

AA1H Calculus Notes
Math1115, Honours 1
1998

John Hutchinson

Author address:

DEPARTMENT OF MATHEMATICS, SCHOOL OF MATHEMATICAL SCIENCES,
AUSTRALIAN NATIONAL UNIVERSITY
E-mail address: `John.Hutchinson@anu.edu.au`

Contents

Chapter 1. Introduction	v
Chapter 2. The Real Number System	1
2.1. Introduction	1
2.2. Important sets of real numbers	2
2.3. Algebraic and order properties	4
2.4. Completeness property	8
2.5. Properties of the rational and irrationals	10
2.6. Functions	13
2.7. Mathematical induction	17
2.8. ★Fields	18
2.9. ★Deductions from the axioms	18
2.10. ★Existence and Uniqueness of the Real Number System	20
Chapter 3. Limits	23
3.1. Introduction	23
3.2. Definition of limit for a function	23
3.3. Properties of limits of functions	30
3.4. ★Functions of more than one variable	35
Chapter 4. Sequences	37
4.1. Examples of sequences	37
4.2. Limit of sequences	38
4.3. Monotone sequences	40
4.4. Limits for functions via limits for sequences	43
Chapter 5. Continuity	45
5.1. Introduction	45
5.2. Definition of continuity	45
5.3. Properties of continuous functions	47
5.4. Deeper properties of continuous functions	49
5.5. ★Pathology and continuity	53
5.6. ★Uniform continuity	54
5.7. ★Functions of two or more variables	57
Chapter 6. Differentiation	59
6.1. Introduction	59
6.2. The derivative of a function	59
6.3. Computing derivatives	61
6.4. Maximum and minimum values	64
6.5. Mean Value Theorem	65
6.6. ★Partial derivatives	67
Chapter 7. Integration	69

7.1. Introduction	69
7.2. The Riemann integral	69
7.3. Riemann sums	76
7.4. Properties of the Riemann integral	77
7.5. Fundamental Theorem of Calculus	79
Chapter 8. Differential Equations	81
8.1. Outline of proof of the Existence and Uniqueness theorem	83
8.2. ★Rigorous proof of the Existence and Uniqueness theorem	85
Bibliography	91
Index	5

CHAPTER 1

Introduction

The aim of the AA1H course is to give an introduction to modern mathematics. In the process you will prove the major results used in the AA1 course and thereby obtain a more fundamental understanding of that material.

Mathematics is the study of pattern and structure. It is studied both for its internal beauty and for its universal applicability. In mathematics we make certain specific assumptions (or axioms) about the objects we study and then develop the consequences of these assumptions in a precise and careful manner. The axioms are chosen because they are “natural” in some sense; it usually happens that these axioms also describe phenomena in other subjects, in which case the mathematical conclusions we draw will also apply to these phenomena.

Areas of mathematics developed for “mathematical” reasons usually turn out to be applicable to a wide variety of subjects; a spectacular recent example being the applications of differential geometry to understanding the fundamental forces of nature studied in physics, and another being the application of partial differential equations and geometric measure theory to the study of visual perception in biology and robotics. There are countless other examples in engineering, economics, and the physical and biological sciences. On the other hand, the study of these disciplines can usually only be done by applying the techniques and language of mathematics. Mathematics is used as a tool in such investigations. But the study of these subjects can also lead to the development of new fields of mathematics and insights into old fields.

In this course we will study the real number system, the concepts of limit and continuity, differentiability and integrability, and differential equations. While most of these terms will be familiar from high school in a more or less informal setting, we will study them in a much more precise way. This is necessary both for applications and as a basis for generalising these concepts to other mathematical settings.

One very important question we investigate is: when do certain types of differential equations have a solution, when is there exactly one solution, and when is there more than one solution? The solution of this problem uses almost all the material that is developed throughout the course. The study of differential equations is of tremendous importance in mathematics and for its applications. Any phenomena that changes with position and/or time is usually represented by one or more such equations.

The ideas we develop are basic to further developments in mathematics. The concepts we study generalise in many ways, such as to functions of more than one variable, and to functions whose variables are themselves functions (!); all these generalisations are fundamental to further applications.

At the end of the first semester you should have a much better understanding of all these ideas.

These Notes are intended so that you can concentrate on the lectures rather than trying to write everything down. Occasionally there may be lecture material not mentioned in the Notes, in which case I will indicate this precisely, but generally

you should *not* take your own notes. So why come to the lectures? Because the Notes are frequently rather formal (this is a consequence of the precision of mathematics) and it is often very difficult to see the underlying concepts. In the lectures I explain the material in a less formal manner, single out and discuss the key and underlying ideas, and generally explain and discuss the subject in a manner which it is not possible to do efficiently in print. It would be a *very big mistake* to skip lectures. If you still do not believe me, ask students from my previous courses.

Do not think that you have covered any of this material in school; the topics may not appear new, but the material certainly will be. Do the assignments, read the lecture notes *before* class. Mathematics is not a body of isolated facts; each lecture will depend on certain previous material and you will understand the lectures much better if you keep up with the course. In the end this approach will be more efficient as you will gain more from the lectures.

Throughout the course I will make various digressions and additional remarks, marked clearly by a star ★. This is *non-examinable* and generally more challenging material. But you should still read and think about it. It is included to put the subject in a broader perspective, to provide an overview, to indicate further directions, and to generally “round out” the subject. In addition, studying this more advanced material will help your understanding of the examinable material.

Moreover, you *will* need to know and understand the statements of the results in the ★ Sections 3.4, 5.6, 5.7 and 6.6, for the proof of the Fundamental Existence and Uniqueness Theorem in Section 8.2, and to a lesser extent in Section 8.1.

The other references for the course are the book [Adams] and the notes [Ward]. Both of these supplement the material here, but are at a less theoretical level. The book [Spivak] is excellent, but at a slightly higher level than the current course. The book [Stromberg] is for the extremely dedicated; it is very terse and essentially only appropriate for later year courses. For interesting discussions and a host of examples on the beauty and utility of mathematics, see [Devlin], [Hildebrandt and Tromba] and [Davis and Hersh].

All these books are on two-day loan through the reserve section of the Hancock library.

If you are having difficulty with some of the concepts, ask your tutor or come and see me during office hours. Do not let things slide!

CHAPTER 2

The Real Number System

The reference for this chapter is [Adams, Chapter P], mainly P1, P2 to page 14, P4 to page 29, and P5. But you should read all of this chapter; it gives a slightly different and somewhat more elementary approach to the material covered here.

2.1. Introduction

Real numbers have decimal expansions. They can be represented as points on an infinite line. The decimal expansions $1.000\dots$ and $.999\dots$ represent the same real number.

Real numbers have decimal expansions, for example:

$$\begin{aligned}2 &= 2.000\dots \\1\frac{1}{2} &= 1.5 = 1.5000\dots \\ \pi &= 3.1459\dots \\ .4527\dot{1}4\dot{6}, & \text{ also written } .4527\overline{146}\end{aligned}$$

The “ \dots ” indicate the expansions go on forever, and the $\dot{1}4\dot{6}$ indicate that the pattern 146 is repeated forever. In the first two case the expansion continues with zeros and in the third case one can compute the expansion to any required degree of accuracy.

Real numbers can also be represented geometrically as points on an infinite line.



In this chapter we will give a careful analysis of what is meant by a real number. (Sometimes we say “real number”, and sometimes we just say “number”. You will also later meet the “complex numbers” (if you have not already done so), these include the real numbers and allow one to give a meaning to $\sqrt{-1}$.)

There is one point that sometimes causes confusion. Is it the case that

$$1 = .\dot{9}?,$$

or is it that $.\dot{9}$ is a “little” less than one? By $.\dot{9}$ we mean, as usual, $.999\dots$, with the 9’s repeated forever.

Any of the approximations to $.\dot{9}$,

$$.9 = \frac{9}{10}, .99 = \frac{99}{100}, .999 = \frac{999}{1000}, .9999 = \frac{9999}{10000}, \dots$$

is certainly strictly less than one.

On the other hand, $.\dot{9}$ is defined to be the “limit” of the above infinite sequence (we discuss limits of sequences in a later chapter). Any *mathematically useful* way in which we define the limit of this sequence will in fact imply that $.\dot{9} = 1$. To see this, let

$$a = .\dot{9} = .999\dots$$

Then, for any reasonable definition of infinite sequence and limit, we would want that

$$10a = 9.999\dots$$

Subtracting, gives $9a = 9$, and hence $a = 1$.

2.2. Important sets of real numbers

Beginning from the set \mathbb{R} of real numbers we define the sets \mathbb{N} of natural numbers, \mathbb{Z} of integers, \mathbb{Q} of rationals, and the set of irrationals. We discuss intervals. We introduce some general notation for describing sets. Finally we discuss n -tuples of numbers and n -dimensional space.

2.2.1. Notation for sets. By a *set* (or *class*, or *family*) we mean a collection, often infinite, of objects of some type.¹ Members of a set are often called *elements* of the set. If a is a member (i.e. element) of the set S , we write

$$a \in S.$$

If a is not a member of S we write

$$a \notin S.$$

If a set is finite, we may describe it by listing its members. For example,

$$A = \{1, 2, 3\}.$$

Note that $\{1, 2, 3\}$, $\{2, 3, 1\}$, $\{1, 1, 2, 2, 2, 3\}$ are different descriptions of exactly the same set. Some infinite sets can also be described by listing their members, provided the pattern is clear. For example, the set of even positive integers is

$$E = \{2, 4, 6, 8, \dots\}.$$

We often use the notation

$$S = \{x : P(x)\},$$

where $P(x)$ is some statement, or “proposition”, involving x . We read this as “ S is the set of all (real numbers) x such that $P(x)$ is true”. It is usually understood from the context of the discussion that x is restricted to the real numbers. But if there is any possible ambiguity, then we write

$$S = \{x \in \mathbb{R} : P(x)\}.$$

Note that this is *exactly* the same set as

$$\{y : P(y)\},$$

or

$$S = \{y \in \mathbb{R} : P(y)\}.$$

The variables x and y are sometimes called “dummy” variables, they are meant to represent any real number with the specified properties.

One also sometimes uses “|” instead of “:” when describing sets.

The *union* of two or more sets is the set of numbers belonging to at least one of the sets. The *intersection* of two or more sets is the set of numbers belonging to all of the sets. We use \cup for union and \cap for intersection.

For example

$$\begin{aligned} \{x : 0 < x < 1 \text{ or } 2 < x \leq 3\} &= (0, 1) \cup (2, 3] \\ (0, 2) &= (0, 1) \cup [1, 2) = (0, 1] \cup [1, 2) = (0, 1) \cup \left(\frac{1}{2}, 2\right) \\ \{x : 0 < x < 2 \text{ and } 1 \leq x \leq 3\} &= [1, 2) \end{aligned}$$

¹★ There is a mathematical theory of sets, and in fact all of mathematics can be formulated within the theory of sets. However, this is normally only useful or practical when considering fundamental questions about the foundations of mathematics.

2.2.2. Different types of real numbers. The set of real numbers is denoted by

$$\mathbb{R}.$$

The set \mathbb{N} of *natural numbers* is defined² to be

$$\mathbb{N} := \{1, 2, 3, \dots\}.$$

The set \mathbb{Z} of *integers* is defined by

$$\mathbb{Z} := \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}.$$

The set \mathbb{Q} of *rational numbers* is defined by

$$\mathbb{Q} := \{m/n : m, n \in \mathbb{Z}, n \neq 0\}.$$

Rational numbers are those whose decimal expansion either terminates after a finite number of places, as for 2 and 1.5, or are recurring, as for $.4527\overline{146}$ or $3/7$. See [Adams, Ex. 1, page 4].

A real number is *irrational* if it is not rational. It can be proved that π and e are irrational, see [Spivak].

2.2.3. Intervals. An *interval* is a set of real numbers with the property that it contains at least two numbers, and moreover it contains all real numbers between any two of its members. (It follows that an interval must contain an infinite number of members.³)

Bounded intervals are intervals of the following type:

$$[a, b] := \{x : a \leq x \leq b\},$$

$$(a, b) := \{x : a < x < b\},$$

$$[a, b) := \{x : a < x \leq b\},$$

$$[a, b) := \{x : a \leq x < b\}.$$

An interval may also be unbounded in either or both directions:

$$[a, \infty) := \{x : a \leq x\},$$

$$(a, \infty) := \{x : a < x\},$$

$$(-\infty, b] := \{x : x \leq b\},$$

$$(-\infty, b) := \{x : x < b\}.$$

Finally, \mathbb{R} is also an interval, which we could write as $(-\infty, \infty)$. Note that ∞ is *not* a number, and by itself does not have any meaning here, just as $\{$ or $:$ does not have any meaning by itself.

An interval is *open* if it does not contain any of its endpoints, and is *closed* if it contains all of its endpoints. Thus open intervals are those of the form (a, b) , (a, ∞) , $(-\infty, b)$ and \mathbb{R} , while closed intervals are those of the form $[a, b]$, $[a, \infty)$, $(-\infty, b]$ and \mathbb{R} . (If this seems confusing, remember that $\pm\infty$ are not numbers, and in particular cannot be endpoints of intervals.) In particular, \mathbb{R} is both open and closed.

The *end-points* or *boundary points* for the previous examples are the points a and b (note that they may or may not belong to the respective interval). The *interior* points are all other points in the interval. Thus in $(1, 2]$ the endpoints are 1 and 2, of which only 2 belongs to the interval, and the interior points are all points in the open interval $(1, 2)$.

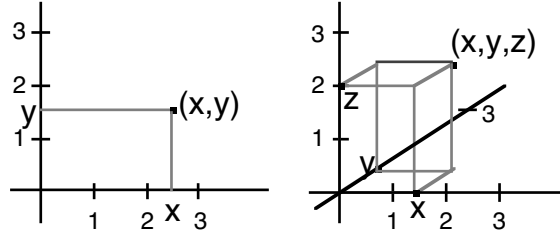
²We often use the notation “:=” to mean “by definition is equal to”.

³If a, b are two numbers in the interval, then, for example,

$$a + \frac{b-a}{2}, a + \frac{b-a}{3}, a + \frac{b-a}{4}, \dots$$

is an infinite set of numbers between a and b which are distinct from one another.

2.2.4. n -tuples of real numbers. We will later work with pairs (x, y) of real numbers, triples of real numbers (x, y, z) , and more generally n -tuples (x_1, \dots, x_n) . Just as real numbers can be represented geometrically by points on a line, so can pairs be represented as points in the plane and triples as points in space (we need to make a choice of origin, coordinate axes, and scales on the axes). We could also “define” n -dimensional space to be the set of n -tuples of real numbers!



2.3. Algebraic and order properties

We introduce the algebraic and order axioms for the real number system and indicate very briefly how all the usual algebraic and order properties follow from these. The rational numbers also satisfy these axioms, but this is not the case for the natural numbers, the integers, or the irrationals.

The absolute value of a real number is defined and the basic properties are proved.

The *real number system* consists of the real numbers, together with the operations *addition* (denoted by $+$) and *multiplication* (denoted by \times) and the *less than* relation (denoted by $<$). One also singles out two particular real numbers, *zero* or 0 and *one* or 1.

If a and b are real numbers, then so are $a + b$ and $a \times b$; and the relation $a < b$ must be either true or false. We will usually write

$$ab \text{ for } a \times b.$$

We will soon see that one can define subtraction and division in terms of $+$ and \times ; and \leq , $>$ etc. can be defined from $<$.

There are three categories of properties of the real number system: the *algebraic* properties, the *order* properties and the *completeness* properties.

2.3.1. Algebraic properties. These are the properties of addition, multiplication, subtraction and division. It turns out that there are certain basic properties, usually called *axioms*, from which we can prove all the other algebraic properties. These axioms are:

AXIOMS (Addition). If a , b and c are real numbers then:

A1:

$$a + b = b + a;$$

A2:

$$(a + b) + c = a + (b + c);$$

A3:

$$a + 0 = 0 + a = a;$$

A4: there is exactly one real number, denoted by $-a$, such that

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

AXIOMS (Multiplication). If a , b and c are real numbers then:

A5:

$$a \times b = b \times a;$$

A6:

$$(a \times b) \times c = a \times (b \times c);$$

A7:

$$a \times 1 = 1 \times a = a, \quad 1 \neq 0;$$

A8: if $a \neq 0$ there is exactly one real number, denoted by a^{-1} , such that

$$a \times a^{-1} = a^{-1} \times a = 1.$$

AXIOMS (The Distributive Property). If a, b and c are real numbers then:**A9:**

$$a \times (b + c) = a \times b + a \times c.$$

NOTE 2.1.

1. We are not really using subtraction in axiom A4; we could just as well have written a^* for $-a$, but it is more consistent with standard notation to write $-a$. We are merely asserting that a unique real number, with a certain property, exists. Similarly, in A8, we could just as well have written $a^\#$, say, for a^{-1} .
2. By the symbol “=” for equality we mean “denotes the same thing as”, or equivalently, “is the same real number as”. We take = to be a *logical notion* and do not write down axioms for it.⁴ Instead, we use any properties of “=” which follow from its logical meaning. For example: $a = a$; if $a = b$ then $b = a$; if $a = b$ and $b = c$ then $a = c$; if $a = b$ and something is true of a then it is also true of b (since a and b denote the same real number!).

When we write $a \neq b$, we mean that a is *not* the same real number as b .

3. Some of the axioms are redundant. For example, from A1 and the property $a + 0 = a$ it follows that $0 + a = a$, (*why?*). Similar comments apply to A4; and because of A5 to A7 and A8.
4. ★ One can show that apart from these cases the axioms are not redundant; in other words that no one axiom follows from the others. More precisely, one can construct examples where, say, A8 is false but all the other axioms are true (see the following section on Fields); and similarly for any of the other axioms.

2.3.2. More algebraic properties. All the usual algebraic properties of the real numbers follow from A1–A9, in particular, one can solve simultaneous systems of linear equations. We will not spend much time on indicating how one deduces other algebraic properties from these axioms, but will continue to use all the usual properties of addition, multiplication, subtraction and division that you have used in the past.

None-the-less, it is useful to have some idea of the methods involved in making deductions from A1–A9. Later in this course, when we discuss vector spaces, you will have more practice at making deductions from “algebraic” sets of axioms somewhat like those above.

The first thing is to *define* subtraction and division. For this, suppose a and b are any⁵ two real numbers (except that $b \neq 0$ in the definition of division). Then we *define*

$$\begin{aligned} a - b &= a + (-b) \\ a \div b &= a \times b^{-1} \quad \text{for } b \neq 0 \end{aligned}$$

This may look like a circular definition; it may appear that we are defining “subtraction” in terms of “subtraction”. But this is not the case. Given b , from A4 there is a certain real number, which we denoted by $-b$, with certain properties. We then define $a - b$ to be the *sum* of a and this real number $-b$.

Similar comments apply to the definition of division. We also write a/b or $\frac{a}{b}$ for $a \div b$.

We can also now define other numbers and operations. For example, we define $2 = 1 + 1$, $3 = 2 + 1$, etc.

Also, we define $x^2 = x \times x$, $x^3 = x \times x \times x$, $x^{-2} = (x^{-1})^2$, etc. etc.

We define \sqrt{b} , for $b \geq 0$, to be that number $a \geq 0$ such that $a^2 = b$. Similarly, if n is a natural number, then $\sqrt[n]{b}$ is that number $a \geq 0$ such that $a^n = b$. To prove there *is* always such a number a requires the “completeness axiom” (see later).

⁴★ One *can* write down basic properties, i.e. axioms, for “=” and the logic we use. See later courses on the foundations of mathematics.

⁵We do not even assume $a \neq b$.

As an example of the way in which one can use A1–A9 to derive other algebraic properties, we prove the cancellation property of addition:

THEOREM 2.2. *If a , b and c are real numbers and $a + c = b + c$, then $a = b$.*

PROOF. Assume

$$a + c = b + c.$$

Since $a + c$ and $b + c$ denote the same real number, we obtain the same result if we add $-c$ to both; i.e.

$$(a + c) + (-c) = (b + c) + (-c).$$

(This used the existence of the number $-c$ from A4.) Hence

$$a + (c + (-c)) = b + (c + (-c))$$

from A2 applied twice, once to each side of the equation. Hence

$$a + 0 = b + 0$$

from A4 again applied twice. Finally,

$$a = b$$

from A3. □

We will not pursue this idea of making deductions from the axioms, but see Section 2.9 if you are interested in knowing more. In future we will forget about the axioms and just use all the usual properties of addition, multiplication, subtraction and division. Here we list a few:

THEOREM 2.3. *If a , b , c , d are real numbers and $c \neq 0$, $d \neq 0$ then*

1. *if $ac = bc$ then $a = b$.*
2. $a0 = 0$
3. $-(-a) = a$
4. $(c^{-1})^{-1} = c$
5. $(-1)a = -a$
6. $a(-b) = -(ab) = (-a)b$
7. $(-a) + (-b) = -(a + b)$
8. $(-a)(-b) = ab$
9. $(a/c)(b/d) = (ab)/(cd)$
10. $(a/c) + (b/d) = (ad + bc)/cd$

★ For those who are interested, I indicate the proofs of these facts from the axioms in Section 2.9.

2.3.3. Order properties. The real numbers have a natural ordering, denoted by “ $<$ ” which we read as “is less than”. Basic properties are:

AXIOMS (Less than). If a , b and c are real numbers then:

A10: one and only one of the following hold

$$a < b \text{ or } a = b \text{ or } b < a;$$

A11:

$$\text{if } a < b \text{ and } b < c, \text{ then } a < c;$$

A12:

$$\text{if } a < b \text{ then } a + c < b + c;$$

A13:

$$\text{if } a < b \text{ and } 0 < c, \text{ then } ac < bc;$$

If $0 < a$ we say a is *positive* and if $a < 0$ we say a is *negative*.

2.3.4. More order properties. One can *define* “ $>$ ”, “ \leq ” and “ \geq ” in terms of $<$ as follows:

$$\begin{aligned} a > b & \text{ if } b < a, \\ a \leq b & \text{ if } (a < b \text{ or } a = b), \\ a \geq b & \text{ if } (a > b \text{ or } a = b). \end{aligned}$$

(Note that the statement $1 \leq 2$, although it is not one we are likely to make, is indeed true, *why?*)

All the usual properties of inequalities can in fact now be proved from A10–A13, (together with A1–A9). For example,

THEOREM 2.4. *If a , b and c are real numbers then*

1. $a < b$ and $c < 0$ implies $ac > bc$
2. $0 < 1$ and $-1 < 0$
3. $a > 0$ implies $1/a > 0$
4. $0 < a < b$ implies $0 < 1/b < 1/a$

Henceforth we will forget about the axioms and use all the properties of inequalities and all the algebraic properties that you have used before.

REMARK 2.5. It is a simple consequence of the standard properties of $<$ that there is no smallest positive number, because if s is any positive number then $s/2$ (for example) is a smaller positive number.

REMARK 2.6. The set \mathbb{Q} of rational numbers, together with 0 and 1, the operations of addition and multiplication, and the $<$ relation, is also a model of the corresponding versions of axioms A1–A13, with “real” replaced by “rational”. The main points to note are that 0 and 1 are of course rational, if a is rational then so are $-a$ and a^{-1} (assuming $a \neq 0$), and the sum and product of rational numbers is rational. Apart from this, the axioms A1–A13 are satisfied for rational numbers because they are satisfied for real numbers (every rational number is certainly real).

The set of irrational numbers does not satisfy the corresponding versions of A1–A13. For example, 0 and 1 are not irrational, and the sum and product of irrational numbers need not be irrational (*examples?*).

Which of the axioms A1–A13 are satisfied by \mathbb{N} ? By \mathbb{Z} ?

DEFINITION 2.7. The *absolute value* of a real number a is defined by

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0 \end{cases}$$

(If we wanted the definition to look more “symmetric” we could have considered the cases $a > 0$ and $a = 0$ separately. But otherwise there is no real point to doing this.)

It follows by considering the cases $a \geq 0$ and $a < 0$ separately that⁶

$$(2.1) \quad |a| = \sqrt{a^2}, \quad -|a| \leq a \leq |a|.$$

Note also that

$$(2.2) \quad |a| < b \quad \text{if and only if} \quad (a < b \text{ and } -a < b).$$

The following properties also follow from the properties of inequalities.

THEOREM 2.8. *If a and b are real numbers then:*

1. $|ab| = |a| |b|$
2. $|a \pm b| \leq |a| + |b|$ (triangle inequality)
3. $||a| - |b|| \leq |a - b|$

PROOF. (We can use any of the usual properties of inequalities, together with the definition of “ $|\cdot|$ ”.)

⁶When we write $p \leq q \leq r$ we mean $p \leq q$ and $q \leq r$. Similarly for $p < q < r$ etc.

1. We consider the various possible cases. If either a or b (or both) are zero, then both sides are zero. If $a > 0$ and $b > 0$ then also $ab > 0$ and both sides equal ab . If $a > 0$ and $b < 0$ then $ab < 0$ and both sides equal $-ab$; similarly if $a < 0$ and $b > 0$. Finally, if $a < 0$ and $b < 0$ then both sides equal ab . Hence one has equality in *all* cases.
2. $|a + b|$ is either $a + b$ or $-(a + b)$, while $|a - b|$ is either $a - b$ or $-(a - b)$. Thus $|a \pm b|$ is one of $a + b$, $-a - b$, $a - b$ or $-a + b$. The result now follows from (2.1), since each of these four quantities is $\leq |a| + |b|$.
3. From (2.2) it is sufficient to prove

$$|a| - |b| \leq |a - b| \quad \text{and} \quad -(|a| - |b|) \leq |a - b|.$$

From the triangle inequality we have

$$|a| = |(a - b) + b| \leq |a - b| + |b|,$$

and so

$$(2.3) \quad |a| - |b| \leq |a - b|.$$

Since this is true for *any* real numbers a and b , we can switch a and b to get

$$|b| - |a| \leq |b - a| (= |a - b|),$$

i.e.

$$(2.4) \quad -(|a| - |b|) \leq |a - b|.$$

It follows from (2.3) and (2.4) that

$$||a| - |b|| \leq |a - b|.$$

□

By repeated applications of the triangle inequality it follows that

$$(2.5) \quad |a_1 + a_2 + \cdots + a_n| \leq |a_1| + |a_2| + \cdots + |a_n|,$$

for any natural number n .

See [Adams, pages 8–11] for more on inequalities.

2.4. Completeness property

The Completeness Axiom is introduced. It is true for the real numbers, but the analogous result is not true for the rationals. We define the notion of upper bound (lower bound) and least upper bound (greatest lower bound) of a set of real numbers.

This is the final axiom for the real number system, and is probably not one you have met before. It is more difficult to understand than the other properties, but it is essential in proving many of the important results in calculus.

AXIOMS (Completeness).

A14: If A is any non-empty set of real numbers with the property that there is some real number x such that $a \leq x$ for every $a \in A$, then there is a *smallest* (or *least*) real number x with this same property.

A is *non-empty* means that A contains at least one number.

Note that the number x in the axiom need not belong to A . For example, if A is the interval $[0, 1)$ then the smallest (or “least”) number x as above is 1, but $1 \notin A$. On the other hand, if $A = [0, 1]$ then the smallest number x as above is again 1, but now $1 \in A$.

There is some useful notation associated with the Completeness axiom.

DEFINITION 2.9. If A is a set of real numbers and x is a real number such that $a \leq x$ for every $a \in A$, then x is called an *upper bound* for A . If x is the smallest upper bound then x is called the *least upper bound* or *supremum* of A . In this case one write

$$x = \text{l. u. b. } A \quad \text{or} \quad x = \sup A.$$

If $x \leq a$ for every $a \in A$, then x is called an *lower bound* for A . If x is the largest lower bound then x is called the *greatest lower bound* or *infimum* of A . In this case one writes

$$x = \text{g.l.b. } A \quad \text{or} \quad x = \inf A.$$

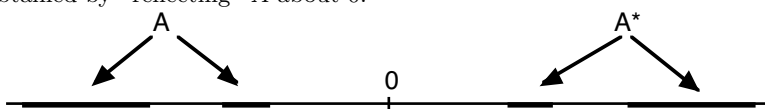
Thus the Completeness Axiom says that *if* a non-empty set A has an upper bound *then* it also has a least upper bound. (Remember that when we say A “has” a least upper bound x we do *not* require that $x \in A$.)

There is an equivalent form of the axiom, which says: *If A is any non-empty set of real numbers with the property that there is some real number x such that $x \leq a$ for every $a \in A$, then there is a largest real number x with this same property.* In other words *if* a non-empty set A has a lower bound *then* it also has a greatest lower bound.

It is not too hard to see that this form does indeed follow from the Completeness axiom. The trick is to consider, instead of A , the set

$$A^* := \{-x : x \in A\},$$

which is obtained by “reflecting” A about 0.



Lowerbounds for A correspond under reflection to upperbounds for A^* , and a g.l.b. corresponds to a l.u.b.. If A is bounded below then A^* is bounded above, and so by the completeness axiom has a l.u.b.. After reflection, this l.u.b. for A^* gives a g.l.b. for A . (To actually write this out carefully needs some care—you need to just show check the relevant definitions and the properties of inequalities that the first two sentences in this paragraph are indeed correct.)

Similarly, the completeness axiom does follow from this equivalent form.

Unlike in the case of A1–A13, we will always indicate when we use the completeness axiom (or “property”).

The completeness axiom implies there are no “gaps” in the real numbers.

For example, the rational numbers are *not* a model of the corresponding version of A14. This is because there are sets of rational numbers A which have the property that there is *some* rational number x such that $a \leq x$ for every $a \in A$, but there is no *smallest* rational number x with this same property. For example, let

$$A = \{a \in \mathbb{Q} : 0 \leq a \text{ and } a^2 < 2\} = \{a \in \mathbb{Q} : 0 \leq a < \sqrt{2}\}.$$

(The first definition for A has the advantage that A is defined without actually referring to the existence of $\sqrt{2}$, even as a real number.) There are certainly rational numbers x such that $a \leq x$ for every $a \in A$, just take $x = 23$. But we claim *there is no smallest such rational number*.

PROOF. This claim basically follows from the fact that $\sqrt{2}$ is not rational, see Theorem 2.11, and so cannot be the required rational number.

On the other hand, the required rational number x cannot be $< \sqrt{2}$, since there is always a rational number between any such x and $\sqrt{2}$ (see Theorem 2.16), contradicting the fact $x \geq a$ for *every* $a \in A$.

Finally, the required x cannot be $> \sqrt{2}$, since there is always a rational number between $\sqrt{2}$ and any such x (see Theorem 2.16), and this rational number means we have contradicted the fact x is the *smallest* rational number such that $x \geq a$ for every $a \in A$. \square

REMARK 2.10. \star We defined an interval to be a set of real numbers with the property that it contains at least two numbers, and moreover it contains all real numbers between any two of its members. It follows from the completeness axiom that any interval is indeed one of the 9 types described in Section 2.2.3. While this may seem obvious, we do need the completeness axiom to prove it, essentially since the upper bound of an interval may not be rational.

For example, suppose the interval I is bounded. Then it has a g.l.b. a and a l.u.b. b , say. If $a \in I$ and $b \in I$ then $I = [a, b]$ since I contains all numbers between a and b , and no others by definition of g.l.b. and l.u.b..

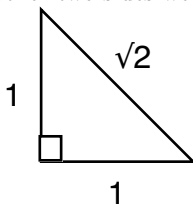
If $a \in I$ and $b \notin I$ then $I = [a, b)$. The main point is to prove that if $a \leq x < b$ then $x \in I$. But if $x \notin I$ then neither can any number greater than x be in I (by definition of "interval") and hence x is an upper bound for I , contradicting the fact that b is the *least* upper bound.

In a similar way, we establish that any interval is of one of the 9 given forms.

2.5. Properties of the rational and irrationals

We prove that $\sqrt{2}$ is irrational. We show the Completeness Axiom implies that there is indeed a real number whose square is 2; that there is no real number greater than every integer; and that for any positive number (no matter how small) there is a smaller number of the form $1/n$ for some natural number n . We prove that between any two real numbers there are an infinite number of rationals and an infinite number of irrationals.

The ancient Greeks, in the school of Pythagoras, first discovered that not all real numbers are rational. This was considered to be a very serious problem, since the Greeks thought in terms of the natural numbers and ratios! They proved that $\sqrt{2}$ is irrational, even though it arose in the very natural manner as the length of the hypotenuse of a right-angled triangle whose other two sides were each of length 1.



THEOREM 2.11. $\sqrt{2}$ is not rational.

PROOF. We argue by contradiction. That is, we *assume*

$$\sqrt{2} = m/n$$

where m and n are integers.

Multiplying numerator and denominator by -1 if necessary, we can take m and n to be positive. By canceling if necessary, we can reduce to the situation where m and n have no common factors. Squaring both sides of the equation, we have for these new m and n that

$$2 = m^2/n^2$$

and hence

$$m^2 = 2n^2.$$

It follows that m is even, since the square of an odd number is odd. (More precisely, if m were odd we could write $m = 2r + 1$ for some integer r ; but then $m^2 = (2r + 1)^2 = 4r^2 + 4r + 1 = 2(2r^2 + 2r) + 1$, which is odd, not even.) But since m is even, we can write

$$m = 2p$$

for some integer p , and hence

$$m^2 = 4p^2.$$

Substituting this into $m^2 = 2n^2$ gives

$$4p^2 = 2n^2,$$

and hence

$$2p^2 = n^2.$$

But now we can argue as we did before for m , and deduce that n is also even. Thus m and n both have the common factor 2, which contradicts the fact they have no common factors.

This contradiction implies that our original assumption was wrong, and so $\sqrt{2}$ is not rational. \square

The fact that there *is* a real number $\sqrt{2}$, i.e. a positive real number x such that $x^2 = 2$, is hardly surprising, but it does not actually follow from A1–A13. This is because the rational numbers themselves satisfy A1–A13, but the previous proof shows there is no *rational* number x with this property.

Although we give the proof of the following theorem at this stage, it follows much more easily later on from the Intermediate Value Theorem, see Theorem 5.17.

THEOREM 2.12. *There exists a positive number x with the property that $x^2 = 2$. We write $x = \sqrt{2}$.*

PROOF. Let

$$A = \{ a \in \mathbb{R} : 0 \leq a \text{ and } a^2 < 2 \}.$$

Let $x = \sup A$, which exists by the Completeness axiom, since A is certainly bounded above, by 2, say.

[The *existence* of such an x requires the completeness axiom. But once we know such a number exists, it has the usual algebraic and order properties of any real number. It is now just a matter of some messy manipulating of inequalities to rule out $x^2 < 2$ and $x^2 > 2$, thereby showing that in fact $x^2 = 2$.]

There are three possibilities:

$$x^2 < 2, \quad x^2 = 2, \quad x^2 > 2.$$

If $x^2 < 2$, then by taking y to be a slightly bigger number than x , we still have $y^2 < 2$.

(\star This is not surprising, but to write it out carefully is a little tricky. To do it, let $y = x + \varepsilon$ where $\varepsilon > 0$ is yet to be chosen. Then $y^2 = x^2 + \varepsilon(2x + \varepsilon)$. Now choose $\varepsilon > 0$ so

$$\varepsilon < \min \left\{ 1, \frac{2 - x^2}{2x + 1} \right\}.$$

Because $\varepsilon < 1$, it follows that $2x + \varepsilon < 2x + 1$.

Because also $\varepsilon < \frac{2 - x^2}{2x + 1}$ it then follows that

$$y^2 = x^2 + \varepsilon(2x + \varepsilon) < x^2 + \varepsilon(2x + 1) < x^2 + (2 - x^2) = 2,$$

and so

$$y^2 < 2. \quad)$$

It follows that $y \in A$ from the definition of A , but then we have a contradiction to $y > x$ since $x \geq$ any member of A .

If $x^2 > 2$, then by taking y to be a slightly *smaller* number than x , we also have $y^2 > 2$ (the proof is similar to the above, *exercise*). But $y^2 > 2$ implies $y > a$ for every $a \in A$ (since if $y \leq a$ and $y \geq 0$, then $y^2 \leq a^2$, which in turn implies $y^2 < 2$, contradiction). Hence y is an upper bound for A , contradicting the fact that x is the *smallest* upper bound for A .

Hence $x^2 = 2$, since we have ruled out $x^2 < 2$ and $x^2 > 2$. \square

One very useful fact is that between any two distinct real numbers there is a rational and an irrational number (in fact an infinite number of each type). We say the set of rationals and the set of irrationals are both *dense* in \mathbb{R} .

But first we need Theorem 2.13, which is logically equivalent to the statement for every real number x there exists a natural number n (which will depend on x) such that⁷ $n > x$.

1. Why does the theorem imply this fact?
2. Why does this fact imply the theorem?

Note that the assertion in Theorem 2.13 is not just an algebraic or inequality property. It is an assertion about the *non-existence* of a real number with a certain property (or equivalently, by the previous paragraph, to the *existence* — for each real number — of a (natural) number with a certain property.) For this reason, it is not surprising that we need the Completeness axiom to prove it.

The proof is a little unusual, but it *is* logically correct.

THEOREM 2.13 (Archimidean Property). *There is no real number x with the property that $x \geq n$ for every natural number n .*

PROOF. ★ (This is another proof by contradiction.) Assume that there is some number, which we denote by x , such that $x \geq n$ for every $n \in \mathbb{N}$. By the Completeness Axiom there is a smallest such x ; consider this particular x .

Since $n + 1$ is a natural number if n is, we must also have that $x \geq n + 1$ for every natural number n . But this is the same as saying $x - 1 \geq n$ for every natural number n .

In other words, x has the property that $x - 1 \geq n$ for every $n \in \mathbb{N}$!! This contradicts the fact x was the *least* number with the property $x \geq n$ for every $n \in \mathbb{N}$. Thus the *assumption* at the beginning of the proof was wrong, and so there is no real number x greater than or equal to every natural number. □

The next result is also important. We could just as well have written x instead of ε , but it is traditional to write ε or δ for a small positive number.

COROLLARY 2.14. *For any real number $\varepsilon > 0$ there is a natural number n such that $1/n < \varepsilon$.*

PROOF. Suppose $\varepsilon > 0$. From the previous remark, there is a natural number n such that $n > 1/\varepsilon$. This implies $1/n < \varepsilon$. □

REMARK 2.15. A *theorem* is an important result, and a *corollary* is something which is a fairly straightforward consequence of a previous theorem.

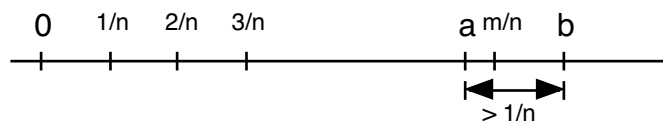
THEOREM 2.16. *Assume $a < b$ are real numbers. Then there is a rational number x and an irrational number y such that $a < x < b$ and $a < y < b$.*

PROOF. We have just seen that there is a natural number n such that

$$\frac{1}{n} < b - a.$$

First assume $a \geq 0$. Consider the sequence

$$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \frac{4}{n}, \dots$$



Since $1/n$ is less than the difference between a and b , it follows that at least one member m/n of this sequence must lie between a and b . This is the required *rational* number x .⁸

⁷In fact, there exists an infinite number of such n .

⁸★ To be more precise, use the Completeness Property to first choose x as the least real number with the property that $x \geq p/n$ whenever $p/n \leq a$ and n natural numbers. One then shows that x must itself be of the form p/n for some p , and then deduces that $a < (p + 1)/n < b$ for this particular p .

If $a < 0$, then a similar proof works with the sequence

$$-\frac{1}{n}, -\frac{2}{n}, -\frac{3}{n}, -\frac{4}{n}, \dots$$

To find an irrational number y between a and b , first choose a natural number n such that $\sqrt{2}/n < b - a$ (*why is this possible?*). Then choose a natural number m as before so that now $a < m(\sqrt{2}/n) < b$. Since $m\sqrt{2}/n$ is irrational (if it were rational, and equaled p/q , say, then it would follow that $\sqrt{2} = np/mq$ and so $\sqrt{2}$ would be rational, contradicting Theorem 2.11) we can take this as the required *irrational* number y .

The proof is again similar if $a < 0$. □

REMARK 2.17. We can now choose another rational number x_1 between a and the rational number x of the theorem, and then another rational number between a and x_2 , etc. This gives an infinite set of rational numbers between a and b .



Similarly there is an infinite number of irrational numbers between any two real numbers.

2.6. Functions

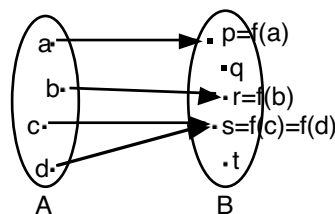
We define the notion of a function, its domain and its range. We discuss the idea of a dependent and of an independent variable. We give a number of examples of functions. We show how functions can be combined algebraically and by composition to give new functions.

One of the most important ideas in mathematics is that of a function.

DEFINITION 2.18. A *function* f from a set A into a set B is something which assigns to every number $x \in A$ a unique (i.e. exactly one) number $f(x) \in B$. We write

$$f : A \rightarrow B,$$

and say f maps (the set) A into (the set) B .



Note that each number in A is mapped to exactly one number in B . *Different* numbers in A may be mapped to the *same* number in B , but a number in A cannot be mapped to more than one number in B . There may be numbers in B which are not of the form $f(a)$ for any $a \in A$.

In this course, A and B will usually (but not always) be sets of numbers, but in later work it is very important to take more general functions where A and B may be much more general sets of objects.

The *domain* of f , written $\mathcal{D}(f)$, is the set A (the set of “input” values of f). The *range* of f , written $\mathcal{R}(f)$, is the set of all numbers of the form $f(x)$ for some $x \in A$ (the set of “output” values of f). Thus

$$\mathcal{D}(f) = A, \quad \mathcal{R}(f) \subset B.$$

Here, $\mathcal{R}(f) = \{p, r, s\}$. (By $S \subset T$, where S and T are two sets, we mean that every member of S is also a member of T . Notice that if S and T are the same sets, which means they have the same members, then it is *also* true that $S \subset T$.)

Many texts say a function from A to B is a “rule” which assigns to each member of A a member of B . But it is necessary to interpret the meaning of the word “rule” in a *very* broad sense—it is not necessary that a function be “described” by some sort of English or

mathematical expression. All that is meant is that to each number x (say) in the domain there corresponds exactly one number, denoted by $f(x)$.

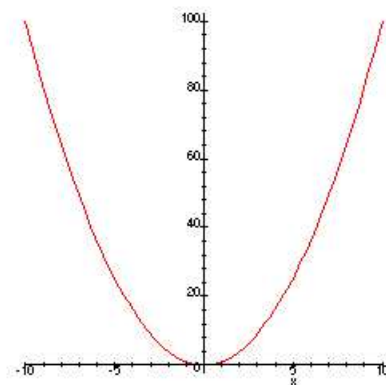
Often a function will only be described in an indirect manner; for example it may be the limit of a sequence of other functions (see the proof of Theorem 8.2). In other cases we may prove a function exists by using argument by contradiction to show that $f(x)$ has a value for every x in some given domain.

We adopt the convention that, unless indicated otherwise, the domain of a function is the largest set of real numbers for which the definition of the function makes sense. Thus the domain of the function f described by $f(x) = 1/(x-2)$ is (unless stated otherwise) the set of all real numbers *other than* 2. We write this set as $\{x : x \neq 2\}$. See also Examples 2,3,5 below.

In order to describe a function completely, we need to give *both* the domain and the rule.

Two functions f_1 and f_2 are said to be the same function, or to be *equal*, if they have the same domain, and if $f_1(x) = f_2(x)$ for all x in the domain.

The *graph* of f is the set of all points $(x, f(x))$ such that $x \in \mathcal{D}(f)$. For example, the graph of the function $f(x) = x^2$ is



NOTATION 2.19. If we denote an arbitrary input value of a function f by x and the corresponding output value $f(x)$ by y , we say x is the *dependent variable* and y is the *independent variable*. We write

$$y = f(x)$$

and say “ y equals f of x ”.

Besides letting y denotes an output value as above, it is also often convenient in computations and applications to let y denote the actual function f itself. In this case we sometimes write $y = y(x)$ to indicate that the “function” y has output value $y(x)$ for input value x . *However, when we are looking at more theoretical questions, this can lead to confusion and ambiguity, and so in those circumstances we will usually avoid this practice.*

EXAMPLE 2.20.

1. Consider the function f defined by

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

Unless we say otherwise the domain here is \mathbb{R} and the range is then $[0, \infty)$. We can write

$$f : \mathbb{R} \rightarrow \mathbb{R}, \quad f : \mathbb{R} \rightarrow [0, \infty), \quad \text{or even } f : \mathbb{R} \rightarrow [-3, \infty).$$

Note that the function f defined by $f(x) = x^2$ is *exactly* the same as the function f defined by $f(y) = y^2$ or by $f(a) = a^2$. We say x, y, a are *dummy variables*. We call f the *squaring function*. It is also the same as the function g defined by $g(x) = x^2$.

Consider the function h defined by

$$h(x) = x^2, \quad x \geq 1.$$

This is the function whose domain is $[1, \infty)$ and which assigns to each x in the domain the number x^2 . Thus the two functions f and h are *not* equal since their domains are not equal.

2. The function f defined by

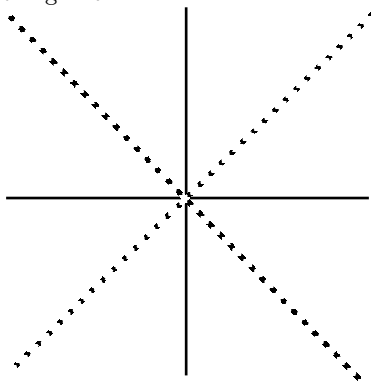
$$f(x) = \frac{x^2 + 1}{x - 1}$$

has domain $\{x : x \neq 1\}$. This is the same set as $(-\infty, 1) \cup (1, \infty)$, the union of $(-\infty, 1)$ and $(1, \infty)$. It is also *exactly* the same set as $\{y : y \neq 1\}$

3. The function f defined by $f(x) = \sqrt{x}$ has domain $\{x : x \geq 0\} = [0, \infty)$, unless otherwise indicated.
4. The function defined by

$$f(x) = \begin{cases} x & x \text{ rational} \\ -x & x \text{ irrational} \end{cases}$$

Its graph looks something like



Of course this is somewhat misleading, as both the rationals and the irrationals are dense in \mathbb{R} .

5. The function defined by

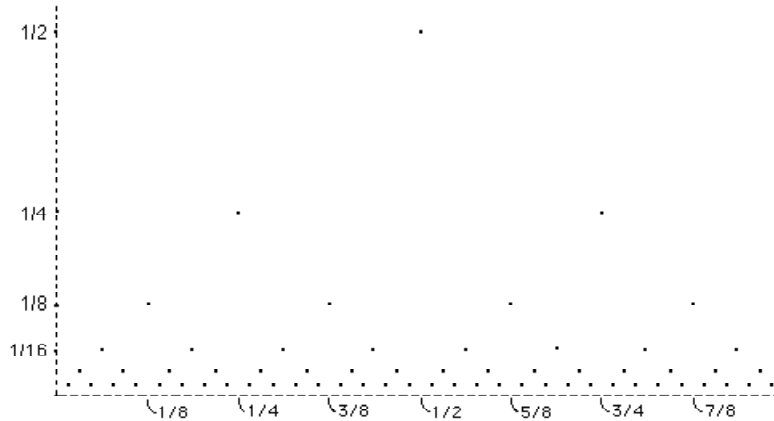
$$f(x) = \frac{x^2 - 1}{x - 1}$$

has domain $\{x : x \neq 1\}$ according to our conventions. For each x in the domain we see that $f(x) = x + 1$. However, f is not the same as the function g defined by

$$g(x) = x + 1,$$

since the domain of g , unless we specify otherwise, is *all* of \mathbb{R} . Of course we could “extend” the domain of f by defining $f(1) = 2$, and the extended function would then be the same as g .

- 6.



An approximation to the graph of g

The function g with domain $[0, 1]$ defined by

$$g(x) = \begin{cases} 1/2 & \text{if } x = 1/2 \\ 1/4 & \text{if } x = 1/4, 3/4 \\ 1/8 & \text{if } x = 1/8, 3/8, 5/8, 7/8 \\ \vdots & \\ 1/2^k & \text{if } x = 1/2^k, 3/2^k, 5/2^k, \dots, (2^k - 1)/2^k \\ \vdots & \\ 0 & \text{otherwise.} \end{cases}$$

In other words,

$$g(x) = \begin{cases} 1/2^k & \text{if } x = p/2^k \text{ where } k \in \mathbb{N}, p \text{ odd and } 1 \leq p < 2^k \\ 0 & \text{otherwise, for } 0 \leq x \leq 1. \end{cases}$$

We can add, subtract, divide and multiply functions to get new functions. We can also multiply a function by a real number to get a new function.

DEFINITION 2.21. Let f and g be functions and c be a real number. Then we define functions $f + g$, $f - g$, fg , f/g and cf by

$$\begin{aligned} (f + g)(x) &= f(x) + g(x) \\ (f - g)(x) &= f(x) - g(x) \\ (fg)(x) &= f(x)g(x) \\ (f/g)(x) &= f(x)/g(x) \\ (cf)(x) &= cf(x). \end{aligned}$$

We restrict x in all cases to be a member of the domains of both f and g , and in the fourth case we also require that $g(x) \neq 0$.

We can also combine two functions by taking the “output” of one to be the “input” of the other.

DEFINITION 2.22. If f and g are two functions, then the *composition* of f and g is defined by

$$(f \circ g)(x) = f(g(x))$$

for all $x \in \mathcal{D}(f)$ such that $f(x) \in \mathcal{D}(g)$.

For example, if $h(x) = |g(x)|$ (for all $x \in \mathcal{D}(f)$) then h is the composition of the “absolute value” function f , given by $f(y) = |y|$, with g . See [Adams, pages 33–36] for further discussion and examples.

2.7. Mathematical induction

Suppose we want to prove some statement is true for every integer n greater than or equal to some fixed integer n_0 . We often do this with the following.

Principle of Mathematical Induction. *Suppose that for some statement $P(n)$ about integers n we know*

- *The statement $P(n_0)$ is true;*
- *if the statement $P(n)$ is true for some integer $n \geq n_0$ then the statement $P(n+1)$ is also true.*

In this case, the statement $P(n)$ is true for all integers $n \geq n_0$.

For an example, see the proof of Theorem 6.6.

The Principle of Mathematical Induction is easy to justify informally. By the first assumption, $P(n_0)$ is true. By the second assumption applied with n replaced by n_0 it follows $P(n_0 + 1)$ is true. By the second assumption applied with n replaced by $n_0 + 1$ it then follows $P(n_0 + 2)$ is true. eEtc.

REMARK 2.23. ★ If one wants to give a more careful proof of the Principle of Mathematical Induction, then one *first* needs to give a more careful definition of the set \mathbb{N} .

One can do this by *defining* \mathbb{N} to be the set of all real numbers which belong to every *inductive* set S , where a set S is defined to be inductive if it has the property that $1 \in S$ and that if $x \in S$ for some real number x , then also $x + 1 \in S$. One then shows that \mathbb{N} itself is inductive and so is the “smallest” inductive set.

In order now to prove the Principle of Mathematical Induction in case $n_0 = 1$, suppose the two assumptions of the Principle are true and let T be the set of all integers n for which $P(n)$ is true. Then T is inductive (*why?*), and so every member of \mathbb{N} is also in T as \mathbb{N} is the smallest inductive set (conversely, every member of T is in \mathbb{N} as T was already assumed to be a set of integers).

The proof in case $n_0 > 1$ is now easy. Just take T to be the set of integers $\{1, \dots, n_0 - 1\}$ together with the set of integers $n \geq n_0$ for which $P(n)$ is true.

There is also a stronger version called the Principle of Complete Mathematical Induction, in which we may assume not only that $P(n)$ is true, but also that $P(n_0), \dots, P(n-1)$ are true.

Principle of Complete Mathematical Induction. *Suppose that for some statement $P(n)$ about integers n we know*

- *The statement $P(n_0)$ is true;*
- *if the statements $P(n_0), \dots, P(n)$ are true for some integer $n \geq n_0$ then the statement $P(n+1)$ is also true.*

In this case, the statement $P(n)$ is true for all integers $n \geq n_0$.

Once again, it is easy to justify informally. By the first assumption, $P(n_0)$ is true. By the second assumption applied with n replaced by n_0 it follows $P(n_0 + 1)$ is true. By the second assumption applied with n replaced by $n_0 + 1$ it then follows $P(n_0 + 2)$ is true. etc.

★ The rigorous proof is similar to that in the case of ordinary induction.

EXERCISE 2.24. The Fibonacci sequence (see Section 4.1) is defined by

$$a_1 = 1, a_2 = 1, a_n = a_{n-1} + a_{n-2} \text{ if } n \geq 2.$$

Prove by complete induction that

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

Fibonacci first came up with this sequence as a model for rabbit population growth. Let a_n be the number of pairs born in the n th month. He assumed there was one pair born the first month and one in the second. After that he assumed that in the n th month there is one pair born for each pair born in the previous month and for each pair born two months ago.

The number of seeds in any ring (layer) of a pine cone is a member of the Fibonacci sequence.

There is an entire journal devoted to Fibonacci sequences, *The Fibonacci Quarterly*.

2.8. ★Fields

The real numbers and the irrationals, as well as the integers modulo a fixed prime number, form a field.

Any set S , together with two operations \oplus and \otimes and two members 0_{\oplus} and 1_{\otimes} of S , which satisfies the corresponding versions of A1–A9, is called a *field*.

Thus \mathbb{R} (together with the operations of addition and multiplication and the two real numbers 0 and 1) is a field. The same is true for \mathbb{Q} , but not for \mathbb{Z} since the analogue of axiom A8 does not hold, *why?*

An interesting example of a field is the set

$$F_p = \{0, 1, \dots, p-1\}$$

for any fixed *prime* p , together with addition and multiplication defined “modulo p ”; i.e. one performs the usual operations of addition and multiplication, but then takes the “remainder” after dividing by p .

Thus for $p = 5$ one has:

\oplus	0	1	2	3	4	\otimes	0	1	2	3	4
0	0	1	2	3	4	0	0	0	0	0	0
1	1	2	3	4	0	1	0	1	2	3	4
2	2	3	4	0	1	2	0	2	4	1	3
3	3	4	0	1	2	3	0	3	1	4	2
4	4	0	1	2	3	4	0	4	3	2	1

It is not too hard to convince yourself that the analogues of axioms A1–A9 hold for any prime p . The axiom which fails if p is not prime is A8, *why?* Note that since F_p is a field, we can solve simultaneous linear equations in F_p .

The fields F_p are very important in coding theory and cryptography.

2.9. ★Deductions from the axioms

We sketch how the usual algebraic and order properties of the real numbers follow rigorously from the axioms A1–A13.

Each line in the following proofs will

1. be an example of one (or occasionally more) of axioms A1–A9;
2. or be a previously proved result;
3. or follow from previously proved results by rules of logic⁹ including the properties of equality.

PROOF OF THEOREM 2.3. *Fill in the missing steps, and go through the proofs line by line and indicate what is used in each step.*

1. Write out your own proof, following the ideas of the proof of the similar result for addition.
2. The trick here is to use the fact $0+0=0$ (from A3), together with the distributive axiom A9. The proof is as follows:

PROOF. One has $a(0+0) = a0$
 But the left side equals $a0 + a0$
 and the right side equals $0 + a0$
 Hence $a0 + a0 = 0 + a0$
 Hence $a0 = 0$. □

⁹For example, if we prove that some statement P implies another statement Q , and if we also prove that P is true, then it follows from rules of logic that Q is true.

3. PROOF. We want to show $-(-a) = a$.
By $-(-a)$ we mean the negative of $-a$, and hence by A4 we know that $-(-a)$ is the *unique* number which when added to $-a$ gives 0.¹⁰ In other words,

$$(-a) + (-(-a)) = 0$$

and if

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

then we *must* have $x = -(-a)$.

Thus if we can also show that

$$(2.6) \quad (-a) + a = 0$$

then it will follow that $a = -(-a)$!!

But (2.6) is just A3, and so we are done. □

The proof can be written more precisely as follows:

PROOF. From the second equality in A4 one has

$$(2.7) \quad (-a) + a = 0.$$

From the first equality in A4 with a there replaced by $(-a)$ one has $(-a) + (-(-a)) = 0$, and moreover,

$$\text{if } (-a) + x = 0 \text{ then } x = -(-a).$$

Hence, from (2.7), $a = -(-a)$. □

4. Write out your own proof, along similar lines to the preceding proof.
5. (As in the proof of 2) it is sufficient to show $(-1)a + a = 0$ (*why?*)

The proof is as follows:

$$\begin{aligned} \text{PROOF. } & (-1)a + a = (-1)a + 1a \\ & = a((-1) + 1) \text{ (two axioms were used for this step)} \\ & = a0 \\ & = 0 \end{aligned}$$

Hence $-a = (-1)a$ from the uniqueness part of A4. □

6. PROOF. $a(-b) = a((-1)b)$
 $= (a(-1))b$
 $= ((-1)a)b$
 $= (-1)(ab)$
 $= -(ab)$

Prove the second equality yourself. □

7. Prove this yourself using, in particular, 4 and A9

8. PROOF. $(-a)(-b) = ((-1)a)(-b)$
 $= (-1)(a(-b))$
 $= -(a(-b))$
 $= -(-ab)$
 $= ab$ □

9. PROOF. First note that $(a/c)(b/d) = (ac^{-1})(bd^{-1})$

and $(ab)/(cd) = (ab)(cd)^{-1}$.

But $(ac^{-1})(bd^{-1}) = (ab)(c^{-1}d^{-1})$

(fill in the steps to prove this equality; which involve a number of applications of A5 and A6).

If we can show that $c^{-1}d^{-1} = (cd)^{-1}$ then we are done.

Since, by A8, $(cd)^{-1}$ is the *unique* real number such that $(cd)(cd)^{-1} = 1$, it is

¹⁰Since a can represent *any* number in A4, we can replace a in A4 by $-a$. This might seem strange at first, but it is quite legitimate.

sufficient to show¹¹ that $(cd)(c^{-1}d^{-1}) = 1$.

Do this; use A5–A8 □

Important Remark: There is a tricky point in the preceding that is easy to overlook; but will introduce some important ideas about logical reasoning.

We used the number $(cd)^{-1}$.

To do this we need to know that $cd \neq 0$.

We know that $c \neq 0$ and $d \neq 0$ and we want to prove that $cd \neq 0$.

This is *equivalent* to proving that if $cd \neq 0$ is false, i.e. if $cd = 0$, then at least one of $c \neq 0$ and $d \neq 0$ is false, i.e. at least one of $c = 0$ or $d = 0$ is true.

In other words, we want to show that if $cd = 0$ then either $c = 0$ or $d = 0$ (possibly both).

The argument is written out as follows:

Claim: If $c \neq 0$ and $d \neq 0$ then $cd \neq 0$

PROOF. We will establish the claim by proving that if $cd = 0$ then $c = 0$ or $d = 0$.¹²

There are two possibilities concerning c ;

either $c = 0$, in which case we are done

or $c \neq 0$. But in this case, since $cd = 0$, it follows

$$c^{-1}(cd) = c^{-1}0 \text{ and so}$$

$$d = 0$$

why?; fill in the steps

Thus we have shown that if $cd = 0$ then $c = 0$ or $d = 0$. Equivalently, if $c \neq 0$ and $d \neq 0$, then $cd \neq 0$. □

10. Exercise

HINT: We want to prove

$$ac^{-1} + bd^{-1} = (ad + bc)(cd)^{-1}.$$

First prove that

$$(ac^{-1} + bd^{-1})(cd) = ad + bc.$$

Then deduce the required result.

And now the proofs of Theorem 2.4

- 1.
- 2.
- 3.
- 4.

2.10. ★Existence and Uniqueness of the Real Number System

We began by assuming that the real number system satisfies Axioms A1–A14. But it is possible to go back even further and begin with axioms for set theory, and then *prove* the existence of a set of objects satisfying Axioms A1–A14.

This is done by first constructing the natural numbers, then the integers, then the rationals, and finally the reals. The natural numbers are constructed as certain types of sets, the negative integers are constructed from the natural numbers, the rationals are constructed as sets of ordered pairs as in [Birkhoff and MacLane, Chapter II-2]. Finally, the reals are then constructed by the method of Dedekind Cuts as in in [Birkhoff and MacLane, Chapter IV-5] or the method of Cauchy Sequences as in [Spivak, Chapter 28].

The real number system is uniquely characterised by Axioms A1–A14, in the sense that any two structures satisfying the axioms are essentially the same up to a renaming of

¹¹When we say “it is sufficient to show . . . ” we mean that if we can show . . . then the result we want will follow.

¹²Note; in mathematics, if we say “***” or “###” (is true) then we *always* include the possibility that *both* “***” and “###” are true.

the members. More precisely, the two systems are *isomorphic*, see [**Birkhoff and MacLane**, Chapter IV-5] or [**Spivak**, Chapter 29].

CHAPTER 3

Limits

The reference for this chapter is generally [Adams, Chapter 1]. In particular and more precisely, Section 1.2, 1.5 to page 87, Appendix III page A-21, and Definition 4 page 644 and the preceding remarks.

3.1. Introduction

Calculus was first developed independently by Newton and Leibniz in the 1660's and 1670's.¹ The notation $\frac{dy}{dx}$ is due to Leibniz. Both thought of y as depending on x via some formula, and of $\frac{dy}{dx}$ as the ratio of “infinitely small quantities” or as the “ultimate ratio of evanescent increments”. They both had a very good intuition (obviously!), and they and their successors developed much of the calculus over the next 150 years, even though they could not fully understand why their methods worked.

To make their ideas precise and to proceed further, one needed to develop a theory of the approximation process which was involved. Newton and Leibniz were unable to do this, and it was not until 1821 that Cauchy introduced a theory of limits and showed how calculus could be based on such a theory.

The difficulty Cauchy overcame was to capture a “dynamic” process by means of a “static” formulation. We may describe a moving particle (a dynamic process) by means of a formula or more generally by means of a function (both of which are static objects), which give the particle's position as a function of time. This was already understood by Newton and Leibniz, but they were not able to “step back” and see that the dynamic process of approximation itself was also one which could be explained in a static way by means of a function. This was done by Cauchy, who worked with the ratio $\frac{y(a+h)-y(a)}{h}$ as a function of h and defined the quantity $\frac{dy}{dx}$ at $x = a$ to be the *limit* of this function as h approached 0.

But Cauchy was unable to give a rigorous definition of the notion of a limit, and it remained for Weierstrass to do this about 15 years later, along the lines of Definition 3.5.

3.2. Definition of limit for a function

We give some examples of limits in order to motivate the formal definition. The notion of a neighbourhood and of a deleted neighbourhood of a number is defined. The definitions of a limit, and of a limit from the right and from the left, are then given. The definition is discussed in terms of always winning a certain game. The definitions are applied to a number of examples.

In order to properly develop a theory of differentiation (as indicated in the previous section), and also of integration and of many other mathematical notions, we need to have a precise theory concerning the notion of limit of a function.

We begin with some examples in order to motivate the definition.

In this chapter and elsewhere, we will use the standard properties of the trigonometric, and occasionally exponential and logarithmic functions, to provide interesting examples. See [Adams, pp 40–52, 209–223] for somewhat informal discussions. But we will not use these functions as part of the rigorous development of the theory.

One can in fact *define* these functions rigorously either as the limit of certain infinite series (after one has developed the theory of infinite series) or as certain definite integrals

¹See [Devlin, pages 79–90], from where I take these notes, and for more discussion.

(after one has developed the theory of integration) or as solutions of certain differential equations (after one has proved Theorem 8.2).

EXAMPLE 3.1 (A Simple One). Let

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

Then we certainly want

$$\lim_{x \rightarrow 2} f(x) = 4.$$

(We say that “the limit as x approaches 2 of $f(x)$ is 4”, or that “ $f(x)$ approaches the limit 4 as x approaches 2”, or that “ f has limit 4 at 2”.) Moreover, if

$$f(x) = \begin{cases} x^2 & x \neq 2 \\ 23 & x = 0 \end{cases},$$

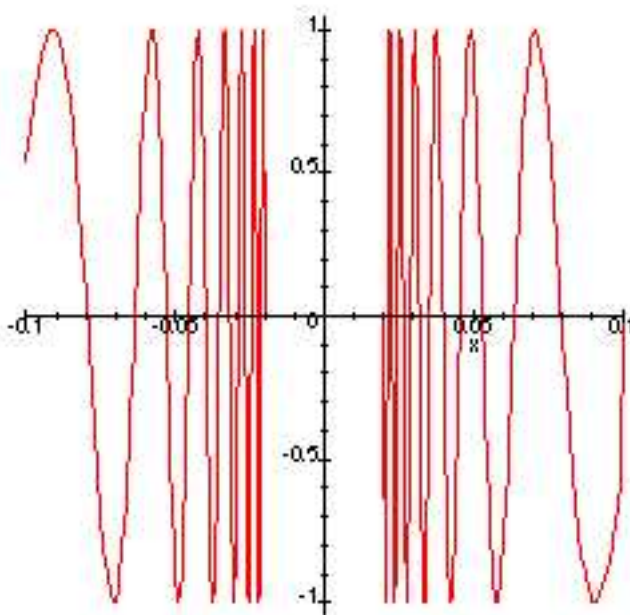
or even if $f(x)$ is defined only for $x \neq 2$, then we still want

$$\lim_{x \rightarrow 2} f(x) = 4.$$

The point is that the *value* of f at 2, or even whether or not f is defined at 2, is irrelevant to the existence and/or value of the *limit* of f at 2.

Thus even though we speak of the limit of f at 2, the notion of the limit at 2 concerns the behaviour of $f(x)$ for x “near” 2 but not when $x = 2$!

EXAMPLE 3.2 (A More Complicated One).



Let

$$f(x) = \begin{cases} \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

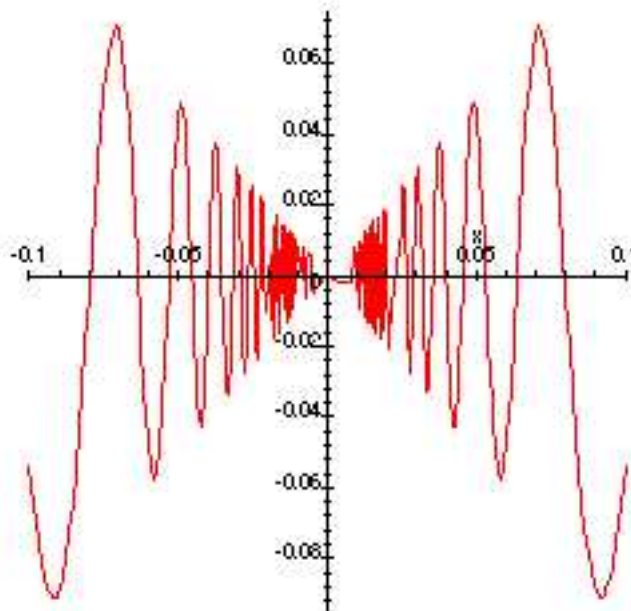
(In the diagram we have not shown the graph in the range $(-.02, .02)$. Sorry about the quality of the diagram!) In this case we want that

$$\lim_{x \rightarrow 0} f(x)$$

does *not* exist, even though there are values of x as close as we wish to 0 such that $f(x) = 0$. (Of course, the same comments apply regardless of the actual value of $f(0)$).

However, if

$$f(x) = \begin{cases} x \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases},$$



then we *do* want that

$$\lim_{x \rightarrow 0} f(x) = 0.$$

Moreover, even if $f(0) = 23$, say, we still want

$$\lim_{x \rightarrow 0} f(x) = 0.$$

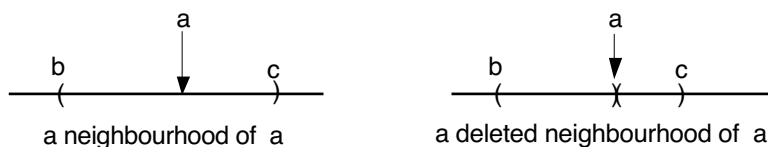
EXAMPLE 3.3 (An Even More Complicated One). If g is the final example of a function in Example 2.20 then we want that

$$\lim_{x \rightarrow a} g(x) = 0$$

for any a such that $0 < a < 1$ (and we also want that the corresponding one-sided limits at $a = 0, 1$ both exist and equal zero). This may seem odd. But the point is that if we take any fixed a , say $a = 1/4$, then (somewhat vaguely) $f(x)$ is as small as we like for *all* x sufficiently close to a (but not equal to a). More precisely, $f(x) < 10^{-23}$ for all $x \neq a$ satisfying $|x - a| < \delta$ (where in this example we can take $\delta = 10^{-23}$ or *any smaller positive number*), and $f(x) < 10^{-100}$ for all $x \neq a$ satisfying $|x - a| < \delta$ (where now, in this example, $\delta = 10^{-100}$ or any smaller positive number).

Before proceeding to the definition of a limit, we need a little notation.

NOTATION 3.4. Let a be a real number. A *neighbourhood* of a is any open interval (b, c) which contains a (thus $b < a < c$).



A *deleted neighbourhood* of a is a neighbourhood of a with a removed. More precisely, a *deleted neighbourhood* of a is a set of numbers of the form

$$N = \{x : b < x < a \text{ or } a < x < c\} = (b, a) \cup (a, c),$$

where again $b < a < c$.

If $\delta > 0$ then the δ -neighbourhood and the δ -deleted-neighbourhood of a are

$$(a - \delta, a + \delta) \quad \text{and} \quad (a - \delta, a + \delta)' := (a - \delta, a) \cup (a, a + \delta)$$

respectively.

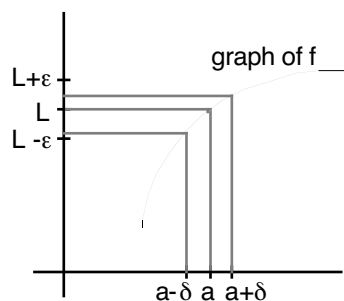
The previous examples lead us to the following definition:

DEFINITION 3.5. Suppose the function $f(x)$ is defined for all x in some deleted neighbourhood of a .² Then we say that *the limit as x approaches a of $f(x)$ is L* ,³ and write

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

if for every $\varepsilon > 0$ there exists a number $\delta > 0$ (which may depend on ε)⁴ such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon.$$



The above diagram is a little misleading, because it is a very simple case. In particular, $f(a) = L$. We could change the value of $f(a)$ to any other number, or even leave f undefined at a , and $\lim_{x \rightarrow a} f(x) = L$ would be unaffected.

Moreover, in the diagram the function f is continuous and increasing in a neighbourhood of a . This is a very simple situation.

We put the statement “which may depend on ε ” in the definition *only* for emphasis. But it is not necessary to do so. In mathematics, when we assert that a number exists (such as δ in the definition), we *always* allow the possibility that it may depend on any previously introduced quantities (such as ε in the case of the definition) unless explicitly stated to the contrary.

It is important to realise that δ *must NOT* depend on x .

NOTE 3.6. The inequalities $0 < |x - a| < \delta$ in Definition 3.5 are equivalent to x belonging to the deleted neighbourhood $(a - \delta, a + \delta)'$. That is

$$\begin{aligned} 0 < |x - a| < \delta \quad \text{if and only if} \quad x \in (a - \delta, a + \delta)' \\ \text{if and only if} \quad (a - \delta < x < a + \delta \text{ and } x \neq a). \end{aligned}$$

Also

$$\begin{aligned} |f(x) - L| < \varepsilon \quad \text{if and only if} \quad f(x) \in (L - \varepsilon, L + \varepsilon) \\ \text{if and only if} \quad L - \varepsilon < f(x) < L + \varepsilon. \end{aligned}$$

Note also that in the definition we are implicitly assuming that x is contained in the original deleted neighbourhood of a (which was part of the domain of f) as well as in

²Of course f may also be defined at a , but this is not required, and is not relevant to the existence or value of the limit of f at a .

³Or “ $f(x)$ approaches L as x approaches a ”, or “the limit of f at a is L , or “ $f(x) \rightarrow L$ as $x \rightarrow a$ ”.

⁴The smaller ε , the smaller δ will normally need to be.

$(a - \delta, a + \delta)'$. This is not a problem, since we can always take a smaller δ if necessary to achieve this.

The definition is not easy to understand; after all it took almost two hundred years to come up with the appropriate ideas.

REMARK 3.7. Another way of thinking about the definition, which may or may not help, is in terms of always being able to win a certain game.

More precisely, your friend sits on the vertical axis and you sit on the horizontal axis. Your friend picks some positive number (let's call it ε) and challenges you to find a positive number (lets call it δ) such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon,$$

i.e.

$$x \in (a - \delta, a + \delta)' \quad \text{implies} \quad f(x) \in (L - \varepsilon, L + \varepsilon).$$

In other words, such that the entire deleted interval $(a - \delta, a + \delta)'$ is mapped by f into the interval $(L - \varepsilon, L + \varepsilon)$. If you can always do this, i.e. *no matter what positive number ε is picked* by your friend you can always find some positive number δ which works for that ε , or in other words if you can *always* win the game, then we say that $\lim_{x \rightarrow a} f(x) = L$.

REMARK 3.8. Suppose $\lim_{x \rightarrow a} f(x) = L$ is *false*, in other words the limit does not exist, or it does exist but equals some number other than L . From the definition, this means that for *some* $\varepsilon > 0$ there is *no* $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon.$$

In other words for this “bad” ε , no matter which $\delta > 0$ you choose, there will be *some* x satisfying

$$0 < |x - a| < \delta \quad \text{and} \quad |f(x) - L| \geq \varepsilon.$$

To show $\lim_{x \rightarrow a} f(x) = L$ is true, we need to show that *every* $\varepsilon > 0$ is “good”. To show $\lim_{x \rightarrow a} f(x) = L$ is false we need to find just *one* “bad” $\varepsilon > 0$.

EXAMPLE 3.9. Consider the function $f(x) = \sin 1/x$ discussed before. Then $\lim_{x \rightarrow a} f(x) = 0$ is false.

To see this, let $\varepsilon = 1/2$ be our candidate for a “bad” ε (any number less than 1 will do). No matter what $\delta > 0$ is chosen (no matter how small!) there will be some x such that

$$0 < |x| < \delta \quad \text{and} \quad |f(x)| \geq \frac{1}{2}.$$

This is clear from the diagram, just choose $x = 1/(n\pi + \frac{1}{2})$ for some sufficiently large integer n , in which case $|f(x)| = 1$.

One needs the precise definition of limit in order to prove general theorems about limits, such as Theorem 3.21, which will apply in all circumstances. On the other hand, once one has proved these theorems, it is usually not necessary in applications to go back to the original definition. So even if you do not not completely understand the definition of limit, that is O.K. Usually you will just need its *consequences*. As we proceed through the course, the definition of limit will become clearer and your understanding of it will increase.

Two very basic examples are the following. Of course they are not surprising. Indeed, if they did not follow from the definition then there would be something wrong with the definition!

THEOREM 3.10.

$$\lim_{x \rightarrow a} x = a \quad \text{and} \quad \lim_{x \rightarrow a} k = k$$

(where k is constant).

PROOF. In order to apply the definition, assume $\varepsilon > 0$.⁵

⁵Thus *all* we are assuming about ε is that it is a positive number. Anything we prove about ε will hence apply to *any* positive number. (Although we are thinking of ε as being small, since this is the important case, the proof will apply to *any* positive number ε .)

We want to show there is some $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |x - a| < \varepsilon.$$

But of course this is true, just take δ to be the same as (or any positive number less than) ε . It now follows from the definition that

$$\lim_{x \rightarrow a} x = a.$$

For the second limit, again assume $\varepsilon > 0$. We want to show there is some $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |k - k| < \varepsilon.$$

This looks silly, but it is in fact true no matter what δ we choose, since $|k - k| = 0$ (regardless of x). Thus it again follows from the definition that

$$\lim_{x \rightarrow a} k = k.$$

□

EXAMPLE 3.11. For the proof from the definition that

$$\lim_{x \rightarrow 2} x^2 = 4$$

(c.f. Example 3.1), see [Adams, page 85].

The proof will probably seem confusing at first, but as indicated before, once we have proved some general theorems, examples like this one are simple consequences.

EXAMPLE 3.12. The fact (see the second diagram in Example 3.2) that

$$\lim_{x \rightarrow 0} x \sin \frac{1}{x} = 0$$

is fairly easy to prove. Take careful note of how one sets out the following proof. I will make certain footnotes about the proof. They would not normally be included, but I have done so here to help understand some of the subtleties involved.

PROOF. Suppose $\varepsilon > 0$.⁶

We want to show there is a $\delta > 0$ such that

$$0 < |x| < \delta \quad \text{implies} \quad \left| x \sin \frac{1}{x} \right| < \varepsilon.$$

But in this example we can simply choose $\delta = \varepsilon$, since if $0 < |x| < \varepsilon$ we have

$$\left| x \sin \frac{1}{x} \right| = |x| \left| \sin \frac{1}{x} \right| \leq \varepsilon \cdot 1 = \varepsilon,$$

i.e. we have

$$\left| x \sin \frac{1}{x} \right| < \varepsilon.$$

To summarise, we have shown

$$0 < |x| < \varepsilon \quad \text{implies} \quad \left| x \sin \frac{1}{x} \right| < \varepsilon.$$

This completes the proof. □

A similar proof (*Exercise*) shows that

$$\lim_{x \rightarrow 2} x^2 \sin \frac{1}{x} = 0.$$

But here one can take $\delta = \sqrt{\varepsilon}$.

One can define *one-sided limits* in a similar manner to ordinary limits.

⁶As before, *all* we are assuming about ε is that it is a positive number. Anything we prove about ε will hence apply to *any* positive number.

DEFINITION 3.13. Suppose the function $f(x)$ is defined for all x in some open interval (a, c) .⁷ Then we say that *the limit, as x approaches a from the right, of $f(x)$ is L* , and we write

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

if for every $\varepsilon > 0$ there exists a number $\delta > 0$ ⁸ such that

$$a < x < a + \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon.$$

Similarly, if the function $f(x)$ is defined for all x in some open interval (b, a) ,

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

if for every $\varepsilon > 0$ there exists a number $\delta > 0$ such that

$$a - \delta < x < a \quad \text{implies} \quad |f(x) - L| < \varepsilon.$$

The symbol $a+$ does *not* mean some number “a little bit bigger than a ”. In fact $a+$ has no meaning by itself, anymore than \rightarrow has a meaning by itself. It is just a way of reminding us that we are taking a limit from the right.

EXAMPLE 3.14. Let

$$f(x) = \begin{cases} 1 & x \leq 0 \\ x^2 & x > 0, \end{cases}$$

Then by arguments similar to those from before

$$\lim_{x \rightarrow 0^-} f(x) = 1, \quad \lim_{x \rightarrow 0^+} f(x) = 0.$$

The following examples can easily be modified to give limits in the usual sense (i.e. not just one-sided).

EXAMPLE 3.15. Consider the following five functions defined on the interval $(0, \infty)$. (Of course we could extend these functions in many different ways to all $x \in \mathbb{R}$.)

$$f_1(x) = x$$

$$f_2(x) = 2x$$

$$f_3(x) = x^2$$

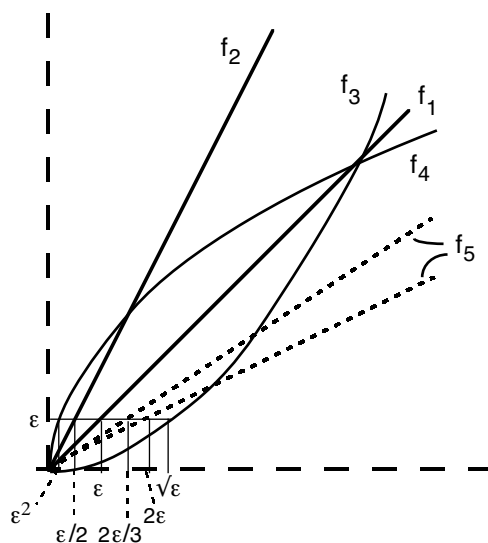
$$f_4(x) = \sqrt{x}$$

$$f_5(x) = \begin{cases} 2x/3 & x \text{ rational} \\ x/2 & x \text{ irrational} \end{cases}$$

⁷We are assuming $a < c$. Of course, $f(x)$ may also be defined for other x .

Note that $f(x)$ may also be defined for $x = a$, and for $x < a$, but this is not required, and is not relevant to the existence or value of the *right* limit of f at a .

⁸As usual, since we do not say otherwise, δ may depend on ε .



The graph of f_5 is, necessarily, somewhat misleading!

It is clear that

$$\lim_{x \rightarrow 0} f_i(x) = 0$$

for each of the functions f_1, \dots, f_5 , since for any given “tolerance” $\varepsilon > 0$ there is some real number $\delta > 0$ such that

$$0 < |x - 0| < \delta \Rightarrow |f(x) - 0| < \varepsilon.$$

For f_1 we take $\delta = \varepsilon$, or any smaller positive number.

For f_2 we take $\delta = \varepsilon/2$, or any smaller positive number.

For f_3 we take $\delta = \sqrt{\varepsilon}$, or any smaller positive number.

For f_4 we take $\delta = \varepsilon^2$, or any smaller positive number.

For f_5 we take $\delta = 2\varepsilon$, or any smaller positive number.

Why?

REMARK 3.16. ★ We can also give a formal definition of what it means for a function to have an infinite limit at a , or to have a limit as $x \rightarrow \infty$ (or $-\infty$). See [Adams, pp 87,88]. This does not introduce any essentially new ideas, but you should have a look in order to reinforce your understanding of the material here.

If we say a function has a limit at a , then we always mean that it has a finite limit unless indicated otherwise.

3.3. Properties of limits of functions

We show that a limit exists if and only if both the two corresponding one-sided limits exist and are equal; that there can be at most one limit at a point; that if a limit is $> K$ (say) then the function is $> K$ in a neighbourhood of the point. We show that limits behave as we expect under addition, multiplication, subtraction and division. We give some simple applications, and finally we prove the Squeeze Theorem.

We now use the formal definition of limit in order to establish the standard properties of limits. The proofs are a little tricky, but the main thing at this stage is to understand the results.

We first show the connection between one-sided limits and ordinary limits.

THEOREM 3.17. *Suppose $f(x)$ is defined for all x in some deleted neighbourhood of a .*

1. *If $\lim_{x \rightarrow a} f(x)$ exists and equals L , then $\lim_{x \rightarrow a^+} f(x)$ and $\lim_{x \rightarrow a^-} f(x)$ both exist and equal L .*
2. *If $\lim_{x \rightarrow a^+} f(x)$ and $\lim_{x \rightarrow a^-} f(x)$ both exist and equal L , then $\lim_{x \rightarrow a} f(x)$ exists and equals L .*

PROOF. Assume

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

In order to show that

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

we have to show that for every number $\varepsilon > 0$ there exists a positive number $\delta > 0$ such that

$$a < x < a + \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon.$$

But the $\delta > 0$ which works in Definition 3.5 will certainly make the above line true (*why?*). This completes the proof for the limit from the right.

Similarly, (*Exercise*)

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

Next assume

$$\lim_{x \rightarrow a^+} f(x) = L, \quad \lim_{x \rightarrow a^-} f(x) = L.$$

In order to show that

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

we have to prove that for every number $\varepsilon > 0$ there exists a positive number $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon.$$

But if $\delta_1, \delta_2 > 0$ work in Definition 3.13 for the limits from the right and the left respectively, then $\delta = \min\{\delta_1, \delta_2\}$ will make the above implication true (*why?*). This completes the proof. \square

But can a function have two *different* limits in the usual sense of limit (i.e. not just one-sided). We do not want this, and another check that our definition correctly captures the idea of a limit is that we can prove this fact.

As in the following proof, I frequently draw a diagram. This is to help construct or understand the proof. But it is *not* part of the proof. *The rigorous proof should always be independent of any diagram.*

THEOREM 3.18. *If*

$$\lim_{x \rightarrow a} f(x) = L_1, \quad \lim_{x \rightarrow a} f(x) = L_2,$$

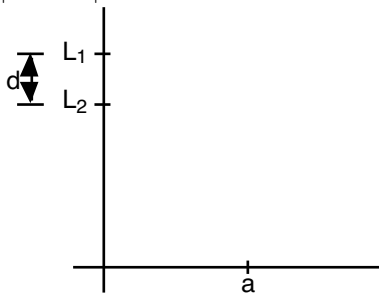
then $L_1 = L_2$. *A similar result applies to limits from the right and to limits from the left.*

PROOF. \star This is an exercise in understanding the definition of limit.

Assume (in order to obtain a contradiction) that

$$\lim_{x \rightarrow a} f(x) = L_1, \quad \lim_{x \rightarrow a} f(x) = L_2,$$

and $L_1 \neq L_2$. Let $d = |L_1 - L_2| > 0$ be the distance between L_1 and L_2 .



Let $\varepsilon = d/3$ in the definition of limit. It follows from the definition that for all x in some deleted neighbourhood of the form

$$(a - \delta_1, a + \delta_1)'$$

we have

$$|f(x) - L_1| < d/3.$$

Similarly, for all x in some deleted neighbourhood of the form

$$(a - \delta_2, a + \delta_2)'$$

we have

$$|f(x) - L_2| < d/3.$$

Thus if we take δ to be the smaller of δ_1 and δ_2 , then

$$x \in (a - \delta, a + \delta)' \quad \text{implies} \quad (|f(x) - L_1| < d/3 \text{ and } |f(x) - L_2| < d/3).$$

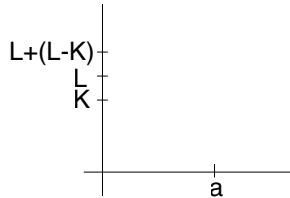
This is impossible (*why?*). Hence our original *assumption* is false and there cannot be two distinct limits. \square

Next we prove the following result.

THEOREM 3.19. *Suppose f is defined in some deleted neighbourhood of a and that $\lim_{x \rightarrow a} f(x) > K$ ($\lim_{x \rightarrow a} f(x) < K$). Then $f(x) > K$ ($f(x) < K$) for all x in some (possibly smaller) deleted neighbourhood of a .*

PROOF. Let L be the limit and first suppose $L > K$. Choose $\varepsilon = L - K$ in the definition of limit. Then there exists some number $\delta > 0$ such that $|f(x) - L| < L - K$ whenever $0 < |x - a| < \delta$. In other words, $L < f(x) < L + (L - K)$ if x is in the deleted neighbourhood $(a - \delta, a + \delta)'$. This proves the result.

If $L < K$ the proof is similar. \square



This theorem is a little more subtle than may at first appear. It requires that f have a limit at a , but f does not need to have a limit anywhere else.

The theorem is not true if $<$ is replaced by \leq , or $>$ is replaced by \geq . For example, $\lim_{x \rightarrow 0} x \sin(1/x) = 0$. But there is no deleted neighbourhood N of 0 such that $x \sin(1/x) \geq 0$, or $x \sin(1/x) \leq 0$, for all $x \in N$.

It follows from the theorem that

COROLLARY 3.20. *If f is defined and $f(x) \geq K$ ($\leq K$) in some deleted neighbourhood of a , and $L = \lim_{x \rightarrow a} f(x)$ exists, then $L \geq K$ ($L \leq K$).*

PROOF. If $L < K$ ($L > K$) then we would get a contradiction by using Theorem 3.19. \square

The previous corollary is not true if \geq (\leq) is replaced by $>$ ($<$). For example, $x^2 > 0$ in any deleted neighbourhood of 0 , but $\lim_{x \rightarrow 0} x^2 = 0$.

Theorem 3.19 is used in the following theorem in order to show that if $\lim_{x \rightarrow a} g(x) = M \neq 0$ then $g(x)$ is nonzero in some deleted neighbourhood of a , and hence the quotient $f(x)/g(x)$ is also defined in some deleted neighbourhood of a .

THEOREM 3.21. Suppose that $f(x)$ and $g(x)$ are defined in some common deleted neighbourhood of a , and that c is a real number. Suppose that

$$\lim_{x \rightarrow a} f(x) = L, \quad \lim_{x \rightarrow a} g(x) = M.$$

Then the following limits exist and have the given values:

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

$$\lim_{x \rightarrow a} cf(x) = cL$$

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

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

In the last case, we require that $M \neq 0$.

PROOF. We begin with the result for the sum of two limits.

Suppose

$$\lim_{x \rightarrow a} f(x) = L, \quad \lim_{x \rightarrow a} g(x) = M.$$

We want to prove that

$$\lim_{x \rightarrow a} f(x) + g(x) = L + M.$$

In order to apply the definition of limit, assume $\varepsilon > 0$. From the definition of limit applied to f and to g , we know there are numbers $\delta_1, \delta_2 > 0$ such that

$$0 < |x - a| < \delta_1 \quad \text{implies} \quad |f(x) - L| < \varepsilon/2,$$

$$0 < |x - a| < \delta_2 \quad \text{implies} \quad |g(x) - M| < \varepsilon/2.$$

(The reason for taking $\varepsilon/2$ instead of ε will be clear in a moment. But note that it is OK to do this, since if $\varepsilon > 0$ also $\varepsilon/2 > 0$.) Using the triangle inequality we have, if $0 < |x - a| < \delta$ where δ is the smaller of δ_1, δ_2 , that⁹

$$\begin{aligned} |(f(x) + g(x)) - (L + M)| &= |(f(x) - L) + (g(x) - M)| \\ &\leq |f(x) - L| + |g(x) - M| < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Since ε was any positive number it follows from the definition of limit that

$$\lim_{x \rightarrow a} f(x) + g(x) = L + M.$$

Exercise: Prove the analogous result for the difference of two limits.

The proof for $cf(x)$ is easier; if δ works for $f(x)$ then $|c|\delta$ works for $cf(x)$. More precisely, suppose

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

If $c = 0$ then the result is immediate, since $0f(x)$ is the constant function which equals zero, and in this case we already know

$$\lim_{x \rightarrow a} 0f(x) = \lim_{x \rightarrow a} 0 = 0 (= 0L).$$

Suppose now that $c \neq 0$. In order to apply the definition of limit, assume $\varepsilon > 0$. From the definition of limit applied to f we know there is a number $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon/|c|.$$

(The reason for taking $\varepsilon/|c|$ instead of ε will be clear in a moment. But note that it is OK to do this, since if $\varepsilon > 0$ also $\varepsilon/|c| > 0$.) It follows that $0 < |x - a| < \delta$ implies

$$|cf(x) - cL| = |c| |f(x) - L| < |c| \varepsilon/|c| = \varepsilon.$$

Since ε was any positive number it follows from the definition of limit that

$$\lim_{x \rightarrow a} cf(x) = cL.$$

⁹Notice how we use “=”, “ \leq ”, “ $<$ ” etc. in the following. The first quantity = the second, which is \leq the third, which is $<$ the fourth, which = the last; hence the first is $<$ the last.

The case for products is more complicated. So I am going to leave it as a tricky exercise, but with hints. See [Adams, Exercises 32,33 page 89].

The case for quotients is even worse, so of course I will leave that also as an exercise; again with hints. See [Adams, Exercises 34–36 page 89]. \square

REMARK 3.22. The previous theorem is also true for left and right limits at a , in which case f and g need only be defined in an interval of the form $(a, a + \delta)$ or $(a - \delta, a)$ respectively.

After that, we can now compute many limits.

THEOREM 3.23. Let $P(x)$ and $Q(x)$ be polynomials¹⁰ such that $Q(a) \neq 0$. Then

$$\begin{aligned}\lim_{x \rightarrow a} P(x) &= P(a) \\ \lim_{x \rightarrow a} \frac{P(x)}{Q(x)} &= \frac{P(a)}{Q(a)}\end{aligned}$$

PROOF. Let

$$P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n, \quad Q(x) = b_0 + b_1x + b_2x^2 + \cdots + b_mx^m.$$

Then the theorem follows from repeated applications of Theorem 3.10 and Theorem 3.21. \square

EXAMPLE 3.24. Compute the derivative of $f(x) = x^4$ at $x = a$.

Solution: The derivative is defined to be (see Definition 6.1)

$$\begin{aligned}\lim_{x \rightarrow a} \frac{x^4 - a^4}{x - a} &= \lim_{x \rightarrow a} \frac{(x - a)(x + a)(x^2 + a^2)}{x - a} \\ &= \lim_{x \rightarrow a} (x + a)(x^2 + a^2) \\ &= 4a^3\end{aligned}$$

This is all that would usually be written down. But this time we will justify each step carefully.

The first expression is just from the definition of the derivative (see later). The first equality comes from factorising the numerator and denominator.

The second equality is valid because if $x \neq a$ the two corresponding expressions (following “ $\lim_{x \rightarrow a}$ ”) are equal, and since the existence and value of the limit at a is unaffected by the value of the corresponding functions at a . The equality should really be read as saying that *if* the limit on the second line exists, *then* so does the limit on the first line, and the two limits are moreover equal.

The third equality comes from Theorem 3.23. More precisely, by Theorem 3.23 the limit on the second line exists (and hence, as just discussed, so does the limit on the first).

There is another important theorem for computing limits. See [Adams, pages 64,65].

THEOREM 3.25 (Squeeze Theorem). Suppose that the functions f , g and h are all defined in some deleted neighbourhood of a and that

$$f(x) \leq g(x) \leq h(x)$$

for every x in this deleted neighbourhood. Suppose also that $\lim_{x \rightarrow a} f(x)$ and $\lim_{x \rightarrow a} h(x)$ exist and both equal L (say). Then $\lim_{x \rightarrow a} g(x)$ exists and

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

PROOF. As usual, let $\varepsilon > 0$ be any positive number.

From the definition of limit applied to f and h , there exist $\delta_1, \delta_2 > 0$ such that

$$\begin{aligned}0 < |x - a| < \delta_1 &\text{ implies } |f(x) - L| < \varepsilon \\ 0 < |x - a| < \delta_2 &\text{ implies } |h(x) - L| < \varepsilon\end{aligned}$$

So if δ is the smaller of δ_1 and δ_2 then $0 < |x - a| < \delta$ implies (see Note 3.6)

$$L - \varepsilon < f(x) \leq g(x) \leq h(x) < L + \varepsilon.$$

¹⁰A *polynomial* is a function of the form $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$. A *rational function* is a function of the form $P(x)/Q(x)$ where P and Q are polynomials.

In particular, $L - \varepsilon < g(x) < L + \varepsilon$, i.e. $|g(x) - L| < \varepsilon$.

It follows from the definition of limit that

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

□

For another proof, see [Adams, Exercise 38 page 89].

All the previous theorems have obvious analogues, with almost exactly the same proofs, for one-sided limits. We will use such results as necessary.

3.4. ★Functions of more than one variable

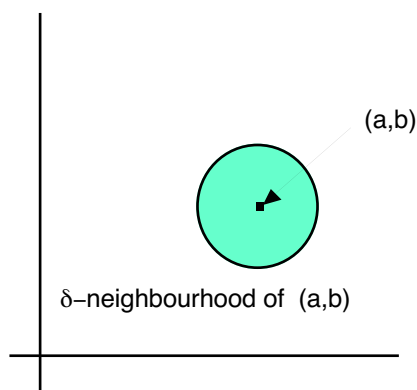
See the Appropriate parts of [Adams, Section 13.2].

If (a, b) is a point in \mathbb{R}^2 and δ is a positive number, then the δ -neighbourhood of (a, b) is the set of all points whose distance to (a, b) is less than δ . In other words, the set

$$\{(x, y) : \sqrt{|x - a|^2 + |y - b|^2} < \delta\}.$$

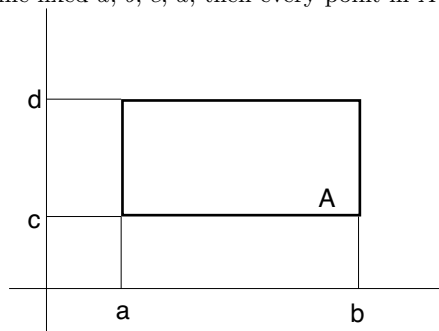
The δ -deleted-neighbourhood is the set of all points, other than (a, b) itself, whose distance to (a, b) is less than δ . In other words

$$\{(x, y) : 0 < \sqrt{|x - a|^2 + |y - b|^2} < \delta\}.$$



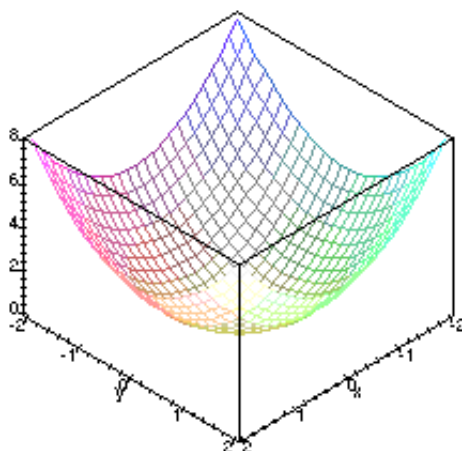
If A is a set of points in \mathbb{R}^2 and (a, b) is a point in A , then (a, b) is an *interior point* of A if all points in some δ -neighbourhood of (a, b) also belong to A .

For example, if (the *open rectangle*) A is the set of points (x, y) such that $a < x < b$ and $c < y < d$ for some fixed a, b, c, d , then every point in A is an interior point.



If A is a subset of \mathbb{R}^2 then a (real-valued) function f with domain A assigns to each pair $(x, y) \in A$ a number which is denoted by $f(x, y)$.

Sometimes such a function is given by a simple formula, such as $f(x, y) = x^2 + y^2$. And it is often possible to sketch the graph, with the aid of Maple, etc., as in the following diagram of this function.



See [Adams, page 643] for further discussion.

Suppose the (real-valued) function $f(x)$ is defined for all x in some deleted neighbourhood of (a, b) . Then we say that *the limit as (x, y) approaches (a, b) of $f(x, y)$ is L* and write

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

if for every $\varepsilon > 0$ there exists a number $\delta > 0$ (which may depend on ε) such that

$$0 < \sqrt{|x - a|^2 + |y - b|^2} < \delta \quad \text{implies} \quad |f(x, y) - L| < \varepsilon.$$

(One could also consider functions whose values are n -tuples of real numbers for some $n \geq 1$.)

(This definition allows us to define the limit at any interior point of the domain of f . One can extend the definition to apply to “boundary” points of the domain. This is straightforward, but we do not need it.. See [Adams, page 647].)

The properties of limits in Theorems 3.18, 3.19, 3.21 and 3.23, Corollary 3.20 and the Squeeze Theorem, are all true, with almost exactly the same proofs.

CHAPTER 4

Sequences

The reference for this chapter is [Adams, Section 10.1], but we do considerably more material than this.

4.1. Examples of sequences

We introduce the idea of a sequence and give a few examples.

A sequence is an infinite list of numbers with a first, but no last, element. Simple examples are

$$1, 2, 1, 3, 1, 4, \dots$$
$$1, \frac{1}{2}, \frac{1}{3}, \dots$$

A sequence can be written in the form

$$a_1, a_2, a_3, \dots, a_n, \dots$$

More precisely, a sequence is a function f whose domain is the set of natural numbers, where in the above example $f(n) = a_n$. We often just write (a_n) or $(a_n)_{n \geq 1}$ to represent the sequence.

If the pattern is clear, we may just write the first few terms, as in

$$2, 4, 6, 8, \dots$$

The general term a_n may instead be given by a formula, such as

$$a_n = \left(1 + \frac{1}{n}\right)^n,$$

which gives the sequence

$$1 + 1, \left(1 + \frac{1}{2}\right)^2, \left(1 + \frac{1}{3}\right)^3, \dots$$

A sequence may be given by a method for calculating each element of the sequence in terms of the preceding elements. One example is the *Fibonacci sequence*

$$a_1 = 1, a_2 = 1, a_n = a_{n-1} + a_{n-2} \text{ if } n \geq 3.$$

Here the sequence is

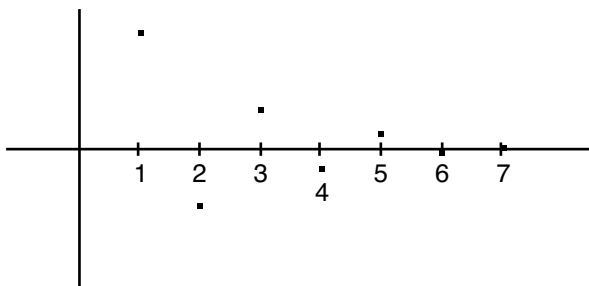
$$1, 1, 2, 3, 5, 8, 13, 21, \dots$$

Sometimes it is convenient to write a sequence in the form

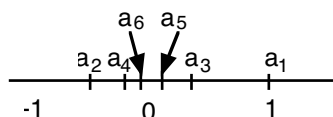
$$a_k, a_{k+1}, a_{k+2}, \dots$$

where k is some other integer than 1 (e.g. 0).

One can represent a sequence by its graph. For example the sequence $((-1)^{n+1}/n^2)$, i.e. $(1, -1/4, 1/9, -1/16, \dots)$ has graph



where the vertical scale is somewhat distorted. However, this is not usually useful. It is often more helpful to think of a sequence as labeled points on the real line.



4.2. Limit of sequences

We define the limit of a sequence and give some examples. We prove that limits of sequences behave as we expect under addition, subtraction, multiplication and division, and the Squeeze Theorem is true. If the terms of a sequence are $\leq p$ then so is the limit. We note that infinite series are also sequences.

We are usually interested in the behaviour of sequences (a_n) for large n . The idea is that “the sequence (a_n) converges to the limit L as n approaches infinity” if the distance between a_n and L approaches zero as n increases. More precisely, if your friend selects a positive number (let’s call it ε) then you can *always* find an integer (let’s call it N , it may depend on ε) such that every member of the sequence after the N th is within ε of L .

For example, the sequence $(1, -1/4, 1/9, -1/16, \dots)$, discussed at the end of the previous section, converges to 0 as n approaches ∞ .

DEFINITION 4.1. We say that *the sequence (a_n) converges to the limit L as n approaches infinity*, and write

$$\lim_{n \rightarrow \infty} a_n = L,$$

if for every positive number ε there exists an integer N (which may depend on ε) such that

$$n > N \quad \text{implies} \quad |a_n - L| < \varepsilon.$$

We sometimes just say a_n *converges to L* and write

$$a_n \rightarrow L, \quad \text{or} \quad a_n \rightarrow L \text{ as } n \rightarrow \infty.$$

(Note that ∞ is *not* a number, and the symbol ∞ by itself here has no meaning, just as \rightarrow has no meaning by itself.)

REMARK 4.2. If $a_n \rightarrow 0$ then $|a_n| \rightarrow 0$, and conversely. This is clear since

$$|a_n - 0| = ||a_n| - 0|.$$

EXAMPLE 4.3. Show that $\lim_{n \rightarrow \infty} \frac{c}{n^p} = 0$ for any real number c and any $p > 0$.

Solution: (See [Adams, Example 4 page 522]). Let $\varepsilon > 0$ be given. Then

$$\left| \frac{c}{n^p} \right| < \varepsilon \quad \text{if} \quad n^p > \frac{|c|}{\varepsilon}, \quad \text{i.e. if} \quad n > \left(\frac{|c|}{\varepsilon} \right)^{1/p}.$$

Thus we can take any integer $N > \left(\frac{|c|}{\varepsilon} \right)^{1/p}$, and it follows that

$$\left| \frac{c}{n^p} \right| < \varepsilon \quad \text{if} \quad n > N.$$

This implies the required limit exists and equals zero.

One can prove the following theorem in the same way as for limits of functions. We will do them in the assignments.

THEOREM 4.4. *Suppose*

$$\lim_{n \rightarrow \infty} a_n = L, \quad \lim_{n \rightarrow \infty} b_n = M,$$

and c is a real number. Then the following limits exist and have the given values.

$$\lim_{n \rightarrow \infty} a_n \pm b_n = L \pm M$$

$$\lim_{n \rightarrow \infty} ca_n = cL$$

$$\lim_{n \rightarrow \infty} a_n b_n = LM$$

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{L}{M}$$

In the last case we require $M \neq 0$.

In the quotient case, the fact $M \neq 0$ implies $b_n \neq 0$ for all $n > N$ (say). The proof is similar to that in Theorem 3.19. This means the expression a_n/b_n is defined for all $n > N$, although it may not be defined for $n \leq N$.

There is also a version of the Squeeze Theorem, proved in a similar manner to the case for functions.

THEOREM 4.5. *Suppose $a_n \leq b_n \leq c_n$ for all n (or at least for all $n \geq N$ for some N). Suppose $a_n \rightarrow L$ and $c_n \rightarrow L$ as $n \rightarrow \infty$. Then $b_n \rightarrow L$ as $n \rightarrow \infty$.*

Suppose $\lim a_n = a$ and $p < a_n < q$ for all n . Then it does not follow that $p < a < q$. For example, let $p = 0$, $q = 2$, and $a_n = 1/n$. But it does follow that $p \leq a \leq q$. In fact we only need assume $p \leq a_n \leq q$ to draw the same conclusion, as we now see.

The following is the analogue of Theorem 3.19.

THEOREM 4.6. *Suppose $\lim a_n = a > K$. Then $a_n > K$ for all sufficiently large n (i.e. there exists an integer N , which may depend on K , such that $n > N$ implies $a_n > K$). A similar result applies for $\lim a_n < K$.*

PROOF. Let a be the limit and suppose $a > K$. Choose $\varepsilon = a - K$ in the definition of limit. Then there exists some integer N such that $|a_n - a| < a - K$ whenever $n > N$. In other words, $a - (a - K) < a_n < a + (a - K)$ if $n > N$, and in particular $a_n > K$. This proves the result.

If $a < K$ the proof is similar. □

It follows from the theorem that

COROLLARY 4.7. *Assume $\lim a_n = a$ and $a_n \geq K$ ($a_n \leq K$) for all n . Then $a \geq K$ ($a \leq K$).*

PROOF. If $a < K$ ($a > K$) then we get a contradiction by using Theorem 3.19. □

It is not true that $a_n < p$ for all n implies $\lim a_n < p$. (Consider the sequence $a_n = -1/n$ and $p = 0$.) But it is true from the previous theorem that $\lim a_n \leq p$.

REMARK 4.8. You have probably seen infinite series before, such as the infinite geometric series

$$\sum_{n \geq 0} r^n \quad (\text{or } \sum_{n=0}^{\infty} r^n) = 1 + r + r^2 + r^3 + \dots$$

You probably know that the sum of the first n terms is $(1 - r^n)/(1 - r)$, see [Adams, page 530], and the “sum” of the infinite series is $\frac{1}{1-r}$ provided $|r| < 1$.

In fact an infinite series is just a particular type of infinite sequence. More precisely, we can replace the infinite geometric series by the sequence of its *partial sums*:

$$1, 1 + r, 1 + r + r^2, \dots$$

When we say that a series such as $\sum_{n \geq 1} a_n$ converges, we mean that the *sequence* of partial sums

$$s_1 = a_1, s_2 = a_1 + a_2, s_3 = a_1 + a_2 + a_3, \dots$$

converges.

4.3. Monotone sequences

Bounded monotone sequences are convergent. If the absolute value of the difference between consecutive members of a sequence decreases at least geometrically fast, then the sequence converges.

A sequence $(a_n)_{n \geq 1}$ is *increasing* if

$$a_1 \leq a_2 \leq \dots \leq \dots \leq a_n \leq \dots,$$

and is *decreasing* if

$$a_1 \geq a_2 \geq \dots \geq \dots \geq a_n \geq \dots$$

A sequence is *monotone* if it is either increasing or decreasing.

A sequence $(a_n)_{n \geq 1}$ is *bounded above* if there is some number K such that

$$a_n \leq K \quad \text{for every } n,$$

and *bounded below* if there is some number J such that

$$a_n \geq J \quad \text{for every } n.$$

Thus the sequence $(1, 2, 3, \dots)$ is bounded below, but not above.

We prove in the next theorem (using the completeness axiom) that every increasing sequence which is bounded above converges to a limit. It is not surprising that we need the completeness axiom, because the analogous result is not true for the rational numbers. For example, consider the sequence

$$1, 1.4, 1.414, 1.4142, 1.41421, 1.414213, 1.4142135, 1.41421356, \dots,$$

which is obtained from the decimal expansion of $\sqrt{2}$. This is an increasing sequence of *rational* numbers, but the limit is *irrational*.

THEOREM 4.9. *Let $(a_n)_{n \geq 1}$ be an increasing sequence which is bounded above, or a decreasing sequence which is bounded below. Then the sequence has a limit.*

PROOF. Let $(a_n)_{n \geq 1}$ be an increasing sequence which is bounded above. The corresponding set¹ S of numbers is bounded above and so there is a l. u. b. K_0 , say.

We claim that

$$a_n \rightarrow K_0.$$

To prove the claim, let $\varepsilon > 0$ be any positive number.

Since K_0 is the *least* upper bound for S , there is some $a_N \in S$ (depending on ε) such that

$$a_N > K_0 - \varepsilon.$$

(Otherwise $K_0 - \varepsilon$ would also be an upper bound for S , contradicting the fact K_0 is the *least* upper bound.)

Since the sequence (a_n) is increasing, it follows that

$$n \geq N \quad \text{implies} \quad a_n > K_0 - \varepsilon.$$

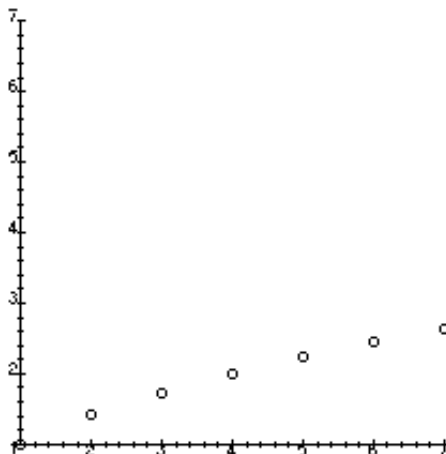
But we also know that $a_n \leq K_0$ for all n , and so $n \geq N$ implies $K_0 - \varepsilon < a_n \leq K_0$, and in particular implies $|a_n - K_0| < \varepsilon$.

It follows that $a_n \rightarrow K_0$. A similar proof applies if (a_n) is decreasing and bounded below. \square

¹There is a difference between a sequence and the corresponding set of numbers. In the first case, the order is important, but not in the second. Thus the sequences $(1, 2, 3, 4, 5, \dots)$ and $(1, 2, 1, 3, 1, 4, 1, 5, \dots)$ are different but have the same corresponding set of real numbers.

It is not necessarily true that if the difference between *consecutive* members of a sequence converges to zero, then the sequence converges. For example, consider the sequence $a_n = \sqrt{n}$, i.e.

$$1, \sqrt{2}, \sqrt{3}, \sqrt{4}, \dots$$



The sequence is unbounded and it is clear from the graph that $a_{n+1} - a_n \rightarrow 0$ as $n \rightarrow \infty$. To prove this analytically, write

$$\begin{aligned} \lim_{n \rightarrow \infty} \sqrt{n+1} - \sqrt{n} &= \lim_{n \rightarrow \infty} (\sqrt{n+1} - \sqrt{n}) \frac{\sqrt{n+1} + \sqrt{n}}{\sqrt{n+1} + \sqrt{n}} \\ &= \lim_{n \rightarrow \infty} \frac{(n+1) - n}{\sqrt{n+1} + \sqrt{n}} = \frac{1}{\sqrt{n+1} + \sqrt{n}} = 0. \end{aligned}$$

The last limit is zero since, for any $\varepsilon > 0$, we can choose N (depending on ε) such that $\frac{1}{\sqrt{n+1} + \sqrt{n}} < \varepsilon$ for all $n > N$ (for example, any integer $N > \varepsilon^{-2}$ will do, *exercise*).

However, if the difference between consecutive members of a sequence converges to zero “sufficiently fast” then the sequence will converge. In particular, we have the following theorem for sequences such that the absolute value of the difference between consecutive terms approaches zero “geometrically” fast. The sequence is *not* required to be monotone, but may “oscillate” around its limit.

THEOREM 4.10. *Suppose the sequence $(a_n)_{n \geq 1}$ satisfies $|a_n - a_{n+1}| \leq Mr^n$ for some real numbers M and r such that $0 \leq r < 1$. Then $a_n \rightarrow L$ for some real number L . Moreover, $|a_n - L| \leq Mr^n \frac{1}{1-r}$.*

PROOF. The difficulty is that the sequence (a_n) may be neither increasing nor decreasing. The (non-obvious) trick is to write

$$a_n = a_1 + (a_2 - a_1) + \dots + (a_n - a_{n-1}).$$

Now consider the corresponding *increasing* sequence whose n th term is

$$c_n = |a_1| + |a_2 - a_1| + \dots + |a_n - a_{n-1}|.$$

Then (c_n) is an increasing sequence as c_{n+1} is obtained from c_n by adding the term $|a_{n+1} - a_n|$. Moreover the sequence is bounded above, since

$$\begin{aligned} c_n &\leq |a_1| + Mr + Mr^2 + \dots + Mr^{n-1} \\ &= |a_1| + Mr(1 + r + \dots + r^{n-2}) = |a_1| + Mr \frac{1 - r^{n-1}}{1 - r} \leq |a_1| + Mr \frac{1}{1 - r} = K_0, \end{aligned}$$

say, where the second equality is from [Adams, page 530]. Since the sequence (c_n) is increasing and bounded above, it follows that $c_n \rightarrow K$, say.

The idea is now to express the sequence (a_n) as the difference of two increasing bounded sequences.

Write

$$\begin{aligned} a_n &= a_1 + (a_2 - a_1) + \cdots + (a_n - a_{n-1}) \\ &= |a_1| - (|a_1| - a_1) \\ &\quad + |a_2 - a_1| - (|a_2 - a_1| - (a_2 - a_1)) \\ &\quad + |a_3 - a_2| - (|a_3 - a_2| - (a_3 - a_2)) \\ &\quad \vdots \\ &\quad + |a_n - a_{n-1}| - (|a_n - a_{n-1}| - (a_n - a_{n-1})) \\ &= c_n - b_n, \end{aligned}$$

where

$$b_n = (|a_1| - a_1) + (|a_2 - a_1| - (a_2 - a_1)) + \cdots + (|a_n - a_{n-1}| - (a_n - a_{n-1})).$$

Since b_{n+1} is obtained from b_n by adding the positive term $(|a_n - a_{n-1}| - (a_n - a_{n-1}))$, it follows that the sequence (b_n) is increasing. It follows that $b_n \rightarrow J$, say.

Finally, since

$$a_n = c_n - b_n,$$

it follows that the sequence (a_n) also converges (to $K - J$).

Let $\lim a_n = L$. To prove the estimate for $|a_n - L|$, suppose $N > n$ and write $N = n + k$. Then

$$\begin{aligned} |a_n - a_N| &= |a_n - a_{n+k}| \\ &= |(a_n - a_{n+1}) + (a_{n+1} - a_{n+2}) + (a_{n+2} - a_{n+3}) + \cdots + (a_{n+k-1} - a_{n+k})| \\ &\leq |a_n - a_{n+1}| + |a_{n+1} - a_{n+2}| + |a_{n+2} - a_{n+3}| + \cdots + |a_{n+k-1} - a_{n+k}| \\ &\leq Mr^n + Mr^{n+1} + Mr^{n+2} + \cdots + Mr^{n+k-1} \\ &= Mr^n(1 + r + r^2 + \cdots + r^{k-1}) \\ &= Mr^n \frac{1 - r^k}{1 - r} \quad \text{see [Adams, p. 530]} \\ &\leq Mr^n \frac{1}{1 - r} \end{aligned}$$

Thus $|a_n - a_N| \leq Mr^n/(1-r)$ for all $N > n$, and so $|a_n - L| \leq Mr^n/(1-r)$ by Corollary 4.7 applied to the sequence

$$|a_n - a_{n+1}|, |a_{n+1} - a_{n+2}|, |a_{n+2} - a_{n+3}|, \dots,$$

(where n is fixed). □

REMARK 4.11. ★ There is a more general result, the *Cauchy convergence criterion*, which implies Theorem 4.10.

Namely, if a sequence (a_n) has the property that for any $\varepsilon > 0$ there exists an integer N (which may depend on ε) such that the difference between *any* two (not necessarily consecutive) members of the sequence after the N th is $< \varepsilon$, then $a_n \rightarrow L$ for some real number L .

That is, suppose for every $\varepsilon > 0$ there exists an integer N (which may depend on ε) such that

$$m, n > N \quad \text{implies} \quad |a_m - a_n| < \varepsilon.$$

Then $a_m \rightarrow L$ for some real number L . A sequence satisfying the above assumption is called a *Cauchy sequence*.

The converse is also true: every convergent sequence is Cauchy.

The advantage of the Cauchy criterion is that we have a criterion for convergence for an arbitrary sequence which does not depend on knowing the limit beforehand.

See [Spivak] for a proof of these (very important) results. We will not give the proofs here (except as a very challenging exercise) since the previous theorem is sufficient for our purposes. We *will* give the proof next year in a more general setting. The proof that if a sequence converges then it is Cauchy is fairly straightforward. The other direction is tricky, and one needs the completeness axiom.

It is not too hard to check that if a sequence satisfies the hypothesis of the previous theorem then it is Cauchy.

4.4. Limits for functions via limits for sequences

We show that we could have defined and developed the theory of limits for functions in terms of limits of sequences.

It is perhaps easier to understand limits for sequences than it is to understand limits for functions. As it happens, there is a nice way to understand limits for functions *in terms of* limits for sequences.

THEOREM 4.12. *Suppose $f(x)$ is defined for all x in some deleted neighbourhood of a . Then the following two statements are equivalent:*

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

$$a_n \rightarrow a \text{ (and } a_n \neq a \text{ for all } n) \quad \text{implies} \quad f(a_n) \rightarrow L$$

REMARK 4.13.

- For example, $\lim_{x \rightarrow 0} x \sin \frac{1}{x} = 0$ because if $a_n \rightarrow 0$ then $a_n \sin 1/a_n \rightarrow 0$. (But to prove this, one essentially has to go through the same argument as in Example 3.12.)
- The restriction in the theorem that $a_n \neq a$ for all n is necessary, since $f(a)$ may not be defined. And even if $f(a)$ were defined, it may not equal the limit L — in this case $f(a_n)$ would not converge to L if there were infinitely many $a_n = a$.

PROOF. [Since we have to prove two statements are equivalent, we have to prove that the first implies the second *and* that the second implies the first.]

- Suppose first that

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

We want to prove that

$$a_n \rightarrow a \text{ (} a_n \neq a \text{)} \quad \text{implies} \quad f(a_n) \rightarrow L.$$

Thus we *assume* that $a_n \rightarrow a$, and $a_n \neq a$ for each n . In order to use the definition to prove the sequence $f(a_n) \rightarrow L$, let $\varepsilon > 0$ be any positive number.

Since $\lim_{x \rightarrow a} f(x) = L$, there exists some real number $\delta > 0$ (which may depend on ε) such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon,$$

and in particular

$$0 < |a_n - a| < \delta \quad \text{implies} \quad |f(a_n) - L| < \varepsilon,$$

But since $a_n \rightarrow a$ and $a_n \neq a$, there exists an integer N (which may depend on δ and hence on ε) such that

$$n > N \quad \text{implies} \quad 0 < |a_n - a| < \delta.$$

Putting these last two implications together, there exists an integer N such that

$$n > N \quad \text{implies} \quad |f(a_n) - L| < \varepsilon.$$

This says $f(a_n) \rightarrow L$, and so we have shown

$$a_n \rightarrow a \text{ (} a_n \neq a \text{)} \quad \text{implies} \quad f(a_n) \rightarrow L.$$

• ★ Next suppose that

$$(4.1) \quad a_n \rightarrow a \ (a_n \neq a) \quad \text{implies} \quad f(a_n) \rightarrow L.$$

We want to prove

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

[This is a little tricky; it turns out that the best way is to obtain a contradiction from assuming that $\lim_{x \rightarrow a} f(x) = L$ is false. In other words, we assume that the limit does not exist, or that it does exist but is something other than L .]

Assume $\lim_{x \rightarrow a} f(x) = L$ is false. This means that there is some “bad” $\varepsilon > 0$ as in Remark 3.8 for which there is *no* $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \text{implies} \quad |f(x) - L| < \varepsilon.$$

Let $\varepsilon > 0$ be “bad”. Thus if we take $\delta = 1/n$ for each natural number n in the above (false) implication, there is some x (which may depend on n) which we denote by a_n , such that

$$0 < |a_n - a| < \frac{1}{n} \quad \text{and} \quad |f(a_n) - L| \geq \varepsilon \quad (n = 1, 2, 3, \dots).$$

But this just means that the sequence (a_n) converges to a , $a_n \neq a$ for each n , and the sequence $(f(a_n))$ does *not* converge to L .

This contradicts (4.1), and so our *assumption* is incorrect.² In other words, $\lim_{x \rightarrow a} f(x) = L$ is true. \square

²There is no logical difference between “supposing” something (as in (4.1)) or assuming something (as in $\lim_{x \rightarrow a} f(x) = L$ is false). But we are interested in what happens under the circumstances that (4.1) *is* true, and so under *this* set of circumstances we have deduced by contradiction that $\lim_{x \rightarrow a} f(x) = L$ is also true.

CHAPTER 5

Continuity

The references for this chapter are Section 1.4 and Appendix 3 of Adams. But we cover somewhat more material.

5.1. Introduction

Intuitively, a function is continuous if its graph does not have any jumps, breaks, or wild oscillations. Although many functions which arise in applications are continuous, there are many which are not. For example, if an electric current is being switched on and off, then it would probably be best modeled by a discontinuous function.

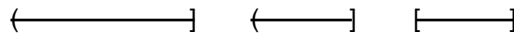
5.2. Definition of continuity

The notion of continuity is defined in terms of limits. Equivalent formulations in terms of $\varepsilon - \delta$'s and sequences are given. Some examples are given.

In future, unless we say otherwise, we make the convention that all functions will be defined on a domain which is a finite union of intervals,

$$I_1 \cup I_2 \cup \cdots \cup I_n.$$

The intervals may be finite or infinite, and open or closed at any finite endpoint. The important case to keep in mind is when the domain is a single interval.



a typical domain

We assume that the intervals have no elements in common with one another. Moreover if two intervals can be combined into a single interval then this is done. For example, we write $[0, 1]$ instead of $[0, 1/2] \cup [1/2, 1]$. But we do allow, for example, $(-1, 0) \cup (0, 1)$, since the union of these intervals is not a single interval.

We also assume that our intervals do *not* consist of a single point.

It follows from our convention that if a is in the domain of the function f (say), then there exists some $\delta > 0$ such that $f(x)$ will be defined for all $x \in (a - \delta, a + \delta)$ (in case a is an interior point of the domain), or $f(x)$ will be defined for all x in some interval of the form $[a, a + \delta)$ or of the form $(a - \delta, a]$ (in case a is an end point of an interval from the domain). We sometimes call $[a, a + \delta)$ and $(a - \delta, a]$ a *one-sided neighbourhood* of a .

It is not difficult to define limits and continuity in a more general setting; the ideas are essentially the same but the notation is a little messier. Moreover, once you understand the ideas in the present setting, it is not difficult to extend them to the more general setting.

DEFINITION 5.1. Suppose a is an interior point of the domain of f . Then f is *continuous at a* if

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

If a is a left (right) endpoint then f is *continuous at a* if

$$\lim_{x \rightarrow a+} f(x) = f(a) \quad \left(\lim_{x \rightarrow a-} f(x) = f(a) \right).$$

REMARK 5.2. In terms of epsilons and deltas the definition says (if a is an interior point): for every $\varepsilon > 0$ there exists a number $\delta > 0$ such that

$$|x - a| < \delta \quad \text{implies} \quad |f(x) - f(a)| < \varepsilon.$$

If a is a left endpoint, the definition says that for every $\varepsilon > 0$ there exists a number $\delta > 0$ such that

$$a \leq x < a + \delta \quad \text{implies} \quad |f(x) - f(a)| < \varepsilon.$$

Similarly for a right endpoint.

This is a little simpler than Definition 3.5 for a limit, since we no longer require $|x - a| > 0$, i.e. that $x \neq a$. The reason is that if $x = a$ we trivially have $f(x) = f(a)$ and so $|f(x) - f(a)| < \varepsilon$ is then trivially true in this case!

We emphasise, that as for limits, $\delta > 0$ will normally depend on ε but must not depend on x .

The following is a useful way to characterise continuity in terms of convergent sequences.

THEOREM 5.3. *The function f is continuous at a point a in its domain if and only if*

$$a_n \rightarrow a \quad \text{implies} \quad f(a_n) \rightarrow f(a)$$

whenever (a_n) is a sequence of points from the domain of f .

PROOF. This essentially follows from the proof of Theorem 4.12.

Suppose f is continuous at a and $a_n \rightarrow a$. We no longer need to exclude the case $a_n = a$ as in the proof of Theorem 4.12, and the proof that $f(a_n) \rightarrow f(a)$ is otherwise the same.

Next suppose that $a_n \rightarrow a$ implies $f(a_n) \rightarrow f(a)$. Then from Theorem 4.12 we have $\lim_{x \rightarrow a} f(x) = f(a)$, which means f is continuous at a . \square

We have defined what it means to be continuous at a point. But the more important situation is for a function to be continuous everywhere on its domain.

DEFINITION 5.4. A function f is *continuous* if it is continuous at every point in its domain.

EXAMPLE 5.5. The function

$$f(x) = \begin{cases} x \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is continuous at 0, because $\lim_{x \rightarrow 0} x \sin \frac{1}{x} = f(0)$ ($0 = 0$). Since f is continuous at *every* point in its domain (by the theorems in the next section), it follows that f is continuous.

However, the function

$$f(x) = \begin{cases} x \sin \frac{1}{x} & x \neq 0 \\ 23 & x = 0 \end{cases}$$

is *not* continuous at 0 because $\lim_{x \rightarrow 0} f(x) (= 0) \neq f(0)$. This function is said to have a *removable* discontinuity at 0 because it is possible to re-define $f(0)$ in such a way that the new function is continuous at 0.

EXAMPLE 5.6. The function $1/|x|$ is continuous on its domain $(-\infty, 0) \cup (0, \infty)$. This may seem surprising, but remember that 0 is not in the domain of the function. There is no way we can extend this function by defining it at 0 so that it becomes continuous on all of \mathbb{R} .

EXAMPLE 5.7. The function $f(x) = |x|$ is continuous on its domain (which is \mathbb{R}). The only point we need to check is continuity at 0.

Here is another example, to show you the need to be cautious. Let (see Example 2.20.4)

$$f(x) = \begin{cases} x & x \text{ rational,} \\ -x & x \text{ irrational.} \end{cases}$$

Then f is continuous at 0, since

$$\lim_{x \rightarrow 0} f(x) = f(0) (= 0).$$

This can be seen directly from the $\varepsilon - \delta$ definition of limit; in fact one can just take $\delta = \varepsilon$ (*why?*).

Alternatively, one can apply the squeeze theorem; just note that

$$-|x| \leq f(x) \leq |x|$$

and use the fact

$$\lim_{x \rightarrow 0} -|x| = \lim_{x \rightarrow 0} |x| = 0.$$

However, $\lim_{x \rightarrow a} f(x)$ does not exist if $a \neq 0$ (*Exercise*), and so f is not continuous at a in this case.

EXERCISE 5.8.

1. For which value(s) of k is the function

$$f(x) = \begin{cases} \frac{x}{|x|} & x \neq 0 \\ k & x = 0 \end{cases}$$

continuous at 0?

2. Suppose $f(x) = n$ if $n \leq x < n + 1$, for every integer n . Show that f is continuous at all non-integer points but discontinuous at every integer point. (It helps to sketch a graph of f .) Such a function is called a *step function*.

5.3. Properties of continuous functions

It is shown that continuity is preserved under algebraic operations and composition

REMARK 5.9. Suppose f is continuous at an interior point a and $f(a) > K$. Then $f(x) > K$ for all x in some neighbourhood of a . This follows immediately from Theorem 3.19. A similar remark applies if $f(a) < K$.

If f is continuous at a left end-point a and $f(a) > K$, then for some $\delta > 0$ one has $f(x) > K$ for all $a \leq x < a + \delta$. Similarly for right endpoints and for $f(a) < K$. The proofs follow from easily modified versions of Theorem 3.19.

Theorem 5.10 and Theorem 5.12 show that if we combine continuous functions in various ways, then the results are also continuous.

THEOREM 5.10. *Suppose f and g are continuous at a , and c is a real number. Then the following are continuous at a :*

$$f \pm g, \quad cf, \quad fg, \quad f/g.$$

In the last case we require that $g(a) \neq 0$.

PROOF. This follows directly from Theorem 3.21. For example, if a is an interior point, then in the quotient case,

$$\lim_{x \rightarrow a} \frac{f}{g}(x) = \lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \frac{f(a)}{g(a)} = \frac{f}{g}(a).$$

The first inequality comes from the fact that $f/g(x)$ is defined to be $f(x)/g(x)$, and similarly for the third equality. The second equality comes from Theorem 3.21. (Recall that if g is continuous at a and $g(a) \neq 0$ then $g(x) \neq 0$ for x in some neighbourhood of a .) It follows that the function is defined in some neighbourhood of a .)

If a is an endpoint the proof is essentially the same, using Example 3.22. \square

Note that if $f(x)/g(x)$ is zero on a complicated set, such as if $g(x) = x \sin(1/x)$, then the domain of f/g may not be a finite union of intervals.

The following is now immediate, but is worth stating separately, since we are usually interested in functions which are continuous on a domain, not just at a point.

THEOREM 5.11. *Suppose f and g are both defined and continuous on the common domain D and c is a real number. Then the following are also continuous on D :*

$$f \pm g, cf, fg, f/g.$$

In the last case we require that $g(x) \neq 0$ for every $x \in D$.

The next theorem proves that the composition $g \circ f$ of two continuous functions f and g is also continuous. For example, the function $|f(x)|$ is continuous if the function $f(x)$ is continuous. This follows from the fact that $|f(x)|$ is the composition of the absolute value function, which we saw was continuous in Example 5.7, with the continuous function f .

In the following, f may be defined in a neighbourhood of a , or perhaps may only be defined in a *one-sided* neighbourhood of a of the form $[a, a + \delta)$ or $(a - \delta, a]$. But g must be defined in an ordinary neighbourhood of $f(a)$. Otherwise, for example, if $f(x) = -x^2$ and $g(y) = \sqrt{y}$, then $(g \circ f)(x) = \sqrt{-x^2}$ is only defined for $x = 0$, and the notion of continuity at 0 is not even defined.

THEOREM 5.12. *Suppose f is a function continuous at a and g is a function continuous at $f(a)$ and which is defined everywhere in some neighbourhood of $f(a)$. Then $g \circ f$ is continuous at a .*

PROOF. To simplify the notation, assume first that a is not an endpoint of the domain of f . Let $\varepsilon > 0$ be any positive number.

[We want to prove there is a $\delta > 0$ (which may depend on ε) such that

$$|x - a| < \delta \quad \text{implies} \quad |g(f(x)) - g(f(a))| < \varepsilon.$$

Remember that $(g \circ f)(x) = g(f(x))$.]

We are given that g is continuous at $f(a)$ and is defined everywhere in some neighbourhood of $f(a)$ (not just in an interval whose endpoint is $f(a)$). Thus there exists $\eta > 0$ ¹ (which may depend on ε) such that

$$|y - f(a)| < \eta \quad \text{implies} \quad |g(y) - g(f(a))| < \varepsilon.$$

[y is just a dummy variable; the meaning would be unchanged if we used z or x throughout. But we do not use x since we are going to use it to represent a number in the domain of f .]

In particular, replacing y by $f(x)$,

$$|f(x) - f(a)| < \eta \quad \text{implies} \quad |g(f(x)) - g(f(a))| < \varepsilon.$$

But since f is continuous at a there exists $\delta > 0$ (which may depend on η and hence on ε) such that

$$|x - a| < \delta \quad \text{implies} \quad |f(x) - f(a)| < \eta.$$

Putting these last two implications together, we have

$$|x - a| < \delta \quad \text{implies} \quad |g(f(x)) - g(f(a))| < \varepsilon.$$

Since ε was any positive number, it follows that $g \circ f$ is continuous at a .

If a is a left endpoint of an interval from the domain of f , we replace $|x - a| < \delta$ in the proof everywhere by $a \leq x < a + \delta$. Similarly for a right endpoint. \square

REMARK 5.13. We have not yet defined the trigonometric, exponential or logarithmic functions. If we assume that such functions *are* continuous on their domains (as is indeed the case) then it is easy to see from Theorem 5.11 and Theorem 5.12 that various functions defined from them by composition, cases, and the usual algebraic operations, are also continuous on their domains.

¹ η is the seventh letter of the Greek alphabet, spelt and pronounced "eta"

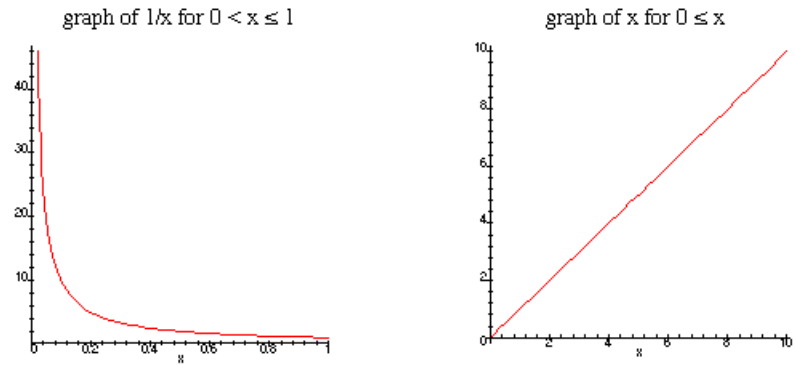
5.4. Deeper properties of continuous functions

Continuous functions defined on closed bounded intervals are bounded and take maximum and minimum values. A continuous function defined on an interval take all values between any two given values of the function.

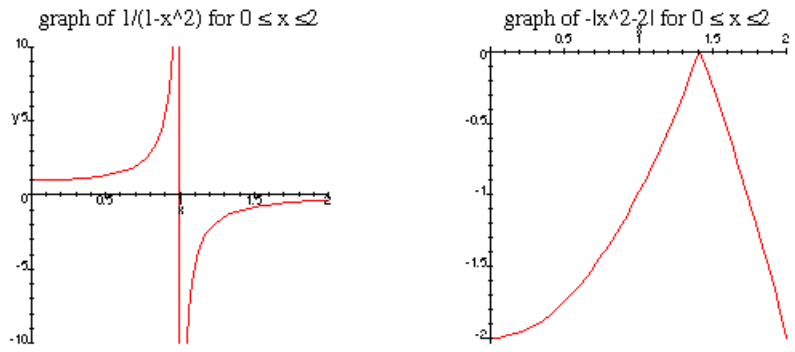
We now can prove some of the most important properties of continuous functions. These results are “global” rather than “local”, in that they say something about the behaviour of continuous functions on a fixed interval, rather than just in *some* neighbourhood of a point.

Although they are not surprising, they are more subtle than may first appear. In particular, they are true on closed bounded intervals and not on arbitrary intervals, as we show by some simple examples. Moreover, they require the completeness axiom for their proofs, since the analogous results are not true over the rationals, as we also explain.

The Max-Min Theorem says that a continuous function defined on a closed bounded interval $[a, b]$ has a maximum (and a minimum) value at some point in $[a, b]$. This is not true on an interval that is open or is infinite at some end, as we see from the examples $f(x) = 1/x$ for $0 < x \leq 1$ and $g(x) = x$ for $0 \leq x < \infty$. Moreover, even if the function is bounded, it need not take a maximum value in the domain, as we see from the example $f(x) = x$ for $0 \leq x < 1$.



The analogous result to the Max-Min Theorem is not true for the rationals. For example, let $f(x) = 1/(1 - x^2)$ for $0 \leq x \leq 2$. This is continuous at every point other than $x = \sqrt{2}$ (and in particular is continuous at every rational point) and it takes rational values at rational points. However, it clearly does not have a maximum value, even when x is restricted to be rational. Moreover, even if a function is bounded and takes rational values at rational points, it need not take a maximum value at some rational point, as we see from the function $f(x) = -|x^2 - 2|$. Since all the axioms other than the completeness axiom are true for the rationals, this indicates that we will need the completeness axiom in the proof of the Max-Min Theorem.



The Intermediate Value Theorem says that if f is a continuous function defined on $[a, b]$ and s is a number between $f(a)$ and $f(b)$, then $f(c) = s$ for some $c \in [a, b]$. The

analogous result is not true for the rational numbers, as we see by considering the function $f(x) = x^2 - 2$. This function is continuous and takes rational values at rational points, $f(1) = -1$ and $f(2) = 2$, but there is no rational number c such that $f(c) = 0$. So we will need the completeness axiom in the proof of the Intermediate Value Theorem.

See [Adams, pages A-25,26] for other proofs of the Min-Max and Intermediate Value Theorem. The proofs there are a little more conceptually difficult as one needs to work with monotone sequences. However, now that we have proved Theorem 4.10 and Theorem 5.3 we can give slightly simpler proofs. We also incorporate the statement and proof of the Boundedness Theorem from [Adams, page A-25] directly into the Min-Max theorem.

See [Spivak] for proofs which use the completeness axiom directly, without invoking sequences at any stage.

THEOREM 5.14 (Max-Min Theorem). *If f is a continuous function defined on an interval $[a, b]$ then there exist numbers $c, d \in [a, b]$ such that $f(c) \leq f(x) \leq f(d)$ for all $x \in [a, b]$. In other words, f takes minimum and maximum values on $[a, b]$. In particular, f is bounded above and below on $[a, b]$.*

PROOF. We first show that f takes a maximum value. To begin, there are two possible cases:

- If the set of numbers $f(x)$ for $x \in [a, b]$ is *not* bounded above, then we choose points $a_n \in [a, b]$ such that $f(a_n) > n$, for each natural number n .
- If the set of numbers $f(x)$ for $x \in [a, b]$ is bounded above, let K be the least upper bound and choose points $a_n \in [a, b]$ such that $f(a_n) > K - 1/n$, for each natural number n . Notice that $f(a_n) \rightarrow K$.

[Our aim is to prove that the second of the above two alternatives holds, *and* that $f(c) = K$ for some $c \in [a, b]$.

We need to be a bit careful in the proof, because there may be more than one maximum point, and the sequence (a_n) may “jump around” by, for example, having an infinite number of terms near one maximum point and also an infinite number of terms near a second maximum point. The sequence (a_n) need not converge, but we will construct a “subsequence” of (a_n) that does converge. Moreover, if this subsequence has limit c , say, it will follow that $f(c) = K$.]

We will now construct a “subsequence” of (a_n) by continually bisecting the interval $I = [a, b]$ as follows.

There must be an *infinite* number of terms from the sequence (a_n) in *at least one* of the two intervals $[a, (a+b)/2]$ and $[(a+b)/2, b]$.² Let I_1 be one such interval and let b_1 be the first (say) member of the sequence (a_n) which is in I_1 .

Now bisect I_1 into two closed sub-intervals of equal length. Again, there must be an infinite number of terms from the sequence (a_n) in *at least one* of these two sub-intervals. Let I_2 be one such interval and let b_2 be the first (say) member of the sequence (a_n) after b_1 which is in I_2 .

Now bisect I_2 into two closed sub-intervals of equal length. Again, there must be an infinite number of terms from the sequence (a_n) in *at least one* of these two sub-intervals. Let I_3 be one such interval and let b_3 be the first (say) member of the sequence (a_n) after b_2 which is in I_3 .

Etc., etc. This process will never stop, since at the k th stage we still have an interval I_k containing an *infinite* number of terms from the original sequence. Thus we can choose b_k which occurs in the original sequence (a_n) after b_{k-1} .

Notice that

$$I \supset I_1 \supset I_2 \supset I_3 \supset \dots \supset I_k \supset \dots,$$

and that the length of I_k is $(b-a)2^{-k}$.

Since both b_k and b_{k+1} belong to I_k (in fact b_{k+1} belongs to $I_{k+1} \subset I_k$), it follows that $|b_k - b_{k+1}| < (b-a)2^{-k}$. It follows from Theorem 4.10 that $b_k \rightarrow c$, say, as $k \rightarrow \infty$.

²Many of the terms in the sequence may have the same value, as for example in the sequence $(1, 2, 3, 1, 1, 1, 1, \dots)$.

Moreover $c \in [a, b]$ from Theorem 4.6. By Theorem 4.12 $f(b_k) \rightarrow f(c)$. This implies that the second of the two alternatives (from the beginning of the proof) must occur, and that also $K = f(c)$ since $f(b_k) \rightarrow K$.

(If the first alternative occurs then the sequence $(f(a_n))$ is monotone increasing with arbitrarily large values, and hence so is the “subsequence” $(f(b_k))$. Thus the second alternative occurs because we have seen that the sequence $(f(b_k))$ converges.

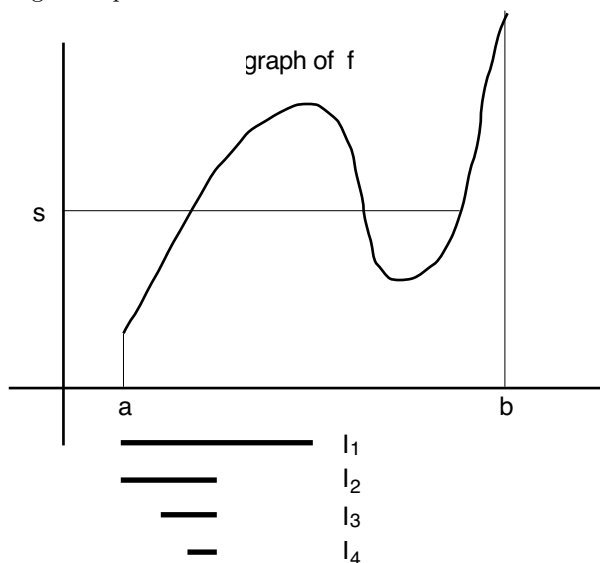
Since $f(a_n) \rightarrow K$, then we also have that the “subsequence” $f(b_k) \rightarrow K$. In general, if a sequence converges to a limit, then any “subsequence” converges to the same limit. This is clear enough, and the proof is just a matter of giving a precise definition of “subsequence” and then using the definition of limit for a sequence.)

The proof that f takes a minimum value is similar. □

THEOREM 5.15 (Intermediate Value Theorem). *If f is a continuous function defined on an interval $[a, b]$ and s is a real number between $f(a)$ and $f(b)$, then there exists some $c \in [a, b]$ such that $f(c) = s$.*

PROOF. We do the case $f(a) < s < f(b)$. If $f(a) > s > f(b)$ the proof is similar, and if $f(a) = s$ or $f(b) = s$ then there is nothing left to prove.

Bisect the interval $I = [a, b]$ into the two intervals $[a, (a+b)/2]$ and $[(a+b)/2, b]$. If $f((a+b)/2) = s$, we are done. If $f((a+b)/2) < s$ let $I_1 = [(a+b)/2, b]$. If $f((a+b)/2) > s$ let $I_1 = [a, (a+b)/2]$. In either case, f takes a value $< s$ at the left end-point of I_1 and a value $> s$ at the right endpoint.



Next consider I_1 . Either f takes the value s at its midpoint, or one of the two (closed) subintervals obtained by bisecting I_1 has the property that f takes a value $< s$ at the left end-point and a value $> s$ at the right endpoint. Call this interval I_2 .

Etc., etc. Either the process stops after a finite number of steps, giving a point where f takes the value s , and we are done.

Or otherwise there is a sequence of closed bounded intervals

$$I \supset I_1 \supset I_2 \supset I_3 \supset \dots \supset I_k \supset \dots,$$

with length of I_k equal to $(b-a)2^{-k}$.

Let $I_n = [a_n, b_n]$ for each n . The sequence (a_n) of left endpoints is increasing, the sequence (b_n) of right endpoints is decreasing, and they both have the same limit c , say.

(This is fairly clear, but needs a bit of thought to write it out carefully.

Certainly the limits exist, as the sequences are increasing or decreasing, and bounded. Since $a_n < b_n$ for all n it follows $\lim a_n \leq \lim b_n$ (if $c_1 := \lim a_n > c_2 := \lim b_n$, then

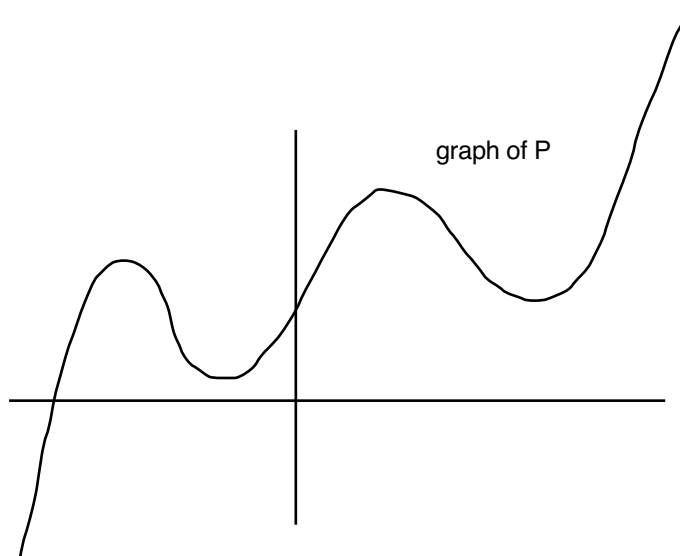
from Theorem 4.6 all a_n will eventually be $> (c_1 + c_2)/2$ and all b_n will eventually be $< (c_1 + c_2)/2$, contradicting $a_n < b_n$.

Thus for any n , $a_n \leq c_1 \leq c_2 \leq b_n$. Since $b_n - a_n \rightarrow 0$, it follows that $c_1 = c_2$.)

Since f is continuous, $a_n \rightarrow c$ implies $f(a_n) \rightarrow f(c)$ from Theorem 4.12. Since $f(a_n) < s$ for all n , $\lim f(a_n) \leq s$ from Corollary 4.7, i.e. $f(c) \leq s$. Similarly, since $f(b_n) > s$ for all n , $f(c) \geq s$. It follows that $f(c) = s$. \square

A nice application is to show that any polynomial of odd degree must have a root.

THEOREM 5.16. *Let $P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ where n is odd and $a_n \neq 0$. Then $P(c) = 0$ for some real number c .*



PROOF. We consider the case $a_n > 0$. The case $a_n < 0$ follows by considering the polynomial $-P(x)$.

We *claim* that for some (large) b , $P(b) > 0$ and for some (large and negative) a , $P(a) < 0$. It then follows from the Intermediate Value Theorem that $P(c) = 0$ for some $a < c < b$.

[The main point in proving the *claim* is to note that for large x , either positive or negative, the “dominant” term is a_nx^n .]

Write

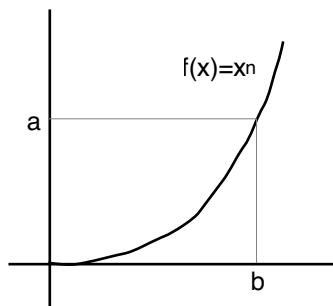
$$P(x) = a_nx^n \left(1 + \frac{a_{n-1}}{a_nx} + \frac{a_{n-2}}{a_nx^2} + \cdots + \frac{a_0}{a_nx^n} \right), \quad \text{for } x \neq 0.$$

By choosing $x = b$ with b sufficiently large and positive, we can make each of the n fractions as small in *absolute value* as we wish, and in particular we can choose b sufficiently large that the term in parentheses is larger than .99, say. Since $a_nb^n > 0$ this means $P(b) > 0$.

Similarly, by choosing a large and *negative*, we can again make the term in parentheses larger than .99, say. But since n is odd, $a_nc^n < 0$ and so $P(c) < 0$. \square

Another application is to show that for any $a > 0$ and any natural number n we can now define $\sqrt[n]{a}$. See [Adams, pp 210, 211].

THEOREM 5.17. *For every $a > 0$ and every natural number n there is a unique number b such that $b^n = a$.*



PROOF. Let f be the function given by $f(x) = x^n$ for $x \geq 0$. We want to show there is a unique number b such that $f(b) = a$.

Now $f(0) = 0$, and we can choose s large enough that $s^n > a$ (e.g. $s = a + 1$ will do). Since f is continuous, by the Intermediate Value Theorem there is a number b such that $f(b) = a$, i.e. such that $b^n = a$.

There is only one such number b because f is an *increasing* function. If $x > b$ then $x^n > b^n = a$ and if $x < b$ then $x^n < b^n = a$; in both cases $x^n \neq a$. \square

Suppose $a > 0$. The b from the theorem is denoted by $\sqrt[n]{a}$ or $a^{1/n}$. For any positive rational number $r = m/n$ we define $a^{m/n} = a^{1/n} \times \dots \times a^{1/n}$ (m times). We also define $a^{-r} = 1/a^r$. Finally, we define $a^0 = 1$.

This defines a^r for $a > 0$ and any rational number r . If x is an arbitrary real number, we could define $a^x = \lim a^{r_n}$ for any sequence of rational numbers $r_n \rightarrow x$.

It is convenient to define $0^r = 0$ if $r > 0$, in particular $0^2 = 0^3 = \dots = 0$. We can also define a^n if $a < 0$ and $n \neq 0$ is an integer (positive or negative). More generally we can define $a^r = -(-a)^r$ if $a < 0$ and $r \neq 0$ is rational with an *odd* denominator, such as $(-5)^{1/3} = -5^{1/3}$, but one does not define $(-5)^{1/2}$ (at least until we introduce the complex numbers.)

The laws for exponents [Adams, page 231] can be proved from these definitions, but they are more difficult to establish in the case of irrational exponents. In particular, in this case, we also would need to show that the definition is independent of the particular sequence of rationals which is chosen.

★ A better way is to first define the logarithmic function by integration and then to define the exponential and power functions. Or one can define these functions as solutions of certain differential equations.

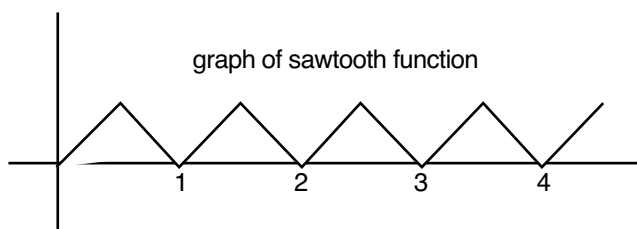
★ We could also show that the function $f(x) = x^r$, where $x \geq 0$ and r is a positive rational number, or $x > 0$ and r is any rational number, is continuous. The main point is to show that if a function is increasing and continuous (such as $g(x) = x^2$ for $x \geq 0$) then its “inverse” function (such as $f(y) = y^{1/2}$ for $y \geq 0$) is also continuous and increasing. This is not so difficult, but we will not do it here.

5.5. ★Pathology and continuity

It is easy to give examples of continuous functions that are not differentiable³ at many points. For example, consider the “saw-tooth function”

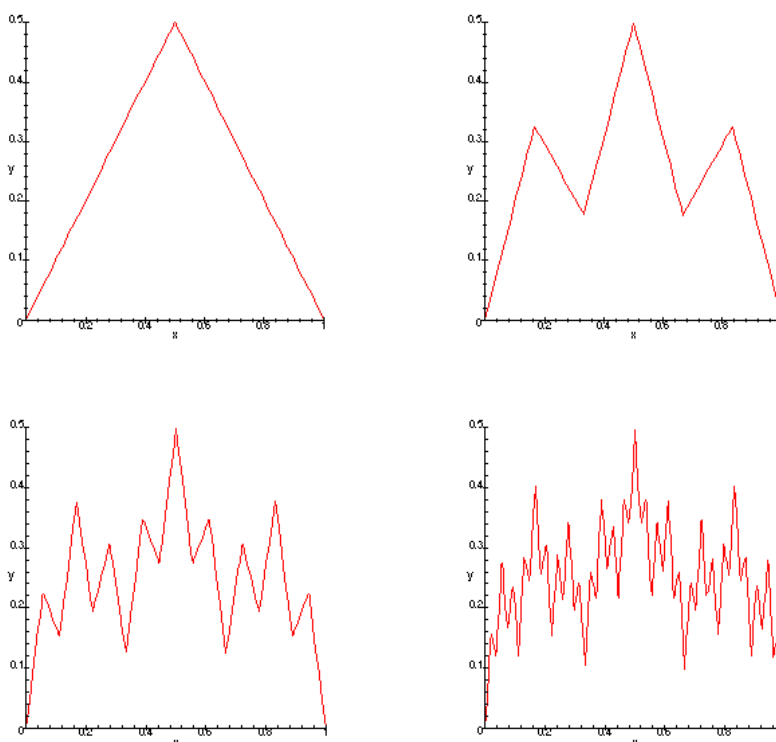
$$f(x) = \begin{cases} x - n & n \in \mathbb{Z} \text{ and } n \leq x < n + \frac{1}{2} \\ (n + 1) - x & n \in \mathbb{Z} \text{ and } n + \frac{1}{2} \leq x < n + 1. \end{cases}$$

³We do not define differentiability formally until Chapter 6. But you already know at least informally what it means.





Then f is not differentiable at n and $n + \frac{1}{2}$ for every integer n .

There are continuous functions which are *nowhere* differentiable. They are, naturally, difficult to draw! But you can get an idea of what one looks like by considering the following sequence of functions which approximate a continuous and nowhere differentiable function.



The idea is that each straight line segment is replaced by a suitable segment of the form

 or  with the same endpoints, when passing to the next approximation.

Such functions actually arise in applications. White noise, Brownian motion, and short term fluctuations on the money markets are best modelled by continuous and nowhere differentiable functions.

5.6. ★Uniform continuity

Any continuous function defined on a closed bounded interval is in fact uniformly continuous.

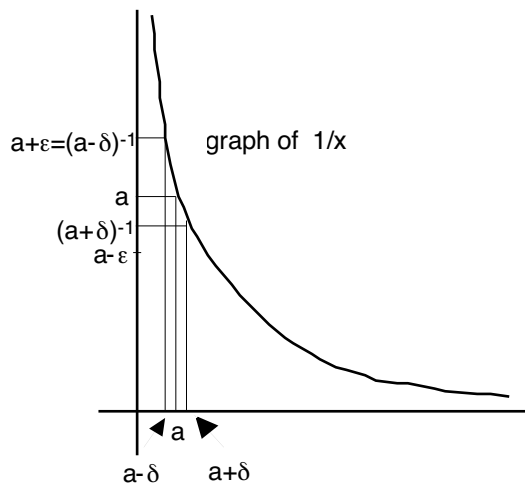
The function $f(x) = x^{-1}$ is continuous on its domain $(-\infty, 0) \cup (0, \infty)$. Of course it is not continuous at zero, since it is not even defined there; and there is no way to define $f(0)$ to make f continuous at 0.

If you play the $\varepsilon - \delta$ game at points a near 0, something interesting happens. If your friend gives you an ε , since f is continuous at a you can always find a δ such that

$$x \in (a - \delta, a + \delta) \quad \text{implies} \quad x^{-1} \in (a^{-1} - \varepsilon, a^{-1} + \varepsilon),$$

i.e.

$$|x - a| < \delta \quad \text{implies} \quad |x^{-1} - a^{-1}| < \varepsilon.$$



It is clear from the diagram, however, that the closer a is to zero, the harder you will have to work to win the game. In particular, the δ you need to win will depend on a as well as ε ; the closer a is to the origin, the smaller δ will need to be for any fixed $\varepsilon > 0$. Given a particular $\varepsilon > 0$ there is no fixed $\delta > 0$, *independent of a* , which will “win” the game.

In fact, algebra shows

$$\begin{aligned} (a - \delta)^{-1} = a^{-1} + \varepsilon & \quad \text{if} \quad \delta = \frac{\varepsilon a^2}{1 + \varepsilon a} \\ (a + \delta)^{-1} = a^{-1} - \varepsilon & \quad \text{if} \quad \delta = \frac{\varepsilon a^2}{1 - \varepsilon a} \end{aligned}$$

It is then clear from the graph, since $a^{-1} - (a + \delta)^{-1} < (a - \delta)^{-1} - a^{-1}$, that in order to have

$$|x - a| < \delta \quad \text{implies} \quad |x^{-1} - a^{-1}| < \varepsilon.$$

you must choose $\delta = \frac{\varepsilon a^2}{1 + \varepsilon a}$ or something smaller. In particular for *fixed* ε you will need to choose $\delta > 0$ closer to 0, the closer a is to 0.

However, when we work with functions f that are continuous on a *closed bounded interval* the situation is much “nicer”. For each $\varepsilon > 0$ there exists a $\delta > 0$, which may depend on ε , but *is independent of a* , such that

$$|x - a| < \delta \quad \text{implies} \quad |f(x) - f(a)| < \varepsilon.$$

In other words, not only can you win the $\varepsilon - \delta$ game at any point a in the domain of f , but you can do it with a δ (which may depend on ε) but will simultaneously work for every a in the domain.

An equivalent but more “symmetric” way of expressing this is to say that for each $\varepsilon > 0$ there exists a $\delta > 0$ (which may depend on ε) such that

$$|x_1 - x_2| < \delta \quad \text{implies} \quad |f(x_1) - f(x_2)| < \varepsilon.$$

In other words, *whenever* two points x_1 and x_2 are distance apart less than δ , *then* the distance between $f(x_1)$ and $f(x_2)$ will be less than ε .

As usual, it is implicitly assumed that both x_1, x_2 are in the domain of f (so that $f(x_1), f(x_2)$ will be defined), and that δ will depend on neither x_1 nor x_2 .

DEFINITION 5.18. A function f is *uniformly continuous* if for every $\varepsilon > 0$ there exists a $\delta > 0$ (which may depend on ε) such that

$$|x_1 - x_2| < \delta \quad \text{implies} \quad |f(x_1) - f(x_2)| < \varepsilon.$$

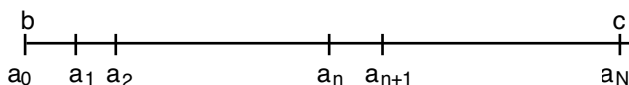
Notice that it does not make sense to say that a function is uniformly continuous at a point (or at least, if it does, then it just means the same as ordinary continuity). Uniform continuity refers to behaviour on the entire domain.

THEOREM 5.19. *Suppose f is a continuous function on a closed bounded interval $[b, c]$. Then f is uniformly continuous on $[b, c]$.*

PROOF. Suppose f is continuous on $[b, c]$. Suppose $\varepsilon > 0$.

We *claim* there is a *finite* strictly increasing sequence of points $b = a_0 < a_1 < a_2 < \dots < a_N = c$ (where N may depend on ε) such that on each of the intervals $[a_n, a_{n+1}]$,

$$(5.1) \quad x \in [a_n, a_{n+1}] \quad \text{implies} \quad |f(x) - f(a_n)| < \varepsilon/3.$$

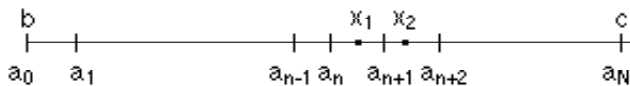


Once we have proved the claim, we obtain the result as follows. Choose $\delta > 0$ such that

$$\delta = \min\{a_1 - a_0, a_2 - a_1, a_3 - a_2, \dots, a_N - a_{N-1}\}$$

Note that since we are taking the minimum of a *finite* set of positive numbers, the minimum δ is also positive.⁴

Let x_1 be any point in $[b, c]$. Then $x_1 \in [a_n, a_{n+1}]$ for some n . Suppose $|x_1 - x_2| < \delta$. It follows that x_2 is in one of the intervals $[a_{n-1}, a_n]$, $[a_n, a_{n+1}]$, $[a_{n+1}, a_{n+2}]$, since the length of each interval is at least δ , x_1 is in the middle interval, and the distance from x_2 to x_1 is less than δ . In other words, x_2 is in the interval $[a_k, a_{k+1}]$ where $k = n - 1, n$ or $n + 1$.



Hence from (5.1)

$$\begin{aligned} |f(x_1) - f(x_2)| &= |f(x_1) - f(a_n) + f(a_n) - f(a_k) + f(a_k) - f(x_2)| \\ &\leq |f(x_1) - f(a_n)| + |f(a_n) - f(a_k)| + |f(a_k) - f(x_2)| \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

In other words we have shown

$$|x_1 - x_2| < \delta \quad \text{implies} \quad |f(x_1) - f(x_2)| < \varepsilon.$$

Since δ did not depend on x_1 or x_2 we have thus established *uniform* continuity, *assuming the claim*.

★ We next *prove the claim*.

Let $a_0 = b$.

Let A_1 be the interval consisting of all those numbers $x \in [a_0, c]$ with the property that every number y , between b and x inclusive⁵, satisfies $|f(y) - f(a_0)| \leq \varepsilon/3$. That is

$$A_1 = \{x \in [a_0, c] : y \in [a_0, x] \text{ implies } |f(y) - f(a_0)| \leq \varepsilon/3\}.$$

It is clear that A_1 is an interval and has left endpoint a_0 (the main point is that if $x \in A_1$ then any number in $[a_0, x]$ is also in A_1). Moreover A_1 is bounded above (by c) and so is of the form $[a_0, a_1)$ or $[a_0, a_1]$ for some number a_1 by the discussion in Remark 2.10. But we must have $a_1 \in A_1$, since if $|f(a_1) - f(a_0)| > \varepsilon/3$ then by continuity of f all numbers

⁴On the other hand, although there is no actual “minimum” of the *infinite* set $1, 1/2, 1/3, \dots$ of positive numbers, the g.l.b. is zero, not positive.

⁵That is, every number $y \in [a, x]$.

x in some neighbourhood of a_1 must also satisfy $|f(x) - f(a_0)| > \varepsilon/3$, and this would contradict the fact that a_1 is the *lub* of A_1 .

Similarly, let A_2 be the interval consisting of all those numbers $x \in [a_1, c]$ with the property that every number $y \in [a_1, x]$ satisfies $|f(y) - f(a_1)| \leq \varepsilon/3$. That is

$$A_2 = \{x \in [a_1, c] : y \in [a_1, x] \text{ implies } |f(y) - f(a_1)| \leq \varepsilon/3\}.$$

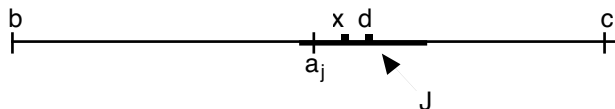
As before, $A_2 = [a_1, a_2]$ for some $a_2 > a_1$.

In this way we construct an increasing sequence $b = a_0 < a_1 < a_2 < \dots$ which either stops after a finite number of steps, or is infinite.

But the sequence cannot be infinite. For if it were infinite, then because it is increasing and bounded above (by c) it must converge to some number d , say. We have

$$a_0 < a_1 < a_2 < \dots < a_n < \dots \rightarrow d.$$

Since f is continuous at d , there is some neighbourhood J of d such that whenever $x \in J$ and $x \in [b, c]$ then $|f(x) - f(d)| < \varepsilon/6$. Now choose a_j so $a_j \in J$ (this is possible since $a_n \rightarrow d$). If $x \in [a_j, d]$ then also $x \in J$ (since J is an interval) and so



$$\begin{aligned} |f(x) - f(a_j)| &= |f(x) - f(d) + f(d) - f(a_j)| \\ &\leq |f(x) - f(d)| + |f(d) - f(a_j)| \\ &< \frac{\varepsilon}{6} + \frac{\varepsilon}{6} \quad \text{since } x, d \in J \\ &= \frac{\varepsilon}{3}. \end{aligned}$$

But from the definition of A_{j+1} this means $d \in A_{j+1}$, contradicting the fact that $a_{j+1} < d$ and a_{j+1} is the l. u. b. of A_{j+1} .

Hence the sequence is finite, and so we can write it as

$$b = a_0 < a_1 < a_2 < \dots < a_N \leq c$$

for some natural number N . If $a_N < c$ then by continuity of f at a_N we could continue the sequence, and so we must have $a_N = c$. This finally proves the *claim*, and hence the theorem. \square

5.7. ★Functions of two or more variables

Suppose (a, b) is an interior point of the domain of f . Then we say that f is *continuous at* (a, b) if

$$\lim_{(x,y) \rightarrow (a,b)} f(x, y) = f(a, b).$$

We could also define continuity at boundary points in a similar manner, but we will not need this.

The analogues of Remark 5.9, and Theorems 5.10, 5.11 and 5.12 hold, with similar proofs. In particular, if h and g are functions of one variable which are continuous at a , and f is a function of two variables which is continuous at $(h(a), g(a))$ and is defined in a neighbourhood of $(h(a), g(a))$, then $f(h(x), g(x))$ is a function of one variables which is continuous at a .

Analogues of the Max-Min Theorem and the Uniform Continuity Theorem hold for continuous functions on any "closed bounded rectangle" A consisting of all points (x, y) such that $a \leq x \leq b$ and $c \leq y \leq d$. The proofs are similar. In particular, any continuous function defined on a closed bounded rectangle is bounded above.

Differentiation

The main references in [Adams] are Sections 2.2, 2.3, 2.5, 5.1 and 5.2.

6.1. Introduction

The theory of differentiation allows us to analyse the concept of the slope of the tangent to the graph of a function.

If we write the function in the form $y = f(x)$ then we can interpret $f'(x)$ in the following way:

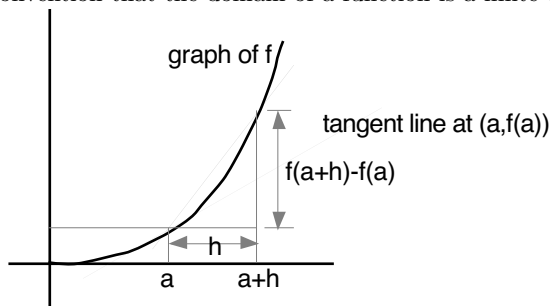
y is changing $f'(x)$ times as fast as x is changing.

There are many other problems that can then be analysed using the techniques of differentiation and extensions of these ideas — for example anything that changes with time or position. Also optimisation problems (e.g. in economics or engineering) and approximation problems. See [Adams, Chapter 3] for a number of examples.

6.2. The derivative of a function

Derivatives are defined and the fact differentiability implies continuity is proved.

Recall our convention that the domain of a function is a finite union of intervals.



The idea from the above diagram is that the derivative $f'(a)$ of f at a should be the slope of the tangent to the graph of f at the point $(a, f(a))$ on the graph.

We make this precise by considering the *slope* of the line through the two points $(a, f(a))$ and $(a + h, f(a + h))$ and considering the limit (if it exists) as $h \rightarrow 0$. (h is allowed to be either positive or negative, except at endpoints a of the domain of f .)

DEFINITION 6.1. If a is an interior point of the domain of f and

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

exists, or if a is an endpoint and the corresponding one-sided limit

$$\lim_{h \rightarrow 0^+} \frac{f(a+h) - f(a)}{h} \quad \text{or} \quad \lim_{h \rightarrow 0^-} \frac{f(a+h) - f(a)}{h}$$

exists, then we say f is *differentiable* at a . The limit is denoted by $f'(a)$ (sometimes $f'_+(a)$ or $f'_-(a)$ in the case of endpoints) and is called the *derivative of f at a* .

The *derivative of f* is the *function f'* whose value at a is the number $f'(a)$ defined above, with domain consisting of all a such that the derivative $f'(a)$ exists. We say f is *differentiable* if it is differentiable at every point in its domain.

The tangent to the graph of f at a has slope $f'(a)$. It follows that the equation of this line is

$$y = f(a) + f'(a)(x - a).$$

See [Adams, Examples 1,2 p.99] to see how the derivatives of the functions x^2 , $1/x$ and \sqrt{x} can be calculated directly from the above definition. But we do not usually need to do this. Instead we normally can use Theorem 6.5 and Theorem 6.7.

NOTATION 6.2. If $y = f(x)$ then we use the dependent variable y to represent the function, and the derivative is denoted in the following various ways:

$$y', \quad \frac{dy}{dx}, \quad \frac{d}{dx}f(x), \quad f'(x),$$

which we read as “ y prime”, “the derivative of y with respect to x ” or “ $dy dx$ ” for short, “the derivative with respect to x of $f(x)$ ” or “ $d dx$ of $f(x)$ ” for short, and “ f prime of x ”, respectively.

In particular, we often write

$$\begin{aligned} \frac{d}{dx}x^3 &= 3x^2, \\ \frac{d}{dt}t^4 &= 4t^3, \end{aligned}$$

etc., and regard $\frac{d}{dx}$ as a “differential operator” which maps one function to another function; such as the function f given by $f(x) = x^3$ to the function g given by $g(x) = 3x^2$.

(★ Thus a differential operator is a function which sends functions to functions, rather than numbers to numbers!)

The value of the derivative of a function at a particular number a can also be written in various ways:

$$y'(a), \quad y' \Big|_a, \quad \frac{dy}{dx} \Big|_a, \quad \frac{d}{dx}f(x) \Big|_a, \quad f'(a).$$

The symbol $\Big|_a$ is the evaluation symbol, and signifies that the function preceding it should be evaluated at a . If there is any doubt as to what is the dependent variable, one replaces $\Big|_a$ by $\Big|_{x=a}$.

The $\frac{dy}{dx}$ type notation is called *Leibniz notation* after its inventor. It is very good for computations and for motivating some results. If one thinks of

$$\Delta y = f(x + h) - f(x)$$

as being the *increment in y* and

$$\Delta x = (x + h) - x = h$$

as being the *increment in x* , then

$$\frac{dy}{dx} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x}.$$

However, *the Leibniz notation should not be used when proving theorems rigorously*. It is often ambiguous in more complicated situations, and this can easily lead to logical errors. See the discussion before Theorem 6.7 for a good example of what can go wrong.

See [Adams, pp 102,103] for more discussion of notation.

The following theorem is important.

THEOREM 6.3. *If f is differentiable at a then f is continuous at a .*

PROOF. Assume f is differentiable at a . We want to show¹ that

$$\lim_{h \rightarrow 0} f(a + h) = f(a).$$

¹ f is continuous at a means $\lim_{x \rightarrow a} f(x) = f(a)$. This is the same as $\lim_{h \rightarrow 0} f(a + h) = f(a)$. The fact this *is* the same is not surprising, but to show it carefully is a matter of writing out the corresponding $\varepsilon - \delta$ definition in each case and checking that each limit means the same thing.

But

$$f(a+h) = f(a) + h \frac{f(a+h) - f(a)}{h}.$$

Taking the limit as $h \rightarrow 0$ of the right side, we see this limit exists and hence so does the limit of the left side, and both are equal. That is

$$\lim_{h \rightarrow 0} f(a+h) = f(a) + 0f'(a) = f(a).$$

A similar proof applies if a is an endpoint of the domain of f . \square

6.3. Computing derivatives

The standard rules for differentiation, including the chain rule, are discussed. Examples are given.

The next result is easy to check from the definition, and is obvious from the relevant diagram. It implies that the slope of the straight line, which is the graph of the function $f(x) = cx + d$, is c .

THEOREM 6.4. *If $f(x) = cx + d$ then $f'(x) = c$.*

PROOF.

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{(c(x+h) + d) - (cx + d)}{h} = \lim_{h \rightarrow 0} \frac{ch}{h} = c.$$

\square

The next theorem follows in a fairly straightforward way from the properties of limits given in Theorem 3.21.

THEOREM 6.5. *If f and g are differentiable at x and c is a real number, then the following functions are differentiable at x with derivatives as shown.*

$$\begin{aligned} (f \pm g)'(x) &= f'(x) \pm g'(x) \\ (cf)'(x) &= cf'(x) \\ (fg)'(x) &= f'(x)g(x) + f(x)g'(x) \\ \left(\frac{1}{g}\right)' &= \frac{-g'(x)}{(g(x))^2} \\ \left(\frac{f}{g}\right)' &= \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2} \end{aligned}$$

In the last two cases we also assume $g(x) \neq 0$.

PROOF. The proof of the first two is given in [Adams, page 108]. The proof of the product rule is in [Adams, page 109]. The proof of the reciprocal and quotients rules is in [Adams, pp 111,112]. \square

The following now follows from the product rule and the Principle of Induction.

THEOREM 6.6. *If $f(x) = x^n$ then $f'(x) = nx^{n-1}$.*

PROOF. The result is true for $n = 1$.

Assume it is true for some integer n , i.e. $(x^n)' = nx^{n-1}$.

Then

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

Thus the corresponding result is true for $n + 1$.

The result is hence true for *all* natural numbers n by the Principle of Induction. \square

It now follows that

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \quad \text{implies}$$

$$f'(x) = a_1 + 2a_2x + \cdots + na_nx^{n-1}.$$

We can also now compute derivatives of rational functions.

One can also show directly (as we noted before for \sqrt{x}), that the derivative of $x^{1/n}$ for $x > 0$ and n a natural number, is $\frac{1}{n}x^{\frac{1}{n}-1}$, see [Adams, Question 56 p. 106]. By using induction on m , one can then show that the derivative of $x^{m/n}$ for $x > 0$ and m, n natural numbers, is $\frac{m}{n}x^{\frac{m}{n}-1}$.

One can also show directly by induction, using the derivative of $1/x$, that for n a natural number the derivative of $x^{-n} = (x^{-1})^n$ is $-nx^{-n-1}$, see [Adams, Question 56 p. 106].

In a similar way, one can prove the general rule $(x^r)' = rx^{r-1}$ for any rational number r wherever the function x^r is defined. The same result is true for any real number r , as we would expect by taking a sequence of rational numbers $r_n \rightarrow r$. But this is best proved by first developing the theory of logarithms, exponential functions and then general power functions. See [Adams, p. 220, before Example 6].

Natural (i.e. to base e) logarithms are defined as integrals in [Adams, page 215] and then the derivative of $\log_e x$ is proved to be $1/x$. From this one can then derive the other usual rules for derivatives of exponentials and other power functions such as a^x and x^a , see [Adams, pp 214–222].

The proofs of the usual rules for the derivatives of the trigonometric functions are given in [Adams, pp 115–120]. They are not completely rigorous, since the definition of \sin and the other trigonometric functions was only given informally, using diagrams, in [Adams, pp 40–51].

At this stage, we will only use derivatives of such functions in the examples, but not in the rigorous development of the subject.

In order to compute the derivatives of functions such as $\sqrt{1+x^2}$ we need the *Chain Rule*. You have probably seen the chain rule in the form

$$\frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx}.$$

Here we have $y = f(u)$ and $u = g(x)$, so $y = f(g(x))$. The rough idea is that

$$\text{at } u, y \text{ is changing } \frac{dy}{du} \text{ times as fast as } u$$

and

$$\text{at } x, u \text{ is changing } \frac{du}{dx} \text{ times as fast as } x,$$

so that

$$\text{at } x, y = f(u) = f(g(x)) \text{ is changing } \frac{dy}{du} \times \frac{du}{dx} \text{ times as fast as } x.$$

In functional notation

$$(f \circ g)'(x) = f'(g(x)) g'(x).$$

An incorrect “proof” along these lines is often given for the chain rule by writing

$$\begin{aligned} \frac{dy}{dx} &= \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta u} \frac{\Delta u}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta u} \lim_{\Delta x \rightarrow 0} \frac{\Delta u}{\Delta x} \\ &= \lim_{\Delta u \rightarrow 0} \frac{\Delta y}{\Delta u} \lim_{\Delta x \rightarrow 0} \frac{\Delta u}{\Delta x} = \frac{dy}{du} \frac{du}{dx} \end{aligned}$$

The second last step is “justified” by saying that $\Delta u \rightarrow 0$ as $\Delta x \rightarrow 0$. This is all rather sloppy, because it is not clear what depends on what.

When one tries to fix it up, there arises a very serious difficulty. Namely, the increment $\Delta u = u(x + \Delta x) - u(x)$ (which depends on Δx) may be zero although $\Delta x \neq 0$. A trivial example is if u is the constant function. There is the same difficulty when u is not constant, but there are points $x + \Delta x$ arbitrarily close to x such that $u(x + \Delta x) = u(x)$ (such as with $u(x) = x^2 \sin(1/x)$ for $x \neq 0$ — see Example 6.8).

This difficulty is not clear with the Leibniz notation, but becomes clearer when we write out the argument in a more precise functional notation. See [Adams, Question 68 p. 127].

We now state the Chain rule precisely, and refer to Adams for a (correct) proof.

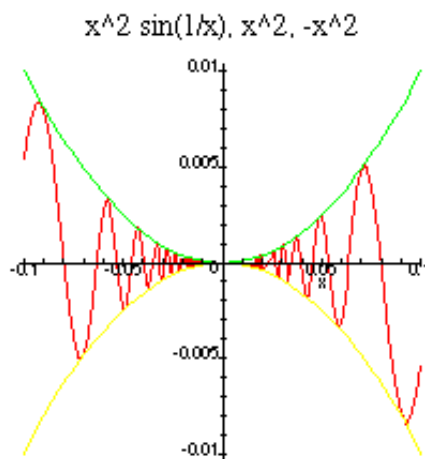
THEOREM 6.7 (Chain Rule). *Assume the function f is differentiable at $g(x)$ and the function g is differentiable at x . Then the composite function $f \circ g$ is differentiable at x and*

$$(f \circ g)'(x) = f'(g(x)) g'(x).$$

(We also assume that g is defined in a neighbourhood of x and f is defined in a neighbourhood of $g(x)$, although this can be generalised a bit.)

PROOF. See [Adams, p. 126]. To help understand the proof, note that the “error” function $E(k)$ is the *difference* between the *slope* of the line through the points $(u, f(u))$ and $(u+k, f(u+k))$ on the graph of f , and the *slope* of the tangent at the point $(u, f(u))$ and $(u+k, f(u+k))$ on the graph of f . Draw a diagram like the first one in this chapter. \square

EXAMPLE 6.8.



We can now compute the derivative of the function

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

If $x \neq 0$ then by the product and chain rules (and using the fact $\sin' y = \cos y$)

$$\begin{aligned} f'(x) &= (x^2)' \sin \frac{1}{x} + x^2 \left(\sin' \frac{1}{x} \right) \left(\frac{1}{x} \right)' \\ &= 2x \sin \frac{1}{x} + x^2 \left(\cos \frac{1}{x} \right) \left(-\frac{1}{x^2} \right) \\ &= 2x \sin \frac{1}{x} - \cos \frac{1}{x} \end{aligned}$$

We see that $f'(x)$ has no limit as $x \rightarrow 0$, since the first term approaches zero but the second “oscillates” between ± 1 .

However, f is differentiable at 0, and in fact $f'(0) = 0$. This is in fact not surprising if we look at the graph. Any line passing through the points $(0, 0)$ and $(h, h^2 \sin \frac{1}{h})$ on the graph lies in the region between the two parabolas corresponding to $\pm x^2$. It is thus geometrically clear that the slope of this line approaches 0 as $h \rightarrow 0$.

Analytically,

$$\begin{aligned} f'(0) &= \lim_{h \rightarrow 0} \frac{f(h) - f(0)}{h} \\ &= \lim_{h \rightarrow 0} \frac{h^2 \sin \frac{1}{h} - 0}{h} \\ &= \lim_{h \rightarrow 0} h \sin \frac{1}{h} = 0 \end{aligned}$$

The last limit was shown in Example 3.12 (it followed easily from the Squeeze Theorem applied with $\pm x$).

Thus f is differentiable for all x , but the derivative is not continuous at 0.

6.4. Maximum and minimum values

The relationship between derivatives and maximum and minimum points is given.

DEFINITION 6.9. A function f has a *maximum value* (*minimum value*) $f(x_0)$ at the *maximum point* (*minimum point*) $x_0 \in \mathcal{D}(f)$ if

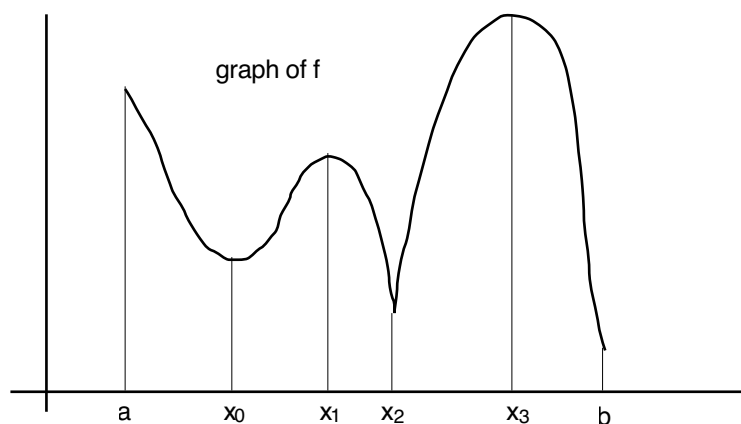
$$f(x) \leq f(x_0) \quad (f(x) \geq f(x_0))$$

for all $x \in \mathcal{D}(f)$.

The function has a *local maximum value* (*local minimum value*) $f(x_0)$ at the *local maximum point* (*local minimum point*) $x_0 \in \mathcal{D}(f)$ if there exists a neighbourhood N of x_0 such that

$$f(x) \leq f(x_0) \quad (f(x) \geq f(x_0))$$

for all x in $N \cap \mathcal{D}(f)$.



In the above diagram, f has a maximum at x_3 , a minimum at b , local maxima at a, x_1, x_3 and local minima at x_0, x_2, b .

We saw in Theorem 5.14 that a *continuous* function f defined on a closed bounded interval has a maximum and a minimum value.

If f has a local maximum or minimum at x then there are three possibilities

- x is an endpoint of the domain of f ;
- x is an interior point and $f'(x)$ does not exist
- x is an interior point and $f'(x)$ does exist

THEOREM 6.10. *Suppose f has a local maximum or minimum at an interior point x_0 and that $f'(x_0)$ exists. Then $f'(x_0) = 0$.*

PROOF. Suppose f has a local maximum at the interior point x_0 (the proof for a local minimum is similar). Then for some $h_0 > 0$,

$$|h| < h_0 \quad \text{implies} \quad f(x_0) \geq f(x_0 + h).$$

Hence,

$$(6.1) \quad \frac{f(x_0 + h) - f(x_0)}{h} \leq 0 \text{ if } 0 < h < h_0.$$

and

$$(6.2) \quad \frac{f(x_0 + h) - f(x_0)}{h} \geq 0 \text{ if } -h_0 < h < 0.$$

We know that the derivative at x_0 exists and hence

$$\lim_{h \rightarrow 0^+} \frac{f(x_0 + h) - f(x_0)}{h} \text{ and } \lim_{h \rightarrow 0^-} \frac{f(x_0 + h) - f(x_0)}{h}$$

both exist and are equal. But the first limit is ≤ 0 from (6.2) and the second is ≥ 0 from (6.1). Hence the derivative must be 0. \square

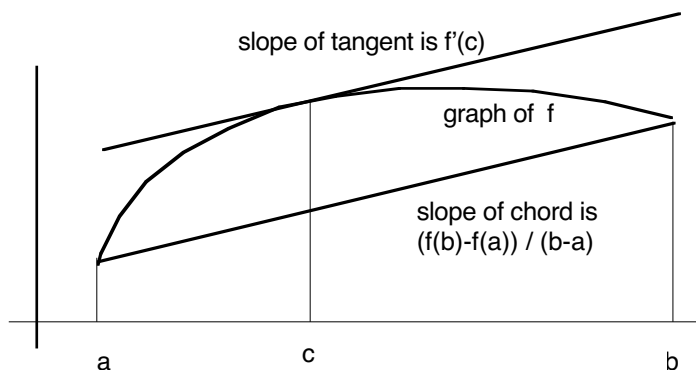
It is not true that if $f'(x_0) = 0$ then f must have a local maximum or minimum at x_0 . Consider $f(x) = x^3$ at 0.

We say a function f has a *critical value* $f(x_0)$ at the *critical point* $x_0 \in \mathcal{D}(f)$ if $f'(x_0) = 0$

6.5. Mean Value Theorem

The Mean Value Theorem is proved. This is used to bound the difference between values of a function, and to prove the Constancy Theorem and Rolle's Theorem. The relationship between the sign of the derivative and the monotone behaviour of a function is developed.

The Mean Value Theorem says that the slope of the line joining two points $(a, f(a))$ and $(b, f(b))$ on the graph of a differentiable function is equal to the slope of the tangent at the point $(c, f(c))$ for some c between a and b . This is geometrically clear for any reasonable function whose graph we can draw. We want to show that it follows rigorously from the definition of differentiable (this then will be another justification that our definition correctly captures our informal notions of differentiability).



From the diagram, we expect that c will correspond to some point on the graph of f at maximum vertical distance from the line joining $(a, f(a))$ and $(b, f(b))$. Since the equation of this line is $y = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$, this vertical distance is given by $f(x) - f(a) - \frac{f(b)-f(a)}{b-a}(x-a)$. This motivates the following proof.

THEOREM 6.11 (Mean Value Theorem). *Suppose f is continuous on the closed bounded interval $[a, b]$ and is differentiable on the open interval (a, b) . Then there exists $c \in (a, b)$ such that*

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

PROOF. Consider the function g given by

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

This is continuous on $[a, b]$ and so (by Theorem 5.14) has a maximum at some point $c \in (a, b)$. By Theorem 6.10 it follows

$$g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} = 0,$$

which gives the result. \square

COROLLARY 6.12. *Suppose f is continuous on an interval and $|f'(x)| \leq K$ at every interior point in the interval. (The interval may be open, closed or unbounded, at either end.) Then*

$$|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$$

for all x_1, x_2 in the interval.

PROOF. Suppose $x_1 < x_2$. (The proof is similar if $x_2 < x_1$ and the result is trivial if $x_1 = x_2$.)

By the Mean Value theorem there exists a number c between x_1 and x_2 such that

$$f(x_1) - f(x_2) = f'(c)(x_1 - x_2),$$

and so

$$|f(x_1) - f(x_2)| = |f'(c)||x_1 - x_2| \leq K|x_1 - x_2|.$$

\square

COROLLARY 6.13 (Constancy Theorem). *If f is continuous on an interval and $f'(x) = 0$ at every interior point in the interval then f is constant on the interval. (The interval I may be open, closed or unbounded, at either end.)*

PROOF. Choose a point $c \in I$ and let $C = f(c)$. We want to show $f(x) = C$ for any $x \in I$.

But $|f(x) - f(c)| = 0$ by the previous corollary, and so $f(x) = C$. \square

The corollary is not true if the domain of f is a finite union of more than one interval. In this case the function is constant on *each* interval, but the constant may depend on the interval.

Rolle's Theorem says that if f is continuous on $[a, b]$ and differentiable on (a, b) , and $f(a) = f(b) = 0$, then $f'(x_0) = 0$ for some $x_0 \in (a, b)$. It is just a particular case of the Mean Value Theorem.

A useful application of Corollary 6.13 is to prove that complicated expressions are equal. For example, to prove that $f(x) = g(x)$ for all x in some interval, it is sufficient to prove that the functions are equal at a single point c and that their derivatives are equal everywhere.

To see this apply the corollary to the function $f(x) - g(x)$. The derivative is zero and so the function is constant; but the constant is zero since $f(c) - g(c) = 0$.

See [Adams, Example 2 p. 256] for an example.

The Mean Value Theorem leads to a result which enables us to decide where a function is increasing or decreasing.

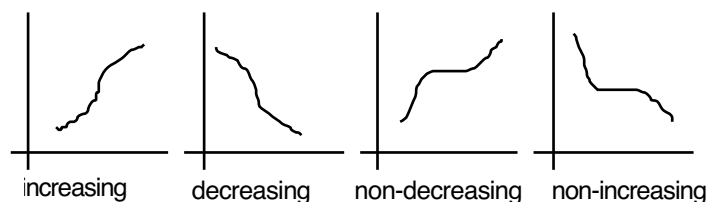
DEFINITION 6.14. We say a function f is

increasing if $x_1 < x_2$ implies $f(x_1) < f(x_2)$

decreasing if $x_1 < x_2$ implies $f(x_1) > f(x_2)$

non-decreasing if $x_1 < x_2$ implies $f(x_1) \leq f(x_2)$

non-increasing if $x_1 < x_2$ implies $f(x_1) \geq f(x_2)$



THEOREM 6.15. *Suppose that f is continuous on an interval and differentiable at every interior point. Then*

$f'(x) > 0$ for every interior point x implies f is increasing on J .

$f'(x) < 0$ for every interior point x implies f is decreasing on J .

$f'(x) \geq 0$ for every interior point x implies f is non-decreasing on J .

$f'(x) \leq 0$ for every interior point x implies f is non-increasing on J .

PROOF. Suppose $x_1 < x_2$ are points in J . By the Mean Value Theorem,

$$f(x_2) - f(x_1) = f'(x_0)(x_2 - x_1)$$

for some x_0 between x_1 and x_2 .

If $f'(x) > 0$ for all $x \in I$ then this implies $f(x_1) < f(x_2)$, and similarly for the other three cases. \square

Note that if a function is increasing on an interval then it does not follow that $f'(x) > 0$ for every interior point x . For example, if $f(x) = x^3$ then f is increasing, but $f'(0) = 0$. However, if f is increasing, or even just non-decreasing, then it does follow that $f'(x) \geq 0$ for all x . This also follows from the Mean value Theorem, (*exercise*).

6.6. ★Partial derivatives

Suppose, for simplicity, we have a function $f(x, y)$ defined on an open rectangle A as in Section 3.4. The *partial derivative with respect to y at (x_0, y_0)* is defined by

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{h \rightarrow 0} \frac{f(x_0, y_0 + h) - f(x_0, y_0)}{h}.$$

Think of the line parallel to the y -axis through the point (x_0, y_0) , and think of f as a function with domain restricted to this line, i.e. f is a function of y with x fixed to be x_0 . Then $\partial f / \partial y(x_0, y_0)$ is just the ordinary derivative with respect to y .

If we know that

$$\left| \frac{\partial f}{\partial y}(x, y) \right| \leq K$$

at every point (x, y) in the open rectangle A , then it follows from Corollary 6.12 that

$$|f(x, y_1) - f(x, y_2)| \leq K|y_1 - y_2|$$

for every $(x, y_1), (x, y_2) \in A$.

We will use this in the proof of Theorem 8.2.

CHAPTER 7

Integration

The main references in [Adams] are Sections 6.1–6.5 and Appendix IV.

Integration allows us to find areas and volumes bounded by curves and surfaces.

It is rather surprising at first, but there is a close relationship between integration and differentiation; each is the inverse of the other. This is known as the Fundamental Theorem of Calculus. It allows us to find areas by doing the reverse of differentiation.

Integrals are also used to express lengths of curves, work, energy and force, probabilities, and various quantities in economics, for example.

7.1. Introduction

The topic of this chapter is the concept of area in a quantitative sense and the elucidation of some of its properties.

Everyone would be happy with the definition: “the area of a rectangle is the product of its length and breadth”. The problem is more difficult with more complicated plane figures. The circle, for example, has “area πr^2 ”; but is this “area” the same concept as that applied to rectangles?

In everyday life one often needs only an approximation to the area of, say, a country or a field. If pressed one would calculate it approximately by filling it as nearly as possible with rectangles and summing their area. This is very close to what we do here in giving a precise definition of the concept of area.

7.2. The Riemann integral

The (definite) Riemann integral is defined in terms of upper and lower sums. It is shown that continuous functions on closed bounded intervals are integrable.

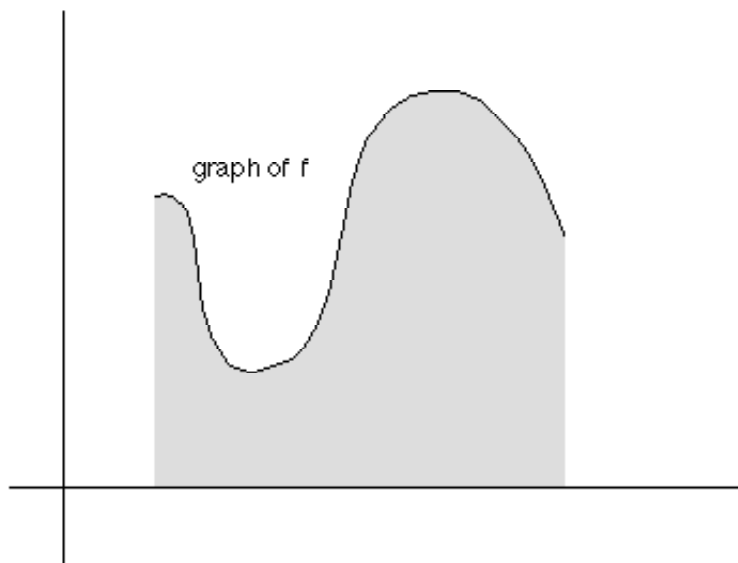
Throughout this section¹, unless stated otherwise, f is a continuous function defined on a closed bounded interval $[a, b]$.

We aim to define the “area under the graph of f ”. That is we wish to attach a number to the shaded region in the following diagram, which is its “area”, and which has the properties that we normally associate with “area”.

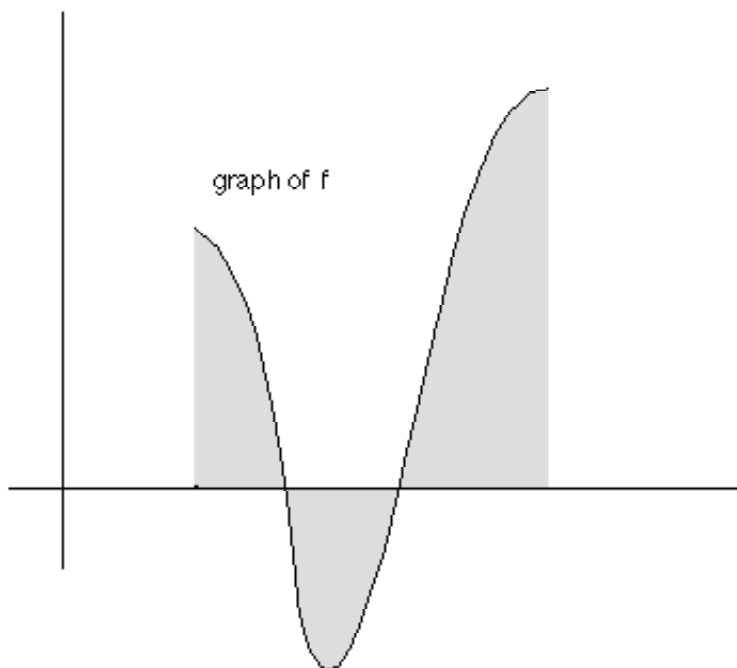
The basic properties that we want of this “area” number are

- the area of a rectangle should be “length times breadth”;
- the area of non-overlapping regions is the sum of their areas;
- if one region is contained in another the area of the first is \leq the area of the second.

¹some of the material in this section closely follows notes of Bob Bryce from a previous first year honours level course.

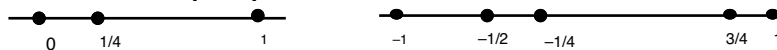


Before we begin, a preliminary comment: a given function f on $[a, b]$ may take values both positive and negative, as in the next diagram.



The concept of area which we are about to define will treat the regions below the x -axis as negative. The concept we define is in this sense not quite what one might expect, though it agrees with our intuition in the case when $f(x) \geq 0$ for all $x \in [a, b]$.

We begin by defining a partition of $[a, b]$; this is simply a finite set of points in $[a, b]$ including a and b . Thus $P = \{0, 1/4, 1\}$ is a partition of $[0, 1]$, and $P = \{-1, -1/2, -1/4, 3/4, 1\}$ is a partition of $[-1, 1]$.



The general notation for a *partition* P of $[a, b]$ with n sub-intervals will be

$$P = \{a = x_0, x_1, x_2, \dots, x_n = b\}.$$

We will assume always that $a = x_0 < x_1 < \dots < x_n = b$.

The length of the i th subinterval is denoted by

$$\Delta x_i := x_i - x_{i-1}.$$

With each partition P of $[a, b]$ we associate the so-called *upper* and *lower sums*. To define these we need the following notation: write

$$M_i = \max \{ f(x) : x_{i-1} \leq x \leq x_i \}, \quad 1 \leq i \leq n$$

$$m_i = \min \{ f(x) : x_{i-1} \leq x \leq x_i \}, \quad 1 \leq i \leq n.$$

That is, M_i is the maximum value and m_i the minimum value of f on the i th sub-interval $[x_{i-1}, x_i]$ of the partition. These exist because f is continuous on the *closed bounded* interval $[x_{i-1}, x_i]$.

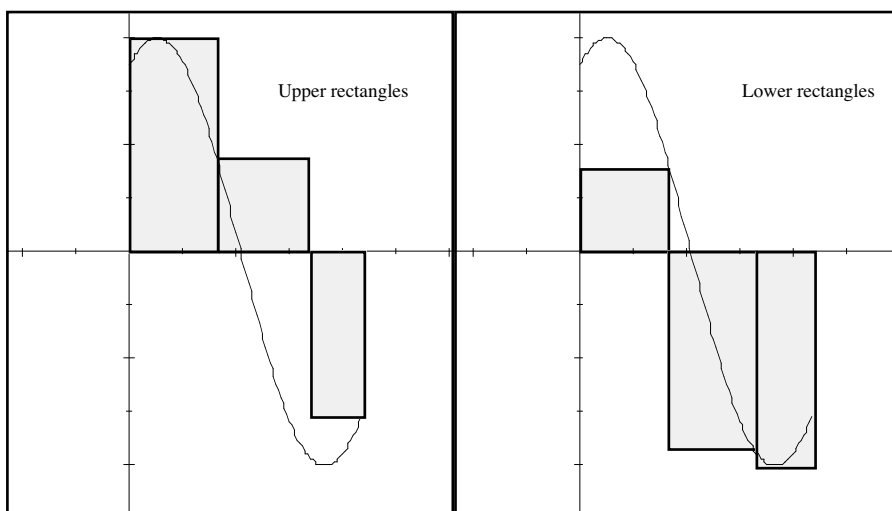
The *upper sum of f over P* is defined by

$$U(P, f) := \sum_{i=1}^n M_i \Delta x_i,$$

and the *lower sum of f over P* is defined by

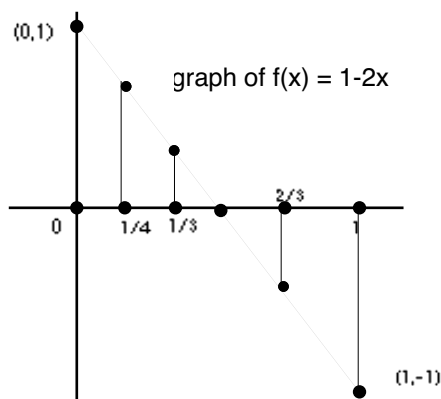
$$L(P, f) := \sum_{i=1}^n m_i \Delta x_i.$$

(See [Adams, pp. 294–7] for a discussion of the summation notation.) Roughly speaking $L(P, f)$ is the sum of the areas of all the rectangles whose bases are the sub-intervals $[x_{i-1}, x_i]$ and which just fit under the graph of f . Similarly $U(P, f)$ is the sum of the areas of all the rectangles whose bases are the sub-intervals $[x_{i-1}, x_i]$ and which just contain the graph of f . At least this is the case when $f(x) \geq 0$. In other cases the interpretation is less simple. Various possibilities are illustrated in the diagrams below.



In [Adams, page A-27] the graph of the function is missing from the diagram. Can you work out what the function looks like?

EXAMPLE 7.1. Let $f(x) = 1 - 2x$ on $[0, 1]$, and let $P = \{0, 1/4, 1/3, 2/3, 1\}$. Find $L(P, f)$ and $U(P, f)$.



Here

$$\begin{array}{lll}
 M_1 = 1 & m_1 = 1/2 & \Delta x_1 = 1/4 \\
 M_2 = 1/2 & m_2 = 1/3 & \Delta x_2 = 1/12 \\
 M_3 = 1/3 & m_3 = -1/3 & \Delta x_3 = 1/3 \\
 M_4 = -1/3 & m_4 = -1 & \Delta x_4 = 1/3
 \end{array}$$

and so

$$\begin{aligned}
 L(P, f) &= \sum_{i=1}^4 m_i \Delta x_i = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{12} + \left(\frac{-1}{3}\right) \frac{1}{3} + \frac{-1}{3} = -\frac{7}{24} \\
 U(P, f) &= \sum_{i=1}^4 M_i \Delta x_i = 1 \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{12} + \frac{1}{3} \cdot \frac{1}{3} + \left(-\frac{1}{3}\right) \frac{1}{3} = \frac{7}{24}
 \end{aligned}$$

EXERCISE 7.2. Let $f(x) = \cos x$ on $[-\pi/2, \pi]$ and $P = \{-\pi/2, -\pi/4, 0, \pi/2, \pi\}$. Show that

$$L(P, f) = \frac{1}{\sqrt{2}} \cdot \frac{\pi}{4} - \frac{\pi}{2}$$

and

$$U(P, f) = \frac{1}{\sqrt{2}} \cdot \frac{\pi}{4} + \frac{3\pi}{4}.$$

We now develop the properties of upper and lower sums that we need.

LEMMA 7.3. Let f be a continuous function on $[a, b]$ and P be a partition of $[a, b]$. Then $L(P, f) \leq U(P, f)$.

PROOF. Since $m_i \leq M_i$ for all i , and since $x_i - x_{i-1} > 0$,

$$m_i(x_i - x_{i-1}) \leq M_i(x_i - x_{i-1}).$$

Adding,

$$L(P, f) \leq U(P, f)$$

as required. \square

Draw a diagram and you will see how obvious this result is.

LEMMA 7.4. Let f be a continuous function on $[a, b]$. Let P_1, P_2 be two partitions of $[a, b]$ with $P_1 \subset P_2$. (We say that P_2 is a refinement of P_1 .) Then

$$L(P_1, f) \leq L(P_2, f) \quad \text{and} \quad U(P_2, f) \leq U(P_1, f).$$

PROOF. We can get P_2 from P_1 by successively adding one new point at a time. If therefore, we can show that adding one new point to a partition has the effect of not decreasing the lower sum and not increasing the upper sum, we will be done. In other words we might as well suppose that P_2 is obtained from P_1 by adding one more point.

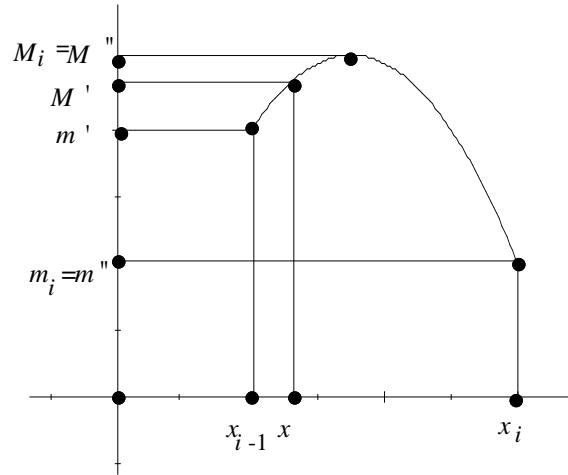
Suppose therefore that $P_1 = \{a = x_0, x_1, x_2, \dots, x_n = b\}$ and that $P_2 = P_1 \cup \{x\}$ ² with $x \in (x_{i-1}, x_i)$. Let M_j, m_j ($1 \leq j \leq n$) be the maximum and minimum values of f on $[x_{j-1}, x_j]$. Let M', m' be the maximum and minimum values for f on $[x_{i-1}, x]$; and M'', m'' be the maximum and minimum values for f on $[x, x_i]$. Note that

$$m' \geq m_i, \quad m'' \geq m_i$$

and

$$M' \leq M_i, \quad M'' \leq M_i$$

because on a sub-interval the minimum value can only increase and the maximum value can only decrease on a subinterval.



Then

$$\begin{aligned} L(P_2, f) - L(P_1, f) &= m'(x - x_{i-1}) + m''(x_i - x) - m_i(x_i - x_{i-1}) \\ &\geq m_i(x - x_{i-1}) + m_i(x_i - x) - m_i(x_i - x_{i-1}) \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} U(P_1, f) - U(P_2, f) &= M_i(x_i - x_{i-1}) - M'(x - x_{i-1}) - M''(x_i - x) \\ &\geq M_i(x_i - x_{i-1}) - M_i(x_i - x) - M_i(x_i - x) \\ &= 0. \end{aligned}$$

That is

$$L(P_1, f) \leq L(P_2, f) \quad \text{and} \quad U(P_2, f) \leq U(P_1, f).$$

□

In words: refining a partition increases lower sums and decreases upper sums.

COROLLARY 7.5. *If f is continuous on $[a, b]$ and if P_1, P_2 are arbitrary partitions of $[a, b]$, then $L(P_1, f) \leq U(P_2, f)$.*

PROOF. The partition P obtained by using *all* the points of P_1 and P_2 together, i.e. P is the union of P_1 and P_2 , is a refinement of both P_1 and P_2 . Hence

$$L(P_1, f) \leq L(P, f) \leq U(P, f) \leq U(P_2, f),$$

by Lemmas Lemma 7.3 and Lemma 7.4, as required. □

²Thus notation just means that P_2 is the union of the set P_1 and the set $\{x\}$ containing the single point x .

In other words : *every lower sum is less than or equal to every upper sum.*

The important consequence we need is this : since the lower sums $L(P, f)$ are all bounded above (by every upper sum in fact) the set of lower sums has a least upper bound. Similarly the set of upper sums is bounded below (by every lower sum) so the set of upper sums has a greatest lower bound. We define the *lower integral of f from a to b* and the *upper integral of f from a to b* by

$$L \int_a^b f := \text{l. u. b.} \{ L(P, f) : P \text{ is a partition of } [a, b] \}$$

$$U \int_a^b f := \text{g. l. b.} \{ U(P, f) : P \text{ is a partition of } [a, b] \}$$

respectively.

The next lemma just uses the fact that every lower sum is \leq every upper sum. It will soon be replaced by the stronger result that (for continuous functions) the lower and upper integrals are in fact equal.

LEMMA 7.6. *Let f be a continuous function on $[a, b]$. Then $L \int_a^b f \leq U \int_a^b f$.*

PROOF. Let P be a partition of $[a, b]$.

Since $U(P, f)$ is an upper bound for all lower sums, and since $L \int_a^b f$ is the *least* upper bound, it follows that

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

Since this is true for every partition P , $L \int_a^b f$ is thus a lower bound for the set of all upper bounds. Since $U \int_a^b f$ is the *greatest* lower bound, it follows that

$$L \int_a^b f \leq U \int_a^b f.$$

□

REMARK 7.7. ★ Everything we have done so far can also be done with an arbitrary *bounded*³ function f defined on $[a, b]$, except that we must define

$$M_i = \text{l. u. b.} \{ f(x) : x_{i-1} \leq x \leq x_i \}, \quad 1 \leq i \leq n,$$

$$m_i = \text{g. l. b.} \{ f(x) : x_{i-1} \leq x \leq x_i \}, \quad 1 \leq i \leq n.$$

Lemma 7.3, Lemma 7.4, Corollary 7.5 and Lemma 7.6 are still valid, with similar proofs as for continuous functions, but with “min” replaced by “g. l. b.” and “max” replaced by “l. u. b.”.

DEFINITION 7.8. A bounded function f defined on $[a, b]$ is *integrable* (in the sense of Riemann) if

$$L \int_a^b f = U \int_a^b f.$$

We call $L \int_a^b f$ and $U \int_a^b f$ respectively the *lower* and *upper integral* of f over $[a, b]$. For an integrable function we denote the common value of upper and lower integral by $\int_a^b f$ and call it the (definite) integral of f over $[a, b]$.

Note that $L \int_a^b f$ and $U \int_a^b f$ are *numbers*.

³A function is bounded if there exist numbers A and B such that $A \leq f(x) \leq B$ for every x in the domain of f . Thus any continuous function defined on $[a, b]$ is bounded. But the function f , with $f(x) = 1/x$ for $x \neq 0$ and $f(0) = 0$, is not bounded on its domain \mathbb{R} .

REMARK 7.9. ★ There is another type of integral called the *Lebesgue integral*. This is much more difficult to define, but it is much more powerful (more functions are integrable) and it has better properties (under very general conditions, if a sequence of functions $f_n(x)$ converges to $f(x)$ for every x , then the Lebesgue integrals of f_n converge to the Lebesgue integral of f). Such a convergence result is true for Riemann integration only if the functions converge in a rather strong sense. If a function is Riemann integrable then it is Lebesgue integrable (and the integrals agree), but the converse is not true.

For many applications, Riemann integration is sufficient, but for more sophisticated applications one needs the Lebesgue integral. There is a course on the *measure theory* and the Lebesgue integral in third year.

The case we will be mainly interested in is when f is continuous. In Theorem 7.10 we prove the important result that every continuous function on a closed bounded interval is integrable.

In general it is not the case that upper and lower integrals are equal. For example, consider the function f defined on $[0, 1]$ by

$$f(x) = \begin{cases} 0 & x \text{ is irrational,} \\ 1 & x \text{ is rational.} \end{cases}$$

Then, whatever partition P of $[0, 1]$ we have, $M_i = 1$ and $m_i = 0$ for every i , since every interval $[x_{i-1}, x_i]$ contains both rational and irrational points. Hence

$$L \int_0^1 f = 0 \quad \text{and} \quad U \int_0^1 f = 1.$$

THEOREM 7.10. *Let f be continuous on $[a, b]$. Then f is integrable.*

PROOF. Suppose f is continuous on $[a, b]$. Suppose $\varepsilon > 0$.

We will first show there exists some partition P (which may depend on ε) such that

$$(7.1) \quad U(P, f) - L(P, f) < \varepsilon.$$

The main point in proving this is to use the fact that by Theorem 5.19, f is *uniformly* continuous on $[a, b]$. Thus we may choose $\delta > 0$ such that

$$|x_1 - x_2| < \delta \quad \text{implies} \quad |f(x_1) - f(x_2)| < \frac{\varepsilon}{b-a}.$$

[We will see the reason for taking $\varepsilon/(b-a)$ in a moment.]

Now let $P = \{a = a_0, a_1, a_2, \dots, a_N = b\}$ be any partition of $[a, b]$ such that the difference between consecutive points in P is $< \delta$. Then by the above implication the difference between the maximum value M_i and the minimum m_i on the i th interval must be $< \varepsilon/(b-a)$. Hence

$$\begin{aligned} U(P, f) - L(P, f) &= \sum_{i=1}^N M_i \Delta x_i - \sum_{i=1}^N m_i \Delta x_i \\ &= \sum_{i=1}^N (M_i - m_i) \Delta x_i \\ &< \frac{\varepsilon}{b-a} (\Delta x_1 + \dots + \Delta x_N) \\ &= \frac{\varepsilon}{b-a} (b-a) = \varepsilon. \end{aligned}$$

This proves (7.1)

From the definition of the lower and upper integrals, and Lemma 7.6,

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

Since the difference between the outer two terms is $< \varepsilon$ by (7.1), the difference between the inner two terms is also $< \varepsilon$. That is

$$U \int_a^b f - L \int_a^b f < \varepsilon.$$

Since this holds for every $\varepsilon > 0$ it follows that $U \int_a^b f = L \int_a^b f$. \square

7.3. Riemann sums

The connection between Riemann sums and the Riemann integral is established.

If f is a continuous function on $[a, b]$ and P is a partition, then the upper and lower sums can be written in the form

$$U(P, f) = \sum_{i=1}^n f(u_i) \Delta x_i,$$

$$L(P, f) = \sum_{i=1}^n f(l_i) \Delta x_i.$$

where u_i and l_i are points in the i th interval $[x_{i-1}, x_i]$ for which f takes its maximum and minimum values respectively. More generally, we can define a *general Riemann sum* corresponding to the partition P by

$$R(P, f) = \sum_{i=1}^n f(c_i) \Delta x_i,$$

where each c_i is an arbitrary point in $[x_{i-1}, x_i]$. Note that this notation is a little imprecise, since $R(P, f)$ depends not only on the partition P , but also on the points c_i chosen in each of the intervals given by P .

Note that

$$(7.2) \quad L(P, f) \leq R(P, f) \leq U(P, f).$$

Let the maximum length of the intervals in a partition P be denoted by $\|P\|$.

THEOREM 7.11.

$$\lim_{\|P\| \rightarrow 0} R(P, f) = \int_a^b f.$$

More precisely, for any $\varepsilon > 0$ there exists a number $\delta > 0$ (which may depend on ε) such that

$$\text{whenever } \|P\| < \delta \text{ then } \left| R(P, f) - \int_a^b f \right| < \varepsilon.$$

PROOF. The proof of Theorem 7.10 in fact showed that if $\|P\| < \delta$ then

$$U(P, f) - L(P, f) < \varepsilon.$$

Since

$$L(P, f) \leq R(P, f) \leq U(P, f)$$

and

$$L(P, f) \leq \int_a^b f \leq U(P, f)$$

it follows (algebra) that

$$\left| R(P, f) - \int_a^b f \right| < \varepsilon.$$

\square

NOTATION 7.12. We often use the notation

$$\int_a^b f(x) dx \quad \text{for} \quad \int_a^b f.$$

Note that this is a number, not a function of x . It has *exactly* the same meaning as $\int_a^b f(y) dy$, just as $\sum_{i=1}^N f(c_i) \Delta x_i$ and $\sum_{j=1}^N f(c_j) \Delta x_j$ mean the same thing. We say x is a “dummy” variable.

You can informally think of $\int_a^b f(x) dx$ as the sum of the “areas” of an infinite number of triangles of height $f(x)$ and “infinitesimal” width “ dx ”. More precisely, from the previous theorem,

$$\int_a^b f(x) dx = \lim_{\|P\| \rightarrow 0} \sum_{i=1}^{N(P)} f(c_i) \Delta x_i.$$

(We write $N(P)$ to emphasise the fact that the number of points in the partition depends on P .)

7.4. Properties of the Riemann integral

The basic linearity and order properties of the Riemann integral are developed. The mean value theorem for integrals is proved. The extension of these results to piecewise continuous functions is noted.

In particular, if f, g are continuous functions on $[a, b]$ and c, d are real numbers, then

$$(7.3) \quad \int_a^b (cf + dg) = c \int_a^b f + d \int_a^b g$$

$$(7.4) \quad f(x) \leq g(x) \text{ for all } x \in [a, b] \quad \text{implies} \quad \int_a^b f \leq \int_a^b g.$$

The main point in the proofs is that similar properties are true for the Riemann sums used to define the integrals.

We also have, if $a < b$,

$$(7.5) \quad \left| \int_a^b f \right| \leq \int_a^b |f|$$

PROOF.

$$-|f(x)| \leq f(x) \leq |f(x)|$$

for all x . From (7.4)

$$\int_a^b -|f| \leq \int_a^b f \leq \int_a^b |f|.$$

From (7.3) this gives

$$-\int_a^b |f| \leq \int_a^b f \leq \int_a^b |f|.$$

This implies (7.5). □

If f is continuous on $[a, b]$ with minimum and maximum values m and M then

$$(7.6) \quad m(b-a) \leq \int_a^b f \leq M(b-a).$$

PROOF. Consider the partition $P = \{a, b\}$ containing just the two points a and b . Since

$$L(P, f) = m(b - a), \quad U(P, f) = M(b - a),$$

and

$$L(P, f) \leq L \int_a^b f = \int_a^b f = U \int_a^b f \leq U(P, f),$$

the result follows. \square

One also has, for $a \leq c \leq b$,

$$(7.7) \quad \int_b^a f = - \int_a^b f$$

$$(7.8) \quad \int_a^a f = 0$$

$$(7.9) \quad \int_a^c f + \int_c^b f = \int_a^b f$$

The first is really a definition. It also follows if we use the same definition of $\int_b^a f$ as in the case $b < a$, but allow “decreasing” partitions where $\Delta x_i < 0$. The second is again by definition. It also follows if we use the same definition as when the endpoints are distinct, except that now the points in any “partition” are all equal and so $\Delta x_i = 0$.

If we allow $b \leq a$ as well as $a < b$, then (7.5) should be replaced by

$$(7.10) \quad \left| \int_a^b f \right| \leq \left| \int_a^b |f| \right|$$

Exercise.

The *Mean Value Theorem for Integrals* says that if f is continuous on $[a, b]$ then there exists some $c \in [a, b]$ such that

$$(7.11) \quad \int_a^b f = (b - a)f(c).$$

PROOF. Choose l and u to be minimum and maximum points for f on $[a, b]$. Then from (7.6) it follows that

$$f(l) \leq \frac{\int_a^b f}{b - a} \leq f(u),$$

By the Intermediate Value Theorem applied to the function f on the interval $[l, u]$ or $[u, l]$ (depending on whether $l \leq u$ or $u \leq l$), there exists c between l and u such that

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

This gives the result. \square

Piecewise continuous functions (see [Adams, p. 316] for the definition) on a closed bounded interval are also integrable, and have the same properties as above. This essentially follows from writing any integrals as a sum of integrals over intervals on which the functions are all continuous.

7.5. Fundamental Theorem of Calculus

The relationship between integration and differentiation is developed.

The following theorem essentially says that differentiation and integration are reverse processes.

In the first part of the theorem we consider the integral $\int_a^x f$ as a function of the endpoint x (we allow $x \leq a$ as well as $x > a$) and prove: *the derivative of the integral of f gives back f .*

In the second part, we are saying that in order to compute $\int_a^b f$ it is sufficient to find a function G whose derivative is f and then compute $G(b) - G(a)$.

To put the second assertion in a form that looks more like the “reverse” of the first, we could write it in the form

$$\int_a^x G' = G(x) - G(a),$$

provided G' is a continuous function on I . We could even use f instead of G and then get

$$\int_a^x f' = f(x) - f(a),$$

provided f' is continuous on I . *The integral of the derivative of f gives back f (up to the constant $f(a)$).*

THEOREM 7.13 (Fundamental Theorem of Calculus). *Suppose that f is continuous on some interval I (not necessarily closed and bounded) and that $a \in I$.*

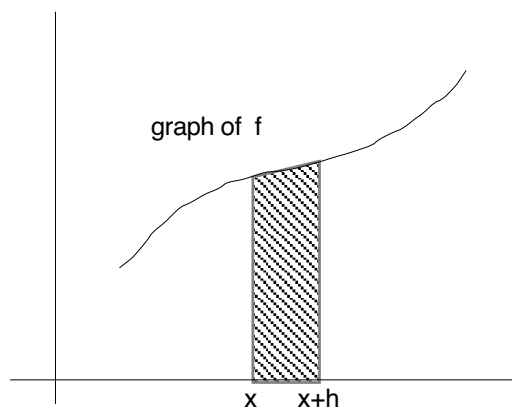
Then

$$\frac{d}{dx} \int_a^x f = f(x).$$

If $G'(x) = f(x)$ for all $x \in I$ then

$$\int_a^b f = G(b) - G(a).$$

PROOF.



⁴We could also write

$$\frac{d}{dx} \int_a^x f(t) dt = f(x).$$

The variable t is a dummy variable, and we could have used y or anything else instead. But it is “good practice” not to use x in this case, since we are already using x here to represent the endpoint of the interval of integration.

For the first assertion we have

$$\frac{d}{dx} \int_a^x f = \lim_{h \rightarrow 0} \frac{\int_a^{x+h} f - \int_a^x f}{h}$$

$$= \lim_{h \rightarrow 0} \frac{\int_x^{x+h} f}{h}$$

$$= \lim_{h \rightarrow 0} \frac{hf(c(h))}{h}$$

$$= \lim_{h \rightarrow 0} f(c(h))$$

$$= f(x)$$

from (7.9)

for some $c = c(h)$ between x and $x + h$,

depending on h , by the Mean Value Theorem for integrals.

since f is continuous at x

and c lies between x and $x + h$.

For the second assertion, suppose $G'(x) = f(x)$ on the interval I .

But we have just seen that the derivative (with respect to the variable x) of the function $\int_a^x f$ is also $f(x)$. It follows that the derivative of the function, given by

$$G(x) - \int_a^x f,$$

is $G'(x) - f(x) = 0$ on the interval I . Thus this function is constant on I by Corollary 6.13.

Setting $x = a$ we see that the constant is $G(a)$. Hence

$$G(x) - \int_a^x f = G(a)$$

for all $x \in I$. Taking $x = b$ now gives the second assertion. \square

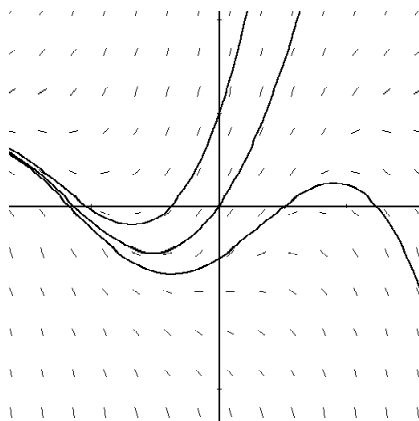
Differential Equations

The differential equation

$$(8.1) \quad \frac{dy}{dx} = f(x, y)$$

requires that the gradient of the function $y = y(x)$ at each point (x, y) on its graph should equal $f(x, y)$ for the given function f .

Suppose that at each point (x, y) on the $x - y$ plane we draw a little line whose slope is $f(x, y)$, this is the *slope field*. Then at every point on the graph of any solution to (8.1), the graph should be tangent to the corresponding little line. In the following diagram we have shown the slope field for $f(x, y) = y + \cos x$ and the graph of three functions satisfying the corresponding differential equation.



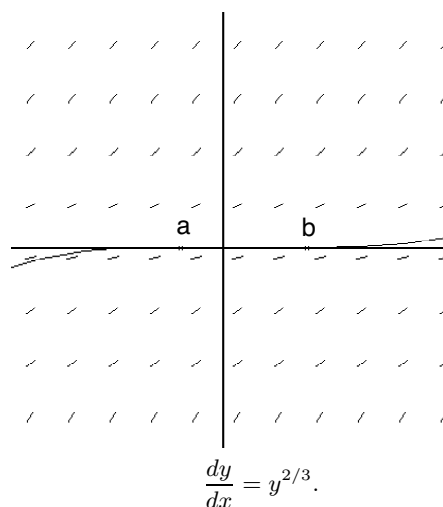
$$\frac{dy}{dx} = y + \cos x$$

It is plausible from the diagram that for any given point (x_0, y_0) there is exactly one solution $y = y(x)$ satisfying $y(x_0) = y_0$. This is indeed the case here, and is true under fairly general conditions.

But it is not always true. For example, if $f(x, y) = y^{2/3}$ then there is an infinite set of solutions satisfying $y(0) = 0$. Namely, for *any* real numbers $a \leq 0 \leq b$,

$$y = \begin{cases} \frac{(x-a)^3}{27} & x \leq a \\ 0 & a \leq x \leq b \\ \frac{(x-b)^3}{27} & x \geq b \end{cases}$$

is a solution, (*check it*). See the following diagram. The problem here is that although $f(x, y)$ is continuous everywhere, $(\partial/\partial y)f(x, y) = 2y^{-1/3}/3$ is not continuous on the x -axis. Notice that the slope lines on the x -axis are horizontal.



If the function f has even worse behaviour, there may be no solution at all.

In the two simple examples we just gave, we could write out the solutions in terms of standard functions. But in practice, this is almost never the case. The solutions of differential equations almost invariably *cannot* be expressed in terms of standard functions. In fact, one of the most useful ways to introduce new and useful functions is to *define* them as the solutions of certain differential equations. But in order to do this, we first need to know that the differential equations have unique solutions if we specify certain “initial conditions”. This is the main result in this chapter.

The point to this chapter is to prove the Fundamental Existence and Uniqueness Theorem for differential equations of the form (8.1). Such an equation is called *first-order*, since only the first order derivative of y occurs in the differential equation. Differential equations of the form (8.1) are essentially the most general first order differential equation.

The following remark justifies that we are about to prove a major result in mathematics!!

REMARK 8.1. ★ A *system of first order differential equations* for the dependent variables y_1, \dots, y_n is a set of differential equations of the form

$$\begin{aligned} \frac{dy_1}{dx} &= f_1(x, y_1, \dots, y_n) \\ \frac{dy_2}{dx} &= f_2(x, y_1, \dots, y_n) \\ &\vdots \\ \frac{dy_n}{dx} &= f_n(x, y_1, \dots, y_n) \end{aligned}$$

which are meant to be satisfied simultaneously by functions $y_1 = y_1(x), y_2 = y_2(x), \dots, y_n = y_n(x)$. Here the functions f_1, f_2, \dots, f_n are given. If $n = 2$ you can visualise this as in the one dimensional case, by considering three axes labeled x, y_1, y_2 . The solution to a differential equation in this case will be represented by the graph (curve) over the x axis which for each point x gives the point $(x, y_1(x), y_2(x))$.

A *very* similar proof as for a single differential equation, gives the analogous Fundamental Existence and Uniqueness Theorem for a system of first-order differential equations.

A differential equation which involves higher derivatives can be reduced to a system of differential equations of first order (essentially by introducing new variables for each of the higher order derivatives). Thus the Existence and Uniqueness Theorem, suitably modified, applies also to higher order differential equations. In fact it even applies to systems of higher order differential equations in a similar manner!.

8.1. Outline of proof of the Existence and Uniqueness theorem

Since I am realistic enough to know that not everyone is going to study the proof in Section 8.2 in detail (but it is not *that* difficult to follow), I will provide you here with an overview. (After that, hopefully you will then be inspired to work through the details.)

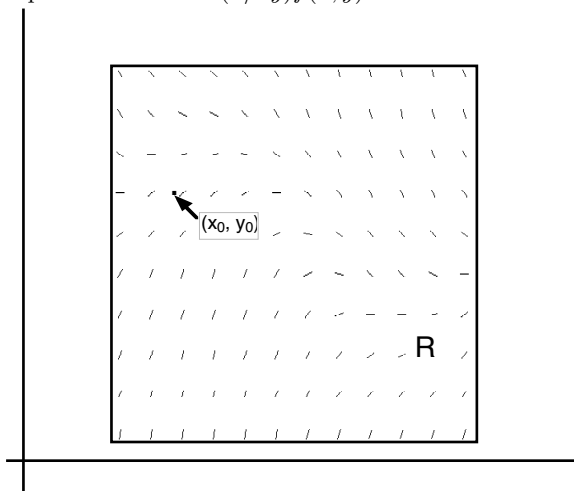
We want to prove that the the initial value problem

$$(8.2) \quad \frac{dy}{dx} = f(x, y)$$

$$(8.3) \quad y(x_0) = y_0$$

has exactly one solution under certain (general) assumptions.

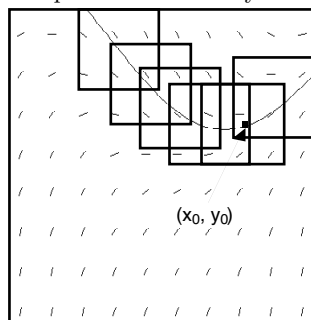
The assumptions are that $f(x, y)$ and $f_2(x, y) = (\partial/\partial y)f(x, y)$ are both *continuous* in some fixed (closed) rectangle R in the $x - y$ plane. Note that in particular, we are assuming that the partial derivative $(\partial/\partial y)f(x, y)$ exists in R .



Slope field for f in the rectangle R

We want to prove there is a unique solution passing through any point (x_0, y_0) in R . In fact the solution will go all the way to the boundary of R — top, bottom or one of the sides.

We will prove there is a solution in some smaller (open) rectangle centred at (x_0, y_0) , which passes through both sides. By then taking a new small rectangle centred at some point further along the solution for the first small rectangle, we can extend the solution. In fact, one can continue this process all the way¹ to the boundary of R .



6 small rectangles here get us all the way to the boundary of R

Thus the main point is to first show that in some (small) rectangle R_δ , whose base is of length 2δ and which is centred at the point (x_0, y_0) , there is a solution which extends to both *sides* of this small rectangle.

The proof proceeds in 6 steps.

¹★ We will be able to compute the size of these small rectangles, and in this way one can show that only a *finite* number of them are needed to “reach” the boundary of R .

Step A The problem is equivalent to showing the “integral equation”

$$(8.4) \quad y(x) = y_0 + \int_{x_0}^x f(t, y(t)) dt$$

has a solution. One sees this by integrating both sides of (8.2) from x_0 to x . Conversely, differentiating the integral equation gives back the differential equation, and clearly $y(x_0) = y_0$ follows from the integral equation.

For our first example

$$(8.5) \quad \frac{dy}{dx} = y + \cos x, \quad y(0) = 0,$$

we get

$$(8.6) \quad y(x) = \int_0^x (y(t) + \cos t) dt.$$

Step B To find the solution of (8.4) we begin with the constant function

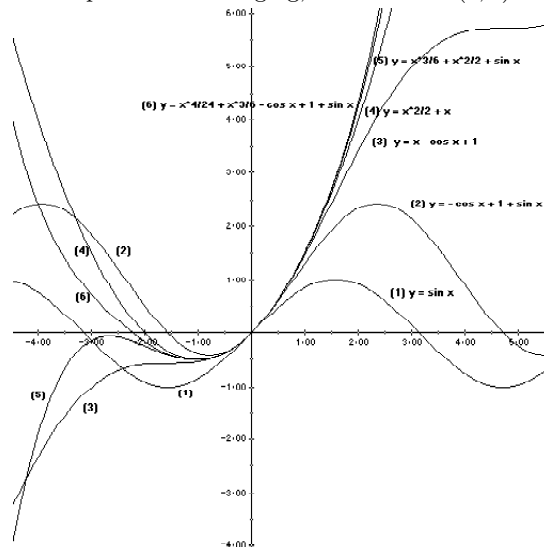
$$y(x) = y_0$$

and plug it into the right side of (8.4) to get a new function of x . We plug this again into the right side to get yet another function of x . And so on.

For example, with (8.5), substituting $y = 0$ in the right side of (8.6), and then repeating, we get

$$\begin{aligned} \int_0^x \cos t dt &\longrightarrow \sin x \\ \int_0^x (\sin t + \cos t) dt &\longrightarrow -\cos x + 1 + \sin x \\ \int_0^x (-\cos t + 1 + \sin t + \cos t) dt &\longrightarrow x - \cos x + 1 \\ \int_0^x (t - \cos t + 1 + \cos t) dt &\longrightarrow \frac{1}{2}x^2 + x \\ \int_0^x \left(\frac{1}{2}t^2 + t + \cos t\right) dt &\longrightarrow \frac{1}{6}x^3 + \frac{1}{2}x^2 + \sin x \\ \int_0^x \left(\frac{1}{6}t^3 + \frac{1}{2}t^2 + \sin t + \cos t\right) dt &\longrightarrow \frac{1}{24}x^4 + \frac{1}{6}x^3 - \cos x + 1 + \sin x \end{aligned}$$

We call this sequence of functions a “sequence of approximate solutions”. We see from the diagram that this sequence is converging, at least near $(0, 0)$.



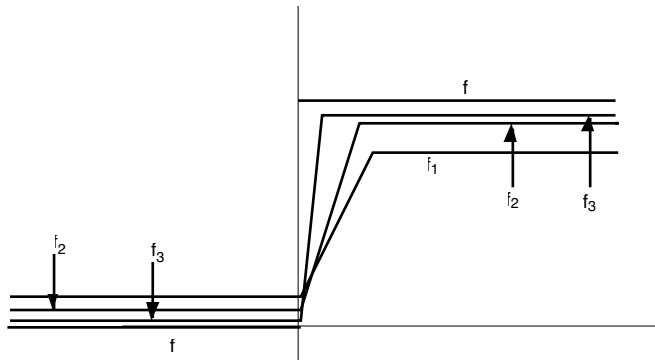
In general, if $y_n(x)$ is the n th approximate solution, then

$$(8.7) \quad y_{n+1}(x) = y_0 + \int_{x_0}^x f(t, y_n(t)) dt.$$

Step C the next step is to show that on some small rectangle around (x_0, y_0) this sequence of “approximate solutions” does indeed converge. The main point in the proof is showing that if the rectangle is sufficiently small then the distance between the n th and $(n + 1)$ th approximate solutions is $< r$ times the distance between the $(n - 1)$ th and the n th approximate solutions, for some fixed $r < 1$.

Thus the distance between consecutive solutions is decreasing “geometrically” fast. This is the main idea in the proof.

Step D Let the limit function for the approximate solutions be denoted by $y = y(x)$. The next step is to show that this limit function is continuous. This is not obvious, although it is not hard to show that the approximate solutions are themselves continuous. The problem is that a sequence of continuous functions can in fact converge to a non-continuous function, as in the following diagram. But in our case the fact that the approximate solutions converge “geometrically” fast is enough to ensure that the limit *is* continuous



Step E The next step is to show that the limit function $y = y(x)$ satisfies the integral equation. The fact it is continuous implies we *can* integrate the right side of (8.4). And the fact that the approximate solutions converge to the function $y = y(x)$ geometrically fast enables us to prove that we *can* take the limit as $n \rightarrow \infty$ on both sides of (8.7) and deduce (8.4)

Step F The final step is to show that any two solutions are equal. We show that if d is the distance between two solutions of (8.4) then $d \leq rd$ for some $r < 1$, by an argument like the one in Step C. This implies that $d = 0$ and so the two solutions agree.

8.2. ★Rigorous proof of the Existence and Uniqueness theorem

THEOREM 8.2. *Suppose that $f(x, y)$ and $f_2(x, y) = (\partial/\partial y)f(x, y)$ are both continuous in the rectangle R consisting of all points (x, y) of the form $a \leq x \leq b$, $c \leq y \leq d$. Suppose (x_0, y_0) is in the interior of R .*

Then there exists a number $\delta > 0$ and a unique function $\phi(x)$, defined and having a continuous derivative on the interval $(x_0 - \delta, x_0 + \delta)$, such that

$$(8.8) \quad \phi'(x) = f(x, \phi(x))$$

$$(8.9) \quad \phi(x_0) = y_0.$$

In other words, $\phi(x)$ solves (i.e. satisfies) the initial value problem

$$\begin{aligned} \frac{dy}{dx} &= f(x, y) \\ y(x_0) &= y_0 \end{aligned}$$

on the interval $(x_0 - \delta, x_0 + \delta)$.

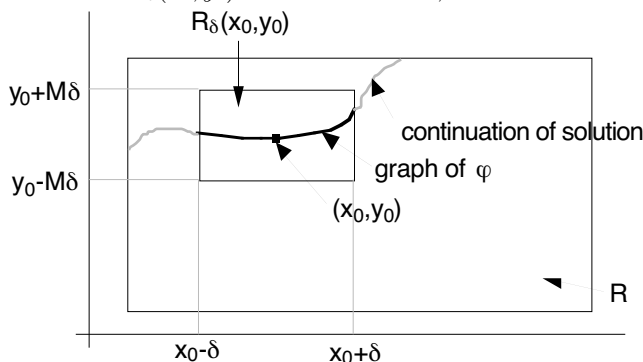
Remark: Let

$$M = \max\{|f(x, y)| : (x, y) \in R\}, \quad K = \max\left\{\left|\frac{\partial}{\partial y}f(x, y)\right| : (x, y) \in R\right\}.$$

We will see in the proof that if we define $R_\delta(x_0, y_0)$ to be the (open) rectangle consisting of all those (x, y) such that $x_0 - \delta < x < x_0 + \delta$ and $y_0 - M\delta < y < y_0 + M\delta$, i.e.

$$(8.10) \quad R_\delta(x_0, y_0) = \left\{ (x, y) : x \in (x_0 - \delta, x_0 + \delta), y \in (y_0 - M\delta, y_0 + M\delta) \right\},$$

then any $\delta > 0$ for which $R_\delta(x_0, y_0) \subset R$ and $\delta < K^{-1}$, will work for the above theorem.



PROOF.

Step A We first *claim* that if $\phi(x)$ is a continuous function defined on some interval $(x_0 - \delta, x_0 + \delta)$, and $(x, \phi(x)) \in R$ for all x , then the following two statements are equivalent:

1. $\phi(x)$ has a continuous derivative on the interval $(x_0 - \delta, x_0 + \delta)$ and solves the given initial value problem there, i.e. (8.8) and (8.9) are true;
2. $\phi(x)$ satisfies the integral equation

$$(8.11) \quad \phi(x) = y_0 + \int_{x_0}^x f(t, \phi(t)) dt.$$

Assume the *first* statement is true. Then both $\phi'(t)$, and $f(t, \phi(t))$ by Section 5.7, are continuous on $(x_0 - \delta, x_0 + \delta)$ (it is convenient to use t here instead of x for the dummy variable). Thus for any x in the interval $(x_0 - \delta, x_0 + \delta)$ the following integrals exist, and from (8.8) they are equal:

$$\int_{x_0}^x \phi'(t) dt = \int_{x_0}^x f(t, \phi(t)) dt.$$

From the Fundamental Theorem of Calculus it follows that

$$\phi(x) - \phi(x_0) = \int_{x_0}^x f(t, \phi(t)) dt,$$

which implies the *second* statement (since we are assuming $\phi(x_0) = y_0$).

Next assume the *second* statement is true. Note that since $\phi(t)$ is continuous, so is $f(t, \phi(t))$ by Section 5.7, and so the integral *does* exist. Setting $x = x_0$ we immediately get (8.9)

Since the right side of (8.11) is differentiable and the derivative equals $f(x, \phi(x))$ (by the Fundamental Theorem of Calculus), the left side must also be differentiable and have the same derivative. That is, (8.8) is true for any x in the interval $(x_0 - \delta, x_0 + \delta)$. Moreover, we see that the derivative $\phi'(x)$ is continuous since $f(x, \phi(x))$ is continuous.

Thus the *first* statement is true.

Step B We now define a sequence of approximations to a solution of (8.11) as follows:

$$\begin{aligned}\phi_0(x) &= y_0 \\ \phi_1(x) &= y_0 + \int_{x_0}^x f(t, \phi_0(t)) dt \\ \phi_2(x) &= y_0 + \int_{x_0}^x f(t, \phi_1(t)) dt \\ &\vdots \\ \phi_{n+1}(x) &= y_0 + \int_{x_0}^x f(t, \phi_n(t)) dt \\ &\vdots\end{aligned}$$

The functions in the above sequence will be defined for all x in some interval $(x_0 - \delta, x_0 + \delta)$, where the δ has yet to be chosen. We will first impose the restriction on δ that

$$(8.12) \quad R_\delta(x_0, y_0) \subset R,$$

where $R_\delta(x_0, y_0)$ was defined in (8.10).

The function $\phi_0(x)$ is just a constant function.

Since the points $(t, \phi_0(t))$ certainly lie in $R_\delta(x_0, y_0)$ if $t \in (x_0 - \delta, x_0 + \delta)$, it follows that $f(t, \phi_0(t))$ makes sense. Also, $f(t, \phi_0(t))$ is a continuous function of t from Section 5.7, being a composition of continuous functions. It follows that the integral used to define $\phi_1(x)$ exists if $x \in (x_0 - \delta, x_0 + \delta)$. In other words the definition of $\phi_1(x)$ makes sense for $x \in (x_0 - \delta, x_0 + \delta)$.

Next, for $x \in (x_0 - \delta, x_0 + \delta)$, we show that $(x, \phi_1(x)) \in R_\delta(x_0, y_0)$ and hence $\in R$. This follows from the fact that

$$\begin{aligned}|\phi_1(x) - y_0| &= \left| \int_{x_0}^x f(t, \phi_0(t)) dt \right| \\ &\leq \left| \int_{x_0}^x |f(t, \phi_0(t))| dt \right| && \text{from (7.10)} \\ &\leq \left| \int_{x_0}^x M dt \right| && \text{since } |f| \leq M \text{ in } R \\ &\leq M\delta && \text{since } |x - x_0| \leq \delta.\end{aligned}$$

It follows as before that the definition of $\phi_2(x)$ makes sense for $x \in (x_0 - \delta, x_0 + \delta)$. (We also need the fact that $f(t, \phi_1(t))$ is continuous. This follows from the fact $\phi_1(t)$ is in fact differentiable by the Fundamental Theorem of Calculus, and hence continuous; and the fact that $f(t, \phi_1(t))$ is thus a composition of continuous functions and hence continuous.)

Etc. etc. (or proof by induction, to be rigorous; but it is clear that it will work).

In this way we have a sequence of continuous functions $\phi_n(x)$ defined on the interval $(x_0 - \delta, x_0 + \delta)$, and for x in this interval we have $(x, \phi_n(x)) \in R_\delta(x_0, y_0)$.

Step C The next step is to prove there exists a function $\phi(x)$ defined on the interval $(x_0 - \delta, x_0 + \delta)$ such that

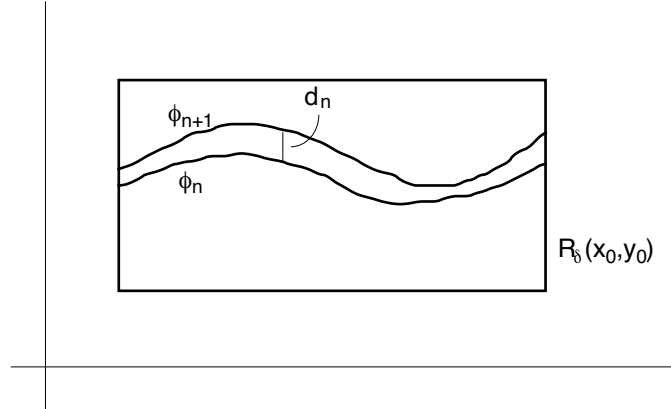
$$\phi_n(x) \rightarrow \phi(x)$$

for all $x \in (x_0 - \delta, x_0 + \delta)$. Let (for $n \geq 0$)

$$d_n = \max |\phi_n(x) - \phi_{n+1}(x)|,$$

where the maximum is taken over the interval $(x_0 - \delta, x_0 + \delta)$.²

²There is a minor technical point here. Since the points $(x, \phi_n(x))$ and $(x, \phi_{n+1}(x))$ both lie in $R_\delta(x_0, y_0)$, it follows that $|\phi_n(x) - \phi_{n+1}(x)| < 2M\delta$. But the maximum may be "achieved" only when $x = x_0 \pm \delta$, which is not actually a point in the (open) interval $(x_0 - \delta, x_0 + \delta)$. To make the argument rigorous, we should replace "max" by "l. u. b." in Step C.



Then for $n \geq 1$

$$\begin{aligned}
 d_n &= \max_{x \in (x_0 - \delta, x_0 + \delta)} |\phi_n(x) - \phi_{n+1}(x)| \\
 &= \max_{x \in (x_0 - \delta, x_0 + \delta)} \left| \int_{x_0}^x f(t, \phi_{n-1}(t)) - f(t, \phi_n(t)) dt \right| \\
 &\leq \max_{x \in (x_0 - \delta, x_0 + \delta)} \left| \int_{x_0}^x |f(t, \phi_{n-1}(t)) - f(t, \phi_n(t))| dt \right| \\
 &\leq \max_{x \in (x_0 - \delta, x_0 + \delta)} \left| \int_{x_0}^x K |\phi_{n-1}(t) - \phi_n(t)| dt \right| && \text{by Section 6.6} \\
 &\leq \max_{x \in (x_0 - \delta, x_0 + \delta)} \left| \int_{x_0}^x K d_{n-1} dt \right| && \text{from the definition of } d_{n-1} \\
 &= K\delta d_{n-1}
 \end{aligned}$$

Repeating this argument we obtain

$$d_n \leq K\delta d_{n-1} \leq (K\delta)^2 d_{n-2} \leq (K\delta)^3 d_{n-3} \leq \dots \leq (K\delta)^n d_0.$$

We now make the further restriction on δ that

$$(8.13) \quad K\delta < 1.$$

Since

$$|\phi_n(x) - \phi_{n+1}(x)| \leq d_n \leq d_0 (K\delta)^n,$$

it follows from Theorem 4.10 that the sequence $\phi_n(x)$ converges for each $x \in (x_0 - \delta, x_0 + \delta)$. We define the function $\phi(x)$ on $(x_0 - \delta, x_0 + \delta)$ by

$$\phi(x) = \lim_{n \rightarrow \infty} \phi_n(x).$$

It also follows from Theorem 4.10 that

$$(8.14) \quad |\phi_n(x) - \phi(x)| \leq \frac{d_0}{1 - K\delta} (K\delta)^n = Ar^n.$$

where $A = d_0/(1 - K\delta)$ and $r = K\delta < 1$. (Note that this is saying that the graph of ϕ_n lies within distance Ar^n of the graph of ϕ .)

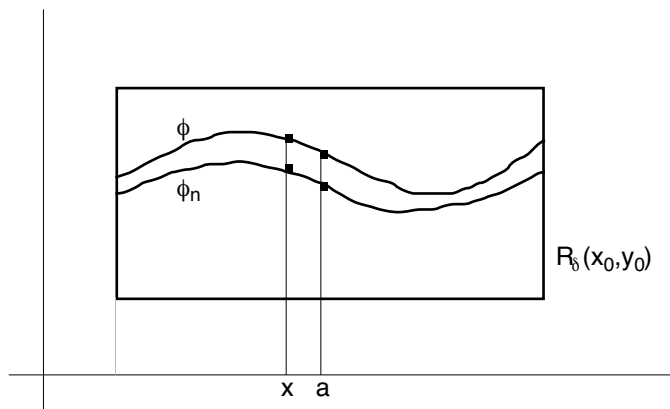
Step D We next claim that $\phi(x)$ is continuous on the interval $(x_0 - \delta, x_0 + \delta)$.

To see this let a be any point in the interval $(x_0 - \delta, x_0 + \delta)$; we will prove that ϕ is continuous at a .

Let $\varepsilon > 0$ be an arbitrary positive number.

First choose n so that $Ar^n < \varepsilon/3$ and hence from (8.14)

$$(8.15) \quad x \in (x_0 - \delta, x_0 + \delta) \quad \text{implies} \quad |\phi_n(x) - \phi(x)| \leq \varepsilon/3.$$



By continuity of ϕ_n there exists $\eta > 0$ (which may depend on n and hence on ε) such that

$$(8.16) \quad |x - a| < \eta \quad \text{implies} \quad |\phi_n(x) - \phi_n(a)| < \varepsilon/3.$$

(We also choose η sufficiently small that if $|x - a| < \eta$ then $x \in (x_0 - \delta, x_0 + \delta)$.)

From (8.15) (applied with x and again with x replaced by a) and (8.16) it follows that if $|x - a| < \eta$ then

$$\begin{aligned} |\phi(x) - \phi(a)| &= |(\phi(x) - \phi_n(x)) + (\phi_n(x) - \phi_n(a)) + (\phi_n(a) - \phi(a))| \\ &\leq |\phi(x) - \phi_n(x)| + |\phi_n(x) - \phi_n(a)| + |\phi_n(a) - \phi(a)| \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Since a was any point in $(x_0 - \delta, x_0 + \delta)$, and ε was any positive number, this proves the *claim* that ϕ is continuous on the interval $(x_0 - \delta, x_0 + \delta)$.

Step E We defined

$$(8.17) \quad \phi_{n+1}(x) = y_0 + \int_{x_0}^x f(t, \phi_n(t)) dt.$$

We have shown in Step C that³

$$\phi_{n+1}(x) \rightarrow \phi(x)$$

for each x in the interval $(x_0 - \delta, x_0 + \delta)$. We next *claim* that for the right side of (8.16) we have

$$y_0 + \int_{x_0}^x f(t, \phi_n(t)) dt \rightarrow y_0 + \int_{x_0}^x f(t, \phi(t)) dt.$$

It then follows from the *claim* that

$$\phi(x) = y_0 + \int_{x_0}^x f(t, \phi(t)) dt,$$

which establishes (8.11) and hence proves the theorem by Step A.

To prove the *claim* we compute

$$\begin{aligned} \left| \int_{x_0}^x f(t, \phi_n(t)) dt - \int_{x_0}^x f(t, \phi(t)) dt \right| &\leq \left| \int_{x_0}^x |f(t, \phi_n(t)) - f(t, \phi(t))| dt \right| \\ &\leq \left| \int_{x_0}^x K |\phi_n(t) - \phi(t)| dt \right| \quad \text{by Section 6.6} \\ &\leq \left| \int_{x_0}^x K A r^n dt \right| \quad \text{by (8.14)} \\ &\leq K \delta A r^n. \end{aligned}$$

³If $a_n \rightarrow a$ for a sequence, then it follows that $a_{n+1} \rightarrow a$, *why*?

Since $0 \leq r < 1$ this establishes the *claim* and hence the theorem.

Step F Finally, we must show that any two solutions of (8.8) and (8.9), or equivalently of (8.11), are equal.

Suppose that $\phi(x)$ and $\psi(x)$ are any two solutions. Then if

$$d = \max |\phi(x) - \psi(x)|,$$

where the maximum is taken over the interval $(x_0 - \delta, x_0 + \delta)$.⁴ Then for any $x \in (x_0 - \delta, x_0 + \delta)$,

$$\begin{aligned} |\phi(x) - \psi(x)| &= \left| \int_{x_0}^x (f(t, \phi(t)) - f(t, \psi(t))) dt \right| \\ &\leq \left| \int_{x_0}^x |f(t, \phi(t)) - f(t, \psi(t))| dt \right| \\ &\leq \left| \int_{x_0}^x K|\phi(t) - \psi(t)| dt \right| \quad \text{from Section 6.6} \\ &\leq K\delta d \end{aligned}$$

Since this is true for *any* $x \in (x_0 - \delta, x_0 + \delta)$, it follows that

$$d \leq K\delta d.$$

Since $K\delta < 1$, this implies $d = 0$!!

Hence $\phi(x) = \psi(x)$ for all $x \in (x_0 - \delta, x_0 + \delta)$. □

End of proof, end of chapter, end of semester. Have a good holiday.

⁴As in Step C we should really write “l. u. b.” instead of “max”. The proof is essentially unchanged.

Bibliography

- [Adams] Robert A. Adams, *Calculus*, Third ed., Addison Wesley, 1995.
- [Birkhoff and MacLane] Garrett Birkhoff and Saunders MacLane, *A Survey of Modern Algebra*, Macmillan, 1965.
- [Davis and Hersh] Philip Davis and Reuben Hersh, *The Mathematical Experience*, Birkhauser, 1981.
- [Devlin] Keith Devlin, *Mathematics, The Science of Patterns*, Scientific American Library, 1994.
- [Hildebrandt and Tromba] Stefan Hildebrandt and Anthony Tromba, *The Parsimonius Universe, Shape and Form in the Natural World*, Springer-Verlag, 1996.
- [Spivak] Michael Spivak, *Calculus*, various editions and publishers.
- [Stromberg] Karl R Stromberg, *An Introduction to Classical Real Analysis*, Chapman and Hall.
- [Ward] Martin Ward, *Mathematics AA1 Calculus Notes*, 1998.