}^{+} = +\infty$. Then for every real number $M > 0$ there exists a finite nonempty subset $\mathcal{W}$ of $\mathcal{F}$ such that $\sum_{\mathcal{W}}\hat{f}^{+} \geq 2M$. Suppose that $\mathcal{W}$ has exactly r distinct elements $S_1, \ldots S_r$, so that

$$\sum\nolimits_{\mathcal{W}}\hat{f}^{+} = \hat{f}^{+}(S_1) + \ldots + \hat{f}^{+}(S_r).$$

By definition of the 'positive part' of a function, for each $j = 1, 2, \ldots r$ one has

$$\hat{f}^{+}(S_j) = \begin{cases} \hat{f}(S_j) & \text{if } \hat{f}(S_j) \geq 0 \\ 0 & \text{if } \hat{f}(S_j) < 0 \end{cases}$$

Since, by definition, $\hat{f}(S_j) = \sum_{S_j} f$, and since $\sum_{S_j} f \leq \sum_{S_j} f^{+}$, it follows easily that $\sum_{S_j} f^{+} \geq \hat{f}^{+}(S_j)$. Thus one has

$$\sum\nolimits_{S_1} f^{+} + \ldots + \sum\nolimits_{S_r} f^{+} \geq \hat{f}(S_1) + \ldots + \hat{f}(S_r) \geq 2M.$$

From the definition of $\sum_{S_j} f^{+}$ there exists a finite nonempty subset W_j of S_j such that

$$\sum\nolimits_{W_j} f^{+} \geq \frac{1}{2}\sum\nolimits_{S_j} f^{+}.$$

Let $W = W_1 \cup \ldots \cup W_r$, so that W is a finite nonempty subset of X. Note that the sets $W_1, \ldots W_r$ in this union are mutually disjoint, since $W_j \subseteq S_j$ and the sets $S_1, \ldots S_r$ are distinct elements of $\mathcal{F}$ and $\mathcal{F}$ satisfies the 'Disjointness Property' for partitions. Thus, Theorem (IX.1.3), the 'Generalized Associative Law' (for finite unordered sums) can be combined with the preceding results to yield

$$\sum\nolimits_{W} f^{+} = \sum\nolimits_{W_1} f^{+} + \ldots + \sum\nolimits_{W_r} f^{+} \geq \frac{1}{2}\left(\sum\nolimits_{S_1} f^{+} + \ldots + \sum\nolimits_{S_r} f^{+}\right) \geq \frac{1}{2}\cdot(2M) = M.$$

Since M can be any positive real number, it follows that $\sum_{W} f^{+} = +\infty$. A similar argument shows that if $\sum_{\mathcal{F}}\hat{f}^{-} = +\infty$ then $\sum_{X} f^{-} = +\infty$. Since, by hypothesis, $\sum_{X} f$ is defined, it cannot be the case that *both* $\sum_{X} f^{+}$ and $\sum_{X} f^{-}$ equal $+\infty$. Thus, by what has just been proved, it cannot happen that both $\sum_{\mathcal{F}}\hat{f}^{+}$ and $\sum_{\mathcal{F}}\hat{f}^{-}$ equal $+\infty$. In other words, the unordered sum $\sum_{\mathcal{F}}\hat{f}$ is defined. Moreover, the preceding also shows that if $\sum_{\mathcal{F}}\hat{f}$ diverges to either $+\infty$ or $-\infty$, then the same holds for $\sum_{X} f$, and one has $\sum_{X} f = \sum_{\mathcal{F}}\hat{f}$, as required.

Case 2 Assume that $\sum_{\mathcal{F}}\hat{f}$ is convergent, and set $C = \sum_{\mathcal{F}}\hat{f}$. Let $\varepsilon > 0$ be given. Then, by Theorem (IX.2.5), there exists a finite nonempty subset $\mathcal{W}_{\varepsilon}$ of $\mathcal{F}$ such that if $\mathcal{W}$ is any finite subset of $\mathcal{F}$ which satisfies $\mathcal{W} \supseteq \mathcal{W}_{\varepsilon}$, then

$$\left|C - \sum\nolimits_{\mathcal{W}}\hat{f}\right| < \frac{\varepsilon}{2}.$$

Suppose that the finite subset $\mathcal{W}_{\varepsilon}$ has exactly r members, namely $S_1, \ldots S_r$. By Theorem (IX.2.5) again, for each $j = 1, 2, \ldots r$ there exists a finite nonempty subset $W_{j\varepsilon}$ of S_j such that if W_j is any finite nonempty subset of S_j satisfying $W_j \supseteq W_{j\varepsilon}$, then

$$\left|\left(\sum\nolimits_{S_j} f\right) - \left(\sum\nolimits_{W_j} f\right)\right| < \frac{\varepsilon}{2r}.$$

Now let $W_{\varepsilon} = W_{1\varepsilon} \cup \ldots \cup W_{r\varepsilon}$, and let W be any finite subset of X such that $W \supseteq W_{\varepsilon}$. For each $j = 1, \ldots r$ let $W_j = W \cap S_j$. Then the sets $W_1, \ldots W_r$ are mutually disjoint (because the sets $S_1, \ldots S_r$ are mutually disjoint) and $W_j \supseteq W_{j\varepsilon} \neq \emptyset$ (because $W \supseteq W_{\varepsilon}$). Note that, by the usual 'cancellation' properties of finite sums, one has $\sum_{W} f - C = \left(\sum_{W_1} f + \ldots + \sum_{W_r} f\right) - C =$

$$\left(\sum\nolimits_{W_1} f - \sum\nolimits_{S_1} f\right) + \left(\sum\nolimits_{W_2} f - \sum\nolimits_{S_2} f\right) + \ldots + \left(\sum\nolimits_{W_r} f - \sum\nolimits_{S_r} f\right) + \left(\sum\nolimits_{S_1} f + \ldots + \sum\nolimits_{S_r} f - C\right).$$

Now apply the 'Extended Triangle Inequality' (see Theorem (II.2.10)) to get

$$|\sum_W f - C| \leq \left|\sum_{W_1} f - \sum_{S_1} f\right| + \left|\sum_{W_2} f - \sum_{S_2} f\right| + \ldots + \left|\sum_{W_r} f - \sum_{S_r} f\right| + \left|\sum_{S_1} f + \ldots + \sum_{S_r} f - C\right|$$

$$< \frac{\varepsilon}{2r} + \ldots + \frac{\varepsilon}{2r} + \frac{\varepsilon}{2} = \varepsilon.$$

The desired result now follows from Theorem (IX.2.5).

IX.2.11 Examples

(1) Let $X = \mathbb{N}\times\mathbb{N}$, and define $f : X \to \mathbb{R}$ by the rule

$$f(i,j) = \frac{1}{2^{i+j}} \text{ for all } (i,j) \text{ in } X$$

It is useful to think of the set X and the function f in terms of the following 'infinite matrix':

$$\begin{bmatrix} \frac{1}{2^2} & \frac{1}{2^3} & \frac{1}{2^4} & \cdots & \frac{1}{2^{1+k}} & \cdots \\ \frac{1}{2^3} & \frac{1}{2^4} & \frac{1}{2^5} & \cdots & \frac{1}{2^{2+k}} & \cdots \\ \frac{1}{2^4} & \frac{1}{2^5} & \frac{1}{2^6} & \cdots & \frac{1}{2^{3+k}} & \cdots \\ \vdots & & & & \vdots & \end{bmatrix}$$

More precisely, the entry in this matrix in the (i,j)-location (i.e., Row #i and Column #j) is the number $f(i,j) = 1/2^{i+j}$. Then the task of computing $\sum_X f$ can be interpreted as 'to add up all the entries of this matrix'.

Suppose that W is a finite nonempty subset of X; thus, it is possible to list the distinct elements of W as

$$W = \{(i_1,j_1), (i_2,j_2), \ldots (i_m,j_m)\}$$

where m is the exact number of elements of the set W. Let $k = \max\{i_1, i_2, \ldots i_m, j_1, j_2, \ldots j_m\}$; that is, k is the largest of all the numbers which appear in the ordered pairs (i,j) in W. Let $Y = \mathbb{N}_k\times\mathbb{N}_k$. Clearly $W \subseteq Y$, and thus one has

$$\sum_W f \leq \sum_Y f.$$

By definition, the number $\sum_Y f$ is the sum of the numbers $1/2^{i+j}$ with $1 \leq i,j \leq k$. By the Generalized Commutative and Associative laws for Addition (see Theorems (II.1.5) and (IX.1.3)), one can obtain this latter sum by grouping together the elements of $\mathbb{N}_k\times\mathbb{N}_k$ 'row-by-row', as follows:

$$\sum_Y f = \left(\frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \ldots + \frac{1}{2^{1+k}}\right) + \left(\frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \ldots + \frac{1}{2^{2+k}}\right) + \left(\frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^6} + \ldots + \frac{1}{2^{3+k}}\right) +$$

$$\ldots + \left(\frac{1}{2^{k+1}} + \frac{1}{2^{k+2}} + \frac{1}{2^{k+3}} + \ldots + \frac{1}{2^{2k}}\right).$$

Note that

$$\frac{1}{2^{j+1}} + \frac{1}{2^{j+2}} + \frac{1}{2^{j+3}} + \ldots + \frac{1}{2^{j+k}} = \left(\frac{1}{2^j}\right)\left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots + \frac{1}{2^k}\right)$$

It follows from the results of Example (III.2.11) (4), with r in that example equalling $1/2$, that

$$\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots + \frac{1}{2^k} < 1, \text{ hence } \left(\frac{1}{2^j}\right)\left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots + \frac{1}{2^k}\right) < 1.$$

Thus one gets

$$\textstyle\sum_Y f \leq \frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^k} < 1$$

for all k in $\mathbb{N}$. Hence,

$$\textstyle\sum_W f < 1 \text{ for every finite nonempty subset } W \text{ of } X.$$

In particular, $\sum_X f \leq 1$.

To compute the actual value of $\sum_X f$, consider the partition of X 'by rows'. That is, let $\mathcal{F}$ denote the infinite family of nonempty subsets S_1, $S_2, \ldots$ of W given by the rule

$$S_i = \{(i,1),(i,2), \ldots (i,k), \ldots\} \text{ for each } i \text{ in } \mathbb{N}.$$

Note that for each such i one has

$$f(i,j) = \frac{1}{2^{i+j}} = \frac{1}{2^{i+1}}\frac{1}{2^{j-1}} = \frac{1}{2^{i+1}}g(j),$$

where $g : \mathbb{N} \to \mathbb{R}$ is the function discussed in Example (IX.2.3) (4) above, with r in that example equalling $1/2$. It is clear that the family $\mathcal{F}$ is a partition of X, and

$$\textstyle\sum_{S_i} f = \frac{1}{2^{i+1}}\sum_{\mathbb{N}} g = \frac{1}{2^i} = \frac{1}{2}g(i).$$

Now use Theorem (IX.2.10), the Generalized Associative Law for Unordered Sums, to conclude

$$\textstyle\sum_X f = \sum_{\mathcal{F}} \hat{f} = \frac{1}{2}\sum_{\mathbb{N}} g = \frac{2}{2} = 1.$$

(2) Let $X = \mathbb{N}\times\mathbb{N}$, as in the preceding example, but now define $f : X \to \mathbb{R}$ by the rule

$$f(i,j) = \begin{cases} 1 & \text{if } j = i \\ -1 & \text{if } j = i+1 \\ 0 & \text{in all other cases} \end{cases}$$

As in the preceding example, it is possible to visualize the set X and the function f as an infinite matrix:

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 1 & -1 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 1 & -1 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 1 & -1 & 0 & \ldots \\ \vdots & & & & & & \vdots \end{bmatrix} \tag{IX.7}$$

It is clear that there are infinitely many points (i,j) in X at which $f(i,j) = +1$; namely, the points corresponding to the main diagonal of this matrix. In particular, one has $\sum_{X_{f;+}} f = +\infty$.

A similar argument shows that $\sum_{X_{f;-}} f = -\infty$. In particular, the unordered sum $\sum_X f$ is *not* defined.

Nevertheless, it is possible to find a partition $\mathcal{F}$ of X such that $\sum_Y f$ *is* defined, and even convergent, for each Y in $\mathcal{F}$, and so that $\sum_{\mathcal{F}} \hat{f}$ is also converent. For instance, if $i \in \mathbb{N}$, let $Y_i = \{(i,1),(i,2),\dots(i,k),\dots\}$. Thus, Y_i corresponds to the i-th row of the matrix above. It is clear that X is the disjoint union of the nonempty subsets $Y_1, Y_2, \dots$, so that the collection $\mathcal{F} = \{Y_1, Y_2, \dots\}$ is a partition of X. It is also clear that each row of Matrix (IX.7) has exactly one entry equal to $+1$, exactly one term equal to -1, and the remaining entries equal to 0. From this one concludes that

$$\sum\nolimits_{Y_i} f = 0 \text{ for each } i \text{ in } \mathbb{N}$$

In particular, for each i the unordered sum $\sum_{Y_i} f$ is convergent; in fact it equals 0. Thus one can write

$$\hat{f}(Y_i) = \sum\nolimits_{Y_i} f = 0 \text{ for all } i \text{ in } \mathbb{N}.$$

By Theorem (IX.2.4) it follows that

$$\sum\nolimits_{\mathcal{F}} \hat{f} = 0.$$

Now consider a second partition, $\mathcal{G}$, of X, obtained from the *columns* of Matrix (IX.7) above. More precisely, for each j in $\mathbb{N}$ let $Z_j = \{(1,j),(2,j),\dots(k,j),\dots\}$. Clearly the sets Z_j, $j \in \mathbb{N}$, are mutually disjoint and $\mathcal{G} = \{Z_1, Z_2, \dots\}$ is a partition of X. Note that, with one exception, every column of the matrix has exactly one entry equal to 1, exactly one entry equal to -1, and all other entries equal to 0. The one exception is the first column, which has exactly one entry equal to 1 and all others equal to 0. It follows easily that for each j the unordered sum $\sum_{Z_j} f$ is convergent. More precisely,

$$\sum\nolimits_{Z_j} f = \begin{cases} 1 & \text{if } j = 1 \\ 0 & \text{if } j \geq 2. \end{cases}$$

Now set $f^{\#} : \mathcal{G} \to \mathbb{R}$ by the rule

$$f^{\#}(Z_j) = \sum\nolimits_{Z_j} f.$$

Clearly $f^{\#}(j) = 1$ if $j = 1$, $f^{\#}(j) = 0$ if $j \geq 2$. Thus one has

$$\sum\nolimits_{\mathcal{G}} f^{\#} = 1.$$

(3) Let $f : \mathbb{N} \to \mathbb{R}$ and $g : \mathbb{N} \to \mathbb{R}$ be real-valued functions defined on $\mathbb{N}$, and suppose that both of the infinite unordered sums $\sum_{\mathbb{N}} f$ and $\sum_{\mathbb{N}} g$ are convergent, with (finite) values A and B, respectively. Define $h : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ by the rule $h(i,j) = f(i)g(j)$ for each (i,j) in $\mathbb{N} \times \mathbb{N}$. Then it is easy to show that $\sum_{\mathbb{N} \times \mathbb{N}} h$ is convergent, and that its value is AB.

It was pointed out after Example (1) above that the quantity $\sum_X f$ can sometimes be defined even when the set X is uncountable. In contrast, the next result explains why we normally restrict our attention to the case in which X is countable.

IX.2.12 Theorem

Suppose that $f : X \to \mathbb{R}$ is a real-valued function whose domain is an uncountable set X. Then there exists a countable nonempty subset Y of X such that the behavior of the unordered sum $\sum_X f$ is the same as the behavior of the unordered sum $\sum_Y f$.

More precisely, there exists a countable nonempty subset Y of X such that

$$\sum_Y f^+ = \sum_X f^+ \text{ and } \sum_Y f^- = \sum_X f^-$$

Proof

For convenience set $A = \sum_X f^+$ and $B = \sum_X f^-$.

Case (i) Suppose that $A = 0$. Then, by Part (a) of Theorem (IX.2.4), one has $f(x) = 0$ for all x in X. In this case, let Y^+ be any nonempty countable subset of X; clearly $\sum_{Y^+} f^+ = 0 = A$.

Case (ii) Suppose $0 < A < +\infty$. For each m in $\mathbb{N}$ let X_m^+ denote the set of all x in X such that $f^+(x) \geq 1/m$.

<u>Claim</u> The set X_m^+ is finite. More precisely, if k is a positive integer such that $k > mA$, then X_m^+ has fewer than k elements.

<u>Proof of Claim</u> Suppose that X_m^+ has at least k elements. Then there must exist a finite subset $W = \{x_1, x_2, \dots x_k\}$ of X_m^+ with exactly k elements. Clearly $f^+(x_j) \geq 1/m$ for each $j = 1, 2, \dots k$, so

$$\sum_W f^+ = f(x_1) + f(x_2) + \dots + f(x_k) \geq \frac{k}{m} > \frac{mA}{m} = A.$$

By Part (b) of Theorem (IX.2.4) one then has $\sum_X f^+ \geq \sum_W f^+ > A$, contrary to the definition of A. That is, assuming that X_m^+ is an infinite set leads to a contradiction; thus, X_m is a finite set, as claimed.

Now let $Y^+ = \bigcup_{m=1}^\infty X_m^+$. Note that Y^+, being the countable union of finite sets, is a countable set; see Theorem (I.8.11). In addition, since $A > 0$ it is clear that there exist at least one x in X such that $f^+(x) > 0$. For each such x there exists m in $\mathbb{N}$ such that $x \in X_m^+$, and thus Y^+ is nonempty. Indeed, the Archimedean Principle guarantees that there exists m such that $m > 1/f^+(x)$, and thus $f^+(x) > 1/m$; that is, $x \in X_m^+$. Finally, it follows from Part (b) of Theorem (IX.2.4) (and the definition of f^+) that $\sum_{Y^+} f^+ = \sum_X f^+ = A$, as required.

Case (iii) Suppose that $A = +\infty$. Then for every m in $\mathbb{N}$ there exists a finite nonempty subset W_m of X such that $\sum_{W_m} f^+ \geq m$. Let $Y^+ = \bigcup_{m=1}^\infty W_m$. The set Y^+ is the countable union of finite sets, hence it is a countable nonempty set. It is clear that $\sum_{Y^+} f^+ = +\infty = A$.

Conclusion: In all cases, there exists a countable nonempty set Y^+ such that $\sum_{Y^+} f^+ = \sum_X f^+$.

In a similar manner one can prove that there exists a countable nonempty subset Y^- of X such that $\sum_{Y^-} f^- = B = \sum_X f^-$. (Alternatively, simply apply what was just proved to the function g^+ when $g = -f$.)

Finally, let $Y = Y^+ \cup Y^-$. Clearly $\sum_Y f^+ = \sum_{Y^+} f^+$, since if $x \in Y \setminus Y^+$, then $f(x) \leq 0$, hence $f^+(x) = 0$. Thus, $\sum_Y f^+ = A = \sum_X f^+$, as required. Likewise, $\sum_Y f^- = B = \sum_X f^-$.

IX.2.13 Remarks

(1) The sense in which 'the behavior of the unordered sum $\sum_X f$ is the same as the behavior of the unordered sum $\sum_Y f$' is this: the sum $\sum_X f$ exists if, and only if, the sum $\sum_Y f$ exists; and if these sums exist, then they are equal.

(2) If $\sum_X f$ is convergent, then the proof above shows that the set $Z = \{x \in X : f(x) \neq 0\}$ must be countable.

In light of the preceding results, it makes sense to narrow our inquiry to the case in which X is countably infinite. The next result shows how to relate unordered sums in that situation to the theory of limits of sequences studied in Chapter (III).

IX.2.14 Theorem

Let $f : X \to \mathbb{R}$ be a real-valued function defined on a countably infinite set X.

(a) Suppose that the unordered sum $\sum_X f$ is defined. Let $\varphi : \mathbb{N} \to X$ be a bijection of $\mathbb{N}$ onto X; and associate with f and φ the real-valued sequence $\sigma_{f;\varphi} = (s_1, s_2, \ldots)$ given by the rule

$$s_k = f(\varphi(1)) + f(\varphi(2)) + \ldots + f(\varphi(k)) \text{ for each } k \text{ in } \mathbb{N} \quad (*)$$

Then the sequence $\sigma_{f;\varphi}$ has a limit, and one has

$$\sum\nolimits_X f = \lim_{k \to \infty} s_k \tag{IX.8}$$

(b) Conversely, for each bijection $\varphi : \mathbb{N} \to X$, define a corresponding sequence $\sigma_{f;\varphi} = (s_1, \ldots s_k, \ldots)$ by means of Equation $(*)$. Assume that there is a quantity L, where L is either a real number or one of the infinities, such that $\lim \sigma_{f;\varphi} = L$ for each such bijection φ. Then the unordered sum $\sum_X f$ exists and equals L.

Proof (a) For convenience let us set $u_i = f(\varphi(i))$ for each i in $\mathbb{N}$. Then one can write $s_k = u_1 + u_2 + \ldots + u_k$ for each index k.

Special Case Suppose that $f(x) \geq 0$ for each x in X. It follows that each of the numbers of the form u_i is nonnegative and thus the sequence $\sigma_{f;\varphi}$ is monotonic up. In particular, the quantity $\lim_{k \to \infty} s_k$ exists. For convenience, we denote this limit by L; of course it is possible that $L = +\infty$. In any event, one certainly has $s_m \leq L$ for all indices m.

By Definition (IX.2.1), $\sum_X f = \sup U_{X;f}$, with $U_{X;f}$ being the set of numbers of the form $f(x_1) + f(x_2) + \cdots + f(x_m)$, and $\{x_1, x_2, \ldots, x_m\}$ ranging over the finite subsets of X. (Of course the notation assumes that the x's are distinct.) For such a set one can write $x_i = \varphi(j_i)$ for certain (distinct) elements $j_1, j_2, \ldots j_m$ in $\mathbb{N}$. Let p be the largest of the indices $j_1, j_2, \ldots j_m$. Then clearly

$$f(x_1) + f(x_2) + \cdots + f(x_m) = u_{j_1} + u_{j_2} + \ldots + u_{j_m} \leq u_1 + u_2 + \ldots + u_p,$$

since the sum $u_1 + u_2 + \ldots + u_p$, which has only nonnegative terms, includes all the terms which appear in the sum $u_{j_1} + u_{j_2} + \ldots + u_{j_m}$. It follows that L is an upper bound for the set $U_{X;f}$, and thus

$$\sum\nolimits_X f \leq L$$

However, if z is any number such that $z < L$, then there exists k such that $z < s_k \leq L$. Since s_k is certainly an element of $U_{X;f}$, it follows that for every such z one has $z < \sum_X f$. This implies that $\sum_X f \geq L$ as well. Thus, $\sum_X f = L$.

The general case now follows by applying the results of the Special Case above to the functions f^+ and f^-, and using the equation $\sum_X f = \sum_X f^+ - \sum_X f^-$.

(b) Assume that the desired conclusion, namely that $\sum_X f = L$, is *not* true. We shall show that this implies the existence of a bijection $\varphi : \mathbb{N} \to X$ for which it is *not* the case that $\lim \sigma_{f;\varphi} = L$, in contradiction to the hypotheses.

Consider first the case in which L is finite. Then, by Theorem (IX.2.5), there would exist $\varepsilon_0 > 0$ such that for every finite subset W' of X there must exist a finite subset W of X such that $W \supseteq W'$ and $|L - \sum_W f| \geq \varepsilon_0$. It is clear that one can choose this W so that W is a *proper* superset of W'; that is, $W \backslash W' \neq \emptyset$.

Express the countably infinite set X as $X = \{x_1, x_2, \ldots\}$, and recursively define a sequence of finite subsets $W_1, W_2, \ldots$ of X as follows:

(1) W_1 is a finite subset of X such that $W_1 \supseteq \{x_1\}$ and $\left|L - \sum_{W_1} f\right| \geq \varepsilon_0$.

(2) If W_k has been defined, let W_{k+1} be a finite subset of X such that W_{k+1} is a *proper* superset of $W_k \cup \{x_{k+1}\}$, and $\left|L - \sum_{W_{k+1}} f\right| \geq \varepsilon_0$

It is clear that $\bigcup_{k=1}^{\infty} W_k = X$; indeed, by construction one has that $x_k \in W_k$ for each k. (Of course it is possible that a given x_k might well be in W_j for some $j < k$.) It then follows that one can write

$$X = W_1 \cup (W_2 \backslash W_1) \cup \ldots \cup (W_k \backslash W_{k-1}) \cup \ldots,$$

and that this is a *disjoint* union; that is, it is a partition of X. Let $m_1 = \#(W_1)$ and for $k \geq 2$ let $m_k = \#(W_k \backslash W_{k-1})$. Note that $m_k \geq 1$ for each k, since the sets W_1 and (if $k \geq 2$) $W_k \backslash W_{k-1}$ have been constructed to be nonempty. Then these numbers determine a corresponding partition of $\mathbb{N}$. Indeed, let $n_1 = m_1$, $n_2 = m_1 + m_2$, $\ldots\ n_k = m_1 + \ldots + m_k$. Then

$$\mathbb{N} = \mathbb{N}_{n_1} \cup (\mathbb{N}_{n_2} \backslash \mathbb{N}_{n_1}) \cup \ldots (\mathbb{N}_{n_k} \backslash \mathbb{N}_{n_{k-1}}) \cup \ldots$$

It follows from Theorem (**??**) that there is a bijection $\varphi : \mathbb{N} \to X$ such that φ maps $\mathbb{N}_{n_1}$ bijectively onto W_1, and if $k \geq 2$ then φ maps $\mathbb{N}_{n_k} \backslash \mathbb{N}_{n_{k-1}}$ bijectively onto $W_k \backslash W_{k-1}$. It is clear that for each k one has

$$f(\varphi(1)) + f(\varphi(2)) + \ldots + f(\varphi(n_k)) = \sum_{W_k} f$$

and thus

$$|L - (f(\varphi(1)) + f(\varphi(2)) + \ldots + f(\varphi(n_k)))| \geq \varepsilon_0.$$

In particular, the sequence $\sigma_{f;\varphi}$ has a subsequence which does not converge to L, namely $(s_{n_1}, s_{n_2}, \ldots)$ so it follows that the equation $\lim\, \sigma_{f;\varphi} = L$ cannot hold, contrary to the hypotheses.

A similar argument works when L is one of the infinities; the details are left as an exercise.

IX.2.15 Remark

The preceding result implies that a sufficient condition for the unordered sum $\sum_X f$ to be defined is that, for *each* bijection $\varphi : \mathbb{N} \to X$, the sequence $\sigma_{f;\varphi}$ has a limit, and the limit is the same each φ. It is natural to ask whether one gets the same conclusions provided one knows only that $\sigma_{f;\varphi}$ has a limit for *at least one* bijection φ. We shall see in the next section that the answer is 'No'.

Let us close this section with a result which says, in effect, that relatively small errors in the values of a function $f : X \to \mathbb{R}$ causes relatively small errors in $\sum_X f$.

The background for this result comes from the problem that in real-life computation, one must often make do with *approximations* of the actual numbers one hoped to deal with. For instance, if one keys in the famous number π on an inexpensive scientific calculator, one may get

$$\pi = 3.1415927$$

Of course, everyone knows that a much more accurate answer would be

$$\pi = 3.14159265358979323846264338327950288419716939937510;$$

but even this is not quite correct. The problem is that in the decimal representation of numbers on a computing device, only a finite numbers of decimal digits can be accomodated, so in general the computer must 'round off' values to a certain number of significant digits. That is, one must approximate the 'true' value of a number y by an 'approximate value' $\hat{y}$ with which one can actually do the computation. For approximations of this type the issue concerns mainly nonzero numbers, and the appropriate type of error to consider is the so-called *relative error*.

IX.2.16 Definition (Relative Error)

Let y be a nonzero real number, and consider an approximation of the form $y \approx \hat{y}$, where $\hat{y}$ is a nonzero number with the same sign as y (ie., both are positive or both are negative). Then the **relative error** in this approximation is $E = \dfrac{\hat{y}-y}{y}$. A number $\varepsilon > 0$ is an **upper bound on the relative error** if $|E| \leq \varepsilon$; that is, if $-\varepsilon \leq E \leq \varepsilon$.

Remark Let $\varepsilon > 0$ be an upper bound on the relative error of the approximation $y \approx \hat{y}$ described above. Then it is easy to see that the requirement that y and $\hat{y}$ be of the same sign implies that $\varepsilon < 1$. More precisely, if $y > 0$ then $(1-\varepsilon)y \leq \hat{y} \leq (1+\varepsilon)y$. Likewise, if $y < 0$ then $(1-\varepsilon)y \geq \hat{y} \geq (1+\varepsilon)y$. If, for instance, the approximation process is 'rounding off after k decimal digits', then one can take $\varepsilon = 1/10^k$.

Note that another way of expressing the fact that the $\hat{y}$ approximates y with relative error less than ε is this: there exists a number z in the interval $[1-\varepsilon, 1+\varepsilon]$ such that $\hat{y} = zy$. This formulation is used in the next result.

IX.2.17 Theorem (The Stability Theorem for Unordered Sums)

Suppose that $f : X \to \mathbb{R}$ is a real-valued function defined on a nonempty set X. For simplicity, assume that for all x in X one has $f(x) \neq 0$. Let ε be such that $0 < \varepsilon < 1$, and suppose that $g : X \to \mathbb{R}$ is a function such that $g(x) \in (1-\varepsilon, 1+\varepsilon)$ for all x in X. Define $\hat{f} : X \to \mathbb{R}$ by the rule $\hat{f}(x) = g(x)f(x)$ for all x in X. Then the following hold:

(a) If the sum $\sum_X f$ is undefined, then so is the sum $\sum_X \hat{f}$.

(b) If the sum $\sum_X f$ is defined, then so is $\sum_X \hat{f}$. More precisely:
(i) if $\sum_X f$ diverges to one of the infinities, then $\sum_X \hat{f}$ diverges to the same infinity;
(ii) if $\sum_X f$ converges, so does $\sum_X \hat{f}$. Futhermore, there exists a constant B, independent of ε, such that

$$\left| \left(\sum_X f\right) - \left(\sum_X \hat{f}\right) \right| \leq \varepsilon B;$$

indeed, one can take $B = \sum_X |f|$.

The simple proof is left as an exercise.

Remark The reason for the name 'Stability Theorem' attached to the preceding theorem should be clear: the theorem says that relatively small changes in the nonzero terms of an unordered sum cannot affect the divergence properties of the sum; and if the sum is convergent, the effects on the value of the sum of these small changes are also under control. Of course, by 'relatively small' here is meant 'with small relative error'.

IX.3 Application: The Measure of Open Sets and of Compact Sets in R

The theory of 'infinite unordered sums' allows one to give a simple treatment of the important concept of the 'size' of a set, at least in the case of open and closed subsets of $\mathbb{R}$. This concept plays an important role in the theory of the Riemann integral; see Chapter (VII). (There is also a much more general theory, associated with the name of the French mathematician Henri Lebesgue, which assigns the notion of 'measure' to a much wider class of sets than these; but we do not need that theory in *This Textbook*.)

The basic problem under consideration is to formulate a concept of 'total length' which applies to subsets of $\mathbb{R}$. As has happened before in *This Textbook*, we begin by providing some guidelines which any 'reasonable' formulation of this concept ought to satisfy:

Guideline 1 If X is a subset of $\mathbb{R}$ for which 'total length' makes sense, then this length must be either a nonnegative number or the quantity $+\infty$.

Guideline 2 If X is a set with a single point, then its 'total length' should be 0.

Guideline 3 If X is an interval in $\mathbb{R}$ – open, closed or half-open – then the total length of X should be its usual length; namely $+\infty$ if the interval is unbounded, $|b-a|$ if it has endpoints a and b which are finite.

Guideline 4 If X can be expressed as the union of disjoint sets A and B for which the notion of 'total length' has been formulated, then the total length of X should be the sum of the total lengths of A and B. (If either A or B has infinite total length, thhis means that X should have infinite total length as well.)

These guidelines, combined with the concept of 'unordered sum', suggest the following definition. In this definition, we use the more common terminology of 'measure' instead of 'total length', and for simplicity we restrict our attention to subsets of $\mathbb{R}$ which are either open in $\mathbb{R}$ or compact (i.e., closed and bounded) in $\mathbb{R}$.

IX.3.1 Definition

(1) Suppose that U is a nonempty open set in $\mathbb{R}$, and let $\mathcal{G}$ be the unique countable family of (nonempty) open intervals whose union equals U. (The existence and uniqueness of such a family is guaranteed by Theorem (**??**).

(a) If the family $\mathcal{G}$ has at least one element which is an *un*bounded open interval, then define the **measure** $\mu(U)$ **of the open set** U to be $+\infty$.

(b) Suppose that each interval I in the family $\mathcal{G}$ is bounded. Let $\lambda : \mathcal{G} \to \mathbb{R}^+$ be given by $\lambda(I) = |b-a|$, where a and b are the endpoints of I. Define the **measure** $\mu(U)$ **of the open set** U to be the unordered sum $\mu(U) = \sum_{\mathcal{G}} \lambda$.

For completeness, define the measure of the empty set by $\mu(\emptyset) = 0$.

(2) Suppose that Y is a nonempty compact set in $\mathbb{R}$ (i.e., a set which is closed and bounded in $\mathbb{R}$), and let $a = \inf Y$, $b = \sup Y$; note that a and b are real numbers (because Y is nonempty and bounded). Let $J = \{x \in \mathbb{R} : a \leq x \leq b\}$, and let $U = J \setminus Y$, so that U is a bounded open set – possibly empty – in $\mathbb{R}$. Then the **measure of the compact set** Y is the quantity $\mu(Y) = |b-a| - \mu(U)$, where $\mu(U)$ is defined as above.

IX.3.2 Examples

(1) Let $Y = \{x_1, x_2, \ldots x_m\}$ be a finite set with exactly $m \geq 2$ elements, written so that $x_1 < x_2 < \ldots < x_m$. Then $\inf Y = x_1$ and $\sup Y = x_m$. One sees that $U = [x_1, x_m \setminus Y$ is the disjoint union of the open intervals (x_1, x_2), $(x_2, x_3), \ldots (x_{m-1}, x_m)$., and thus $\mu(U)$ equals the

finite (collapsing) sum $\sum_{k=2}^{m}(x_k - x_{k-1}) = x_m - x_1$. Thus $\mu(Y) = (x_m - x_1) - (x_m - x_1) = 0$. That is, every finite subset of $\mathbb{R}$ has measure 0.

(2) In a similar manner, one can easily show that if Y is the closed bounded set $\{1, 1/2, 1/3, \lambda 1/k, \dots\} \cup \{0\}$, then $\mu(Y) = 0$.

(3) The set of rational numbers that lie in the interval $[0,1]$ is certainly a countable set. However, it is neither open nor closed, so the definition above does not apply to it.

(4) Let C be the standard Cantor Ternary Set (see Definition (**??**)); thus C is a compact subset with $\inf C = 0$ and $\sup C = 1$. Let $U = [0,1]\backslash C$, so that U is an open set; indeed, U is the union of the open intervals which are removed from $[0,1]$ during the 'Middle Thirds' construction of C (see Remark (**??**)).

Since the open intervals removed in the 'Middle Thirds' process are clearly mutually disjoint, it is easy to determine $\mu(U)$. Indeed, in the first step of the process, one open interval, $(1/3, 2/3)$ is removed; call it U_{11}. In the second step, two open intervals are removed from $[0,1]\backslash U_{11}$; call then U_{21} and U_{22}. More generally, in the k-th step, 2^{k-1} open intervals are removed; call them U_{k1}, U_{k2}, $\dots U_{k2^{k-1}}$. Finally, let $\mathcal{G}$ be the family whose elements are the open sets U_{km}, with k in $\mathbb{N}$ and $1 \leq m \leq 2^{k-1}$. Then U is the (disjoint) union of the sets in the family $\mathcal{G}$, and $\mu(U) = \sum_{\mathcal{G}\lambda}$, where as before $\lambda(U_{km})$ is the ordinary length of the interval U_{km}. To compute this infinite unordered sum, define subsets $S_1, S_2, \dots S_k, \dots$ of $\mathcal{G}$ by

$$S_k = \{U_{k1}, U_{k2}, \dots U_{k2^{k-1}})\} \text{ for each } k \text{ in } \mathbb{N},$$

and set $\mathcal{F} = \{S_1, S_2, \dots S_k, \dots)\}$. It is clear that $\mathcal{F}$ is a partition of the set $\mathcal{G}$. Furthermore, each element S_k of the partition $\mathcal{F}$ is a itself a collection of 2^{k-1} mutually disjoint open intervals, each of length $1/3^k$. Thus the Generalized Associative Law for Infinite Unordered Sums implies that

$$\sum_{\mathcal{G}} \lambda = \sum_{S\in\mathcal{F}} \sum_{S} \lambda = \sum_{k\in\mathbb{N}} \sum_{S_k} \lambda = \sum_{k\in\mathbb{N}} 2^{k-1} \cdot \left(\frac{1}{3^k}\right) = \frac{1}{3} \sum_{k\in\mathbb{N}} \left(\frac{2}{3}\right)^{k-1}$$

The final unordered sum can be computed using Example (IX.2.3) with r in that example set to $r = 2/3$. One then gets

$$\mu(U) = \frac{1}{3}\left(\frac{1}{1-(2/3)}\right) = \left(\frac{1}{3}\right)\left(\frac{1}{1/3}\right) = 1.$$

It follows that $\mu(C) = 1 - \mu(U) = 0$; that is, the Cantor Ternary Set has measure 0.

Remark It may appear that the underlying reason the Cantor Ternary Set turns out to have measure 0 is that it is a closed set which has no nonempty subsets which are open in $\mathbb{R}$. However, there exist closed bounded subsets of $\mathbb{R}$, which also have no open subsets, which have positive measure. An example of one such so-called ''is constructed in the exercises.

IX.3.3 Theorem

(a)Suppose that U is an open set in $\mathbb{R}$ such that $U \subseteq [a,b]$ for some (finite) numbers a, b with $a < b$. Then $\mu(U) \leq b - a$.

(b) Suppose that U and V are open sets in $\mathbb{R}$ such that $U \subseteq V$. Then $\mu(U) \leq \mu(V)$.

(c) Suppose that $U_1, U_2, \dots U_k, \dots$ is an infinite sequence of open bounded intervals, and $V \subseteq U_1 \cup U_2 \cup \dots \cup U_k \cup \dots$. Let $f : \mathbb{N} \to \mathbb{R}$ be given by $f(j) = \mu(U_j)$ for each j. Then $\mu(V) \leq \sum_{\mathbb{N}} f$.

(d) Suppose that X and Y are compact subsets of $\mathbb{R}$ such that $X \subseteq Y$. Then $\mu(X) \leq \mu(Y)$.

The simple proof is left as an exercise. ■

The extension of the concept of 'measure', to apply to a much wider class of subsets of $\mathbb{R}$, is surprisingly difficult to carry out, and is beyond the scope of *This Textbook.* However, there is one partial extension which is easy to carry out and which we shall need in Chapter (VII):9

IX.3.4 Definition

A subset X of $\mathbb{R}$ is said to be of **measure** 0, or a **null set in** $\mathbb{R}$, provided that for every $\varepsilon > 0$ there exists an open subset U of $\mathbb{R}$ such that $X \subseteq U$ and $\mu(U) < \varepsilon$.

IX.3.5 Examples

(1) If X is a countable subset of $\mathbb{R}$, then X is a null set. Indeed, when X is finite the result follows from an earlier example. Thus, assume that X is countably infinite, and express X as $\{x_1, x_2, \dots x_k, \dots\}$, with $x_i \neq x_j$ if $i \neq j$. For each $j = 1, 2, \dots$, let $\varepsilon_j = \varepsilon/2^{j+2}$, and let U_j be the open interval $(x_j - \varepsilon_j, x_j + \varepsilon_j)$. Note that X is contained in the union V of the intervals U_j, $j = 1, 2, \dots$. Also $\mu(U_j) = \varepsilon/2^{j+1}$, so by Part (c) of Theorem (IX.3.3) $\mu(V)$ and Part (4) of Example (IX.2.3) one has $\mu(V) \leq \varepsilon$. The claimed result follows.

(2) If a set X is a null set, then so is every subset of X.

(3) If Y is a compact set such that $\mu(Y) = 0$, then Y is a null set.

IX.4 Ordered Infinite Sums; that is, Infinite Series

The standard approach to 'infinite sums', as found in most freshman-calculus texts, is to combine the concept of 'successive addition', as described in Definition (II.1.4), with the idea of 'convergent sequence'; the 'unordered sum' approach is probably not even mentioned. In practice, however, it is customary to use the word 'series' instead of 'sum' when following the standard approach.

NOTE: The use of the word 'series' in this context is absolutely standard in mathematics. Nevertheless, there are a few down-sides to that usage. For example, in ordinary English the word 'series' is often used to mean a 'succession' or 'ordered list' of objects. Thus, in nonmathematical contexts 'series' means 'sequence'! This, combined with the fact that both 'sequence' and 'series' begin with the letters 'se', causes no end of confusion for calculus students.

A second problem with the word 'series' is that the plural of 'series' is 'series' (and *not* 'serieses', as some freshmen believe). Thus, one has to pay close attention to know that an author may be referring to more than one series in a given context.

IX.4.1 Definition

Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be an infinite sequence of real numbers.

(1) The expression $x_1 + x_2 + \dots$, in which the 'dots' indicate that the additions are to go on indefinitely, is called the **infinite series associated with** ξ; the number x_k is then called the **k-th term** of this infinite series, and ξ is called the **sequence of terms associated with the infinite series** $\sum_{k=1}^{\infty} x_k$. One often writes such an expression using the well-known 'Sigma' notation; for

instance, as $\sum_{j=1}^{\infty} x_j$. On occasion we may write $\sum_{\xi}$ as a shorthand for the expression $\sum_{k=1}^{\infty} x_k$; however, one must then be careful to not confuse the expression $\sum_{\xi}$ with the unordered sum $\sum_{\mathbb{N}} \xi$.

(2) For each k in $\mathbb{N}$ the **k-th partial sum associated with the infinite series** $\sum_{j=1}^{\infty} x_j$ is the number s_k given by the finite sum

$$s_k = x_1 + x_2 + \ldots + x_k$$

The sequence $\sigma_{\xi} = (s_1, s_2, \ldots)$ is then called the **sequence of partial sums** associated with the given infinite series $\sum_{\xi}$.

Equivalently, the sequence σ_{ξ} can be described recursively by the rule

$$s_1 = x_1, \quad s_{k+1} = s_k + x_{k+1} \text{ for each } k \text{ in } \mathbb{N}.$$

Note: If the context makes it clear which sequence ξ of terms we are using, we may write σ instead of the more proper σ_{ξ}.

(3) If the sequence σ of partial sums is convergent, with limit L in $\mathbb{R}$, then one says that **series $\sum_{k=1}^{\infty} x_k$ is a convergent series**, and that series **converges to** L, or that **the value of the series is** L. One then assigns the value L to the series and writes $\sum_{k=1}^{\infty} x_k = L$.

(4) If the sequence σ of partial sums is divergent, then one says that **the series $\sum_{k=1}^{\infty} x_k$ is divergent**.

Special Case Suppose that $\lim_{k \to \infty} s_k = L$, where L equals either $+\infty$ or $-\infty$. Then one says that **the series** $\sum_{\xi}$ **diverges to** L, and one writes $\sum_{k=1}^{\infty} x_k = L$.

(5) If there exists a quantity L, either a real number or one of the infinities, such that $\sum_{k=1} x_k = L$ in the sense described in (3) and (4) above, then one says that the series $\sum_{k=1}^{\infty} x_k$ **exists** or **is defined**.

IX.4.2 Examples

(1) An infinite series $\sum_{k=1}^{\infty} x_k$ for which at most finitely many terms are nonzero is certainly convergent. Indeed, let m be a natural number such that $x_k = 0$ for all indices k such that $k > m$. Then it is clear that the corresponding sequence of partial sums $\sigma = (s_1, s_2, \ldots)$ satisfies $s_k = s_m$ for all $k \geq m$, and thus the sequence σ converges to the finite sum $s_m = x_1 + x_2 + \ldots + x_m$. It is sometimes convenient to refer to such an infinite series as a **trivial infinite series**; likewise, one then can refer to a series in which infinitely many of the terms are nonzero as a **nontrivial infinite series**.

(2) An infinite series of the form $A + Ar + Ar^2 + Ar^3 + \ldots + Ar^k + \ldots$ is called a **geometric series**. The quantities A and r are called, respectively, the **initial term** and the **common ratio** of the series. (The reason for the name 'initial term' should be obvious. As for the 'common ratio' terminology, suppose that the series is nontrivial – see (1) above – so that A and r are nonzero. Then each term of the series is nonzero, and the ratio $(Ar^{k+1})/(Ar^k)$ of any two consecutive terms of the series is the same, namely r.)

It follows from Part (d) of Proposition (II.2.16) and Part (b) of Corollary (III.2.10) that a nontrivial geometric series $\sum_{k=1}^{\infty} Ar^{k-1}$ converges to $A/(1-r)$ if $|r| < 1$. In contrast, it is an easy exercise to show that such a series is divergent if $|r| \geq 1$. (As for a *trivial* geometric series, that is, one for which either A or r equals 0, it is clear that the series always converges to the initial term A.)

(3) The **harmonic series** is the infinite sum $1+\dfrac{1}{2}+\dfrac{1}{3}+\dfrac{1}{4}+\ldots$; it is so called because its terms form the harmonic sequence (see Example (I.9.2) (4)). It is easy to see that this series diverges to $+\infty$. Indeed, there are several well-known ways to show this; here is an approach which is based on the properties of the natural-logarithm function:

Let $f(x) = \ln(x)$ for all $x > 0$; then $f'(x) = 1/x$ for all $x > 0$. In particular, if $0 < a < b$ then the maximum value of f' on the interval $[a,b]$ is $1/a$. It then follows from the Mean-Value Theorem, that $f(b) - f(a) \leq (b-a)/a$. Now let j be a natural number and set $a = j$, $b = j+1$, so that $b - a = 1$. Then preceding inequality takes the form

$$\ln(j+1) - \ln(j) \leq \frac{1}{j} \text{ for all } j \text{ in } \mathbb{N}.$$

Add up the terms appearing on each side of the preceding inequalities for $j = 1, 2, \ldots k$, and note that the sum of the left sides will involve lots of cancellation (and that $\ln 1 = 0$), to get

$$\ln(k+1) \leq 1 + \frac{1}{2} + \ldots + \frac{1}{k}.$$

That is, the k-th partial sum of the harmonic series is bounded below by $\ln(k+1)$. Since $\lim_{k\to\infty} \ln(k+1) = +\infty$, it follows that the harmonic series diverges to $+\infty$, as claimed.

(4) In Section (**??**) we obtained an expression for the standard exponential function as the limit of certain polynomial functions:

$$\exp\, x = \lim_{k\to\infty} p_{k-1}(x) \text{ where } p_{k-1}(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^{k-1}}{(k-1)!}$$

This result has a cleaner expression in terms of an infinite series. Indeed, for any given x in $\mathbb{R}$, consider the infinite sequence $\rho = (r_1, r_2, \ldots)$ given by

$$r_k = x^{k-1}/(k-1)! \text{ for each index } k \text{ in } \mathbb{N};$$

note that $r_1 = 1$, and $p_{k-1}(x) = r_1 + r_2 + \ldots + r_k$ for each k in $\mathbb{N}$. Then the results of Section (**??**) imply that the infinite series $\sum_{k=1}^{\infty} r_k$ converges to $\exp\, x$. Thus the preceding 'limit' expression for $\exp x$ can be written as an infinite series:

$$\exp\, x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots + \frac{x^k}{k!} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

(5) In a like manner, the results of Section (**??**) allow us to express the standard sine and cosine functions as the values of the following infinite series:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + (-1)^{k-1}\frac{x^{2k-1}}{(2k-1)!} + \ldots$$

and

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + (-1)^{k-1}\frac{x^{2k}}{(2k)!} + \ldots.$$

Both formulas are valid for all x in $\mathbb{R}$.

(6) Consider the 'Natural Logarithm' function $\ln : \mathbb{R}^+ \to \mathbb{R}$. Then one has, for all $x > 0$,

$$\ln'(x) = \frac{1}{x},\ \ln''(x) = -\frac{1}{x^2},\ \ln'''(x) = \frac{2}{x^3},$$

and so on. The general formula is

$$\ln^{(k)}(x) \;=\; (-1)^{k-1}\frac{(k-1)!}{x^k} \text{ for each } k \;=\; 1, 2, \ldots .$$

Now let $h_k : \mathbb{R} \to \mathbb{R}$ denote the k-th Taylor polynomial of the function ln about the center point $c \,=\, 1$, so that

$$h_0(x) \;=\; 0, \qquad h_k(x) \;=\; \frac{x-1}{1} - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \ldots + (-1)^{k-1}\frac{x^k}{k}$$

It follows from Taylor's Formula with Remainder (see Corollary (**??**)) that if $x \,>\, 1$ then

$$\ln x - h_k(x) \;=\; (-1)^k \frac{(x-1)^{k+1}}{kc^{k+1}} \text{ for some } c \text{ such that } 1 \,<\, c \,<\, x.$$

Since for such c one has $c^k \,>\, 1$, it follows easily that

$$|\ln x - h_k(x)| \le \frac{|x-1|^{k+1}}{k+1} \le \frac{1}{k+1} \text{ for all } x \text{ such that } 1 \le x \le 2$$

In particular, one has

$$\ln x \;=\; \lim_{k \to \infty} h_k(x) \text{ for all } x \text{ such that } 1 \le x \le 2.$$

Since $h_k(x)$ is the k-th partial sum for the infinite series

$$(x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \ldots + (-1)^{k-1}\frac{(x-1)^k}{k} + \ldots,$$

we can now write: if $1 \le x \le 2$, then

$$\ln x \;=\; (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \ldots + (-1)^{k-1}\frac{(x-1)^k}{k} + \ldots \qquad \text{(IX.9)}$$

NOTE: With a little more work one can show that the preceding equation is also correct if $0 \,<\, x \,<\, 1$, but that it fails to hold if $x \,>\, 2$.

The formula in Equation (IX.9) is a slight variation of a formula obtained independently by several mathematicians during the latter part of the 17-th century; the first publication was apparently by the famous cartographer Mercator, so it is often called the **Mercator series**. In the special case $x \,=\, 2$ the formula reduces to the following striking result:

$$\ln 2 \;=\; 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + (-1)^{k-1}\frac{1}{k} + \ldots \qquad \text{(IX.10)}$$

The infinite series which appears on the right side of this equation is called the **alternating harmonic series**. This is because it can be obtained from the harmonic series (see Example (3) above)

$$1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{k} + \ldots$$

by alternating the signs of the terms. We shall see that both the harmonic series and the alternating harmonic series have interesting properties which makes them particularly useful for illustrating some of the features of the theory of infinite series (i.e., ordered infinite sums).

The main idea behind the standard definition of 'convergent infinite series' given above is to introduce the concept of 'sequence of partial sums', and thereby reduce the discussion to the theory of convergent sequences that has already been treated in Chapter (III). It is worth noting, however, that *every* sequence of numbers can be throught of as a 'sequence of partial sums'.

IX.4.3 Proposition

Let $\sigma = (s_1, s_2, \ldots)$ be an infinite sequence of real numbers. Then there exists a unique infinite series $\sum_{j=1}^{\infty} x_j$ for which σ is its sequence of partial sums. More precisely, $x_1 = s_1$, and if $k \geq 1$ then $x_{k+1} = s_{k+1} - s_k$.

Proof Uniqueness Suppose that σ is the sequence of partial sums for some infinite series $\sum_{j=1}^{\infty} x_j$. Then certainly $s_1 = x_1$, $s_2 = x_1 + x_2$, and so on. The first of these equations can be read as $x_1 = s_1$, so the first term of the desired series is determined by the given sequence σ. The second equation can then be rewritten in the equivalent form $x_2 = s_2 - x_1 = s_2 - s_1$; thus the second term of the desired series is also determined by σ. More generally, it is easy to show by Mathematical Induction that if $\sum_{j=1}^{\infty} x_j$ is an infinite series whose sequence of partial sums is σ, then $x_{k+1} = s_{k+1} - s_k$ for each $k \geq 1$. In other words, if such a series exists, it is unique, and the expression given in the statement of the proposition is correct.

Existence Based on the results just obtained, set $x_1 = s_1$, $x_2 = s_2 - s_1$, $x_3 = s_3 - s_2$, and so on; the general rule is given resursively by $x_{k+1} = s_{k+1} - s_k$ for each $k \geq 1$. It is easy to show by Mathematical Induction that $s_k = x_1 + x_2 + \ldots + x_k$ for each k.

IX.4.4 Remark

In light of the preceding result, it should be the case that every concept associated with infinite sequences ought to have a corresponding concept associated with infinite series. Indeed, we have already carried out such a correspondence in Definition (IX.4.1): the concepts of ‘limit of a sequence’ and ‘convergent sequence, when applied to the sequence of partial sums, correspond to the concepts of ‘value of an infinite series’ and ‘convergent infinite series’. The next result provides more examples of such correpondences between ‘sequence’ concepts and ‘series’ concepts. In each case, a standard concept for the sequence σ is translated to the equivalent property for the series $\sum_{k=1} \infty x_k$; the name of this concept used in the theory of infinite series is then written in boldface. If the sequence concept under discussion admits slight variations, the slight variations of the corresponding series concepts are usually left to the reader to write down.

IX.4.5 Theorem

General Situation In the next several examples, $\xi = (x_1, x_2, \ldots)$ is a sequence of terms, $\sum_{k=1}^{\infty} x_k$ is the corresponding infinite series $\sum_{\xi} = x_1 + x_2 + \ldots + x_k + \ldots$, and $\sigma = (s_1, s_2, \ldots)$ is the corresponding sequence of partial sums; that is, $s_k = x_1 + x_2 + \ldots + x_k$ for each index k.

(1) Sequence Concept: ‘The Cauchy Property’

A necessary and sufficient condition for the sequence σ to have the standard Cauchy Property is that the series $\sum_{k=1} x_k$ satisfy the following **Cauchy Property for Infinite Series**:

For every $\varepsilon > 0$ there exists B such that if n in $\mathbb{N}$ satisfies $n \geq B$, then $|x_{n+1} + x_{n+2} + \ldots + x_{n+k}| < \varepsilon$.

Note Cauchy’s original formulation of what we now call the ‘Cauchy Property’ was the series version, not the one for sequences.

(2) Sequence Concept: ‘Subsequence’

Suppose that $\tau = (t_1, t_2, \ldots)$ is a subsequence of the sequence σ of partial sums. That is, there exist indices k_1, $k_2, \ldots$, with $1 \leq k_1 < k_2 < \ldots$ such that $t_j = s_{k_j}$ for each j in $\mathbb{N}$. Then an infinite series $\sum_{j=1}^{\infty} z_j$ is the series whose sequence of partial sums is τ if, and only if, one has

$$z_1 = x_1 + x_2 + \ldots + x_{k_1}; \quad z_{j+1} = x_{k_j+1} + \ldots + x_{k_{j+1}} \text{ for each } j \in \mathbb{N}.$$

Note If one replaces each z_j in the infinite series $z_1 + z_2 + \ldots$ by the value given above, one gets

$$\sum_{j=1}^{\infty} z_j = (x_1 + x_2 + \ldots + x_{k_1}) + (x_{k_1+1} + x_{k_1+2} + \ldots + x_{k_2}) + (x_{k_2+1} + \ldots + x_{k_3}) + \ldots$$

In the expression on the right, the numbers x_k appear in the same order as in the original series $x_1 + x_2 + \ldots + x_k + \ldots$. Indeed, the only difference is the presence of the parentheses in the new series. Thus, one says that the series $\sum_{j=1}^{\infty} z_j$ is **obtained from $\sum_{k=1}^{\infty} x_k$ by inserting parentheses at the locations** $1, k_1, k_2, \ldots$. Likewise, one says that $\sum_{k=1}^{\infty} x_k$ is **obtained from the series $\sum_{j=1}^{\infty} z_j$ by removing parentheses**.

(3) Sequence Concept: 'Monotonicity'

A necessary and sufficient condition for the sequence σ to be monotonic up is that the terms x_k are all nonnegative. One then says that the series $\sum_{k=1}^{\infty} x_k$ is a **series of nonnegative terms**. (Technically, we need only that $x_k \geq 0$ for $k \geq 2$, but it is fairly conventional – and easy – to reduce to the case in which all the terms of the series, even the first, are nonnegative.) A similar remark relates 'monontonic-down sequences' and **series of nonpositive terms**, and for the the variants 'eventually monotonic up' and 'eventually monotonic down'.

The next result uses the correspondences described above to translate known facts about sequences into the corresponding facts about infinite series.

IX.4.6 Theorem

Let $\sum_{k=1}^{\infty} x_k$ be an infinite series of real numbers, and let $\sigma = (s_1, s_2, \ldots)$ be the corresponding sequence of partial sums.

(a) The series $\sum_{k=1}^{\infty} x_k$ is convergent if, and only if, it has the Cauchy Property.

(b) If the series $\sum_{k=1}^{\infty} x_k$ is convergent, then so is any series obtained from $\sum_{k=1}^{\infty}$ by inserting parentheses, and the values of the series are equal.

(c) If the terms x_k of the series $\sum_{k=1}^{\infty} x_k$ are all nonnegative, then either the series is convergent, or it diverges to $+\infty$. The former case occurs provided the sequence σ is bounded above, the latter if σ is *not* bounded above.

(d) A necessary condition for the series $\sum_{k=1}^{\infty} x_k$ to be convergent is is that $\lim_{k \to \infty} x_k = 0$; however, this condition is not sufficient.

The proofs of (a), (b) and (c) above follow from Theorem (IX.4.5) above, combined with Theorems (III.5.5), (III.2.1) (b) and (III.2.5) (b). Part (d) follows from Theorem (III.5.1), combined with Example (III.5.3).

IX.4.7 Examples

(1) Consider the infinite series $1 - 1 + 1 - 1 + \ldots$; that is, $\sum_{k=1}^{\infty} (-1)^{k+1}$. One sees that the corresponding sequence σ of partial sums is given by

$$s_k = \begin{cases} 1 & \text{if } k \text{ is odd} \\ 0 & \text{if } k \text{ is even} \end{cases}$$

In particular, the sequence σ does not have a limit, hence the infinite series is not defined. In particular, this shows that an infinite series with bounded partial sums need not be convergent. (In light of Part (c) of the preceding theorem, this conclusion would *not* be valid if the terms of the series were all nonnegative.)

(2) Consider the same series as in (1). Insert parentheses into this series at the locations 1, 3, 5,... $2m-1,\ldots$ to obtain

$$(1-1)+(1-1)+(1-1)+\ \ldots\ = 0+0+\ \ldots$$

Clearly this new series does converge, and has value 0. This shows that the converse of Part (b) of the preceding theorem is not true. Otherwise stated, inserting parentheses in a convergent series does not change the convergence properties; but removing parentheses from a convergent series may yield a nonconvergent series.

(3) Consider again the same series as in (1), but now insert parentheses at the locations 2, 4, 6,... $2m,\ldots$ to obtain

$$1+(-1+1)+(-1+1)+(-1+1)+\ \ldots\ = 1+0+0+\ \ldots$$

Clearly this new series does converge, and has value 1. This also shows that the converse of Part (b) of the preceding theorem is not true. However, things are even worse than we thought: by comparing with the result of the preceding example, we see that it is possible to insert parentheses in two different ways, and obtain different values for the resulting pair of series.

Remark We know from the Generalized Associative Law for Addition that in a finite sum inserting or deleting parentheses has no effect on the final result. Examples (2) and (3) above tell us that we cannot expect all the usual laws of addition that work for finite ordered sums to necessarily work for convergent infinite ordered sums. We shall soon see that the situation is even worse: not even the Generalized Commutative Law extends to convergent infinite ordered sums. Fortunately, *some* facts that hold for finite sums do extend to convergent infinite ordered sums:

IX.4.8 Theorem

(a) Suppose that $\sum_{j=1}^{\infty} x_j$ is an infinite series which converges to a number L. Then for every real number c the series $\sum_{j=1}^{\infty}(cx_j)$ converges to cL. Symbolically:

$$\sum_{j=1}^{\infty} cx_j\ =\ c\left(\sum_{j=1}^{\infty} x_j\right).$$

(b) Suppose that $\sum_{j=1}^{\infty} x_j$ and $\sum_{j=1}^{\infty} y_j$ are infinite series which converge to the numbers L and M, respectively. Then the series $\sum_{j=1}^{\infty}(x_j+y_j)$ converges to $L+M$.

(c) Suppose that $x_1+x_2+\ \ldots\ +x_k+\ \ldots$ is a convergent series with value L. Then both of the following infinite series are convergent and have value L:

$$x_1+0+x_2+0+x_3+\ \ldots\ +0+x_k+\ \ldots\ \text{ and } 0+x_1+0+x_2+0+x_3+\ \ldots\ +0+x_k+\ \ldots$$

That is, inserting a 'zero' term between each term of the original series has no effect on either the convergence or the value of the series.

(d) Suppose that $\sum_{j=1}^{\infty} x_j$ and $\sum_{j=1}^{\infty} y_j$ are infinite series which converge to the numbers L and M, respectively. Let $\sum_{m=1}^{\infty} z_m$ be the series $x_1 + y_1 + x_2 + y_2 + \ldots$, which is obtained from the original pair of series by 'interlacing' their terms. Then $\sum_{m=1}^{\infty} z_k$ converges to $L + M$.

Proof Parts (a), (c) and (d) are left as exercises.

Proof of (b) Let $\xi = (x_1, x_2, \ldots)$ and $\eta = (y_1, y_2, \ldots)$ denote the sequences of terms from which these series are formed, and let σ_ξ and σ_η denote the corresponding sequences of partial sums of the series; thus,

$$\sigma_\xi = (x_1, x_1 + x_2, \ldots x_1 + x_2 + \ldots + x_k, \ldots) \text{ and } \sigma_\eta = (y_1, y_1 + y_2, \ldots y_1 + y_2 + \ldots + y_k, \ldots)$$

Now let $\zeta = (z_1, z_2, \ldots)$ denote the sequence given by adding corresponding terms of ξ and η, so that $z_j = x_j + y_j$ for all j. Then the k-th term of σ_ζ is

$$z_1 + z_2 + \ldots + z_k = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_k + y_k).$$

It follows by applying the Generalized Associative Law for Addition, Theorem (IX.1.3), to the right side of this last equation, that one can write

$$z_1 + z_2 + \ldots + z_k = (x_1 + x_2 + \ldots + x_k) + (y_1 + y_2 + \ldots + y_k).$$

In other words, $\sigma_\zeta = \sigma_\xi + \sigma_\eta$. The desired result now follows by applying the results of Theorem (III.3.2).

IX.4.9 Remarks

(1) At first glance Part (d) may appear to be just a restatement of Part (b). However, there is a subtle difference: The series under consideration in Part (b) involves lots of inserted parentheses, while the series in Part (d) does not. In light of the fact that one cannot freely remove parentheses in ordered infinite sums, Part (d) does not follow directly from Part (b).

(2) One can easily extend the statement of Part (c) to allow the insertion of arbitrarily many zeros between consecutive terms of the original series. The details are left as an exercise.

IX.4.10 Important Example

If one multiplies both sides of Equation (IX.10) by 1/2, one gets

$$\frac{1}{2}\ln(2) = \tfrac{1}{2} - \tfrac{1}{4} + \tfrac{1}{6} - \tfrac{1}{8} + \ldots \quad (*)$$

(To see that this equation does follows from Equation (IX.10), use Part (a) of Theorem (IX.4.8).)

Next, insert a zero term before the first term in the series on the right side of Equation ($*$); likewise, insert a zero term between each of the terms of that series. Then one gets

$$\frac{1}{2}\ln(2) = 0 + \tfrac{1}{2} + 0 - \tfrac{1}{4} + 0 + \tfrac{1}{6} + 0 - \tfrac{1}{8} + \ldots \quad (**)$$

(The justification for Equation ($**$) is Part (d) of Theorem (IX.4.8).)

Now add each term of the preceding series with the corresponding term of the Alternating Harmonic Series (see Equation (IX.10)), and use Part (b) of Theorem (IX.4.8), to get $\frac{1}{2}\ln(2) + \ln(2) =$

$$(0+1) \ + \ \left(\tfrac{1}{2}-\tfrac{1}{2}\right) \ + \ \left(0+\tfrac{1}{3}\right) \ - \ \left(\tfrac{1}{4}+\tfrac{1}{4}\right) \ + \ \left(0+\tfrac{1}{5}\right) \ + \ \left(\tfrac{1}{6}-\tfrac{1}{6}\right) \ + \ \left(0+\tfrac{1}{7}\right) \ - \ \left(\tfrac{1}{8}+\tfrac{1}{8}\right) \ + \ \ldots$$

After simplifying this becomes

$$\frac{3}{2}\ln(2) \ = \ 1 \ + \ 0 \ + \ \tfrac{1}{3} \ - \ \tfrac{1}{2} \ + \ \tfrac{1}{5} \ + \ 0 \ + \ \tfrac{1}{7} \ - \ \tfrac{1}{4} \ + \ \ldots$$

Now use Part (d) of Theorem (IX.4.8) once again to justify deleting all the zero terms above to finally get

$$\frac{3}{2}\ln(2) \ = \ 1 \ + \ \tfrac{1}{3} \ - \ \tfrac{1}{2} \ + \ \tfrac{1}{5} \ + \ \tfrac{1}{7} \ - \ \tfrac{1}{4} \ + \ \ldots$$

The final series has exactly the same terms as the Alternating Harmonic Series, including the appropriate signs, but written in a different order. And it turns out that adding up the same terms but in this new order changes the value of the infinite series: we obtain the value $(3/2)\ln(2)$ for this rearranged series, compared to the value $\ln(2)$ for the original Alternating Harmonic Series.

Conclusion The Generalized Commutative Law for Addition, which certainly holds for ordered finite sums, does not extend to ordered infinite sums.

We investigate this phenomenon further in the next section.

The next result seems somewhat specialized, but it turns out to be surprisingly useful.

IX.4.11 Theorem (Alternating-Series Test)

Hypotheses Suppose that $\alpha = (a_1, a_2, \ldots)$ is a sequence of real numbers such that

(i) $a_k > 0$ for each k in $\mathbb{N}$;
(ii) α is monotonic down
(iii) $\lim_{k\to\infty} a_k = 0$.

Conclusions

(a) The infinite series $a_1 - a_2 + a_3 - a_4 + \ldots + (-1)^{k-1}a_k + \ldots$ is convegent. More precisely, let s_k denotes the k-th partial sum of this series and L is the value of the series. Then one has

$$0 \le L \le a_1, \text{ and } 0 \le (-1)^k(L - s_k) \le a_{k+1} \text{ for all } k \text{ in } \mathbb{N} \qquad \text{(IX.11)}$$

(b) Similarly, the infinite series $-a_1 + a_2 - a_3 + \ldots + (-1)^k a_k + \ldots$ is convergent. Moreover, if s_k denotes the k-th partial sum of this series and L is the value of the series, then

$$-a_1 \le L \le 0, \text{ and } 0 \le (-1)^{k-1}(L - s_k) \le a_{k+1} \text{ for all } k \text{ in } \mathbb{N} \qquad \text{(IX.12)}$$

(c) If the sequence α is actually *strictly* decreasing, then all the inequalities in (IX.11) and (IX.12) can be replaced by *strict* inequalities.

Proof

(a) Note that $s_{2m+1} = s_{2m} + a_{2m+1} \ge s_{2m}$ and $s_{2m} = s_{2m-1} - a_{2m} \le s_{2m-1}$ for each m in $\mathbb{N}$, by Condition (i). Likewise, $s_{2m+2} = s_{2m} + a_{2m+1} - a_{2m+2} \ge s_{2m}$ and $s_{2m+1} = s_{2m-1} - a_{2m} + a_{2m+1} \le s_{2m-1}$ for each m in $\mathbb{N}$, by Condition (ii). Combining these results, one sees that the segments $\mathrm{Seg}\,[s_{k+1}, s_k]$, $k = 1, 2, \ldots$, forms a nested sequence of segments. Finally, Condition (iii) implies that $\lim_{k\to\infty} |s_{k+1} - s_k| = 0$. One can now apply the Nested-Segments Theorem (Part (b) Theorem (II.4.31)) to conclude that the intersection of these segments is a singleton set $\{L\}$ and that the sequence σ converges to L. It is also easy to see from the above that $0 \leq a_1 - a_2 = s_2 \leq s_1 = a_1$. Since $L \in \mathrm{Seg}\,[s_2, s_1]$, it follows that $0 \leq L \leq a_1$, as required. Also, since $s_{2m} \leq s_{2m+1}$ and $L \in \mathrm{Seg}\,[s_{2m}, s_{2m+1}]$ it follows that $0 \leq s_k \leq L \leq s_{k+1}$ if k is even, so $0 \leq L - s_k \leq s_{k+1} - s_k = a_{k+1}$ if k is even. Likewise, if k is odd then $s_{k+1} \leq L \leq s_{k+2} \leq s_k$, where the final inequality uses the fact that the segments are nested. It follows that if k is odd then $-a_{k+1} \leq s_{k+1} - s_k \leq L - s_k \leq 0$. Since $(-1)^k = -1$ when k is odd, this last can be written $0 \leq (-1)^k (L - s_k) \leq a_{k+1}$ when k is odd. Inequalities (IX.11) now follow.

The simple proofs of Parts (b) and (c) are left as exercises.

Among the most important examples of infinite series are the so-called 'Power Series'.

IX.4.12 Definition (Power Series)

(1) An expression of the form $a_0 + a_1 u + a_2 u^2 + \ldots a_j u^j + \ldots$, or, briefly, $\sum_{j=0}^{\infty} a_j u^j$, is called a **power series in the quantity u with corresponding coefficients a_0, $a_1, \ldots$.**

(2) If one differentiates each term of a power series $a_0 + a_1 u + a_2 u^2 + \ldots + a_j u^j + \ldots$ 'with respect to u', in the sense of elementary calculus, one obtains a second power series $b_0 + b_1 u + b_2 u^2 + \ldots + b_k u^k + \ldots$, where

$$b_0 = a_1,\ b_1 = 2a_2,\ \ldots b_k = (k+1)a_{k+1}, \ldots$$

One says that the power series $\sum_{k=0}^{\infty} b_k u^k$ is obtained from the original series, $\sum_{j=0}^{\infty} a_j u^j$, using **term-by-term differentiation**.

(3) Similarly, let $\sum_{m=0}^{\infty} c_m u^m$ be a power series whose coefficients are of the form

$$c_0 = C \text{ for some constant } C;\ c_m = \frac{a_{m-1}}{m} \text{ for each } m = 1, 2, 3, \ldots.$$

That is,

$$\sum_{m=0}^{\infty} c_m u^m = C + a_0 u + \frac{a_1}{2} u^2 + \ldots + \frac{a_{m-1}}{m} u^m + \ldots$$

One says that the power series $\sum_{m=0}^{\infty} c_m u^m$ is obtained from the original series, $\sum_{j=0}^{\infty} a_j u^j$, using **term-by-term antidifferentiation**. (Or, if one uses older terminology, one refers to this process as **term-by-term integration**.)

IX.4.13 Remarks

(1) The use of the word 'power' in the name 'power series' is clear: the dominant feature of such a series is the presence of the successive powers of the quantity u.

(2) We are already familiar with examples of power series; namely, the Taylor series $\displaystyle\sum_{j=0}^{\infty} \frac{f^{(j)}(c)}{j!}(x-c)^j$ about a center point c of a C^{∞} function f. Note that this is a power series in the quantity $u = (x - c)$. It is easy to show that the Taylor series about c of f' is obtained from the corresponding series for f by using term-by-term differentiation. Likewise, the Taylor series about c for an antiderivative of f is obtained using term-by-term antidifferentiation on the series for f.

(3) One should normally think of the role played by the quantity u in the power series $\sum_{j=0}^{\infty} a_j u^j$ as follows: for each choice of u in $\mathbb{R}$ one gets a particular infnite series. For instance, consider the power series

$$1 + u + u^2 + u^3 + \ldots + u^{k-1} +$$

If we set $u = 1/2$ we get the convergent geometric series

$$1 + \frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^{k-1}} + \ldots$$

whose value is 2. If, in contrast, we set $u = 2$, we get the divergent geometric series

$$1 + 2 + 2^2 + \ldots + 2^{k-1} + \ldots$$

The preceding examples suggest a couple if initial questions one might ask about a power series $\sum_{j=0}^{\infty} a_j u^j$:

(i) For which values of u does the series converge, and for which does it diverge?

(ii) What is the nature of a function that is defined by the values of a power series?

The principle convergence/divergence properties of power series are established in Section (IX.6) below.

(4) In principle, *every* infinite series can be viewed as arising from a power series. For example, consider a series $x_1 + x_2 + \ldots + x_k + \ldots$. If we set $a_j = x_{j+1}$ for each $j = 0, 1, 2, \ldots$, then the given series can be viewed as the result of setting $u = 1$ in the power series $\sum_{j=0}^{\infty} a_j u^j$. Sometimes this viewpoint turns out to be useful; sometimes, not.

(5) The particular labeling of the coefficients used here, with the initial index being 0 instead of the usual 1, is conventional: one wishes to allow the zero-th power of u to appear in such sums, and it then becomes convenient to have the label assigned to a coefficient match the exponent of the corresponding power of u. However, this index convention is frequently violated. In any event, note that with this convention the k-th term of the series $\sum_{j=0}^{\infty} a_j u^j$ is $a_{j-1} u^{j-1}$, not $a_j u^j$.

(6) In some contexts one is forced to consider similar infinite series which involve negative powers of the quantity u, or perhaps even fractional powers of u. Normally it is easy to modify the preceding definition, and any associated facts, to take care of these new situations.

IX.5 Absolute Convergence, Conditional Convergence

If $\xi = (x_1, x_2, \ldots)$ is an infinite sequence of numbers, then ξ is a real-valued function whose domain is the set $X = \mathbb{N}$. Thus the results of the preceding sections tell us that there are (at least) two ways of defining the concept of the 'sum of the values of the function ξ':

(i) the unordered sum $\sum_{\mathbb{N}} \xi$, as defined in Section (IX.2);

(ii) the infinite series $\sum_{k=1}^{\infty} x_k$, as defined in Section (IX.4).

Theorem (IX.2.14) shows that the concept of (ordered) infinite series is at least as general as that of unordered sum (over $\mathbb{N}$), and that these two concepts agree when both apply. However, Theorem (IX.2.10) shows that that the theory of unordered sums retains one of the main properties of finite sums, namely the Generalized Associative Law holds for unordered infinite sums. In contrast, Example (IX.4.7) shows that the theory of (ordered) infnite series does *not* have a general Associative Law; likewise, Example (IX.4.10) shows that the Generalized Commutative Law, valid for finite sums, does not hold for (ordered) infinite series.

In this section we explore in a little more depth the connections between the two types of infinite sums.

IX.5.1 Theorem

Suppose that $\xi = (x_1, x_2, \dots)$ is an infinite sequence of real numbers; that is, ξ is a real-valued function defined on $\mathbb{N}$.

(a) The unordered sum $\sum_{\mathbb{N}} \xi^+$ diverges to $+\infty$ if, and only if, the corresponding (ordered) infinite series $\sum_{k=1}^{\infty} x_k^+$ diverges to $+\infty$.

Likewise, the unordered sum $\sum_{\mathbb{N}} \xi^-$ diverges to $+\infty$ if, and only if the infinite series $\sum_{k=1}^{\infty} x_k^-$ diverges to $+\infty$.

(b) The following statements are equivalent:

(i) The unordered sum $\sum_{\mathbb{N}} \xi$ is convergent.

(ii) Each of the (ordered) infinite series $\sum_{k=1}^{+} x_k^+$ and $\sum_{k=1}^{\infty} x_k^-$ is convergent.

(iii) The (ordered) infinite series $\sum_{k=1}^{\infty} |x_k|$ is convergent.

(c) The unordered $\sum_{\mathbb{N}} \xi$ has value $+\infty$ if, and only if, the infinite series $\sum_{k=1}^{\infty} x_k^+$ diverges to $+\infty$, while the infinite series $\sum_{k=1}^{\infty} x_k^-$ is convergent.

Likewise, the unordered $\sum_{\mathbb{N}} \xi$ has value $-\infty$ if, and only if, the infinite series $\sum_{k=1}^{\infty} x_k^-$ diverges to $+\infty$, while the infinite series $\sum_{k=1}^{\infty} x_k^+$ is convergent.

(d) The unordered sum $\sum_X \xi$ is not defined if, and only if, each of the (ordered) series $\sum_{k=1}^{\infty} x_k^+$ and $\sum_{k=1}^{\infty} x_k^-$ diverges to $+\infty$.

Proof

(a) The 'only if' part follows from Theorem (IX.2.14). As for the converse, suppose that $\sum_{k=1}^{\infty} x_k^+ = +\infty$. Let $M > 0$ be given. Then there exists N such that if n in $\mathbb{N}$ satisfies $n \geq N$ then $\sum_{k=1}^{n} x_k^+ \geq M$. In particular, $\sum_W \xi^+ \geq M$ for the finite subset $W = \mathbb{N}_n$ of $\mathbb{N}$. It follows easily that $\sum_{\mathbb{N}} \xi^+ = +\infty$.

To prove the analogous result for ξ^-, apply what was just proved to the sequence $\tau = -\xi$.

(b) The equivalence of Statements (i) and (ii) follows directly from Part (a) and the definition of 'convergence' for unordered sums. To see that Statements (ii) and (iii) are equivalent, recall that for each index k one has

$$x_k^+ \leq x_k^+ + x_k^- = |x_k| \text{ and } x_k^- \leq x_k^+ + x_k^- \leq |x_k|$$

Thus if $\sum_{k=1}^{\infty} |x_k|$ is convergent then it is clear that $\sum_{k=1}^{\infty} x_k^+ \leq \sum_{k=1}^{\infty} |x_k| < +\infty$; likewise, $\sum_{k=1}^{\infty} x_k^- \leq \sum_{k=1}^{\infty} < +\infty$. In particular, both of the series $\sum_{k=1}^{\infty} x_k^+$ and $\sum_{k=1}^{\infty} x_k^-$ converge. Conversely, if each of the series $\sum_{k=1}^{\infty} x_k^+$ and $\sum_{k=1}^{\infty} x_k^-$ is convergent, then by Theorem (IX.4.8) one knows that $\sum_{k=1}^{\infty}(x_k^+ + x_k^-)$ converges; that is, $\sum_{k=1}^{\infty} |x_k|$ is convergent.

The simple proofs of Parts (c) and (d) are left as an exercise.

IX.5.2 Corollary

Let $\xi = (x_1, x_2, \dots) : \mathbb{N} \to \mathbb{R}$ is a sequence of numbers, and suppose that the infinite series $\sum_{k=1}^{\infty} |x_k|$ is convergent. Then:

(a) The series $\sum_{k=1}^{\infty} x_k$ is convergent, and

$$\sum_{k=1}^{\infty} x_k = \left(\sum_{k=1}^{\infty} x_k^+\right) - \left(\sum_{k=1}^{\infty} x_k^-\right)$$

(By Part (b) of the preceding theorem, both series on the right side of the equation are convergent.)

(b) If $\varphi : \mathbb{N} \to \mathbb{N}$ is a bijection of $\mathbb{N}$ onto $\mathbb{N}$, and if $\zeta = (z_1, z_2, \dots) = \xi \circ \varphi$, so $z_j = x_{\varphi(j)}$ for each j in $\mathbb{N}$,j=1 then $\sum_{j=1}^{\infty} z_j = \sum_{k=1}^{\infty} x_k$.

Proof

(a) By Part (b) of the preceding theorem one knows that the unordered sum $\sum_{\mathbb{N}} \xi$ is convergent. Then Theorem (IX.2.14), with the bijection φ in that result given by the fomula $\gamma(j) = j$ for all j in $\mathbb{N}$, implies that the (ordered) infinite series $\sum_{k=1}^{\infty} x_k$ is convergent and has the same value as $\sum_{\mathbb{N}} \xi$. This latter value equals $\sum_{\mathbb{N}} \xi^+ - \sum_{\mathbb{N}} \xi^-$, which in light of Part (a) of the preceding theorem, combined with Theorem (IX.2.14) again, implies the desired formula.

(b) This follows immediately by using Theorem (IX.2.14).

Remark The properties of the Alternating Harmonic Series discussed in the preceding section show that the converse of this corollary is not true. That is, if an infinite series $\sum_{k=1}^{\infty} x_k$ is convergent, the corresponding series $\sum_{k=1}^{\infty} |x_k|$ of absolute values need not be convergent. Likewise, rearrangements of the terms of an infinite series can affect the value of the series.

The preceding results lead to the following definition.

IX.5.3 Definition

Let $\sum_{k=1}^{\infty} x_k$ be a $\underline{\text{convergent}}$ infinite series. This series is said to be **absolutely convergent** provided the series $\sum_{k=1}^{\infty} |x_k|$ also is convergent. In contrast, the convergent series $\sum_{k=1}^{\infty} x_k$ is said to be **conditionally convergent** provided the series $\sum_{k=1}^{\infty} |x_k|$ is *not* convergent.

It is convenient to repeat some of the earlier results using the 'absolute' and 'conditional' terminology.

IX.5.4 Theorem

(a) A convergent infinite series $\sum_{k=1}^{\infty} x_k$ is $\underline{\text{absolutely}}$ convergent if, and only if, each of the series $\sum_{k=1}^{\infty} x^+$ and $\sum_{k=1}^{\infty} x_k^-$ is convergent. In that case one has

$$\sum_{k=1}^{\infty} x_k = \left(\sum_{k=1}^{\infty} x_k^+\right) - \left(\sum_{k=1}^{k} x_k^-\right) \text{ and } \sum_{k=1}^{\infty} |x_k| = \left(\sum_{k=1}^{\infty} x_k^+\right) + \left(\sum_{k=1}^{k} x_k^-\right)$$

(b) Suppose that $\sum_{k=1}^{\infty} x_k$ is absolutely convergent. If $(c_1, c_2, \dots)$ is a bounded sequence of real numbers, then the series $\sum_{k=1}^{\infty} c_k x_k$ is also absolutely convergent.

Likewise, if $\sum_{k=1}^{\infty} x_k$ and $\sum_{k=1}^{\infty} y_k$ are absolutely convergent, then so is are the series $\sum_{k=1}^{\infty}(x_k + y_k)$ and $\sum_{k=1}^{\infty}(x_k \cdot y_k)$.

(c) (‘Rearrangement Property’) Suppose that $\sum_{k=1}^{\infty} x_k$ is an absolutely convergent series, and that $\sum_{j=1}^{\infty} z_j$ is a rearrangement of the series $\sum_{k=1}^{\infty} x_k$; that is, there is a bijection $\varphi : \mathbb{N} \to \mathbb{N}$ of $\mathbb{N}$ onto $\mathbb{N}$ such that $z_j = x_{\varphi(j)}$ for each j in $\mathbb{N}$. Then $\sum_{j=1}^{\infty} z_j$ is also absolutely convergent, and one has $\sum_{j=1}^{\infty} z_j = \sum_{k=1}^{\infty} x_k$.

(d) (‘Stability Property’) Suppose that $\sum_{k=1}^{\infty} x_k$ is absolutely convergent. Then there is a constant B such that if ε satisfies $0 < \varepsilon < 1$, and if $\tau = (t_1, t_2, \dots)$ is a sequence of numbers such that $1 - \varepsilon < t_k < 1 + \varepsilon$ for each k, then

$$\left| \sum_{k=1}^{\infty} x_k - \sum_{k=1}^{\infty} t_k x_k \right| \leq \varepsilon B.$$

The proof consists essentially in using Theorem (IX.5.1) to relate the statements above to known properties of unordered sums, and is left as an exercise.

In contrast, some of the properties of *conditionally* convergent series are quite different.

IX.5.5 Theorem

(a) A convergent infinite series $\sum_{k=1}^{\infty} x_k$ is conditionally convergent if, and only if, each of the series $\sum_{k=1}^{\infty} x^{+}$ and $\sum_{k=1}^{\infty} x_k^{-}$ is divergent to $+\infty$. In particular, one cannot write $\sum_{k=1}^{\infty} x_k = \left(\sum_{k=1}^{\infty} x_k^{+}\right) - \left(\sum_{k=1}^{\infty} x_k^{-}\right)$, since the expression $\infty - \infty$ is not defined.

(b) Suppose that $\sum_{k=1}^{\infty} x_k$ is conditionally convergent. Then there exists a bounded sequence $(c_1, c_2, \dots)$ of real numbers such that the series $\sum_{k=1}^{\infty} c_k x_k$ is not convergent.

Likewise, there exist conditionally convergent series $\sum_{k=1}^{\infty} x_k$ and $\sum_{k=1}^{\infty} y_k$ such that the series $\sum_{k=1}^{\infty}(x_k + y_k)$ and $\sum_{k=1}^{\infty}(x_k \cdot y_k)$ are not conditionally convergent.

(c) (‘Riemann’s Rearrangement Theorem’) Suppose that $\sum_{k=1}^{\infty} x_k$ is a conditionally convergent series. Then:

(i) For each quantity L, with $-\infty \leq L \leq +\infty$, there exists a rearrangement of the given series whose value is L. That is, there is a bijection $\varphi : \mathbb{N} \to \mathbb{N}$ such that if one sets $z_k = x_{\varphi(k)}$ for each index k, then $\sum_{k=1}^{\infty} z_k = L$.

(ii) There exist rearrangements of the original series $\sum_{k=1}^{\infty}$ which do not have a value, finite or infinite.

(d) (‘Instability Property’) Suppose that $\sum_{k=1}^{\infty} x_k$ is conditionally convergent. Then there is no constant B such that if ε satisfies $0 < \varepsilon < 1$, and $\tau = (t_1, t_2, \dots)$ is a sequence of numbers such that $1 - \varepsilon < t_k < 1 + \varepsilon$ for each k, then

$$\left| \sum_{k=1}^{\infty} x_k - \sum_{k=1}^{\infty} t_k x_k \right| \leq \varepsilon B.$$

The simple proofs are left as an exercise.

IX.5.6 Remarks

Students in calculus courses are often mystified by the ‘absolute convergence’ and ‘conditional convergence’ terminology:

(1) Most texts define 'absolute convergence' thusly:

'A series $\sum_{k=1}^{\infty} x_k$ is said to be **absolutely convergent** provided the related series $\sum_{k=1}^{\infty} |a_k|$ is convergent.'

The text then proves what, to the student, sounds like a vacuous result:

'If a series converges absolutely, then it converges'.

This statement sounds vacuous to students because in ordinary language similar statements *would* be vacuous:

'If a woman is walking quickly, then she is walking'. 'If a man is very tall then he is tall.'

(2) In contrast, the definition of a series $\sum_{k=1}^{\infty} x_k$ being *conditionally* convergent assumes, from the start, that the series is convergent, and that the 'conditional' describes the nature of that convergence. We do not prove, for example, the theorem that 'A conditionally convergent series is convergent'.

(3) However, the phrase 'conditionally convergent' has its own problems for students. In fact, it makes it sound like there is doubt that the series is 'truly' convergent. It gets worse when we explain to the students that the 'conditional' refers to the fact that we might lose convergence if we rearranged the terms – as if the students had any intention of rearranging the terms at all!

Of course the underlying issue is that these words are being used in a technical sense, peculiar to mathematics; thus it is irrelevant that the 'ordinary language' sense of these words is different. Nevertheless, it might help teachers of mathematics if the creaters of mathematical jargon thought more about their choice of words.

For example, the 'Stability' and 'Instability' properties described in Theorems (IX.5.4) and (IX.5.5) might suggest using the phrases '*stably* convergent' and '*unstably* convergent' instead of the 'absolute' and 'conditional' terminology.

It might appear reasonable to divide the *divergent* infinite series into a pair of categories: 'absolutely divergent' (i.e., no rearrangement is convergent) and 'conditionally divergent' (i.e., the series diverges but some rearrangement converges.) However, that terminology is not of use here. Instead, we *do* find the following terminology to be of some help:

IX.5.7 Definition

A divergent infinite series $\sum_{k=1}^{\infty} x_k$ is said to **diverge badly** if it is not the case that $\lim_{k\to\infty} x_k = 0$. Such a series is said to **diverge very badly** if, in fact, the set $\{x_1, x_2, \ldots\}$ is not even bounded.

Remarks

(1) Note that if a series diverges badly or very badly, then there is no way to 'make it converge' by simply rearranging its terms or changing the signs of some of its terms. In particular, if $\sum_{k=1}^{\infty} |x_k|$ diverges badly, then so does the series $\sum_{k=1}^{\infty} x_k$.

(2) The technical meaning of the 'diverges badly' and 'diverges very badly' terminology introduced here is not standard in analysis.

We conclude this section with a pair of test for convergence that are particularly elegant.

IX.5.8 Theorem (Dirichlet's Test)

Suppose that $\sum_{k=1}^{\infty} x_k$ is an infinite series whose partial sums, $\sigma_1 = x_1$, $\sigma_2 = x_1 + x_2, \dots$, form a *bounded* sequence of numbers; note that it is *not* assumed that this series is convergent. Let $\alpha = (a_1, a_2, \dots)$ be a monotonic sequence which converges to 0. Then the series $\sum_{k=1}^{\infty} a_k x_k$ is convergent.

Proof Without loss of generality one may assume that the sequence α is monotonic *down.* (If α is monotonic up, replace α by $\alpha' = -\alpha$, and note that α' also converges to 0.) In particular, one must then have $a_k \geq 0$ for each index k. Consider now the infinite sum

$$a_1 - a_2 + a_2 - a_3 + a_3 - a_4 + \dots + a_k - a_{k+1} + a_{k+1} \dots \qquad (*)$$

It is clear that the sequence of partial sums for the series $(*)$ is

$$a_1,\, a_1 - a_2,\, a_1,\, a_1 - a_3,\, a_1,\, a_1 - a_4, a_1,\, \dots a_1,\, a_1 - a_k,\, \dots$$

Since, by hypthesis, one has $\lim_{m \to \infty} a_m = 0$, it follows from Theorem (III.2.14), the 'Odd/Even Convergence Theorem', that the series $(*)$ is convergent. Indeed, the monotonicity hypothesis on α implies that this series is *absolutely* convergent. Now apply Part (b) of Theorem (IX.5.4) to conclude that the series $\sum_{k=1}^{\infty} \sigma_k (a_k - a_{k+1})$ is absolutely convergent. Consider the partial sums $\sigma'_1,\, \sigma'_2, \dots$ of this last series:

$$\sigma'_1 = \sigma_1(a_1 - a_2),\, \sigma'_2 = \sigma_1(a_1 - a_2) + \sigma_2(a_2 - a_3),\, \sigma'_3 = \sigma_1(a_1 - a_2) + \sigma_2(a_2 - a_3) + \sigma_3(a_3 - a_4),$$

and so on. The general formula is

$$\sigma'_k = \sigma_1(a_1 - a_2) + \sigma_2(a_2 - a_3) + \sigma_3(a_3 - a_4) + \dots + \sigma_k(a_k - a_{k+1}) \qquad (**)$$

Now regroup these terms (which is allowed because this is a *finite* sum) to get

$$\sigma'_k = \sigma_1 a_1 + (\sigma_2 - \sigma_1)a_2 + (\sigma_3 - \sigma_2)a_3 + \dots + (\sigma_k - \sigma_{k-1})a_k - \sigma_k a_{k+1} \qquad (***)$$

However, one has $x_1 = \sigma_1$, and $x_k = \sigma_k - \sigma_{k-1}$ if $k \geq 2$. Thus Equation $(***)$ takes the form

$$\sigma'_k = a_1 x_1 + a_2 x_2 + \dots + a_k x_k - \sigma_k a_{k+1} \qquad (****)$$

Since, by hypothesis, the partial sums $\sigma_1, \sigma_2, \dots$ form a bounded set, and the numbers a_k approach 0 as k goes to ∞, it follows that the sequence $(\sigma'_1, \sigma'_2, \dots)$ is convergent; that is, the series $\sum_{k=1}^{\infty} a_k x_k$ is convergent, as claimed.

IX.5.9 Remark

The process of going from Equation $(**)$ to Equation $(****)$ proves the following formula:

$$a_1 x_1 + a_2 x_2 + \dots + a_k x_k = \sigma_k a_{k+1} + \sigma_1(a_1 - a_2) + \sigma_2(a_2 - a_3) + \sigma_3(a_3 - a_4) + \dots + \sigma_k(a_k - a_{k+1}) \qquad \text{(IX.13)}$$

This equation is called **Abel's Partial Summation Formula**.

IX.5.10 Corollary (Abel's Test)

Suppose that $\sum_{k=1}^{\infty} x_k$ is a *convergent* infinite series. Let $\beta = (b_1, b_2, \ldots)$ be a convergent monotonic sequence; note that we do not insist that β converge to *zero*. Then the series $\sum_{k=1}^{\infty} b_k x_k$ is convergent.

Proof Let $B = \lim_{k\to\infty} b_k$. Define $\alpha = (a_1, a_2, \ldots)$ by the rule $a_k = B - b_k$. Then it is clear that α is monotonic and converges to 0. It is also clear that the sequence of partial sums of the series $\sum_{k=1}^{\infty} x_k$, being convergent, is bounded. Thus Dirichlet's Test can be applied to conclude that the series $\sum_{k=1}^{\infty} a_k x_k$ is convergent. However, one also has $b_k = B - a_k$, and the series $\sum_{k=1} B x_k$ and $\sum_{k=1}^{\infty} a_k x_k$ both are convergent. Thus it follows from the usual algebraic rules for infinite series that the series $\sum_{k=1}^{\infty} b_k x_k$ is convergent. Indeed, one has

$$\sum_{k=1}^{\infty} B x_k - \sum_{k=1}^{\infty} a_k x_k = \sum_{k=1}^{\infty} (B - a_k) x_k = \sum_{k=1}^{\infty} b_k x_k.$$

IX.6 Convergence Tests for Series of Nonnegative Terms

In light of the results in the preceding section, it is important to study the properties of those infinite series whose terms are all nonnegative.

The first such property is given by the familiar 'comparison tests' from elementary calculus.

IX.6.1 Theorem (The Basic Comparison Tests for Convergence/Divergence)

Suppose that $\alpha = (a_1, a_2, \ldots)$ and $\beta = (b_1, b_2, \ldots)$ are sequences of real numbers such that for some real number $c > 0$ one has

$$0 \leq a_k \leq c b_k \text{ for all } k \text{ in } \mathbb{N} \quad (*)$$

(a) If $\sum_{k=1}^{\infty} b_k$ is convergent, then so is $\sum_{k=1}^{\infty} a_k$. More precisely, one has

$$0 \leq \sum_{k=1}^{\infty} a_k \leq c \left(\sum_{k=1}^{\infty} b_k \right), \text{ with equality if, and only if, } a_k = c b_k \text{ for all } k.$$

(b) If $\sum_{k=1}^{\infty} a_k$ is divergent, then so is $\sum_{k=1}^{\infty} b_k$. More precisely, one has

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} b_k = +\infty.$$

The simple proofs are left as an exercise.

The next results study the convergence properties for the special class of power series – see Definition (IX.4.12) above – whose terms are nonnegative.

IX.6.2 Theorem

Consider a power series $\sigma = \sum_{j=0}^{\infty} a_j r^j$, in which the quantities a_j and r are all assumed to be nonnegative. Let S_σ be the set of all real numbers $r \geq 0$ such that the sequence $(a_0, a_1 r, a_2 r^2, \ldots a_j r^j, \ldots)$ is bounded; note that $S_\sigma \neq \emptyset$ since clearly $0 {\in} S$. Let $R = \sup S_\sigma$ (which exists, since $S_\sigma \neq \emptyset$). Then the series $\sum_{j=0}^{\infty} a_j r^j$ is convergent if $r < R$ and is divergent if $r > R$. More precisely:

(a) Suppose that $R = 0$. Then the series σ converges if $r = 0$, but diverges very badly (see Definition (IX.5.7)) if $r > 0$.

(b) Suppose that $R > 0$. Then for each r such that $0 \leq r < R$, there exists a number ρ, with $0 \leq \rho < 1$, and a real number $B > 0$, such that

$$a_j t^j \leq B\rho^j \text{ for each } j = 0, 1, 2, \ldots \text{ and each } t \text{ in } [0, r].$$

In constrast, if $r > R$ then the series σ diverges very badly. (Note that if $R = +\infty$ then this part of the statement is vacuously true, since the condition $r > +\infty$ can never occur.)

Proof

(a) The hypothesis that $R = 0$ implies that $S_\sigma = \{0\}$. If $r = 0$ then the series reduces to the trivial series $a_0 + 0 + 0 + \ldots$, which of course converges to a_0. In contrast, if $r > 0$ then $r \notin S$, hence (by definition of S_σ) the sequence $(a_0, a_1 r, \cdots a_k r^k, \ldots)$ is unbounded. It follows easily that the series $\sum_{j=0}^{\infty} a_j r^j$ diverges very badly.

(b) Suppose $0 \leq r < R$. Then, by the fact that R is the supremum of the set S_σ, there exists r_1 in S such that $r < r_1 \leq R$. Thus, there exists a number $B > 0$ such that $0 \leq a_j r_1^j \leq B$ for all j. Now suppose that $0 \leq t \leq r$. One gets

$$0 \leq a_j t^j \leq a_j r^j = a_j r_1^j \left(\frac{r}{r_1}\right)^j \leq B\rho^j,$$

for all j, where $\rho = r/r_1$. Since $0 \leq r < r_1$, it follows that $0 \leq \rho < 1$, as required.

Suppose instead that $r > R$. Then, since R is an upper bound of the set S_σ, it follows that r is not in S_σ. In particular, the sequence $(a_0, a_1 r, \ldots a_j r^j, \ldots)$ is not bounded above. It follows that the series $\sum_{j=0}^{\infty} a_j r^j$ is very badly divergent.

IX.6.3 Definition (Radius of Convergence of a Power Series)

Let $\sum_{j=0}^{\infty} a_j r^j$ and R be as in the preceding theorem. Then one calls R the **radius of convergence** of this power series.

More generally, if $\sum_{j=0}^{\infty} b_j u^j$ is a *general* power series, i.e., no restrictions that the coefficients or u be nonnegative, then the radius of convergence of the associated series $\sum_{j=0}^{\infty} |b_j||u|^j$ is also called the radius of convergence of the (general) series $\sum_{j=0}^{\infty} b_j u^j$.

IX.6.4 Remarks

(1) The 'radius' terminology comes from the archetype of the concept of 'power series'; namely, the Taylor series $\sum_{j=0}^{\infty} a_j (x-c)^j$, about a point c, of a C^∞ function f; here one has $a_j = f^{(j)}(c)/j!$. In that context one thinks of c as the 'center point' of the series, so that $r = |x-c|$ is the distance from the center.

(2) The preceding theorem can be interpreted as saying that a power series is convergent (in a rather strong sense) *inside* the radius of convergence, i.e., when $r < R$, while it is divergent (in a rather strong sense) *outside* the radius of convergence, i.e., when $r > R$. In particular, the theorem says nothing when $r = R$; that is, *at* the radius of convergence; and as some examples below illustrate, nothing general *can* be said when $r = R$.

Note that this last statement holds even for a 'general' power series; that is, a series $\sum_{j=0}^{\infty} b_j u^j$ for which there is no restriction that b_j or u be nonnegative. Indeed, suppose that R is the radius of convergence of such a series, so that R is the radius of the series $\sum_{j=0}^{\infty} |b_j||u|^j$. This latter series is convergent when $|u| < R$, and thus by Corollary (IX.5.2) the original series $\sum_{j=0}^{\infty} b_k u^k$ also converges. Likewise, if $|u| > R$, then the series $\sum_{j=0}^{\infty} |b_j||u|^j$ diverges *very badly*, and thus the same is true for the series $\sum_{j=0}^{\infty} b_j u^j$.

(3) The preceding example tells us that we could have defined the quantity R described in Theorem (IX.6.2) directly in terms of the convergence/divergence of the given power series: R is the unique quantity such that the series is convergent when $r < R$ and divergent when $r > R$. Many texts do define R that way, with no reference to the boundedness of the sequence of terms of the power series. The advantage of focusing on the boundedness of the sequence of terms of the power series is that often it is easier to check for that boundedness than it is to check directly for the convergence of the series.

IX.6.5 Examples

(1) The power series $\sigma = \sum\limits_{j=0}^{\infty} \dfrac{r^j}{j!}$ has radius of convergence $R = +\infty$. Indeed, as is shown in Example (III.2.11) (3), the sequence $(1, r, r^2/2!, r^3/3!, \dots)$ converges to 0, and thus is bounded, for each $r \geq 0$. Thus the set S_σ referred to in Theorem (IX.6.2) equals the interval $[0, +\infty)$, so $\sup S_\sigma = +\infty$.

(2) Consider the series $\sigma = \sum_{j=0}^{\infty} (j!) r^j$. Then it is easy to see that if $r > 0$ then the sequence $(1, r, (2!)r^2, (3!)r^3, \dots)$ diverges to $+\infty$. It follows that $S_\sigma = \{0\}$, hence $R = 0$.

(3) The geometric series $\sigma = 1 + r + r^2 + r^3 + \dots$ clearly has radius of convergence $R = 1$. Indeed, the sequence $(1, r, r^2, \dots)$ is bounded above by 1 if $0 \leq r \leq 1$, and is unbounded if $r > 1$. Thus, $S_\sigma = [0, 1]$, so $R = \sup S_\sigma = 1$.

More generally, if $R_0 > 0$, then one can show that the geometric series $\sigma = \sum\limits_{j=0}^{\infty} \left(\dfrac{r}{R_0}\right)^j$ has $R = R_0$.

Notice that with such a geometric series the series fails to converge when $r = R$. Indeed, when $r = R$ the series reduces to $1 + 1 + 1 + \dots$. In particular, it diverge badly, but not *very* badly, when $r = R$.

(4) Consider the series $\sigma = \sum\limits_{k=1}^{\infty} \dfrac{r^k}{k}$; this series is obtained from the origial geometric series using term-by-term antidifferentiation (see Part (3) of Definition (IX.4.12)). It is obvious that if $0 \leq r \leq 1$ then $r \in S_\sigma$.

The situation for $r > 1$ is not so obvious. However, the following analysis provides the answer: Let $x_k = \ln(r^k/k)$. Then, by the usual properties of logarithms, one has

$$x_k = k \ln(r) - \ln(k) = k \cdot \left(\ln(r) - \frac{\ln k}{k}\right)$$

One can easily show that $\lim_{k\to\infty}(\ln k)/k = 0$; for example, use L'Hôpital's Rule. Since $\ln r > 0$ when $r > 1$, it then follows easily, using the Archimedean Principle, that $\lim_{k\to\infty} x_k = +\infty$. Thus one gets

$$\lim_{k\to\infty} \frac{r^k}{k} = \lim_{k\to\infty} e^{x_k} = +\infty \text{ when } r > 1.$$

Combining this with what was shown before, one sees that $S_\sigma = [0,1]$, and so the radius of convergence of the given series is $R = 1$. Note that in this case the series reduces to the Harmonic Series $1 + 1/2 + 1/3 + \ldots$ when $r = R = 1$. In particular, the series diverges at the radius of convergence itself; however, since the terms $1/k$ of the Harmonic Series converge to 0, it turns out in this example that the divergence of the power series when $r = R$ is *not* 'bad'.

(5) More generally, it is easy to modify the preceding argument to show that if m is any fixed natural number such that $m \geq 2$, then the power series $\displaystyle\sum_{k=1}^{\infty} \frac{r^k}{k^m}$ has radius of convergence $R = 1$. However, in this case the power series is convergent when $r = R$.

(6) Now let m be a fixed positive integer, and consider the power series $\sigma = \sum_{k=1}^{\infty} k^m r^{k-1}$; note that if $m = 1$ then this series is obtained from the Geometric Series using term-by-term differentiation (see Part (2) of Definition (IX.4.12)). It is clear that if $r \geq 1$ then the sequence of terms is unbounded, but it is less clear what happens when $0 < r < 1$. However, by an argument similar to that used in Example (4) above one can show that if $0 < r < 1$ then $\lim_{k\to\infty} k^m r^{k-1} = 0$. In particular, $r \in S_\sigma$ when $0 \leq r < 1$. Thus once again one has $R = 1$. Note, however, that in this case the power series diverges very badly when $r = R$.

Remark The preceding examples illustrate the fact that the convergence behavior of a power series of nonnegative terms at the radius of convergence can vary wildly from case to case: the series may converge when $r = R$, or it may diverge but not badly, or it may diverge badly but not very badly, or it may diverge very badly. Compare this to the behavior just inside the radius, where the series always converges, and the behavior just outside the radius, where the divegence is always very bad. The moral of the story is: be very careful when dealing with a power series at its radius of convergence.

Examples (4) and (6) above can be used to prove the following important result.

IX.6.6 Theorem (Radius of Convergence and Term-by-Term Differentiation and Integration)

Suppose that $\sum_{j=0}^{\infty} a_j r^j$ is a power series for which the coefficients a_j and the quantity r are all nonnegative. Let R denote the radius of convergence of this power series. Then R is also the radius of convergence of any power series obtained from $\sum_{j=0}^{\infty} a_j r^j$ using term-by-term differentiation or term-by-term integration a finite number of times.

Proof Suppose that $\sum_{k=0}^{\infty} b_k r^k$ is obtained from the original series using term-by-term differentiation a single time; thus $b_k = (k+1)a_{k+1}$ for each $k = 0, 1, 2, \ldots$. Let R' be the radius of convergence of this derived power series. By the definition of the radius of convergence R of the original series, one knows that the sequence $(a_0, a_1 r, a_2 r^2, \ldots)$ is unbounded if $r > R$. Since $b_k r^k = \Big((k+1)a_{k+1} r^{k+1}\Big)$, it follows that the sequence $(b_0, b_1 r, b_2 r^2, \ldots)$ is also unbounded when $r > R$. In particular, $R' \leq R$.

Now suppose that $0 < r < R$, and let r_1 be a number such that $0 < r < r_1 < R$. By the definition of the radius of convergence R, one knows that the sequence $(a_0, a_1 r_1, a_2, r_1^2, \ldots)$ is bounded. Let M be a real number such that $a_j r_1^j \leq M$ for all $j = 0, 1, 2, \ldots$. Note that

$$b_k r^k = (k+1)a_{k+1}r^k = a_{k+1}r_1^{k+1}\left((k+1)\frac{r}{r_1}\right)^{k+1}\frac{1}{r} \leq \frac{M}{r}\left((k+1)\rho^{k+1}\right),$$

where $\rho = r/r_1 < 1$. It follows from the results obtained in Example (IX.6.5) (6) above that $\lim_{k\to\infty}(k+1)\rho^{k+1} = 0$. Thus one has $\lim_{k\to\infty} b_k r^k = 0$; in particular, the terms of this sequence are bounded. It follows that $r \leq R'$. Since this is true for every r in the interval $(0, R)$, it follows that $R \leq R'$. Combining these results yields $R' = R$, as claimed. The fact that the radius of convergence remains unchanged under repeated differentiations of the original poser series now follows by using Mathematical Induction.

Next, let $\sum_{m=0}^{\infty} c_m r^m$ be obtained from the original power series using term-by-term antidifferentiation. Then the original power series $\sum_{j=0}^{\infty} a_k r^k$ can be obtained by differentiating the series $\sum_{m=0}^{\infty} c_m r^m$ term-by-term. Thus, by what was just proved for term-by-term differentiation, the series $\sum_{m=0}^{\infty} c_m r^m$ and $\sum_{j=0}^{\infty} a_j r^j$ have the same radius of convergence. And once again the result for repeated term-by-term antidifferentiation of the original series follows by Mathematical Induction.

The next result provides a 'formula' for the radius of convergence R in terms of the coefficients a_j.

IX.6.7 Theorem (Cauchy's Power-Series Theorem)

Suppose that $\sum_{j=0}^{\infty} a_j r^j$ is a power series in which the quantity r and the coefficients a_j, $j = 0, 1, 2, \ldots$ are all nonnegative. Let R denote the corresponding radius of convergence of this series, and let $L = \limsup_{k\to\infty} \sqrt[k]{a_k}$. Then $R = 1/L$.

Note: As is customary, we use the convention here that $R = +\infty$ when $L = 0$ and $R = 0$ when $L = +\infty$.

Proof Suppose that $c > L$. Then, by the basic properties of 'lim sup', there exists N in $\mathbb{N}$ so that if k is an index such that $k \geq N$, then $\sqrt[k]{a_k} < c$. Thus for $k \geq N$ one has $a_k < c^k$, hence $a_k\left(\frac{1}{c}\right)^k < 1$. It follows that the sequence $\left(a_0, a_1\left(\frac{1}{c}\right), a_2\left(\frac{1}{c}\right)^2, \ldots\right)$ is bounded, hence $\frac{1}{c} \leq R$. Since this is true for *every* $c > L$, it follows that $\frac{1}{L} \leq R$.

Next, suppose that $0 < c_1 < c_2 < L$. Then there exist infinitely many indices k such that $\sqrt[k]{a_k} > c_2$. That is, there exist infinitely many k such that $a_k > c_2^k$. Divide both sides by c_1^k to get

$$a_k\left(\frac{1}{c_1}\right)^k > \left(\frac{c_2}{c_1}\right)^k \qquad (*)$$

Since $c_2/c_1 > 1$, and $(*)$ holds for infinitely many values of k, it follows that the sequence $\left(a_0, a_1\left(\frac{1}{c_1}\right), \ldots\right)$ is unbounded. Thus $1/c_1 \geq R$. Since this is true for every c_1 such that $0 < c_1 < L$, it follows that $1/L \geq R$.

Combine the preceding results to get $R = \frac{1}{L}$, as required.

Remark Cauchy introduced the concept of the limit supremum of a sequence in his treatment of the preceding result; he called it the 'upper limit' (but in French, of course).

Cauchy's Power Series Theorem can be used to prove a pair of well-known tests for convergence.

IX.6.8 Theorem (The Root Test)

Suppose that $\sum_{j=0}^{\infty} a_j$ is a series of nonnegative terms. Let $L = \limsup_{k\to\infty} \sqrt[k]{a_k}$.

(a) If $L < 1$ then the series $\sum_{j=0}^{\infty} a_j$ is convergent.

(b) If $L > 1$ then the series $\sum_{j=0}^{\infty} a_j$ is badly divergent.

Proof

Note that, by Cauchy's Power Series Theorem, the radius of convergence of the power series $\sum_{j=0}^{\infty} a_j r^j$ is $R = 1/L$.

(a) If $L < 1$ then $R = 1/L > 1$. Then it follows from Theorem (IX.6.2) that the power series is convergent at $r = 1$.

(b) Similarly, if $L > 1$ then $R = 1/L < 1$, so $1 > R$. Then it follows from the same theorem that the series diverges badly when $r = 1$.

IX.6.9 Theorem (The Ratio Test)

Suppose that $\sum_{j=0}^{\infty} a_j$ is a series of *positive* terms. Let $L^+ = \limsup_{j\to\infty} \dfrac{a_{j+1}}{a_j}$, and let $L^- = \liminf_{j\to\infty} \dfrac{a_{j+1}}{a_j}$.

(a) If $L^+ < 1$ then the series converges.

(b) If $L^- > 1$ then the series diverges badly.

Proof

(a) By the basic properties of 'limsup', if $L^+ < c < 1$, there exists N in $\mathbb{N}$ so that if $j \geq N$ then $a_{j+1}/a_j < c$. It follows that for every m in $\mathbb{N}$ one has $a_{N+m} < c^m a_N$. Divide both sides by c^{N+m} to get $a_{N+m}(1/c)^{N+m} < a_N(1/c)^N$ for all m. In particular, the sequence $(a_0, a_1(1/c), a_2(1/c)^2, \ldots)$ is bounded, and thus $1/c \leq R$, where R is the radius of convergence of the power series $\sum_{j=0}^{\infty} a_j r^j$. Since $R \geq 1/c > 1$, it follows that this power series is convergent when $r = 1$; that is, the original series converges, as claimed.

(b) Left as an exercise.

There is one more standard convergence test for series of nonnegative terms that should be mentioned.

IX.6.10 Theorem (The Integral Test)

Suppose that $f : [m,\infty) \to \mathbb{R}$ is a continuous function which is monotonic down on $[m,\infty)$, where $m \in \{0,1,2,\ldots\}$. Let $F : [m,\infty) \to \mathbb{R}$ be the antiderivative of f on $[m,\infty)$ such that $F(m) = 0$; that is, $F = D_m^{-1} f$. Let $L = \lim_{x\to+\infty} F(x)$. (It is easy to see that this limit exists.)

(a) If L is finite then the series $\sum_{k=m}^{\infty} f(k)$ converges, and one has

$$L + f(m) \geq \sum_{k=m}^{\infty} f(k) \geq L.$$

(b) If $L = +\infty$ then $\sum_{k=1}^{\infty} f(k) = +\infty$.

The simple proof is left as an exercise.

IX.6.11 Examples

(1) (The 'p-series Test') Let p be a positive real number. Then the series $\displaystyle\sum_{k=1}^{\infty}\frac{1}{k^p}$ is called the **p-series**.

Case 1 Suppose that $p = 1$. Let $f(x) = 1/x$ for all $x > 0$, so that $D_1^{-1}f = \ln$. Note that $\lim_{x\to\infty}\ln x = +\infty$. Since f is strictly decreasing on $[1,+\infty)$, one can use the integral test to conclude that the series $\sum_{k=1}^{\infty} 1/k$ diverges to $+\infty$; that is, the p-series diverges if $p = 1$. Of course we alredy know this from before: this is the Harmonic Series.

Case 2 Suppose that $p > 1$. Let $f : (0,+\infty)\to\mathbb{R}$ be given by $f(x) = 1/x^p$; clearly this function is continuous and strictly decreasing on $(0,+\infty)$. One computes that $D_1^{-1}f(x) = \dfrac{-1}{(p-1)x^{p-1}}$, from which one gets $L = 1/(p-1)$. In particular, the p-series is convergent if $p > 1$.

Case 3 A similar argument shows that the p-series is divergent if $0 < p < 1$.

(2) Consider the series $\displaystyle\sum_{k=2}^{\infty}\frac{1}{(k\ln k})$. Define $f : [2,+\infty)\to\mathbb{R}$ by the rule $f(x) = 1/(x\ln x)$; note that f is continuous and monotonic down on the interval $[2,+\infty)$. It is easy to verify that $D_2^{-1}f(x) = \ln(\ln x) - \ln(\ln 2)$. Clearly this last function diverges to $+\infty$. It follows from the Integral Test that the given series diverges.

Remark The use of 'integral' in the phrase 'integral text' reflects the older meaning of integral as 'antiderivative'.

IX.7 Series of Functions

In Section (VIII.2) we discussed convergence properties of *sequences* of functions defined on a set. Since the convergence of an infinite series can be reduced to facts about convergence of sequences, it should come as no surprise that many of the results in Section (VIII.2) have analogs in the context of infnite series.

IX.7.1 Definition

Let $\varphi = (f_1, f_2, \ldots f_k, \ldots)$ be a sequence of real-valued functions defined on a nonempty subset X of $\mathbb{R}$. Associated with the expression $\sum_{k=1}^{\infty} f_k$ is the corresponding sequence $\sigma = (s_1, s_2, \ldots)$ of **partial sums**, where for each index k one has

$$s_k(x) = f_1(x) + \ldots + f_k(x) \text{ for each } x \text{ in } X.$$

(1) The infinite series $\sum_{k=1}^{\infty} f_k$ is said to **converge at a point** x **of** X provided the numerical sequence $(s_1(x), s_2(x), \ldots)$ is convergent.

(2) The series $\sum_{k=1}^{\infty} f_k$ is said to **converge pointwise on** X to a function $g : X\to\mathbb{R}$ provided the corresponding sequence of partial sums of the series converges pointwise on X to g, in the sense of Definition (VIII.2.1) (1); eqivalently: provided $\sum_{k=1}^{\infty} f_k(x) = g(x)$ for every x in X.

(3) The series $\sum_{k=1}^{\infty} f_k$ is said to **converge uniformly on** X to a function $g : X\to\infty$ provided the corresponding sequence of partial sums of the series converges uniformly on X to g, in the sense of Definition (VIII.2.1) (2).

(4) The series $\sum_{k=1}^{\infty} f_k$ is said to be **pointwise Cauchy on** X provided that the corresponding sequence of partial sums is pointwise Cauchy on X, in the sense of Definition (VIII.2.7) (1). Likewise, the series $\sum_{k=1}^{\infty} f_k$ is said to be **uniformly Cauchy on** X provided that the corresponding sequence of partial sums is uniformly Cauchy on X, in the sense of Definition (VIII.2.7) (2).

IX.7.2 Theorem

Let X be a nonempty subset of $\mathbb{R}$ and let $\sum_{k=1}^{\infty} f_k$ be an infinite series of functions with domain X.

(a) A necessary and sufficient condition for this series to converge pointwise on X to some function is that it be pointwise Cauchy on X.

(b) A necessary and sufficient condition for this series to converge uniformly on X to some function is that it be uniformly Cauchy on X.

Proof Apply Theorem (VIII.2.8) to the sequence of partial sums of the given series.

IX.7.3 Theorem (The 'Uniform-Convergence-Preserves-Continuity' Theorem for Series)

Let X be a nonempty subset of $\mathbb{R}$.

(a) Let $\sum_{k=1}^{\infty} f_k$ be an infinite series of functions, each with domain X. Assume that the series converges uniformly on X to a function $g : X \to \mathbb{R}$. If each summand f_k is continuous at a point c of X, then g is also continuous at c. Likewise, if each function f_k is continuous on X, then g is continuous on X.

(b) Let $\sum_{k=1}^{\infty} f_k$ be an infinite series of functions which are defined and contiuous on X. Assume that the series converges pointwise on X to a function $g : X \to \mathbb{R}$, and that for each interval $[a, b]$ for which $X \cap [a, b] \neq \emptyset$ the convergence is uniform. Then g is continuous on X.

Proof Apply Theorem (VIII.2.4) to the sequence of partial sums associated with the series $\sum_{k=1}^{\infty} f_k$.

IX.7.4 Theorem (The 'Uniform-Convergence and Term-by-Term Antidifferentiation' Theorem for Series)

Let $\sum_{k=1}^{\infty} f_k$ be a sequence of real-valued functions defined on an open interval I in $\mathbb{R}$. Assume that this series converges pointwise on I to a function $g : I \to \mathbb{R}$, and that on each closed bounded subinterval $[a, b]$ of I the convergence is uniform. If each of the functions f_k has an antiderivative on I, then g has an antiderivative on I. More precisely, fix a point c in I, and set $F_k = D_c^{-1} f_k$. The the series $\sum_{k=1}^{\infty} F_k$ converges pointwise on I to a function $G : I \to \mathbb{R}$ such that $G'(x) = g(x)$ for all x in I, and $G(c) = 0$. Moreover, the series $\sum_{k=1}^{\infty} F_k$ converges uniformly to G on every closed bounded subinterval $[a, b]$ of I.

Proof Apply Theorem (VIII.2.10) to the sequence of partial sums of the given series.

The next result does not seem to follow by simply applying a result from Section (VIII.2) directly to the sequence of partial sums of a series of functions.

IX.7.5 Theorem (The Weierstrass 'M-Test')

Let $(f_1, f_2, \ldots)$ be a sequence of real-valued functions defined on a nonempty subset X of $\mathbb{R}$. Suppose that there exists a sequence of nonnegative numbers $M_1, M_2, \ldots$ such that for each index k one has $|f_k(x)| \leq M_k$ for all x in X. If the numerical series $\sum_{k=1}^{\infty} M_k$ is convergent, then the series $\sum_{k=1}^{\infty} f_k$ is uniformly convergent on X.

Proof Let $\varepsilon > 0$ be given, and let N be large enough that if $n \geq N$, then $0 \leq M_{n+1} + M_{n+2} + \ldots + M_{n+k} < \varepsilon$ for every k in $\mathbb{N}$. Then one has

$$|f_{n+1}(x) + f_{n+2}(x) + \ldots + f_{n+k}(x)| \leq |f_{n+1}(x)| + |f_{n+2}(x)| + \ldots + |f_{n+k}(x)| \leq$$

$$M_{n+1} + M_{n+2} + \ldots + M_{n+k} < \varepsilon \text{ for all } x \text{ in } X.$$

Thus, the series $\sum_{k=1}^{\infty} f_k$ is uniformly Cauchy on X. The desired result now follows easily.

Remark The name 'M Test' for this result is completely standard in analysis, so the reader should certainly be familiar with this name and know to which result it refers. Nevertheless, it is often a bad idea to tie the name of a result or concept too closely with irrelevant features such as the specific notation used by a particular author. For instance, suppose one wished to apply the Weierstrass test in a context in which the letters M_k were already being used for something else. Then to refer to the 'M Test' but use letters other than M would seem a bit strange, and might even cause confusion. Perhaps a better name for the preceding result would have been 'The Weierstrass Uniform-Comparison Test'.

Important Case – Functions Defined by Power Series

IX.7.6 Definition

(1) Suppose that $\sum_{j=0}^{\infty} a_j(x-c)^j$ is a power series whose radius of convergence is $R > 0$. This series defines a function $f : (c-R, c+R) \to \mathbb{R}$ by the rule $f(x) = \sum_{j=0}^{\infty} a_j(x-c)^j$ for each x in the interval $(c-R, c+R)$. (If $R = +\infty$, then we interpret $(c-R, c+R)$ to be the interval $(-\infty, +\infty)$; that is, the interval $\mathbb{R}$.) We refer to f as **the function determined by the series** $\sum_{j=0}^{\infty} a_j(x-c)^j$.

(2) Suppose that $f : (c-R, c+R) \to \mathbb{R}$ is a function defined on an interval of the form $(c-R, c+R)$, where $c \in \mathbb{R}$ and $0 < R \leq +\infty$; we allow the full domain of f to be a proper superset of this interval. Then one says that f **is represented on** $(c-R, c+R)$ **by the power series** $\sum_{j=0}^{\infty} a_j(x-c)^j$ provided the radius of convergence of this power series is at least R, and on the interval $(c-R, c+R)$ the function f agrees with the function determined by this power series.

Functions of the type described in the preceding definition has some very pleasant properties.

IX.7.7 Theorem

Suppose that a function $f : (c-R, c+R) \to \mathbb{R}$ is represented on a nonempty open interval $(c-R, c+R)$ by a power series $\sum_{j=0}^{\infty} a_k(x-c)^j$. Then:

(a) For each x in the interval $(c-R, c+R)$ the series $\sum_{j=0}^{\infty} a_j(x-c)^j$ converges absolutely to $f(x)$. Furthermore, for each r such that $0 < r < R$ the series converges uniformly on the subinterval $[c-r, c+r]$ to f.

(b) The function f is differentiable on $(c-R, c+R)$, and $f' : (c-R, c+R) \to \mathbb{R}$ is represented on the interval by the power series $\sum_{j=0}^{\infty} j a_j (x-c)^{j-1}$.

(c) More generally, f is of class C^{∞} on $(c-R, c+R)$. In fact, if $k \in \mathbb{N}$ then $f^{(k)}$ is represented on the interval by the power series obtained by differentiating the original series $\sum_{j=0}^{\infty} a_j (x-c)^j$ term-by-term k times.

(d) Similarly, the k-th order antiderivative $D_c^{-k} f$ is defined on $(c-R, c+R)$, and can be obtained by repeatedly antidifferentiating the power series for f term-by-term.

(e) The power series $\sum_{j=0}^{\infty} a_j (x-c)^j$ which represents f on the interval $(c-R, c+R)$ is unique. In fact, it is the Taylor series of f about the center point c.

Proof

(a) This follows directly from Part (b) of Theorem (IX.6.2).

(b), (c) and (d) These follow directly from Theorem (IX.6.6).

(e) If $f(x) = \sum_{j=0}^{\infty} a_j (x-c)^j$ on $(c-R, c+R)$, then set $x = c$ to get $a_0 = f(c)$. More generally, for k in $\mathbb{N}$ compute $f^{(k)}$ by differentiating the series k times term-by-term to get (after using Part (c))

$$f^{(k)}(x) = k! a_k + (k+1)k \ldots 2 a_{k+1}(x-c) + (k+2)(k+1)k \ldots 3 a_{k+2}(x-c)^2 + \ldots$$

In particular when $x = c$ one gets $f^{(k)}(c) = k! a_k$, so that $a_k = f^{(k)}(c)/k!$. Thus the series $\sum_{j=0}^{\infty} a_j (x-c)^j$ has the same coefficients as the Taylor series of f about the center point c, and the desired result follows.

The class of functions represented by power series, about a given center, behaves nicely under the usual algebraic operations.

IX.7.8 Theorem

Let c be a real number, let R be a quantity such that $0 < R \leq +\infty$, and let $I = (c-R, c+R)$; note that if $R = +\infty$ then $I = \mathbb{R}$. Suppose that $f_1, f_2, \ldots f_m : I \to \mathbb{R}$ are functions which can be represented by power series on the interval I; that is, for each $k = 1, 2, \ldots m$ and each $j = 0, 1, 2, \ldots$, there are coefficients $a_j^{(k)}$ such that

$$f_k(x) = \sum_{j=0}^{\infty} a_j^{(k)} (x-c)^j \text{ for each } x \text{ in } I \qquad (*)$$

Let R_k denote the radius of convergence of the power series $\sum_{j=0}^{\infty} a_j^{(k)} (x-c)^j$, so that one has $R < R_k$ for each $k = 1, 2, \ldots m$. Then:

(a) Every function $F : I \to \mathbb{R}$ of the form $F = c_1 f_1 + c_2 f_2 + \ldots + c_m f_m$, where $c_1, c_2, \ldots c_m$ are in $\mathbb{R}$, can be represented by a power series on I. More precisely, one has

$$F(x) = \sum_{j=0}^{\infty} b_j (x-c)^j, \text{ for all } x \text{ in } I, \text{ where } b_j = \sum_{k=1}^{m} c_k a_j^{(k)} \text{ for each } j = 0, 1, 2, \ldots.$$

If R' denotes the radius of convergence of the series $\sum_{j=0}^{\infty} b_j (x-c)^j$, then $R' \geq \min\{R_1, \ldots R_m\}$.

(b) The product function $P : I \to \mathbb{R}$, given by the rule

$$P(x) = f_1(x)\cdot f_2(x)\cdot \ldots \cdot f_m(x) \text{ for all } x \text{ in } I,$$

can be represented by a power series in I. More precisely, define coefficients p_k, for $k = 0, 1, 2, \ldots,$ by the rule

$$p_k = \sum_{i_1+i_2+\ldots+i_m = k} a^{(1)}_{i_1}\cdot a^{(2)}_{i_2}\cdot \ldots \cdot a^{(m)}_{i_m} \qquad \text{(IX.14)}$$

If R'' denotes the radius of convergence of the series $\sum_{k=0}^{\infty} p_k(x-c)^k$, then $R'' \geq \min\{R_1, \ldots R_m\}$.

Proof

(a) The simple proof is left to the reader.

(b) Choose x such that $|x - c| < R$. Then x lies within the radius of convergence of each of the given power series. In particular, each quantity $M_k = \sum_{j=0}^{\infty} |a^{(k)}_j (x-c)^j|$ is finite.

Next, let X be the set of all m-tuples of the form $(j_1, j_2, \ldots j_m)$, with each j_k in $\mathbb{N} \cup \{0\}$ for each $k = 1, 2, \ldots m$. Choose a real number x such that $|x - c| < R$, and define a corresponding function $G : X \to \mathbb{R}$ by the following rule:

$$G((i_1, i_2, \ldots i_m)) = (a^{(1)}_{i_1} x^{i_1})\cdot(a^{(2)}_{i_2} x^{i_2})\cdot \ldots \cdot(a^{(m)}_{i_m} x^{i_m}) \qquad (**)$$

<u>Claim 1</u> The unordered sum $\sum_X |G|$ is convergent.

<u>Proof of Claim 1</u> Let W be a finite nonempty subset of X. Then there exists a positive integer k such that if $(i_1, i_2, \ldots i_m) \in W$, then $k \geq i_j$ for each $j = 1, 2, \ldots m$. Let W_k be the set of all elements $(i_1, i_2, \ldots i_m)$ of X such that $i_j \leq k$ for each index j. Thus, W_k is the m-fold Cartesian product of the set $\{0, 1, 2, \ldots k\}$ with itself. It is clear that W_k is a finite set (it has exactly $(k+1)^m$ elements), and that W is a subset of W_k. Since $|G|$ is a nonnegative function, it follows that $\sum_W |G| \leq \sum_{W_k} |G|$. However, it follows from the basic laws of finite sums that

$$\sum_{W_k} |G| = \sum_{0 \leq i_1, i_2, \ldots i_m \leq k} |a^{(1)}_{i_1}(x-c)^{i_1}|\cdot|a^{(2)}_{i_2}(x-c)^{i_2}|\cdot \ldots \cdot|a^{(m)}_{i_m}(x-c)^{i_m}|$$

The finite sum on the right can be formulated as 'iterated ordered sums' to yield

$$\sum_W |G| \leq \sum_{W_k} |G| = \sum_{i_1=0}^{k}\sum_{i_2=0}^{k} \cdots \sum_{i_m=0}^{k} |a^{(1)}_{i_1}(x-c)^{i_1}|\cdot|a^{(2)}_{i_2}(x-c)^{i_2}|\cdot \ldots \cdot|a^{(m)}_{i_m}(x-c)^{i_m}| =$$

$$\left(\sum_{i_1=0}^{k} |a^{(1)}_{i_1}(x-c)^{i_1}|\right)\cdot\left(\sum_{i_2=0}^{k} |a^{(2)}_{i_2}(x-c)^{i_2}|\right)\cdot \ldots \cdot\left(\sum_{i_m=0}^{k} |a^{(m)}_{i_m}(x-c)^{i_m}|\right)$$

It follows that there is a real constant M such that $\sum_W |G| \leq M$ for all finite subsets W of X. Indeed, let $M = M_1\cdot M_2\cdot \ldots \cdot M_m$, where the finite quantities M_j are defined above. In particular, $\sup U_{X;|G|} \leq M < +\infty$, so the unordered sum $\sum_X |G|$ is convergent, as claimed.

Note that it follows from Claim 1, combined with Theorem (IX.2.8), that the unordered sum $\sum_X G$ is also convergent.

<u>Claim 2</u> The unordered sum $\sum_X G$ converges to $P(x)$.

<u>Proof of Claim 2</u> This is done is by induction on the number m of factors in the product $P = f_1\cdot f_2\cdot \ldots \cdot f_m$.

Indeed, the desired result is trivially true if $m = 1$, since in this case the result reduces to saying that if f_1 can be expressed as a power series on I, then the same is true for $F = f_1$.

Now suppose that $m \geq 2$ and that the desired result is true for products involving fewer than m power series. Define a partition $\mathcal{F} = \{X_0, X_1, X_2, \dots\}$ of X as follows: for each $k = 0, 1, 2, \dots$, X_k is the set of all m-tuples $(j_1, j_2, \dots j_m)$ such that $j_m = k$. Indeed, this is the partition of X determined by the surjective function $\varphi(i_0, i_1, \dots i_m) = i_m$. It now follows from Theorem (IX.2.10), the Generalized Associative Law for Infinite Unordered Sums, that

$$\sum_X G = \sum_{\mathcal{F}} \hat{G},$$

where

$$\hat{G}(X_j) = \sum_{X_j} G.$$

However, each term in the latter sum has the factor $a_j^{(m)}(x-c)^j$. It is easy to see that one can then write

$$\sum_{X_j} G = \left(\sum_{0 \leq i_1, i_2, \dots i_{m-1}} a_{i_1}^{(1)}(x-c)^{i_1} a_{i_2}^{(2)}(x-c)^{i_2} \dots a_{i_{m-1}}^{(m-1)}(x-c)^{i_{m-1}} \right) (a_j^{(m)}(x-c)^j).$$

By the induction hypothesis, the sum which multiplies the factor $a_j^{(m)}(x-c)^j$ equals $f_1(x)\cdot \dots \cdot f_{m-1}(x)$, independent of j. Thus, since the chosen number x lies inside the radius of convergence of each of the functions f_1, $f_2, \dots$ f_m, one can write

$$\sum_X G = (f_1(x)\cdot \dots \cdot f_{m-1}(x)) \left(\sum_{j=0}^{\infty} a_j^{(m)}(x-c)^j \right) = f_1(x)\cdot \dots \cdot f_{m-1}(x)\cdot f_m(x).$$

The desired power series expansion for the product function P now follows easily. Indeed, let $\tilde{\varphi} : X \to \mathbb{N} \cup \{0\}$ be the surjective function given by the rule

$$\tilde{\varphi}(j_1, j_2, \dots j_m) = j_1 + j_2 + \dots + j_m,$$

and let $\tilde{\mathcal{F}}$ denote the corresponding partition of X. Thus, $\tilde{\mathcal{F}} = \{Y_0, Y_1, \dots\}$, where for each $k = 0, 2, \dots$ one has $Y_k = \{(j_1, j_2, \dots j_m) \in X : j_1 + j_2 + \dots + j_m = k\}$. It then follows from the Generalized Associative Law for Unordered Infinite Sums that

$$\sum_X G = \sum_{\tilde{\mathcal{F}}} \tilde{f} = \sum_{k=0}^{\infty} \tilde{f}(Y_k),$$

where

$$\tilde{f}(Y_k) = \sum_{Y_k} G = \sum_{j_1 + j_2 + \dots + j_m = k} \left((a_{j_1}^{(1)}(x-c)^{j_1})\cdot(a_{j_2}^{(2)}(x-c)^{j_2})\cdot \dots \cdot(a_{j_m}^{(m)}(x-c)^{j_m}) \right) =$$

$$\left(\sum_{j_1 + j_2 + \dots + j_m = k} a_{j_1}^{(1)}\cdot a_{j_2}^{(2)}\cdot \dots \cdot a_{j_m}^{(m)} \right) (x-c)^k = p_k(x-c)^k$$

Combining this with the results of Claim 2 one finally gets

$$P(x) = \sum_{k=0}^{\infty} p_k(x-c)^k,$$

where p_k is given by Equation (IX.14), as required.

IX.7.9 Remarks

(1) Many texts state the preceding theorem explicitly only in the case $m = 2$. In that case one normally writes f and g instead $f^{(1)}$ and $f^{(2)}$, and one writes a_j and b_j instead of the more complicated $a_j^{(1)}$ and $a_j^{(2)}$. In this simpler context Equation (IX.14) can be written in the simpler form

$$p_k = a_0 b_k + a_1 b_{k-1} + \ldots + a_{k-1} b_1 + a_k b_0. \tag{IX.15}$$

The expression $\sum_{k=0}^{\infty} p_k (x-c)^k$, in which p_k is given by Equation (IX.15), is called the **Cauchy Product of the series $\sum_{j=0}^{\infty} a_j (x-c)^j$ with the series $\sum_{j=0}^{\infty} b_j (x-c)^j$**. It then makes sense to refer to the series obtained using Equation (IX.14) above as the **Extended Cauchy Product of the m given power series**.

(2) It is clear that the radius of convergence of the power series $\sum_{k=0}^{\infty} p_k (x-c)^k$, obtained in the preceding theorem, is at least as large as the smallest of the radii $R_1, R_2, \ldots R_m$ of the original series. However, it is possible that the radius of convergence of this product series might be much larger than $\min\{R_1, R_2, \ldots R_m\}$. For instance, consider the product $P(x) = f(x) \cdot g(x)$, where

$$f(x) = \frac{1+x^4}{1+x^2} \text{ and } g(x) = \frac{1+x^2}{1+x^4} \text{ for all } x \text{ in } \mathbb{R}.$$

Each of these functions is defined for all x in $\mathbb{R}$, and the Taylor series about the center 0 of each has radius of convergence $R = 1$. Their product, $P(x) = f(x)g(x)$, equals the constant 1 for all x in $\mathbb{R}$, and thus its Taylor series about 0 has radius of convergence $+\infty$.

Before discussing several more properties enjoyed by functions which can be represented by power series, it is useful to introduce some terminology.

IX.7.10 Definition (Real-Analytic Functions)

Let $f : I \to \mathbb{R}$ be a real-valued function defined on an open set U, and let c be a point in U. One says that f is **real analytic at** c provided that f can be represented on some subinterval $(c-R, c+R)$ of U, with $R > 0$, by a power series $\sum_{j=0}^{\infty} a_k (x-c)^k$. The function f is said to be **real analytic on** U provided it is real analytic at each point of U.

IX.7.11 Example

It is an instructive exercise to prove that if $f : \mathbb{R}\backslash\{0\} \to \mathbb{R}$ is the reciprocal function, given by the rule $f(x) = 1/x$, then f is analytic on its domain.

IX.7.12 Remarks

(1) It is understood in the preceding definition that the radius of convergence R' of the power series $\sum_{j=0}^{\infty} a_j (x-c)^j$ is at least as large as the positive number R, but that we allow the possibility that $R' > R$. In particular, for each x in the interval $(c-R, c+R)$ the power series converges to the corresponding value $f(x)$ of f.

(2) Some texts give the following alternate, but equivalent, definition: The function f is said to be real analytic at c provided that f is C^{∞} at c and the Taylor series $\displaystyle\sum_{j=0}^{\infty} \frac{f^{(j)}(c)}{j!}(x-c)^j$ converges to $f(x)$ in some nonempty open interval $(c-R, c+R)$ about c.

(3) Consider the polynomial $f(x) = -3 + 5(x-1) + 2(x-1)^2$.

Question Is this function real analytic on $\mathbb{R}$? (Think about your response before reading further.)

A natural response to this question is to say

'Obviously it *is* real analytic. In fact, it is already expressed as a power series about $c = 1$, with coefficients $a_0 = -3$, $a_1 = 5$, $a_2 = 2$, and $a_k = 0$ if $k \geq 3$. Clearly this series has infinite radius of convergence.'

However, this response misses a subtle feature of the definition of 'real analytic'. Namely, the definition requires that the function f be expressible as a convergent power series about *each* point c of the interval I. The polynomial above is expressed as a convergent power series about $c = 1$, but what about other values of c?

In this example the solution is simple: use a technique which we employed, in Section (**??**), when discussing ways of calculating Taylor polynomials without computing derivatives. Namely, write $x-1$ in the form $(c-1)+(x-c)$ and expand the powers of $(x-1)$ in terms of powers of $x-c$, using the standard Binomial Theorem:

$$x-1 = (c-1)+(x-c),\ (x-1)^2 = ((c-1)+(x-c))^2 = (c-1)^2 + 2(c-1)(x-c) + (x-c)^2.$$

Substitute these expressions into the original formula for f and do some obvious simplifications to get

$$f(x) = b_0 + b_1(x-c) + b_2(x-c)^2, \text{ where } b_0 = 2c^2 + c - 6,\ b_1 = 4c+1, \text{ and } b_2 = 2.$$

The final remark above illustrates a more general question: if a function can be represented by a convergent power series about a particular point c, is it real analytic? The next theorem answers this question by using the 're-expand about other points' method illustrated in that remark.

IX.7.13 Theorem

Suppose that $f : (c-R, c+R) \rightarrow \mathbb{R}$ is a real-valued function which can be represented as a convergent power series $\sum_{j=0}^{\infty} a_j(x-c)^j$ on a nonempty open interval $I = (c-R, c+R)$. Then f is real analytic on I. More precisely, if p is any point in I then there is a power series $\sum_{j=0}^{\infty} b_j(x-p)^j$ which represents f on the open interval $(b-R', b+R')$, where $R' = R - |c-p|$. (If $R = +\infty$ this equation should be interpreted to mean $R' = +\infty$. In either case, it is clear thaT $R' > 0$)

Proof Let x be any point of the interval I. For each index $j \geq 1$ express the quantity $(x-c)^j$ as a polynomial in $x-p$:

$$(x-c)^j = ((p-c)+(x-p))^j = \sum_{l=0}^{j} C(j,l)(p-c)^l(x-p)^{j-l}$$

where $C(j,l)$ denotes the **binary coefficient** $\dfrac{j!}{l!(j-l)!}$. Note that one has

$$|x-c|^j \leq \sum_{l=0}^{j} C(j,l)|p-c|^l|x-p|^{j-l}$$

Now let X be the set of all ordered pairs of the form (j, l) with j, l nonnegative integers such that $0 \leq l \leq j$. Define $F : X \to \mathbb{R}$ by the rule $F(j,l) = a_j C(j,l)(p-c)^l(x-p)^{j-l}$.

Claim If $|x-p| < R - |p-c|$, then the unordered sum $\sum_X |F|$ is convergent.

Proof of Claim For convenience, set $r = |x-p| + |p-c|$, so that $0 \leq r < R$. Let W be a nonempty finite subset of X, and let k be the largest first entry of any pair (j,l) in W. Then it is clear that W is a subset of the finite set $W_k = \{(j,l) \in X : 0 \leq l \leq j \leq k\}$. Thus one has

$$\sum_W |F| \leq \sum_{W_k} |F| = \sum_{j=0}^{k}\sum_{l=0}^{j} |a_j| C(j,l) |p-c|^l |x-p|^{j-l} = \sum_{j=0}^{k} |a_j| \left(|p-c| + |x-p|\right)^j = \sum_{j=0}^{k} |a_j| r^j \leq \sum_{j=0}^{\infty} |a_j| r^j.$$

The last sum, which is independent of k, is finite because $r < R$ and the original series $\sum_{j=0}^{\infty} a_k (x-c)^j$ is absolutely convergent inside the radius of convergence. In other words, $\sum_W |F| \leq \sum_{j=0}^{\infty} |a_j| r^j < +\infty$ for every finite subset W of X. Thus $\sum_X |F|$ is convergent, as claimed, and it follows as usual that the unordered sum $\sum_X F$ is also convergent.

Finally, let $\varphi : X \to \{0, 1, 2, \ldots\}$ be the surjective function given by $\varphi(j,l) = j - l$, and let $\mathcal{F} = \{Z_0, Z_1, \ldots\}$ be the corresponding partition of X. Thus, Z_k consists of all the pairs (j,l) for which $j - l = k$. Apply the Generalized Associative Law for Unordered Infinite Sums with this partition to get

$$\sum_X F = \sum_{k=0}^{\infty} \hat{F}(Z_k).$$

However, $\sum\limits_{Z_k} F = b_k (x-p)^k$ where $b_k = \sum_{l=0}^{k} a_{k+l} C(k+l, l)(p-c)^l$. Thus

$$\sum_X F = \sum_{k=0}^{\infty} b_k (x-p)^k.$$

However, if one lets $\psi : X \to \{0, 1, 2 \ldots\}$ be the surjection given by $\psi(j,l) = j$, one gets the partition $\tilde{\mathcal{F}} = \{\tilde{Z}_0, \tilde{Z}_1, \ldots\}$, where $\tilde{Z}_j = \{(j,l) : 0 \leq l \leq j\}$. One easily computes that

$$\sum_{\tilde{Z}_j} F = \sum_{l=0}^{j} a_j C(j,l)(p-c)^l (x-p)^{j-l} = a_j \sum_{l=0}^{j} C(j,l)(p-c)^l (x-p)^{j-l} = a_j((p-c)+(x-p))^j = a_j (x-c)^j.$$

Thus one can use the Generalized Associative Law again, but this time with the partition $\tilde{\mathcal{F}}$, to show that

$$\sum_X F = \sum_{j=0}^{\infty} a_j (x-c)^j = f(x).$$

Combine these result to get $f(x) = \sum_{k=0}^{\infty} b_k (x-p)^k$ if $|x-p| < R'$.

IX.7.14 Theorem

Suppose that $f : U \to \mathbb{R}$ and $g : V \to \mathbb{R}$ are real analytic functions on open sets U and V, respectively. Suppose further that $f[U] \subseteq V$. Then the composition $h = g \circ f : U \to \mathbb{R}$ is real analytic on U.

Outline of Proof Let c be a point of U and set $d = f(c)$. The 'analyticity' hypothesis on f guarantees that there is a power series $\sum_{j=0}^{\infty} a_j (x-c)^j$, and a positive quantity R, such that $f(x) = \sum_{j=0}^{\infty} a_j(x-c)^j$ for all x such that $|x-c| < R$. Likewise, there is a power series $\sum_{k=0}^{\infty} b_k(y-d)^k$, and a positive number R', so that $g(y) = \sum_{k=0}^{\infty} b_k(y-d)^k$ for all y such that $|y-d| < R'$. Without loss of generality we may assume that if $|x-c| < R$ then $f(x) - d < R'$: simply replace R by a smaller positive quantity if needed. Note that for $|x-c| < R$ one has

$$h(x) = g(f(x)) = \sum_{k=0}^{\infty} b_k(f(x))^k = \sum_{k=0}^{\infty} b_k \left(a_0 + a_1(x-c) + a_2(x-c)^2 + \cdots\right)^k$$

In accordance with Theorem (IX.7.8), one can express the the quantity $\left(a_0 + a_1(x-c) + a_2(x-c)^2 + \cdots\right)^k$ as a power series $a_0^{(k)} + a_1^{(k)}(x-c) + \ldots + a_j^{(k)}(x-c)^j + \ldots$, where the coefficients $a_j^{(k)}$ are given by Equation (IX.14). Now let X be the set of all pairs (k,j) with k and j nonegative integers, and fix x such that $|x-c| < R$. Define $F : X \to \mathbb{R}$ by the rule $F(k,j) = b_k a_j^{(k)}(x-c)^j$. The desired result then follows by a 'Generalized Associative Law' argument similar to the ones used above. The details are left as an exercise.

IX.7.15 Corollary

Suppose that $f : U \to \mathbb{R}$ is real analytic on an open subset U of $\mathbb{R}$, and assume that for all x in U one has $f(x) \neq 0$. Then the function $1/f$ is also real analytic on U.

Proof Apply the preceding theorem with $g(y) = 1/y$ for all $y \neq 0$; see Example (IX.7.11).

IX.8 EXERCISES FOR CHAPTER IX

IX - 1 Prove Part (a) of Corollary G.1.6 by using the definition of unordered sums given in the Notes. (The proof of this result given in the Notes is based on the 'epsilon' approach to unordered sums that is outlined in Statement (ii) in Theorem G.1.4. Thus, this exercise asks you to prove the same result, but based directly on the '$f = f^+ - f^-$' approach to ordered sums.)

IX - 2 Prove Corollary G.1.8.

IX - 3 It is mentioned in Remark G.1.5 on Page 362 that many authors use the 'ε' characterization of 'unordered sum' as their defnition of this concept instead of our '$f = f^+ - f^-$' approach. One minor difficulty in doing so is that one must then show that the value one assigns to $\sum_X f$ is unique, an issue which the '$f = f^+ - f^-$' approach avoids.

Problem Using the 'ε' approach to unordered sums, and without using the results of Theorem G.1.4, prove that the value C which that approach assigns to a convergent unordered sum $\sum_X f$ is unique.

IX - 4 Suppose that f and g are real-valued functions defined on an infinite set X, and suppose that $|f(x)| \leq |g(x)|$ for all x in X. Prove that if the unordered sum $\sum_X g$ is defined then so is $\sum_X f$.

Note This result is called the **Comparison Test for Unordered Sums**

IX - 5 Suppose that f and g are real-valued functions defined on $\mathbb{N}$; assume that the unordered sums $\sum_{\mathbb{N}} f$ and $\sum_{\mathbb{N}} g$ are convergent, with values A and B, respectively. Define $P : \mathbb{N}\times\mathbb{N} \to \mathbb{R}$ by the rule $P(i,j) = f(i)\cdot g(j)$ for each (i,j) in $\mathbb{N}\times\mathbb{N}$.

(a) Show that the unordered sum $\sum_{\mathbb{N}\times\mathbb{N}} P$ is also convergent, and that $\displaystyle\sum_{\mathbb{N}\times\mathbb{N}} P = A\cdot B$.

(b) Define a third function $h : \mathbb{N} \to \mathbb{R}$ by the rule

$$h(k) = f(1)\cdot g(k) + f(2)\cdot g(k-1) + \ldots + f(k)\cdot g(1) \text{ for each } k \text{ in } \mathbb{N}.$$

(Thus, $h(1) = f(1)g(1)$, $h(2) = f(1)g(2) + f(2)g(1)$, $h(3) = f(1)g(3) + f(2)g(2) + f(3)g(1)$, and so on.)

Show that the unordered sum $\sum_{\mathbb{N}} h$ is convergent and that $\sum_{\mathbb{N}} h = A\cdot B$.

IX - 6 In this exercise we generalize the 'Middle-Thirds' characterization of the Cantor Ternary Set.

Let r be a number such that $0 < r < 1$. If I is any closed interval of length L, the 'middle r-portion of I' is the *open* subinterval J of I which is symmetric about the midpoint of I and whose length is rL. This concept leads to the following construction:

Step 1 Remove the middle r-portion of the unit interval $[0,1]$. What remains is the disjoint union of two closed subintervals of $[0,1]$, each of length $(1-r)/2$.

Step 2 Remove the middle r-portion of each of the subintervals obtained in Step 1, obtaining 4 disjoint closed subintervals of $[0,1]$, all of equal length.

General Step Continue this process of removing middle r-portions.

Let C_r be the set of points of $[0,1]$ which do *not* get removed during this process.

Problem

(a) Show that the set C_r is closed in $\mathbb{R}$, and that this set has no nonempty open subsets.

(b) Show that C_r is an uncountable set.

(c) Show that C_r has measure 0.

IX - 6 In this exercise we modify the construction described in the preceding exercise.

Step 1 Remove the middle $1/2$-portion of the unit interval $[0,1]$. What remains is the disjoint union of two closed subintervals of $[0,1]$, each of length $1/4$.

Step 2 Remove the middle $1/2^2$-portion of each of the subintervals obtained in Step 1, obtaining 4 disjoint closed subintervals of $[0,1]$, all of equal length.

General Step Continue this process of removing middle $1/2^k$-portion, increasing k by 1 at each stage.

Let $\hat{C}$ be the set of points of $[0,1]$ which do *not* get removed during this process.

Problem

(a) Show that the set $\hat{C}$ is closed in $\mathbb{R}$, and that this set has no nonempty open subsets.

(b) Show that $\hat{C}$ is an uncountable set.

(c) Show that the set $\hat{C}$ has positive measure.

Remark Because $\hat{C}$ has these properties, it is sometimes called a **fat Cantor set**.

IX - 8 Prove or Disprove If $\sum_{k=1}^{\infty} x_k$ is a convergent infinite series and $\alpha = (a_1, a_2, \dots a_k, \dots)$ is a sequence of positive numbers which converges to 0, then the series $\sum_{k=1}^{\infty} a_k x_k$ is also convergent.

IX - 9 Suppose that $\alpha = (a_1, a_2, \dots)$ is a sequence as in the statement of the Alternating-Series Test. That is, $a_k > 0$ for each index k, the sequence α is monotonic down, and $\lim_{k\to\infty} a_k = 0$. For each k in $\mathbb{N}$ let s_k denote the k-th partial sum of the alternating series $\sum_{k=1}^{\infty}(-1)^{k-1}a_k$, and let L be the sum of that series.

(a) Prove that if $\varepsilon > 0$ and if $a_{k+1} \leq \varepsilon$ then $|L - s_k| \leq \varepsilon$.

(b) Suppose that, in addition, the sequence α has the property that $a_k - a_{k+1} \geq a_{k+1} - a_{k+2}$ for each k in $\mathbb{N}$; that is, the *differences* of consecutive terms form a monotonic-down sequence. Prove that if $\varepsilon > 0$ and $a_k \leq 2\varepsilon$ then $|L - s_k| \leq \varepsilon$.

(c) Use the results of Parts (a) and (b) to determine how many terms of the Alternating Harmonic Series one should use to estimate the value of $\ln 2$ with error less than $1/1000$ in magnitude. Compare the results.

IX - 10 Let $\xi = (x_1, x_2, \dots x_k, \dots)$ be given by the rule

$$x_k = \begin{cases} \dfrac{1.0000001}{k} & \text{if } k \text{ is odd} \\[2ex] -\dfrac{0.9999999}{k} & \text{if } k \text{ is even} \end{cases}$$

(a) Show that the series $\sum_{k=1}^{\infty} x_k$ is divergent.

(b) Explain why this does not contradict the Alternating Series Test. (Or does it ???)

IX - 11 Suppose that $\xi = (x_1, x_2, \dots)$ is a monotonically decreasing sequence of nonnegative numbers. Prove that the series $\sum_{j=1}^{\infty} x_j$ is convergent if, and only if, the series $\sum_{j=0}^{\infty} 2^j x_{2^j}$ is convergent.

IX - 12 Prove or Disprove If the series $\sum_{k=1}^{\infty} x_k$ is divergent, then so is the series $\sum_{k=1}^{\infty} kx_k$.

IX - 13 (a) Show that if $n \geq 2$ then $\frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} < \ln(n) < 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1}$.

(b) Show that the sequence $(a_n)_{n \geq 1}$, defined by the rule $a_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} - \ln n$, is bounded and strictly decreasing.

Remark: It follows from the preceding that the sequence $\alpha = (a_1, a_2, \ldots)$ given here is convergent. The limit of this sequence is an important number called *Euler's Constant*; its standard symbol is γ. The value of γ, to 10 decimal places, is 0.5772156649.

(c) For each $k \in \mathbb{N}$ let $b_k = 1/2 + 1/4 + 1/6 + \ldots 1/(2k)$ denote the sum of the reciprocals of the first k *even* positive integers; likewise, let $c_k = 1 + 1/3 + 1/5 + \ldots + 1/(2k-1)$ denote the sum of the reciprocals of the first k *odd* positive integers;

Problem: Use results of Parts (a) and (b) to show that

$$\lim_{k \to \infty} \left(b_k - \frac{\ln(k)}{2} \right) = \frac{\gamma}{2}, \text{ and } \lim_{k \to \infty} \left(c_k - \frac{\ln(k)}{2} \right) = \frac{\gamma}{2} + \ln(2)$$

(d) Use the results of Part (c) to give an alternate proof that the Alternating Harmonic Series converges to $\ln(2)$.

IX - 14 Suppose that $\sum_{k=1}^{\infty} x_k$ is an infinite series such that $\sum_{k=1}^{\infty} x_k^+$ and $\sum_{k=1}^{\infty} x_k^-$ both diverge to $+\infty$, and such that $\lim_{k \to \infty} x_k = 0$. Prove that some rearrangement of $\sum_{k=1}^{\infty} x_k$ is conditionally convergent.

IX - 15 Prove that if $f : (a,b) \to \mathbb{R}$ is real-analytic on an open interval (a,b), and if $f(x) = 0$ for all x in some subinterval of (a,b), then $f(x) = 0$ for all x in (a,b).

IX - 16 Prove that the exponential function satisfies the usual 'Law of Exponents', namely $e^{x+y} = e^x \cdot e^y$, by using the product of the Maclaurin series for for e^x and e^y.

IX - 17 Suppose that $\alpha = (a_0, a_1, \ldots)$ is a sequence such that $a_0 = 1$. Assume that r is a positive number such that $\sum_{k=1}^{\infty} |a_k| r^k < 1$. Define a second sequence $\beta = (b_0, b_1, b_2, \ldots)$ recursively by the rule

$$b_0 = 1; \; b_k = -(a_k + a_{k-1}b_1 + \ldots + a_1 b_{k-1}) \text{ if } k \geq 1$$

Problem Show that the radius of convergence of the power series $\sum_{j=0}^{\infty} b_j x^j$ is at least as large as r.

IX - 18 (a) Determine what the Root Test tells us about the convergence or divergence of the following series:

$$\frac{1}{2} + 0 + 0 + \frac{1}{2^2} + 0 + 0 + 0 + 0 + \frac{1}{2^3} + 0 + 0 + 0 + 0 + 0 + 0 + \frac{1}{2^4} + \ldots;$$

the general pattern is that between the consecutive *non*zero terms $1/2^k$ and $1/2^{k+1}$ there are $2k$ zeroes.

(b) Determine what the Ratio Test tells us about the convergence or divergence of the following series of positive terms:

$$\frac{1}{2^2} + \frac{1}{2} + \frac{1}{2^4} + \frac{1}{2^3} + \frac{1}{2^6} + \frac{1}{2^5} + \ldots;$$

the general pattern is that if $k = 2m-1$, then the k-th term is $1/2^{2m}$, while if $k = 2m$, then the k-th term is $1/2^{2m-1}$.

(c) Determine the sums of the series discussed in Parts (a) and (b).

IX - 19 Dirichlet's Test for Uniform Convergence Consider an infinite series $\sum_{j=1}^{\infty} f_j$ of real-valued functions $f_1,\, f_2,\, \dots$ defined on a nonempty set $X \subseteq \mathbb{R}$, and for each k let $s_k : X \to \mathbb{R}$ denote the corresponding k-th partial sum; that is, $s_k = f_1 + f_2 + \dots + f_k$. Suppose that the sequence $(s_1, s_2, \dots)$ is uniformly bounded on X; that is, there is a positive real number M such that for all x in X and all k in $\mathbb{N}$ one has $|s_k(x)| \leq M$. Let $\gamma = (g_1, g_2, \dots)$ be a sequence of real-valued functions on X such that $g_{k+1}(x) \leq g_k(x)$ for all x in X. Assume further that the sequence γ converges uniformly on X to the zero function. Then the series $\sum_{j=1}^{\infty} f_j g_j$ converges uniformly on X.

IX - 20 Abel's Test for Uniform Convergence Consider an infinite series $\sum_{j=1}^{\infty} f_j$ of real-valued functions $f_1,\, f_2,\, \dots$ defined on a nonempty set $X \subseteq \mathbb{R}$; suppose that this series converges uniformly on X. Let $\gamma = (g_1, g_2, \dots)$ be a sequence of real-valued functions on X such that $g_{k+1}(x) \leq g_k(x)$ for all x in X. Assume further that the sequence γ is uniformly bounded on X; that is, there is a positive real number M such that for all x in X and all k in $\mathbb{N}$ one has $|g_k(x)| \leq M$.

Then the series $\sum_{j=1}^{\infty} f_j g_j$ converges uniformly on X.

IX - 21 Prove that if $\sum_{k=1}^{\infty} a_k$ is an absolutely convergent series then the series $\sum_{k=1}^{\infty} a_k \sin(kx)$ is uniformly convergent on $\mathbb{R}$.

IX - 22 Prove that the series $\displaystyle\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{\sqrt{k}} \sin\left(1 + \frac{x}{k}\right)$ converges uniformly on every closed bounded interval of $\mathbb{R}$.

IX - 23 Let $\alpha = (a_1, a_2, \dots)$ be a monotonic down sequence of positive numbers. Prove that if the series $\sum_{k=1}^{\infty} a_k \sin(kx)$ converges uniformly on $\mathbb{R}$, then $\lim_{k \to \infty} k a_k = 0$. The converse is also true, but is definitely ha

IX - 24 Recall that the Cantor set can be constructed by removing certain open intervals – the 'middle thirds' – from the unit interval $[0,1]$. Let X be the set of those open intervals, and define a function $f : X \to \mathbb{R}$ by the rule that if $I = (a,b)$ is in X then $f(I) = b - a$.

Problem Compute the unordered sum $\sum_X f$.

Appendix A

The Dedekind-Peano Axioms for ℕ

Quotes for Appendix A:

> (1) ‘Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk.’
> (‘The dear God has made the integers; all the rest is man’s work.’)
> Statement quoted of Leopold Kronecker.

In this appendix we present an axiomatic approach to the concept of ‘counting’. The usual custom is to refer to the axioms discussed here as the ‘Peano Axioms’. However, since they apparently were discovered earlier (and independently) by Dedekind, in *This Textbook* we follow the usage of several authors and refer to them as the *Dedekind-Peano axioms*. The approach in this Appendix largely follows that of Dedekind in his famous essay *Was sind und was sollen die Zahlen*; of course, Dedekind’s terminology and notation have been upgraded to reflect the more modern usage.

Note If you have never worked through the ‘Dedekind-Peano’ axiomatic approach to the natural numbers, it is worth your time to at least skim through the material in this appendix; however, doing so is not required for reading the preceding chapters.

In Definition (I.7.1) we characterize what it means for a finite nonempty set A to have (exactly) k elements, for some $k \in \mathbb{N}$, in terms of a complete pairing of A with the subset $\mathbb{N}_k$ of the ‘standard counting set $\mathbb{N}$; that is, the set of standard counting numbers. Of course other sets could have been chosen to play the role of ‘standard counting set’; for example, in many computer spreadsheets the columns are labeled by letters, not natural numbers:

$$A, B, \ldots Z, A\,A, A\,B, \ldots A\,Z, B\,A, B\,B \ldots$$

In this appendix we characterize axiomatically the properties which any such comparison set ought to enjoy.

Add1-(I) 1: Definition A **counting structure** is a pair of objects (X, σ) consisting of a nonempty set X together with a function $\sigma : X \to X$ which satisfies the following **Dedekind-Peano Axioms**:

(a) The function σ is one-to-one on X, and there is exactly one element of X which is *not* in the image $\sigma(X)$. This element is denoted u_σ; the letter u stands for ‘unit’.

(b) Suppose that A is any subset of X with the following properties:

(i) The element u_σ is in the set A;

(ii) For every element x in A the element $\sigma(x)$ is also in A; that is, $\sigma[A] \subseteq A$.

Then $A = X$.

Add1-(I) 2: Remarks

(1) Concerning Axiom (a) The function σ associated with the counting structure (X, σ) is usually called the **successor function** of that structure, and if x is any element of X then $\sigma(x)$ is called the **successor of** x. The requirement that σ be one-to-one then can be phrased as 'No element of X can be the successor of more than one element'. Likewise, the requirement on the special element u_σ can be phrased as 'The special element u_σ is not the successor of any element, and it is the *unique* element of X which is not a successor'. Because of this last fact, one often refers to u_σ as the **initial element of** X. When there can be no confusion, one normally writes the simpler u instead of u_σ.

Note that most authors assume the existence, but not the uniqueness, of an element u in X which is not the successor of any element. They can do this because the uniqueness can be trivially deduced later on using Axiom (b). The choice, whether to treat the 'uniqueness' as part of an axiom or as a theorem to be proved later on, is thus largely a matter of taste.

(2) Concerning Axiom (b) For obvious reasons, this axiom is called the **Induction Axiom**, and any nonempty subset of X which satisfies Part (ii) of this condition is called an **inductive subset of** X. Peano uses essentially this axiom in his formulation; that is, he *assumes* as an axiom the 'Principle of Mathematical Induction'.

In contrast, the analogous axiom in Dedekind's formulation can be phrased as follows:

'The set X is the intersection of the family of all the inductive subsets of X containing the element u.'

(Note that X itself is an inductive subset of X containing u, so the family in question is not empty.) In particular, Dedekind does *not* assume the Principle of Mathematical Induction as an axiom; instead he *proves* that Principle as a consequence of his axioms. Likewise, one can prove that the Peano axiom implies the Dedekind version. (Both proofs are trivial.) In any event, one needs both the 'Principle of Mathematical Induction' and Dedekind's 'intersection of inductive subsets' construction to work out the theory; the choice, of which is to be an axiom and which is to be a theorem, is thus largely a matter of taste.

Add1-(I) 3: Examples

(1) Let X be the standard set $\mathbb{N} = \{1, 2, 3, \ldots\}$, whose elements are described by Arabic (decimal) numerals. Let the function $\sigma : \mathbb{N} \to \mathbb{N}$ be given by the usual rule for finding the successor of a natural number k: if the right-most (decimal) numeral of a number k is one of the digits 0, 1,… 8, then replace that digit by the next higher one, and leave the other digits alone. If the decimal expression for k ends with a string of one or more '9's on the right, to get $\sigma(k+1)$ replace each such 9 with the digit 0, and increase the right-most non-9 digit to the next higher digit; if *all* the numerals of k are 9, then $\sigma(k+1)$ has initial numeral 1 followed by as many 0s as k has 9s. The initial element for this counting structure is 1.

Note A faster way of describing the successor function in this example would be to simply say $\sigma(k) = k + 1$. The reason for using the wordier description given above is to emphasize that one does not have to know about 'addition' – not even the special case of 'addition with 1' – to characterize the successor function. Because of this, when we later *define* addition for arbitrary counting structures in terms of the successor function, we can avoid the sensation that the definition is somehow 'circular' – we define addition in terms of the successor function, which itself is often defined in terms of 'addition with 1'. Having done so in this key example, however, we take the easy way out in some of the following examples and describe the successor function there in terms of addition.

(2) The set X is $\hat{\mathbb{N}} = \{0, 1, 2, \dots\}$, the function $\sigma : \hat{\mathbb{N}} \to \hat{\mathbb{N}}$ is given by $\sigma(k) = k + 1$. Then $u = 0$.

(3) The set X consists of all elements of $\mathbb{N}$ starting with 7: $X = \{7, 8, 9, \dots\}$. The function σ is given as usual by $\sigma(k) = k + 1$. In this case the initial element is $u = 7$.

(4) The set X consists of all the *even* natural numbers; that is, $X = \{2, 4, 6, \dots\}$. The function σ is given by the rule $\sigma(k) = k + 2$, so that $u = 2$.

(5) The set X consists of the standard Roman numerals

$$I, II, III, IV, V, \dots X, \dots C, \dots M, \dots$$

Since we shall not be using these numerals extensively, we shall leave to the reader the happy task of determining the corresponding successor function σ.

The reader should be able to easily figure out the pattern and from that to determine the corresponding 'successsor function' σ.

Add1-(I) 4: Theorem Suppose that (X, σ) is a counting structure with initial element u. Then for all x in X one has $\sigma(x) \neq x$.

Proof Let A be the set of all x in X such that $\sigma(x) \neq x$. Certainly the initial element u is in A; for if not then one would have $\sigma(u) = u$, contrary to the fact that the initial element is not in the image of the function σ.

Next, suppose that $x \in A$. If $\sigma(x)$ were not in A, then one would have $\sigma(\sigma(x)) = \sigma(x)$. But by the fact that σ is one-to-one, one would then also have $\sigma(x) = x$, contrary to the induction hypothesis that $x \in A$. Thus, if $x \in A$ then $\sigma(x) \in A$ as well. By the Induction Axiom it follows that $A = X$, and the desired result follows.

Example (3) above illustrates a way of obtaining new counting structures from a given such structure.

Add1-(I) 5: Theorem Suppose that (X, σ) is a counting structure with initial element u. Let x_1 be an element of X, and let X_1 denote the intersection of the family $\mathcal{F}_1$ of all the inductive subsets of X which contain the element x_1. Then X_1 is a nonempty subset of X. Furthermore, if σ_1 denotes the restriction of the function σ to the subset X_1, then (X_1, σ_1) is a counting structure, and its initial element is x_1.

Proof Let A be the set of all x_1 in X for which the conclusion of the theorem is true. It suffices to show that $A = X$.

Initial Step Clearly $u \in A$, since by the Induction Axiom the only inductive subset of X containing u is X itself, and thus when $x_1 = u$ one has $X_1 = X$ and $\sigma_1 = \sigma$.

Inductive Step Now suppose that $w \in A$, and let W denote the intersection of all inductive subsets of X containing w, and let σ_w denote the restriction of σ to W. Then, by the definition of the set A, the pair (W, σ_w) is a counting structure with initial element w. Let $x_1 = \sigma(w)$, and let $\mathcal{F}_1$, X_1 and σ_1 be as in the statement of the theorem. Note that the family $\mathcal{F}_1$ is nonempty, since X itself is an inductive set containing x_1. In adddition, by definition of $\mathcal{F}_1$, every set in the family $\mathcal{F}_1$ contains x_1, hence so does the intersection X_1. Thus, $x_1 \in X_1$, so in particular $X_1 \neq \emptyset$. Next, suppose that $y \in X_1$. Then for every set Y in the family $\mathcal{F}_1$ one has $y \in Y$; and since each such set Y is an inductive subset of X, it follows that $\sigma(y) \in Y$ as well. Thus, $\sigma(y)$ is also in the intersection X_1, hence X_1 is an inductive set. In particular, the restriction σ_1 of σ to X_1 maps X_1 into itself. It is clear that σ_1 is one-to-one on X_1, since it is the resriction to X_1 of the one-to-one function σ on X. Furthermore, if z is a point of X_1 such that $z \neq x_1$, then z must be of the form $\sigma_1(x)$ for some x in X_1. Indeed, if this were not the case, consider the set $Y = X_1\backslash\{z\}$. Since $z \neq x_1$ it is clear that $x_1{\in}Y$. And since z is not in $\sigma_1(X_1)$, and X_1 is an inductive subset of X, it is clear that Y is a nonempty inductive subset of X containing x_1. Thus, X_1, being the intersection of all such sets, must be a subset of Y. However, this is impossible since $z{\in}X_1$ but z is *not* in Y.

All that is left to show is that x_1 is not in $\sigma_1(X_1)$. Since, by definition, $x_1 = \sigma(w)$, and σ is one-to-one on X, this means one need only show that w is not in X_1. But if w were in X_1, then X_1 would be an inductive subset of X containing w, hence one would have $W \subseteq X_1$, since W is the intersection of all such subsets. On the other hand, w is in W, and W is an inductive subset of X, so $x_1 = \sigma(w)$ is also in W, so W is an inductive subset of X containing x_1. Since X_1 is the intersection of all such subsets of X, it follows that $X_1 \subseteq W$. Combining these results, one sees that if $w{\in}X_1$, then $X_1 = W$. However, since $w{\in}A$ it follows that w is the initial element of W, hence if one removes w from $W = X_1$ one would get a proper subset of X_1 which contains $x_1 = \sigma(w)$ and is inductive. (Note that $\sigma(w) \neq w$ because w is the initial element of W; that is, $x_1 \neq w$. Thus, removing w from W does not remove x_1 from W.) This would contradict the definition of X_1 as the intersection of all such subsets of X.

Add1-(I) 6: Definition Let (X, σ) be a counting structure with initial element u. Let x_1 be an element of X. Then the counting structure (X_1, σ_1) with initial element x_1 described in the preceding theorem is called the **counting substructure of** (X, σ) **determined by** x_1. The set X_1 so described is denoted $< x_1 >_\sigma$, and the correspondng function σ_1 is denoted σ_{x_1}.

By a similar line of reasoning one can prove the following result; the details are left as an exercise.

Add1-(I) 7: Corollary Let (X, σ) be a counting structure with initial element u. Let x_1 be an element of X, and let (X_1, σ_1) be the counting substructure of (X, σ) determined by x_1. Then the counting substructure of (X, σ) determined by $x_2 = \sigma(x_1)$ is of the form (X_2, σ_2), where $X_2 = \sigma(X_1) = X_1\backslash\{x_1\}$.

Add1-(I) 8: Theorem Let (X, σ) be a counting structure with initial element u. Let x_1 and x_2 be elements of X such that $x_1 \neq x_2$, and let (X_1, σ_1) and (X_2, σ_2) be the corresponding counting substructures of (X, σ). Then exactly one of the following statements is true:

(i) x_1 is in X_2

(ii) x_2 is in X_1

Proof To see that *at most* one of these properties can hold for such x_1 and x_2, suppose that x_1 is in X_2. Then X_2 is an inductive subset of X which contain x_1. Since $x_2 \neq x_1$, and (by the preceding theorem) x_2 is the initial element of the set X_2, it follows that the set $Y = X_2\backslash\{x_2\}$ obtained by removing x_2 from X_2 is also an inductive subset of X which contains x_1. Since X_1 is, by definition, the intersection of such subsets, it follows that $X_1 \subseteq Y$. But Y does not contain x_2, hence neither does X_1, as claimed. A similar argument shows that if x_2 is in X_1 then x_1 cannot be an element of X_2.

To see that *at least* one of these properties must hold, let A be the set of all x_1 in X such that if x_2 is an element of X not equal to x_1, then either (i) or (ii) holds.

Initial Step Certainly $x_1 = u$ is in A, since in this case $X_1 = X$ and thus (ii) holds for every x_2.

Induction Step Suppose that $w{\in}A$, and let (W, σ_w) denote the counting substructure of (X, σ) determined by w. Let $x_1 = \sigma(w)$, and suppose that x_2 is an element of X with $x_2 \neq x_1$. Let X_1 and X_2 be as in the statement of the theorem. Note that, by the preceding corollary, one has $X_1 = \sigma[W]$. If $x_2 = w$ then $x_1 = \sigma(x_2)$, and thus x_1 is an element of W_2, so (i) holds. Thus, suppose $x_2 \neq w$. Then, by the induction hypothesis that w is in A, it follows that either w is in X_2 or x_2 is in W. In the former case it follows (from the fact that X_2 is an inductive set) that $x_1 = \sigma(w)$ is also in X_2, and thus (i) holds. In the latter case, since $x_2 \neq w$, it follows from the fact that $\sigma[W] = W\backslash\{w\}$, that $x_2 \in \sigma[W]$; that is, since $\sigma[W] = X_1$, $x_2 \in X_1$, and thus in this case (ii) holds. Thus, in all the cases either (i) or (ii) holds, hence $\sigma(w)$ is also in A.

It now follows from the Induction Axiom that $A = X$, and the desired theorem follows.

In light of the preceding theorem, it is now easy to introduce the concept of 'greater than' (or equivalently, 'less than') for any counting structure.

Add1-(I) 9: Definition Let (X, σ) be a counting structure. Let x_1 and x_2 be elements of X with $x_1 \neq x_2$, and let (X_1, σ_1) and (X_2, σ_2) denote the counting substructures of (X, σ) determined by x_1 and x_2, respectively. One says that x_2 **is greater than** x_1 or, equivalently, x_1 **is less than** x_2, with respect to the given structure (X, σ), provided x_2 is in X_1. In this case one writes $x_2 >_{(X,\sigma)} x_1$ (equivalently, $x_1 <_{(X,\sigma)} x_2$); or, if the choice of counting structure (X, σ) is understood from the context, simply $x_2 > x_1$ (equivalently, $x_1 < x_2$).

More generally, one writes $x_2 \geq x_1$, or, equivalently, $x_1 \leq x_2$, if either $x_1 = x_2$ or $x_2 > x_1$.

If w is any element of X, then X_w denotes the set of all elements x of X such that $x \leq w$; this set is called the **initial section of** X. More generally, if w_1 and w_2 are elements of X such that $w_1 \leq w_2$, then $X_{[x_1,x_2]}$ denotes the set of all x in X such that $w_1 \leq x \leq w_2$. Note that $X_w = X_{[u,w]}$.

Add1-(I) 10: Examples $X_u = \{u\}$; $X_{\sigma(u)} = \{u, \sigma(u)\}$. More generally, for each x in X one has $X_{\sigma(x)} = X_x \cup \{\sigma(x)\}$.

The next result follows easily from what precedes, so the proof is left as an exercise.

Add1-(I) 11: Theorem Let (X, σ) be a counting structure, and let $<$ denote the corresponding 'less than' relation. Then:

(a) ('The Trichotomy Law') If x_1 and x_2 are elements of X, then exactly one of the following statements is true:

(i) $x_1 = x_2$
(ii) $x_1 < x_2$
(iii) $x_2 < x_1$

(b) ('The Transitivity Law') If x_1, x_2 and x_3 are elements of X such that $x_1 < x_2$ and $x_2 < x_3$, then $x_1 < x_3$.

Add1-(I) 12: Theorem Let (X, σ) be a counting structure. If x_1 is any element of X, then $x_1 < \sigma(x_1)$. Moreover, then there is no element w of X such that $x_1 < w < \sigma(x_1)$. (This is often phrased: 'There is no element of X between two consecutive elements of X'.)

Proof Let $x_2 = \sigma(x_1)$, and as usual denote the counting substructures of (X, σ) determined by x_1 and x_2, respectively, by (X_1, σ_1) and (X_2, σ_2). It follows from the corollary above that $X_2 = X_1\backslash\{x_1\}$; that is, X_2 is obtained by removing from X_1 the single element x_1. Now suppose that there exists w in X such that $x_1 < w < x_2$. Denote the counting substructure of (X, σ) determined by w as (X_w, σ_w). Then, by the definition of the relation '$<$', the element w must be in the set X_1 but not in the set X_2. However, the only element of X_1 which is not in X_2 is x_1, which would imply $w = x_1$. This would contradict the Trichotomy Law, since it is given that $x_1 < w$. Thus, no such w can exist, which proves the desired result.

Add1-(I) 13: Theorem Let (X, σ) be a counting structure. Then every nonempty subset Y of X has a least element; that is, there is an element y_0 in X such that $y_0 \in Y$ and $y_0 \leq y$ for all y

in Y.

This is essentially just the 'Least-Natural-Number Principle', but expressed in the more general context of counting structures. The proof is left as an exercise.

The next result says, in effect, that all counting structures are equivalent in a natural sense, and thus it does not matter which specific example one uses as one's 'standard counting structure'.

Add1-(I) 14: Theorem Let (X, σ) and (Y, τ) be counting structures, with initial elements u and v, respectively. Then there exists a unique bijection $F : X \to Y$ of X onto Y such that
(i) $F(u) = v$;
(ii) $F(\sigma(x)) = \tau(F(x))$ for all x in X.

Proof The proof of the stated result follows directly from the following.

Claim For each w in X there exists a unique function $F_w : X_{\sigma(w)} \to Y$ such that $F_w(u) = v$ and $F_w(\sigma(x)) = \tau(F_w(x))$ for all x in X_w.

Proof ('by Contradiction') of the Claim Let C denote the set of all w in X for which it is *not* the case that there exists a unique function with the indicated properties. If $C \neq \emptyset$, then by the Theorem of the Least Element the set C must have a least element; call it m. Clearly $m \neq u$, since it is obvious that the function $F_u : X_{\sigma(u)} \to Y$ given by the rule $F_u(u) = v$, $F_u(\sigma(u)) = \tau(v)$ has the desired properties, and is the only such function. Thus, m must be of the form $m = \sigma(w_1)$ for some (unique) w_1 in X; clearly $w_1 < m$. Since, by hypothesis, m is the *least* element of C, it follows that w_1 is not in C, and thus there is exactly one function $F_{w_1} : X_{\sigma(w_1)} \to Y$ which has the desired properties. Note that the domain of F_{w_1} can also be written as X_m since $m = \sigma(w_1)$. Now define $F_m : X_{\sigma(m)} \to Y$ by the rule

$$F_m(x) = F_{w_1}(x) \text{ if } x \in X_m; \quad F_m(\sigma(m)) = \tau(F_m(m)).$$

It is clear that this function also satisfies the conditions of the 'Claim', and thus m is not in C. That is, assuming that $C \neq \emptyset$ leads to an element which is simultaneously in C and not in C, which is impossible. Thus, $C = \emptyset$, and the claim follows.

The theorem now follows easily. Indeed, it is clear that the set X is the union of the nonempty subsets of the form X_x for x in X. Moreover, since the intersection of sets X_{x_1} and X_{x_2} is of the form X_x, with x being the larger of x_1 and x_2, it follows from the uniqueness properties enjoyed by the functions F_w described in the Claim that Theorem (I.6.6) can be applied to conclude that there is a unique function $F : X \to Y$ whose restriction to each set $X_{\sigma(w)}$ equals F_w. In light of the results of the 'Claim', this function clearly satisfies Conditions (i) and (ii) of the theorem, and is the only function that does. To see that F is a bijection of X onto Y, first note that in the special case $(Y, \tau) = (X, \sigma)$ one gets that there is a unique function $H : X \to X$ such that $H(u) = u$ and $H(\sigma(x)) = \sigma(H(x))$ for all x in X. However, it is clear that the identity map I_X on X has these properties, so in this case $H = I_X$. Next, note that by reversing the roles of (X, σ) and (Y, τ) in the theorem, it follows that there exists a unique function $G : Y = X$ such that $G(v) = u$ and $G(\tau(z)) = \sigma(G(z))$ for all z in V. Now let $H : X \to X$ be the composition $H = G \circ F$. Then $H(u) = G(F(u)) = G(v) = u$, and $H(\sigma(x)) = G((F(\sigma(x)))) = G(\tau(F(x))) = \sigma(G(F(x))) = \sigma(H(x))$ for all x. Thus by what was observed above, it follows that $G = I_X$. In a similar way one gets $F \circ G = I_Y$. It now follows from Part (c) of Theorem (I.6.5) that F is a bijection of X onto Y, and that $G = F^{-1}$.

The Peano Axioms are strong enough to characterize all the main features of the standard comparison sets for counting. For example, here is how to define 'addition'.

Add1-(I) 15: Theorem Let (X, σ) be a counting structure with initial element u. Then there is a unique function $S_\sigma : X\times X \to X$, called the **sum function associated with the structure** (X, σ), with the following properties: If x is any element of X, then

(a) $S_\sigma(x, u) = S_\sigma(u, x) = \sigma(x)$;

(b) $S_\sigma(x, \sigma(y)) = \sigma(S_\sigma(x, y))$ for all y in X.

If the context makes clear which counting structure (X, σ) is under discussion, one normally drops the explicit reference to σ and writes S instead of the more proper S_σ.

NOTE To understand how these formulas arise, first think of the expression $S(x, y)$ as a shorthand for 'the *sum* of x and y'. Then in the context of $\mathbb{N}$, for which $u = 1$ and $\sigma(x) = x + 1$, the formulas become

$$\text{(i) } S(x, 1) = x + 1 \text{ and (ii) } S(x, y + 1) = S(x, y) + 1.$$

Equation (i) says that the 'sum' S agrees with the usual addition $+$ in $\mathbb{N}$ when applied to x and 1; that is, the 'sum of x and 1', as defined by the function S, equals the usual $x+1$ in $\mathbb{N}$. Equation (ii) says, in effect, that if $S(x, y)$ agrees with the usual $x + y$ in $\mathbb{N}$ for a particular second summand y, then it continues to agree with the usual addition for the next larger second summand $y + 1$. Indeed, if $S(x, y) = x + y$, then $S(x, y) + 1 = (x + y) + 1 = x + (y + 1)$, where the last equation reflects the associative law for ordinary addition of natural numbers. Thus, Equation (ii) becomes $S(x, y + 1) = x + (y + 1)$. Combining these observations with induction on the second summand y then yields the fact that, in the case of $\mathbb{N}$, the sum S satisfies $S(x, y) = x + y$ for *all* x and y.

Proof of Theorem

Uniqueness of S Suppose that S_1 and S_2 are both functions which satisfy the conditions stated in the theorem. Then certainly $S_1(x, u) = S_2(x, u)$ for all x in X, since, by (a), both quantities equal $\sigma(x)$.

Next, let A denote the set of y in X such that $S_1(x, y) = S_2(x, y)$ for all x in X. By what was just proved, it is clear that $u\in A$. Furthermore, if $y\in A$, then by (b) one has

$$S_1(x, \sigma(y)) = \sigma(S_1(x, y)) \text{ and } S_2(x, \sigma(y)) = \sigma(S_2(x, y)) \text{ for all } x \text{ in } X.$$

However, by the hypothesis that $y\in A$ one has $S_1(x, y) = S_2(x, y)$, so that $\sigma(y)$ is also in A. Now the Induction Axiom implies that $A = X$. Thus, $S_1(x, y) = S_2(x, y)$ for all (x, y) in $X\times X$, so $S_1 = S_2$, as claimed.

Existence of S According to Definition (I.5.1), a function with domain $X\times X$ and values in X is a subset of the Cartesian product $Z = (X\times X)\times X$ which satisfies certain properties. The approach followed here is to construct the subset of Z corresponding to the desired function S – viewed as a subset of Z – as the disjoint union of smaller subsets of Z. In effect, we use the fact that Z can be expressed as the disjoint union

$$Z = \bigcup_{x\in X} (\{x\}\times X) \times X$$

Let A be the set of all x in X such that the following holds: There exists a nonempty subset Y_x of Z such that

(i) Every element of Y_x is of the form $((x, y), z)$ for some y and z in X. That is, Y_x is a subset of $(\{x\}\times X) \times X$.

(ii) For every y in X there is exactly one element of the form $((x, y), z)$ in Y_x. That is, Y_x is a function (in the sense of Definition (I.5.1)) with domain $\{x\}\times X$ and values in X.

(iii) The points $((x, u), \sigma(x))$ and $((u, x), \sigma(x))$ are in Y_x. Likewise, for every y in X if $((x, y), z)$ is in Y_x then $((x, \sigma(y)), \sigma(z))$ is also in Y_x. That is, Y_x satisfies Conditions (a) and (b) for the desired function S.

It is easy to show, by Mathematical Induction, that $A = X$:

Initial Step To see that $u \in A$, define Y_u to be the set of all $((x,y),z)$ of the form $((u,y),\sigma(y))$ for y in X. Since σ is a function with domain X and with values in X, it is clear that Conditions (i) and (ii) for the set Y_u hold. As for Condition (iii), note that when $y = u$ then the element $((u,y),\sigma(y))$ becomes $((u,u),\sigma(u))$, so that the first portion of the condition is satisfied. Likewise, if $((u,y),z)$ is in Y_u, then (by definition of Y_u) one has $z = \sigma(y)$. But (again be the definition of Y_u) one also has $((u,\sigma(y)),\sigma(\sigma(y)))$ in Y_u; that is, the second part of Condition (iii) also holds, so $u \in A$.

Induction Step Suppose that $x \in A$, and let Y_x be a set which satisfies Conditions (i), (ii) and (iii). Then for each y in X there is a unique element y' in X such that $((x,y),y')$ is in Y_x. (Of course y' depends on both x and y.) Using this notation, define $Y_{\sigma(x)}$ to be the set of all elements of Z of the form $((\sigma(x),y),\sigma(y'))$ with y in X. It is easy to see, from the properties of Y_x, that $Y_{\sigma(x)}$ also satisfies Conditions (i), (ii) and (iii), and thus $\sigma(x)$ is in A.

Finally, define S to be the union of the sets Y_x, with x in X, constructed above. The sets Y_x are clearly disjoint, so it is easy to see that the set S is a function with domain $X \times X$ and with values in X.

And since each set Y_x satisfies Conditions (i), (ii) and (iii), it is clear that the function S satisfies Conditions (a) and (b) of the theorem.

In a like manner one can define 'multiplication'.

Add1-(I) 16: Theorem Let (X,σ) be a counting structure with initial element u. Then there is a unique function $P_\sigma : X \times X \to X$, called the **product function associated with the structure** (X,σ), with the following properties: If x is any element of X, then

(a) $P_\sigma(x,u) = P_\sigma(u,x) = x$;

(b) $P_\sigma(x,\sigma(y)) = S(P_\sigma(x,y),x)$ for all y in X, where S denotes the sum function S_σ associated with (X,σ).

As usual, if there is no possibility of confusion one normally writes P instead of P_σ.

Proof Left to the reader.

Add1-(I) 17: Examples

(1) Let $X = \mathbb{N} = \{1,2,3,\dots\}$, with the usual rule for the successor function σ; see Example P3 (1) above. Then the successor operation can be written

$$\sigma(x) = x + 1 \text{ for all } x \text{ in } \mathbb{N},$$

where throughout this example the symbol $+$ denote the usual addition of natural numbers. Likewise, the operations S and P associated with this counting structure are clearly the usual addition and multiplication $+$ and $\cdot$ on $\mathbb{N}$. For example, the condition $S(x,1) = \sigma(x)$ then takes the form

$$S(x,1) = x + 1.$$

Thus, by induction on y, if x and y are elements of $\mathbb{N}$ such that $S(x,y) = x+y$, then the condition $S(x,\sigma(y)) = \sigma(S(x,y))$ takes the form

$$S(x,y+1) = (x+y)+1 = x+(y+1),$$

in which the final equation reflects the usual properties of arithmetic in $\mathbb{N}$. That is, $S(x,y) = x+y$ holds for all x, y in $\mathbb{N}$. A similar argument shows that $P(x,y) = x \cdot y$ for all x, y in $\mathbb{N}$.

(2) Now let $X = \hat{\mathbb{N}} = \{0, 1, 2, \dots\}$, so that the role of initial element u is now played by the number 0. In this case the 'addition' and 'multiplication' functions associated with the given counting structure do *not* agree with the usual $+$ and $\cdot$ of numbers. For example, $x + u = x + 0 = x \neq \sigma(x)$, since in $\hat{\mathbb{N}}$ one has $\sigma(x) = x + 1$. Thus, it is not the case that $S(x, y) = x + y$ for all x, y. Likewise, $x{\cdot}0 = 0 \neq x$ when $x \neq 0$, so it is not the case that $P(x, y) = x{\cdot}y$ for all x, y. Indeed, one can easily show that in $\hat{\mathbb{N}}$ one has

$$S(x, y) = x + y + 1 \text{ and } P(x, y) = x(y + 1) + y \text{ for all } x,\ y \text{ in } \hat{\mathbb{N}}$$

The fact that in the example of $\hat{\mathbb{N}}$ the operations S and P as define above do not agree with the usual addition and multiplication on $\hat{\mathbb{N}}$ is not a flaw. It simply reflects the fact that, in any choice of addition and multiplication on a counting structure (X, σ), one must decide the role to be played by the initial element u. The choice taken in *This Textbook* reflects the fact that we shall use $\mathbb{N}$, not $\hat{\mathbb{N}}$, as our primary model of a counting structure. It is an easy exercise to provide an alternative formulation of 'addition' and 'multiplication' axioms for those who prefer that the initial element u behave like 0.

Add1-(I) 18: Modified Notation It is customary to use the more familiar notation 1_σ, $x +_\sigma y$, $x{\cdot}_\sigma y$ instead of u_σ, $S_\sigma(x, y)$ and $P_\sigma(x, y)$ when dealing with a general counting structure (X, σ). And, as usual, if the context makes clear which σ is under consideration, the subscript σ is normally omitted and one writes 1, $x + y$ and $x{\cdot}y$ instead. With this notation, the basic defining properties of addition and multiplication take the following more familiar form for all x (or, where appropriate, all x and y) of X:

$$(i)\,\sigma(x) = x{+}1 = 1{+}x;\ (ii)\,x{+}(y{+}1) = (x{+}y){+}1;\ (iii)\,x{\cdot}1 = 1{\cdot}x = x;\ (iv)\,x{\cdot}(x{+}1) = x{\cdot}y{+}x \qquad (*)$$

The operations of addition and multiplication for a counting structure obey the usual rules of grade-school arithmetic.

Add1-(I) 19: Theorem Let (X, σ) be a counting structure, with initial element 1, addition operation $+$ and multiplication $\cdot$. Then the following facts hold.

(a) (Commutative Laws) If x and y are in X then

$$(i)\,x + y = y + x \text{ and } (ii)\,x{\cdot}y = y{\cdot}x.$$

(b) (Associative Laws) If x, y and z are in X then

$$(i)\,(x + y) + z = x + (y + z) \text{ and } (ii)\,(x{\cdot}y){\cdot}z = x{\cdot}(y{\cdot}z).$$

(c) (Distributive Laws) If x, y and z are in X then

$$(i)\,x{\cdot}(y + z) = x{\cdot}y + x{\cdot}z \text{ and } (ii)\,(x + y){\cdot}z = x{\cdot}y + x{\cdot}z.$$

Partial Proof The laws above are grouped to emphasize the similarities shared by addition and multiplication, and to ease the task of learning their statements; for example, the two commutative laws are joined together, and are stated before the two (more complicated) associative laws. The proofs of the laws as given here, however, require handling them in a different order. For instance, we use the Commutative and Associative Laws for Addition, together with the Distributive Laws, in the proof of the Associative Law for Multiplication. As usual, since the structure of the proofs of the various laws are so similar, we leave some of the proofs as exercises.

Part (i) of (b) (Associative Law for Addition) Let A_1 denote the set of z in X such that $(x+y)+z = x+(y+z)$ for all x and y in X. It is clear that it suffices to show that $A_1 = X$.

Initial Step Property (ii) of $(*)$ above states that $1{\in}A_1$.

Inductive Step Suppose that $z{\in}A_1$. Then for all x and y in X one has

$$x+(y+(z+1)) \overset{(1)}{=} x+((y+z)+1) \overset{(2)}{=} (x+(y+z))+1 \overset{(3)}{=} ((x+y)+z)+1 \overset{(4)}{=} (x+y)+(z+1)$$

That is, $(z+1)$ is also in A_1, so by mathematical induction $A_2 = X$, as desired.

Here are the justifications for the numbered equations above:

Equation (1): This comes by applying Property (ii) of $(*)$ to the expression $y+(z+1)$.
Equation (2): This follows from the fact, proved above, that $1{\in}A_1$.
Equation (3): This reflects the induction hypothesis that $z{\in}A_1$.
Equation (4): This comes by applying Property (ii) of $(*)$ to the expression $((x+y)+z)+1$.

Part (i) of (a) (Commutative Law for Addition) Let A_2 be the set of all y in X such that $x+y = y+x$ for all x in X.

Initial Step It is clear, from the defining proeprties of 'addition', that $x+1 = 1+x = \sigma(x)$ for all x in X, so certainly $1{\in}A_2$.

Inductive Step Suppose that $y{\in}A_2$. Then for all x in X one has

$$x+(y+1) \overset{(1)}{=} (x+y)+1 \overset{(2)}{=} (y+x)+1 \overset{(3)}{=} y+(x+1) \overset{(4)}{=} y+(1+x) \overset{(5)}{=} (y+1)+x.$$

That is, $(y+1){\in}A_2$, so $A_2 = X$, and the desired result follows.

Here are the justifications of the preceding equations:

Equations (1), (3) and (5): the Associative Law for Addition
Equations (2) and (4): the fact that 1 and (by the Induction Hypothesis) y are both in A_2.

Part (i) of(c) (The First Distributive Law) Let A_3 be the set of all z in X such that $x{\cdot}(y+z) = x{\cdot}y + x{\cdot}z$ for all x, y in X.

Initial Step Note that, from the defining properties of 'multiplication', one has

$$x{\cdot}(y+1) = x{\cdot}y + x = x{\cdot}y + x{\cdot}1.$$

That is, $x{\cdot}(y+1) = x{\cdot}y + x{\cdot}1$ for all x and y in X, hence $1{\in}A_3$.

Induction Step Suppose that $z{\in}A_3$. Then for all x and y in X one has

$$x{\cdot}(y+(z+1)) \overset{(1)}{=} x{\cdot}((y+z)+1) \overset{(2)}{=} x{\cdot}(y+z)+x \overset{(3)}{=} (x{\cdot}y+x{\cdot}z)+x \overset{(4)}{=} x{\cdot}y+(x{\cdot}z+x) \overset{(5)}{=} x{\cdot}y+x{\cdot}(z+1).$$

Thus, $z+1$ is also in A_3. It now follows that $A_3 = X$, so the First Distributative Law, $x{\cdot}(y+z) = x{\cdot}y + x{\cdot}z$, follows.

The justifications of the preceding equations are as follows:

Equations (1) and (4): The Associative Law for Addition
Equations (2) and (5): Defining properties of 'multiplication'
Equation (3): The induction hypothesis that $z{\in}A_3$.

Part (ii) of(c) (The Second Distributive Law) Let A_4 be the set of all z in X such that $(x+y){\cdot}z = x{\cdot}z + y{\cdot}z$ for all x, y in X.

Initial Step Since $(x+y){\cdot}1 = x+y = x{\cdot}1 + y{\cdot}1$ by one of the defining properties of 'multiplication', it follows that $1{\in}A_4$.

Induction Step Suppose that $z \in A_4$. Then for all x and y in X one has

$$(x+y)\cdot(z+1) \overset{(1)}{=} (x+y)\cdot z+(x+y) \overset{(2)}{=} (x\cdot z+y\cdot z)+(x+y) \overset{(3)}{=} (x\cdot z+x)+(y\cdot z+y) \overset{(4)}{=} x\cdot(z+1)+y\cdot(z+1).$$

Thus $z+1$ is also in A_4. It follows that $A_4 = X$, so the desired result holds.

The reader should be able to figure out the justifications for the preceding equations. Note that Equation (3) uses several applications of the commutative and associative laws for addition; be sure you can break it down to the individual applications of those laws. Also, be sure you can determine where the induction hypothesis gets used.

Part *(ii)* of (a) (The Commutative Law for Multiplication) Let A_5 be the set of all y in X such that $x\cdot y = y\cdot x$ for all x in X.

Initial Step One of the defining proerties of 'multiplication' is that $x\cdot 1 = 1\cdot x = x$ for all x in X, so clearly $1 \in A_5$.

Induction Step Suppose that $y \in X$. Then for all x in X one has

$$x\cdot(y+1) = x\cdot y+x = y\cdot x+x = y\cdot x+1\cdot x = (y+1)\cdot x.$$

(The reader is invited to explain why each of these equations is valid.) In particular, $y+1$ is also in A_5, and thus $A_5 = X$, as desired.

The proofs of the remaining laws are left as exercises.

There is a simple relation between the operation of 'addition' and the concept of 'greater than'.

Add1-(I) 20: Theorem Let (X,σ) be a counting structure, with associated initial element 1, addition operation + and 'greater than' ordering $>$. Then a pair of elements x and y in X satisfy the condition $y > x$ if, and only if, there exists z in X such that $y = x+z$.

Proof

(The 'If' Part) Let A be the set of all z in X such that for all x in X the elements x and $x+z$ satisfy $x+z > x$. Certainly $1 \in A_1$, by Theorem Add1-(I) 12. Next, suppose that $z \in A$. Then, for all x in X, one has $(x+z)+1 > x+z$ (since $1 \in A$), and $x+z > x$ (since, by the induction hypothesis, $z \in A_1$). Then, from the Transitive Law one obtains $(x+z)+1 > x$; that is, in light of the Associative Law, $x+(z+1) > x$. It follows that $z+1$ is also in A. Now apply the Principle of Mathematical Induction to conclude that $A = X$, as required.

(The 'Only if' Part) Suppose that x and y are elements of X such that $y > x$. Let C be the set of all w in X such that $x+w \geq y$. By what was just proved, it is clear that $y \in C$, since $x+y > y$. In particular, C is a nonempty subset of X, so, by Theorem Add1-(I) 13, there is a smallest element of C; call it z. Clearly $x+z \geq y$, since z is in C. If $z = 1$ then one has $x+1 \geq y > x$, which implies $x+1 = y$, since there are no elements of X which lie between x and $x+1$. Thus, suppose $z \neq 1$, so that $z = v+1$ for some v in X. Since z is the *smallest* element of C, and clearly $v < z$, it follows that v is *not* in C. That is, it is not the case that $x+v \geq y$, hence one has $x+v < y$. Thus, one has $x+v < y \leq (x+v)+1$. Since there are no elements of X between $(x+v)$ and $(x+v)+1$, it must be the case that $y = (x+v)+1 = x+(v+1) = x+z$, as claimed.

There is a similar theorem involving multiplication and the ordering 'greater than'.

Add1-(I) 21: Theorem Let (X,σ) be a counting structure, with associated initial element 1, multiplication operation $\cdot$ and 'greater than' ordering $>$. If x and y are elements of X such that $y > x$, then $y\cdot z > x\cdot z$ for all x in X.

The simple proof is left as an exercise.

There is a result, *not* one of the standard facts from grade-school arithmetic, that is worth mentioning here.

Add1-(I) 22: Theorem Let (X, σ) and (Y, τ) be counting structures with initial elements 1_σ and 1_τ, addition operations $+_\sigma$ and $+_\tau$, multiplication operations $\cdot_\sigma$ and $\cdot_\tau$, and orderings $>_\sigma$ and $>_\tau$, respectively. Furthemore, let $F : X \rightarrow Y$ be the bijection described in Theorem Add1-(I) 14. Then the bijection F preserves addition and multiplication, in the following sense:

$$F(x_1 +_\sigma x_2) = F(x_1) +_\tau F(x_2) \text{ and } F(x_1 \cdot_\sigma x_2) = F(x_1) \cdot_\tau F(x_2) \text{ for all } x_1, x_2 \text{ in } X.$$

Likewise, F preserves the ordering, in the following sense:

$$\text{If } x_2 >_\sigma x_1 \text{ then } F(x_2) >_\tau F(x_1).$$

(For readers with a background in modern algebra: the map F is an *isomorphism* between the two counting structures.)

The simple proof of this theorem is left as an exercise; it boils down to noticing that the definitions of addition, multiplication and the order all come from the properties of the successor functions σ and τ; and the bijection F 'preserves' those functions because of the condition $F(\sigma(x)) = \tau(F(x))$.

The import of this result is that it does not matter which counting structure one elects to use: they are equivalent. From now on we shall use the standard example $\mathbb{N}$ from grade-school arithmetic, for which the successor function is 'addition by 1'.

It is useful to prove some of the results that were accepted without proof earlier in *This Textbook*. For example, here is a more complete treatment of Theorem (I.7.7)

Add1-(I) 23: Theorem

(a) Let X be a nonempty finite set. If X has the same cardinality as $\mathbb{N}_k$ and the same cardinality as $\mathbb{N}_m$ for natural numbers k and m, then $k = m$.

(b) If Y is a subset of a finite set X, then Y is a finite set, and $\#(Y) \leq \#(X)$. Moreover, the only time one gets $\#(Y) = \#(X)$ is when $Y = X$. In particular, X cannot have the same cardinality as one of its proper subsets (i.e., there is no 'Galileo Paradox' for finite sets.)

(c) Suppose that $\{X_1, X_2, \ldots X_n\}$ is a finite collection of finite sets. Then the union $X_1 \cup X_2 \cup \ldots \cup X_n$ is also a finite set. More precisely,

$$\#(X_1 \cup X_2 \cup \ldots \cup X_n) \leq \#(X_1) + \#(X_2) + \ldots + \#(X_n).$$

One gets equality in this last relation if, and only if, the sets are mutually disjoint, in the sense that $X_i \cap X_j = \emptyset$ whenever $i \neq j$.

Proof

(a) It follows easily from Part (c) of Theorem (I.7.4) that the issue to be proved reduces to this: if k and m are natural numbers such that $\mathbb{N}_k$ has the same cardinality as $\mathbb{N}_m$, then $k = m$; equivalently: if $k \neq m$, then $\mathbb{N}_k$ does *not* have the same cardinality as $\mathbb{N}_m$. Then in light of Part (b) of the same theorem, the problem reduces to proving the following statement:

For every j in $\mathbb{N}$, if k in $\mathbb{N}$ satisfies the condition $k > j$, then $\mathbb{N}_j$ does not have the same cardinality as $\mathbb{N}_k$.

We shall prove a slightly more precise result: For every j in $\mathbb{N}$. if $k{\in}\mathbb{N}$ satisfies $k > j$, then there is no surjection of $\mathbb{N}_j$ onto $\mathbb{N}_k$.

Indeed, let A be the set of j in $\mathbb{N}$ for which this last statement is true.

Initial Step It is clear that $1{\in}A$. Indeed, suppose that $k > 1$, so that $k \geq 2$ and thus 1 and 2 are both elements of $\mathbb{N}_k$. If $F : \mathbb{N}_1 \to \mathbb{N}_k$ were a surjection onto $\mathbb{N}_k$, then there would have to exist x_1 and x_2 in $\mathbb{N}_1$ such that $F(x_1) = 1$ and $F(x_2) = 2$. However, the only element of $\mathbb{N}_1$ is 1, so this would require $x_1 = x_1 = 1$, and thus $F(1) = 1$ and $F(1) = 2$. Viewing F as a set of ordered pairs, this would mean that $(1,1){\in}F$ and $(1,2){\in}F$, contrary to the definition of 'function'.

Inductive Step Suppose that $j{\in}A$. If $j+1$ were *not* in A, then there would have to exist a surjection F from $\mathbb{N}_{j+1}$ onto $\mathbb{N}_m$ for some $m > j+1$. Note that in this case one would certainly have $k = m-1 > j$. There are two cases to consider:

Case 1: Suppose that $F(j+1) = m$. Define $G : \mathbb{N}_j \to \mathbb{N}_k$ by the rule

$$G(i) = \begin{cases} F(i) & \text{if } F(i) \neq m \\ 1 & \text{if } F(i) = m \end{cases}$$

It is easy to see that G would have to map $\mathbb{N}_j$ onto $\mathbb{N}_k$ with $k > j$. This would contradict the induction hypothesis that $j{\in}A$.

Case 2: Suppose that $F(j+1) \neq m$. Let $p = F(j+1)$, so that $1 \leq p \leq k$. Now define $G : \mathbb{N}_j \to \mathbb{N}_k$ by the rule

$$G(i) = \begin{cases} F(i) & \text{if } F(i) \neq m \\ p & \text{if } F(i) = m \end{cases}$$

It is easy to see that G maps $\mathbb{N}_j$ onto $\mathbb{N}_k$, contrary to the hypothesis that $j{\in}A$.

This argument shows that $j+1$ is also in A. Thus, by the Principle of Mathematical Induction one concludes that $A = \mathbb{N}$. The desired result now follows.

(b) and (c): These follow easily from Part (a) combined with Mathematical Induction. The details are left as exercises.

There are many more results about $\mathbb{N}$ (or, if you prefer, about counting structures) that one could list. However, the main purpose of this appendix is to convince the reader that the standard properties of $\mathbb{N}$ can be derived from the Dedekind-Peano axioms. The results already given here are sufficient for that purpose.

Appendix B

Further Results in Set Theory

Quotes for Appendix B:

(1) Fake Quote

In this appendix we consider some standard results in Set Theory with which every mathematician should eventually become familiar; however, this material is not required for reading the preceding chapters.

B.0.1 Theorem (DeMorgan's Laws for Sets)

Let $\mathcal{A}$ be a nonempty family of sets, and let X be a set; we do *not* assume that X is a member of the family $\mathcal{A}$. Let $\mathcal{B}$ be the family of all sets Y which can be expressed in the form $Y = X \backslash A$ for at least one set A in the family $\mathcal{A}$. Then

(i) $X \backslash (\bigcup \mathcal{A}) = \bigcap \mathcal{B}$. (ii) $X \backslash (\bigcap \mathcal{A}) = \bigcup \mathcal{B}$.

Using the alternate notation introduced in Part (d) of Definition (I.2.8) above, these set identities can be written, without needing to introduce the family $\mathcal{B}$, as follows:

$$(i)\ X \backslash \left(\bigcup_{A \in \mathcal{A}} A\right) = \bigcap_{A \in \mathcal{A}} (X \backslash A) \quad (ii)\ X \backslash \left(\bigcap_{A \in \mathcal{A}} A\right) = \bigcup_{A \in \mathcal{A}} (X \backslash A).$$

Proof of (i) (The proof of (ii) is similar, and is left as an exercise): Suppose that c is an element of the set $X \backslash (\bigcup \mathcal{A})$. Then c is in X but c is not in $\bigcup \mathcal{A}$. To say that c is not in $\bigcup \mathcal{A}$ means that for each A in the family $\mathcal{A}$, c is not in A. Thus, for each such A one has c in X but c not in A; hence one has $c \in (X \backslash A)$. Thus, c is in each of the sets of the family $\mathcal{B}$, hence c is in $\bigcap \mathcal{B}$. It follows that

$$X \backslash \left(\bigcup \mathcal{A}\right) \subseteq \bigcap \mathcal{B} \quad (*)$$

Likewise, suppose that p is an element of $\bigcap \mathcal{B}$. Then p is in each of the sets of the family $\mathcal{B}$; that is, p is in every set of the form $X \backslash A$ with A in the family $\mathcal{A}$. That is, for each such A the point p is in X but not in A. Since p fails to be in any of these sets A, it follows that p is also not in $\bigcup \mathcal{A}$; and since p is in X, it then follows that p is in $X \backslash (\bigcup \mathcal{A})$. That is,

$$\bigcap \mathcal{B} \subseteq X \backslash \left(\bigcup \mathcal{A}\right) \quad (**)$$

The set relations $(*)$ and $(**)$ then imply, by Theorem (I.2.7), that $X \backslash (\bigcup \mathcal{A}) = \bigcap \mathcal{B}$, as claimed.

Remark The preceding theorem is sometimes expressed informally as follows:
(i) The complement of a union is the intersection of the complements;
(ii) The complement of an intersection is the union of the complements.

B.0.2 Remark Russell's Paradox

The theory of sets is a very deep subject, and is still an area of active research. Fortunately, the use of sets in *This Textbook* does not require a deep knowledge of this theory. Nevertheless, it is appropriate to mention one topic from the wider theory of sets.

The basic issue arises directly from the definition of a 'set' as being simultaneously a collection of objects and an object in its own right. As has already been seen, this allows the construction of sets whose individual elements are themselves sets.

Example: Let us say that a set Y has Property R provided that the set Y, thought of as an object in its own right, is *not* an element of the set Y, thought of as a collection of objects. (Note that all the sets we have considered so far do have Property R.) Now define the set X be the family of all sets Y such that Y has Property R.

Question Does the set X have Property R?

Analysis Suppose that the set X *does* have Property R. Then, by the very definition of the set X just given, the elements of X include *every* set having Property R and thus include X itself. That is, X, viewed as an object, is an element of X, viewed as a set. However, by definition of Property R, this last fact means that X does not have Property R, which contradicts the supposition that X does have Property R.

Now suppose that X does *not* have Property R. Then, by definition of 'Property R', the set X, thought of as an object, *is* an element of the set X. But by the very definition of X given above, each element of the set X – including now X itself – is a set which has Property R, contradicting the supposition that X does not have Property R.

In summary: assuming that X *does* have Property R implies that X does *not* have Property R; while assuming X does *not* have Property R implies that X *does* have Property R. This mysterious state of affairs is called **Russell's Paradox**, in honor of the British philosopher Bertrand Russell, who discovered it in 1901. It is well beyond the scope of *This Textbook* to consider a proper analysis of such set-theoretic paradoxes; for that, the reader should pick up a book on the logical foundations of mathematics. Instead, we follow the common approach in Analysis and assume that the sets we deal with are all subsets of some fixed, but unspecified, 'universal set' which is big enough for our purposes but 'small enough' to avoid such paradoxes. In particular, this approach allows us to use only sets X which satisfy the relation $X \notin X$.

(4) Suppose that X is a nonempty set and let k be a natural number. If $X_1, X_2, \ldots X_k$ are nonempty subsets of X, then the set of all k-tuples $g : \mathbb{N}_k \to X$ such that $g(j) \in X_j$ for each $j = 1, 2, \ldots k$ is denoted $X_1 \times X_2 \times \ldots \times X_k$; it is called the **Cartesian product** of these sets. This is sometimes written more briefly by expressions such as $\prod_{j=1}^{n} X_j$, where, as usual, the symbol $\prod$ stands for 'product'.

More generally, suppose that $\varphi : Z \to \mathcal{P}(X) \backslash \{\emptyset\}$ is a function defined on a nonempty set Z, with values being nonempty subsets of X. For each i in Z let $X_i = \varphi(i)$. Then the corresponding **Cartesian product of the indexed family $\boldsymbol{X}_i$**, denoted $\prod_{i \in Z} X_i$, consists of the functions $g : Z \to X$ such that $g(i) \in X_i$ for each i in Z. The set X_i is called the ***i*-th factor** in this product.

It is convenient to extend these ideas to cases in which one of the 'factors' is the empty set by declaring that such a product to also be the empty set.

Remark For small values of k it is customary to use older language. For instance, an ordered 2-tuple is often called an **ordered pair**. Of course this conflicts with the primitive (i.e., Kuratowski) concept of 'ordered pair. The way to get around this ambiguity is to think of the definition of 2-tuple (as a function defined on the set $\{1, 2\}$) as the 'new, improved' type of ordered pair, while the Kuratowski definition describes the 'primitive ordered pair'. From this point on in *This Textbook* the phrase 'ordered pair' means this new, improved version unless stated otherwise.

Associated with a set X is a second set whose elements are precisely the subsets of X, each such subset being thought of as a single object in its own right.

B.0.3 Definition

Let X be a set. The **power set of X** is the collection whose elements are precisely the subsets of X. That is, to say that an object C is an element of the power set of X means that C is a subset of X. Some authors denote the power set of X by the symbol $\mathcal{P}(X)$; some denote it by the symbol 2^X (whence the name '*power* set'). We shall use the $\mathcal{P}(X)$ notation in *This Textbook.* (See Example (5) below for the origin of the 'exponential' notation 2^X mentioned here.)

B.0.4 Examples

(1) Suppose that $X = \{c\}$ for some object c, so that X is a singleton set. It is obvious that $\mathcal{P}(X)$ has exactly two (different) subsets, namely X itself and the empty set $\emptyset$. (You should try to provide a correct proof before going on. You might also wish to read the Side Comment 'on the power set of a singleton set'.)

(2) A similar analysis, which is left as an exercise, shows that if $X = \{c, d\}$ where $c \neq d$, (i.e., if X is a doubleton set), then $\mathcal{P}(X) = \{\emptyset, \{c\}, \{d\}, X\}$. In particular, if X has exactly two elements, then $\mathcal{P}(X)$ has exactly four elements.

(3) The empty set $\emptyset$ has a subset, namely itself; see Part (b) of Theorem (I.2.9). It is also clear that $\emptyset$ has no other subsets, and thus that $\mathcal{P}(\emptyset) = \{\emptyset\}$. In particular, the power set of the empty set is not the empty set, since $\mathcal{P}(\emptyset)$ does have an element, namely the set $\emptyset$, thought of as an object in its own right.

(4) It is possible that X and $\mathcal{P}(X)$ can have one or more elements in common. For instance, let c be an object, and let $X = \{c, \{c\}\}$. Thus, X is a set with two distinct elements, namely the object c and the singleton set $\{c\}$, thought of as an object in its own right. It follows from the results of Example (2) above that the power set $\mathcal{P}(X)$ consists of the four elements $\emptyset$, $\{c\}$, $\{\{c\}\}$ and X. Notice that the second entry in this list is the set $\{c\}$, which is also an element of X.

A simpler example of this type is $X = \{\emptyset\}$; that is, X is as in Example (1) above, with c equal to the empty set $\emptyset$, thought of as an object in its own right. Then the result of Example (1) takes the form $\mathcal{P}(X) = \{\emptyset, X\}$. In particular, the object $\emptyset$ is both an element of X and an element of $\mathcal{P}(X)$.

(5) We shall see later that if X is a set with exactly n members, where n is some natural number, then $\mathcal{P}(X)$ has exactly 2^n members; and of course if X has no members (i.e., if $X = \emptyset$), then $\mathcal{P}(X)$ has $1 = 2^0$ members. The exponential notation 2^X grew out of these facts.

Side Comment (on the power set of a singleton set): In Example (1) above it is asserted that the fact that $X = \{c\}$ has exactly two subsets, namely, X itself and $\emptyset$. Whenever you see such an assertion of 'obviousness' in a math context, beware: the author may be trying to slide something by you. And even if you agree that the stated result is 'obvious', you should always be prepared to give a *proper* proof; that is, a proof which arises from definitions and previously-accepted, results using correct logic. For example, the following is a 'fake proof':

'The sets X and $\emptyset$ are certainly subsets of X, and I can't think of any others.'

Here is a proper proof:

First note that, by Part (b)(i) of Theorem (I.2.9), the sets $\emptyset$ and X are, indeed, subsets of X; in particular, by the definition of 'power set' and 'subset', $\{\emptyset, X\} \subseteq \mathcal{P}(X)$. (This is essentially how the 'fake proof' above starts.)

To see that there are no other subsets of X, i.e., no other elements of $\mathcal{P}(X)$, let Y be any nonempty subset of X. Since, by hypothesis, $Y \neq \emptyset$, it follows, by the definition of 'empty set', that there must be at least one element in the set Y. Let d be any element of Y. Since, by hypothesis, $Y \subseteq X$, it follows (from the definition of 'subset') that d is an element of X. However, by the Fundamental Principle of Set Theory, combined with the hypothesis that $X = \{c\}$, so that c is the *only* element of X, it follows that $d = c$. That is, every element of Y equals c, so Y and X have exactly the same elements, namely the single element c. Thus, by the Fundamental Principle of Set Theory, one has $Y = X$. That is, if $Y \subseteq X$ and $Y \neq \emptyset$, then $Y = X$. It follows that $\mathcal{P}(X) \subseteq \{\emptyset, X\}$. However, it was already noted above that $\mathcal{P}(X) \supseteq \{\emptyset, X\}$.

In the latter case, Y must have at least one element (by definition of 'not the empty set'), and any such element must be an element of X (by definition of Y being a *subset* of X). If d is any element of Y, then, as just shown, d must be an element of X and thus, by the hypothesis that $X = \{c\}$, one must have $d = c$.

Preliminary Comment One knows that if X is a finite set with exactly n elements, then the corresponding power set $\mathcal{P}(X)$ has exactly 2^n elements. In particular, the power set $\mathcal{P}(X)$ has more elements than the original set X; moreover, the larger n is, the greater the difference in the sizes of the sets. This observation may make the next result seem almost 'obvious'.

B.0.5 Theorem (Cantor's Power-Set Theorem)

Let X be a set, and let $Y = \mathcal{P}(X)$ be the corresponding power set of X.

Then X does *not* have the same cardinality as Y; that is, there does not exist a bijection of X onto $\mathcal{P}(X)$. Indeed, even more can be said; namely, if $X \neq \emptyset$ and $F : X \to \mathcal{P}(X)$ is any function with domain X and values in $\mathcal{P}(X)$, then F does *not* map X onto $\mathcal{P}(X)$; that is, F is *not* a surjection of X onto $\mathcal{P}(X)$. In particular, anyone who claims to have constructed an example of a bijection of X onto $\mathcal{P}(X)$ is wrong, since there does not exist a surjection, much less a bijection.

Proof: Case 1 Suppose that X is the empty set. Recall that $\mathcal{P}(\emptyset) = \{\emptyset\}$, so $\mathcal{P}(\emptyset)$ is a nonempty set. Since, as has already been noted, the empty set does not have the same cardinality as a nonempty set, it follows that the claimed result is true when X is empty. (More intuitively: The empty set has no elements, while $\mathcal{P}(\emptyset)$ has one element.)

Case 2 Suppose that X is nonempty, and suppose that $F : X \to \mathcal{P}(X)$ is a function with domain X and values in $\mathcal{P}(X)$. Let S be the set of all x in X such that x is *not* an element of the set $F(x)$.

Claim The set S is not in the image of the function F.

Proof of Claim Suppose, to the contrary, that S can be expressed as $F(x_0)$ for some x_0 in X. If x_0 is in S, then (by definition of S) x_0 is *not* an element of $F(x_0)$; that is, since S is supposed to equal $F(x_0)$, one has x_0 *not* in S. Similarly, if x_0 is *not* in S, then (again by the definition of S) x_0 must be an element of $F(x_0)$; that is, x_0 is an element of S. Thus, assuming that there exists x_0 in X such that $S = F(x_0)$ leads to the existsence of an element, namely this x_0, which is simultaneously an element of S and *not* an element of S. No such x_0 can exist, hence S is not in the image of F.

B.0.6 Corollary

There exist sets which are uncountable.

Proof

Let $Y = \mathcal{P}(\mathbb{N})$, the power set of $\mathbb{N}$. It is clear that Y is not finite since it has the infinite subset $\{\{n\} : n{\in}\mathbb{N}\}$ whose elements are the singleton sets of natural numbers. In addition, by Cantor's Power-Set Theorem above, Y does not have the same cardinality as $\mathbb{N}$; that is, Y is not countably infinite. It follows that the set Y is not countable.

The uncountable set $\mathcal{P}(\mathbb{N})$ is difficult to visualize. Fortunately, it is possible to interpret it geometrically as a subset of the real number line. The basic idea is to use the representation of subsets of $\mathbb{N}$, in terms of sequences of 0s and 1s, discussed in Example (I.9.2) (6). For convenience we break up the discussion into several parts.

B.1 'Ordered Pairs' in Terms of Sets (Kuratowski)

In the current section we illustrate the 'reduction to set theory' process by expressing the important concept of 'an ordered pair of objects' purely in terms of sets.

Most students first encounter the idea of an 'ordered pair of objects' in analytical geometry: one characterizes the location of a point (i.e., a 'dot') in the xy-plane in terms of a pair of numbers, the 'Cartesian coordinates' of the point (relative to a choice of origin and coordinate axes). The order in which these numbers are written makes a difference; for instance, the pair $(2, -3)$ corresponds to a geometric point in the fourth quadrant of the xy-plane, while the pair $(-3, 2)$ corresponds to a geometric point in the second quadrant.

More generally, if x and y are any objects, one seeks a purely set-theoretic way to encode the main information contained in intuitive concept of 'the ordered pair (x, y)':

(i) The concept should be expressed purely in terms of sets formed from the objects x and y.

(ii) If $x \neq y$, then the definition should allow one to distinguish which of these objects is to be the 'first' and which is to be the 'second'.

Notice that the 'obvious' solution, namely to define the desired ordered pair as the set $\{x, y\}$, does not satisfy Condition (ii); indeed, one has $\{x, y\} = \{y, x\}$, so this doubleton set does not single out one of the objects x or y as somehow being 'preferred'.

Around 1914 the American mathematician Norbert Wiener developed a suitable set-theoretic treatment of 'ordered pairs'; this was simplified by the Polish mathematician Kasimir Kuratowski in 1921. We follow Kuratowski's approach here.

B.1.1 Definition (Kuratowski Ordered Pairs)

Let x and y be objects. Then the **Kuratowski ordered pair whose first entry is x and whose second entry is y** is the set $\{\{x\},\{x,y\}\}$. If $x = y$ one says that the ordered pair is of **Type I**, while if $x \neq y$ then it is of **Type II**.

Remarks

(1) For convenience one normally denotes the ordered pair whose first entry is x and whose second entry is y by the standard symbol (x, y), where the first entry of the ordered pair appears on the left in this notation and the second on the right; in the spoken form, one then refers to 'the ordered pair x y', with the first entry of the pair spoken first. Note that neither of these conventions distinguishes 'first entry' in a purely set-theoretic way: the phrase 'on the left' introduces a visual aspect, while the phrase 'spoken first' introduces a temporal aspect.

(2) In the Kuratowski formulation, the ordered pair with first entry x and second entry y is either a singleton set, if $x = y$, or a doubleton set, if $x \neq y$.

(3) As was indicated above, other set-theoretic definitions of 'ordered pairs' have been proposed. These are not simply rewordings of the Kuratowski definition which describe the same objects. For instance, in the Kuratowski formulation the ordered pair whose first entry is the number x and whose second entry is the number y is the set $\{\{x\},\{x,y\}\}$, thought of as a single object. In the earlier formulation of Norbert Wiener, however, the same ordered pair is defined to be the set $\{\{\{x\},\emptyset\},\{\{y\}\}\}$. In particular, the Wiener version of the ordered pair (x, y) is a set which includes the object $\{\{y\}\}$ as an element; in contrast, the Kuratowski version is a set which does *not* have $\{\{y\}\}$ as an element. It follows from the Fundamental Principle of Set Theory that these objects (i.e., sets) cannot be equal, so the definitions themselves, while superficially similar, are fundamentally different. More precisely, one cannot take, say, the Kuratowski definition of the ordered pair (x, y) and then prove that (x, y) is also given by the Wiener definition. (Compare this situation with the discussion above on defining even and odd integers.) Instead, the purpose of such constructions in modern mathematics is to provide a precise construction, purely in terms of sets, of concrete objects whose set-theoretic properties correspond precisely to the properties our understanding of the given intuitive concept.

Side Comment (on Kuratowski ordered pairs): The symbols used to formulate the preceding definition, when combined with the standard (x, y) notation, may make it appear that almost nothing has really happened. However, this perception changes when one tries to use this definition 'with one's eyes closed'. Indeed, the true test of whether the Kuratowski definition 'works', i.e., whether it does accurately encode the intuitive idea of 'ordered pair' in a purely set-theoretic way, is this: given any object Z, one ought to be able to tell, using set-theoretic ideas alone, whether Z is an ordered pair in the sense of Kuratowski; if it is, one ought to be able to determine the original objects from which the ordered pair Z is formed, and also to determine which of these objects is to be the 'first'. This determination should involve only set-theoretic ideas, and not involve spatial notions such as 'the object on the left', or temporal notions such as 'the first object mentioned'. Here is how the Kuratowski formulation carries this out:

Step 1: Given an object Z, determine whether it is a set. If it is not a set, then it cannot be an ordered pair, so stop. Otherwise:

Step 2: If Z is a set, determine whether it is either a singleton or a doubleton set. If Z is neither a singleton set nor a doubleton set, then (Remark 2 above) it cannot be an ordered pair, so stop. Otherwise:

Step 3:

(a) If Z is a singleton, determine whether its unique element is itself a singleton set. If it is not, then Z cannot be an ordered pair, so stop. However, if the unique element of Z *is* a singleton set A, let x be the unique element of A. Then this analysis shows that Z is the ordered pair $\{\{x\}\} = \{\{x\}, \{x\}\} = \{\{x\}, \{x, x\}\}$ whose first entry is x and whose second entry is also x. Success!

(b) If Z is a doubleton set, determine whether one of its two elements is a singleton set and the other is a doubleton set. If this is not the case, then Z cannot be an ordered pair, so stop. Otherwise:

Step 4: For convenience, call the element of Z that is a singleton set B, and call the element of Z that is a doubleton set C. If B is not a subset of C, then Z cannot be an ordered pair, so stop. Otherwise:

Step 5: Let x be the unique element of B, and let $D = C \setminus B$. Since B is a subset of C, it is clear that D is obtained by removing x from C. It follows that D is also a singleton set whose unique element does *not* equal x. Now let y be the unique element of D. Then it is clear that $Z = \{\{x\}, \{x, y\}\}$. Success!

Note The preceding analysis may seem overly fussy, but it is forced on us because the object Z may be quite complicated.

Example: Consider the following object:

$$Z = \{\{\{\{\{\{a\},\{a,b\}\}\},\{\{\{a\},\{a,b\}\},c\}\}\},\{\{\{\{\{a\},\{a,b\}\}\},\{\{\{a\},\{a,b\}\},c\}\},\{\{a\},\{a,\{\{b\},\{b,c\}\}\}\}\}\}\}$$

where a, b and c are distinct objects. The reader is encouraged to determine whether Z is an ordered pair in the Kuratowski sense.

It is fairly straight forward to generalize the Kuratowski construction to define 'ordered triples', 'ordered quadruples', and so on. However, the results become increasingly complicated. Instead we handle such extensions in an alternate way in Section (I.5). At that time it also will be convenient to provide a 'new, improved' formulation for ordered pairs. The formulation there will ultimately be based on the Kuratowski construction given above, but from then on the Kuratowski ordered pair will be treated as the 'primitive' version. This process of evolving from a 'primitive definition' of ordered pair to an 'improved definition' should sound familiar: a similar evolution process was described in 'Warning (b)' above for 'Definitions'.

One of the advantages of formulating concepts in terms of sets is that one has a precise way of determining when two sets are equal to each other; namely, when they have exactly the same elements (Axiom of Extension). In particular, the Kuratowski definition gives actual content to the following statement:

B.1.2 Theorem

A necessary and sufficient condition for ordered pairs Z and W to be equal is that the first entry of Z equal the first entry of W and the second entry of Z equal the second entry of W.

Proof:

Let a be the first entry of the ordered pair Z and let b be the second entry of Z. Likewise, let c and d be the first and second entries, respectively, of W. Then, in accordance with Kuratowski's definition, the statement to be proved is this:

A necessary and sufficient condition for $\{\{a\}, \{a, b\}\} = \{\{c\}, \{c, d\}\}$ is that $a = c$ and $b = d$.

'Sufficient' Half Suppose that $a = c$ and $b = d$. Then $\{a\} = \{c\}$ and $\{a, b\} = \{c, d\}$, so $\{\{a\}, \{a, b\}\} = \{\{c\}, \{c, d\}\}$. (We have simply replaced 'equals by equals'.) That is, $Z = W$, as claimed.

'Necessary' Half: Suppose that $Z = W$. Then, by the Kuratowski definition of 'ordered pairs', this can be written

$$\{\{a\},\{a,b\}\} = \{\{c\},\{c,d\}\} \quad (*)$$

Case 1: Suppose that $a = b$, so that (a,b) is an ordered pair of Type I (see Definition (B.1.1) above). Then $\{\{a\},\{a,b\}\} = \{\{a\}\}$; in particular, the left side of Equation $(*)$ is a singleton set. Thus, from the 'Axiom of Extension', it follows that the right side of Equation $(*)$ must also be a singleton. That is, (c,d) is also an ordered pair of Type I, so $c = d$. Thus,

$$\{\{a\}\} = \{\{c\}\}.$$

Apply the Axiom of Extension to the singleton sets $\{\{a\}\}$ and $\{\{c\}\}$ to get $\{a\} = \{c\}$; Apply the axiom again to the singleton sets $\{a\}$ and $\{c\}$ to get $a = c$. Since in this case one also has $a = b$ and $c = d$, it follows that $a = c$ and $b = d$, as claimed.

Case 2: Suppose $a \neq b$. Then $Z = (a,b)$ is a Type II ordered pair, hence $W = (c,d)$ is also of Type II (since, by hypothesis, $Z = W$), and thus $c \neq d$. The equality $Z = W$ also implies that the element of Z which is a singleton set must equal the element of W which is a singleton set. That is, $\{a\} = \{c\}$, so by the Axiom of Extension again one has $a = c$. Likewise, the element of Z which is a doubleton set must equal the element of W which is a doubleton set, so $\{a,b\} = \{c,d\}$. Since, as was just shown, one has $a = c$, one can then write $\{a,b\} = \{a,d\}$; just replace c by a in the set $\{c,d\}$. The fact that $\{a,b\} = \{a,d\}$ means that b must be one of the elements of the doubleton set $\{a,d\}$. Since we are assuming $a \neq b$, it follows that $b = d$, as required, and the desired result follows.

In analysis, as elsewhere in mathematics, one encounters sequences which are constructed in a step-by-step manner, using the early entries to determine the later ones. For example, in the definition of the quantity $k!$, where k is a natural number, one first sets $1! = 1$, and then $(k+1)! = k!(k+1)$. Such definitions are said to be **recursive**. Likewise, the construction of the canonical bijection Ψ_C in Definition (**??**) is given recursively. The main theoretical fact about recursive definitions is the following result.

B.1.3 Theorem (Dedekind's Theorem on Recursive Definitions)

Let Y be a nonempty set, let y_0 be a point of Y, and let $G : Y \rightarrow Y$ be a function with domain Y and values in Y. Then the exists a unique function $f : \mathbb{N} \rightarrow Y$ such that $f(1) = y_0$ and $f(k+1) = G(f(k))$ for all k in $\mathbb{N}$.

Proof The 'uniqueness' portion is a simple consequence of the Principle of Mathematical Induction, and its proof is left as an exercise. To show that such f exists, we use Theorem (I.6.6) to reduce the problem to the existence of analogous functions on the subsets $\mathbb{N}_k$ of $\mathbb{N}$.

Claim For each k in $\mathbb{N}$ there is a function $f_k : \mathbb{N}_k \rightarrow Y$ such that $f(1) = y_0$ and if $j \in \mathbb{N}_k$ and $j < k$ then $f_k(j+1) = G(f_k(j))$.

Proof of Claim Let A be the set of k in $\mathbb{N}$ such that such f_k exists. Clearly $1 \in A$: just define $f_1 : \mathbb{N}_1 \rightarrow Y$ by the rule $f_1(1) = y_0$.

Next, suppose that $k \in A$. Define $f_{k+1} : \mathbb{N}_{k+1} \rightarrow Y$ by the rule

$$f_{k+1}(j) = f_k(j) \text{ if } 1 \leq j \leq k; f_{k+1}(k+1) = G(f_k(k)).$$

It is clear that f_{k+1} has the desired properties, so that $k+1$ is also in A. Now the Principle of Mathematical Induction implies that $A = \mathbb{N}$, so the Claim follows.

It is also clear from the construction above that if i and j are elements of $\mathbb{N}$ with $i < j$, then f_i is the restriction to $\mathbb{N}_i$ of f_j. Hence it follows from Theorem (I.6.6), the 'Union-of-Functions Theorem', that the union of the functions f_k for k in $\mathbb{N}$ is a function $f : \mathbb{N} \to Y$; it is also clear that this function has the desired properties.

B.1.4 Example

The statement of Dedekind's Theorem above is phrased along the lines of the original Principle of Mathematical Induction. There is an alternate version which follows the phrasing of the Strong Principle of Mathematical Induction ((**??**)).

B.1.5 Theorem (Strong Form of Dedekind's Theorem on Recursive Definitions)

Let Y be a nonempty set, let y_0 be a point of Y, and let $\mathcal{B}$ be the set of all tuples in Y. Let $g : \mathcal{B} \to Y$ be a function with domain $\mathcal{B}$ and values in Y, and let $G : \mathcal{B} \to \mathcal{B}$ be the function defined by the following rule:

If $\sigma = (y_1, y_2, \dots y_k)$ is an element of $\mathcal{B}$, then $G(\sigma)$ is the tuple $(\sigma, g(\sigma))$; more precisely,

$$G((y_1, y_2, \dots y_k)) = (y_1, x_2, \dots y_k, g((y_1, y_2, \dots y_k))).$$

Then the exists a unique function $f : \mathbb{N} \to Y$ such that $f(1) = y_0$ and $f(k{+}1) = g((f(1), f(2), \dots f(k))$ for all k in $\mathbb{N}$.

The simple proof is left as an exercise. (Hint: Apply the original form of the Dedekind Theorem to the map $G : \mathcal{B} \to \mathcal{B}$.)

In *This Textbook* we have many occasions to construct sequences of objects. Often a proper treatment of these constructions should involve one or the other of the preceding Dedekind Theorems. In most of these situations, however, we leave the details of such a proper treatment to the reader, and follow a more informal approach in order to simplify the discussion.

> Side Comment (on the Schröder-Bernstein Theorem): The next result is not needed for the rest of *This Textbook* – which is why it appears in a Side Comment– but it is a powerful tool in set theory.
>
> **The Schröder-Bernstein Theorem**
>
> Suppose that A and B are nonempty sets such that A has a subset X with the same cardinality as B, and B has a subset Y with the same cardinality as A. Then A has the same cardinality as B.
>
> **Proof** Let $f : A \to Y$ and $g : B \to X$ and be bijections; such bijections exist because of the 'equal cardinality' hypotheses. The key idea is given in the following analysis:
>
> The Basic Schröder-Bernstein Construction
> Let $A' = g[Y]$. Clearly A' is a subset of X, and g maps Y one-to-one onto A'; that is, the restriction of g to Y is a bijection of Y onto A'. It follows that the composition $g{\circ}f$ maps A bijectively onto A'. Likewise, let $B' = f[X]$. Then by a similar argument one sees that the composition $f{\circ}g$ maps B bijectively onto B'
>
> Now let $C = A\backslash X$ and $D = X\backslash A'$; note that C and D are mutually disjoint. Since $A' \subseteq X \subseteq A$, it is also clear that $A = A' \cup C \cup D$, a disjoint union. Likewise, let $E = B\backslash Y$ and $F = Y\backslash B'$. Then it is clear that B is the disjoint union $B = B' \cup E \cup F$ of the subsets B', E and F.

It is also easy to see that the sets $C = A\backslash X$ and $F = Y\backslash B'$ have the same cardinality. More precisely, the function f maps A one-to-one onto Y and f maps X one-to-one onto B', so f maps $A\backslash X$ one-to-one onto $Y\backslash B'$. Similarly, the function g maps $E = B\backslash Y$ bijectively onto $D = X\backslash A'$.

Finally, notice that f maps A' into $B' = f[X]$ since $A' \subseteq X$; likewise, g maps B' into A'. In addition, the restrictions of f to A' and g to B' are one-to-one functions, since the original f and g are one-to-one. In particular, A' has a subset X' with the same cardinality as B', namely $X' = f[A']$; likewise, B' has a subset Y' with the same cardinality as A', namely $Y' = g[B']$. Thus, the 'Basic Construction' can be repeated on the sets A', B', X' and Y' to express A' and B' as disjoint unions of the form $A' = A'' \cup C' \cup D'$ and $B' = B'' \cup E' \cup F'$.

The rest of the proof consists primarily of carrying out the 'Basic Construction' above infinitely many times. To do this, however, it is convenient to introduce numerical subscripts instead of 'primes'. Thus, let us write A_0, X_0, and A_1 instead of A, X, A'; and write $C_0 = A_0\backslash X_0$ and $D_0 = X_0\backslash A_1$ instead of $C = A\backslash X$ and $D = X\backslash A'$. Likewise, write B_0, Y_0 and B_1 instead of B, Y and B'; and write E_0 and F_0 instead of E and F. And when one repeats the 'Basic Construction', simply increase the indices by 1 instead of slapping on another 'prime'.

By doing this, for each $k = 0, 1, 2, \ldots$ one obtains sets A_k, X_k, C_k, D_k, and B_k, Y_k, E_k and F_k, which satisfy the following relations:

(a) A_0, B_0, X_0 and Y_0 equal the original sets A, B, X and Y.

(b) $Y_k = f[A_k]$ and $X_k = g[B_k]$.

(c) $C_k = A_k\backslash X_k$ and $D_k = X_k\backslash A_{k+1}$; likewise, $E_k = B_k\backslash Y_k$ and $F_k = Y_k\backslash B_{k+1}$.

(d) A_k is the disjoint union of A_{k+1}, C_k and D_k; likewise, B_k is the disjoint union of B_{k+1}, E_k and F_k.

(e) C_k and F_k have the same cardinality; likewise, D_k and E_k have the same cardinality.

From the preceding one then can write

$$A = A_0 = A_1 \cup C_0 \cup D_0 = A_2 \cup C_1 \cup D_1 \cup C_0 \cup D_0 = A_3 \cup C_2 \cup D_2 \cup C_1 \cup D_1 \cup C_0 \cup D_0 = \ldots.$$

Likewise,

$$B = B_0 = B_1 \cup E_0 \cup F_0 = B_2 \cup E_1 \cup F_1 \cup E_0 \cup F_0 = B_3 \cup E_2 \cup F_2 \cup E_1 \cup F_1 \cup E_0 \cup F_0 = \ldots.$$

All the unions shown here are *disjoint* unions.

Since $A_{k+1} \subseteq A_k$ for each k, one can conclude that

$$A = P \cup C_0 \cup D_0 \cup C_1 \cup D_1 \cup \ldots \cup C_k \cup D_k \cup \ldots \qquad (*)$$

where $P = A_0 \cap A_1 \cap \ldots \cap A_k \cap \ldots$. Likewise,

$$B = Q \cup E_0 \cup F_0 \cup E_1 \cup F_1 \cup \ldots \cup E_k \cup F_k \cup \ldots \qquad (**)$$

where $Q = B_0 \cap B_1 \cap \ldots \cap B_k \cap \ldots$.

Claim The sets P and Q have the same cardinality.

Proof of Claim First note that $A_{k+1} \subseteq X_k \subseteq A_k$, so $P = X_1 \cap X_2 \cap \ldots \cap X_k \cap \ldots$. Likewise, $B_{k+1} \subseteq Y_k \subseteq B_k$, so $Q = Y_1 \cap Y_2 \cap \ldots \cap Y_k \cap \ldots$.

Suppose first that P is nonempty, and let x be an element of P. Then for each k, $x \in A_k$, hence $f(x) \in Y_k$ (since f maps A_k to Y_k). Thus, $f(x)$ is in the intersection of the sets $Y_1, Y_2, \ldots$, hence $f(x) \in Q$. That is, f maps P *into* Q. To see that f maps P *onto* Q, let y be any element of Q. Then for each k, one has $y \in Y_k$; hence, since $Y_k = f[A_k]$, one has $f(x_k)$ for some element x_k in A_k. And since f is a bijection on X, one has $x_1 = x_2 = \ldots$. Call this common value x, so $x \in A_k$ for each k and thus $x \in P$. That is f maps P *onto* Q. It follows that f maps P bijectively onto Q, hence P and Q have the same cardinality, as claimed.

It is easy to verify by a similar argument that if $P = \emptyset$ then $Q = \emptyset$; thus P and Q have the same cardinality in this case as well.

Now consider the right sides of Equations $(*)$ and $(**)$ above, which express $A = A_0$ and $B = B_0$ as disjoint unions of sets. We have already proved that for each k the sets C_k and F_k have the same cardinality, as do the sets D_k and E_k. And since P and Q also have the same cardinality, Theorem (**??**) now implies that A and B have the same cardinality, as claimed.

The reader is encouraged to review the examples considered in Sections (I.7), (I.8) and (**??**) and see whether some of them could have been obtained more easily using the Schröder-Bernstein theorem.

Remark The reader is encouraged to review the examples considered in Sections (I.7), (I.8) and (**??**) and see how the results obtained there could have been proved using the Schröder-Bernstein theorem.

Problem Precisely define the concept of 'relation'.

As a starting point, consider the definition of 'relation' given in one dictionary:

'A relation is a logical or natural association between two or more things.' (Of course in the special case of a *binary* relation, the phrase 'or more' would be omitted.)

That is, a relation is defined to be a type of association. However, the same dictionary goes on to define 'association' as follows:

'An association is a mental connection or relation between thoughts, feelings, ideas, or sensations.'

One can combine these to say, in effect, that

'A relations is an association, and an association is a relation'.

This is a 'circular definition'. It appears that the concept of 'relation' must fall under Justice Stewart's dictum:

'I can't define it, but I know it when I see it'.

Nevertheless, mathematicians have formulated a precise definition, based purely on set-theoretic concepts; this definition is given below. In principle it could be given here, with no further introduction, since all the words which appear in it already make sense. However,as was observed above in the discussion of the 'Definition-Theorem-Proof' style of mathematical exposition, in reality formal definitions arise only *after* one understands what the relevant features of the situation being described. One normally gets that understanding through a supply of examples which indicate what features any reasonable formal definition ought to have. Here is a short list of such examples of relations, both from within mathematics and from ordinary life:

(1) 'John is a son of Jane.' (Relation: 'is a son of')

(2) 'Jane is the mother of Jill.' (Relation: is the mother of')

(3) 'John is a son of Jane and George.' (Relation: is a son of')

(4) 'The number x is greater than the number y. (Relation: 'is greater than')

(5) 'The number y is greater than the number x'. (Relation: 'is greater than')

(6) 'The number x is the square of the number y.' (Relation: 'is the square of')

These examples indicate some features that any reasonable definition of the concept 'relation' ought to have, where by 'reasonable' is meant a definition which corresponds well to one's intuitive notion of the concept.

(a) A 'relation' should relate two or more objects; for instance, Example (3) relates three objects (John, Jane and George), while the other examples relate pairs of objects. In *This Textbook* the focus is on relations of the latter type, called *binary* relations. (This is partially because such relations are the most common in mathematics, and partially because more general relations can often be reduced to the binary case.)

(b) The order in which the objects are listed may make a difference. For instance, in Example (2) the statements 'Jane is the mother of Jill' and 'Jill is the mother of Jane' both make grammatical sense, but they mean very different things. Likewise, the only difference between Statements (4) and (5) is the order in which the numbers x and y are listed. If $x \neq y$ then the order makes a difference in the truth of the stated relationship.

(c) The sets from which the objects being related are drawn must be specified, and these sets need not be the same. For instance, in Example (1) the context suggests that, in the relation 'x is a son of y', the object x ('John' here) should be chosen from, say, the set of all human males; in contrast, it is not so clear whether the y should be drawn fron the set of, say, all humans, or all human females, or all human parents. Likewise, in Example (6) it makes a difference whether the numbers x and y are to be chosen from the natural numbers or from the real numbers.

(d) Of course the real question for any such binary relation is this: which pairs are in the given relationship? (Once that is answered, the issue of which pairs are *not* in the relationship is automatically settled.)

(e) The definition must allow for the possibility that there are no pairs in the given relationship. For instance, the relation described by the inequality $x^2 < -y^2$' is satisfied by no pair of real numbers.

Having said all that, here is the standard definition of 'binary relation' in mathematics:

B.1.6 Definition

(1) Let X and Y be a pair of nonempty sets. A **binary relation from *X* to *Y*** is a (possibly empty) subset of the Cartesian product $X \times Y$. If R is such a relation, i.e., if $R \subseteq X \times Y$, and if (x, y) is an element of R, one says that ***x* and *y* are in the relation** R. One frequently indicates this state of affairs by writing $x\,R\,y$ instead of the more usual $(x, y) \in R$.

If $Y = X$, then one normally speaks of a **relation on *X*** instead of a relation from X to X.

(2) If R is a binary relation from X to Y, as above, then the corresponding **negated binary relation** from X to Y, often denoted $\not R$, is the complement $X \times Y \setminus R$; that is, it is the set of all ordered pairs (x, y), with x in X and y in Y, such that (x, y) is *not* in the set R.

(3) A **formal binary relation** is an ordered pair (U, R), where U is an ordered pair (X, Y) of nonempty sets and R is a subset of $X \times Y$.

B.1.7 Some Nonmathematical Examples of Binary Relations

(1) Let W be the set consisting of the following women; the numbers following their names are their ages as of January 1, 2000:

Abigail (85), Betty (65), Carrie (22), Debbie (45), Edie (34) and Francine (42)

Here are the familial relations between these women:

Betty is the daughter of Abigail; Carrie is the daughter of Francine; Debbie is the daughter of Betty.

Problem Let $X = Y = W$, and consider the binary relation on W, 'is the daughter of'. Describe this relation in terms of the preceding definition; that is, as a set of ordered pairs.

Solution To save a little writing, abbreviate the names of the six women to their first initials, so that $W = \{A, B, C, D, E, F\}$. Then the relation is the following set R of ordered pairs:

$$R = \{\text{(B,A), (C,F), (D,B)}\}.$$

(2) Let W be the same set as in the preceding example.

Problem As before, let $X = Y = W$, but now consider the binary relation, again on W, described by 'is exactly twenty years younger than' (on January 1, 2000, of course). Describe this relation in terms of a set of ordered pairs.

Solution It is easy to see that the answer is exactly the same set R as in the preceding example.

(3) Let $\hat{W} = W \setminus \{E\} = \{A, B, C, D, F\}$, and once again consider the binary relation described by 'is the daughter of', but now on the set $\hat{W}$. It is clear that the corresponding set of ordered pairs is the same as in the preceding examples. Nevertheless, in light of Part (3) of Definition (B.1.6) they are considered to be different. Indeed, as 'formal' binary relations the first is the ordered pair $((W, W), R)$ while the second is $(\hat{W}, \hat{W}, R)$. Since $W \neq \hat{W}$, these two ordered pairs are physically different objects.

Remarks (a) The relations discussed in Examples (1) amd (2) above are, intuitively speaking, quite dissimilar: one concerns the specific relation of daughters to mothers, while the other concerns age differences. Nevertheless, from the viewpoint of Definition (B.1.6) they are literally the same object.

In contrast, the binary relation described in Example (3) appears to be the same as that in Example (1), even when speaking intuitively. Nevertheless, in terms of Definition (B.1.6) they must be considered to be different, because the underlying sets W and $\hat{W}$ are not equal. If this seems strange, consider the alternative: if these relations were literally the same, then the corresponding negated relations would be equal. However, the set of ordered pairs associated with the *negated* relation for Example (1) includes, among several others, the pair (E, A): Edie is not the daughter of Abigail. In contrast, the pair (E, A) cannot be part of the negated relation for Example (3), since E is not in the set W'.

B.1.8 Some Mathematical Examples of Binary Relations

(1) Some of the standard well-known examples of mathematical binary relations have been alluded to before: 'greater than' 'less than', 'equal to' and so on. To formulate these precisely in the terms of Definition (B.1.6), one needs to specify the sets which pay the role of X and Y in that definition. For instance, the relation 'is greater than', denoted '$>$', certainly applies when $X = Y = \mathbb{R}$, the set of all real numbers. In this case, the corresponding set of ordered pairs is $R = \{(x, y) : x \in \mathbb{R}, y \in \mathbb{R}, x > y\}$.

The corresponding negated relation, denoted $\not>$ and pronounced 'not greater than', consists of those elements (x, y) in $\mathbb{R} \times \mathbb{R}$ for which $x \not> y$. This is equivalent to the relation 'less than, or equal to', denoted '$\leq$' on the same set.

(2) Let $X = \{1, 2, 3, 4, 5\}$ and let R be the binary relation on X decribed as the set of all ordered pairs (x, y) in $X \times X$ such that $x \neq y$ and $x - y$ is a whole-number multiple of 3. One easily verifies that $R = \{(4, 1), (5, 2)\}$.

(3) Let X be as in the preceding example but now let R' be the binary relation on X decribed as the set of all ordered pairs (x, y) in $X \times X$ such that $x - y$ is a whole-number multiple of 6; that is, $x - y = 6\,k$ for some k in $\mathbb{N}$. Clearly, $R' = \emptyset$.

(4) Let $\hat{X} = \{4, 5, 6, 7, 8\}$ and let R'' be described as the set of all ordered pairs (x, y) in $\hat{X} \times \hat{X}$ such that $x - y$ is a whole-number multiple of 10 Clearly $R'' = \emptyset$, so R'' equals the corresponding set R' from the preceding example. Nevertheless, the binary relations R' and R'' are treated as being different: the sets X and $\hat{X}$ on which these relations are defined are not the same.

Remarks (1) The phrasing 'relation *from* X *to* Y' emphasizes the requirement, already mentioned above, that the definition should allow the possibility that the order (X is first, Y is second) in a relation might make a difference. Indeed, the phrases 'from X to Y' and 'from Y to X' certainly do not signify the same ideas. However, some authors use instead the phrasing 'relation of X *with* Y'. Linguistically speaking, this means the same as 'relation of Y with X', which hides the order dependence. Such authors rely on the convention that by writing $X \times Y$, as opposed to $Y \times X$, one tacitly treats X as the first set, Y as the second. (See the Side Comment below on the 'left-to-right bias' in mathematics.) This usage is mildly sloppy, but seems to cause no confusion.

(2) Some authors abbreviate the content of Definition (B.1.6) to the following:

'A binary relation is a set R of ordered pairs of objects'

In this formulation there is no reference to any given sets from which the entries of the ordered pairs in R are to be drawn. Stated this way, the binary relation *is* the set R itself. With that viewpoint, it would follow that if $R \subseteq X \times Y$, and X' and Y' are proper supersets of X and Y, respectively, then the same object, R, would be a binary relation from X to Y and, simultaneously, a binary relation from X' to Y'. This is *not* the intent of Definition (B.1.6), as is made clear in the examples.

(3) The restriction that X and Y be nonempty is to avoid situations which are of no interest whatsoever. In contrast, the definition *does* allow the possibility that the subset R might be empty, since that fact might reflect something of interest; namely, that no pairs (x, y) in $X \times Y$ happen to be in the given relation.

(4) The preceding definition uses only subsets of the Cartesian product $X \times Y$, and the latter concept is defined purely in set-theoretic terms. Thus, Definition (B.1.6) does formulate the concept of 'relation' purely in terms of set theory. However, by identifying a relation from X to Y with a subset of $X \times Y$, the definition does allow the possibility of two relations, which intuitively to be seem very different, being treated as the same; see below.

One of the most important types of binary relations on a set is a so-called 'equivalence relation'.

B.1.9 Definition

Let X be a nonempty set. An **equivalence relation on** X is a subset W of $X \times X$ with the following properties:

(i) (Reflexivity) If $x \in X$ then $(x, x) \in W$.

(ii) (Symmetry) If $(x, y) \in W$ then $(y, x) \in W$.

(iii)(Transitivity) If $(x, y) \in W$ and $(y, z) \in W$ then $(x, z) \in W$.

One often indicates that $(x, y) \in W$ by the notation $x \overset{W}{\sim} y$; the expression $x \overset{W}{\sim} y$ is pronounced 'x is equivalent to y with respect to W'. If the choice of equivalence relation W is understood from the context, then this notation may be simplified to $x \sim y$, which is pronounced 'x is equivalent to y'; in this situation we may also refer to 'the equivalence relation $\sim$'.

If x is any element of X then $[x]_W$ denotes the set of all y in X such that $x \overset{W}{\sim} y$. The set $[x]_W$ is called the **equivalence class of x relative to** W; as usual, the explicit reference to the equivalence relation W may be dropped if it is clear from the context.

B.1.10 Examples

(1) The relation of 'equality' is an equivalence relation on any nonempty set. For this equivalence relation, the equivalence class $[x]$ of x is simply the singleton set $\{x\}$.

(2) Suppose that $\mathcal{F}$ is a partition of a nonempty set X. Then $\mathcal{F}$ determines an equivalence relation $W_{\mathcal{F}}$ on X by the rule that $x \overset{W_{\mathcal{F}}}{\sim} y$ if, and only if, x and y lie in the same element of the family $\mathcal{F}$. Indeed, suppose that $x \in X$. Then the fact that $\bigcup \mathcal{F} = X$ implies that there is an element S in the family $\mathcal{F}$ such that $x \in S$. Clearly x and x are in S. Likewise, if x and y are both in the same element S of $\mathcal{F}$ then y and x are both in S. Finally, suppose that x and y are both in the same element S of $\mathcal{F}$ and that y and z are both in the same element T of $\mathcal{F}$. Then S and T have an element in common, namely y, so $S \cap T \neq \emptyset$. By the fact that $\mathcal{F}$ is a partition it follows that $S = T$, and thus x and z are both in T.

It is clear that if $x \in X$ then the equivalence class $[x]$ is precisely the unique set A in the partition $\mathcal{F}$ such that $x \in A$.

(3) Suppose that $f : X \to Y$ is a surjection of a set X onto a set Y. Then f determines an equivalence relation W on X by the rule $x \overset{W}{\sim} y$ if, and only if, $f(x) = f(y)$. We refer to this as the **equivalence relation on X determined by f**, and we denote it by W_f. We also sometimes write $x \sim_f y$ instead of the more proper $x \overset{W_f}{\sim} y$.

The next result summarizes the relations between these concepts.

B.1.11 Theorem

Let X be a nonempty set.

(1) A subset W of $X \times X$ is an equivalence relation on X if, and only if, there exists a partition $\mathcal{F}$ on X such that $W = W_{\mathcal{F}}$, where $W_{\mathcal{F}}$ is described above. The partition $\mathcal{F}$ is unique; that is, if $\mathcal{F}$ and $\mathcal{G}$ are partitions of X such that $W_{\mathcal{F}} = W_{\mathcal{G}}$, then $\mathcal{F} = \mathcal{G}$.

(2) A subset W of $X \times X$ is an equivalence relation on X if, and only if, there exists a surjection $f : X \to Y$ of X onto some set Y such that such that $W = W_f$. (The set Y is *not* at all unique, and therefore neither is the map f.)

Proof: (1) The 'if' portion is essentially the content of Example (2) above. The rest of the proof is left as a straight-forward exercise.

(2) This follows from Part (1) together with the results of Theorem (**??**).

The next result says, in effect, that an equivalence relation on a set X determines corresponding equivalence relations on the nonempty subsets of X. Likewise, a partition on X determines corresponding partitions on the nonempty subsets of X.

B.1.12 Theorem

Let X be a nonempty set, and let Y be a nonempty subset of X.

(a) Suppose that W is an equivalence on X, and let $W|_Y$ denote the set $W \cap (Y \times Y)$. Then $W|_Y$ is an equivalence relation on Y.

(b) Let $\mathcal{F}$ be a partition of X, and let $\mathcal{F}_Y$ denote the family of all nonempty sets of the form $S \cap Y$, where S is in the family $\mathcal{F}$. Then $\mathcal{F}_Y$ is a partition of Y.

(c) Suppose that W is an equivalence relation on X, and let $\mathcal{F}$ be the partition of X such that $W = W_{\mathcal{F}}$ (see Part (a) of Theorem (B.1.11)). Then $W|_Y = W_{\mathcal{F}_Y}$.

Proof

(a) Suppose that $y{\in}Y$. Then $y{\in}X$ (since Y is a subset of X), hence $(y,y){\in}W$ (since W satisfies the Reflexivity condition). But $(y,y){\in}Y{\times}Y$ (by definition of Cartesian Product), so $(y,y){\in}W \cap Y{\times}Y$ (by definition of intersection). In particular, W_Y satisfies the Reflexivity condition on Y. Similar arguments can be used to show that W_Y also satisfies the Symmetry and Transitivity conditions on Y.

(b) First note that the family $\mathcal{F}_Y$ is nonempty. Indeed, let y be an element of Y; such y exists because Y is assumed to be nonempty. Then (by definition of 'partition') there exists a unique set S in the family $\mathcal{F}$ such that $y{\in}S$. Thus y is in both S and Y, so $y{\in}(S \cap Y)$, hence the intersection $S \cap Y$ is nonempty. Thus $S \cap Y$ is an element of $\mathcal{F}_Y$, which implies that the family $\mathcal{F}_Y$ is nonempty.

Next, consider a pair of sets U and V in the family $\mathcal{F}_Y$, with $U \neq V$. Then, by definition of this family, there exists elements S and T of $\mathcal{F}$ such that $U = S \cap Y$ and $V = T \cap Y$. Since $U \neq V$, it follows that $S \neq T$. Thus $S \cap T = \emptyset$, by the 'Disjointness Property' of partitions. It follows that

$$U \cap V = (S \cap Y) \cap (T \cap Y) = (S \cap T) \cap (Y \cap Y) = \emptyset \cap Y = \emptyset,$$

by basic properties of 'intersection'. That is, the family $\mathcal{F}_Y$ satisfies the 'Disjointness Property' as well.

Finally, it follows from the fact that $X = \bigcup_{S\in\mathcal{F}} S$, the definition of $\mathcal{F}_Y$, and basic properties of 'union' and 'intersection', that

$$Y = X \cap Y = \left(\bigcup_{S\in\mathcal{F}} S\right) \cap Y = \bigcup_{S\in\mathcal{F}} (S \cap Y) = \bigcup_{U\in\mathcal{F}_Y} U.$$

(In the final equation any sets of the form $S \cap Y$ for which this intersection is empty can be ignored, since they do not contribute to the union.) Thus, $\mathcal{F}_Y$ satisfies the 'Union Property' for partitions.

(c) The simple proof is left to the reader as an exercise.

B.1.13 Definition

Let X be a nonempty set and let Y be a nonempty subset of X.

(a) If W is an equivalence relation on X, then the **equivalence relation on Y induced from W** is the equivalence relation $W|_Y = Y \cap (Y{\times}Y)$ discussed in the preceding theorem.

(b) If $\mathcal{F}$ is a partition of X, then the **partition of Y induced from $\mathcal{F}$** is the partition $\mathcal{F}_Y$ described in the preceding theorem.

B.1.14 Definition

Let W be an equivalence relation on a nonempty set X.

(1) For each x in X the set $W[x] = \{y{\in}X : (x,y){\in}W\})$ is called the **equivalence class of x with respect to W**. (If the equivalence relation W under consideration is clear from the context, we may abbreviate the notation $W[x]$ to $[x]$.) If S is an equivalence class for the equivalence relation W then each element of S is called a **representative of S** (relative to the given equivalence relation). (The element x 'represents' the equivalence class S in the sense that

one can write $S = [x]$; but of course the choice of x used to represent S this way is not unique unless S happens to be a singleton set.)

(2) The set whose elements are the equivalence sets of W is called the **quotient set of X with respect to W**; this set is denoted X/W.

Note: It is clear that the quotient set X/W is the same as the partition $\mathcal{F}$ described in Part (a) of Theorem (B.1.11). It is also clear that $X/W = X/(f)$, where f is the map described in Part (b) of the same theorem.

The next result ties together much of what we have just seen.

B.1.15 Theorem

Suppose that $f : X \to Y$ is a surjection of a nonempty set X onto a set Y, and let $Z = X/(f)$ be the corresponding quotient set. Define a function $\hat{f} : Z \to Y$ as follows: if S is an element of Z, so $S = f^{-1}[y]$ for some element y in Y, set $\hat{f}(S) = y$. Then the function $\hat{f}$ is a bijection from Z onto Y.

The simple proof is left as an exercise.

B.1.16 Definition

The bijection $\hat{f}$ described in the preceding theorem is called the **canonical bijection from W onto Y determined by the bijection** $f : X \to Y$.

B.2 Construction of $\mathbb{Q}$ from $\mathbb{N}$

Introduction In Chapter (I) we accepted as 'primitive truths' basic properties of the natural numbers and the rational numbers; but there was a promise to delve more deeply into the structure of such systems. Indeed, in Appendix A the internal structure of the natural numbers is reduced to a consideration of the Dedekind-Peano axioms. Of course, the actual nature of these numbers, that is, the answer to the question 'Exactly what is a natural number in some philosophical sense', is left open: We continue to treat these numbers as 'primitive objects' which we do not define, but 'we know them when we see them'. In particular, we do not prove that there exists a system of objects which satifies the Dedekind-Peano axioms; we accept our intuitive notions of $\mathbb{N}$ as a given.

Having done this, however, it is now easy to derive from this primitive notion of 'Natural Number', in a rigorous manner which uses only these accepted facts about $\mathbb{N}$ and accepted set-theoretic concepts, the rigorous construction of systems which correspond in every important way to our intuitive ideas of 'integer' and 'rational number'. The process lets the intuitive notion lead lead us, in a fairly natural way, to a rigorous construction of $\mathbb{Q}$. The goal of the present section is such a contruction.

Note The analogous questions for the real number system are postponed to Chapter (II).

B.2.1 Construction 1: The Positive Rationals from the Natural Numbers

Let X be the set $\mathbb{N}\times\mathbb{N}$ of all ordered pairs of natural numbers. The set X is 'rigorously defined', in the sense that we are taking both the set $\mathbb{N}$ and the concept of 'ordered pairs' as primitive concepts.. Let Y be the set $\mathbb{Q}^+$ of all positive rational numbers', viewed as an intuitively understood system which eventually must be defined rigorously. In terms of this intuition, there is a well-known surjective map $\rho : X \to \mathbb{Q}^+$ defined by $\rho(j,k) = j/k$ for each ordered pair (j,k) in X; the Greek lettre 'ρ' stands for 'ratio'. It is easy to see that the equivalence relation determined on X by the surjection ρ is this:

$$(j_1,k_1) \sim (j_2,k_2) \text{ if, and only if, } j_1{\cdot}k_2 \,=\, j_2{\cdot}k_1 \qquad (*)$$

This equivalence relation, even though it arises in this discussion from an intitive understanding of 'rational number', which includes the map ρ, is formulated in $(*)$ purely in terms of the primitive notions of ordered pairs and the algebraic system $\mathbb{N}$ as described by the Dedekind-Peano axioms. In particular, there is no mention in Condition $(*)$ of either the 'intuitive' concept of the set $\mathbb{Q}^+$ or the 'intuitive' map ρ. For that reason we use the symbol $\sim$ instead of the more proper $\sim_\rho$ in Condition $(*)$.

Return now to the intuitive discussion of $\mathbb{Q}^+$. Let $A = X/(\rho)$ be the set of equivalence classes associated with the equivalence relation described above, and let $\hat{\rho} : A \to \mathbb{Q}^+$ be the corresponding canonical bijection; see Definition (B.1.16) above. Note also that the inverse map $\hat{\rho}^{-1} : \mathbb{Q}^+ \to A$ is easy to compute: if $r\in\mathbb{Q}^+$, express r in the form j/k as above; then $\hat{\rho}^{-1}(r)$ is the equivalence class with representative (j,k). Note that there is no need at this stage to verify that if r is expressed as j'/k' then $(j',k')\sim(j,k)$, since that is built into the definition of the equivalence relation determined by the surjection ρ.

The bijection $\hat{\rho}$, together with the binary operation of 'addition' already (intuitively) defined on the set $\mathbb{Q}^+$, determines a corresponding operation of 'addition' on A for which the map φ is an isomorphism; see Example (I.10.7) (4). More precisely, suppose that z_1 and z_2 are elements of A, so that $z_1 = \hat{\rho}^{-1}[\{r_1\}]$ and $z_2 = \hat{\rho}^{-1}[\{r_2\}]$ for positive rationals r_1 and r_2. Then define the sum $z_1 + z_2$ by the rule

$$z_1 + z_2 \,=\, \hat{\rho}^{-1}[\{r_1 + r_2\}],$$

where the addition on the right is the (intuitive) addition of positive rational numbers. It is clear from the description of $\hat{\rho}^{-1}$ just given that if $r_1 = j_1/k_1$ and $r_2 = j_2/k_2$, then

$$\hat{\rho}^{-1}(r_1 + r_2) \,=\, \hat{\rho}^{-1}\left(\frac{j_1{\cdot}k_2 + j_2{\cdot}k_1}{k_1{\cdot}k_2}\right),$$

which is the equivalence class with representative $(j_1{\cdot}k_2 + j_2{\cdot}k_1, k_1{\cdot}k_2)$. Likewise, one sees that the bijection ρ, together with the (intuitive) concept of 'multiplication' on $\mathbb{Q}^+$, determines a 'multiplication' on A, given by

$$z_1{\cdot}z_2 \,=\, \hat{\rho}^{-1}[\{r_1{\cdot}r_2\}] \text{ for all } z_1 \text{ and } z_2 \text{ in } A.$$

Using the same notation as above, one sees that $z_1{\cdot}z_2$ is the equivalence class with the representative $(j_1{\cdot}j_2, k_1{\cdot}k_2)$. Furthermore, the 'order' relation on $\mathbb{Q}^+$ likewise determines a corresponding order relation on A:

$$z_1 \,<\, z_2 \text{ if, and only if, } \rho(z_1) \,<\, \rho(z_2).$$

In terms of fractions used above, one has $z_1 < z_2$ if, and only if, $j_1{\cdot}k_2 < j_2{\cdot}k_1$. It is clear that the bijection $\hat{\rho}$ is an isomorphism between the operations of addition and multiplication, just defined on the set A, and the corresponding 'intuitive' operations on $\mathbb{Q}^+$; likewise, the bijection $\hat{\rho}$ is an order-preserving map from A onto $\mathbb{Q}^+$. It follows that the algebraic properties of A are the same as those on $\mathbb{Q}^+$. For example, the 'intuitive' understanding of $\mathbb{Q}^+$ implies that there is a unique element in $\mathbb{Q}^+$, the number 1, such that $r{\cdot}1 = r$ for all r in $\mathbb{Q}^+$. Likewise, for each r in $\mathbb{Q}^+$ there is a unique s in $\mathbb{Q}^+$ such that $r{\cdot}s = 1$. The existence of the isomorphism $\hat{\rho}$ then implies that the analogous properties hold for A.

As the formulas given above for $+$, $\cdot$ and $<$ on A indicate, the definitions of these concepts can also be obtained directly using only the equivalence relation $\sim$, given by $(*)$, and properties of $\mathbb{N}$. That is, by using only the *results* of the intuitive discussion as a new starting point, and eliminating the intuitive discussion itself entirely, one can give a rigorous development of $\mathbb{Q}^+$ directly from the properties of $\mathbb{N}$ (and elementary set theory), as if one had never heard of $\mathbb{Q}^+$ before. Here is how such a rigorous construction would normally be carried out:

Given the set $X = \mathbb{N}{\times}\mathbb{N}$ and the equivalence relation $(\sim)$ described above in Condition $(*)$, let A being the corresponding quotient set $X/\sim$. Define the binary operations of $+$ and $\cdot$ on A, and the order $<$, as follows: Let z_1 and z_2 be elements of A, i.e., equivalence classes associated with $\sim$, and let (j_1, k_1) and (j_2, k_2) in X be representatives of z_1 and z_2, respectively. Define $z_1 + z_2$ to be the equivalence class which is represented by the ordered pair $(j_1{\cdot}k_2 + j_2{\cdot}k_1, k_1{\cdot}k_2)$. Likewise, define $z_1{\cdot}z_2$ to be the equivalence class represented by the ordered pair $(j_1{\cdot}j_2, k_1{\cdot}k_2)$. Finally, define the relation $z_1 < z_2$ to mean $j_1{\cdot}k_2 < j_2{\cdot}k_1$. Of course these definitions might seem somewhat arbitrary without the earlier intuitive discussion using the surjection ρ; but technically speaking they do not refer to that discussion, and thus are ultimately based purely on the rigorously defined $\mathbb{N}{\times}\mathbb{N}$. Using the standard notation $[(j,k)]$ for the equivalence class with repressentative (j,k), one can write

$$[(j_1,k_1)] + [(j_2,k_2)] = [(j_1{\cdot}k_2 + j_2{\cdot}k_1, k_1{\cdot}k_2)], \qquad [(j_1,k_1)]{\cdot}[(j_2,k_2)] = [(j_1{\cdot}j_2, {\cdot}k_1{\cdot}k_2)],$$

$$[(j_1,k_1)] < (j_2,k_2) \text{ if, and only if, } j_1{\cdot}k_2 < j_2{\cdot}k_1.$$

The first need is to check that these constructions are **well defined**, in the sense that they does not depend on the choice of representatives of the classes z_1 and z_2 used. Consider in more detail, for example, the definition of 'multiplication' just given. Suppose that (m_1, n_1) and (m_2, n_2) are also representatives of z_1 and z_2, respectively, so that by $(\sim)$ one has $j_1{\cdot}n_1 = k_1{\cdot}m_1$ and $j_2{\cdot}n_2 = k_2{\cdot}m_2$. Then $(m_1{\cdot}m_2, n_1{\cdot}n_2)$ is equivalent to $((j_1{\cdot}j_2, k_1{\cdot}k_2))$, since $(m_1{\cdot}m_2){\cdot}(k_1{\cdot}k_2) = (n_1{\cdot}n_2{\cdot}(k_1{\cdot}k_2))$. That is, using different representatives for the equivalence classes z_1 and z_2 produce the same equivalence class $z_1{\cdot}z_2$, as required. The proofs that the operation of addition and the order relation are also well-defined are left as exercises.

The set A of equivalence classes just constructed from $\mathbb{N}$, together with the operations $+$ and $\cdot$, and the order relation $<$, can now be considered as our our 'official' definition of the system $\mathbb{Q}^+$ of positive rational numbers, at least for the time being. (Later on we'll 'improve' the construction.)

Notice that system just constructed has a subsystem which behaves just like our original primitive system $\mathbb{N}$ of natural numbers, namely the subset $\mathbb{N}'$ of $\mathbb{Q}^+$ consisting of all equivalence classes z in A of the form $[(j,1)]$ with j in $\mathbb{N}$. It is easy to see that $[(j_1,1)] + [(j_2,1)] = [((j_1+j_2),1)$ and $[(j_1,1)]{\cdot}[(j_2,1)] = [(j_1{\cdot}j_2,1)]$. Likewise, $[(j_1,1)] < [(j_2,1)]$ if, and only if, $j_1 < j_2$ in the original set $\mathbb{N}$. We think of the set $\mathbb{N}'$ so obtained as the 'new and improved natural numbers': 'new' because obviously the set $\mathbb{N}'$ is different from the original set $\mathbb{N}$; 'improved', because these new numbers extend the original algebra of $\mathbb{N}$ to the more inclusive arithmetic of $\mathbb{Q}^+$. It is easy to see that in this arithmetic one has $[(1,1)]{\cdot}[(j,k)] = [(j,k)]$ and $[(k,j)]{\cdot}[(j,k)] = [(k{\cdot}j, k{\cdot}j)] = [(1,1)]$. In particular, $[(1,1)]$ is the multiplicative unit, while $[(k,j)]$ is the multipicative inverse of $[(j,k)]$. By introducing the obvious 'division' operation by $[(j_1,k_1)/(j_2,k_2)] = [(j_1,k_1)]{\cdot}[(k_2,j_2)]$, one gets the formula $[(j,k)] = [(j,1)]/[(k,1)]$. This is the 'new and improved' version of the formula $\rho(j,k) = j/k$ used in the intuitive discussion above.

Note that, by the preceding constructions, a positive rational number is an equivalence class of ordered pairs of the primitive natural numbers. In particular, the 'new and improved' natural numbers are special types of such classes.

B.2.2 Construction 2: The Rationals from the Positive Rationals

The next step is to rigorously extend the system $\mathbb{Q}^+$, just obtained, to the system consisting of *all* the rational numbers, not just the positives. The procedure is, in spirit, similar to – but, in the details, rather different from – that used in Construction (1), so we can be somewhat briefer. Indeed, if we let $\mathbb{Q}$ denote for now the rational numbers of our intuition, there is a natural surjection $\delta : \mathbb{Q}^+ \times \mathbb{Q}^+ \to \mathbb{Q}$ given by

$$\delta(z_1, z_2) = z_1 - z_2.$$

(The symbol δ here stands for 'difference'.) The equivalence relation on $\mathbb{Q}^+ \times \mathbb{Q}^+$ determined by this surjection is

$$(z_1, z_2) \sim (z_1', z_2') \text{ if, and only if, } z_1 + z_2' = z_1' + z_2$$

Let $\mathbb{Q}$ denote the set of equivalence classes arising from this equivalence relation on $\mathbb{Q}^+ \times \mathbb{Q}^+$. It is easy to define the concepts of addition, multiplication and order on the set $\mathbb{Q}$, using only the structures on $\mathbb{Q}^+$, but not the intuitive map Δ. For example, if $s = [(z_1, z_2))]$ and $s' = [(z_1', z_2')]$ are elements of $\mathbb{Q}$, then $s + s' = [(z_1 + z_1', z_2 + z_2')]$. Likewise, $s \cdot s' = [(z_1 \cdot z_1' + z_2 \cdot z_2', z_1 \cdot z_2' + z_2 \cdot z_1')]$. Likewise, $s < s'$ if, and only if, $z_1 + z_2' < z_1' + z_2$. These constructions use only the rigorously defined structure of $\mathbb{Q}^+$.

There is a zero element in $\mathbb{Q}$, namely the equivalence class with representative $(1,1)$, where 1 now denotes the multiplicative identity in $\mathbb{Q}^+$; any representative of the form (c,c) with c in $\mathbb{Q}^+$ works too. One then defined the *positive rationals* to be those elements $[(z_1, z_2)]$ of $\mathbb{Q}$ such that $[(z_1, z_2)] > [(1,1)]$. It is easy to check that this is equivalent to $z_1 > z_2$ in $\mathbb{Q}^+$, and that this set of 'positive rationals' is isomorphic to $\mathbb{Q}^+$ in the usual sense. Likewise, there is a subset of the set $\mathbb{Q}$ just constructed which corresponds to the set $\mathbb{Z}$ of integers.

Of course one needs to verify that these concepts do not depend on choices of representatives of equivalence classes in $\mathbb{Q}^+ \times \mathbb{Q}^+$, and one needs to show that the copy of $\mathbb{Q}^+$ in $\mathbb{Q}$ just described is equivalent to the $\mathbb{Q}^+$ already described. The details are tedious, so the usual custom, gladly followed here, is to leave the details to the reader.

Index