FOUNDATIONS of CALCULUS

John Hutchinson

email: John.Hutchinson@anu.edu.au

Contents

Introduction

These notes present the theoretical foundations of Calculus. As such they are an introduction to the mathematical field of Analysis. More generally, they are an introduction to the methods used in modern mathematics. In the process of working through this material you will prove the major theorems used in the Calculus section of MATH1115 and to a lesser extent in MATH1116, and thereby obtain a more fundamental understanding of that material.

Some of the material here will be part of your first year courses, some will be supplementary and covered in later courses. It is all important mainstream mathematics.

There are essentially two parts to the Calculus section of MATH1115 — theoretical foundations on the one hand; methods, techniques and applications on the other. The former is treated here, more thoroughly than in Adams. The latter is covered in Adams, and is an extension of the material in the higher level school/college calculus courses.

Mathematics is the study of pattern and structure. It is studied both for its universal applicability and its internal beauty. 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. The study of these disciplines can usually only be done by applying the techniques and language of mathematics. On the other hand, their study often leads to the development of new fields of mathematics and gives insights into old fields.

In these notes you 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, you 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 important question to be investigated in the last chapter 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 earlier material. The study of differential equations is of tremendous importance in mathematics and for its applications. Any phenomenon that changes with position and/or time is usually represented by one or more differential equations.

The ideas developed here are basic to further developments in mathematics. The concepts 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 relevant lectures rather than trying to write everything down. There may occcasionally be lecture material on this part of the course which is not mentioned in the notes here, in which case that will be indicated.

There are quite a few footnotes. This can be annoying. You should read the footnotes when you initially study the material. But after you have noted and understood the footnotes, you do not need to reread them every time. Instead, you should concentrate on the main ideas in the main body of the notes.

References are to the seventh edition of the text *Calculus* text by Adams, and occasionally to the book *Calculus* by Michael Spivak.

Why go 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 the material is explained in a less formal manner, the key and underlying ideas are singled out and discussed, and generally the subject is explained and discussed in a manner which it is not possible to do efficiently in print. It would be a *very big mistake* to skip lectures.

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 these notes I make various digressions and additional remarks, marked clearly by a star ★. This would generally be *non-examinable* and is (even) 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.

There are a number of places where I ask *why?* Don't just convince yourself informally that it is indeed so. Write down a careful proof and then copy it into the margin of these notes.

Some of the proofs of theorems are quite tricky, certainly at first. In this case just try to understand what the theorem is saying, think about some examples, and think why the various hypotheses are necessary, and think about how they are used in the proof. There are examples which discuss some of these points before or after some of the more difficult theorems.

Studying mathematics is not like reading a book in other subjects. It may take many hours to understand just one sentence or one paragraph. When you get stuck, it will often help to eventually continue on, and then later come back to the difficult points. Also, ask your tutor, your lecturer, a fellow student, or an assistant in the mathematics "drop in" centre. Do not let things slide!

The study of Mathematics is not easy, but it is challenging, rewarding and enjoyable.

CHAPTER 1

The Real Numbers

You should begin by first reviewing the material in Chapter P1 of Adams, pages 3–9; and particularly pages 3 and 4.

We begin with a brief discussion of a few properties of the real numbers. We then discuss the algebraic and order properties and indicate how they follow from 13 basic properties called the *Algebraic and Order Axioms*. Finally we discuss the *Completeness Axiom*. All properties of the real numbers follow from these 14 axioms.

1.1. Preliminary remarks

1.1.1. Decimal expansions. Real numbers have decimal expansions, for example:

$$2 = 2.000\ldots$$
$$1\frac{1}{2} = 1.5 = 1.5000\ldots$$
$$\pi = 3.14159\ldots$$
$$.4527\dot{1}4\dot{6}, \quad \text{also written } .4527\overline{146}.$$

The "..." 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 cases the expansion continues with zeros and in the third case one can compute the expansion to any required degree of accuracy.

Instead of counting with base 10, we could count with any other integer base $b \geq 2$. In this case, for integers we write

$$b_1 b_2 \ldots b_n = b^{n-1}b_1 + \cdots + b^2 b_{n-2} + b b_{n-1} + b_n,$$

where each b_i takes values in the set $\{0, 1, \ldots, b-1\}$.(Some societies did count with other than base 10.)

How is 123 *written in base* 16 *and in base* 2*?*

We can also write nonintegers using base b. In particular,

$$\cdot b_1 b_2 \ldots b_n = \frac{b_1}{b} + \frac{b_2}{b^2} + \cdots + \frac{b_n}{b^n}.$$

1.1.2. Geometric representation. Real numbers can be represented geometrically as points on an infinite line.

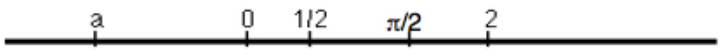

FIGURE 1. Geometric representation of real numbers as points on an (infinite) line.

The ancient Greeks thought of real numbers as *lengths* of lines, and they knew that if x is the length of the hypotenuse of the following right angled triangle, then its square must satisfy $x^2 = 1^2 + 1^2 = 2$ (Pythagoras's theorem). We write $\sqrt{2}$ for this number x.

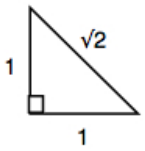

FIGURE 2. Is the length of the hypotenuse given by a rational number? No!

The Greeks also thought of real numbers as *ratios* of integers (or what we now call *rational numbers*).

So when they discovered the following theorem, they were very upset.

THEOREM 1.1.1. *$\sqrt{2}$ 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 cancelling 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.)

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 that they have no common factors.

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

Use a similar argument to prove $\sqrt{3}$ *is irrational.* HINT: Instead of considering even and odd integers (i.e. remainder 0 or 1 after dividing by 2), you will need to consider integers with remainders 0, 1 or 2 after dividing by 3.

1.1.3. Different decimal expansions for the same number. 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, $.999\ldots$, with the 9's repeated forever.

Each of the following *approximations* to $.\dot{9}$,

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

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

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

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

$$10a = 9.999\ldots.$$

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

The *only* way a real number can have two decimal expansions is for it to be of the form

$$.a_1a_2\ldots a_{n-1}a_n = .a_1a_2\ldots a_{n-1}(a_n - 1)\dot{9}\ .$$

For example,

$$.2356 = .2355\dot{9}\ .$$

1.1.4. Density of the rationals and the irrationals. We claim that between any two real numbers there is a rational number (in fact infinitely many of them) and an irrational number (in fact infinitely many).

To see this, first suppose $0 < a < b$.

Choose an integer n such that $\frac{1}{n} < b - a$. Then at least one member $\frac{m}{n}$ of the sequence

$$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \frac{4}{n}, \frac{5}{n}, \ldots$$

will lie between a and b. To see this, take the *first* integer m such that $a < \frac{m}{n}$. It follows that $\frac{m}{n} < b$. *Why?*[1]

Since we can similarly obtain another rational between $\frac{m}{n}$ and b, and yet another rational between *this* rational and b, etc., etc., there is in fact an infinite number of rationals between a and b.

[1] First try to understand this geometrically. Then write out an algebraic proof, which should only be a couple of lines long!

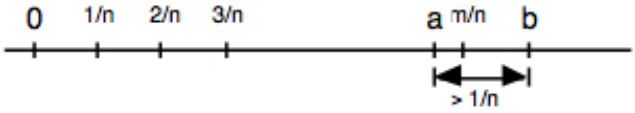

FIGURE 3. Since $b - a > 1/n$, some integer multiple of $1/n$ must lie between a and b.

If $a < 0$, a similar argument works with the sequence

$$-\frac{1}{n}, -\frac{2}{n}, -\frac{3}{n}, -\frac{4}{n}, -\frac{5}{n}, \dots$$

Finally, choosing n so $\frac{\sqrt{2}}{n} < b - a$ and applying a similar argument to the sequence

$$\frac{\sqrt{2}}{n}, \frac{2\sqrt{2}}{n}, \frac{3\sqrt{2}}{n}, \frac{4\sqrt{2}}{n}, \frac{5\sqrt{2}}{n}, \dots$$

gives the result for irrational[2] numbers.

1.2. Algebraic and Order Properties

We introduce the Algebraic and Order Axioms for the real number system and indicate how all the usual algebraic and order properties follow from these.

In a later section we discuss the Completeness Axiom.

1.2.1. Algebraic and Order Axioms. The *real number system* consists of the real numbers, together with the two 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 $a \times b$. We say that the real numbers are *closed* under addition and multiplication. We usually write

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

For any two real numbers a and b, the statement $a < b$ is either true or false.

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

Algebraic Axioms. For all real numbers a, b and c:

(1) $a + b = b + a$ (commutative axiom for addition)
(2) $(a + b) + c = a + (b + c)$ (associative axiom for addition)
(3) $a + 0 = 0 + a = a$ (additive identity axiom)
(4) there is a real number, denoted $-a$, such that $a + (-a) = (-a) + a = 0$ (additive inverse axiom)
(5) $a \times b = b \times a$ (commutative axiom for multiplication)
(6) $(a \times b) \times c = a \times (b \times c)$ (associative axiom for multiplication)

[2]The number $\frac{m\sqrt{2}}{n}$ is irrational, because if it were rational then on multiplying by $\frac{n}{m}$ we would get that $\sqrt{2}$ is rational, which we know is not the case.

(7) $a \times 1 = 1 \times a = a$, moreover $0 \neq 1$ (multiplicative identity axiom)
(8) if $a \neq 0$ there is a real number, denoted a^{-1}, such that $a \times a^{-1} = a^{-1} \times a = 1$ (multiplicative inverse axiom)
(9) $a \times (b + c) = a \times b + a \times c$ (distributive axiom)

Order Axioms. For all real numbers a, b and c:

(10) exactly one of the following holds: $a < b$ or $a = b$ or $b < a$ (trichotomy axiom)
(11) if $a < b$ and $b < c$, then $a < c$ (transitivity axiom)
(12) if $a < b$ then $a + c < b + c$ (addition and order axiom)
(13) if $a < b$ and $0 < c$, then $a \times c < b \times c$ (multiplication and order axiom)

There are a number of points that need to be made at this stage, before we proceed to discuss the consequences of these axioms.

- By the symbol "=" for equality we mean "denotes the same thing as", or equivalently, "represents the same real number as". We take "=" to be a *logical notion* and do not write down axioms for it.[3] 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 just mean that a does *not* denote the same real number as b.
- We are not really using subtraction in the algebraic Axiom 4; we are merely asserting that a real number, with a certain property, exists. It is convenient to denote this number by $-a$. A similar remark applies for Axiom 8.
- The assertion $0 \neq 1$ in Axiom 7 may seem silly. But it does not follow from the other axioms, since all the other axioms hold for the set containing just the number 0.
- Parts of some of the axioms are redundant. For example, from Axiom 1 and the property $a + 0 = a$ it follows that $0 + a = a$. Similar comments apply to Axiom 4; and because of Axiom 5 to Axioms 7 and 8.

1.2.2. Algebraic consequences. All the usual algebraic properties of the real numbers follow from Axioms 1–9. In particular, one can solve simultaneous systems of linear equations. We will not spend much time on indicating how one deduces 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.

Nonetheless, it is useful to have some idea of the methods involved in making deductions from axioms.

[3] ★One *can* write down basic properties, i.e. axioms, for "=" and the logic we use. See later courses on the foundations of mathematics.

1.2.2.1. *Sum of three or more numbers.* The expression $a + b + c$ is at first ambiguous. Does it mean $(a+b)+c$ or $a+(b+c)$? The first expression means add a to b, then add c to the result; the second means add a to (the result of adding b to c). By the associative axiom, the result is the same in either case, and so we can define $a + b + c$ to be *either* $(a + b) + c$ or $a + (b + c)$.[4]

Using also the commutative axiom we also have

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

etc.

Similar remarks apply to the product of three or more numbers.

1.2.2.2. *Subtraction and Division.* We use the axioms to *define* these operations in terms of addition and multiplication by

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

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 Axiom 4 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 $a \div b$. Division by 0 is never defined.

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

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

1.2.2.4. *Cancellation property of addition.* As an example of how to use Axioms 1–9 to derive other algebraic properties, we prove the cancellation property of addition, which says informally that if $a + c = b + c$ then we can "cancel" the number c.

THEOREM 1.2.1 (Cancellation Theorem). *If a, b and c are real numbers and $a + c = b + c$, then $a = b$. Similarly, if $c + a = c + b$ 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 Axiom 4.) Hence

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

from Axiom 2 applied twice, once to each side of the equation. Hence

$$a + 0 = b + 0$$

[4] A similar remark is not true for subtraction, since $(a - b) - c$ and $a - (b - c)$ are in general not equal.

from Axiom 4 again applied twice. Finally,

$$a = b$$

from Axiom 3.

If $c + a = c + b$, then from the commutative axiom $a + c = b + c$, and we have just seen that this implies $a = b$. □

1.2.2.5. *Characterisation of* 0 *and of* $-a$. One of the axioms is that 0 has the property:

$$a + 0 = 0 + a = a \tag{1}$$

for every real number a. Does this property characterise 0? In other words, is there any other real number x with the property that

$$a + x = x + a = a \tag{2}$$

for every real number a?

Of course we know the answer is NO, but the point here is that the answer follows from the axioms.

In fact from (1) and (2) we have $a + 0 = a + x$, and so from (the second part of) Theorem 1.2.1 with c, a, b there replaced by $a, 0, x$ respectively, it follows that $0 = x$.

One of the axioms asserts that for each number a there is a number (denoted by $-a$) which when added to a gives zero. But could there be another number which when added to a also gives zero? We know the answer is NO, but this fact does not need to be asserted as a separate axiom, because it also follows from the existing axioms.

In fact, suppose $x + a = 0$. Because we already know $(-a) + a = 0$ for some specific number $-a$, it follows that $x + a = (-a) + a$. We can "cancel" the a by Theorem 1.2.1, giving $x = -a$.

1.2.2.6. *More algebraic consequences.* Certain not so obvious "rules", such as "the product of minus and minus is plus" and the rule for adding two fractions, follow from the axioms. If we want the properties given by Axioms 1–9 to be true for the real numbers (and we do), then there is no choice other than to have $(-a)(-b) = ab$ and $(a/c) + (b/d) = (ad + bc)/cd$ (see the following theorem).

We will not emphasise the idea of making deductions from the axioms, and for this reason I have marked the proofs of the assertions in the following theorem as ★ material. Nonetheless, you should have some appreciation of the ideas involved, and so you should work through a couple of the proofs.

THEOREM 1.2.2. *If a, b, c, d are real numbers and $c \neq 0$, $d \neq 0$ then*

(1) $ac = bc$ *implies* $a = b$.
(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) $(a/c) + (b/d) = (ad + bc)/cd$

PROOF. ★ Each line in the following proofs will be

(1) an example of one (or occasionally more) of axioms 1–9;
(2) a previously proved result;
(3) or follow from previously proved results by rules of logic[5] (which include the properties of equality).

Fill in any missing steps. Go through the proofs line by line and indicate what is used to justify 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. The proof is as follows:
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 $a0 = 0$.

3. We want to show $-(-a) = a$.
By $-(-a)$ we mean the negative of $-a$, and hence by Axiom 4 we know that[6]

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

But from Axiom 3

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

Hence

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

and so $-(-a) = a$ from the Cancellation Theorem.

4. Write out your own proof, along similar lines to the preceding proof. You should first prove a cancellation theorem for multiplication.

5. (As in the proof of **2.**) it is sufficient to show $(-1)a + a = 0$, because then

$$(-1)a + a = (-a) + a \quad \text{(additive inverse axiom)}$$

and so

$$(-1)a = -a$$

by the Cancellation Theorem.

---

[5]For 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.

[6]Since a can represent *any* number in Axiom 4, we can replace a in Axiom 4 by $-a$. This might seem strange at first, but it is quite legitimate.

The proof is as follows:

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

This completes the proof.

6.

$$\begin{aligned}a(-b) &= a((-1)b) \\ &= (a(-1))b \\ &= ((-1)a)b \\ &= (-1)(ab) \\ &= -(ab)\end{aligned}$$

Prove the second equality yourself.

7. Prove this yourself using, in particular, use **4** and Axiom 9.

8.

$$\begin{aligned}(-a)(-b) &= ((-1)a)(-b) \\ &= (-1)(a(-b)) \\ &= -(a(-b)) \\ &= -(-(ab)) \\ &= ab\end{aligned}$$

9. 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 Axioms 5 and 6).

If we can show that $c^{-1}d^{-1} = (cd)^{-1}$ then we are done.
Since, by Axiom 8, $(cd)^{-1}$ is the *unique* real number such that $(cd)(cd)^{-1} = 1$, it is sufficient to show[7] that $(cd)(c^{-1}d^{-1}) = 1$.
Do this; use A5–A8.

This completes the proof.

Important Remark: There is a tricky point in what we havew just done 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

[7]When we say "it is sufficient to show ..." we mean that if we can show ... then the result we want will follow.

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 $cd \neq 0$

We will establish the *claim* by proving that if $cd = 0$ then $c = 0$ or $d = 0$.[8]

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

$c^{-1}(cd) = c^{-1}0$ 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 $cd \neq 0$. This completes the proof of the claim.

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.

□

1.2.3. Order consequences. All the standard properties of inequalities for the real numbers follow from Axioms 1–13.

1.2.3.1. *More definitions.* One defines "$>$", "$\leq$" and "$\geq$" in terms of $<$ as follows:

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

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

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). (To prove that there is a *unique* such number a requires the order axioms.)

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

[8] *Note*; in mathematics, if we say P or Q (is true) then we *always* include the possibility that *both* P and Q are true.

1.2.3.2. *Some properties of inequalities.* The following are consequences of the axioms, although we will not stop to prove them.

THEOREM 1.2.3. *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* $0 < 1/b < 1/a$
(5) $|a + b| \leq |a| + |b|$ *(triangle inequality)*
(6) $\big|\, |a| - |b| \,\big| \leq |a - b|$ *(a consequence of the triangle inequality)*

1.2.4. Our approach henceforth. From now on, unless specifically noted or asked otherwise, we will use all the standard algebraic and order properties of the real numbers that you have used before. In particular, we will use any of the definitions and results in Adams Section P1.

1.2.5. Natural and Rational numbers. The set $\mathbb{N}$ of natural numbers is defined by

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

Here $2 = 1+1$, $3 = 2+1$, $\dots$. (Thus $\mathbb{N}$ is described by listing its members.)

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

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

(We read the right side after the equality as "the set of m/n such that m, n are members of $\mathbb{Z}$ and $n \neq 0$.)

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

The set $\mathbb{N}$ is not a model of Axiom 3, as 0 is not a member. The set $\mathbb{Z}$ is a model of all of Axioms 1–13, except for Axiom 8, since the multiplicative inverse a^{-1} of an integer is not usually an integer.

The set $\mathbb{Q}$ is a model of all of Axioms 1–13 but not of the Completeness Axiom (see later).

1.2.6. ★Fields.

The real numbers and the rationals, 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 Axioms 1–9, 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 Axiom 8 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
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\otimes$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

It is not too hard to convince yourself that the analogues of Axioms 1–9 hold for any prime p. The axiom which fails if p is not prime is Axiom 8, *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.

1.2.7. ★Groups. Any set S, together with an operation $\otimes$ and a particular member $e \in S$, which satisfies:

for all $a, b, c \in S$:

(1) $(a \otimes b) \otimes c = a \otimes (b \otimes c)$ (associative axiom)
(2) $a \otimes e = e \times a = a$ (identity axiom)
(3) there is a member of S, denoted a^{-1}, such that $a \otimes a^{-1} = a^{-1} \otimes a = e$ (inverse axiom)

Examples are the reals or rationals, with $\otimes$ replaced by $\times$ and e replaced by 1, or with $\otimes$ replaced by $+$ and e replaced by 0. Another example is $\mathbb{Z}$ with $\otimes$ and e replaced by $+$ and 0.

The notion of a group pervades much of mathematics and its applications. Important examples are groups of transformations and their application to classification of crystals.

As a simple case, consider a square.

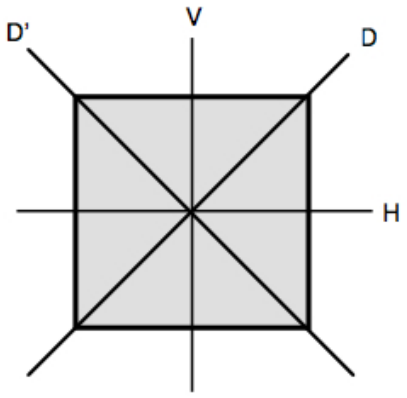

FIGURE 4. Reflection in any of the axes D, D', V, H maps the shaded square onto itself.

If the square is reflected in any of the four axes V (vertical axis), H (horizontal axis), D or D' (two diagonals), then the image coincides with the original square. We denote these four reflection transformations by V, H, D, D' respectively.

We could also rotate the square through 90^0, 180^0 or 270^0; these operations are denoted by R, R' and R'' respectively. Finally, we could rotate through 360^0, which has the same effect as doing nothing; this operation is called the *identity* operation and is denoted by I.

We say the square is *invariant* under the 8 transformations I, V, H, D, D', R, R' and R''.

If we first apply R and then apply H, the result is written as $H \otimes R$ (read from right to left, just as for the composition of functions in general). This is not a new operation — it has the same effect as applying D'. One way to see this is to check what happens to each of the four vertices of the square. For example, if we apply $H \otimes R$ the top right vertex is first mapped to the top left vertex and then to the bottom right vertex. This is the same as applying D'. Similarly for the other three vertices.

However, if we apply H and R in the reverse order, i.e. $R \otimes H$, then the result is different. This time it is the same as D. In particular, the operation $\otimes$ is not commutative.

If we draw up a table we obtain

$\otimes$	I	V	H	D	D′	R	R′	R″
I	I	V	H	D	D′	R	R′	R″
V	V	I	R′	R	R″	D	H	D′
H	H	R′	I	R″	R	D′	V	D
D	D	R″	R	I	R	H	D′	V
D′	D′	R	R″	R′	I	V	D	H
R	R	D′	D	V	H	R′	R″	I
R′	R′	H	V	D′	D	R″	I	R
R″	R″	D	D′	H	V	I	R	R′

A crystal is classified by the group of transformations which leaves it invariant. These ideas are treated in your later Algebra courses.

Other important examples of groups are certain sets of 2×2 matrices, where multiplication is matrix multiplication and the identity is the matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. In fact this is related to the previous example, the 8 operations correspond to certain 2×2 matrices (Can you find them? See *Linear Algebra* by Lay.). More generally, important examples are various groups of $n \times n$ matrices.

See "A Survey of Modern Algebra" by Birkhofff and MacLane for an introduction to fields and groups. (An old classic.)

1.3. Completeness Axiom

The Completeness Axiom is introduced. It is true for the real numbers, but the analogous axiom 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.

See Adams page 4 for a few brief remarks, page A22 in the Appendices, and an application to sequences on page A23.

The *Completeness Axiom* 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.

1.3.1. Statement of the Axiom.

Axiom 14 (Completeness Axiom): *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 1.3.1. 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 in addition $x \leq b$ for every upper bound b, then x is called the *least upper bound* or *supremum* of A. In this case one writes

$$x = \operatorname{lub} A \quad \text{or} \quad x = \sup A.$$

If $x \leq a$ for every $a \in A$, then x is called a *lower bound* for A. If also $x \geq c$ for every lower bound c, then x is called the *greatest lower bound* or *infimum* of A. In this case one write

$$x = \operatorname{glb} A \quad \text{or} \quad x = \inf A.$$

Thus we have:

Axiom 14 (Completeness Axiom): If a non-empty set A has an upper bound then it has a least upper bound.

(Remember that when we say A "has" an upper bound or a least upper bound x we do *not* require that $x \in A$.)

See Adams page A22, Example 1.

1.3.2. An equivalent formulation. 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 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.

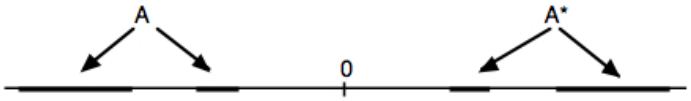

FIGURE 5. Reflecting A through the origin sends A to A^*, lower bounds of A to upper bounds of A^* and the *g.l.b.* of A to the *l.u.b.* of A^*.

Lower bounds for A correspond under reflection to upper bounds for A^*, and a glb corresponds to a lub. If A is bounded below then A^* is bounded above, and so by the Completeness Axiom has a lub. After reflection, this lub for A^* gives a glb for A. (To actually write this out carefully needs some care—you need to check from the relevant definitions and the properties of inequalities that the first three sentences in this paragraph are indeed correct.)

(Similarly, the Completeness Axiom follows from the above version.)

Unlike in the case of Axioms 1–13, we will always indicate when we use the Completeness Axiom.

1.3.3. Interpretation of the Completeness Axiom. 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 Axiom 14. In other words, Axiom 14 is not true if, in the first statement of the axiom, the three occurence of the word "real" are replaced by "rational".

For example, let

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

(The first definition for A has the advantage that A is defined without actually referring to the existence of the irrational number $\sqrt{2}$.) There are certainly rational numbers x which are upper bounds for A, i.e. such that $a \leq x$ for every $a \in A$, just take $x = 23$. But we claim *there is no rational number b which is a* least *upper bound for A.*

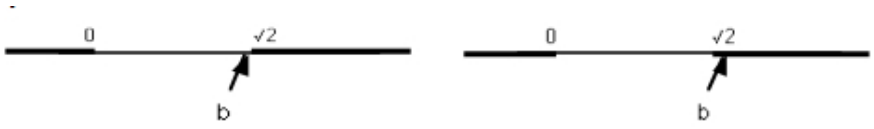

FIGURE 6. In each of these diagrams the set A is represented by the *thin* line. In the left diagram we see b is not an upper bound for A as there are members of A greater than b. In the right diagram we see b is not the *l.u.b.* for A as there are other upper bounds that are less than b.

PROOF OF CLAIM. Since $\sqrt{2}$ is not rational it cannot be the required rational number b.

On the other hand, if $b < \sqrt{2}$, since there is always a rational number between b and $\sqrt{2}$ this gives a member of A between b and $\sqrt{2}$, and so b cannot be an upper bound.

Finally, if $b > \sqrt{2}$, there is always a smaller rational number between $\sqrt{2}$ and b, and so b cannot be the *least* rational number which is an upper bound for A.

We have ruled out the three possibilities $b = \sqrt{2}$, $b < \sqrt{2}$ and $b > \sqrt{2}$. This completes the proof of the claim. Hence there is no rational number which is a least upper bound for A. □

1.3.4. Archimedean Property. The following property of the real numbers is not surprising, but it does not follow from the algebraic and order axioms alone. It says, informally, that there are no real numbers beyond all the natural numbers.

THEOREM 1.3.2 (Archimedean Property). *For every real number x there is a natural number n such that $x < n$. Equivalently, the set $\mathbb{N}$ is not bounded above.*

★ PROOF. Suppose that the theorem were false. Then there would be a real number x with the property that $n < x$ for all $n \in \mathbb{N}$. This implies $\mathbb{N}$ is bounded above and so there must be a *least* upper bound b (say) by the Completeness Axiom.

In other words,

$$n \leq b \text{ for } \textit{every } n \in \mathbb{N}.$$

It follows that

$$n + 1 \leq b \text{ for } \textit{every } n \in \mathbb{N},$$

since $n + 1 \in \mathbb{N}$ if $n \in \mathbb{N}$. But this implies

$$n \leq b - 1 \text{ for } \textit{every } n \in \mathbb{N}.$$

In other words, $b - 1$ is also an upper bound for $\mathbb{N}$, which contradicts the fact that b is the *least* upper bound.

Since we have obtained a contradiction by assuming the statment of the theorem is false, the statement must in fact be true. □

The only surprising thing about the Archimedean property is that it needs the Completeness Axiom to prove it. But there are in fact models of the algebraic and order axioms in which the Archimedean property is false. They are sometimes called the *Hyperreals*! See the next starred section.

The following corollary says that between zero and any positive number (no matter how small) there is always a number of the form $1/n$, where $n \in \mathbb{N}$. This is the same type of result as in Section 1.1.4, which stated that between any two different real numbers there is always a rational number. But in Section 1.1.4 we were not very careful, and did not go back to the axioms to prove the result (as we actually do need to!).

The symbol ε in the following is called "epsilon" and is a letter of the Greek alphabet. You could replace ε by any other symbol such as x or a, and the Corollary would have *exactly* the same meaning. However, it is traditional in mathematics to use ε when we are thinking of a very small

positive number. Sometimes we use the symbol δ or "delta", another letter of the Greek alphabet, in the same way.

COROLLARY 1.3.3. *For any real number $\varepsilon > 0$ there is a natural number n such that $\frac{1}{n} < \varepsilon$.*

PROOF. By the Archimedean Property there is a natural number n such that $n > \frac{1}{\varepsilon}$. But then $\frac{1}{n} < \varepsilon$ (by standard properties for manipulating inequalities). □

Note that this corollary was actually used in the proof of the results in Section 1.1.4, *where*?

It is probably confusing as to why the Archimedean property and its Corollary should rely on the Completeness Axiom for their proofs. What is the difference between the Archimedean Property and the previous Corollary on the one hand, and the other standard properties of inequalities such as in Theorem 1.2.3?

Well, the main difference is that the usual properties of inequalities, such as in Theorem 1.2.3, essentially tell us how to *manipulate* inequalities. The Archimedean Property is essentially a *non-existence* property concerning the infinite set $\mathbb{N}$ of real numbers, namely that $\mathbb{N}$ has no upper bound.

Don't worry! There will not be any more surprises like this. There will be important situations where we rely on the Completeness Axiom, such as in proving certain properties of continuous functions, but these applications will not be so surprising.

1.3.5. ★★Hyperreals. *Part* of any model of the hyperreals looks like a "fattened up" copy of $\mathbb{R}$, in the sense that it contains a copy of $\mathbb{R}$ together with "infinitesimals" squeezed between each real a and all reals greater than a. (In particular there are hyperreals bigger than 0 and less than any positive real number!) This part is followed and preceded by infinitely many "copies" of itself, and between any two copies there are infinitely many other copies. See the following crude diagram.

In particular there are hyperreals bigger than any natural number, as the natural numbers will all lie on the fattened up copy of $\mathbb{R}$.

The hyperreals were discovered by Abraham Robinson in the 1960's. They turn out to be quite useful in proving results about the usual real numbers and have been used in many areas of mathematics (for example in probability theory and stochastic processes). They are one way of giving a rigorous meaning to the notion of an infinitesimal — i.e. it is possible to interpret the expression $\frac{dy}{dx}$ as the ratio of two hyperreal numbers dy and dx.

There are even courses on Calculus that are based on using the hyperreals (also called "nonstandard" numbers). See "Elementary Calculus" by H. Jerome Keisler, and the corresponding Instructors Manual "Foundations of Infinitesimal Calculus". However, this approach to teaching Calculus has not been particularly popular!

A full analysis of the hyperreals and their properties requires a study of the underlying formal logic and set theory used in mathematics.

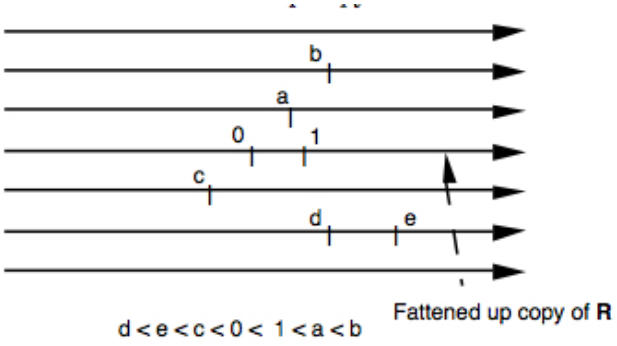

FIGURE 7. The line containing 0 and 1 is a "fattened up" copy of $\mathbb{R}$. It contains the infinitesimals, and for all each numbers a it contains the hyperreals $a + \epsilon$ where ϵ is an arbitrary infinitesimal. All points on each line are less than all points on any higher line.

You will not use the hyperreals. You certainly should not refer to them in any of your proofs.

1.4. Sets

The notion of a set is basic in mathematics. We will not need to study the theory of sets, but we will need to know some notation and a few basic properties.

1.4.1. Notation for sets. By a *set* (sometimes called a *class* or *family*) we mean a collection, often infinite, of (usually mathematical) objects.[9]

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 $\{1, 2, 3\}$, $\{2, 3, 1\}$, $\{1, 1, 1, 2, 3\}$ 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

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

We often use the notation

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

[9]★ 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.

where $P(x)$ is some statement involving x. We read this as "S is the set of all x such that $P(x)$ is true". It is often understood from the context of the discussion that x is restricted to be a real number. But if there is any possible ambiguity, then we write

$$S = \{\, x \in \mathbb{R} \mid P(x) \,\},$$

which we read as "S is the set of elements x in $\mathbb{R}$ such that $P(x)$ is true". Note that this is *exactly* the same set as

$$\{\, y \mid P(y) \,\} \quad \text{or equivalently} \quad \{\, y \in \mathbb{R} \mid P(y) \,\}.$$

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 "$\mid$" 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. Thus if A and B are sets, then

$$A \cap B = \{\, x \mid x \in A \text{ and } x \in B \,\},$$
$$A \cup B = \{\, x \mid x \in A \text{ or } x \in B \,\}.$$

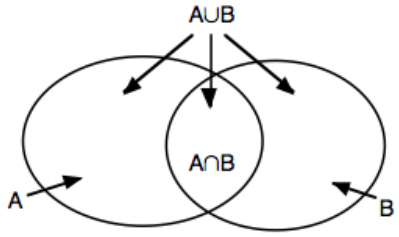

FIGURE 8. A Venn diagram representing the sets A, B, $A \cap B$ an $A \cup B$.

The set A is a *subset* of B, and we write $A \subseteq B$, if every element of A is also an element of B. In symbols

$$A \subseteq B \quad \text{iff} \quad x \in A \Rightarrow x \in B.$$

(Where $\Rightarrow$ is shorthand for "implies".)

For example

$$\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}.$$

It is also true that $\mathbb{N} \subseteq \mathbb{N}$, etc., although we would not normally make this statement.

In fact it is very common to write $A \subset B$ instead of $A \subseteq B$. The most common convention nowadays is that $A \subset A$ is a *true* statement. If you want to indicate that every element in A is also in B, and that there is at least one element in B that is not in A, then you should write $A \subsetneq B$.

Two sets are *equal* iff they have the same elements. It follows that

$$A = B \quad \text{iff} \quad A \subseteq B \text{ and } B \subseteq A.$$

In particular, we frequently prove two sets A and B are equal by first proving that every member of A is a member of B (i.e. $A \subseteq B$) and then proving that every member of B is a member of A (i.e. $B \subseteq A$).

For example, using the standard notation for intervals of real numbers in Section **??**,

$$\{\, x \mid 0 < x < 2 \text{ and } 1 \le x \le 3 \,\} = (0,2) \cap [1,3] = [1,2),$$
$$\{\, x \mid 0 < x < 1 \text{ or } 2 < x \le 3 \,\} = (0,1) \cup (2,3].$$

Also

$$(0,2) = (0,1) \cup [1,2) = (0,1] \cup [1,2) = (0,1) \cup (\tfrac{1}{2},2),$$

etc.

1.4.2. Ordered pairs of real numbers. In the Linear Algebra course you use both the notation $\begin{bmatrix} a \\ b \end{bmatrix}$ and (a,b) to represent vectors in $\mathbb{R}^2$, which we also regard as ordered pairs, or 2-tuples, of real numbers. Of course, (a,b) and (b,a) are distinct, unless $a = b$. This is different from the situation for the *set* containing a and b; i.e. $\{a,b\}$ and $\{b,a\}$ are just different ways of describing the same set.

We also have ordered triples (a,b,c), and more generally ordered n-tuples $(a_1,\ldots,a_n)$, of real numbers.

REMARK 1.4.1. ★ It is sometimes useful to know that we can define ordered pairs in terms of sets. The only property we require of ordered pairs is that

$$(a,b) = (c,d) \quad \text{iff} \quad (a = c \text{ and } b = d). \tag{3}$$

We could not define $(a,b) = \{a,b\}$, because we would not be able to distinguish between (a,b) and (b,a). But there are a number of ways that we can define ordered pairs in terms of sets. The standard definition is

$$(a,b) := \{\{a\},\{a,b\}\}.$$

To show this is a good definition, we need to prove (3).

PROOF. It is immediate from the definition that if $a = c$ and $b = d$ then $(a,b) = (c,d)$.

Next suppose $(a,b) = (c,d)$, i.e. $\{\{a\},\{a,b\}\} = \{\{c\},\{c,d\}\}$. We consider the two cases $a = b$ and $a \neq b$ separately.

If $a = b$ then $\{\{a\},\{a,b\}\}$ contains exactly one member, namely $\{a\}$, and so $\{\{c\},\{c,d\}\}$ also contains exactly the one member $\{a\}$. This means $\{a\} = \{c\} = \{c,d\}$. Hence $a = c$ and $c = d$. In conclusion, $a = b = c = d$.

If $a \neq b$ then $\{\{a\},\{a,b\}\}$ contains exactly two (distinct) members, namely $\{a\}$ and $\{a,b\}$. Since $\{\{a\},\{a,b\}\} = \{\{c\},\{c,d\}\}$ it follows $\{c\} \in \{\{a\},\{a,b\}\}$ and so $\{c\} = \{a\}$ or $\{c\} = \{a,b\}$. The second equality cannot be true since $\{a,b\}$ contains two members whereas $\{c\}$ contains one member, and so $\{c\} = \{a\}$, and so $c = a$.

Since also $\{c,d\} \in \{\{a\},\{a,b\}\}$ it now follows that $\{c,d\} = \{a,b\}$ (otherwise $\{c,d\} = \{a\}$, but since also $\{c\} = \{a\}$ this would imply $\{\{c\},\{c,d\}\}$ and hence $\{\{a\},\{a,b\}\}$ has only one member, and we have seen this is not

so). Since a and b are distinct and $\{c, d\} = \{a, b\}$, it follows c and d are distinct; since $a = c$ it then follows $b = d$. In conclusion, $a = c$ and $b = d$.

This completes the proof. □

REMARK 1.4.2. In Adams page 24 there is the definition: *a **function** f on a set D into a set S is a rule that assigns a* unique *element $f(x)$ in S to each element x in D.*

This is indeed the way to think of a function. But there is a problem. What do we mean by a "rule"? Is it something that can be ultimately written down by some algorithmic procedure, or in some sense can be programmed? This underestimates the collection of all possible functions.[10]

★ If we take the notion of a set as basic then we can define the notion of a function as follows: *A function f on a set D into a set S is a set f of ordered pairs of the form (x, y) satisfying the conditions*

(1) *if $(x, y) \in f$ then $x \in D$ and $y \in S$;*
(2) *for each $x \in D$ there is exactly one $y \in S$ such that $(x, y) \in f$ — this y is denoted by $f(x)$.*

This definition is useful because it shows that the notion of a function can be reduced to that of a set. In fact all of mathematics can be formulated within set theory. This is fundamental for the study of the foundations of mathematics, which is done later in a third year honours level course.

[10]★ In a manner that can be made precise, and will be in later courses, it is possible to define a notion of *countably infinite* and *uncountably infinite*. The integers, the rationals, and the set of possible rules, are all countably infinite. The reals, the irrationals and the set $[0, 1]$, are uncountably infinite. The set of functions from $\mathbb{R}$ into $\mathbb{R}$, or even from $[0, 1]$ into $[0, 1]$, are both uncountably infinite. In fact, these sets of functions can be shown to be, in a precise manner, a larger infinity than the infinity of real numbers.

CHAPTER 2

Sequences

The reference here is Adams Section 9.1, and the material on pages A-22 and A-23, but we do considerably more. Another reference is *Calculus* by M. Spivak.

2.1. Examples and Notation

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

$$1, 2, 1, 3, 1, 4, \ldots$$
$$1, \frac{1}{2}, \frac{1}{3}, \ldots$$
$$1, -1, 1, -1, 1, \ldots$$

A sequence can be written in the form

$$a_1, a_2, a_3, \ldots, a_n, \ldots.$$

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.

See Adams, page 496, Example 1.

See Adams pages 496 and 497 for the definitions of the following terms:

(1) bounded below, lower bound; bounded above, upper bound; bounded;
(2) positive, negative;
(3) increasing, decreasing, monotonic;
(4) alternating;
(5) ultimately (or "eventually").

See Examples 2 and 3 in Adams page 497.

2.2. Convergence of Sequences

The most fundamental concept in the study of sequences is the notion of *convergence* of a sequence.

The informal idea is that a sequence (a_n) converges to a, and we write

$$\lim a_n = a,$$

if no matter how small a positive number is chosen, the distance between a_n and a, i.e. $|a_n - a|$, will ultimately be less than this positive number. (The smaller the positive number, the further out in the sequence we will need to go.)

It is important to note that this is a condition that must be satisfied by *any* positive number. For example, we may have

$$|a_n - a| < .1 \text{ if } n > 50, \quad \text{(here the positive number is .1)}$$
$$|a_n - a| < .01 \text{ if } n > 300, \quad \text{(here the positive number is .01)}$$
$$|a_n - a| < .001 \text{ if } n > 780, \quad \text{(here the positive number is .001)}$$

etc.

DEFINITION 2.2.1. We say that *the sequence* (a_n) *converges to a limit* a, and write

$$\lim_n a_n = a, \quad \lim a_n = a, \quad a_n \to a, \quad \text{or} \quad \lim_{n\to\infty} a_n = a,$$

if for every positive number ε there exists an integer N such that

$$n > N \quad \text{implies} \quad |a_n - a| < \varepsilon. \tag{4}$$

See Adams page 498, Figure 9.1, for a graphical illustration of limit.

EXAMPLE 2.2.2. Show that the sequence given by $a_n = 1 + \frac{1}{n^2}$ converges to 1 according to the definition.

SOLUTION. Let $\varepsilon > 0$ be given.
We want to find N such that (4) is true with $a = 1$.
We have

$$|a_n - 1| = \frac{1}{n^2}.$$

Since

$$\frac{1}{n^2} < \varepsilon \quad \text{if} \quad n^2 > \frac{1}{\varepsilon},$$

i.e.

$$\text{if} \quad n > \frac{1}{\sqrt{\varepsilon}},$$

we can take

$$N = \left[\frac{1}{\sqrt{\varepsilon}}\right],$$

or any larger integer, where [ ] denotes "the integer part of". □

Thus if $\varepsilon = .1$ we can take any integer $N > 1/\sqrt{.1}$, for example $N = 4$ (or anything larger). If $\varepsilon = .01$ we can take $N = 10$ (or anything larger). If $\varepsilon = .001$ we can take $N = 32$ (or anything larger). But the above proof works of course for *any* $\varepsilon > 0$.[1]

---

[1]* In the example we took $N = \left[\frac{1}{\sqrt{\varepsilon}}\right]$, the integer part of $\left[\frac{1}{\sqrt{\varepsilon}}\right]$, or equivalently the smallest integer greater than $\frac{1}{\sqrt{\varepsilon}} - 1$.

The general statement

$$\forall z > 0 \; \exists N \in \mathbb{N} \; (N > z),$$

is just the Archimedean Property, which follows from the Completeness Axiom as we saw before. Thus we are usually using the Archimedean Property when we prove the existence of limits.

A similar remark applies to the following examples, but we will not usually explicitly state this fact.

EXAMPLE 2.2.3. Consider the sequence defined by $a_1 = 1$, and $a_{n+1} = \frac{1}{2}a_n + 2$ for $n \geq 1$.

It is easy to calculate that the first few terms are

$$1, 2.5, 3.25, 3.625, 3.8125, 3.90625, 3.953125, 3.9765625, \ldots.$$

It seems reasonable that the sequence is converging to 4. One way to prove this is as follows.

PROOF. Let $\varepsilon > 0$ be given.

We want to find N such that[2]

$$n > N \quad \Rightarrow \quad |a_n - 4| < \varepsilon. \tag{5}$$

We have a formula for a_{n+1} in terms of a_n, and we first use this to get a formula for $|a_{n+1} - 4|$ in terms of $|a_n - 4|$. Thus

$$|a_{n+1} - 4| = \left|\frac{1}{2}a_n + 2 - 4\right| = \left|\frac{1}{2}a_n - 2\right| = \left|\frac{1}{2}(a_n - 4)\right| = \frac{1}{2}|a_n - 4|.$$

Thus $|a_1 - 4| = 3$, $|a_2 - 4| = 3/2$, $|a_3 - 4| = 3/2^2$, $|a_4 - 4| = 3/2^3$, ... In general[3] $|a_n - 4| = 3/2^{n-1}$.

It follows that

$$|a_n - 4| < \varepsilon \quad \text{for those } n \text{ such that } \quad \frac{3}{2^{n-1}} < \varepsilon.$$

This last inequality is equivalent to $2^{n-1}/3 < 1/\varepsilon$, i.e. $2^{n-1} > 3/\varepsilon$, i.e. $(n-1)\ln 2 > \ln(3/\varepsilon)$, i.e. $n > 1 + \ln(3/\varepsilon)/\ln 2$.[4]

Hence (5) is true for

$$N = 1 + \left[\frac{\ln \frac{3}{\varepsilon}}{\ln 2}\right].$$

□

You may object that we used ln, the natural logarithm, in the previous example, but we have not yet shown how to define logarithms and establish their properties from the axioms. This is a valid criticism. But in order to have interesting examples, we will often do this sort of thing.

However, we will not do it when we are establishing the underlying theory. In particular, the development of the theory will not depend on the examples.

See Adams page 498 Example 4.

DEFINITION 2.2.4. If a sequence (a_n) does not converge, then we say that it *diverges*.

We say that the sequence (a_n) *diverges to* $+\infty$ (or to ∞),[5] and write

$$\lim_n a_n = \infty, \quad \lim a_n = \infty, \quad a_n \to \infty, \quad \text{or} \quad \lim_{n\to\infty} a_n = \infty,$$

[2]We will often write "⇒" for "implies".

[3]This could easily be proved by induction, but it is not necessary to do so.

[4]★ As in the previous example, to prove the existence of such an n requires, strictly speaking, the Archimedean Property, which is a consequence of the Completeness Axiom.

[5]Note that $+\infty$ is *not* a real number. In fact we do *not* here define an object denoted by "∞". We just define a certain concept which we denote by "$\lim a_n = \infty$" or by something similar.

if for each real number M there exists an integer N such that

$$n > N \quad \text{implies} \quad a_n > M.$$

(The interesting situation is M large and positive.)

Similarly, we say (a_n) to *diverges to* $-\infty$ if for each real number M there exists an integer N such that

$$n > N \quad \text{implies} \quad a_n < M.$$

(The interesting situation is M large and negative.)

A sequence may diverge, without diverging to $\pm\infty$. See Adams page 498 Example 5.

2.3. Properties of Sequence Limits

It is normally not very efficient to use the definition of a limit in order to prove that a sequence converges. Instead, we prove a number of theorems which will make things much easier.

The first theorem shows that if two sequences converge, then so does their sum, and moreover the limit of the new sequence is just the sum of the limits of the original sequences. Similar results are true for products and quotients, and if we multiply all terms in a sequence by the same real number.

In (9) we need to assume $b \neq 0$. This will imply that ultimately $b_n \neq 0$ (i.e. $b_n \neq 0$ for all sufficiently large n), and hence that the sequence (a_n/b_n) is defined for all sufficiently large n.

THEOREM 2.3.1. *Suppose*

$$\lim a_n = a, \quad \lim b_n = b,$$

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

$$\lim(a_n \pm b_n) = a \pm b, \tag{6}$$

$$\lim ca_n = ca, \tag{7}$$

$$\lim a_n b_n = ab, \tag{8}$$

$$\lim \frac{a_n}{b_n} = \frac{a}{b}, \qquad \textit{assuming } b \neq 0. \tag{9}$$

We will give the proofs in the next section. The theorem is more justification that Definition 2.2.1 does indeed capture the informal notion of a limit.

The results are not very surprising. For example, if a_n is getting close to a and b_n is getting close to b then we expect that $a_n + b_n$ is getting close to $a + b$.

EXAMPLE 2.3.2. Let $a_n = \left(1 + \frac{1}{\sqrt{n}}\right)^2 - (1 + 2^{-n})$.

We can prove directly from the definition of convergence that $\frac{1}{\sqrt{n}} \to 0$ and $2^{-n} \to 0$. It then follows from the previous theorem that $1 + \frac{1}{\sqrt{n}} \to 1$ (since we can think of $1 + \frac{1}{\sqrt{n}}$ as obtained by adding the term 1 from

the constant sequence (1) to the term $\frac{1}{\sqrt{n}}$). Applying the theorem again, $\left(1+\frac{1}{\sqrt{n}}\right)^2 \to 1$. Similarly, $1+2^{-n} \to 1$.

Hence (again from the theorem) $a_n \to 1-1=0$.

EXAMPLE 2.3.3. Let $a_n = \frac{2n^2-1}{3n^2-7n+1}$.

Write

$$\frac{2n^2-1}{3n^2-7n+1} = \frac{2-\frac{1}{n^2}}{3-\frac{7}{n}+\frac{1}{n^2}}.$$

Since the numerator and denominator converge to 2 and 3 respectively, it follows $a_n \to 2/3$.

See also Adams, pages 499, Example 6.

Before we prove Theorem 2.3.1 there is a technical point. We should prove that a convergent sequence cannot have two different limits. This is an easy consequence of the definition of convergence.

This is done in the next section, but try it yourself first of all.

THEOREM 2.3.4. *If (a_n) is a convergent sequence such that $a_n \to a$ and $a_n \to b$ then $a = b$.*

The next easy result is useful in a number of situations.

THEOREM 2.3.5. *Suppose $a_n \to a$. Then the sequence is bounded; i.e. there is a real number M such that $|a_n| \le M$ for all n.*

The next theorem is not true if we replace both occurrences of "$\le$" by "$<$". For example $-1/n < 1/n$ for all n, but the sequences $(1/n)$ and $(-1/n)$ have the same limit 0.

THEOREM 2.3.6. *Suppose $a_n \to a$, $b_n \to b$, and $a_n \le b_n$ ultimately. Then $a \le b$.*

The following theorem says that if a sequence is "squeezed" between two sequences which both converge to the same limit, then the original sequence also converges, and it converges to the same limit.

THEOREM 2.3.7 (Squeeze Theorem). *Suppose $a_n \le b_n \le c_n$ ultimately. Suppose $a_n \to L$ and $c_n \to L$. Then $b_n \to L$.*

EXAMPLE 2.3.8. Consider the sequence $3 + (\sin\cos n)/n$. Since $-1 \le \sin x \le 1$, it follows that $3 - 1/n \le 3 + (\sin\cos n)/n \le 3 + 1/n$. But $3 - 1/n \to 3$ and $3 + 1/n \to 3$. Hence $3 + (\sin\cos n)/n \to 3$.

2.4. Proofs of limit properties

I have starred some of these proofs, as they are a bit technical. But you should aim to have some understanding of the ideas involved.

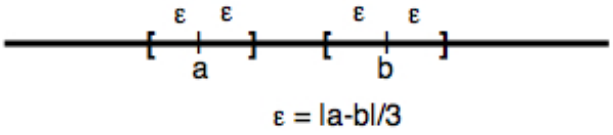

FIGURE 1.

★ PROOF OF THEOREM 2.3.4. Suppose $a_n \to a$ and $a_n \to b$.

Assume (in order to obtain a contradiction) that $a \neq b$.

Take $\varepsilon = |a-b|/3$ in the definition of a limit, Definition 2.2.1. (For motivation, look at the following diagram).

Since $a_n \to a$, it follows that

$$a_n \in (a-\varepsilon, a+\varepsilon) \tag{10}$$

for all sufficiently large n, say for $n > N_1$.

Since $a_n \to b$, it follows that

$$a_n \in (b-\varepsilon, b+\varepsilon) \tag{11}$$

for all sufficiently large n, say for $n > N_2$.

But this implies

$$a_n \in (a-\varepsilon, a+\varepsilon) \quad \text{and} \quad a_n \in (b-\varepsilon, b+\varepsilon)$$

for all $n > \max\{N_1, N_2\}$.

However this is impossible since $\varepsilon = |a-b|/3$, as we can see geometrically or show algebraically.

Thus the assumption $a \neq b$ leads to a contradiction, and so $a = b$. □

★ PROOF OF THEOREM 2.3.5. Assume $a_n \to a$.

From the definition of convergence, taking $\varepsilon = 1$, there is an integer N such that

$$a - 1 < a_n < a + 1 \quad \text{for all } n > N. \tag{12}$$

Fix this N. Since the set of terms

$$a_1, a_2, \ldots, a_N$$

is *finite*, it follows that there exist real numbers M_1 and M_2 such that

$$M_1 \leq a_n \leq M_2 \quad \text{for all } n \leq N. \tag{13}$$

(Just take $M_1 = \min\{a_1, a_2, \ldots, a_N\}$ and $M_2 = \max\{a_1, a_2, \ldots, a_N\}$.)

From (12) and (13),

$$M_1^* \leq a_n \leq M_2^* \quad \text{for } \textit{all } n,$$

where $M_1^* = \min\{a-1, M_1\}$, $M_2^* = \max\{a+1, M_2\}$.

Hence $|a_n| \leq M$ for all n where $M = \max\{|M_1^*|, |M_2^*|\}$. □

PROOF OF THEOREM 2.3.1 (6) ("+" CASE). Suppose $a_n \to a$ and $b_n \to b$.

Let $\varepsilon > 0$ be given.

Since $a_n \to a$ there exists an integer N_1 (by Definition 2.2.1) such that

$$n > N_1 \quad \text{implies} \quad |a_n - a| < \varepsilon/2. \tag{14}$$

Since $b_n \to b$ there exists an integer N_2 (again by Definition 2.2.1) such that

$$n > N_2 \quad \text{implies} \quad |b_n - b| < \varepsilon/2. \tag{15}$$

It follows that if $n > \max\{N_1, N_2\}$ then

$$\begin{aligned} |(a_n + b_n) - (a + b)| &= |(a_n - a) + (b_n - b)| \\ &< |a_n - a| + |b_n - b| \quad \text{by the triangle inequality} \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \quad \text{by (14) and (15)} \\ &= \varepsilon. \end{aligned}$$

It follows from Definition 2.2.1, with $N = \max\{N_1, N_2\}$, that $(a_n + b_n)$ converges and the limit is $a + b$. □

Notice in the proof how the definition of a limit is used three times; once to get information from the fact $a_n \to a$, once to get information from the fact $b_n \to b$, and finally to deduce that $a_n + b_n \to a + b$.

By the way, why do we use $\varepsilon/2$ in (14) and (15), and why is this justifiable by Definition 2.2.1?

PROOF OF THEOREM 2.3.1 (7). Suppose $a_n \to a$ and c is a real number.

Let $\varepsilon > 0$ be any positive number. We want to show

$$|ca_n - ca| < \varepsilon$$

for all sufficiently large n.

Since $a_n \to a$ there exists an integer N such that

$$|a_n - a| < \varepsilon/|c| \quad \text{for all} \quad n > N.$$

(This assumes $c \neq 0$. But if $c = 0$, then the sequence (ca_n) is the sequence all of whose terms are 0, and this sequence certainly converges to $ca = 0$.) Multiplying both sides of the inequality by $|c|$ we see

$$|c|\,|a_n - a| < \varepsilon \quad \text{for all} \quad n > N,$$

i.e.

$$|ca_n - ca| < \varepsilon \quad \text{for all} \quad n > N,$$

and so $ca_n \to ca$ by the definition of convergence. □

PROOF OF THEOREM 2.3.1 (6) ("−" CASE). Suppose $a_n \to a$ and $b_n \to b$.

We can write

$$a_n - b_n = a_n + (-1)b_n.$$

But $(-1)b_n \to (-1)b$ by the previous result with $c = -1$, and so the result now follows from (6) for the *sum* of two sequences. □

★ PROOF OF THEOREM 2.3.1 (8). Suppose $a_n \to a$ and $b_n \to b$.

As usual, let $\varepsilon > 0$ be an arbitrary positive number.

We want to show there is an integer N such that

$$|a_n b_n - ab| < \varepsilon$$

for all $n > N$.

To see how to choose N, write

$$\begin{aligned} |a_n b_n - ab| &= |a_n b_n - a_n b + a_n b - ab| \\ &= |a_n(b_n - b) + b(a_n - a)| \\ &\le |a_n(b_n - b)| + |b(a_n - a)| \\ &= |a_n|\,|b_n - b| + |b|\,|a_n - a|. \end{aligned} \tag{16}$$

(This trick of adding and subtracting the same term, here it is $a_n b$, is often very useful.) We will show that both terms are $< \varepsilon/2$ for all sufficiently large n.

For the second term $|b|\,|a_n - a|$, the result is certainly true if $b = 0$, since the term is then 0. If $b \neq 0$, since $a_n \to a$, we can choose N_1 such that

$$|a_n - a| < \frac{\varepsilon}{2|b|} \quad \text{for all} \quad n > N_1,$$

and so

$$|b|\,|a_n - a| < \frac{\varepsilon}{2} \quad \text{for all} \quad n > N_1. \tag{17}$$

For the first term $|a_n|\,|b_n - b|$, we use Theorem 2.3.5 to deduce for some M that $|a_n| \le M$ for all n. By the same argument as for the second term, we can choose N_2 such that

$$M\,|b_n - b| < \frac{\varepsilon}{2} \quad \text{for all} \quad n > N_2,$$

and so

$$|a_n|\,|b_n - b| < \frac{\varepsilon}{2} \quad \text{for all} \quad n > N_2. \tag{18}$$

Putting (16), (17) and (18) together, it follows that if $n > N$, where $N = \max\{N_1, N_2\}$, then

$$|a_n b_n - ab| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this proves $a_n b_n \to ab$. □

★ PROOF OF THEOREM 2.3.1 (9). Suppose $a_n \to a$ and $b_n \to b$ where $b \neq 0$.

We will prove that $a_n/b_n \to a/b$ by first showing $1/b_n \to 1/b$ and then using the previous result about products of sequences.

We first prove

$$|b_n| > |b|/2 \text{ ultimately.} \tag{19}$$

The proof is similar to that in Theorem 2.3.5, and goes as follows:

First assume $b > 0$. Choose $\varepsilon = |b|/2$ (> 0) in the definition of convergence and deduce that for some integer N,

$$n > N \Rightarrow |b_n - b| < b/2,$$

and so in particular

$$n > N \Rightarrow |b_n| > b/2.$$

This proves (19) in case $b > 0$.

In case $b < 0$ we similarly prove that ultimately $b_n < b/2$, and so ultimately $|b_n| > |b|/2$. Ths completes the proof of (19).

We now proceed with the proof that $1/b_n \to 1/b$. For this let $\varepsilon > 0$ be any positive number.

In order to see how to choose N in the definition of convergence, we compute

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| = \frac{|b - b_n|}{|b_n|\,|b|} \leq \frac{2|b - b_n|}{|b|^2}, \tag{20}$$

ultimately (this uses (19)). (The only reason for "$\leq$" instead of "$<$" is that perhaps $|b - b_n| = 0$.)

Since $b_n \to b$ we can find an integer N such that for all $n > N$,

$$|b - b_n| < \frac{|b|^2}{2}\varepsilon.$$

It follows from (20) that if $n > N$ then

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, it follows that $1/b_n \to 1/b$.

Since $a_n \to a$, it now follows from the result for products that $a_n/b_n \to a/b$. □

PROOF OF THEOREM 2.3.6. (The proof is similar to that for Theorem 2.3.4.)

Suppose $a_n \to a$, $b_n \to b$, and ultimately $a_n \leq b_n$.

Assume (in order to obtain a contradiction) that $a > b$. Let $\varepsilon = \frac{1}{3}(a-b)$.

Then

$$a_n \in (a - \varepsilon, a + \varepsilon) \text{ ultimately,}$$

and in particular

$$a_n > a - \varepsilon \text{ ultimately.}$$

Similarly

$$b_n < b + \varepsilon \text{ ultimately.}$$

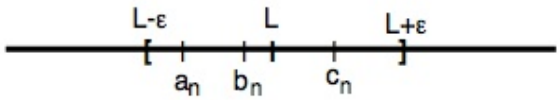

FIGURE 2.

(Draw a diagram.) Since $\varepsilon = \frac{1}{3}(a-b)$, this implies

$$a_n > b_n \text{ ultimately.}$$

But this contradicts $a_n \leq b_n$, and so the assumption is false. Thus $a \leq b$. □

PROOF OF THEOREM 2.3.7. Suppose $a_n \leq b_n \leq c_n$ ultimately. Suppose $a_n \to L$ and $c_n \to L$.

Let $\varepsilon > 0$ be given. (For motivation, look at the following diagram).

Since $a_n \to L$ there is some integer N_1 such that

$$n > N_1 \Rightarrow a_n \in (L-\varepsilon, L+\varepsilon). \tag{21}$$

Since $c_n \to L$ there is some integer N_2 such that

$$n > N_2 \Rightarrow c_n \in (L-\varepsilon, L+\varepsilon). \tag{22}$$

Let $N = \max\{N_1, N_2\}$. Then since $a_n \leq b_n \leq c_n$ it follows from (21) and (22) that

$$n > N \Rightarrow b_n \in (L-\varepsilon, L+\varepsilon).$$

But ε was an arbitrary positive number, and so it follows that $b_n \to L$. □

2.5. More results on sequences

We have seen that a convergent sequence is bounded.

The converse is false. For example, the sequence

$$1, -1, 1, -1, \ldots$$

is bounded but does not converge.

However, a bounded monotone sequence *does* converge. The proof needs the Completeness Axiom in order to give a "candidate" for the limit. (See also Adams, Appendix III page A-23, Theorem 2.)

THEOREM 2.5.1. *If a sequence is bounded and ultimately monotone (i.e. either ultimately increasing or ultimately decreasing), then it converges.*

PROOF. We do the "ultimately decreasing case", the other is similar.

Suppose the sequence decreases from the Nth term onwards. Considering just those terms from this point, and changing notation, we may assume $a_1 \geq a_2 \geq a_3 \geq \ldots a_n \geq \cdots$.

This set of terms is bounded below, and so by the Completeness Axiom this set has a *glb* which we denote by L. We will prove that $\lim a_n = L$.

Suppose $\epsilon > 0$ is an arbitrary positive number.

Because L is a lower bound, $L \leq a_n$ for all n.

Because L is the *greatest* lower bound, $L+\epsilon$ is *not* a lower bound, and so there is a k (possibly depending on ϵ)[6] such that $a_k < L+\epsilon$.[7]

By the decreasing property of the sequence, it follows $a_n < L+\epsilon$ for all $n \geq k$.

We have shown that $L \leq a_n < L+\epsilon$ for all $n \geq k$, where k may depend on ϵ. Since $\epsilon > 0$ was arbitrary, it follows from the definition of convergence that $a_n \to L$ as $n \to \infty$. □

EXAMPLE 2.5.2. Prove the sequence (a_n) defined by

$$a_1 = 1, \quad a_{n+1} = \sqrt{6+a_n}$$

is convergent, and find the limit.

SOLUTION. (See Adams Example 8, page 501, for details).

The idea is to show by induction that

(1) (a_n) is monotone increasing,
(2) $a_n \leq 3$.

It follows from the previous theorem that $a_n \to a$, say.

In order to find a, we use the facts that if $a_n \to a$ then $a_{n+1} \to a$ [8] and $\sqrt{6+a_n} \to \sqrt{6+a}$ [9]. By uniqueness of the limit of a sequence, $a = \sqrt{6+a}$. Solving gives $a = -2$ or 3. The former is impossible, as it it easy to see $a_n \geq 1$ for every n. □

The following limits are often useful.

THEOREM 2.5.3.

(1) *If* $|x| < 1$ *then* $\lim x^n = 0$.
(2) *If* x *is any real number, then* $\lim \frac{x^n}{n!} = 0$.

PROOF. (Adams gives a proof (see Theorem 3 section 9.1 page 501) which uses continuity and properties of ln for the first part — here is another proof that does not use this.)

Since $|x| < 1$ the sequence $|x|^n$ is decreasing[10] and all terms are ≥ 0. Hence $|x|^n \to a$ (say) by Theorem 2.5.1.

Since $|x|^n \to a$, also $|x|^{n+1} \to a$ (*Why?*). But $|x|^{n+1} = |x|\,|x|^n \to |x|\,a$. Hence $a = |x|a$ by uniqueness of limits, and so $a = 0$ as $|x| \neq 1$.

---

[6] The statement "possibly depending on ϵ" is redundant. We include it here for emphasis, but normally would not include it. Whenever we introduce a constant ϵ and then say there exists a k such that "blah blah involving k and ϵ is true" we *always* mean that k may, and indeed it almost always does, depend on ϵ.

[7] In fact there are infinitely many such k, as follows from the decreasing property which we next use. But at this point we are just using the properties of greatest lower bound, and so just get the existence of one k.

[8] It follows easily from the definition of convergence that if $a_n \to a$, then also $a_{n+1} \to a$ (*Exercise*). This is frequently a useful fact.

[9] This can either be proved from the definition of a limit (*Exercise*). Later it will follow easily from the fact that the function f given by $f(x) = \sqrt{6+x}$ is continuous.

[10] We could prove this by induction, but that is not really required at this level as it is routine and assumed you can give a proof if asked.

Because $-|x|^n \leq x^n \leq |x|^n$ and since both $|x|^n \to 0$ and $-|x|^n \to 0$, it follows from the Squeeze Theorem that $x^n \to 0$.

The second result follows from the first, see Adams. □

EXAMPLE 2.5.4. Find $\lim \frac{3^n+4^n+5^n}{5^n}$.

SOLUTION. Example 10 page 502 of Adams. □

The following theorem is easy and useful.

THEOREM 2.5.5. *Suppose S is a set of real numbers which has lub (glb) equal to L. Then there is a sequence (x_n) from S such that $\lim_n x_n = L$.*

PROOF. We do the *glb* case.

If $L \in S$ then we can take the constant sequence $L, L, L, \ldots$. In general, for each positive integer n, since L is a lower bound and $L + 1/n$ is not a lower bound, there is an element $x_n \in S$ such that $L \leq x_n < L + 1/n$. It follows that $x_n \to L$. □

REMARK 2.5.6. It is easy to obtain a sequence (x_n) as in the theorem, such that (x_n) is decreasing in the *glb* case and increasing in the *lub* case. *Exercise.*

2.6. Bolzano Weierstrass Theorem

DEFINITION 2.6.1. A *subsequence* of the sequence $a_1, a_2, a_3, \ldots, a_n, \ldots$ is a sequence $a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_n}, \ldots$, where $i_1 < i_2 < i_3 < \cdots < i_n < \cdots$.

Subsequences of $1, 1/2, 1/3, \ldots, 1/n, \ldots$ are the sequence $1, 1/3, 1/5, \ldots$, the original sequence itself, and infinitely many other possibilities. Subsequences of the sequence $1, -1, 1, -1, \ldots$ are the sequence $1, 1, 1, \ldots$, the sequence $-1, -1, -1, \ldots$ and infinitely many other possibilites.

If a sequence does not converge, it may or may not be the case that some subsequence converges.

For example, the sequence $1, 2, 3, 4, \ldots$, does not converge, nor does any subsequence.[11]

The sequence $1, -1, 1, -1, \ldots$ does not converge, but the subsequence $1, 1, 1, \ldots$ does converge, as does any subsequence of the original sequence for which all terms are eventually equal to 1. Similarly, the subsequence $-1, -1, -1, \ldots$ converges as does any subsequence of the original sequence for which all terms are eventually equal to -1.

The following theorem is needed to prove Theorem 4.5.2 on uniform continuity, and this in turn is used to prove Theorem 6.2.10 showing that a continuous function on a closed bounded interval is Riemann integrable. These results are all very important.

[11] This is clear. More precisely, if $(a_{i_n}) = a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_n}, \ldots$ is a subsequence then $a_{i_n} \geq n$ as we could show by induction. This rules out $\lim_{n\to\infty} a_{i_n} = a$ for any real number a, *why?*.

THEOREM 2.6.2 (Bolzano Weierstrass Theorem). *Let (c_n) be a sequence of real numbers all of which are contained in the closed bounded interval $[a, b]$.*[12]

Then some subsequence converges, and the limit also belongs to $[a, b]$.

REMARK 2.6.3. The conclusion of the theorem is not true for the non-bounded interval $[1, \infty)$ or for the nonclosed interval $[0, 1)$. In the first case consider the sequence $1, 2, 3, \ldots, n, \ldots$ and in the second consider the sequence $1 - \frac{1}{2}, 1 - \frac{1}{3}, \ldots, 1 - \frac{1}{n}, \ldots$.

Where does the following proof break down in each of these two cases?

REMARK 2.6.4. The idea behind the following proof is straightforward. Divide the interval $[a, b]$ into halves, keeping one subinterval which contains an (infinite) subsequence of (c_n). Keep subdividing the subintervals into halves, so that each new subinterval contains an (infinite) subsequence of the previous (infinite) subsequence. Then define a new subsequence (x_n) of the original sequence in such a way that x_n is in the nth subinterval.

Checking that this all works is a little tedious, but not difficult.

PROOF OF THEOREM. Divide the interval $[a, b]$ into two closed bounded intervals $[a, (a + b)/2]$ and $[(a + b)/2, b]$ each of equal length and with the common endpoint $(a + b)/2$. At least one of these two subintervals contains an (*infinite*)[13] subsequence of the original sequence (c_n). Choose one such subinterval and denote it by $[a_1, b_1]$.

Similarly subdivide $[a_1, b_1]$ and chose a subinterval which contains an (infinite) subsequence of the infinite subsequence in $[a_1, b_1]$. Denote this interval by $[a_2, b_2]$.

Similarly subdivide $[a_2, b_2]$ to obtain $[a_3, b_3]$ which contains a subsequence of the subsequence of the original sequence. Etc., etc.

Now define a convergent subsequence (x_n) from the original sequence (c_n) as follows. First choose x_1 to be any element from the (infinite) subsequence corresponding to $[a_1, b_1]$. Next choose some x_2 from the subsequence corresponding to $[a_2, b_2]$ which occurs in the sequence (c_n) after x_1. (*Why is this possible?*) Next choose some x_3 from the subsequence corresponding to $[a_3, b_3]$ which occurs in the sequence (c_n) after x_2. (*Why is this possible?*) Etc., etc.

We now have

$$a_1 \leq b_1, \quad a_1 \leq a_2 \leq b_2 \leq b_1, \quad a_1 \leq a_2 \leq a_3 \leq b_3 \leq b_2 \leq b_1, \quad \ldots \quad ,$$
$$x_1 \in [a_1, b_1], \quad x_2 \in [a_2, b_2], \quad x_3 \in [a_3, b_3], \quad \ldots \quad ,$$

[12] A *bounded* interval is one for which there is both an upper and lower bound, not necessarily in the interval. In particular, $(0, 1)$, $(0, 1]$ and $[0, 1]$ are all bounded. However, $(-\infty, 0]$, and $[1, \infty)$ are not bounded.

A *closed* interval is one which contains all of its "finite" endpoints. Thus $[0, \infty)$ and $[0, 1]$ are both closed. The only *closed bounded* intervals are those of the form $[a, b]$, where a and b are both real numbers (and $a < b$ in cases of interest!).

[13] We use the word "infinite" only for emphasis. By our definition, any sequence is infinite in the sense it contains an infinite number of terms. Of course, some or even all of the terms may be equal. Consider the sequence $1, 1, 1, \ldots, 1, \ldots$.

and (x_n) is a subsequence of (c_n).

Since the sequence (a_n) is increasing and bounded above it has a *lub* L by Theorem 2.5.1. Similarly, (b_n) has a *glb* M. Since each a_n is a lower bound for *every* b_m (*why?*)[14], it follows that $a_n \leq M$ for all n. Hence $L \leq M$ (*why?*).

Since $a_n \leq L \leq M \leq b_n$ for every n, and since $b_n - a_n = 2^{-n}(b-a) \to 0$, it follows that $L = M$.

It follows that $\lim x_n = L(= M)$, *why?* (And $\lim a_n = \lim b_n = L$, *why?*) □

2.7. ★Cauchy Sequences

You might think that if a sequence $(a_n)_{n\geq 1}$ satisfies $\lim_n(a_n - a_{n+1}) = 0$ then the sequence converges (to a finite limit). This is false. Consider the sequence $(a_n)_{n\geq 1}$ given by

$$0, \tfrac{1}{2}, 1, \tfrac{2}{3}, \tfrac{1}{3}, 0, \tfrac{1}{4}, \tfrac{2}{4}, \tfrac{3}{4}, 1, \tfrac{4}{5}, \tfrac{3}{5}, \tfrac{2}{5}, \tfrac{1}{5}, 0, \tfrac{1}{6}, \tfrac{2}{6}, \tfrac{3}{6}, \tfrac{4}{6}, \tfrac{5}{6}, 1, \tfrac{6}{7}, \dots .$$

The sequence does not converge although $a_{n+1} - a_n \to 0$.

Another example is given by $a_n = \sqrt{n}$. Clearly $\sqrt{n} \to \infty$ as $n \to \infty$. However,

$$\sqrt{n+1} - \sqrt{n} = \frac{(n+1) - n}{\sqrt{n+1} + \sqrt{n}} = \frac{1}{\sqrt{n+1} + \sqrt{n}} \to 0$$

as $n \to \infty$. (To obtain the first equality we "rationalised the numerator" by multiplying by $(\sqrt{n+1} + \sqrt{n})/(\sqrt{n+1} + \sqrt{n})$.)

However, there is a stronger requirement than $\lim_n(a_n - a_{n+1}) = 0$ which *does* imply a sequence $(a_n)_{n\geq 1}$ converges, see Theorem 2.7.2. First we need a definition.

DEFINITION 2.7.1. A sequence $(a_n)_{n\geq 1}$ is a *Cauchy sequence* if for every $\epsilon > 0$ there exists an integer N such that

$$n, m > N \Longrightarrow |a_n - a_m| < \epsilon.$$

We write $\lim_{m,n}(a_n - a_m) = 0$, or write $a_n - a_m \to 0$ *as* $m, n \to \infty$.

The difference between $a_n - a_{n+1} \to 0$ and $a_n - a_m \to 0$ is that the first requires the distance between every two consecutive members beyond a certain point should be $< \epsilon$, whereas the second requires the distance between *every* two members (*not* necessarily consecutive) beyond a certain point should be $< \epsilon$.

THEOREM 2.7.2. *A sequence $(a_n)_{n\geq 1}$ converges if and only if it is Cauchy.*

PROOF. First assume (a_n) converges and so $a_n \to a$ for some a.

Suppose $\epsilon > 0$. Then there exists an integer N such that

$$n > N \Longrightarrow |a_n - a| < \epsilon/2.$$

Hence if $m, n > N$ then

$$|a_n - a_m| = |a_n - a + a - a_m| \leq |a_n - a| + |a - a_m| < \epsilon/2 + \epsilon/2 = \epsilon.$$

[14]Think first about a_1. Then try a_2. Then a_3.

Hence $(a_n)_{n\geq 1}$ is Cauchy.

Next assume (a_n) is Cauchy.

The first problem is to identify an a to which the sequence might converge. For this note that the sequence (a_n) is bounded. To see this apply the Cauchy definition with $\epsilon = 1$. It follows there exists an integer N such that

$$m, n > N \Longrightarrow |a_m - a_n| < 1.$$

Hence for any n,

$$|a_n| \leq \max\{|a_1|, \ldots, |a_N|, |a_{N+1}| + 1\}.$$

Why? This shows the sequence is bounded.

By the Bolzano Weierstrass Theorem 2.6.2 there is a subsequence (a_{i_n}) such that $a_{i_n} \to a$. We *claim* that for the full sequence we similarly have $a_n \to a$.

To see this suppose $\epsilon > 0$.

Since (a_n) is Cauchy there is an integer N such that

$$m, n > N \Longrightarrow |a_n - a_m| < \epsilon/2.$$

Since $a_{i_n} \to a$ there is an integer k such that

$$i_k > N \text{ and } |a_{i_k} - a| < \epsilon/2.$$

Using both the previous inequalities it follows that

$$n > N \Longrightarrow |a_n - a| \leq |a_n - a_{i_k}| + |a_{i_k} - a| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Since $\epsilon > 0$ was arbitrary, it follows that $a_n \to a$. □

REMARK 2.7.3. The idea behind the second part of the above proof, from "By the Bolzano Weierstrass ..." onwards, is as follows:

(1) if we go out far enough in the original sequence (a_n) then every two elements of the sequence are within $\epsilon/2$ of one another;
(2) if we go out far enough in the subsequence then every element is within $\epsilon/2$ of a.

Putting these two facts together, if we go out far enough in the original sequence then every element is within ϵ of a.

The following gives a condition on consecutive members of a sequence which implies the sequence is Cauchy, and hence converges.

THEOREM 2.7.4. *Suppose a sequence $(a_n)_{n\geq 1}$ satisfies*

$$|a_n - a_{n+1}| \leq Kr^n \quad \textit{for all } n,$$

where K is a positive real and $0 \leq r < 1$. Then the sequence is Cauchy and hence converges.

PROOF. Suppose $m > n$. Then

$$\begin{aligned}|a_n - a_m| &\leq |a_n - a_{n+1}| + |a_{n+1} - a_{n+2}| + \cdots + |a_{m-1} - a_m| \\ &\leq K(r^n + r^{n+1} + \cdots + r^{m-1}) \\ &\leq K(r^n + r^{n+1} + \cdots + r^{m-1} + \dots) \\ &= K\frac{r^n}{1-r}.\end{aligned}$$

Hence $|a_n - a_m| \to 0$ as $m, n \to \infty$, and so the sequence is Cauchy. □

Note that if a is the limit of the sequence (a_n) in the previous theorem then $|a_n - a| \leq Ar^n$ where $A = K/(1-r)$. *Why?*

CHAPTER 3

Continuous Functions

The intuitive idea of a continuous function is one whose graph can be drawn without lifting pen from paper. But this is too vague to develop a useful theory.

3.1. The Approach in these Notes

In this Chapter we first use sequences, and properties of their limits, in order to define continuity and establish its properties. This approach is more intuitive and easier to follow for most students that what is called the $\epsilon - \delta$ approach used in Adams,[1] see also Section 4.4 here. The sequence approach also enables us to proceed quickly to the proof of the major deep theorems about continuous functions in Section 3.4.

In the next Chapter we first treat limits in a similar manner to continuity, by again using sequences. In fact, the properties of continuity follow easily from the properties of limits, as we see in the next chapter, but the reason for treating continuity first is to make this Chapter brief and self-contained.

In the next Chapter we also show that the definitions of continuity and of limit via sequences are equivalent to the analogous $\epsilon-\delta$ definitions.[2] Thus we get to the same point in the theory by either route. Understanding both approaches will give you much more insight into the subject.

There is one other small difference between the approach here and that in Adams. In the latter the domain of a function is always an interval or the union of a finite number of intervals. This is the example case you should always think about, but here we consider more general domains.

3.2. Definition via Sequences and Examples

Recall[3] that the domain of a function f, denoted by $\mathcal{D}(f)$, is the set of numbers x such that $f(x)$ is defined. We will usually be interested in functions whose domains are intervals $[a, b]$, (a, b), (a, ∞)[4], etc. These are the cases you should think of when you study the material in these notes, unless it is indicated otherwise.

---

[1] See Adams Chapter 1 and Appendix III.

[2] In Adams Appendix III page A-20 the fact that the $\epsilon - \delta$ definitions of limit and continuity imply the sequence definitions is proved. However, the converse direction, that the sequence definitions implies the $\epsilon - \delta$ definitions is not proved.

[3] See Adams page 24.

[4] Note that ∞ is *not* a number, and that for us the symbol ∞ has no meaning by itself. The interval (a, ∞) is just the set of real numbers strictly greater than a.

But it is possible for the domain to be a more complicated set of real numbers. In fact, a sequence is just a function whose domain is the set $\mathbb{N}$ of natural numbers.

We will define the notion of continuity of a function in terms of convergence of sequences. The informal idea of "continuity of a function f at a point c" is that "as x approaches c then $f(x)$ approaches $f(c)$".

More precisely, we have the following natural definition.

DEFINITION 3.2.1. A function f *is continuous at a point* $c \in \mathcal{D}(f)$ if for every sequence $(x_n)_{n\geq 1}$ from $\mathcal{D}(f)$ such that $x_n \to c$, we have $f(x_n) \to f(c)$.

The function f is *continuous* (on its domain) if f is continuous at every point in its domain.

The first paragraph in the definition can be slightly rewritten as follows: f *is continuous at* $c \in \mathcal{D}(f)$ if:

$$x_n \in \mathcal{D}(f) \text{ and } x_n \to c \quad \Longrightarrow \quad f(x_n) \to f(c).$$

We will often not specifically write $x_n \in \mathcal{D}(f)$, although this is always understood in order that $f(x_n)$ be defined. Note that $\big(f(x_n)\big)_{n\geq 1}$ is also a sequence of real numbers.

In order to show f *is* continuous at c, we have to show that for *every* sequence $(x_n)_{n\geq 1}$ (from $\mathcal{D}(f)$) such that $x_n \to c$, one has $f(x_n) \to f(c)$.

In order to show f *is not* continuous at c, we only have to show there is *one* ("bad") sequence (from $\mathcal{D}(f)$) such that $x_n \to c$ and $f(x_n) \not\to f(c)$.[5]

EXAMPLE 3.2.2. Suppose

$$f(x) = \begin{cases} x & 0 \leq x < 1 \\ \frac{1}{2}x^2 & 1 \leq x \leq \frac{3}{2} \end{cases}$$

The domain of f is $[0, \frac{3}{2}]$. The following is an attempt to sketch the graph of f.

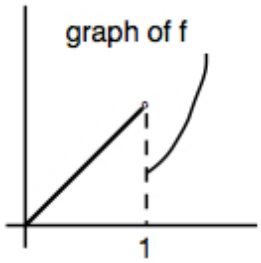

FIGURE 1. Sketch of the graph of f.

It is clear that f is not continuous at 1. For example, take the sequence $x_n = 1 - \frac{1}{n}$. Then $x_n \to 1$ but $f(x_n)$ $(= 1 - \frac{1}{n}) \not\to f(1)$ since $f(1) = \frac{1}{2}$.

[5]If there is one, there will in fact be many such "bad" sequences — we can always change the first million or so terms — but the point of logic is that to show continuity fails it is sufficient to just prove there is one "bad" sequence.

On the other hand, if $c \neq 1$ and $c \in \mathcal{D}(f)$ then

$$x_n \to c \quad \Rightarrow \quad f(x_n) \to f(c).$$

To see this, first suppose $x_n \to c$ and $1 < c \leq \frac{3}{2}$. Then $x_n \geq 1$ for all sufficiently large n, and so $f(x_n) = \frac{1}{2}x_n^2$ for all sufficiently large n. From properties of sequences if $x_n \to c$ then $x_n^2 \to c^2$ and so $\frac{1}{2}x_n^2 \to \frac{1}{2}c^2$. But $f(x_n) = \frac{1}{2}x_n^2$ for all sufficiently large n, and so $\lim f(x_n) = \lim \frac{1}{2}x_n^2 = \frac{1}{2}c^2$.

The case $0 \leq c < 1$ is similar, and easier.

In particular, f is not continuous on its domain since it fails to be continuous at $c = 1$.

If we vary this example a little, and define

$$g(x) = \begin{cases} x & 0 \leq x < 1 \\ \frac{1}{2}x^2 & 1 < x \leq \frac{3}{2}, \end{cases}$$

then the domain of g is $[0, 1) \cup (1, \frac{3}{2}]$. The function g is continuous at each $c \in \mathcal{D}(g)$, and so *is* continuous on its domain.

However, there is no *extension* of g to a continuous function defined on all of $[0, 3/2]$.

EXAMPLE 3.2.3. *The absolute value function f (given by $f(x) = |x|$) is continuous.*

We first show continuity at 0. For this, suppose $x_n \to 0$. Then $|x_n| \to 0$ (this is immediate from the definition of convergence, since $|x_n - 0| \leq \epsilon$ iff $| \, |x_n| - 0 \, | \leq \epsilon$), i.e. $f(x_n) \to f(0)$.

To prove continuity at $c \neq 0$ is similar to the previous example.

The following result is established directly from the properties of convergent sequences.

PROPOSITION 3.2.4. *Every polynomial function is continuous.*

PROOF. Let

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k.$$

To show that f is continuous at some point c, suppose $x_n \to c$.

Then $x_n^2 \to c^2$, $x_n^3 \to c^3$, etc., by the theorem about products of convergent sequences. It follows that $a_1 x_n \to a_1 c$, $a_2 x_n^2 \to a_2 c^2$, $a_3 x_n^3 \to a_3 c^3$, etc., by the theorem about multiplying a convergent sequence by a constant. Finally,

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k \to a_0 + a_1 c + a_2 c^2 + \cdots + a_k c^k$$

by repeated applications of the theorem about sums of convergent sequences (a_0 is here regarded as a constant sequence). □

EXAMPLE 3.2.5.★ Here is a surprising example.

Let

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

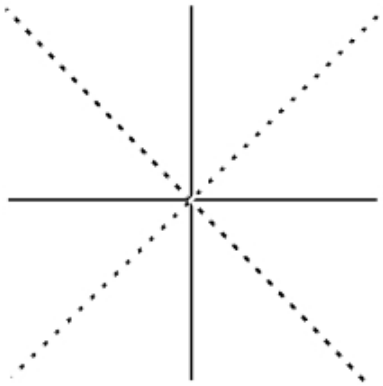

FIGURE 2. An attempt to sketch the graph of f.

The following diagram is misleading, since between any two real numbers there is both a rational and an irrational number.

The function f is continuous at 0. To see this, suppose $x_n \to 0$. Then $|x_n| \to 0$ (this follows from the definition of a limit). Since $-|x_n| \leq f(x_n) \leq |x_n|$, it follows from the Squeeze Theorem that $f(x_n) \to 0$, i.e. $f(x_n) \to f(0)$.

On the other hand, *f is not continuous at c if $c \neq 0$*. For example if c is irrational then we can choose a sequence of rationals x_n such that $x_n \to c$ (by repeated applications of the remark above in italics). It follows that $f(x_n) = x_n \to c \neq f(c)$. Similarly if c is irrational.

We will later define the exponential, logarithm, and trigonometric functions, and show they are continuous. Meanwhile, we will use them in examples (but not in the development of the theory).

3.3. Basic Properties of Continuous Functions

These properties follow easily from the analogous properties of sequences.

3.3.1. Combining Continuous Functions.

THEOREM 3.3.1. *Let f and g be continuous functions and let $D = \mathcal{D}(f) \cap \mathcal{D}(g)$. Then*

1. *$f + g$ is continuous on D,*
2. *fg is continuous on D,*
3. *αf is continuous on $\mathcal{D}(f)$ (α any real number),*
4. *f/g is continuous at any point $c \in D$ such that $g(c) \neq 0$.*

PROOF. Suppose $c \in D$. Let (x_n) be any sequence such that $x_n \to c$ (and as usual, $x_n \in D$).

Then $f(x_n) \to f(c)$ and $g(x_n) \to g(c)$, since f and g are continuous at c. It follows

$$f(x_n) + g(x_n) \to f(c) + g(c)$$

by Theorem 2.3.1 about sums of convergent sequences. That is,

$$(f + g)(x_n) \to (f + g)(c).$$

It follows that $f + g$ is continuous at c.

The proof in the other cases is similar. Just note for the case f/g that if $x_n \to c$ and $g(c) \neq 0$, then $g(x_n) \neq 0$ for all sufficiently large n.[6] □

The composition of two continuous functions is continuous. (See Adams page 35 for a discussion about the composition of functions.)

THEOREM 3.3.2. *Suppose f and g are continuous. Then $f \circ g$ is continuous.*

PROOF. The domain D of $f \circ g$ is the set of numbers x such that both $x \in \mathcal{D}(g)$ and $g(x) \in \mathcal{D}(f)$.

Suppose $c \in D$. Let $x_n \to c$ and $x_n \in D$. It follows that $g(x_n) \to g(c)$ since g is continuous at c. It then follows that $f(g(x_n)) \to f(g(c))$ since f is continuous at $g(c)$ (note that $g(x_n) \in \mathcal{D}(f)$). In other words, $(f \circ g)(x_n) \to (f \circ g)(c)$, and so $f \circ g$ is continuous at c. □

It follows from our results so far that rational functions (quotients of polynomials) and in general functions defined from other continuous functions by means of algebraic operations and composition, will be continuous on their domains.

3.3.2. Analogous Results at a Point. As in the previous section, we will usually be interested in functions that are continuous everywhere on their domain. However, occasionally we may not have continuity at every point in the domain. However, the following analogues of the previous theorems are proved with almost exactly the same proofs as before. *Check this yourself.*

THEOREM 3.3.3. *Suppose the functions f and g are both continuous at a, and c is any real number. Then the following are also continuous at a:*

$$f + g, \quad fg, \quad cf, \quad f/g \text{ provided } g(a) \neq 0.$$

If g is continuous at a and f is continuous at $g(a)$ then $f \circ g$ is continuous at a.

3.3.3. Removable and Non-Removable Singularities. The function

$$f_1(x) = \sin \frac{1}{x}$$

is the composition of the two continuous functions $\sin(x)$ and $1/x$ [7] and so is continuous. The domain of f_1 is the set of real numbers x such that $x \neq 0$. That is, $\mathcal{D}(f_1) = \{\, x \mid x \neq 0 \,\}$.

---

[6] If $g(c) > 0$, by continuity of g at c and the definition of convergence of a sequence, $g(x_n) \in [\frac{1}{2}g(c), \frac{3}{2}g(c)]$ for all sufficiently large n and so it is positive. The argument in case $g(c) < 0$ is similar.

[7] The notation may seem a bit confusing. You may ask "is it the same x in both cases"? But this is not the right way to look at it. By the function $\sin x$, is meant the function which assigns to each real number x (say) the real number $\sin x$. If we said the function $\sin y$, or just sin, we would mean the same thing.

Similarly, the function $1/x$, or $1/y$, or "the reciprocal function", all mean the same thing.

Similarly, the function

$$f_2(x) = x \sin \frac{1}{x}$$

is continuous on its domain, which is the *same* domain as for f_1.

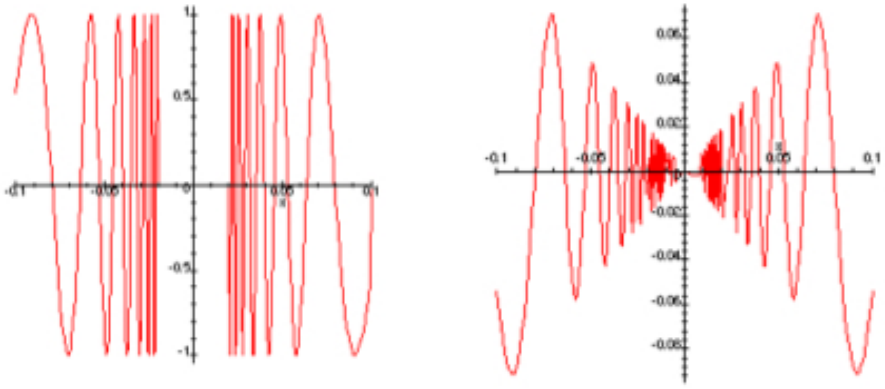

FIGURE 3. Graphs of $f_1(x) = \sin(1/x)$ and $f_2(x) = x \sin(1/x)$.

However, there is an important difference between f_1 and f_2, even though they both have the same domain and are continuous on this domain. In the case of f_2 we can define a new function g_2 by

$$g_2(x) = \begin{cases} x \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Then $\mathcal{D}(g_2) = \mathbb{R}$ and $g_2(x) = f_2(x)$ if $x \neq 0$, i.e. if $x \in \mathcal{D}(f_2)$. Moreover, g_2 is continuous on its domain $\mathbb{R}$.

To show continuity of g_2 at $x \neq 0$, take any sequence $x_n \to x$. For all sufficiently large n, $x_n \in \mathcal{D}(f_2)$, and so $g_2(x_n) = f_2(x_n)$. It follows that $g_2(x_n) \to g_2(x)$ since $f_2(x_n) \to f(x)$ by the continuity of f. This means g_2 is continuous at x if $x \neq 0$.

To show continuity of g_2 at $x = 0$, take any sequence $x_n \to 0$. Then

$$-|x_n| \leq g_2(x_n) \leq |x_n|$$

since $|\sin t| \leq 1$, and so $g_2(x_n) \to 0$ $(= g_2(0))$ by the Squeeze Theorem. (We need to be a bit careful since some of the x_n may equal zero.) This means g_2 is continuous at 0.

In the case of f_1 there is *no* way of extending the function to a continuous function g_1 defined on all of $\mathbb{R}$. This is essentially because there is no number y such that $f_1(x_n) \to y$ for *every* sequence $x_n \to 0$ (with $x_n \neq 0$.)

We sometimes say that f_2 has a *removable singularity* at 0, and that the singularity of f_1 at 0 is *not removable*.

3.4. Continuous Functions on a Closed Bounded Interval

The following two theorems are "deep" and require the Completeness Axiom for their proof.

Adams gives different proofs in Appendix III. But here we have easier proofs using sequences and the Bolzano Weierstrass Theorem.

See also Adams pp 80–85 for some discussion of these matters.

THEOREM 3.4.1 (Boundedness and Max-Min Theorems). *If f is continuous on $[a, b]$, then it is bounded there (i.e. there exists a constant K such that $|f(x)| \leq K$ if $a \leq x \leq b$).*

Moreover, there exist points $u, v \in [a, b]$ such that for any $x \in [a, b]$ we have

$$f(v) \leq f(x) \leq f(u).$$

That is, f assumes maximum and minimum values on $[a, b]$.

PROOF. Suppose in order to obtain a contradiction that f is not bounded above.

Then for each positive integer n there is some $x_n \in [a, b]$ such that $f(x_n) > n$. By the Bolzano Weierstrass Theorem there is a subsequence of (x_n), which we denote by (x'_n), such that (x'_n) converges to some $c \in [a, b]$. By continuity, $f(x'_n) \to f(c)$. But $f(x'_n) \to \infty$ since $f(x_n) \to \infty$.

This is a contradiction. Hence f is bounded above. Similarly f is bounded below.

Now let

$$L = lub\{f(x) : x \in [a, b]\}.$$

By what we have proved, L is finite. It follows from Theorem 2.5.5 applied to the set $S = \{f(x) : x \in [a, b]\}$ that there is a sequence (x_n) for which $f(x_n) \to L$.

By the Bolzano Weierstrass Theorem there is a subsequence of (x_n), which we denote by (x'_n), such that $x'_n \to u$ for some $u \in [a, b]$. By the continuity of f, $f(x'_n) \to f(u)$.

Because $f(x_n) \to L$ it follows $f(x'_n) \to L$. (*Why?*) By uniqueness of limits (see Theorem 2.3.4) it follows $f(u) = L$.

Since $f(x) \leq L$ for all $x \in [a, b]$, it immediately follows that $f(x) \leq f(u)$ for all $x \in [a, b]$.

Similarly there is a minimum point v. *Write out the proof.* □

REMARK 3.4.2. The Completeness Axiom is used in the proof, since it is used in the proof of the Bolzano Weierstrass Theorem.

In fact the Completeness Axiom is required for the theorem to hold. To see this, first note that all the other axioms hold if we restrict to the "universe of rational numbers".

Now consider the function $f(x) = x - x^3$ on the domain $[0, 1]$. The maximum occurs at $x = 1/\sqrt{3}$, as a little calculus shows. (One can also show this directly, *how?*) It follows there is no rational maximum point, and so the previous theorem is not true in the "universe of rational numbers".

REMARK 3.4.3. The next theorem implies that if a continuous function defined on an interval I (not necessarily closed or bounded) takes two particular values, then it must take all values between. In other words, for any two points $a, b \in I$ and any γ between $f(a)$ and $f(b)$, there is a $c \in [a, b]$ such that $f(c) = \gamma$.

In order to understand the proof, look at Figure 4. It is clear in this simple case that $f(c) = \gamma$. But the function f needs to be continuous, the proof needs to follow from the original defintion of continuity, and the Completeness Axiom has to come into the proof. It also should be noted that continuous functions can be pretty wild, there are examples which are not differentiable at any point!

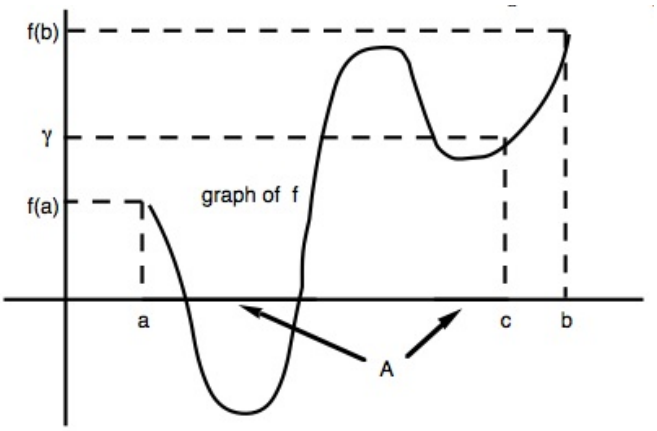

FIGURE 4. f is continuous on $[a, b]$, $A = \{x \in [a, b] : f(x) \leq \gamma\}$, $c = lub\, A$.

THEOREM 3.4.4 (Intermediate Value Theorem). *Suppose f is continuous on $[a, b]$. Then for any γ between $f(a)$ and $f(b)$ there exists $c \in [a, b]$ such that $f(c) = \gamma$.*

PROOF. We do the case where $f(a) < \gamma < f(b)$.

Let

$$A = \{x \in [a, b] \mid f(x) \leq \gamma\}.$$

Because A is bounded, and nonempty since $a \in A$, it follows that it has a supremum c, say.

We want to show that $f(c) = \gamma$.

There is a sequence $x_n \in A$ such that $x_n \to c$ (Theorem 2.5.5). By continuity, $f(x_n) \to f(c)$. Since $f(x_n) \leq \gamma$ for all n, it follows $f(c) \leq \gamma$ (special case of Theorem 2.3.6). So, in particular, $c \in A$.

Since $c \neq b$ (because $f(c) \leq \gamma$ but $f(b) > \gamma$), there is a sequence $x_n' \to c$ and $c < x_n' < b$. But $f(x_n') > \gamma$ (since otherwise $x_n' \in A$, which contradicts $c = \sup A$) and so $f(c) \geq \gamma$ (Theorem 2.3.6 again).

Because $f(c) \leq \gamma$ and $f(c) \geq \gamma$ it follows $f(c) = \gamma$. □

EXAMPLE 3.4.5. See Example 11 in Section 1.4 of Adams. Here the Intermediate Value Theorem is used to justify the existence of a solution of the equation $x^3 - x - 1 = 0$.

One can also prove the existence of a number x such that $x^2 = 2$ in this manner. Just note that if $f(x) = x^2$ then $f(1) = 1$, $f(2) = 4$, and since f is continuous it follows by the Intermediate Value Theorem that $f(x) = 2$ for some x between 1 and 2. Thus we have justified the existence of $\sqrt{2}$, i.e. a positive number whose square is 2.

EXAMPLE 3.4.6. In order to improve/test your understanding give examples to show that all the hypotheses in the previous two theorems are necessary. Make your examples as simple as possible.

A first example might be of a simple function f which satisfies all the hypotheses of Theorem 3.4.1 except that f is not continuous and the theorem's conclusion of being bounded does not hold. Another example would give a discontinuous f which is bounded but does not have a maximum value. Others would give examples where the domain is $(0, 1]$ (hence not a closed bounded interval) and the various conclusions of the theorem do not hold.

EXAMPLE 3.4.7. *An interesting problem.* Use the Intermediate Value Theorem to prove that at any fixed point in time, there are two antipodal points on the equator with the same temperature. Assume the temperature is a continuous function of position.

3.5. ★Functions of two or more variables

Suppose $f : A\ (\subset)\mathbb{R}^2 \to \mathbb{R}$. Think of the case of a "closed bounded rectangle"

$$A = \{(x, y) : a \leq x \leq b,\ c \leq y \leq d\}.$$

We say f *is continuous at* $(u, v) \in A$ if, for every sequence $\big((x_n, y_n)\big)_{n\geq 1}$ from A such that $(x_n, y_n) \to (u, v)$, we have $f(x_n, y_n) \to f(u, v)$.[8] This is completely analogous to Definition 3.2.1.

Analogues of Theorems 3.3.1, 3.3.2 and 3.3.3 hold, with similar proofs. In particular, if h and g are functions of one variable which is continuous at a (and hence defined at a), and f is a function of two variables which is continuous at $(h(a), g(a))$ (and hence defind there), then $f(h(x), g(x))$ is a function of one variable which is continuous at a.

Analogues of the Bolzano Weierstrass Theorem 2.6.2, Boundedness and Max-Min Theorem 3.4.1 and the Uniform Continuity Theorem 4.5.2 hold for continuous functions on any closed bounded rectangle A. The proofs are similar.

[8] We define $(x_n, y_n) \to (u, v)$ to mean that $|(x_n, y_n) - (u, v)| \to 0$. This can be easily shown to be equivalent to $x_n \to u$ and $y_n \to v$. *Exercise.*

CHAPTER 4

Limits

4.1. Definition via Sequences, and Examples

We often need to consider limits of a function f at a point c where f may not be continuous or even defined. For example, consider the two cases

$$f(x) = \begin{cases} 2x & \text{if } x \neq 1 \\ \text{undefined} & \text{if } x = 1 \end{cases}, \qquad f(x) = \begin{cases} 2x & \text{if } x \neq 1 \\ 3 & \text{if } x = 1 \end{cases}.$$

In either case we want the limit of f at $c = 1$ to equal 2, that is we want a definition of limit such that $\lim_{x\to 1} f(x) = 2$.

Important points to take into account when defining $\lim_{x\to c} f(x)$ are that

(1) c may not be in $\mathcal{D}(f)$;
(2) if $c \in \mathcal{D}(f)$ then the value of $f(c)$ is not relevant to the existence or value of the limit at c;
(3) we need at least one sequence (x_n) from $\mathcal{D}(f)$ such that $x_n \to c$ and such that $x_n \neq c$ for all n.

We now first define the set of points at which we consider the existence of a limit, and then we define what we mean by a limit. The following definition is a natural consequence of the previous discussion.

DEFINITION 4.1.1. Suppose f is a function with domain $\mathcal{D}(f)$. Then c is a *limit point* of $\mathcal{D}(f)$ if there exists a sequence (x_n) from $\mathcal{D}(f)$ such that $x_n \to c$ and such that $x_n \neq c$ for all n.

Suppose c is a limit point of $\mathcal{D}(f)$. If $f(x_n) \to L$ for every sequence (x_n) from $\mathcal{D}(f)$ such that $x_n \to c$ and $x_n \neq c$ for all n, then we say that f *has limit* L *at* c. We write

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

REMARK 4.1.2. If $c \in \mathcal{D}(f)$ and c is not a limit point of $\mathcal{D}(f)$ we say c is an *isolated* point of $\mathcal{D}(f)$. For example, if $\mathcal{D}(f) = [0, 1] \cup \{3\}$ then 3 is an *isolated* point of $\mathcal{D}(f)$. See Example 4.1.3. □

EXAMPLE 4.1.3. For the two examples at the beginning of this Chapter it follows from Definition 4.1.1 that $\lim_{x\to 1} f(x) = 2$. *Why?* In fact we have in both cases that $\lim_{x\to a} f(x) = 2a$ for *every* real number a. *Why?*

Other simple examples are

$$f(x) = 2x \text{ for } x \in \mathcal{D}(f) = [0, 1], \qquad f(x) = 2x \text{ for } x \in \mathcal{D}(f) = [0, 1).$$

In both cases $\lim_{x\to 1} f(x) = 2$.

Finally, if

$$f(x) = 2x \text{ for } x \in \mathcal{D}(f) = [0,1] \cup \{3\},$$

then $\lim_{x\to 3} f(x)$ is *not* defined, *why?* In particular, $\lim_{x\to 3} f(x) = 6$ is not a true statement, since there is no limit of f at 3.

4.2. Calculating Limits

The following rules are listed on page 68 of Adams.

The proofs for us are easy, because we defined limits via sequences and we have already proved similar properties for sequences in Theorem 2.3.1 and Section 2.4.

Since Adams defines limits using the $\epsilon - \delta$ approach, as we later also do here in Section 4.4, he needs to prove the following Theorem from "scratch". See Adams page 89 for the sum and the Exercises on page 92 for the other cases.

THEOREM 4.2.1. *Suppose that a is a limit point of $\mathcal{D}(f) \cap \mathcal{D}(g)$.*[1] *Suppose $\lim_{x\to a} f(x) = L$, $\lim_{x\to a} g(x) = M$ and k is a constant. Then*

(1) $\lim_{x\to a} \big(f(x) + g(x)\big) = L + M$;
(2) $\lim_{x\to a} \big(f(x)g(x)\big) = LM$;
(3) $\lim_{x\to a} \big(kf(x)\big) = kL$;
(4) $\lim_{x\to a} \big(f(x)/g(x)\big) = L/M$, *if* $M \neq 0$.

PROOF. (We just do the quotient. The other cases are similar and slightly easier. *Do them yourself!*)

Suppose $x_n \to a$ where $(x_n)_{n\geq 1}$ is *any* sequence such that for all n, $x_n \in \mathcal{D}(f) \cap \mathcal{D}(g)$.

Then $f(x_n) \to L$ and $g(x_n) \to M$. (*Why?*) Moreover, since $M \neq 0$ it follows that $g(x_n) \neq 0$ for all sufficiently large n (*why?*) and so $f(x_n)/g(x_n)$ is a real number for such n.

Since $M \neq 0$ it follows from Theorem 2.3.1 that $f(x_n)/g(x_n) \to L/M$.

The fourth claim in the Theorem now follows from Definition 4.1.1. □

4.3. Limits and Continuity

In the following Proposition, particularly the second paragraph, we see that once we have the definition of the limit of a function at a point c, we can define the notion of continuity at c in terms of such a limit.

The proof is a straightforward matter of using the relevant definitions. It is just a little messy and tedious to write out! The main point is that in Definition 3.2.1 of continuity we consider *all* sequences $x_n \to c$, whereas in Definition 4.1.1 of a limit we only consider sequences $(x_n)_{n\geq 1}$ satisfying $x_n \neq c$. (In both cases, as usual, $x_n \in \mathcal{D}(f)$.)

[1] It follows that a is a limit point of both $\mathcal{D}(f)$ and $\mathcal{D}(g)$. *Why?* The important and simple case to keep in mind is where f and g have the same domain and this domain is an interval.

Note that Adams does not need to consider the first paragraph in the Proposition, since the type of domain he considers does not contain isolated points.[2]

PROPOSITION 4.3.1. *Suppose f is a function and $c \in \mathcal{D}(f)$. If c is not a limit point of $\mathcal{D}(f)$, i.e. c is an isolated point of $\mathcal{D}$, then f is continuous at c.*

If c is a limit point of $\mathcal{D}(f)$ then f is continuous at c iff $\lim_{x\to c} f(x) = f(c)$.

PROOF. First suppose $c \in \mathcal{D}(f)$ and c is an isolated point in $\mathcal{D}(f)$. (Think of the last example in Example 4.1.3.)

In order to apply Definition 3.2.1 let (x_n) be any sequence from $\mathcal{D}(f)$ such that $x_n \to c$. It follows that $x_n = c$ for all sufficiently large n (*why?*). Hence $f(x_n) = f(c)$ for all sufficiently large n, and so in particular $f(x_n) \to f(c)$. Hence f is continuous at c.

Next suppose $c \in \mathcal{D}(f)$ and c is a limit point of $\mathcal{D}(f)$.

First assume f is continuous at c. By Definition 3.2.1, for any sequence (x_n) from $\mathcal{D}(f)$ such that $x_n \to c$ (even if $x_n = c$ for some n) it follows that $f(x_n) \to f(c)$. In particular it follows that $\lim_{x\to c} f(x) = f(c)$.

Next assume that $\lim_{x\to c} f(x) = f(c)$. (Remember that Definition 4.1.1 only considers sequences (x_n) satisfying $x_n \neq c$.) Now consider *any* sequence (x_n) from $\mathcal{D}(f)$ such that $x_n \to c$. We want to show that $f(x_n) \to f(c)$. Split the sequence (x_n) into one subsequence (x'_n) consisting of those terms $x_n \neq c$ and the remaining subsequence (x''_n) consisting of those terms $x_n = c$. At least one of these subsequences is infinite.

If (x'_n) is an infinite sequence, we have that $f(x'_n) \to f(c)$ by the Definition 4.1.1 of a limit. If (x''_n) is an infinite sequence, we have $f(x''_n) \to f(c)$ since $f(x''_n) = f(c)$ for all n. Putting these two facts together it follows that for the original sequence (x_n) we have $f(x_n) \to f(c)$. □

4.4. Definitions of Limit and Continuity via $\epsilon - \delta$

See Adams, Section 1.5 pages 87 and 88. The following definition is just a little different from Definition 8 on page 88 in Adams. It is written in exactly the same format. *What is the difference?*

Note that Definition 8 in Adams applies only to domains which are a finite union of intervals, and does not usually apply to endpoints of such an interval.[3] Adams uses Definition 9 on page 90 of a right limit (and similarly for left limit), to take care of endpoints.

DEFINITION 4.4.1. We say that $f(x)$ approaches the limit L as x approaches a, and we write

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

if the following conditions are satisfied:

[2]Isolated points are just a bit of a nuisance and not particularly significant when we consider limits and continuity.

[3]If the domain is $(0,1)$, $[0,1]$, $(0,1]$ or $[0,1)$, then Adams' definition of a limit does not apply to either 0 or 1, whereas the (standard) definition given here does apply. If the domain is $[0,1) \cup (1,2]$ then Adams' definition does apply at 1, as does the one here.

(1) a is a limit point of $\mathcal{D}(f)$,
(2) for every number $\epsilon > 0$ there exists a number $\delta > 0$, possibly depending on ϵ, such that if $0 < |x - a| < \delta$ and $x \in \mathcal{D}(f)$, then
$$|f(x) - L| < \epsilon.$$

The idea of the definition is that for any given "tolerance" $\epsilon > 0$, there is a corresponding "tolerance" $\delta > 0$, such that any x in the domain of f within distance δ of a, and not equal to a, gives an output (or value) $f(x)$ which is within ϵ of L.

Inputs and Outputs Analogy. For any allowable x (i.e. $x \in \mathcal{D}(f)$), input x into the f machine gives output $f(x)$.

Suppose there are allowable inputs arbitrarily close to a. Then $\lim_{x\to a} f(x) = L$ is equivalent to the following:

For every output tolerance $\epsilon > 0$ for the deviation of $f(x)$ from L, there is a corresponding input tolerance $\delta > 0$ (normally depending on ϵ) for the deviation of x from a, such that whenever the input tolerance is satisfied by $x \neq a$, then the output tolerance is satisfied by $f(x)$.

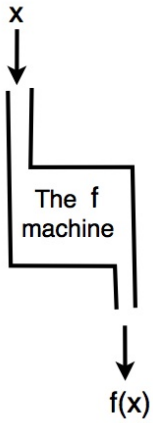

FIGURE 1. $\lim_{x\to a} f(x) = L$ means: for each $\epsilon > 0$ there is a corresponding $\delta > 0$ such that if $x \neq a$ and x is δ-tolerant from a, then $f(x)$ is ϵ-tolerant from L.

Now we come to the important theorem showing the approach to limits in Definition 4.1.1 via sequences, and the $\epsilon - \delta$ approach in Definition 4.4.1, are equivalent.

The proof in one direction is easy.[4] The other direction is more subtle and is not covered in Adams, even in the Appendices. If you find it tricky, don't be too concerned. (Just think about it for a couple of hours each day.)

THEOREM 4.4.2. *Suppose f is a function and c is a limit point of $\mathcal{D}(f)$. Then $\lim_{x\to c} f(x) = L$ according to Definition 4.1.1 iff $\lim_{x\to c} f(x) = L$ according to Definition 4.4.1.*

[4]See Adams Theorem 1(b) and Theorem 4 in Appendix III for similar results. But it is easier in Adams as he just covers special cases.

PROOF. First suppose $\lim_{x\to c} f(x) = L$ according to Definition 4.4.1. We want to show that

$$x_n \neq c \;\&\; x_n \to c \quad \Longrightarrow \quad f(x_n) \to L. \tag{23}$$

In order to do this, suppose $\epsilon > 0$ is given. We need to show there is a corresponding N such that

$$n > N \quad \Longrightarrow \quad |f(x_n) - L| < \epsilon. \tag{24}$$

First note from Definition 4.4.1 that there exists $\delta > 0$ such that

$$0 < |x - c| < \delta \quad \Longrightarrow \quad |f(x) - L| < \epsilon.$$

But if $x_n \neq c$ and $x_n \to c$, then there is an N corresponding to δ such that

$$n > N \quad \Longrightarrow \quad 0 < |x_n - c| < \delta.$$

Combining the last two implications gives (24). Since $\epsilon > 0$ was arbitrary, this gives (23). In other words, f is continuous at c according to Definition 4.1.1.

Conversely, suppose f is continuous at c according to Definition 4.1.1. Suppose $\epsilon > 0$ is given. We want to show there is some corresponding $\delta > 0$ such that

$$0 < |x - c| < \delta \quad \Rightarrow \quad |f(x) - L| < \epsilon. \tag{25}$$

Assume there is no such δ (in order to obtain a contradiction). This means that for each $\delta > 0$ there is a real number x (with $x \in \mathcal{D}(f)$) such that

$$0 < |x - c| < \delta \quad \text{but} \quad |f(x) - L| \geq \epsilon.$$

Let $\delta = \frac{1}{n}$ and denote some x as above by x_n. Thus we have for each natural number n a real number x_n such that

$$0 < |x_n - c| < \frac{1}{n} \quad \text{but} \quad |f(x_n) - L| \geq \epsilon.$$

It follows that $x_n \neq c$ and $x_n \to c$ but $f(x_n) \not\to L$.

This contradicts the fact that f is continuous at c according to Definition 3.2.1, and so the *assumption* is false. In other words, there *is* a $\delta > 0$ corresponding to ϵ such that (25) is true. Since $\epsilon > 0$ was arbitrary, it follows that f is continuous at c according to Definition 4.4.1. □

There is also an $\epsilon - \delta$ definition of continuity at c and a corresponding theorem showing the equivalence with the sequence definition. The ideas are essentially the same as for limits, but easier because we do not need to restrict to $x \neq c$ and to $x_n \neq c$.

DEFINITION 4.4.3. A function f *is* continuous at a point $c \in \mathcal{D}(f)$ if for every $\epsilon > 0$ there is a corresponding $\delta > 0$ such that

$$x \in \mathcal{D}(f) \text{ and } |x - c| < \delta \quad \Rightarrow \quad |f(x) - f(c)| < \epsilon.$$

The function f is *continuous* (on its domain) if f is continuous at every point in its domain.

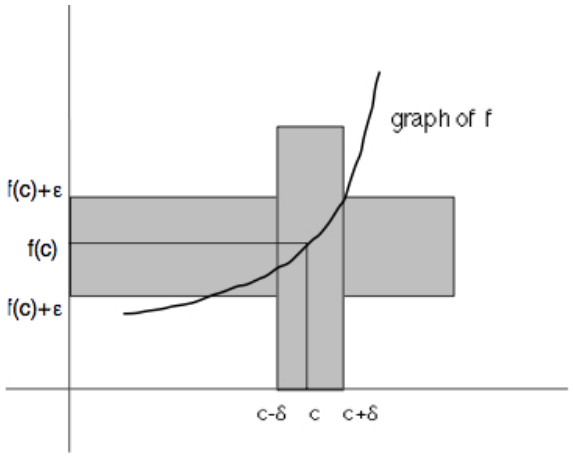

FIGURE 2. In this diagram we see that

$$|x - c| < \delta \Longrightarrow |f(x) - f(c)| < \epsilon.$$

For ϵ as shown, I have drawn the largest δ for which this implication holds. Can you explain to a neighbour why a larger δ will not work? Any smaller $\delta > 0$ will also work.

See Figure 2 for a diagram similar to that in Adams on page 87 for limits.

Finally, we have the following theorem.

THEOREM 4.4.4. *Suppose f is a function and $c \in \mathcal{D}(f)$.*
Then f is continuous at c according to Definition 3.2.1 iff it is continuous at c according to Definition 4.4.3.

PROOF. Almost exactly the same as for the proof of Theorem 4.4.2. Write it out yourself to consolidate your understanding. □

4.5. Uniform Continuity

In the Definition 4.4.3 of continuity at a point $c \in \mathcal{D}(f)$, we require that for every $\epsilon > 0$ there is a corresponding $\delta > 0$ with a certain property. The number δ will normally depend on ϵ, but δ may also depend on c in an essential manner.

As a simple example, consider the function $f(x) = x^2$ defined on $\mathbb{R}$. For simplicity consider the case where c is large and positive and $x \in (c-1, c+1)$. Then

$$|f(x) - f(c)| = |x^2 - c^2| = |x - c|\,|x + c| \begin{cases} \leq |x - c|\,(2c + 1) \\ \geq |x - c|\,(2c - 1) \end{cases}.$$

This implies that the $\delta > 0$ needed in Definition 4.4.3 will be roughly of the order $\epsilon/2c$ or less. More precisely, any δ less than $\epsilon/(2c + 1)$ will work, but δ must not be more than $\epsilon/(2c - 1)$.

Thus for a given $\epsilon > 0$, no single δ will work for every c. For a fixed $\epsilon > 0$, the larger we take c the smaller we need to take δ.

If for each $\epsilon > 0$ there is a $\delta > 0$ which works for all $c \in \mathcal{D}(f)$ then we say that f is uniformly continuous on its domain.

DEFINITION 4.5.1. A function f is *uniformly* continuous on its domain if for every $\epsilon > 0$ there is a corresponding $\delta > 0$ such that, for every $c \in \mathcal{D}(f)$,

$$x \in \mathcal{D}(f) \text{ and } |x - c| < \delta \quad \Longrightarrow \quad |f(x) - f(c)| < \epsilon.$$

THEOREM 4.5.2. *If f is continuous on a closed bounded interval $[a, b]$ then it is uniformly continuous on $[a, b]$.*

PROOF. We argue by contradiction.

Assume f is *not* uniformly continuous on its domain. By Definition 4.5.1 this is equivalent to claiming there is a "bad" $\epsilon > 0$, for which there is *no* $\delta > 0$ such that for every $c \in \mathcal{D}(f)$,

$$x \in \mathcal{D}(f) \text{ and } |x - c| < \delta \quad \Longrightarrow \quad |f(x) - f(c)| < \epsilon.$$

In particular, for each positive integer n, we see by taking $\delta = 1/n$ that there exists $c_n \in \mathcal{D}(f)$ and $x_n \in \mathcal{D}(f)$, such that

$$|x_n - c_n| < \frac{1}{n} \quad \text{and} \quad |f(x_n) - f(c_n)| \geq \epsilon. \tag{26}$$

By the Bolzano Weierstrass Theorem there is a subsequence of (c_n), which we denote by (c'_n), such that $c'_n \to c$ for some $c \in [a, b]$. Let (x'_n) be the corresponding subsequence of (x_n). It follows that $x'_n \to c$ since $x'_n - c'_n \to 0$ (*why?*) and so

$$x'_n = c'_n + (x'_n - c'_n) \to c + 0 = c.$$

Since f is continuous at c, and since both $x'_n \to c$ and $c'_n \to c$, it follows that $f(x'_n) \to f(c)$ and $f(c'_n) \to f(c)$. Hence $|f(x'_n) - f(c'_n)| \to 0$, which contradicts the fact from (26) that $|f(x'_n) - f(c'_n)| \geq \epsilon$ for all n.

This contradiction shows our *assumption* is false and so f *is* uniformly continuous. □

REMARK 4.5.3. We saw an example at the beginning of this section showing the theorem would not hold if we dropped just the assumption that the domain of f is bounded.

If we drop just the assumption that the domain be closed, the theorem will not hold. A counterexample is obtained by defining $f(x) = 1/x$ for $x \in (0, 1]$. *Try and write out the details.*

REMARK 4.5.4. If $f(x) = \sqrt{x}$ for $x \in [0, 1]$ then the function f is uniformly continuous on its domain by the Theorem. This may seem surprising, since $f'(x) \to \infty$ as $x \to 0$.

The fact that $f(x) = x^2$ is not uniformly continuous on $\mathbb{R}$ is *related* to the fact that $f'(x) \to \infty$ as $x \to \infty$. But this is not a complete explanation!

CHAPTER 5

Differentiation

The main references in Adams is Chapter 2.

Convention and Notation In this and the following chapters, unless stated otherwise, *the domain of a function is an interval.* The interval may be open or closed at either end, and may be bounded or unbounded.[1]

A point in the domain is an *interior point* if it is not an endpoint.

The results we prove usually extend in a straightforward manner to more general cases, in particular if the domain is a finite union of intervals.

5.1. Introduction

The theory of differentiation allows us to analyse the concept of the slope of the tangent to the graph of a function. Similarly, it allows us to find the best linear approximation to a function near a given point.

If we write $y = f(x)$ then we can interpret $f'(a)$ in the following way: "for x near a, y is changing approximately $f'(a)$ times as fast as x is changing".

Alternatively, for $x \approx a$ ("x approximately equal to a") we have

$$f(x) \approx f(a) + f'(a)(x - a).$$

See Adams Chapter 4.9, where this is made more precise.

There are many problems that can then be analysed using the ideas of differentiation and extensions of these ideas. For example, anything that changes with time or position, as well as optimisation problems (e.g. in economics or engineering) and approximation problems. See Adams Chapter 3 for a number of examples.

5.2. The Derivative of a Function

Derivatives are defined and the fact that differentiability implies continuity is proved.

The idea from Figure 1 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 $h \to 0$ (where h is allowed to be either positive or negative, except at endpoints a of an interval from the domain of f).

[1] Thus the allowable domains are (a, b), $[a, b]$, $(a, b]$, $[a, b)$, (the four bounded possibilities); (a, ∞), $[a, \infty)$, $(-\infty, b)$, $(-\infty, b]$, $(-\infty, \infty)$, (the five unbounded possibilities).

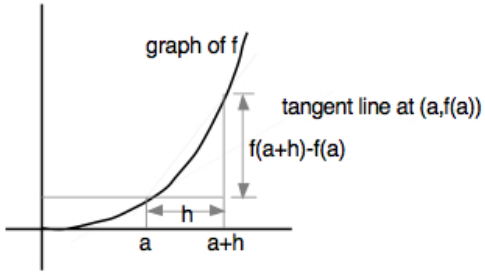

FIGURE 1. Draw in the tangent line at $(a, f(a))$.

DEFINITION 5.2.1. If a is an interior point of the domain of f [2] and

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

exists, or if $a \in \mathcal{D}(f)$ and is an endpoint and the corresponding one-sided limit

$$\lim_{h\to 0+} \frac{f(a+h)-f(a)}{h} \quad \text{or} \quad \lim_{h\to 0-} \frac{f(a+h)-f(a)}{h}$$

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.

The function f is *differentiable* if it is differentiable at every point in its domain.

An *alternative way* of writing the same limit is

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

and similarly for the one-sided limits. *Why* is this equivalent?[3]

The tangent to the graph of f at a has slope $f'(a)$. It follows that the equation of the tangent line is

$$y = f(a) + f'(a)(x-a).$$

See the examples on pages 96 and 97 of Adams, where the derivatives of some simple functions x^2, $1/x$ and $\sqrt[3]{x}$ are calculated directly from the above definition. But we do not usually need to do this. Instead we normally can use the methods in Section 5.3.

NOTATION 5.2.2. 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:

$$y', \quad \frac{dy}{dx}, \quad \frac{d}{dx}f(x), \quad f'(x),$$

[2] This, of course, implies that $a \in \mathcal{D}(f)$.

[3] Just write out the corresponding $\varepsilon - \delta$ definition in each case and checking that each limit means the same thing.

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

$$\frac{d}{dx}x^3 = 3x^2,$$
$$\frac{d}{dt}t^4 = 4t^3,$$

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 *fixed* real 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 $\Big|$ 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

$$\Delta x = (x+h) - h = h$$

as being the *increment in* x, then

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$

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 5.3.4 for a good example of what can go wrong.

See Adams pp 103-105 for more discussion of notation.

The following theorem is important.

THEOREM 5.2.3. *If f is differentiable at a then f is continuous at a.*

PROOF. Assume f is differentiable at a, and a is an interior point of $\mathcal{D}(f)$. We want to show[4] that

$$\lim_{h \to 0} f(a+h) = f(a).$$

[4] f is continuous at a means $\lim_{x \to a} f(x) = f(a)$. This is the same as $\lim_{h \to 0} f(a+h) = f(a)$. See Footnote 3 for a similar but slightly more complicated situation.

But

$$f(a+h) = f(a) + h\frac{f(a+h)-f(a)}{h}.$$

Taking the limit as $h \to 0$ 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

$$\lim_{h\to 0} f(a+h) = f(a) + 0f'(a) = f(a).$$

A similar proof applies if a is an endpoint of the domain of f. □

5.3. Computing Derivatives

The standard rules for differentiation, including the chain rule, are discussed. Examples are given.

5.3.1. Sums, Products and Quotients. The next result is easy to check from the definitions, 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 5.3.1. *If $f(x) = cx + d$ then $f'(x) = c$.*

PROOF.

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{(c(x+h)+d)-(cx+d)}{h} = \lim_{h\to 0} \frac{ch}{h} = c.$$

□

The next theorem follows in a fairly straightforward way from the properties of limits given in Theorem 4.2.1.

THEOREM 5.3.2. *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.*

$$\begin{aligned}
(f \pm g)'(x) &= f'(x) \pm g'(x) \\
(cf)'(x) &= cf'(x) \\
(fg)'(x) &= f'(x)g(x) + f(x)g'(x) \\
\left(\frac{1}{g}\right)'(x) &= \frac{-g'(x)}{(g(x))^2} \\
\left(\frac{f}{g}\right)'(x) &= \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}
\end{aligned}$$

In the last two cases we also assume $g(x) \neq 0$.

PROOF. See Adams, Section 3.3. □

The following now follows from the product rule and the Principle of Induction.

THEOREM 5.3.3. *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

$$(x^{n+1})' = (xx^n)' = x'x^n + xnx^{n-1} = x^n + xnx^{n-1} = (n+1)x^n.$$

Thus the corresponding result is true for $n + 1$.

This gives the result for *all* natural numbers n by the Principle of Induction. □

It follows that if $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ then $f'(x) = a_1 + 2a_2x + \cdots + na_nx^{n-1}$.

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 p107 Question 54. 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 p112, just before Example 8.

In a similar way, one can prove the general rule $(x^r)' = rx^{r-1}$ for any *rational* number r whenever the function x^r is defined. The same result is also true for *any* real number r, as we would expect by taking a sequence of rational numbers $r_n \to r$. But this is best proved by first developing the theory of logarithms and exponential functions. See in particular Adams p179, just before Example 6.

In fact, natural (i.e. to base e) logarithms are defined as areas under the graph of the function $1/x$, see Adams p174.[5] Then the derivative of $\log_e x$ is proved to be $1/x$. From this one can then derive the usual rules for derivatives of exponentials a^x and general power functions x^a, see Adams p179 between Examples 5 and 6.

The proofs of the usual rules for the derivatives of the trigonometric functions are given in Adams Section 2.5. They are not completely rigorous, since the definition of sin and the other trigonometric functions was only given informally, using diagrams, in Section P7.[6]

At this stage, we only use derivatives of such functions in the examples, but not in the rigorous development of the subject.

---

[5] In Chapter 5 the area is defined more precisely as a Riemann integral.

[6] ★To do things rigorously, without using diagrams other than for motivation, one approach is to *define* $\sin x$ by means of the power series in Adams p 540. Another is to use complex numbers and *define* $\sin x$ as in Adams Exercise 17 page A-19.

The reason for rigour is not a matter of being obsessive compulsive about such things. Diagrams and intuition are extremely useful and important, but they can be misleading. Moreover, passing into the realm of complex numbers turns out to be extremely useful. The analysis of periodic behaviour is fundamental throughout physics, engineering, meteorology, ..., and is essentially intractable if one does not pass through the world of complex numbers!

5.3.2. The Chain Rule. 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

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx},$$

where $u = g(x)$, $y = f(u)$ and $y = f(g(x))$.

The informally stated motivation is that for a given value of $x = a$ and corresponding value $u = g(a)$,

$$\text{at } x = a,\ u \text{ is changing } \frac{du}{dx} \text{ times as fast as } x,$$

and

$$\text{at } u = g(a),\ y \text{ is changing } \frac{dy}{du} \text{ times as fast as } u$$

(where $\dfrac{du}{dx}$ is evaluated at $x = a$ and $\dfrac{dy}{du}$ is evaluated at $u = g(a)$). So that

$$\text{at } x = a,\ y \text{ is changing } \frac{dy}{du} \times \frac{du}{dx} \text{ times as fast as } x.$$

In functional notation

$$(f \circ g)'(a)) = f'(g(a))\, g'(a) \quad \text{or} \quad (f \circ g)'(x)) = f'(g(x))\, g'(x).$$

An incorrect "proof" along these lines is often given for the chain rule by writing

$$\begin{aligned}\frac{dy}{dx} &= \lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x} = \lim_{\Delta x\to 0}\frac{\Delta y}{\Delta u}\frac{\Delta u}{\Delta x} = \lim_{\Delta x\to 0}\frac{\Delta y}{\Delta u}\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x}\\ &= \lim_{\Delta u\to 0}\frac{\Delta y}{\Delta u}\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x} = \frac{dy}{du}\frac{dy}{dx}\end{aligned}$$

The second last step is "justified" by saying that $\Delta u \to 0$ as $\Delta x \to 0$. This is all rather sloppy, because it is not clear what depends on what.

When one tries to fix it up, there arises a 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 5.3.5).

This becomes clearer when we write out the argument in a more precise functional notation. See Adams Q46 p119.

We now state the Chain Rule precisely, and refer to Adams for a (correct) proof.

THEOREM 5.3.4 (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 some open interval containing x and f is defined in some open interval containing $g(x)$, although this can be generalised a bit.)

PROOF. See Adams p118. 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.) □

EXAMPLE 5.3.5.

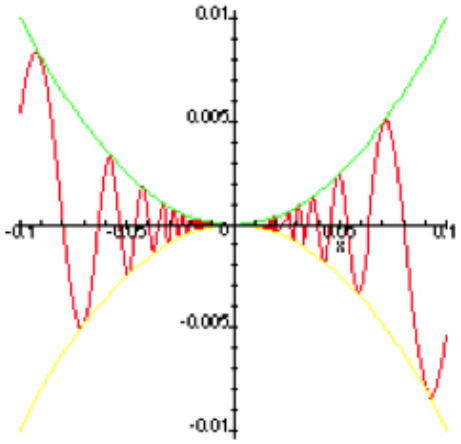

FIGURE 2. The graph of f if $f(x) = x^2 \sin(1/x)$ at $x \neq 0$ and $f(0) = 0$.

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 $x \to 0$, 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 $h \to 0$.

Analytically,

$$f'(0) = \lim_{h \to 0} \frac{f(h) - f(0)}{h} = \lim_{h \to 0} \frac{h^2 \sin \frac{1}{h} - 0}{h} = \lim_{h \to 0} h \sin \frac{1}{h} = 0.$$

The last limit follows easily from the Squeeze Theorem in Adams p69 applied with $\pm x$. (This Squeeze Theorem for limits follows easily from the Squeeze Theorem 2.3.7 here, for sequences. *Exercise.*)

Thus f is differentiable for all x, but the derivative is not continuous at 0.

5.4. Maximum and Minimum Values

The relationship between derivatives and maximum and minimum points is given.

DEFINITION 5.4.1. 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 an open interval N containing x_0 such that

$$f(x) \leq f(x_0) \quad (f(x) \geq f(x_0))$$

for all x in $N \cap \mathcal{D}(f)$.

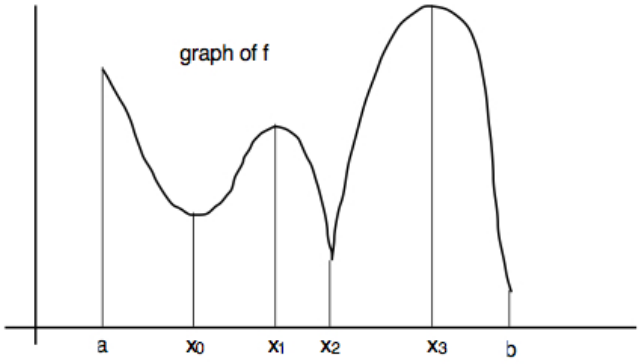

FIGURE 3. Which are the (local) maximum and minimum points?

In Figure 3, 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 3.4.1 that a *continuous* function f defined on a closed bounded interval always has a maximum and a minimum value.

If the domain of f is an interval and f has a local maximum or minimum at x there are three logical possibilities:

- x is an endpoint of the domain of f;
- x is not an endpoint and $f'(x)$ does not exist ;
- x is not an endpoint and $f'(x)$ does exist.

THEOREM 5.4.2. *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 non endpoint 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,

$$\frac{f(x_0 + h) - f(x_0)}{h} \leq 0 \text{ if } 0 < h < h_0. \tag{27}$$

and

$$\frac{f(x_0 + h) - f(x_0)}{h} \geq 0 \text{ if } -h_0 < h < 0. \tag{28}$$

We know that the derivative at x_0 exists and hence

$$\lim_{h \to 0+} \frac{f(x_0 + h) - f(x_0)}{h} \text{ and } \lim_{h \to 0-} \frac{f(x_0 + h) - f(x_0)}{h}$$

both exist and are equal. But the first limit is ≤ 0 from(28) and the second is ≥ 0 from(27). Hence the derivative must be 0. □

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 $f'(x_0) = 0$.

It is not true that if f has a local maximum or minimum at an endpoint then the derivative is zero there. Just consider $f(x) = x$ on $[0, 1]$.

5.5. 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 f 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 5.5.1 (Mean Value Theorem). *Suppose f is continuous on a closed bounded interval $[a, b]$ and is differentiable on the open interval (a, b). Then there exists $c \in (a, b)$ such that*

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

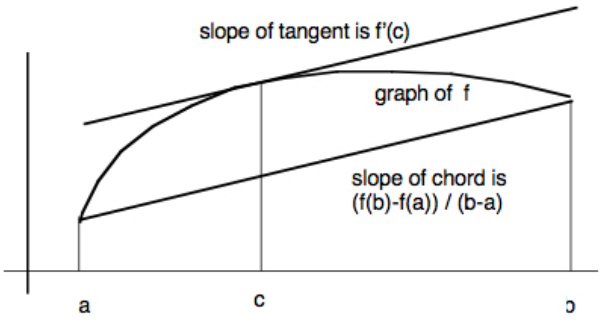

FIGURE 4. Example of the Mean Value Theorem.

PROOF. Consider the function g given by

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

Then $g(a) = g(b) = 0$ and g is continuous on $[a, b]$. (*Why?*)

We *claim* that $g'(c) = 0$ for some *interior* point c. To prove the claim consider three cases, at least one of which must occur:

(1) $g(x) = 0$ for all x in $[a, b]$. Since g is a constant functon it follows immediately from the definition of differentiation that $g'(c) = 0$ for every c in (a, b).
(2) $g(x) > 0$ for some x in $[a, b]$. By Theorem 3.4.1 the function g has at least one *maximum* point c in $[a, b]$. Since $g(c) > 0$ (*why?*), and since $g(a) = g(b) = 0$, if follows that $c \neq a$ and $c \neq b$. Hence c is an interior point and so $g'(c) = 0$ by Theorem 5.4.2.
(3) $g(x) < 0$ for some x in $[a, b]$. Then g has at least one *minimum* point c in $[a, b]$, this is an interior point, and $g'(c) = 0$, by a similar argument to case (2). *Write it out yourself!*

Since $g'(c) = 0$, and by the rules for differentiation

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

this proves the theorem. □

REMARK 5.5.2. The proof used the Max-Min Theorem 3.4.1, which in turn required the Completeness Axiom. Give an example that shows the result would not be true if we did not assume the Completeness Axiom. (The fact that our proof relied on the Completeness Axiom does answer the question, *why?*)

HINT: See Remark 3.4.2. □

REMARK 5.5.3. *Rolle's Theorem*, see Adams p140, 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 a particular case of the Mean Value Theorem.

We gave a direct proof of the Mean Value Theorem, whereas Adams first proved Rolle's Theorem. □

COROLLARY 5.5.4. *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, bounded or unbounded.) 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

$$|f(x_1) - f(x_2)| = |f'(c)|\,|x_1 - x_2| \leq K\,|x_1 - x_2|.$$

□

COROLLARY 5.5.5 (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 every $x \in I$.

But $|f(x) - f(c)| = 0$ by the previous corollary, and so $f(x) = C$. □

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.

A useful application of Corollary 5.5.5 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 f and g 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 $f(c) - g(c) = 0$.

The Mean Value Theorem leads to a result which enables us to decide where a function is increasing or decreasing.

DEFINITION 5.5.6. 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 $f(x_1) \geq f(x_2)$.

See Adams p139 for some diagrams.

THEOREM 5.5.7. *Suppose that f is continuous on an interval I and differentiable at every interior point of I . Then*

$f'(x) > 0$ for every interior point x implies f is increasing on I,

$f'(x) < 0$ for every interior point x implies f is decreasing on I,

$f'(x) \geq 0$ for every interior point x implies f is non-decreasing on I,

$f'(x) \leq 0$ for every interior point x implies f is non-increasing on I.

PROOF. Suppose $x_1 < x_2$ are points in I. 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. □

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 on the interval $\mathbb{R}$, but $f'(0) = 0$.

However, if f is increasing, or even just non-decreasing, on an interval, then it does follow that $f'(x) \geq 0$ for all x in the interval. This also follows from the Mean Value Theorem. (*Exercise*).

Note also that the first of the four cases in the theorem applies to $f(x) = x^3$ on the interval $[0, 1]$. *Why?*

5.6. ★Partial derivatives

Suppose, for simplicity, we have a function $f(x, y)$ defined on an open rectangle A, where

$$A = (a, b) \times (c, d) = \big\{(x, y) \in \mathbb{R}^2 : a < x < b,\ c < y < d\big\}.$$

Draw a diagram.

The *partial derivative with respect to y at $(x_0, y_0) \in A$* is defined by

$$\frac{\partial f}{\partial y}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0, y_0 + h) - f(x_0, y_0)}{h}.$$

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. *Draw a diagram.* 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 5.5.4 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 7.3.1.

CHAPTER 6

Integration

The main references in Adams are Sections 5.1–5.5 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.

6.1. Introduction

The topic of this chapter is the concept of "area under a graph" 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.

6.2. 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 section,[1] *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" 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.

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

[1] Some of the material in this section closely follows notes of Bob Bryce from a previous first year honours level course.

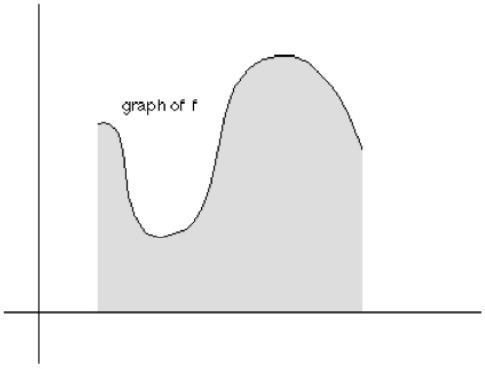

FIGURE 1. Suppose the domain of f is the interval $[a, b]$. Then $\int_a^b f$ is the area of the shaded region.

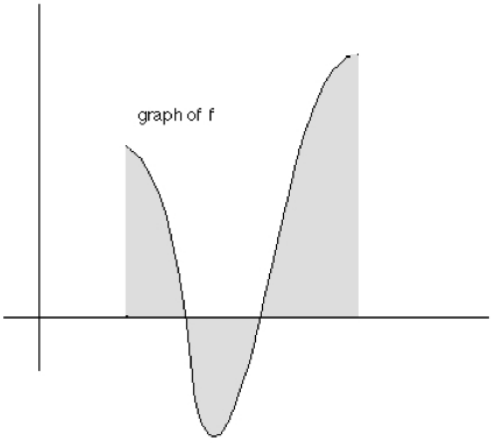

FIGURE 2. Suppose the domain of f is the interval $[a, b]$. Then $\int_a^b f$ is the area of the shaded region above the axis minus the area below the axis.

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 from $[a, b]$ which includes 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 $[a, b]$ is

$$P = \{a = x_0, x_1, x_2, \ldots, x_n = b\}.$$

We will assume always that $a = x_0 < x_1 < \ldots < x_n = b$. The ith subinterval is $[x_{i-1}, x_i]$ and its length is defined to be

$$\Delta x_i := x_i - x_{i-1}.$$

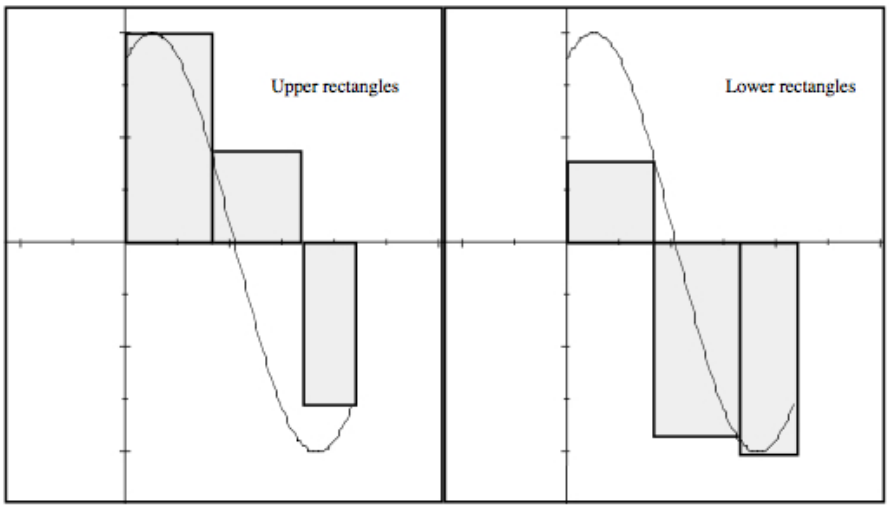

FIGURE 3. In this case the partition P divides the domain into three intervals. The upper sum $U(P, f)$ is the sum of the areas of the three rectangles, but with the first two counted positively and the third negatively. *What about* $L(P, f)$*?*

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

$$M_i = \max\{\, f(x) : x_{i-1} \le x \le x_i \,\}, \quad 1 \le i \le n;$$
$$m_i = \min\{\, f(x) : x_{i-1} \le x \le x_i \,\}, \quad 1 \le i \le n.$$

That is, M_i is the maximum value and m_i the minimum value of f on the ith 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 p289 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 $f(x) \ge 0$. In other cases the interpretation is less simple. Various possibilities are illustrated in the diagrams below.

EXAMPLE 6.2.1. Let $f(x) = 1 - 2x$ on $[0, 1]$, and let $P = \{0, 1/4, 1/3, 2/3, 1\}$. Find $L(P, f)$ and $U(P, f)$.

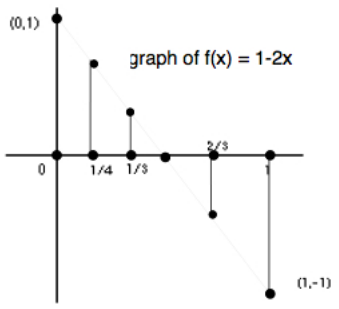

FIGURE 4. Sketch the graph of the function $f(x) = 1 - 2x$.

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

$$L(P,f) = \sum_{i=1}^{4} m_i\,\Delta x_i = \frac{1}{2}\cdot\frac{1}{4} + \frac{1}{3}\cdot\frac{1}{12} + \left(\frac{-1}{3}\right)\frac{1}{3} + \frac{-1}{3} = -\frac{7}{24}$$

$$U(P,f) = \sum_{i=1}^{4} M_i\,\Delta x_i = 1\cdot\frac{1}{4} + \frac{1}{2}\cdot\frac{1}{12} + \frac{1}{3}\cdot\frac{1}{3} + \left(-\frac{1}{3}\right)\frac{1}{3} = \frac{7}{24}$$

EXERCISE 6.2.2. 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 6.2.3. *Let f be a continuous function on $[a,b]$ and P be a partition of $[a,b]$. Then $L(P,f) \le U(P,f)$.*

PROOF. Since $m_i \le M_i$ for all i, and since $x_i - x_{i-1} > 0$,

$$m_i(x_i - x_{i-1}) \le M_i(x_i - x_{i-1}).$$

Summing,

$$L(P,f) \le U(P,f)$$

as required. □

Draw a diagram and you will see how obvious this result is.

The following lemma is obvious from a few diagrams. We need a proof that does not rely on particular examples. This is straightforward.

LEMMA 6.2.4. *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*

$$L(P_1,f) \le L(P_2,f) \quad \textit{and} \quad U(P_2,f) \le U(P_1,f).$$

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 $P_1 = \{a = x_0, x_1, x_2, \ldots, x_n = b\}$ and that $P_2 = P_1 \cup \{x\}$ [2] with $x \in (x_{i-1}, x_i)$. Let M_j, m_j $(1 \le j \le n)$ 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

$$m' \ge m_i, \qquad m'' \ge m_i$$

and

$$M' \le M_i, \qquad M'' \le M_i$$

because when passing from an interval to a subinterval, the minimum value cannot decrease and the maximum value cannot increase.

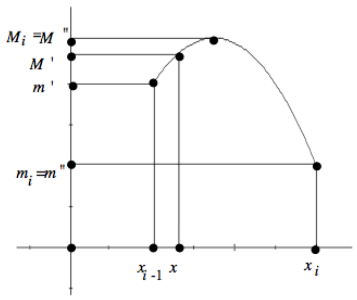

FIGURE 5. The minimum and maximum values of f on $[x_{i-1}, x]$ are m' and M' respectively, and on $[x, x_i]$ are m'' and M'' respectively.

Then

$$\begin{aligned} L(P_2,f) - L(P_1,f) &= m'(x - x_{i-1}) + m''(x_i - x) - m_i(x_i - x_{i-1}) \\ &\ge m_i(x - x_{i-1}) + m_i(x_i - x) - m_i(x_i - x_{i-1}) \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} U(P_1,f) - U(P_2,f) &= M_i(x_i - x_{i-1}) - M'(x - x_{i-1}) - M''(x_i - x) \\ &\ge M_i(x_i - x_{i-1}) - M_i(x_i - x) - M_i(x_i - x) \\ &= 0. \end{aligned}$$

That is

$$L(P_1,f) \le L(P_2,f) \quad \text{and} \quad U(P_2,f) \le U(P_1,f).$$

□

[2] This notation just means that P_2 is the union of the set P_1 and the set $\{x\}$ containing the single point x.

In words: *refining a partition increases lower sums and decreases upper sums.*

COROLLARY 6.2.5. *If f is continuous on $[a,b]$ and if P_1, P_2 are arbitrary partitions of $[a,b]$, then $L(P_1,f) \le U(P_2,f)$.*

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

$$L(P_1,f) \le L(P,f) \le U(P,f) \le U(P_2,f),$$

by Lemma 6.2.3 and Lemma 6.2.4. □

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 any 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

$$L\int_a^b f := \mathit{lub}\{\, L(P,f) : P \text{ is a partition of } [a,b]\,\}$$
$$U\int_a^b f := \mathit{glb}\{\, U(P,f) : P \text{ is a partition of } [a,b]\,\}$$

respectively.

The next lemma just uses the fact that every lower sum is $\le$ 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 6.2.6. *Let f be a continuous function on $[a,b]$. Then $L\int_a^b f \le U\int_a^b f$.*

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

$$L\int_a^b f \le U(P,f).$$

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

$$L\int_a^b f \le U\int_a^b f.$$

□

REMARK 6.2.7. ★ 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

$$M_i = \mathit{lub}\{\, f(x) : x_{i-1} \le x \le x_i \,\}, \quad 1 \le i \le n,$$
$$m_i = \mathit{glb}\{\, f(x) : x_{i-1} \le x \le x_i \,\}, \quad 1 \le i \le n.$$

Lemma 6.2.3, Lemma 6.2.4, Corollary 6.2.5 and Lemma 6.2.6 are still valid, with similar proofs as for continuous functions, but with "min" replaced by "*glb*" and "max" replaced by "*lub*".

[3]A function is bounded if there exist numbers A and B such that $A \le f(x) \le B$ 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 $x \ne 0$ and $f(0) = 0$, is not bounded on its domain $\mathbb{R}$.

DEFINITION 6.2.8. 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 the upper and lower integrals 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*.

REMARK 6.2.9. ★ 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 third year course on *measure theory* and *Lebesgue integration*.

The case we will be mainly interested in is when f is continuous. In Theorem 6.2.10 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 6.2.10. *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

$$U(P,f) - L(P,f) < \varepsilon. \tag{29}$$

Since f is *uniformly* continuous on $[a,b]$ by Theorem 4.5.2, there exists $\delta > 0$ such that

$$|x_1 - x_2| < \delta \Longrightarrow |f(x_1) - f(x_2)| < \frac{\varepsilon}{b-a}.$$

(We will see the reason for taking $\varepsilon/(b-a)$ in a moment.)

Now let $P = \{a = a_0, a_1, a_2, \ldots, a_N = b\}$ 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 of the function f on the ith interval must be $< \varepsilon/(b-a)$. Hence

$$\begin{aligned} U(P,f) - L(P,f) &= \sum_{i=1}^{N} M_i\,\Delta x_i - \sum_{i=1}^{N} m_i\,\Delta x_i \\ &= \sum_{i=1}^{N} (M_i - m_i)\,\Delta x_i \\ &< \frac{\varepsilon}{b-a}(\Delta x_1 + \cdots + \Delta x_N) \\ &= \frac{\varepsilon}{b-a}(b-a) = \varepsilon. \end{aligned}$$

This proves (29).

From the definition of the lower and upper integrals, and Lemma 6.2.6,

$$L(P,f) \le L\int_a^b f \le U\int_a^b f \le U(P,f).$$

Since the difference between the outer two terms is $< \varepsilon$ by (29), 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 $U\int_a^b f = L\int_a^b f$. □

6.3. 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

$$\begin{aligned} U(P,f) &= \sum_{i=1}^{n} f(u_i)\,\Delta x_i, \\ L(P,f) &= \sum_{i=1}^{n} f(l_i)\,\Delta x_i. \end{aligned}$$

where u_i and l_i are points in the ith 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

$$L(P,f) \le R(P,f) \le U(P,f). \tag{30}$$

Let the maximum length of the intervals in a partition P be denoted by $\|P\|$.

THEOREM 6.3.1.

$$\lim_{\|P\|\to 0} R(P,f) = \int_a^b f.$$

More precisely, for any $\varepsilon > 0$ there exists a number $\delta > 0$ (which may depend on ε) such that

$$\textit{whenever} \quad \|P\| < \delta \quad \textit{then} \quad \left| R(P,f) - \int_a^b f \right| < \varepsilon.$$

PROOF. The proof of Theorem 6.2.10 in fact showed that if $\|P\| < \delta$ then

$$U(P,f) - L(P,f) < \varepsilon.$$

Since

$$L(P,f) \le R(P,f) \le U(P,f)$$

and

$$L(P,f) \le \int_a^b f \le U(P,f)$$

it follows that

$$\left| R(P,f) - \int_a^b f \right| < \varepsilon.$$

□

NOTATION 6.3.2. We often use the notation

$$\int_a^b f(x)\,dx \quad \text{for} \quad \int_a^b f.$$

Note that $\int_a^b f(x)\,dx$ 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 rectangles of height $f(x)$ and "infinitesimal" width "dx". More precisely, from the previous theorem,

$$\int_a^b f(x)\,dx = \lim_{\|P\|\to 0} \sum_{i=1}^{N(P)} f(c_i)\,\Delta x_i.$$

(We write $N(P)$ to emphasise the fact that the number of points in the partition depends on P.)

6.4. Properties of the Riemann integral

The basic linearity and order properties of the Riemann integal are developed. The mean value theorem for integrals is proved. The extension of these results to piecewise continuous functions is noted.

REMARK 6.4.1. The (easy) theorems in this section apply more generally with minor modifications, provided the functions are Riemann integrable, and not necessarily continuous. For example, in Theorem 6.4.3 we need to replace the minimum m by the *glb* of the set of values, and similarly for the maximum M.

In particular, *piecewise continuous functions* (see Adams p309) on a closed bounded interval are integrable, and have the same properties as below. This essentially follows from writing each integral as a sum of integrals over intervals on which all the relevant functions are continuous.

THEOREM 6.4.2. *If f, g are continuous functions on $[a, b]$ and c, d are real numbers, then*

$$\int_a^b (cf + dg) = c\int_a^b f + d\int_a^b g, \tag{31}$$

$$f(x) \le g(x) \text{ for all } x \in [a, b] \implies \int_a^b f \le \int_a^b g, \tag{32}$$

$$\left|\int_a^b f\right| \le \int_a^b |f|. \tag{33}$$

PROOF. The main point in the proofs of (31) and (32) is that similar properties are true for the Riemann sums used to define the integrals. See Adams, A-30, Exercise 6.

To prove (33), note that

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

for all x. From (32)

$$\int_a^b -|f| \le \int_a^b f \le \int_a^b |f|.$$

From (31) this gives

$$-\int_a^b |f| \le \int_a^b f \le \int_a^b |f|.$$

This implies (33). □

THEOREM 6.4.3. *If f is continuous on $[a, b]$ with minimum and maximum values m and M then*

$$m(b - a) \le \int_a^b f \le M(b - a). \tag{34}$$

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

$$L(P, f) \le L\int_a^b f = \int_a^b f = U\int_a^b f \le U(P, f),$$

the result follows. □

THEOREM 6.4.4. *Suppose $a \le c \le b$. Then*

$$\int_b^a f = -\int_a^b f, \tag{35}$$

$$\int_a^a f = 0, \tag{36}$$

$$\int_a^c f + \int_c^b f = \int_a^b f. \tag{37}$$

PROOF. 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 of the integral as when the endpoints are distinct, except that now the points in any "partition" are all equal and so $\Delta x_i = 0$.

The third is straightforward. See **** for details. □

REMARK 6.4.5. If we allow $b \le a$ as well as $a < b$, then (33) should be replaced by

$$\left| \int_a^b f \right| \le \left| \int_a^b |f| \right| \tag{38}$$

Exercise.

THEOREM 6.4.6 (Mean Value Theorem for Integrals). *If f is continuous on $[a, b]$ then there exists $c \in [a, b]$ such that*

$$\int_a^b f = (b - a) f(c). \tag{39}$$

PROOF. Choose l and u to be minimum and maximum points for f on $[a, b]$. Then from (34) it follows that

$$f(l) \le \frac{\int_a^b f}{b - a} \le f(u),$$

By the Intermediate Value Theorem applied to the function f on the interval $[l, u]$ or $[u, l]$ (depending on whether $l \le u$ or $u \le l$), there exists c between l and u such that

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

This gives the result. □

6.5. 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 $\int_a^x f$ [4] as a function of the endpoint x (we allow $x \le a$ 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 $G(b) - G(a)$.

[4] In the theorem we 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 instead of t in this case, since we are already using x here to represent the endpoint of the interval of integration.

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 6.5.1 (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.

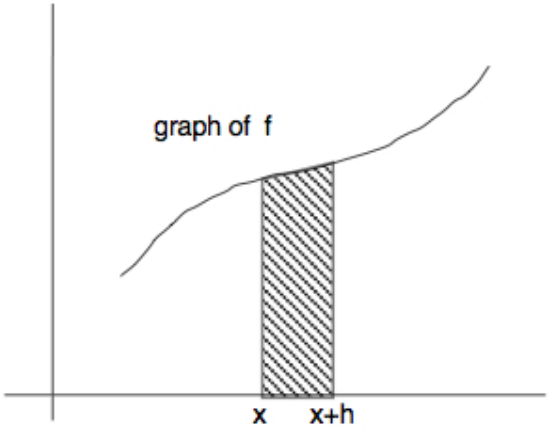

FIGURE 6. The area of the shaded region is $\int_x^{x+h} f$.

For the first assertion we have

$$\begin{aligned}
\frac{d}{dx}\int_a^x f &= \lim_{h\to 0}\frac{\int_a^{x+h} f - \int_a^x f}{h} \\
&= \lim_{h\to 0}\frac{\int_x^{x+h} f}{h} && \text{from (37)} \\
&= \lim_{h\to 0}\frac{hf(c(h))}{h} && \text{for some } c = c(h) \text{ between } x \text{ and } x+h, \\
& && \text{depending on } h, \text{ by the Mean Value Theorem} \\
& && \text{for integrals.} \\
&= \lim_{h\to 0} f(c(h)) \\
&= f(x) && \text{since } f \text{ is continuous at } x \\
& && \text{and } c \text{ lies between } x \text{ and } x+h.
\end{aligned}$$

For the second assertion, suppose $\dfrac{d}{dx}G(x) = f(x)$ on the interval I.

But we have just seen that $\dfrac{d}{dx}\displaystyle\int_a^x f = f(x)$. 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 5.5.5.

Setting $x = 0$ 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. □

CHAPTER 7

★Differential Equations

7.1. Overview

The differential equation

$$\frac{dy}{dx} = f(x, y) \tag{40}$$

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 (40), 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.

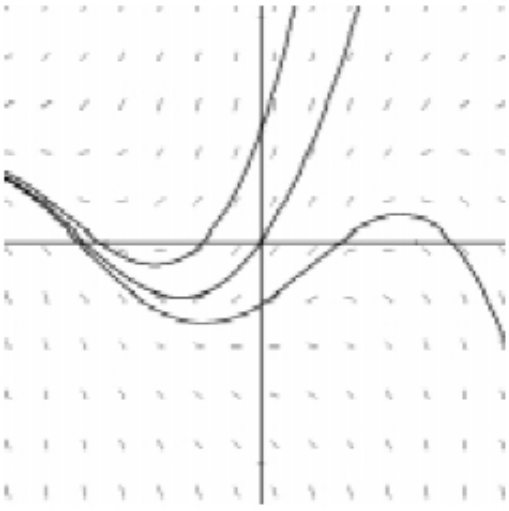

FIGURE 1. Slope field for the function $y + \cos x$, and the graph of three funcions $y = y(x)$ satisfying $\dfrac{dy}{dx} = y + \cos x$.

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.

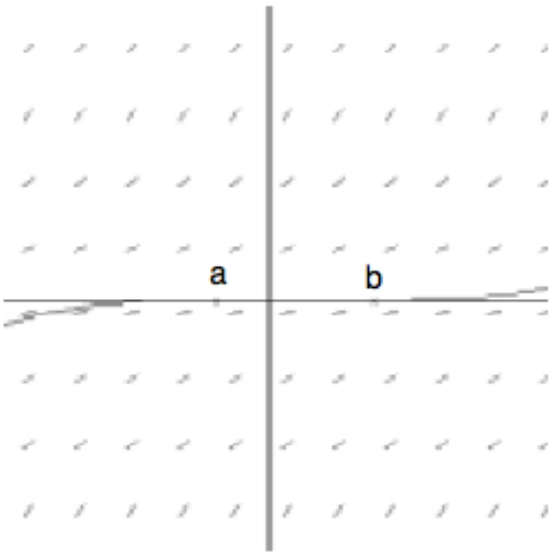

FIGURE 2. Slope field for the function $(x,y) \mapsto y^{2/3}$. (In this case the slope field does not depend on x.)

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 (40). 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 (40) 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 7.1.1. ★ A *system of first order differential equations* for the dependent variables $y_1, \ldots, y_n$ is a set of differential equations of the form

$$\begin{aligned}\frac{dy_1}{dx} &= f_1(x, y_1, \ldots, y_n)\\ \frac{dy_2}{dx} &= f_2(x, y_1, \ldots, y_n)\\ &\vdots\\ \frac{dy_n}{dx} &= f_n(x, y_1, \ldots, y_n)\end{aligned}$$

which are meant to be satisfied simultaneously by functions $y_1 = y_1(x), y_2 = y_2(x), \ldots, y_n = y_n(x)$. Here the functions $f_1, f_2, \ldots, f_n$ 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!.

7.2. 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 7.3 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

$$\begin{aligned}\frac{dy}{dx} &= f(x, y)\\ y(x_0) &= y_0\end{aligned} \tag{41}$$

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.

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 rectangle centred at (x_0, y_0), which passes through both of its 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 way[1] to the boundary of R.

[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.

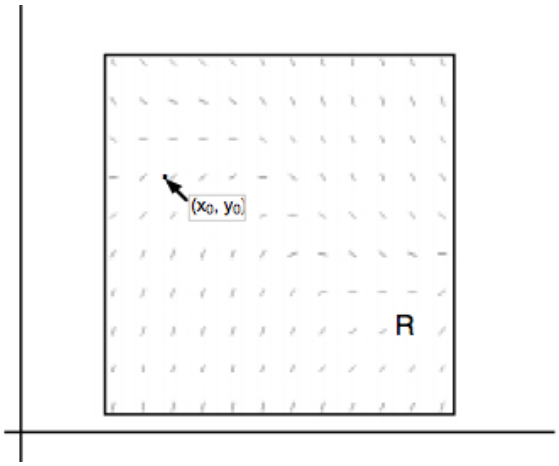

FIGURE 3. Slope field for f in the rectangle R.

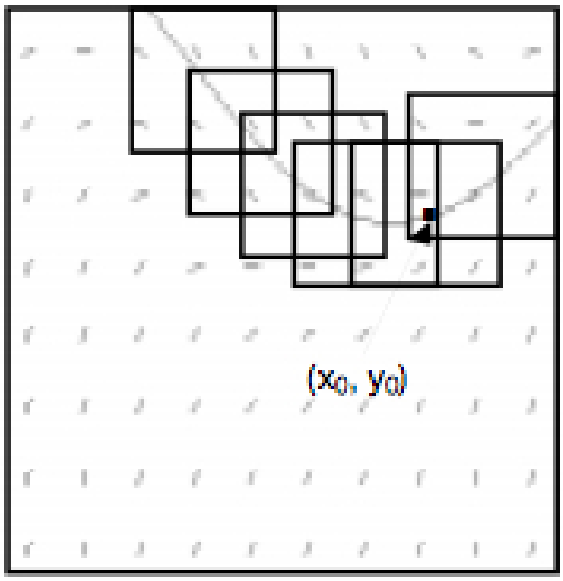

FIGURE 4. 6 small rectangles here get us all the way to the boundary of R. The point (x_0, y_0) is indicated by a small square black dot with an arrowhead below it.

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 7 steps.

Step A Problem (41) is equivalent to showing the "integral equation"

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\, dt \tag{42}$$

has a solution. One sees this by integrating both sides of (41) from x_0 to x. Conversely, differentiating the integral equation (42) gives back the differential equation, and clearly $y(x_0) = x_0$ also follows from the integral equation.

For our first example

$$\frac{dy}{dx} = y + \cos x, \quad y(0) = 0, \tag{43}$$

we get

$$y(x) = \int_0^x (y(t) + \cos t)\, dt. \tag{44}$$

Step B To find the solution of (42) we begin with the constant function

$$y(x) = y_0$$

and plug it into the right side of (42) 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 (43), substituting the constant function $y = 0$ in the right side of (44), and then repeating by plugging in the new function obtained after each step, we get

$$\begin{aligned}
\int_0^x \cos t\, dt &\longrightarrow \sin x \\
\int_0^x \Big(\sin t + \cos t\Big)\, dt &\longrightarrow -\cos x + 1 + \sin x \\
\int_0^x \Big(-\cos t + 1 + \sin t + \cos t\Big)\, dt &\longrightarrow x - \cos x + 1 \\
\int_0^x \Big(t - \cos t + 1 + \cos t\Big)\, dt &\longrightarrow \frac{1}{2}x^2 + x \\
\int_0^x \Big(\frac{1}{2}t^2 + t + \cos t\Big)\, dt &\longrightarrow \frac{1}{6}x^3 + \frac{1}{2}x^2 + \sin x \\
\int_0^x \Big(\frac{1}{6}t^3 + \frac{1}{2}t^2 + \sin t + \cos t\Big)\, dt &\longrightarrow \frac{1}{24}x^4 + \frac{1}{6}x^3 - \cos x + 1 + \sin x
\end{aligned}$$

We call this sequence of functions a "sequence of approximate solutions". We see from the diagram that this sequence is converging, at least near $(0, 0)$.

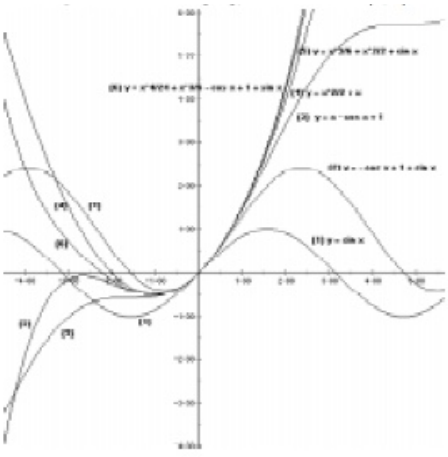

FIGURE 5. Apologies for the poor quality diagram.

In general, if $y_n(x)$ is the nth approximate solution, then

$$y_{n+1}(x) = y_0 + \int_{x_0}^x f(t, y_n(t))\, dt \tag{45}$$

is the $(n+1)$th approximate solution.

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 nth and $(n+1)$th approximate solutions is $< r$ times the distance between the $(n-1)$th and the nth 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

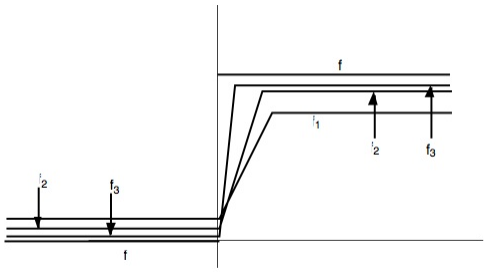

FIGURE 6. A sequence of continuous functions converging at each point to a discontinuous limit function.

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 (42). 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 $n \to \infty$ on both sides of (45) and deduce (42)

Step F The next step is to show that any two solutions are equal. We show that if d is the distance between two solutions of (42) 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.

Step G The final step is to extend the solution from the small rectangle in step C up to the boundary of R. We do this by essentially starting the process again at a new point on the graph near the boundary of the small rectangle, getting a new rectangle centred at the new point, and extending the solution out into the new rectangle. We can show this process stops after a finite number of steps, when the solution reaches the boundary of R.

7.3. ★Rigorous proof of the Existence and Uniqueness theorem

THEOREM 7.3.1. *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*

$$\phi'(x) = f(x, \phi(x)) \tag{46}$$

$$\phi(x_0) = y_0. \tag{47}$$

In other words, $\phi(x)$ *solves (i.e. satisfies) the initial value problem*

$$\frac{dy}{dx} = f(x, y)$$
$$y(x_0) = y_0$$

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.

$$R_\delta(x_0, y_0) = \left\{ (x,y) : x \in (x_0 - \delta, x_0 + \delta),\ y \in (y_0 - M\delta, y_0 + M\delta) \right\}, \tag{48}$$

then any $\delta > 0$ for which $R_\delta(x_0, y_0) \subset R$ and $\delta < K^{-1}$, will work for the above theorem.

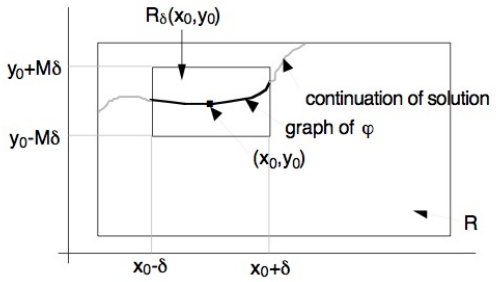

FIGURE 7. The rectangle R_δ from (48). (The small "wiggles" in the graph of ϕ should not be present. The shafts of the arrow heads are barely discernible.)

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. (46) and (47) are true;
(2) $\phi(x)$ satisfies the integral equation

$$\phi(x) = y_0 + \int_{x_0}^{x} f(t, \phi(t))\, dt. \tag{49}$$

Assume the *first* statement is true. Then both $\phi'(t)$, and $f(t, \phi(t))$ by Section 3.5, 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 (46) 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 3.5, and so the integral *does* exist. Setting $x = x_0$ we immediately get (47)

Since the right side of (49) *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, (46) 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 (49) 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

$$R_\delta(x_0, y_0) \subset R, \tag{50}$$

where $R_\delta(x_0, y_0)$ was defined in (48).

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 fnction of t from Section 3.5, 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

$$\begin{aligned} |\phi_1(x) - y_0| &= \left| \int_{x_0}^{x} f(t, \phi_0(t))\, dt \right| \\ &\leq \left| \int_{x_0}^{x} |f(t, \phi_0(t))|\, dt \right| && \text{from (38)} \\ &\leq \left| \int_{x_0}^{x} M\, dt \right| && \text{since } |f| \leq M \text{ in } R \\ &\leq M\delta && \text{since } |x - x_0| \leq \delta. \end{aligned}$$

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

$$\phi_n(x) \to \phi(x)$$

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]

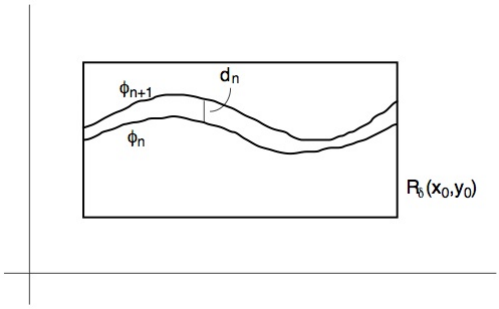

FIGURE 8. d_n is the distance between the functions ϕ_n and ϕ_{n+1}.

[2]We should be a little more careful 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 "*lub*" in Step C.

Then for $n \geq 1$

$$\begin{aligned}
d_n &= \max_{x\in(x_0-\delta,x_0+\delta)} |\phi_n(x) - \phi_{n+1}(x)| \\
&= \max_{x\in(x_0-\delta,x_0+\delta)} \left| \int_{x_0}^{x} f(t,\phi_{n-1}(t)) - f(t,\phi_n(t))\, dt \right| \\
&\leq \max_{x\in(x_0-\delta,x_0+\delta)} \left| \int_{x_0}^{x} \left|f(t,\phi_{n-1}(t)) - f(t,\phi_n(t))\right| dt \right| \\
&\leq \max_{x\in(x_0-\delta,x_0+\delta)} \left| \int_{x_0}^{x} K\left|\phi_{n-1}(t) - \phi_n(t)\right| dt \right| && \text{by Section 5.6} \\
&\leq \max_{x\in(x_0-\delta,x_0+\delta)} \left| \int_{x_0}^{x} K d_{n-1}\, dt \right| && \text{from the definition of } d_{n-1} \\
&= K\delta\, d_{n-1}
\end{aligned}$$

Repeating this argument we obtain

$$d_n \leq K\delta\, d_{n-1} \leq (K\delta)^2 d_{n-2} \leq (K\delta)^3 d_{n-3} \leq \cdots \leq (K\delta)^n d_0.$$

We now make the further restriction on δ that

$$K\delta < 1. \tag{51}$$

Since

$$|\phi_n(x) - \phi_{n+1}(x)| \leq d_n \leq d_0(K\delta)^n,$$

it follows from Theorem 2.7.4 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

$$\phi(x) = \lim_{n\to\infty} \phi_n(x).$$

Moreover, by the commt following that theorem,

$$|\phi_n(x) - \phi(x)| \leq Ar^n, \tag{52}$$

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, see Figure 9.)

Step D (See Figure 9.) 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 (52)

$$x \in (x_0 - \delta, x_0 + \delta) \quad \text{implies} \quad |\phi_n(x) - \phi(x)| \leq \varepsilon/3. \tag{53}$$

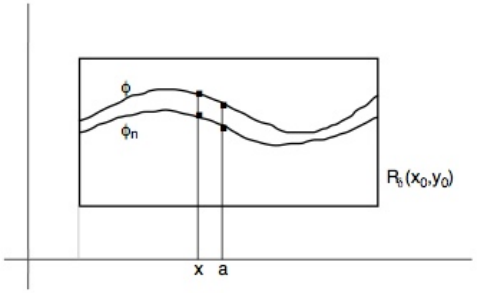

FIGURE 9. Figure for Step D. (The small squiggles in the graph are the fault of the graphics program!)

By continuity of ϕ_n there exists $\eta > 0$ (which may depend on n and hence on ε) such that

$$|x - a| < \eta \quad \text{implies} \quad |\phi_n(x) - \phi_n(a)| < \varepsilon/3. \tag{54}$$

(We also choose η sufficiently small that if $|x - a| < \eta$ then $x \in (x_0 - \delta, x_0 + \delta)$.

From (53) (applied with x and again with x replaced by a) and (54) it follows that if $|x - a| < \eta$ then

$$\begin{aligned} |\phi(x) - \phi(a)| &= |(\phi(x) - \phi_n(x)) + (\phi_n(x) - \phi_n(a)) + (\phi_n(a) - \phi(a))| \\ &\le |\phi(x) - \phi_n(x)| + |\phi_n(x) - \phi_n(a)| + |\phi_n(a) - \phi(a)| \\ &\le \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

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

$$\phi_{n+1}(x) = y_0 + \int_{x_0}^{x} f(t, \phi_n(t))\, dt. \tag{55}$$

We have shown in Step C that[3]

$$\phi_{n+1}(x) \to \phi(x)$$

for each x in the interval $(x_0 - \delta, x_0 + \delta)$. We next *claim* that for the right side of (54) we have

$$y_0 + \int_{x_0}^{x} f(t, \phi_n(t))\, dt \to y_0 + \int_{x_0}^{x} f(t, \phi(t))\, dt.$$

It then follows from the *claim* that

$$\phi(x) = y_0 + \int_{x_0}^{x} f(t, \phi(t))\, dt,$$

which establishes (49) and hence proves the theorem by Step A.

To prove the *claim* we compute

$$\begin{aligned} \left| \int_{x_0}^{x} f(t, \phi_n(t))\, dt - \int_{x_0}^{x} f(t, \phi(t))\, dt \right| &\le \left| \int_{x_0}^{x} \big| f(t, \phi_n(t)) - f(t, \phi(t)) \big|\, dt \right| \\ &\le \left| \int_{x_0}^{x} K \big| \phi_n(t) - \phi(t) \big|\, dt \right| \qquad \text{by Section 5.6} \\ &\le \left| \int_{x_0}^{x} K A r^n\, dt \right| \qquad \text{by (52)} \\ &\le K \delta A r^n. \end{aligned}$$

Since $0 \le r < 1$ this establishes the *claim* and hence the theorem.

Step F Next, we must show that any two solutions of (46) and (47), or equivalently of (49), are equal.

Suppose that $\phi(x)$ and $\psi(x)$ are any two solutions. Then if

$$d = \max |\phi(x) - \psi(x)|,$$

[3] If $a_n \to a$ for a sequence, then it follows that $a_{n+1} \to a$, *why?*

where the maximum is taken over the interval $(x_0 - \delta, x_0 + \delta)$.[4] Then for any $x \in (x_0 - \delta, x_0 + \delta)$,

$$\begin{aligned} |\phi(x) - \psi(x)| &= \left| \int_{x_0}^{x} \big(f(t, \phi(t)) - f(t, \psi(t))\big)\, dt \right| \\ &\leq \left| \int_{x_0}^{x} \big|f(t, \phi(t)) - f(t, \psi(t))\big|\, dt \right| \\ &\leq \left| \int_{x_0}^{x} K|\phi(t)) - \psi(t)|\ \ dt \right| \qquad \text{from Section 5.6} \\ &\leq K\delta d \end{aligned}$$

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 $x \in (x_0 - \delta, x_0 + \delta)$. □

Step G So far we have a solution, and it is unique, whose graph lies in a rectangle R_δ. The dimensions of R_δ *depend only on* K *an* M, and otherwise not on the initial point, except that we also require $R_\delta \subset R$. By starting the process again at a new point close to the boundary of R_δ we can extend the solution outside R_δ, and after a finite number of steps extend the solution up to the boundary of R.

End of proof, end of chapter, end of semester. Have a good holiday.

[4] As in Step C we should really write "*lub*" instead of "max". The proof is essentially unchanged.