An Introduction to Contemporary Mathematics

John Hutchinson
(suggestions and comments to:
John.Hutchinson@anu.edu.au)

March 21, 2010

Mathematical Sciences Institute
College of Science
Australian National University

A text for the ANU secondary college course
"An Introduction to Contemporary Mathematics"

I wish to dedicate this text:

- *to the memory of my father George Hutchinson and to my mother Ellen Hutchinson for their moral and financial support over many years of my interest in mathematics;*
- *to my mentor Kevin Friel for being such an inspirational high school teacher of mathematics;*
- *and to my partner and wife Malise Arnstein for her unflagging support and encouragement, despite her insight from the beginning that this project was going to take far more time than I ever anticipated.*

Contents

Introduction

For Whom are these Notes?

These notes, together with the book *The Heart of Mathematics* [HM] by Burger and Starbird, are the texts for the ANU College Mathematics Minor for Years 11 and 12 students. If you are doing this course you will have a strong interest in mathematics, and probably be in the top 5% or so of students academically.

What is Mathematics?

Mathematics is the study of pattern and structure. Mathematics is fundamental to the physical and biological sciences, engineering and information technology, to economics and increasingly to the social sciences.

The patterns and structures we study in mathematics are universal. It is perhaps possible to imagine a universe in which the biology and physics are different, it is much more difficult to imagine a universe in which the mathematics is different.

Philosophy of this Course

The goal is to introduce you to contemporary mainstream 20th and 21st century mathematics.

This is not an easy task. Mathematics is like a giant scaffolding. You need to build the superstructure before you can ascend for the view. The calculus and algebra you will learn in college is an essential part of this scaffolding and is fundamental for your further mathematics, but most of it was discovered in the 18th century.

We will take a few short cuts and only use calculus later in this course. We will investigate some very exciting and useful modern mathematics and get a feeling for "what mathematics is all about". The mathematics you will see in this course is usually not seen until higher level courses in second or third year at University.

Of course, you will not cover the mathematics in the same depth or generality as you will if you pursue mathematics as a part of your University studies (as I hope most of you will do). The way we will proceed is by studying carefully chosen parts and representative examples from various areas of mathematics which illustrate important and general key concepts. In the process you will

gain a real understanding and feeling for the beauty, utility and breadth of mathematics.

These Notes and *The Heart of Mathematics*

[HM] is an excellent book. It is one of a small number of texts intended to give you, the reader, a feeling for the theory and applications of contemporary mathematics at an early stage in your mathematical studies. However, [HM] is directed at a different group of students — undergraduate students in the United States with little mathematics background (e.g. no calculus) who might take no other mathematics courses in their studies.

Despite its apparently informal style, [HM] develops a significant amount of interesting contemporary mathematics. The arguments are usually complete (and if not, this is indicated), correct and well motivated. They are often done by means of studying particular but important examples which cover the main ideas in the general case.

However, you might find that the language is a little verbose at times (and you may or may not find the jokes tedious!). After first studying the arguments in [HM] you may then find the more precisely written mathematical arguments in these Notes more helpful in understanding "how it all hangs together".

So here is a suggested procedure:

1. Look very briefly at these notes both to see what parts of [HM] you should study and to gain an overview.
2. Study (= read, think about, cogitate over) the relevant section in [HM].
3. Then study the relevant section in these Notes.

You may want to change the order, do what is best for you.

In the Notes we:

- Follow the same chapter and section numbering as in [HM]
- Discuss and extend the material in [HM] and fill in some gaps
- Often write out more succinct and general arguments
- Indicate which parts of [HM] are to be studied and sometimes recommend questions to attempt
- Include some more difficult and challenging questions

What is Covered in this Course?

There are four parts to the course. Each will take approximately 1.5 terms. You will study the first 2 parts in terms 2,3,4 of year 11 and the second 2 parts in terms 1,2,3 of year 12.

Part 1 *An introduction to number theory and its application to cryptography.* Essentially Chapter 2 from [HM] and supplementary material from these Notes. The RSA cryptography we discuss is essential to internet security and the method was discovered in 1977. The 3 mathematicians involved started a company which they sold for about $600,000,000(US).

Part 2 *A Hierarchy of Infinities.* Essentially Chapter 3 from [HM] and supplementary material from these Notes. What is infinity? Can one infinite

set be larger than another (Yes). If you remove 23 objects from an infinite set is the resulting set "smaller" (No). These ideas are interesting, but are they important or useful? (Yes).

Part 3 *Dynamical Processes, Chaos and Fractals.* Modelling change by dynamical processes, how chaos can arise out of simple processes, how fractal sets have fractional dimensions. Some of the ideas here on fractals were first developed by the present writer (iterated function systems) and other ideas (the chaos game) by another colleague now at the ANU, Michael Barnsley. Barnsley applied these ideas to image compression and was a founder of the company "Iterated Systems", at one stage valued at $200,000,000(US), later known as "Media Bin" and then acquired by "Interwoven".

Part 4 *Geometry and Topology.* Parts of Chapters 4 and 5 from [HM] and supplementary material from these Notes. Platonic solids, visualising higher dimensions, topology, classifying surfaces, and more. This is beautiful mathematics and it is fundamental to our understanding of the universe in which we live — some current theories model our universe by 10 dimensional curved geometry

I suggest you also

- read ix–xiv of [HM] in order to understand the philosophy of that book;
- read xv–xxi of [HM] to gain an idea of the material you will be investigating over the next 2 years.

Studying Mathematics

This takes time and effort but it is very interesting material and intellectually rewarding. Do lots of Questions from [HM] and from these Notes, answer the questions here marked with a  and keep your solutions and comments in a folder.

Material marked $\star$ is not in [HM] and is more advanced. Some is a little more advanced and some is a lot more advanced. It is included to give you an idea of further connections. Don't worry if it does not make complete sense or you don't fully understand. Just relax and realise it is not examinable, except in those cases where your teacher specifically says so, in which case you will also be told how and to what extent it is examinable.

Acknowledgements

I would like to thank Richard Brent, Tim Brook, Clare Byrne, Jonathan Manton, Neil Montgomery, Phoebe Moore, Simon Olivero, Raiph McPherson, Jeremy Reading, Bob Scealy, Lisa Walker and Chris Wetherell, for comments and suggestions on various drafts of these notes.

Quotations

Philosophy is written in this grand book—I mean the universe— which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics, and its characters are triangles, circles, and other mathematical figures, without which it is humanly impossible to understand a single word of it; without these one is wandering about in a dark labyrinth.

Galileo Galilei *Il Saggiatore* [1623]

Life is good for only two things, discovering mathematics and teaching mathematics.[1]

Siméon Poisson [1781-1840]

Mathematics is the queen of the sciences.

Carl Friedrich Gauss [1856]

Mathematics takes us still further from what is human, into the region of absolute necessity, to which not only the actual world, but every possible world, must conform.

Bertrand Russell *The Study of Mathematics* [1902]

Mathematics, rightly viewed, possesses not only truth, but supreme beauty — a beauty cold and austere, like that of a sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of perfection such as only the greatest art can show.

Bertrand Russell *The Study of Mathematics* [1902]

The science of pure mathematics, in its modern developments, may claim to be the most original creation of the human spirit.

Alfred North Whitehead *Science and the Modern World* [1925]

All the pictures which science now draws of nature and which alone seem capable of according with observational facts are mathematical pictures .... From the intrinsic evidence of his creation, the Great Architect of the Universe now begins to appear as a pure mathematician.

Sir James Hopwood Jeans *The Mysterious Universe* [1930]

[1] Simeon Poisson was the thesis adviser of the thesis adviser of ... of my thesis adviser, back 9 generations. See `www.genealogy.math.ndsu.nodak.edu` . I do not agree with Poisson's statement!

The language of mathematics reveals itself unreasonably effective in the natural sciences..., a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure even though perhaps to our bafflement, to wide branches of learning.

Eugene Wigner [1960]

The same pathological structures that mathematicians invented to break loose from 19th naturalism turn out to be inherent in familiar objects all around us in nature.

Freeman Dyson *Characterising Irregularity,* Science 200 [1978]

Mathematics is like a flight of fancy, but one in which the fanciful turns out to be real and to have been present all along. Doing mathematics has the feel of fanciful invention, but it is really a process for sharpening our perception so that we discover patterns that are everywhere around. ... To share in the delight and the intellectual experience of mathematics – to fly where before we walked – that is the goal of mathematical education.

One feature of mathematics which requires special care ... is its "height", that is, the extent to which concepts build on previous concepts. Reasoning in mathematics can be very clear and certain, and, once a principle is established, it can be relied upon. This means that it is possible to build conceptual structures at once very tall, very reliable, and extremely powerful. The structure is not like a tree, but more like a scaffolding, with many interconnecting supports. Once the scaffolding is solidly in place, it is not hard to build up higher, but it is impossible to build a layer before the previous layers are in place.

William Thurston Notices Amer. Math. Soc. [1990]

Chapter 1

Fun and Games

In this Chapter in [HM, §1.1] there are 9 puzzles/questions — most are a "lead in" to topics in later chapters. The relevant ones for us are *[HM, 2–28]*

Story 3	Part 1 of Course
Story 5	Part 2
Story 2 & 4	Part 3
Story 6	Part 4

In [HM, §1.2] there are some gentle hints. You will learn more if you do not look at the hints until after you have expended some real thought on the questions.

In [HM, §1.3] the solutions are given and discussed.

Chapter 2

Numbers and Cryptography

Important Note The material in the Notes corresponds to and often extends that in *The Heart of Mathematics* [HM]. See also the comments on page iv. The corresponding page numbers in [HM] are noted here in the margin. First study the material in [HM], then study the more concentrated and extended treatment here.

Additional material beyond that in [HM] is noted as such in the margin, and is not necessarily a required part of the course. Your teacher will let you know.

In any case I hope you look at this additional material. It is there to set the course in a broader context, to indicate future directions, to introduce important techniques and methods, and to provide some additional challenges!

Similar remarks apply to the other Chapters in these Notes.

Contents

[1] Anything marked with ⋆ is either not in [HM] or is only treated lightly there, and is more advanced material. Some is a little more advanced and some is a lot more advanced. It is included to give you an idea of further connections. Don't worry if it does not make complete sense or you don't fully understand. Just relax and realise it is not examinable, except in those cases where your teacher specifically says so, in which case you will also be told how and to what extent it is examinable.

2.1 Counting[2]

Using estimation to move from qualitative to quantitative thinking and reasoning is a powerful tool.

Overview

The Pigeon Hole Principle is used in [HM, §2.1] to show that *there are at least 2 people on the earth with exactly the same number of hairs on their body.*

A whimsical argument is also given to show that all natural numbers are "interesting", or perhaps more accurately to show that "interesting" is not a well defined mathematical concept. This argument is essentially the *Principle of Mathematical Induction*, which we will discuss later.

Types of Numbers

[HM, 39–41]

Natural Numbers and Integers For future reference we note:

Definition. The *natural numbers* are the numbers $1, 2, 3 \ldots$. The *integers* are the numbers $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$.

Real Numbers and Their Properties Later we will discuss in some detail the *real numbers*, often just called *numbers*.[3] The real numbers include the integers and in particular the natural numbers.

At this stage we will use the usual properties of addition, multiplication, subtraction and division for the real numbers such as:

- $x + y = y + x$ and $x(y + z) = xy + xz$ for any real numbers x, y, z;
- the properties of 0 and 1 such as $x + 0 = x$ and $x \times 1 = x$ for any number x, and that for any real number x there is another real numbers written $-x$ such that $x + (-x) = 0$;
- the properties of inequalities such as $x < y$ implies $x + z < y + z$ for any numbers x, y, z.

The natural numbers are the real numbers $1, 1 + 1, 1 + 1 + 1, \ldots$, which we write as $1, 2, 3, \ldots$. The integers are the real numbers $\ldots, -(1 + 1 + 1), -(1 + 1), -1, 0, 1, 1 + 1, 1 + 1 + 1, \ldots$ which we write as $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$.

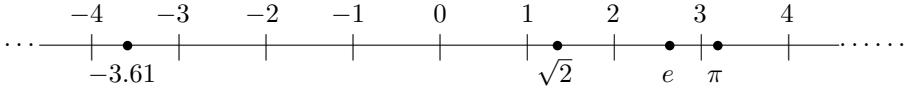
We will also use all the standard properties of the natural numbers and the integers such as the sum and product of two natural numbers is a natural number.

It is common to use symbols like i, j, k, m, n, N to denote natural numbers and symbols like x, y, z, u, v to denote real numbers in general.

[2] The epigrams in each Section are from [HM] and its supporting material.

[3] Even later we will also discuss *complex* numbers, which involve the square root of -1.

Geometric Representation of Numbers Sometimes it helps to think of real numbers as being represented by points on an infinite straight line as follows:



Most numbers do not have simple names as do $\sqrt{2}$, e and π.

There is nothing special about -3.61.

The number $\sqrt{2}$ is $1.41421356237309504880168872420 9\ldots$ to 30 decimal places, and is the number which when multiplied by itself gives 2.

The number π is the ratio of the circumference of a circle to its diameter and is $3.14159265358979323846264338327 9\ldots$ to 30 decimal places.

The number e is one of the most important numbers in mathematics and is $2.71828182845904523536028747135 2\ldots$ to 30 decimal places. You will come across it later when you study calculus. It arises naturally in the study of logarithms, in growth and decay models, even in understanding compound interest[4].

The Pigeon Hole Principle

[HM, 41–43]

The following simple result has interesting and often surprising conclusions.

Theorem 2.1.1. *If N objects are put into n boxes and $N > n$, then at least one box will contain more than one object.*

Proof. [5] *Assume* no box has more than one object in it. Since the number of boxes is n this implies there are at most n objects. But we know there are N objects and N is greater than n.

This contradiction implies the assumption is false. Hence at least one box has *more* than one object in it. □

The idea is that if you have more pigeons than pigeon holes, then at least one pigeon hole must contain more than one pigeon.

---

[4] If you take \$1 and let it earn 100% interest you will have \$2 after a year.

If you calculate the interest each 6 months you will have $\$(1+\frac{1}{2})$ after 6 months and then $\$(1+\frac{1}{2})^2 = \$2.25$ after a year.

Why?

If you calculate the interest every month you will have $\$(1+\frac{1}{12})$, $\$(1+\frac{1}{12})^2$, and $\$(1+\frac{1}{12})^3$ after each of the first 3 months, and finally $\$(1+\frac{1}{12})^{12} \approx \$2.61$ after a year.

Why?

If you calculate the interest every week (supposing there are exactly 52 weeks in the year) you will have $\$(1+\frac{1}{52})^{52} \approx \$2.69$ after a year.

If you calculate the interest every day (supposing there are exactly 365 days in the year) you will have $\$(1+\frac{1}{365})^{365} \approx \$2.7146$.

And as you compound more and more frequently the number of dollars you have after a year will *not* increase without bound, but will instead get closer and closer to the number e.

[5] Later we will discuss more carefully what is meant by a "proof". In particular we will discuss what one can assume and what methods of argument one can use. At this stage by a "proof" we mean essentially an argument which uses only (i) basic properties of numbers including those about addition, multiplication, and inequalities; (ii) facts we may have previously proved; and (iii) logical reasoning.

The Theorem was proved by assuming it to be false and from this deriving a contradiction. This method of *proof by contradiction* is a very powerful one in Mathematics.[6]

[HM] uses the Pigeon Hole principle to show that at least two people on the earth are equally hairy!

In Questions 4 and 5 we give two tricky applications, *with* Hints.

⋆*The Principle of Mathematical Induction*[7]

[HM, 43–45]

This is only discussed in [HM] in a very light way, to show that "all numbers are interesting". Here we discuss and give some more serious examples.

Sum of First n Natural Numbers You may know the formula

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}. \tag{2.1}$$

One way to prove this is to write

$$\begin{aligned} S &= 1 + \quad 2 \quad + \quad 3 \quad + \cdots + n - 1 + n, \\ \therefore S &= n + (n-1) + (n-2) + \cdots + \quad 2 \quad + 1, \end{aligned}$$

by reversing the order of addition. Adding first terms together, second terms together, etc.,

$$\begin{aligned} 2S &= (1+n) + (2+n-1) + (3+n-2) + \cdots + (n-1+2) + (n+1) \\ &= (1+n) + (1+n) + (1+n) + \cdots + (1+n) + (1+n) \quad (n \text{ times}) \\ &= n(n+1). \end{aligned}$$

It follows that $S = n(n+1)/2$.

For example,

$$1 + 2 + 3 + \cdots + 100 = 100 \cdot 101/2 = 5050.$$

[6] Sometimes students tend to overuse *proof by contradiction*. There is no logical reason not to use it as often as you like, after all it is certainly a valid method of proof. However a direct proof, if it is not too long, will usually give someone a better idea and more insight as to "why" a Theorem is true.

[7] Anything marked with ⋆ is either not in [HM] or is only treated lightly there, and is more advanced material. Some is a little more advanced and some is a lot more advanced. It is included to give you an idea of further connections. Don't worry if it does not make complete sense or you don't fully understand. Just relax and realise it is not examinable, except in those cases where your teacher specifically says so, in which case you will also be told how and to what extent it is examinable.

Sum of First n Squares, Cubes, etc. Here are some formulae.

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6} = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}, \tag{2.2}$$

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4} = \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4}, \tag{2.3}$$

$$1^4 + 2^4 + 3^4 + \cdots + n^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30} = \frac{n^5}{5} + \frac{n^4}{2} + \frac{n^3}{3} - \frac{n}{30}. \tag{2.4}$$

Suppose we were able to guess one of these formulae by a bit of trial and error, or perhaps someone told you that they saw one of the formulae somewhere. You can readily check that it is true for $n = 1$ and $n = 2$. But is there a systematic way of proving it for every n?

The answer is YES, and it is by the method of Mathematical Induction, which we now state and prove.

Statement & Proof of Induction

Theorem 2.1.2 (Principle of Mathematical Induction)**.** *Let $P(n)$ be a statement about n, for each natural number n. Suppose we know:*

1. *$P(1)$ is true,* ***(basic step)***
2. *Whenever $P(k)$ is true for a natural number k, it follows that $P(k+1)$ is also true.* ***(inductive step)***

Then the statement $P(n)$ is true for every *natural number n.*

Proof.

- By the first assumption, $P(1)$ is true.
- By the second assumption, since $P(1)$ is true it follows that $P(2)$ is true.
- By the second assumption, since $P(2)$ is true it follows that $P(3)$ is true.
- By the second assumption, since $P(3)$ is true it follows that $P(4)$ is true.
- By the second assumption, since $P(4)$ is true it follows that $P(5)$ is true.
- etc.

In this way we see that for *every* natural number n, $P(n)$ is true. □

Remark. This is more of an informal justification than a proof, essentially because of the "etc.". In fact, some form of the Principle of Mathematical Induction is usually taken as one of the *axioms* of arithmetic.

Application to Sums of Squares, Cubes, etc. We now use mathematical induction to prove (2.3).

Solution. Let $P(n)$ be the statement

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4}. \tag{2.5}$$

In order to show that $P(n)$ is true for all natural numbers n, we need to show:

1. $P(1)$ is true. **(basic step)**
2. Whenever $P(k)$ is true then $P(k+1)$ is true. **(inductive step)**

Clearly $P(1)$ is true since both sides of (2.5) then equal 1. This means we have shown the *basic step*.

Next assume $P(k)$ is true for some natural number k, i.e.

$$1^3 + 2^3 + 3^3 + \cdots + k^3 = \frac{k^4}{4} + \frac{k^3}{2} + \frac{k^2}{4}. \tag{2.6}$$

We want to show it follows that $P(k+1)$ is true. In other words, we want to show it follows that

$$1^3 + 2^3 + 3^3 + \cdots + k^3 + (k+1)^3 = \frac{(k+1)^4}{4} + \frac{(k+1)^3}{2} + \frac{(k+1)^2}{4} \tag{2.7}$$

Here is the argument:

$$\begin{aligned}
&1^3 + 2^3 + 3^3 + \cdots + k^3 + (k+1)^3 \\
&\quad = \frac{k^4}{4} + \frac{k^3}{2} + \frac{k^2}{4} + (k+1)^3 \text{ because we assumed } P(k) \text{ is true} \\
&\quad = \frac{k^4}{4} + \frac{k^3}{2} + \frac{k^2}{4} + (k^3 + 3k^2 + 3k + 1) \text{ check it!} \\
&\quad = \frac{k^4}{4} + \frac{3k^3}{2} + \frac{13k^2}{4} + 3k + 1 \\
&\quad = \frac{(k+1)^4}{4} + \frac{(k+1)^3}{2} + \frac{(k+1)^2}{4} \text{ check it!}
\end{aligned}$$

which is what we wanted to show. This means we have shown the *inductive step*, since we have shown $P(k)$ implies $P(k+1)$ for every k.

It now follows from the Principle of Mathematical Induction that $P(n)$ is true for all natural numbers n. □

⋆*Finding the Sum of First n Squares, Cubes, etc.*

We saw in the previous Section how to prove the formulae for the sum of the first n squares, cubes etc. But is there a systematic way for finding these formulae in the first case? Yes, and here is how to do it.

For the sum $1^2 + 2^2 + 3^2 + \cdots + n^2$ we use the formula

$$k^3 - (k-1)^3 = 3k^2 - 3k + 1,$$

which you should check.

Setting $k = 1$, $k = 2$, $k = 3$, $\ldots$, $k = n$ we get

$$\begin{aligned}
1^3 - 0^3 &= 3 \times 1^2 - 3 \times 1 + 1 \\
2^3 - 1^3 &= 3 \times 2^2 - 3 \times 2 + 1 \\
3^3 - 2^3 &= 3 \times 3^2 - 3 \times 3 + 1 \\
4^3 - 3^3 &= 3 \times 4^2 - 3 \times 4 + 1 \\
&\vdots \\
n^3 - (n-1)^3 &= 3 \times n^2 - 3 \times n + 1
\end{aligned}$$

Add all this together and notice how on the left the terms 1^3 and -1^3 cancel, as do 2^3 and -2^3, 3^3 and -3^3, etc. This gives

$$\begin{aligned} n^3 - 0^3 &= 3(1^2 + 2^2 + 3^2 + \cdots + n^2) - 3(1 + 2 + 3 + \cdots + n) \\ &\quad + (1 + 1 + 1 + \cdots + 1) \ (n \text{ terms}) \\ \therefore n^3 &= 3(1^2 + 2^2 + 3^2 + \cdots + n^2) - 3\frac{n(n+1)}{2} + n \end{aligned}$$

$$\therefore 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n^3}{3} + \frac{n(n+1)}{2} - \frac{n}{3} = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$$

This is formula (2.2).

See Questions 8, 9, 10 for finding the sum of the first n cubes, fourth powers and fifth powers.

Questions

The following questions are to test your understanding of the method of induction.

1 Prove (2.1) by the method of mathematical induction. Use the Example on page 9 as a template for your proof.

2 Similarly prove (2.2).

3 Similarly prove (2.4).

Here are two tricky applications of the Pigeon Hole Principle. If you really want a challenge, try them before looking at the HINTS which follow. Before you begin you may want to make up and try out a few test examples.

4 Prove that among any 10 natural numbers (not necessarily all distinct) there are two numbers whose difference is divisible by 9. See[8] for a Hint.

5 Prove that in any list $a_1, a_2, \ldots, a_{10}$ of 10 natural numbers (not necessarily all distinct) there is always a string (of one or more numbers) of the form $a_k, a_{k+1}, \ldots, a_n$ whose sum is divisible by 10. See[9] for Hints.

Now try these generalisations.

6 Replace"10" by "N" and "9" by "N-1" in Question 4. State and prove a general theorem.

7 Replace"10" by "N" in Question 5. State and prove a general theorem.

Next we find formulae for the sum of the first n cubes, fourth powers and even fifth powers.

---

[8]HINT: Imagine there are 9 boxes marked $0, 1, 2, \ldots, 8$. Put each of the 10 given natural numbers into the box corresponding to its remainder after dividing by 9.

What does the pigeon hole principle tell you and what can you deduce?

[9]Consider the sums $a_1, a_1 + a_2, a_1 + a_2 + a_3, \ldots, a_1 + a_2 + a_3 + \cdots + a_{10}$. Imagine there are 10 boxes marked $0, 1, 2, \ldots, 9$. Put each of the sums into the box corresponding to its remainder after dividing by 10.

What happens if one sum is in the box marked 0?

If all the sums are in the boxes marked $1, 2, \ldots, 9$ what does the pigeon hole principle tell you?

If 2 sums are in the same box what do you know about their difference?

8 Find $1^3 + 2^3 + 3^3 + \cdots + n^3$ by using the formula

$$k^4 - (k-1)^4 = 4k^3 - 6k^2 + 4k - 1,$$

and proceed in a similar way to that used on page 10. You will need to use the formulae for the sum of the n natural numbers *and* the sum of their squares, which we have already found. Check against (2.3).

9 Find $1^4 + 2^4 + 3^4 + \cdots + n^4$. Check against (2.4).

10 Find $1^5 + 2^5 + 3^5 + \cdots + n^5$. Here is the answer.[10]

[10] Answer to Question 10: $1^5 + 2^5 + 3^5 + \cdots + n^5 = \dfrac{n^6}{6} + \dfrac{n^5}{2} + \dfrac{5n^4}{12} - \dfrac{n^2}{12}$.

2.2 The Fibonacci Sequence

Looking at simple things deeply, finding a pattern, and using the pattern to gain new insights provides great value.

Overview

In the remainder of this Chapter we will often say "number" when we mean an integer rather than a general real number. We do this to be consistent with [HM]. It should be clear from the context what we mean.

The *Fibonacci sequence* is

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181,$$
$$6765, 10946, 17711, 28657, 46368, 75025, 121393, 196418, 317811, 514229,$$
$$832040, 1346269, 2178309, 3524578, 5702887, 9227465, 14930352, \ldots \quad (2.8)$$

The first 2 numbers are 1 and subsequent numbers are obtained by adding the previous two numbers.

This sequence arose originally as a model of rabbit population growth and also arises in spiral counts in pinecones and various flowers. Have a look at [HM, p57 Q6].

The methods we use to study the Fibonacci sequence include continued fractions, characteristic equations and mathematical induction, all of which will be explained later. They are very important and are used in many areas of mathematics.

Sequences of Numbers

[HM, 49,50]

We usually write an (infinite) sequence of numbers in the form

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

Thus for the Fibonacci sequence, $a_1 = 1$, $a_2 = 1$, $a_3 = 2$, $a_4 = 3$, $a_5 = 5$, $a_6 = 8$, etc.

Occasionally it is convenient to write a sequence as

$$a_0, a_1, a_2, \ldots, a_n, \ldots$$

One could even call the first term a_3 or a_7 or even a_{-23}, but this is not very common!

Definition of the Fibonacci Sequence

[HM, 51]

Definition 2.2.1. The Fibonacci sequence is defined by

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

Notice this Definition says that *every* term from the third term onwards is the sum of the previous two.

It also implies

$$a_{n+1} = a_n + a_{n-1} \text{ if } n+1 \geq 3, \text{ i.e. if } n \geq 2$$
$$a_{n+2} = a_{n+1} + a_n \text{ if } n+2 \geq 3, \text{ i.e. if } n \geq 1,$$
$$a_{n-1} = a_{n-2} + a_{n-3} \text{ if } n-1 \geq 3, \text{ i.e. if } n \geq 4, \text{ etc.}$$

Converging Quotients of Fibonacci Numbers

[HM, 51–55]

Calculating Successive Quotients It is interesting to investigate what happens to the ratio (i.e. quotient) of successive terms a_n/a_{n-1} when n becomes large.

In [HM, pp 51,52] you will see by using a calculator that it looks like the ratio might be getting closer and closer to a number around 1.6. Do a few calculations!

In [HM, pp 53] you see, or just look at (2.8), that the ratio of the 13th and 12th terms is

$$\frac{233}{144} = \frac{144+89}{144} = 1 + \frac{89}{144} = 1 + \frac{1}{\frac{144}{89}},$$

where $\dfrac{144}{89}$ is the ratio of the 12th and 11th terms.

Similarly, the ratio of the 14th and 13th terms is

$$\frac{377}{233} = \frac{233+144}{233} = 1 + \frac{144}{233} = 1 + \frac{1}{\frac{233}{144}},$$

where $\dfrac{233}{144}$ is the ratio of the 13th and 12th terms.

✍ *Do a similar analysis for the ratio of the 15th and 14th terms.*[11]

The General Result

Theorem 2.2.2. *If a_n is the nth term in the Fibonacci sequence and $n \geq 3$ then*

$$\frac{a_n}{a_{n-1}} = 1 + \frac{1}{\frac{a_{n-1}}{a_{n-2}}}.$$

Proof. [12]

[11] Recall the symbol ✍ indicates an example you should do, a question you should answer, etc. Write out your working neatly along the style of these notes and keep it. This is an extremely helpful way to increase your understanding of the material.

[12] This satisfies the requirements for a proof as discussed in Footnote 5. We have used only things we already know, namely the Definition of the Fibonacci sequence, a consequence of the Definition discussed immediately after the Definition, and some simple properties of addition and division. We have also briefly justified the important steps.

This is how you should try to write out your proofs.

From Definition 2.2.1 and the comments following this Definition, if $n \geq 3$ then

$$\frac{a_n}{a_{n-1}} = \frac{a_{n-1} + a_{n-2}}{a_{n-1}} = 1 + \frac{a_{n-2}}{a_{n-1}} = 1 + \frac{1}{\frac{a_{n-1}}{a_{n-2}}}.$$

□

The Limit of the Quotients If we *assume*[13] that the ratio $\frac{a_n}{a_{n-1}}$ "converges to a limit" (which is true) and assume certain properties of limits (which are true), then we can actually calculate the limit in this case.

Theorem 2.2.3. *If a_n is the nth term in the Fibonacci sequence then $\frac{a_n}{a_{n-1}}$ converges to $\frac{1+\sqrt{5}}{2}$ as n becomes arbitrarily large.*

Proof. [14] From Theorem 2.2.2, $\frac{a_n}{a_{n-1}} = 1 + \frac{1}{\frac{a_{n-1}}{a_{n-2}}}$.

Assume that $\frac{a_n}{a_{n-1}}$ converges to a limit ϕ as n becomes arbitrarily large. Then $\frac{a_{n-1}}{a_{n-2}}$ also converges to ϕ (this uses properties of limits, but it is not surprising).

It follows that

$$\phi = 1 + \frac{1}{\phi}.$$

(Notice that ϕ cannot be zero since the Fibonacci sequence is increasing and so $\frac{a_n}{a_{n+1}}$ is always at least one, and so also ϕ is at least one.) Hence

$$\phi^2 = \phi + 1.$$

and so

$$\phi^2 - \phi - 1 = 0.$$

The formula for solving a quadratic gives the two solutions

$$\phi_+ = \frac{1+\sqrt{5}}{2} \approx 1.618033988, \quad \phi_- = \frac{1-\sqrt{5}}{2} \approx -0.618033988. \tag{2.9}$$

Since the terms in the Fibonacci sequence are all positive we must have

$$\phi = \phi_+.$$

□

[13] We will discuss limits and their properties later in the course. The informal idea of a limit was known to mathematicians in the 1600's, but it caused much philosophical debate. The precise definition was not obtained until the 1800s. It took over 100 years to clarify the ideas.

[14] This is **not** really a "Proof" in the precise sense of Footnote 5. We have not given a careful definition of "converges" or "becomes arbitrarily large". We are assuming in the proof that $\frac{a_n}{a_{n-1}}$ does indeed converge to some limit and that limits have certain natural properties. All this is OK in this particular, but needs to eventually be justified. We will address these issues later in the course.

The Golden Ratio The number $\phi = (1+\sqrt{5})/2$ is called the *Golden Ratio* and will arise a number of times in the course.

Both numbers ϕ_+ and ϕ_- arise later in the formula for the nth term of the Fibonacci sequence. See Theorem 2.2.5.

Here is another way the Golden Ratio arises. Suppose we partition a line segment into two parts of lengths a (the larger) and b (the smaller).

If we require the ratio of the larger to the smaller to equal the ratio of the whole to the larger, i.e.

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

then we get $a^2 = ab + b^2$. This gives

$$\left(\frac{a}{b}\right)^2 = \frac{a}{b} + 1.$$

It follows that the ratio a/b is just the golden ratio ϕ.

Fibonacci Numbers and Continued Fractions

[HM, 52–54]

The Golden Ratio as a Continued Fraction In [HM, p 52] the formula in Theorem 2.2.2 is used to show that the ratios of successive Fibonacci numbers are

$$1,\ 1+\frac{1}{1},\ 1+\cfrac{1}{1+\frac{1}{1}},\ 1+\cfrac{1}{1+\cfrac{1}{1+\frac{1}{1}}},\ 1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\frac{1}{1}}}},\ \ldots$$

✍ *Explain how this follows from Theorem 2.2.2.*

Fractions written in this manner are called *continued fractions*. The limit of these numbers is the Golden Ratio and it is written as the *infinite continued fraction*

$$1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{\ddots}}}$$

This material is not in [HM]

⋆Properties of Continued Fractions Continued fractions and infinite continued fractions are important in number theory, approximation theory and chaos theory — all of which are subjects in mathematics.

You probably know that every real number has a (possibly infinite) decimal expansion, e.g.

$$\begin{aligned}\pi &= 3.14159265358979323846264338327950288419716939937510\ldots \\ &= 3 + \frac{1}{10} + \frac{4}{100} + \frac{1}{1,000} + \frac{5}{10,000} + \frac{9}{100,000} + \frac{2}{1,000,000} + \cdots \\ &= 3 + 1\cdot 10^{-1} + 4\cdot 10^{-2} + 1\cdot 10^{-3} + 5\cdot 10^{-4} + 9\cdot 10^{-5} + 2\cdot 10^{-6} + \cdots .\end{aligned}$$

It is also true that every real number has a (possibly infinite) continued fraction expansion and can be approximated by finite continued fractions. For example

$$\pi = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}}}}$$

In some ways continued fractions are "better" and more "natural" than a decimal expansion. For example, decimal expansions use the base number ten. But why do we count in multiples of ten? The answer is in biology. Because we have ten fingers and ten toes (usually).[15]

But continued fractions do not favour any particular base. They are more "pure" in this respect. And finite continued fractions usually give "better" approximations than finite decimal expansions of the same length.

Sums of Fibonacci numbers

[HM, 55,56]

Theorem 2.2.4. *Every natural number is either a Fibonacci number, or is a sum of Fibonacci numbers where none are adjacent Fibonacci numbers.*

We won't give the proof here since it is in [HM, pp 55,56]. A method of actually finding the Fibonacci numbers in the sum is also given there.

⋆Formula for the nth Fibonacci Number

The remainder of the mate
You may prefer to just look

There is a formula for the nth Fibonacci number. This is tricky and is not done in [HM].

[15] The Babylonians about 3000 BC counted in multiples of the base number sixty.

Binary systems with base two are used by computers. In this case two is written as 10 and the natural numbers are 1, 10, 11, 100, 101, 110, 111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111, 10000, 100001, etc.

Hexadecimal systems with base sixteen are also used. In this case the digits are 0, 1, 2, ..., 9, A, B, C, D, E, F. The next numbers after this are 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 1A, 1B, 1C, 1D, 1E, 1F, 20, ..., FF, 100. In this system A is ten and 10 is sixteen, i.e. 16 in our usual way of counting.

Theorem 2.2.5. *If a_n is the nth Fibonacci number then*

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

(See also (2.9) for the definition of ϕ_+ and ϕ_-.) This is a pretty amazing formula. It is not even obvious that it gives a natural number.

There are many ways of proving this result, and we will give two methods.

One method is the method of *mathematical induction*, one version of which we discuss earlier. The disadvantage of this method is that you need to guess the answer ahead of time, but then the method of mathematical induction allows you to prove your guess is correct. However, with Fibonacci numbers it is very far from clear how you might guess the correct formula.

Another method is the *method of characteristic equations* (see (2.14)), which we look at now.

Proof via the Characteristic Equation This method works in *many* similar situations.

The following argument will be a bit tricky. But after you have worked through it you should try Questions 1 and 2. These Questions involve other examples, with Hints as you proceed, and will help reinforce the ideas.

First we will discuss the method in detail. Then we will write it out again more briefly in the Proof of Theorem 2.2.5 on page 21.

We first break Definition 2.2.1 on page 13 of the Fibonacci sequence into 2 parts.

The *first part* consists of the *initial conditions*:

$$a_1 = 1,\ a_2 = 1. \tag{2.10}$$

The *second part* is the *recurrence relation* (or *recurrence equation*) for later terms:

$$a_n = a_{n-1} + a_{n-2} \text{ if } n \geq 3. \tag{2.11}$$

We could change just the initial conditions. For example,

$$a_1 = 2,\ a_2 = 1.$$

Then the recurrence relation gives the sequence

$$2, 1, 3, 4, 7, 11, 18, 29, 47, 76, 123, 199, 322, 521, 843, \ldots \tag{2.12}$$

This is called the *Lucas sequence* in [HM, p59 Q10].

Dealing with the Recurrence Relation Let us first think about the recurrence relation (2.11) by itself, without considering the initial conditions (2.10). Because of the way powers of numbers behave, it is going to be a good idea to *fix* a number r (which we will later find) and test if $a_n = r^n$ for each natural number n satisfies the recurrence relation. It is not at all obvious that this will work. The main reason it will is that equation (2.13) is equivalent to the characteristic equation (2.14) which no longer involves n.

Note that if $a_n = r^n$ for every n then $a_{n-1} = r^{n-1}$ and $a_{n-2} = r^{n-2}$ (to be precise, in the first case for $n - 1 \geq 1$ and so $n \geq 2$, while in the second case for $n \geq 3$).

From (2.11) we see $a_n = r^n$ satisfies the recurrence relation *if and only if*

$$r^n = r^{n-1} + r^{n-2} \text{ for } n \geq 3. \tag{2.13}$$

This is true if and only if $r = 0$ or, dividing through by r^{n-2},

$$r^2 = r + 1, \text{ i.e. } r^2 - r - 1 = 0. \tag{2.14}$$

This is sometimes called the *characteristic equation*.

Notice that something very interesting has happened! There is no longer an n in the last equation. This is because of the way powers of a number behave when substituted into the recurrence relation.

This last equation is a quadratic and is satisfied by r if and only if

$$r = \phi_+ = \frac{1+\sqrt{5}}{2} \quad \text{or} \quad r = \phi_- = \frac{1-\sqrt{5}}{2}.$$

(We saw the same equation in the proof of Theorem 2.2.3.)

Thus we have shown that both $a_n = (\phi_+)^n$ and $a_n = (\phi_-)^n$ (as well as $a_n = 0$), are solutions of the recurrence relation. But you will easily see that these a_n do not satisfy the initial conditions, just try $n = 1$ or $n = 2$.

Are we stuck? No!

Notice that if $a_n = r^n$ satisfies the recurrence relation then so does $a_n = 2r^n$ or $a_n = 7r^n$ or even $a_n = -23.57r^n$. The main point is that if (2.11) is true then it remains true when we multiply through by 2 or 7 or even by -23.57.

In words: *Any constant multiple of a solution of the recurrence relation is itself a solution.*

So now we have many solutions of the recurrence equation. Namely

$$a_n = A(\phi_+)^n \text{ and } a_n = B(\phi_-)^n,$$

where A and B can be any two real numbers.

The next important observation is that if we have one sequence of numbers satisfying the recurrence relation and a second sequence of numbers satisfying the recurrence relation, then the sequence obtained by adding corresponding terms also satisfies the recurrence relation. *Why is this?* ✍

In words: *The sum of any two solutions of the recurrence relation is itself a solution.*

So putting all this together we have shown that

$$a_n = A(\phi_+)^n + B(\phi_-)^n$$

is a solution of the recurrence relation for any A and B. (Notice that the uninteresting solution $a_n = 0$ is also included, just set $A = B = 0$.)

Dealing with the Initial Conditions Now we come back to the initial conditions. Since there are 2 numbers A and B at our disposal, and 2 initial conditions, it seems likely, and is true, that we can choose A and B so the initial conditions are satisfied.

In fact, we have $a_n = A(\phi_+)^n + B(\phi_-)^n$ satisfies the initial conditions if and only if

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

These are just 2 simultaneous equations in 2 unknowns. To minimise the amount of calculation it is probably best to multiply the first equation by $\left(\frac{1+\sqrt{5}}{2}\right)$ and subtract the second, and then multiply the first equation by $\left(\frac{1-\sqrt{5}}{2}\right)$ and subtract the second. This gives *(Check it!)*:

$$\left(\frac{-1+\sqrt{5}}{2}\right) = B\left(\frac{-5+\sqrt{5}}{2}\right)$$
$$\left(\frac{-1-\sqrt{5}}{2}\right) = A\left(\frac{-5-\sqrt{5}}{2}\right).$$

In order to solve, write this as

$$\left(\frac{-1+\sqrt{5}}{2}\right) = -B\sqrt{5}\left(\frac{\sqrt{5}-1}{2}\right)$$
$$\left(\frac{-1-\sqrt{5}}{2}\right) = A\sqrt{5}\left(\frac{-\sqrt{5}-1}{2}\right),$$

and so

$$A = 1/\sqrt{5},\ B = -1/\sqrt{5}.$$

Putting it all together,

$$a_n = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n}{\sqrt{5}}. \tag{2.15}$$

One Final Point We have seen that if a_n is defined by (2.15) for each integer $n \geq 1$ then this gives a solution of (2.10) and (2.11). Are there other solutions? That is, are there other possible values for $a_1, a_2, a_3, \ldots$ which satisfy (2.10) and (2.11)?

The answer is NO for the following reason.

From (2.10) there is *exactly one* possible value for each of a_1 and a_2, namely 1 and 1 respectively. Then from (2.11) there is *exactly one* possible value for a_3 (namely 2), *exactly one* possible value for a_4, *exactly one* possible value for a_5, etc. Moreover, all these values are natural numbers. (We could use mathematical induction to justify all this, but I think that would be overkill).

So, to summarise, we have found in (2.15) *one* sequence of numbers which satisfies (2.10) and (2.11). But we have also shown that there is *one and only one* sequence of numbers (and that they are in fact natural numbers) which satisfies (2.10) and (2.11). It follows that since the sequence given by (2.15) is *one* solution of (2.10) and (2.11), it is *the one and only* solution, and moreover all members of the sequence given by (2.15) are natural numbers.

We will not normally repeat this type of argument each time, but you should at least see it once! (Then perhaps stop worrying and forget about it.)

Setting out the Proof in a Compact Manner OK, that was pretty heavy going. So I will now write it out again more briefly.

Proof of Theorem 2.2.5. $a_n = r^n$ is a solution of the recurrence relation

$$a_n = a_{n-1} + a_{n-2} \text{ for } n \geq 3$$

if and only if

$$r^n = r^{n-1} + r^{n-2}, \text{ i.e. } r^2 = r + 1 \text{ (or } r = 0).$$

This gives

$$r = \phi_+ = \frac{1+\sqrt{5}}{2} \quad \text{or} \quad r = \phi_- = \frac{1-\sqrt{5}}{2}.$$

A constant multiple of a solution of the recurrence relation is a solution, and the sum of solutions is a solution, so

$$a_n = A(\phi_+)^n + B(\phi_-)^n, \quad n \geq 1,$$

is a solution of the recurrence relation for any numbers A and B (not necessarily integers).

This satisfies the initial conditions $a_1 = 1$ and $a_2 = 1$ if and only if

$$\begin{aligned} 1 &= A\phi_+ + B\phi_- \\ 1 &= A(\phi_+)^2 + B(\phi_-)^2. \end{aligned}$$

Solving these simultaneous equations as before gives

$$A = 1/\sqrt{5}, \ B = -1/\sqrt{5}.$$

Thus the required solution is

$$a_n = \frac{(\phi_+)^n - (\phi_-)^n}{\sqrt{5}}$$

for all natural numbers n. □

⋆*Proof by Induction of the Formula*

Discussion Recall that the Fibonacci sequence in (2.8) is defined by the relations given in Definition 2.2.1, i.e.

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

We then proved in Theorem 2.2.5 that

$$a_n = \frac{(\phi_+)^n - (\phi_-)^n}{\sqrt{5}}.$$

If we were able to guess this formula (by some devious means) then we could prove it rigorously by the Principle of Mathematical Induction.

To be more precise, we have to use an extension of this Principle. In the previous applications we proved $P(1)$ was true and also proved that whenever $P(k)$ is true then $P(k+1)$ is true. It then follows that $P(n)$ is true for all n.

Strong Principle of Mathematical Induction

Theorem 2.2.6 (Strong Principle of Mathematical Induction)**.** *Suppose that $P(n)$ is a statement about n, for each natural number n. Assume we know:*

1. *$P(1), \ldots, P(a)$ are all true for some natural number a,* ***(basic step)***
2. *Whenever $P(1), \ldots, P(k)$ are true for a natural number $k \geq a$, it follows that $P(k+1)$ is also true.* ***(inductive step)***

Then the statement $P(n)$ is true for every *natural number n.*

Proof.

- By the first assumption, $P(1), \ldots, P(a)$ are true.
- By the second assumption, since $P(1), \ldots, P(a)$ are true it follows that $P(a+1)$ is true.
- By the second assumption, since $P(1), \ldots, P(a+1)$ are true it follows that $P(a+2)$ is true.
- By the second assumption, since $P(1), \ldots, P(a+2)$ are true it follows that $P(a+3)$ is true.
- By the second assumption, since $P(1), \ldots, P(a+3)$ are true it follows that $P(a+4)$ is true.
- etc.

In this way we see that for *any* natural number n, $P(n)$ is true. □

Proof of the Formula

Theorem 2.2.7. *Suppose*

$$a_1 = 1,\ a_2 = 1,\ a_n = a_{n-1} + a_{n-2} \text{ if } n \geq 3. \tag{2.16}$$

Then

$$a_n = \frac{(\phi_+)^n - (\phi_-)^n}{\sqrt{5}}. \tag{2.17}$$

Proof. We use the strong principle of mathematical induction (with $a = 2$).

For the *basic step*, just check that (2.17) is true for $n = 1$ and $n = 2$, i.e.

$$1 = \frac{\phi_+ - \phi_-}{\sqrt{5}} \text{ and } 1 = \frac{(\phi_+)^2 - (\phi_-)^2}{\sqrt{5}}.$$

For the *inductive step* take any $k \geq 2$ and *assume* that (2.17) is true for all n from 1 up to k. Now

$$a_{k+1} = a_{k-1} + a_k$$

(from (2.16) by setting $n = k + 1$)

$$= \frac{(\phi_+)^{k-1} - (\phi_-)^{k-1}}{\sqrt{5}} + \frac{(\phi_+)^k - (\phi_-)^k}{\sqrt{5}}$$

(because we are assuming that (2.17) is true for $n = k - 1$ and for $n = k$)

$$= \frac{(\phi_+)^{k-1}(1 + \phi_+)}{\sqrt{5}} - \frac{(\phi_-)^{k-1}(1 + \phi_-)}{\sqrt{5}}$$
$$= \frac{(\phi_+)^{k-1}(\phi_+)^2}{\sqrt{5}} - \frac{(\phi_-)^{k-1}(\phi_-)^2}{\sqrt{5}}$$

(since both ϕ_+ and ϕ_- are solutions of $r^2 = r + 1$)

$$= \frac{(\phi_+)^{k+1} - (\phi_-)^{k+1}}{\sqrt{5}}.$$

Thus (2.17) is true for $n = k + 1$.

It follows from the Strong Principle of Mathematical Induction that (2.17) is true for all natural numbers n. □

Questions

Here are two examples of the method used in the proof of Theorem 2.2.5. If you do them you will understand the method much better!

1 Find a formula for the nth Lucas number, see (2.12).

(The argument is similar to that in the proof of Theorem 2.2.5 on page 21. You may save yourself some calculation effort if you look carefully at the way we did the calculations in the discussion before the proof. Check that your answer really works for the cases $n = 1, 2, 3$.)

DON'T LOOK NOW but the answer is in footnote[16] below.

2 Consider the sequence

$$1, 1, 3, 5, 11, 21, 43, 85, 171, \ldots$$

The first 2 terms are $a_1 = 1$ and $a_2 = 1$. For $n \geq 2$, $a_n = a_{n-1} + 2a_{n-2}$.

Find a formula for the nth term.

Check your answer for $n = 1, 2, 3$.

HINT: The recurrence relation will be different from that for the Fibonacci and Lucas sequences. However, it will have nicer solutions and this will make the arithmetic easier.

DON'T LOOK NOW but the answer is in footnote[17] below.

Next try the same 2 examples using induction.

3 Prove the formula in footnote 16 by the strong principle of induction. Use the method on page 22 as a template for your argument.

4 Prove the formula in footnote 17 by the strong principle of induction.

---

[16] The answer is $\left(\frac{1+\sqrt{5}}{2}\right)^{n-1} + \left(\frac{1-\sqrt{5}}{2}\right)^{n-1}$, i.e. $(\phi_+)^{n-1} + (\phi_-)^{n-1}$.

[17] The answer is $-\frac{1}{3}(-1)^n + \frac{1}{3}2^n$.

2.3 Prime Numbers

Examining the building blocks of a complex structure answers old questions, invites new questions, and leads to greater understanding.

Overview

In [HM] it is shown that every natural number can be written as a product of prime numbers and that there are infinitely many primes. The prime number theorem is discussed (it estimates how "dense" the primes are in the set of all natural numbers). Some famous theorems and conjectures in number theory are discussed briefly (Fermat's last Theorem, The Twin Prime Conjecture and the Goldbach Conjecture).

In addition, in these Notes the greatest common divisor of two numbers is discussed and the Euclidean algorithm is developed — this is important material in general and in particular in Section 2.5 on RSA codes. We also prove that the factorisation into primes is unique and show some consequences that are important in understanding RSA encryption.

The Division Algorithm

[HM, 64–66]

Examples We know that 23 divided by 7 gives the quotient 3 and the remainder 2. That is

$$23 = 3 \times 7 + 2.$$

Similarly

$$14 = 2 \times 7 + 0, \quad 3 = 0 \times 7 + 3, \quad 19 = 2 \times 7 + 5, \quad 13 = 1 \times 7 + 6,$$

etc.

Sometimes it will be convenient to divide a negative integer by a natural number (and of course we can also divide by a negative integer but not by 0). For example

$$-3 = (-1) \times 7 + 4, \quad -16 = (-3) \times 7 + 5, \quad -28 = (-4) \times 7.$$

In general if we divide an integer m by a natural number n then we get a *quotient* q which may be positive or negative or 0, and a *remainder* r which is in the range $0, 1, 2, \ldots, n-1$. In symbols,

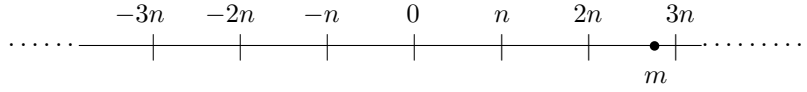
$$m = qn + r.$$

Geometric Picture and Theorem Think of multiples of n being marked off by points on the real line. The number m will either lie on one of these points or between 2 consecutive points. See the diagram below.

In the first case $m = qn$ for some integer q. In the second case m will lie between qn and $(q+1)n$ for some integer q.

In the picture below $q = 2$ and $r = m - 2n$.

The proof is motivated by the picture.



Theorem 2.3.1 (The Division Algorithm)**.** *Suppose n is a natural number and m is any integer. Then there exists a unique integer q, and a unique integer r in the range $0, 1, 2, \ldots, n-1$, such that*

$$m = qn + r.$$

Proof. [18] Consider the integers $\ldots, -3n, -2n, -n, 0, n, 2n, 3n, 4n, \ldots$. The number m will either (see the previous diagram)

- equal some multiple of n, let's call it qn, for a unique ("exactly one") integer q; *or*
- will lie strictly between some qn and $(q+1)n$, i.e. $qn < m < (q+1)n$ for a unique ("exactly one") integer q.

In the first case, $m = qn + r$ where $r = 0$. In the second case $m = qn + r$ where $r = m - qn$. □

Dividing a Number We say 3 *divides* 21 (or equivalently, 3 *is a factor of* 21) because the remainder is 0 after dividing 21 by 3. We write $3 \mid 21$. In general, we have the following Definition.

Definition 2.3.2. Suppose n is a natural number and m is an integer.

We say n *divides* m if $m = qn$ for some integer q, i.e. if the remainder in the Division Algorithm is 0.

We write $n \mid m$ and say "n divides m". The integers q and n are called *factors* of m.

For example, $3 \mid 27$, $4 \mid 12$, $7 \mid -21$, but $4 \nmid 6$ (which we read as "4 does *not* divide 6").

Dividing Sums and Products It follows that if n divides m, then n also divides the product mj where j is any natural number.

For example, $6 \mid 18$. It follows that $6 \mid (18 \times 5)$, $6 \mid (18 \times 4)$, $6 \mid (18 \times 23)$, $6 \mid 18^2$, etc.

It also follows that if n divides k and n divides m then n divides the sum $k + m$.

For example, $6 \mid 18$ and $6 \mid 24$ so $6 \mid (18 + 24)$.

The above are not surprising when you think about a few examples. It is also possible to write out a short proof, and there are some Hints in Questions 1 and 2.

[18] In this proof we are using some simple properties about inequalities. See Footnote 5.

You may find this particular proof a little unsatisfactory. And in some ways it is. You may think that the result we are "proving" here is just as obvious as the facts we are using in the proof. I would not quite agree, but I think the difference is not great.

If you prefer, in this case you can just take the result as one of the basic facts we assume about integers. Later we will prove results which are far less obvious.

[HM, 66,67]

Prime Factorisation

Definition of Prime Numbers

Definition 2.3.3. A natural number p greater than one is called a *prime number* if it is not the product of 2 smaller natural numbers.

In other words, $p > 1$ is *prime* if the only natural numbers which divide p are 1 and p itself.

We do not include 1 as a prime number.

Examples of Prime Numbers The primes less than 500 are

$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89,$
$97, 101, 103, 107, 109, 113, 127, 131, 137, 139, 149, 151, 157, 163, 167, 173, 179,$
$181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, 241, 251, 257, 263, 269, 271,$
$277, 281, 283, 293, 307, 311, 313, 317, 331, 337, 347, 349, 353, 359, 367, 373, 379,$
$383, 389, 397, 401, 409, 419, 421, 431, 433, 439, 443, 449, 457, 461, 463, 467, 479,$
$487, 491, 499.$

Natural Numbers are a Product of Primes The following Theorem is also called the *Fundamental Theorem of Arithmetic*.

Theorem 2.3.4 (Prime Factorisation Theorem)**.** *Each natural number n greater than 1 is either a prime or a product of primes.*

Moreover, n can only be expressed as a product of primes in one way, except for a reordering of factors.

Proof. Suppose n is a natural number greater than 1.

If n is a prime then we are done.

If n is not a prime this means n can be divided by some other natural number larger than 1 but less than n and so $n = a \times b$, say. If either a or b is not prime it can be factored as a product of 2 smaller numbers. Continuing in this way we get smaller and smaller factors and so after a finite number of steps we get a factorisation of n where all the factors are prime.

The proof that n can only be expressed as a product of primes in one way, except for a reordering of factors, is not done in [HM]. We discuss and prove it here in the Section "Prime Factorisations are Unique" beginning on page 32, see Theorem 2.3.13 on page 34. □

For example: $9857934 = 2 \times 3^2 \times 547663$ and the numbers $2, 3, 547663$ are primes. Also $988788377878738398 = 2 \times 3^2 \times 13 \times 541 \times 7810704913967$ and all the factors are prime[19].

There are Infinitely Many Primes

67–71]

This is discussed carefully in [HM]. Here I will briefly write out first the proof in [HM] and then write a slightly different proof.

Theorem 2.3.5. *There are infinitely many prime numbers.*

First Proof. We will show that for every natural number n there is a prime which is larger than n.

This is clearly true if $n = 1$, just take the prime 2.

So now we assume that $n \geq 2$. Let

$$N = (1 \times 2 \times 3 \times \cdots \times n) + 1.$$

Then $N > n$.

If N is itself prime then we have a prime larger than n and we are done.

If N is not prime then it must have prime factors by the Prime Factorisation Theorem. But none of these prime factors can be $\leq n$, because any number from 1 to n when divided into N gives the remainder 1. It follows that the prime factors of N must be larger than n.

Thus whether or not N itself is prime, we have shown there is a prime larger than n. □

Second proof. Assume there are only finitely many primes and write them in a list as

$$p_1, p_2, \ldots, p_k.$$

Let

$$M = (p_1 \cdot p_2 \cdot \ldots \cdot p_k) + 1.$$

Since M is larger than any of $p_1, \ldots, p_k$ in the list of *all* primes, M itself is not a prime.

This means that M has prime factors by the Prime Factorisation Theorem. But each prime in the list of *all* primes $p_1, p_2, \ldots, p_k$ when divided into M give a remainder equal to 1. This means there are *no* primes which divide M, contradicting the fact that M must be a product of primes.

This contradiction means the *assumption* that there are only finitely many primes is wrong. □

How Dense are the Primes?

[HM, 71–73]

Numerical Experimentation It is conventional to let $\pi(n)$[20] denote the number of primes up to and including n.

In [HM] page 72 there is a table which shows n, $\pi(n)$ and $\pi(n)/n$ in its first three columns for various values of n. You should think of $\pi(n)/n$ as the *density* of primes among the first n natural numbers.

[19] I did these factorisations by using the MAPLE program, which you will use in this course.

[20] The "π" here is *not* the same as the usual "π" which is the ratio of the circumference of a circle to its diameter.

In column four page 72 of [HM] the corresponding values of $1/\ln(n)$ are calculated. You may or may not have yet seen logarithms. By $\ln(n)$ is meant something a little more complicated — it is the logarithm of n to the base e instead of to the base 10. The number $e \approx 2.7182818284590452354$ arises naturally in many ways (calculus, compound interest, number theory, ...) and was mentioned before on page 7. However, for our purposes, it is sufficient to use the LN (or similar) key on your calculator to find $\ln(n)$.

It turns out that $1/\ln(n)$ is a very good approximation to the density (i.e. proportion) of primes near n when n is very large.

An interesting example for us is when n has about 150 digits. This is because in RSA cryptography in Section 2.5 we will be looking for primes of this size. If we take $n = 10^{150}$ and use Maple we find that $1/\ln(10^{150}) \approx 0.0029 \approx 1/345$. Which means about one in every 345 natural numbers with 150 digits is prime. This means there are an awful lot of primes out there to choose from — and in fact it is very easy to find them using Maple! See also Question 4.

The Prime Number Theorem

Theorem 2.3.6. *The number $\pi(n)$ of primes less than or equal to n is asymptotic to $n/\ln(n)$ as n gets larger and larger.*

By "asymptotic" we mean that the ratio $\pi(n)\Big/\dfrac{n}{\ln(n)}$ gets as close as we wish to 1 (i.e. converges to 1) as n gets larger and larger. We sometimes write this as $\pi(n) \sim n/\ln(n)$. Even though $\pi(n)\Big/\dfrac{n}{\ln(n)}$ is getting closer and closer to 1, it does so very slowly. You can see this in the right hand diagram, top graph, of Fig. 2.1.

The fact $\pi(n)\Big/\dfrac{n}{\ln(n)}$ gets closer and closer to 1 is the same as saying the ratio $\dfrac{\pi(n)}{n}\Big/\dfrac{1}{\ln(n)}$ gets closer and closer to 1 (i.e. converges to 1) as n gets larger and larger. We write this as $\pi(n)/n \sim 1/\ln(n)$. Remember that $\pi(n)/n$ is the density of primes in the first n integers.

The fact $\pi(n)\Big/\dfrac{n}{\ln(n)}$ gets closer and closer to 1 does not mean that the difference between $\pi(n)$ and $\dfrac{n}{\ln(n)}$ gets closer and closer to 0 as n gets larger and larger,[21] and in fact this is not true. For example, $n^2 \sim (n^2 + n)$ (*why?*), but the difference between n^2 and $(n^2 + n)$ is n and this is certainly not getting close to 0 as n gets larger and larger.

Another point worth noting is that it is also true that $\pi(n) \sim n/(\ln(n) - 1)$, and this gives a better approximation to $\pi(n)$ than $n/\ln(n)$. See Fig 2.1.

The first proof of the Prime Number Theorem was given in 1896. There are a number of different proofs, all complicated. The easiest way to do the proof

[21] For this reason the statement of the Prime Number Theorem on [HM, page 73] is too vague and even misleading.

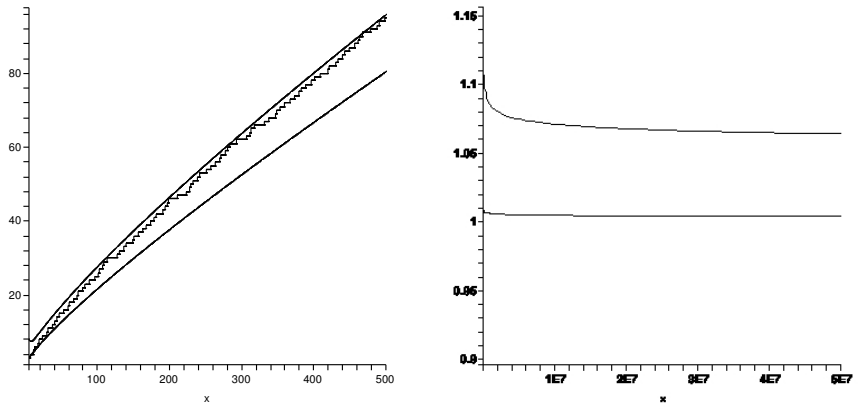


Figure 2.1: From top to bottom on the left, graphs of $n/(ln(n)-1)$, $\pi(n)$ and $n/ln(n)$. From top to bottom on the right, graphs of $\pi(n)\Big/\dfrac{n}{\ln(n)}$ and $\pi(n)\Big/\dfrac{n}{(\ln(n)-1)}$. By 4E7 is meant 4×10^7, etc.

involves some very deep properties of complex numbers.[22] Unfortunately we do not have nearly enough tools to prove this theorem at this stage.

Big Theorems and Big Conjectures

[HM, 73–76]

In [HM] there is a discussion of Fermat's Last Theorem, which was finally proved after 350 years by Andrew Wiles (Princeton) in 1994. I think it fair to assert that all experts in the field would agree that Fermat was mistaken in his claim that he had a proof of the Theorem. (Andrew Wiles was a PhD student of John Coates. John Coates was an honours student at ANU, later on the ANU faculty, now at Cambridge.)

There is also mention in [HM] of the Twin Prime Question (are there infinitely many pairs of primes differing by 2 — such as 11 and 13, 17 and 19, 29 and 31, 41 and 43, ...) and the Goldbach Conjecture (every even number greater than 2 is a sum of 2 primes), which have been open questions for centuries.

Finally, I would like to mention a famous problem that has been around for over 200 years and was solved in 2004. Ben Green and Terry Tao showed for every natural number k that there are arithmetic progressions[23] of length k which consist of prime numbers. For example, 199, 409, 619, 829, 1039, 1249, 1459, 1669, 1879, 2089 is an arithmetic progression of primes of length 10, with difference 210 between any two successive primes in this sequence.

[22] I hope that by now you have some feeling for the fact that the different parts of mathematics have wonderful, deep and initially surprising connections with each other.

[23] An arithmetic progression is an increasing sequence of numbers such that the difference between any 2 consecutive numbers in the sequence is the same.

Terry Tao is an Australian mathematician from Adelaide, who recently spent some time at ANU and is now at the University of California in Los Angeles. In 2006 he received the Fields Medal in mathematics for the above work and much else. He is one of the youngest, and the only Australian, to receive this award. The Fields medal is considered to be the Nobel Prize equivalent in mathematics.

⋆Greatest Common Divisor

The material in this Section is not in [HM]. You may prefer to look at it briefly and come back to it later. It is important. It is the key to the RSA encryption method.

Definition 2.3.7. The *greatest common divisor* of two natural numbers a and b is the largest natural number which divides both a and b.

If d is the greatest common divisor of a and b we write $d = \gcd(a, b)$.

Another terminology is *highest common factor*.

For example, $\gcd(3, 6) = 3$, $\gcd(1, 6) = 1$, $\gcd(4, 10) = 2$. What about $\gcd(2261, 1275)$? See below.

Definition 2.3.8. If the greatest common divisor of two numbers is 1 then we say the two numbers are *relatively prime*.

In other words, two natural numbers are relatively prime if there is no natural number which is a common factor of both other than 1.

For example, 1 and 6, 14 and 15, 5 and 12, 6 and 25, are relatively prime.

The Euclidean Algorithm[24] There is a mechanical procedure for finding any gcd, called the *Euclidean Algorithm*[25] which we now describe.

If we divide 1275 into 2261 we get $2261 = 1 \cdot 1275 + 986$.

It follows from this equation that if d divides both 2261 and 1275 then d divides both 1275 (of course) and 986. Moreover, it also follows from the equation that if d divides both both 1275 and 986 then d divides both 2261 and 1275.

This implies in particular that $\gcd(2261, 1275) = \gcd(1275, 986)$. In this way we will keep reducing the problem to finding the gcd of smaller and smaller pairs of numbers.

Two Worked Examples This is how we set out the *Euclidean Algorithm* to find $\gcd(2261, 1275)$ in the example we were looking at:

$$\begin{aligned}
2261 &= 1 \times 1275 + 986 && \text{so } \gcd(2261, 1275) = \gcd(1275, 986),\\
1275 &= 1 \times 986 + 289 && \text{so } \gcd(1275, 986) = \gcd(986, 289),\\
986 &= 3 \times 289 + 119 && \text{so } \gcd(986, 289) = \gcd(289, 119),\\
289 &= 2 \times 119 + 51 && \text{so } \gcd(289, 119) = \gcd(119, 51),\\
119 &= 2 \times 51 + 17 && \text{so } \gcd(119, 51) = \gcd(51, 17),\\
51 &= 3 \times 17 + 0 && \text{so } \gcd(51, 17) = 17.
\end{aligned} \tag{2.18}$$

[24] An algorithm is a "mechanical" routine which can be programmed into a computer and which will eventually stop and give the required answer.

[25] Same Euclid as in geometry. His books, written about 300 BC, contain the algorithm.

The idea is to continue until the remainder is 0. This will eventually occur and the Euclidean Algorithm will stop as we explain below.

It follows that $\gcd(2261, 1275) = \gcd(1275, 986) = \cdots = \gcd(51, 17) = 17$.

Here is another example. Find $\gcd(245, 24)$.

$$\begin{aligned} 245 &= 10 \times 24 + 5 \quad \text{so } \gcd(245, 24) = \gcd(24, 5), \\ 24 &= 4 \times 5 + 4 \quad \text{so } \gcd(24, 5) = \gcd(5, 4), \\ 5 &= 1 \times 4 + 1 \quad \text{so } \gcd(5, 4) = \gcd(4, 1) \\ 4 &= 1 \times 4 + 0 \quad \text{so } \gcd(4, 1) = 1. \end{aligned} \tag{2.19}$$

It follows that $\gcd(245, 24) = \gcd(24, 5) = \gcd(5, 4) = \gcd(4, 1) = 1$.

Again we continued until the remainder was 0.

To make sure you understand the Euclidean Algorithm do Question 5. ✍

Programming the Euclidean Algorithm Because the method is quite mechanical, one can program a computer to do the Euclidean Algorithm.

The Euclidean Algorithm Eventually Stops First arrange the pair of natural numbers so that the larger number comes first. (In the very uninteresting case where the 2 numbers are the same then the algorithm stops after one step since we get a quotient one and a remainder zero. The gcd is then the same as the two numbers.)

After each step we end up with a smaller pair of natural numbers. The first (i.e. larger) number in the new pair is the same as the second (i.e. smaller) number from the previous pair, and the second number in the new pair is the remainder that was obtained.

So long as the remainder is larger than 0, each pair of natural numbers is smaller than the previous pair. Eventually we will come to a situation where the remainder is 0. Otherwise the pair of natural numbers will keep decreasing until it is just (1,1), but then at the next step the remainder is also 0 in this case.

Since the pair of natural numbers cannot go below (1,1) we eventually obtain a remainder which is 0, and then the gcd is the smaller number in the pair which gave this remainder.

The Extended Euclidean Algorithm By working backwards in (2.18) from the second last line to line 1 we can use the calculations to express the greatest common divisor 17 as a multiple of 119 + a (negative) multiple of 51, and then as a multiple of 289 + a multiple of 119, and then as a multiple of 986 + a multiple of 289, and then as a multiple of 1275 + a multiple of 9986, and finally as a multiple of 2261 + a multiple of 1275. The "multiple" in each case may be a positive or a negative integer.

Here we do the calculation.

$$\begin{aligned}
17 &= 119 - 2 \times 51 && \text{from line 5 of (2.18)}\\
&= 119 - 2 \times (289 - 2 \times 119) && \text{from line 4 of (2.18)}\\
&= -2 \times 289 + 5 \times 119 && \text{by simplifying}\\
&= -2 \times 289 + 5 \times (986 - 3 \times 289) && \text{from line 3 of (2.18)}\\
&= 5 \times 986 - 17 \times 289 && \text{by simplifying}\\
&= 5 \times 986 - 17 \times (1275 - 1 \times 986) && \text{from line 2 of (2.18)}\\
&= -17 \times 1275 + 22 \times 986 && \text{by simplifying}\\
&= -17 \times 1275 + 22 \times (2261 - 1 \times 1275) && \text{from line 1 of (2.18)}\\
&= 22 \times 2261 - 39 \times 1275 && \text{by simplifying.}
\end{aligned}$$

Thus we have found that $17 = 22 \times 2261 - 39 \times 1275$ (*check the answer!*) and so expressed $\gcd(2261, 1275)$ (which is 17) as a multiple of 2261 + a multiple of 1275.

In a similar way we can express $\gcd(245, 24)$ (which we saw was 1) as a multiple of 245 + a multiple of 24.

$$\begin{aligned}
1 &= 5 - 1 \times 4 && \text{from line 3 of (2.19)}\\
&= 5 - 1 \times (24 - 4 \times 5) && \text{from line 2 of (2.19)}\\
&= -1 \times 24 + 5 \times 5 && \text{by simplifying}\\
&= -1 \times 24 + 5 \times (245 - 10 \times 24) && \text{from line 1 of (2.19)}\\
&= 5 \times 245 - 51 \times 24 && \text{by simplifying.}
\end{aligned}$$

Thus we have found that $1 = 5 \times 245 - 51 \times 24$ (*check the answer!*) and so expressed $\gcd(245, 24)$ (which is 1) as a multiple of 245 + a multiple of 24.

The above argument gives a general result.

Theorem 2.3.9 (The Extended Euclidean Algorithm)**.** *Suppose a and b are natural numbers and $d = \gcd(a, b)$. Then there exist integers s and t such that $d = sa + tb$.*

Moreover, there is an algorithm for finding s and t.

Proof. The method used above first for $a = 2261$ and $b = 1275$, and then for $a = 245$ and $b = 24$, gives an algorithm which works for any a and b. □

✍ *To make sure you understand the Extended Euclidean Algorithm now do Question 6.*

⋆ *Prime Factorisations are Unique*

The material in this Section is not in [HM]. You may prefer to look at it briefly and come back to it later.

Discussion of The Result We will prove the second half of Theorem 2.3.4, that *if you write a number as a product of prime factors in 2 ways then after perhaps changing the order of the factors the factorisations are the same.* In other words, each prime that occurs in one factorisation occurs exactly the same number of times in the other factorisation. See Theorem 2.3.13.

You possibly already knew this result. But it is not as obvious as might first appear — see the following discussion.

However, don't get concerned if the following is confusing. The main point is just that you know the result.

Two Questions

Q1 We know that $3 \times 5 \neq 2 \times 7$ and $7 \times 13 \neq 3 \times 31$. *Are there some gargantuan primes which are all* different*, let's call them p, q, r and s, such that $p \times q = r \times s$?*

The answer is NO, but it is not obvious why this is so.

Q2 Here is a related Question. *If p is a prime and $p \mid (a \times b)$ where a and b are natural numbers, does it follow that p divides at least one of a or b?*

We will show that the answer is YES. But suppose we could actually find different primes p, q, r and s as in the first question so that $pq = rs$. Then we would have an example where $p \mid r \times s$ but $p \nmid r$ and $p \nmid s$.

A Division Property of Primes We will first answer **Q2**.

Theorem 2.3.10. *If p is prime and $p \mid ab$ where a and b are natural numbers, then $p \mid a$ or $p \mid b$.*[26]

Proof. Suppose p is prime and $p \mid ab$.

If $p \mid a$ then we are done. So assume that $p \nmid a$. It follows that $\gcd(p, a) = 1$.

(Reason: Which natural numbers are divisors of both p and a? Because such a number divides p and p is prime, any divisor must be p or 1. But we have seen the divisor cannot be p since we are assuming that $p \nmid a$. So the only divisor of both p and a is 1.)

By the Extended Euclidean Algorithm there are integers s and t such that

$$1 = sa + tp.$$

Multiplying through by b we get

$$b = sab + tpb.$$

Because $p \mid ab$ and certainly $p \mid p$ it follows that $p \mid (sab + tpb)$. (See the discussion "Dividing Sums and Products" on page 25.) But this means $p \mid b$ and so we are done. □

For example, if someone (who you trust not to make a mistake!) told you that 7 divides 1,683,136 and that $1{,}683{,}136 = 952 \times 1768$, then you would know immediately that 7 divides at least one of 952 and 1768 (in fact, 7 divides 952 but not 1768).

We now see that the answer to **Q1** is NO.

For suppose p, q, r and s are primes and $p \times q = r \times s$. Then $p \mid rs$ and so, by Theorem 2.3.10, $p \mid r$ or $p \mid s$. But because r is prime the only divisors of r are 1 and r, and since any prime p is larger than one, it follows that if $p \mid r$ then $p = r$. Similarly if $p \mid s$ then $p = s$.

[26] In mathematics, if P and Q are statements and "P or Q" is true, then we mean that at least one of P or Q is true, but we also allow the possibility that both statements are true.

Thus p must equal one of r or s. Similarly, q must equal one of r or s.

Theorem 2.3.10 can be extended to the product of more than two natural numbers.

Corollary 2.3.11. [27] *If p is prime and $p \mid a_1 \cdot a_2 \cdot \ldots \cdot a_k$ where $a_1, \ldots, a_k$ are natural numbers, then p divides at least one of $a_1, \ldots, a_k$.*

Proof. If $k = 1$ then the result is of course immediate!

So suppose $k \geq 2$ and suppose p is prime and $p \mid a_1 a_2 \cdot \ldots \cdot a_k$.

From the previous Theorem $p \mid a_1$ or $p \mid a_2 \cdot \ldots \cdot a_k$. If $p \mid a_1$ we are done.

If $p \mid a_2 \cdot \ldots \cdot a_k$ then from the previous Theorem $p \mid a_2$ or $p \mid a_3 \cdot \ldots \cdot a_k$. If $p \mid a_2$ we are done.

Continuing in this way the last thing that can happen is $p \mid a_{k-1} a_k$. Then from the previous Theorem $p \mid a_{k-1}$ or $p \mid a_k$.

Thus p divides at least one of $a_1, a_2, \ldots, a_k$. □

In Theorem 2.3.10 we really showed something more general.

Theorem 2.3.12. *Suppose n, a, b are natural numbers, suppose $n \mid ab$ and suppose n and a are relatively prime. Then $n \mid b$.*

Proof. In the proof of Theorem 2.3.10 we only used the fact that p and a were relatively prime, but otherwise not the fact that p itself was prime. It then followed that $p \mid b$. The only difference here is that we are using the letter n instead of p. □

For example, if someone (who once again you trust not to make a mistake) told you that 12 divides 973,128,240 and that $973,128,240 = 35 \times 27,803,664$, then because 12 and 35 are relatively prime you would know immediately that 12 divides 27,803,664.

Uniqueness of Prime Factorisation We proved in Theorem 2.3.4 on page 26 that every natural number larger than 1 can be written as a product of primes.

For example $9857934 = 2 \times 3 \times 3 \times 547663$ and $988788377878738398 = 2 \times 3 \times 3 \times 13 \times 541 \times 7810704913967$.

We will now complete the proof of Theorem 2.3.4 by showing that we can write a natural number as a product of primes in essentially just *one* way.

Theorem 2.3.13. *Let n be a natural number and suppose*

$$n = p_1 p_2 \cdot \ldots \cdot p_j = q_1 q_2 \cdot \ldots \cdot q_k, \tag{2.20}$$

where $p_1, \ldots, p_j, q_1, \ldots, q_k$ are all primes. Then $j = k$ and both factorisations are the same except possibly for a reordering of the factors.

[27] A corollary is a theorem which is a relatively easy consequence or extension of a previous theorem.

Proof. Since $p_1 \mid n$ this means that $p_1 \mid q_1 q_2 \cdot \ldots \cdot q_k$. From Corollary 2.3.11 it follows that $p_1 \mid q_i$ for some i. But q_i is prime and so $p_1 = q_i$.

If we cancel p_1 and q_i from the left and right respectively of the second equality in (2.20) we get

$$p_2 \cdot \ldots \cdot p_j = q_1' q_2' \cdot \ldots \cdot q_{k-1}',$$

where $q_1', q_2', \ldots, q_{k-1}'$ are the primes remaining after cancelling q_i.

If we keep cancelling in this way we see that the number of factors on each side are the same and each factor occurs the same number of times on each side. This means that both factorisations are the same except possibly for a reordering of the factors. □

Questions

1 Prove that if $n \mid m$ and $n \mid k$ then $n \mid (m + k)$.
Here is a HINT.[28]

2 Prove that if $n \mid m$ and j is an integer then $n \mid mj$.
Here is a HINT.[29]

3 Prove that if $n \mid m$ and $n \mid k$ then $n \mid (m - k)$.

4 There are exactly 10^{150} natural numbers from 1 up to 10^{150}. What fraction of these natural numbers have exactly 150 digits?
Make life easier and first do the 10^2 case, the 10^3 case and the 10^4 case.

5 Use the Euclidean Algorithm to find by hand
a) $\gcd(1188, 385)$,
b) $\gcd(1177, 509)$.
DON'T LOOK NOW but the answer is in footnote[30] below.

6 Use the Euclidean algorithm calculations from the previous Question to
a) Express $\gcd(1188, 385)$ as a sum of multiples of 1188 and 385. Check your answer.
b) Express $\gcd(1177, 509)$ as a sum of multiples of 1177 and 509. Check your answer.

[28] We know $m = qn$ and $k = sn$ for certain integers q and s. What can you say about $m+k$?
[29] We know $m = qn$ for some integer q. What can you say about mj?
[30] The answers are 11 and 1 respectively.

2.4 Modular Arithmetic

Generalising a simple idea like telling time on a clock can lead to important applications.

Overview

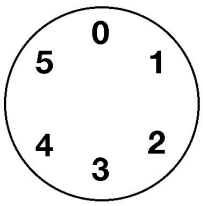
A new type of arithmetic, modular arithmetic, is discussed. For example, mod 6 arithmetic is done by just using the integers from $0, 1, 2, 3, 4, 5$. The sum and product of two integers mod 6 is the remainder after dividing the usual sum and product by 6.

Applications to bar codes and bank accounts are discussed. In Section 2.5 we will see that modular arithmetic is basic for internet security.

We prove some important Theorems for mod arithmetic which are needed for RSA cryptography and are basic in number theory. The most important of these is Fermat's Little Theorem

Examples of Modular Arithemetic

[HM, pp 82–87]



On Being Equivalent Mod 6 We say two integers are *equivalent mod 6* if they have the same remainder after division by 6. For example, 17 and 5 are equivalent mod 6 since $17 = 6 \times 2 + 5$. We write $17 \equiv 5 \bmod 6$ and say "17 is equivalent to 5 mod 6".

Often we just write $17 \equiv 5$ and say "17 is equivalent to 5", provided it is understood that (in this case) we mean "mod 6".

We can also do this for negative integers. For example, $-4 = 6 \times (-1) + 2$, and so $-4 \equiv 2 \bmod 6$.

The clock with the numbers $0, 1, 2, 3, 4, 5$ is a good way to visualise what is happening. As we move around the clock through $0, 1, 2, 3, 4, 5$ the next number is 0, and this corresponds to the fact $6 \equiv 0$. Similarly $7 \equiv 1$, $8 \equiv 2$. We can also move backwards (i.e. anticlockwise) from 0 and so we get $-1 \equiv 5$, $-2 \equiv 4, \ldots, -5 \equiv 1$, $-6 \equiv 0$, $-7 \equiv -1 \equiv 5$, etc.

More generally

$$
\begin{aligned}
-6 \equiv \mathbf{0} \equiv 6 &\equiv 12 \equiv 18 \equiv \ldots \bmod 6 \\
-5 \equiv \mathbf{1} \equiv 7 &\equiv 13 \equiv 19 \equiv \ldots \bmod 6 \\
-4 \equiv \mathbf{2} \equiv 8 &\equiv 14 \equiv 20 \equiv \ldots \bmod 6 \\
-3 \equiv \mathbf{3} \equiv 9 &\equiv 15 \equiv 21 \equiv \ldots \bmod 6 \\
-2 \equiv \mathbf{4} \equiv 10 \equiv 16 \equiv 22 &\equiv \ldots \bmod 6 \\
-1 \equiv \mathbf{5} \equiv 11 \equiv 17 \equiv 23 &\equiv \ldots \bmod 6
\end{aligned}
$$

Every integer, positive or negative, is equivalent mod 6 to one of the integers $0, 1, 2, 3, 4, 5$.

Adding and Multiplying Mod 6 What happens if we add 2 large integers and we want the result mod 6? For example, what is 2137 + 512 mod 6? We can work out the answer as follows:

$$2137 + 512 = (6 \times 356 + 1) + (6 \times 85 + 2) \equiv (0 + 1) + (0 + 2) = 3 \bmod 6.$$

In general, if we are adding two or more integers and want the result mod 6 we just replace each number by its remainder after dividing by 6, then add the remainders, then take the remainder mod 6 once again if necessary.

For example, working mod 6

$$135 + 91 + 46 \equiv 3 + 1 + 4 = 8 \equiv 2 \bmod 6.$$

Likewise if we multiply 2137 and 512 and want the answer mod 6:

$$\begin{aligned} 2137 \times 512 &= (6 \times 356 + 1) \times (6 \times 85 + 2) = (6 \times 356 \times 6 \times 85) \\ &\quad + (6 \times 356 \times 2) + (1 \times 6 \times 85) + 1 \times 2 \equiv 1 \times 2 = 2 \bmod 6. \end{aligned}$$

In general, if we are multiplying two integers and want the answer mod 6 we can replace any multiple of 6 by 0. So we would usually shorten the previous calculation to

$$2137 \times 512 = (6 \times 356 + 1) \times (6 \times 85 + 2) \equiv 1 \times 2 = 2 \bmod 6.$$

Likewise, if we are multiplying more than two integers and want the answer mod 6 we can replace any multiple of 6 by 0. Equivalently, we replace each number by its remainder after dividing by 6, and *then* multiply.

Finally, a similar thing happens with subtraction. For example, working mod 6,

$$182 - 93 \equiv 2 - 3 = -1 \equiv 5 \bmod 6.$$

Now try Question 1 on page 50 to test your understanding.

⋆Exponentiating Mod Wise

The method in Q2 here is n

Q1. What is 2136^{1035} mod 7? Now it would not be very smart to first multiply 2136 by itself 1035 times (the result is 3447 digits long) and then divide by 7. It would take quite a while and you would probably make a mistake. Here is a better way:

$$2136^{1035} = (7 \times 305 + 1)^{1035} \equiv 1^{1035} = 1 \bmod 7$$

Next comes a trickier question. It is the type of calculation that needs to be done for RSA cryptography, see Section 2.5. In practice the calculation will be done by a computer, but you should try to understand the idea.

Q2. What is 507^{107} mod 14?

$$507^{107} = (14 \times 36 + 3)^{107} \equiv 3^{107} \bmod 14.$$

What next? We don't really want to multiply 3 by itself 107 times.

In cases like this we write the exponent 107 as a sum of powers of 2.

First note that

$$2^1 = 1,\ 2^2 = 4,\ 2^3 = 8,\ 2^4 = 16,\ 2^5 = 32,\ 2^6 = 64,\ 2^7 = 128,\ 2^8 = 256,$$
$$2^9 = 512,\ 2^{10} = 1024,\ 2^{11} = 2048,\ 2^{12} = 4096, \ldots$$

By subtracting off the highest possible power of 2 at each step we get

$$107 = 2^6 + 43 = 2^6 + 2^5 + 11 = 2^6 + 2^5 + 2^3 + 3 = 2^6 + 2^5 + 2^3 + 2^1 + 1.$$

So we can write

$$3^{107} = 3^{2^6+2^5+2^3+2^1+1} = 3^{64+32+8+2+1} = 3^{64} \times 3^{32} \times 3^8 \times 3^2 \times 3.$$

Next successively compute $3^2, 3^4, 3^8, 3^{16}, 3^{32}, 3^{32}, 3^{64}$ mod 14.

$$\begin{aligned}
3^2 &= 9 \\
3^4 &= (3^2)^2 \equiv 9^2 \text{ (by previous step) } \equiv 11 \\
3^8 &= (3^4)^2 \equiv 11^2 \text{ (by previous step) } = 121 \equiv 9 \\
3^{16} &= (3^8)^2 \equiv 9^2 \text{ (by previous step) } = 81 \equiv 11 \\
3^{32} &= (3^{16})^2 \equiv 11^2 \text{ (by previous step) } = 121 \equiv 9 \\
3^{64} &= (3^{32})^2 \equiv 9^2 \text{ (by previous step) } = 81 \equiv 11.
\end{aligned}$$

(See how the calculation follows a pattern.)

Putting this all together we get mod 14 that

$$\begin{aligned}
3^{107} &= 3^{64+32+8+2+1} = 3^{64} \times 3^{32} \times 3^8 \times 3^2 \times 3 \\
&\equiv 11 \times 9 \times 9 \times 9 \times 3 = 99 \times 81 \times 3 \text{ (for example) } \equiv 1 \times 11 \times 3 = 33 \equiv 5.
\end{aligned}$$

So the answer to our question is that

$$507^{107} \equiv 5 \bmod 14.$$

Now try Questions 2 and 3 to test your understanding.

Tables for Mod Arithmetic Here are the tables for mod 1 up to mod 8 arithmetic. *Make sure you check them.*

When we write "$\oplus$" (or "$\otimes$") we mean first do ordinary addition (or multiplication) and then take the remainder mod n. With this new type of addition and multiplication we often just operate on numbers from 0 up to $n-1$, in any case the result is a number in this range.

mod 1

$\oplus$	0
0	0

$\otimes$	0
0	0

mod 2

$\oplus$	0	1
0	0	1
1	1	0

$\otimes$	0	1
0	0	0
1	0	1

mod 3

$\oplus$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\otimes$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

mod 4

$\oplus$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

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

mod 5

$\oplus$	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

mod 6

$\oplus$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

$\otimes$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

mod 7

$\oplus$	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

$\otimes$	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

mod 8

$\oplus$	0	1	2	3	4	5	6	7
0	0	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7	0
2	2	3	4	5	6	7	0	1
3	3	4	5	6	7	0	1	2
4	4	5	6	7	0	1	2	3
5	5	6	7	0	1	2	3	4
6	6	7	0	1	2	3	4	5
7	7	0	1	2	3	4	5	6

$\otimes$	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7
2	0	2	4	6	0	2	4	6
3	0	3	6	1	4	7	2	5
4	0	4	0	4	0	4	0	4
5	0	5	2	7	4	1	6	3
6	0	6	4	2	0	6	4	2
7	0	7	6	5	4	3	2	1

See also Table 2.1 on page 64 for the mod 20 multiplication table.

Patterns in the Mod Tables Looking at the tables we observe:

- If two numbers are added or multiplied, the *order of addition and multiplication does not matter.*

  That is, $a \oplus b = b \oplus a$ and $a \otimes b = b \otimes a$. This is not surprising, since $a + b = b + a$ (ordinary addition) and so $a + b$ and $b + a$ each have the same remained when divided by n. This means that $a \oplus b = b \oplus a$.

  Similarly, $a \times b = b \times a$ (ordinary multiplication) and so $a \times b$ and $b \times a$ each have the same remainder when divided by n. This means that $a \otimes b = b \otimes a$.

- In the *addition* tables each row is a cyclic rearrangement of the top row.
  In each row of the mod n addition table, every number $0, 1, 2, \ldots, n-1$ occurs exactly once.
- Next look at the *multiplication* tables mod n.
  If we multiply by 0 then every entry in that row is 0.
  What happens if we multiply by a number other than 0?
  1. *Multiplication by r where r and n have no common factor other than 1.* In other words $\gcd(r, n) = 1$. This always happens if n is prime and sometimes happens for other n; such as $n = 8$ and $r = 1, 3, 5$; and such as $n = 6$ and $r = 1, 5$.
     In this case the numbers in the row corresponding to r each occur exactly once.
  2. *Multiplication by r where r and n have a common factor larger than 1.*
     In this case some numbers do not occur in the corresponding row while others occur more than once.

✍ *Check every row in every multiplication table.*

- Here is an important consequence of the preceding observation.
  Suppose r is a natural number less than n, and r and n have no common factor other than 1, or in other words $\gcd(r, n) = 1$. Then there is exactly one natural number s less than n such that $rs \equiv 1 \bmod n$.
  We will prove this in Theorem 2.4.1.
  We say that s is the *multiplicative inverse* of r mod n.

✍ *Have a look at the mod 5 and mod 6 multiplication tables and find the multiplicative inverse of r if $r = 1, 2, 3, 4$ in the mod 5 case and if $r = 1, 5$ in the mod 6 case.*

⋆*Properties of Mod Arithmetic*

This material is not in [HM].

Addition and Multiplication Properties Mod n arithmetic satisfies the following properties which are also satisfied by ordinary arithmetic with integers. Notice that the analogue of 1(d) is true for integers but not for the natural numbers.

Here we only work with integers a, b, c in the range $0, 1, 2, \ldots, n-1$. When we "add" or "multiply" them with $\oplus$ or $\otimes$ we again get integers in the range $0, 1, 2, \ldots, n-1$.

1. Addition
   a) $a \oplus b = b \oplus a$
   b) $a \oplus (b \oplus c) = (a \oplus b) \oplus c$
   c) $a \oplus 0 = a$
   d) For each number a (from $0, 1, \ldots, n-1$) there is exactly one number b (from $0, 1, \ldots, n-1$) such that $a \oplus b = 0$
2. Multiplication
   a) $a \otimes b = b \otimes a$
   b) $a \otimes (b \otimes c) = (a \otimes b) \otimes c$
   c) $a \otimes 1 = a$

3. Connection between addition and multiplication
   $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$

You can check that these properties are true for each table. But that is very tedious and will not tell you what happens in general. However, we will now explain why these properties are *always* true for mod arithmetic.

In the case of 1(a), 1(b), 1(c), 2(a), 2(b), 2(c), and 3, we know that these properties are true for ordinary addition and multiplication. For example, $a \times b = b \times a$ for ordinary multiplication. It follows that $a \times b$ and $b \times a$ have the same remainder after dividing by n. In other words $a \otimes b = b \otimes a$.

A similar argument works for the other properties mentioned.
Write out the similar argument for 3.

In the case of 1(d) $b = 0$ if $a = 0$ and $b = n - a$ if $a = 1, 2, \ldots, n - 1$.
Convince yourself this is true.

Modular Inverses Here is the Theorem mentioned in the last dot point on page 40 concerning Patterns in the Mod Tables. Moreover, we will see how we can calculate the multiplicative inverse.

In the following Theorem, s is called the *multiplicative inverse of r mod n*. Usually $r < n$, but this is not actually needed.

Theorem 2.4.1. *Suppose that r and n are natural numbers which are relatively prime. Then there is exactly one natural number $s < n$ such that $rs \equiv 1 \bmod n$.*

Proof. There are two facts we need to prove:
1. There is at least one natural number $s < n$ such that $rs \equiv 1 \bmod n$.
2. If s_1 and s_2 are two natural numbers both less than n such that $rs_1 \equiv 1 \bmod n$ and $rs_2 \equiv 1 \bmod n$, then $s_1 = s_2$.

Proof of 1. Because r and n are relatively prime, $\gcd(r, n) = 1$.

We know from Theorem 2.3.9 that there are integers $\hat{s}$ and t such that

$$1 = \hat{s}r + tn.$$

Because $n \mid tn$,

$$1 \equiv \hat{s}r \bmod n. \tag{2.21}$$

We only know $\hat{s}$ is an integer, it may be negative or $\geq n$.[31]

However, if we add or subtract enough multiples of n to $\hat{s}$ we will get a number $s = \hat{s} + kn$ in the range $0, \ldots, n - 1$. The number k may be a positive or negative integer. For example, in the following diagram $k = 2$:

[31] For example, in (2.19) we saw that $\gcd(24, 245) = 1$, and in the discussion just before Theorem 2.3.9 we obtained

$$1 = (-51) \times 24 + 5 \times 245.$$

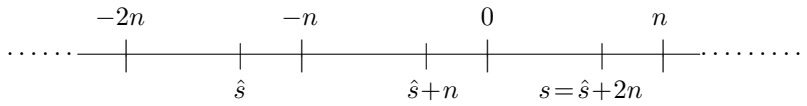
It follows that

$$1 \equiv (-51) \times 24 \bmod 245.$$

The problem is that -51 is certainly not in the range $1, \ldots, 244$. However, if we add 245 to -51 we get 194. It follows that

$$1 \equiv (-51) \times 24 \equiv (-51 + 245) \times 24 = 194 \times 24 \bmod 245.$$

So the multiplicative inverse of 24 mod 245 is 194.



It follows that working mod n

$$\begin{aligned} sr &= (\hat{s} + kn)r \\ &\equiv \hat{s}r \text{ (since } n \mid kn) \\ &\equiv 1 \bmod n \text{ (from (2.21)).} \end{aligned}$$

We know s is in the range $0, \ldots, n-1$. But we cannot have $s = 0$ since then $sr \equiv 0 \bmod n$. So s is actually in the range $1, \ldots, n-1$.

This completes the proof of 1.

Proof of 2. Suppose s_1 and s_2 are two natural numbers, both less than n, such that

$$rs_1 \equiv 1 \bmod n, \quad rs_2 \equiv 1 \bmod n.$$

For convenience, if s_1 and s_2 are not equal we let s_1 be the larger of the two numbers. In any case, $s_1 \geq s_2$.

By subtraction, $rs_1 - rs_2 \equiv 0 \bmod n$.

It follows that $r(s_1 - s_2) \equiv 0 \bmod n$.

This means $n \mid r(s_1 - s_2)$. But r and n are relatively prime, and so from Theorem 2.3.12 on page 34 it follows that $n \mid (s_1 - s_2)$. That is, $s_1 - s_2$ is a multiple of n.

But $s_1 \geq s_2$ and both are less than n, and so $0 \leq s_1 - s_2 < n$.

These two facts together imply $s_1 - s_2 = 0$, i.e. $s_1 = s_2$. □

Applications of Modular Arithmetic

[HM, 87–88]

Barcodes The barcodes used on goods in Australia appear to mostly follow the European Article Numbering Code EAN-13, a different scheme from that discussed in [HM].

In these barcodes there are 13 digits. A typical example is 9 315999 091207, i.e. one digit then six digits and then another six digits. The first two digits "93" are reserved for Australia. The next five (here 15999) identify the manufacturer. The next five (here 09120) identify the product. The last digit (here 7) is called the *check digit*.

If we write the digits as $d_1, d_2, d_3, \ldots, d_{13}$ then the *weighted sum* S is defined by

$$\begin{aligned} S &= d_1 + 3d_2 + d_3 + 3d_4 + d_5 + 3d_6 + d_7 + 3d_8 + d_9 + 3d_{10} + d_{11} + 3d_{12} + d_{13} \\ &= (\text{sum of odd numbered digits}) \quad + \quad 3 \times (\text{sum of even numbered digits}). \end{aligned} \tag{2.22}$$

We say that the odd digits have *weight 1* and the even digits have *weight 3*.

The *check digit* d_{13} is always selected so

$$S \equiv 0 \bmod 10. \tag{2.23}$$

With d_{13} chosen in this way we say the bar code is *valid*. A computer can easily and automatically check the validity.

In the example 9 315999 091207, working mod 10 we get

$$S = (9+1+9+9+9+2+7) + 3 \times (3+5+9+0+1+0) \equiv 6 + 3 \times 8 \equiv 0.$$

(It is easy to do this in your head working mod 10: for example $9 + 1 = 0$, adding another 9 gives 9, another 9 gives 8, another 9 gives 7, adding 2 gives 9, adding 7 gives 6, etc.) So this is a valid barcode.

Detecting Barcode Errors We now examine some common errors that can be picked up by checking if (2.23) is true.

Changing One Digit Suppose the second digit in 9 315999 091207 is mistakenly increased by 4 from 3 to 7. The original weighted sum S satisfies

$$S \equiv 0 \bmod 10.$$

Since the digit which was changed was the second it is an even numbered digit. From (2.22) the new weighted sum is $S + 3 \times 4 = S + 12$, and this satisfies

$$S + 12 \equiv 0 + 2 = 2 \bmod 10.$$

So we see the new bar code is *not* a valid bar code.

In general, suppose a single *even* numbered digit is mistakenly changed by n, perhaps increased and perhaps decreased.

The original weighted sum was S and this must be equivalent to 0 mod 10. The new weighted sum will be $S \pm 3n$. So the new weighted sum will be equivalent to 0 mod 10 if and only if $3n \equiv 0 \bmod 10$, which is the same as requiring that the remainder after dividing $3n$ by 10 be 0.

We can draw up a table for the different possible values of n.

n	1	2	3	4	5	6	7	8	9
3n	3	6	9	12	15	18	21	24	27
remainder after dividing 3n by 10	3	6	9	2	5	8	1	4	7

This means the new weighted sum $S \pm 3n$ is *never* equivalent to 0 mod 10 and so the new bar code is not valid.

If a single *odd* numbered digit is changed by n then it is easier to see what happens. The new weighted sum will be $S \pm n$. This will be equivalent to 0 mod 10 if and only if n is a multiple of 10, which never happens as n will be one of the numbers $1, 2, \ldots, 9$. This again means the new weighted sum $S \pm n$ is *never* equivalent to 0 mod 10 and so the new bar code is not valid.

The final conclusion is that if any single digit in a barcode is inadvertently changed then the bar code will no longer be a valid bar code.

Transposing Two Digits Suppose the first digit d_1 and the *following* digit d_2 in a valid barcode are different and they are accidently switched. (In our example this would mean changing 9 315999 091207 to 3 915999 091207.) Then the weighted sum will change from

$$S = (d_1 + 3d_2) + d_3 + 3d_4 + d_5 + 3d_6 + d_7 + 3d_8 + d_9 + 3d_{10} + d_{11} + 3d_{12} + d_{13}$$

to

$$S^* = (d_2 + 3d_1) + d_3 + 3d_4 + d_5 + 3d_6 + d_7 + 3d_8 + d_9 + 3d_{10} + d_{11} + 3d_{12} + d_{13}.$$

The change in the weighted sum is $S^* - S = 2d_1 - 2d_2$, i.e. $S^* = S + 2(d_1 - d_2)$.

Remember that the original barcode is a valid one and so the weighted sum S is divisible by 10. The new weighted sum S^* will be divisible by 10 if and only if $2(d_1 - d_2)$ is divisible by 10. The only way this can happen is if d_1 and d_2 differ by 5 (remember that we are assuming d_1 and d_2 are different, since if they are equal there is no problem with switching them).

Thus by doing a barcode check we can see there is an error if the first 2 digits are switched, *unless* these 2 digits differ by exactly 5.

Almost the same argument works if we switch any 2 different *adjacent* digits. The new bar code will be not be valid *unless* the 2 switched digits differ by 5.

In the particular case of our example 9 315999 091207, *every* switch of different adjacent digits would be picked up as an error. But in the example 9 315999 097207 if the 72 was switched to 27 then the fact that a mistake was made will not be detected.

Other Error Checking Methods For cheques see [HM, page 88]. The last 9 digits are the ones to look at.

For ISBN numbers (on books) see [HM, pp 92,93, Q 32–34]. For airline tickets [HM, pp 92, Q 29].

Now try Questions 5 and 6.

⋆*More Properties of Modular Arithmetic*

Most of the remaining material on Modular Arithmetic is not in [HM]. But it will help you understand why RSA cryptography works.

We will need the ideas and Theorem in this Section when we explain the mathematics behind RSA cryptography in Section 2.5.

Tables of Powers Here are tables of powers for mod 2 up to mod 8.

For example, in the mod 5 table in the column under $a = 3$ we have computed the powers $3^1, 3^2, 3^3, 3^4, 3^5, 3^6, 3^7$ and taken the remainder after dividing each by 5. This gives $3, 4, 2, 1, 3, 4, 2$. Likewise, in the column under $a = 4$ we have computed the powers $4^1, 4^2, 4^3, 4^4, 4^5, 4^6, 4^7$ and taken the remainder after dividing each by 5. This gives $4, 1, 4, 1, 4, 1, 4$.

Again in the mod 5 table, in the row corresponding to a^4 we have the powers $0^4, 1^4, 2^4, 3^4, 4^4, 5^4, 6^4, 7^4$ after dividing each by 5 and taking the remainder. This gives $0, 1, 1, 1, 1, 0, 1, 1$. Likewise in the row corresponding to a^5 we have the powers $0^5, 1^5, 2^5, 3^5, 4^5, 5^5, 6^5, 7^5$ after dividing each by 5 and taking the remainder. This gives $0, 1, 2, 3, 4, 0, 1, 2$.

powers mod 3

a	0	1	2	3	4	5	6
a^1	0	1	2	0	1	2	0
a^2	0	1	1	0	1	1	0
a^3	0	1	2	0	1	2	0
a^4	0	1	1	0	1	1	0
a^5	0	1	2	0	1	2	0

powers mod 4

a	0	1	2	3	4	5	6
a^1	0	1	2	3	0	1	2
a^2	0	1	0	1	0	1	0
a^3	0	1	0	3	0	1	0
a^4	0	1	0	1	0	1	0
a^5	0	1	0	3	0	1	0
a^5	0	1	0	1	0	1	0

powers mod 5

a	0	1	2	3	4	5	6	7
a^1	0	1	2	3	4	0	1	2
a^2	0	1	4	4	1	0	1	4
a^3	0	1	3	2	4	0	1	3
a^4	0	1	1	1	1	0	1	1
a^5	0	1	2	3	4	0	1	2
a^6	0	1	4	4	1	0	1	4
a^7	0	1	3	2	4	0	1	3

powers mod 6

a	0	1	2	3	4	5	6	7	8
a^1	0	1	2	3	4	5	0	1	2
a^2	0	1	4	3	4	1	0	1	4
a^3	0	1	2	3	4	5	0	1	2
a^4	0	1	4	3	4	1	0	1	4
a^5	0	1	2	3	4	5	0	1	2
a^6	0	1	4	3	4	1	0	1	4
a^7	0	1	2	3	4	5	0	1	2
a^8	0	1	4	3	4	1	0	1	4

powers mod 7

a	0	1	2	3	4	5	6	7	8	9
a^1	0	1	2	3	4	5	6	0	1	2
a^2	0	1	4	2	2	4	1	0	1	4
a^3	0	1	1	6	1	6	6	0	1	1
a^4	0	1	2	4	4	2	1	0	1	2
a^5	0	1	4	5	2	3	6	0	1	4
a^6	0	1	1	1	1	1	1	0	1	1
a^7	0	1	2	3	4	5	6	0	1	2
a^8	0	1	4	2	2	4	1	0	1	4
a^9	0	1	1	6	1	6	6	0	1	1

powers mod 8

a	0	1	2	3	4	5	6	7	8	9	10
a^1	0	1	2	3	4	5	6	7	0	1	2
a^2	0	1	4	1	0	1	4	1	0	1	4
a^3	0	1	0	3	0	5	0	7	0	1	0
a^4	0	1	0	1	0	1	0	1	0	1	0
a^5	0	1	0	3	0	5	0	7	0	1	0
a^6	0	1	0	1	0	1	0	1	0	1	0
a^7	0	1	0	3	0	5	0	7	0	1	0
a^8	0	1	0	1	0	1	0	1	0	1	0
a^9	0	1	0	3	0	5	0	7	0	1	0
a^{10}	0	1	0	1	0	1	0	1	0	1	0

See also Table 2.2 on page 65 for the mod 33 power table.

There are simpler ways to calculate the tables than by first finding every power and then finding the remainder. For example in the mod 7 table with the column corresponding to powers of 4 we have, *working mod 7*,

$$
\begin{aligned}
4^1 &= 4 \equiv 4,\\
4^2 &= 16 \equiv 2,\\
4^3 &= 4^2 \times 4 \equiv 2 \times 4 \text{ (from previous line)} = 8 \equiv 1,\\
4^4 &= 4^3 \times 4 \equiv 1 \times 4 \text{ (from previous line)} = 4 \equiv 4,\\
4^5 &= 4^4 \times 4 \equiv 4 \times 4 \text{ (from previous line)} = 16 \equiv 2,\\
4^6 &= 4^5 \times 4 \equiv 2 \times 4 \text{ (from previous line)} = 8 \equiv 1,\\
4^7 &= 4^6 \times 4 \equiv 1 \times 4 \text{ (from previous line)} = 4 \equiv 4,\\
4^8 &= 4^7 \times 4 \equiv 4 \times 4 \text{ (from previous line)} = 16 \equiv 2,\\
4^9 &= 4^8 \times 4 \equiv 2 \times 4 \text{ (from previous line)} = 8 \equiv 1.
\end{aligned}
$$

In effect we just go down the column under 4, beginning with 4 and at each stage multiplying the previous entry by 4 and taking the remainder mod 7. The entries are $4, 2, 1, 4, 2, 1, 4, 2, 1$. As soon as you come to a number which has already occurred, in this case 4, the next number under the second 4 will be the same as the next number under the first 4 (namely 2). Look at the column in the mod 7 table under 6 and under 5 and see how the same idea occurs. Look at the column under 6 in the mod 8 table for another example. *Can you explain why this happens as you move down a column?*

There are further ways to simplify the calculations. Since $0 \equiv 7 \bmod 7$, the powers of 0 and the powers of 7 are equal mod 7. This means that in the mod 7 table the column under 7 is the same as the column under 0. Similarly the column under 8 is the same as the column under 1, the column under 9 is the same as the column under 2, the column under 10 (which we do not show) is the same as the column under 3, etc. Similarly in the mod 8 table the columns repeat themselves from the column under 8 onwards to the right.

In summary, in the mod n table we only need to compute the columns $0, 1, 2, 3, \ldots, n-1$. After this the columns repeat themselves. We do not need to go across the table any further to the right than $n-1$. Notice also that the

column under 0 always consists just of 0's, and the column under 1 consists just of 1's.

In each table there is a box around the "interesting" cases from which all other cases can be computed easily. We could just as well have left out the column corresponding to powers of 1, although it is more common to include it.

Work out and write down the mod 9 table for a, a^2, ..., a^9 where $a = 0, 1, 2, \ldots, 8$. Explain what extra rows and columns would look like.

Patterns in the Power Tables If you look at the mod n tables where n *is a prime* you will see two things:

- The last row in each box is the same as the first and consists of the numbers $1, 2, \ldots, n-1$.
- The second last row in each box consists entirely of 1's.

Neither of these observations hold for the mod n tables in general, but they do hold whenever n is a prime number. We will prove this in Fermat's Little Theorem, Theorem 2.4.2.

Fermat's Little Theorem (Not to be confused with Fermat's Last Theorem, which we have already mentioned.)

Explain to another student what Fermat's Little Theorem, Theorem 2.4.2, says about the power mod tables for 3, 5 and 7.

Notice that there are different but equivalent ways of stating the hypotheses and the conclusions of the Theorem. For example, "a is not a multiple of p" is *equivalent* to "p is not a factor of a". Also, "$a^{p-1} \equiv 1 \bmod p$" is *equivalent* to "$p \mid (a^{p-1} - 1)$".

We will give two different proofs[32] of the Theorem. We will also see that (2.24) follows easily from (2.25) (end of first proof) and that (2.25) follows easily from (2.24) (end of second proof).

The two proofs of Fermat's just take a brief look at this that is excellent.

Theorem 2.4.2. *Suppose p is a prime and a is a natural number. Then*

$$a^p \equiv a \mod p \tag{2.24}$$

$$a^{p-1} \equiv \begin{cases} 1 \mod p, & \text{if } a \text{ is not a multiple of } p \\ 0 \mod p, & \text{if } a \text{ is a multiple of } p \end{cases} \tag{2.25}$$

First Proof. We will prove (2.25) first.

If a is a multiple of p then

$$a^{p-1} \equiv 0^{p-1} = 0 \bmod p.$$

So next we consider the case a is not a multiple of p.

We begin by looking at the remainders obtained by dividing each of the numbers

$$a, 2a, 3a, \ldots, (p-1)a$$

[32] The proofs are challenging but are included for completeness.

by p.

The first thing to notice is that *the remainder in each case is not 0.* The reason is that if p divides sa then p must divide at least one of s or a (by Theorem 2.3.10). But we are assuming $p \nmid a$ and we know $p \nmid s$ since s is a natural number less than p. So $p \nmid sa$.

The next thing to notice is that *the remainders are different in each case.* The reason for this is that if sa and ta (let s be the larger of s and t) have the same remainder after being divided by p then it follows that $p \mid (sa - ta)$ and so $p \mid (s-t)a$. But this implies that p divides at least one of $s-t$ and a (again by Theorem 2.3.10). However, we are assuming $p \nmid a$. And since $s-t$ is a natural number less than p it follows that $p \nmid (s-t)$. Hence $p \nmid (s-t)a$.

We now know that:

> *The remainders obtained after dividing the $p-1$ numbers a, $2a$, $3a$, ..., $(p-1)a$ by p are all different. None equal 0 and so they take each of the $p-1$ values $1, 2, 3, \ldots, p-1$ exactly once.*

We don't know actually know which remainder occurs in each case, but this will not matter.[33]

From this we make the following *very cunning observation.* If we multiply $a, 2a, 3a, \ldots, (p-1)a$ together and take the remainder after dividing by p, then we get the same remainder as if we multiply $1, 2, 3, \ldots, p-1$ together and take the remainder after dividing by p.[34]

In symbols:

$$a \times 2a \times 3a \times \cdots \times (p-1)a \equiv 1 \times 2 \times 3 \times \cdots \times (p-1) \bmod p, \qquad (2.26)$$

In (2.26) we are multiplying $p-1$ terms together on each side and so

$$(p-1)!\, a^{p-1} \equiv (p-1)! \bmod p.$$

This means both sides have the same remainder after division by p. It follows their difference has remainder 0, i.e. their difference is divisible by p:

$$(p-1)!\,(a^{p-1} - 1) \equiv 0 \bmod p.$$

---

[33] Here are two examples.

In the case $p = 7$ and $a = 4$ the remainders after dividing 4, 8, 12, 16, 20 and 24 by 7 are 4, 1, 5, 2, 6 and 3 respectively. The remainders take the values 1, 2, 3, 4, 5, 6 each exactly one, but in a very mixed up way.

In the case $p = 5$ and $a = 3$ the remainders after dividing 3, 6, 9 and 12 by 5 are 3, 1, 4 and 2 respectively. In the second case the remainders take the values 1, 2, 3, 4 each exactly one, but again in a very mixed up way.

Do a similar examination for the case $p = 7$ and $a = 5$.

[34] Let's look at the previous examples.

In the first case, working mod 7,

$$\begin{aligned} 4 \times 8 \times 12 \times 16 \times 20 \times 24 &\equiv 4 \times 1 \times 5 \times 2 \times 6 \times 3 \\ &\equiv 1 \times 2 \times 3 \times 4 \times 5 \times 6 \bmod 7. \end{aligned}$$

In the second case, working mod 5,

$$3 \times 6 \times 9 \times 12 \equiv 3 \times 1 \times 4 \times 2 \equiv 1 \times 2 \times 3 \times 4 \bmod 5.$$

Do a similar examination for the case $p = 7$ and $a = 5$.

Since the prime p does not divide any of the numbers 1, 2, ..., $p-1$, it follows that p divides $(a^{p-1}-1)$. In other words,

$$a^{p-1} \equiv 1 \mod p.$$

This completes the proof of (2.25).

In order to prove (2.24) in case p does not divide a, multiply both sides of the first equation in (2.25) by a. It follows that

$$a^p \equiv a \bmod p$$

if $p \nmid a$.

The only case remaining is to show (2.24) is true when a is a multiple of p. But in this case both sides of (2.24) equal 0 mod p and so (2.24) is true. □

Second Proof. We will use mathematical induction on a to first prove (2.24).[35]

Let p be a fixed prime number. Let $P(a)$ be the statement "$a^p \equiv a \bmod p$".

The *basic step* is to prove that $P(1)$ is true. But this is certainly the case because if $a = 1$ both a^p and a are equal to 1.

For the *inductive step* assume that $P(a)$ is true for some natural number a. Our goal is to deduce from this that $P(a+1)$ is true.

By the binomial theorem[36]

$$\begin{aligned}(a+1)^p =& a^p + pa^{p-1} + \frac{p(p-1)}{2!}a^{p-2} + \frac{p(p-1)(p-2)}{3!}a^{p-3} + \cdots \\ &+ \frac{p(p-1)(p-2)\cdot \ldots \cdot 2}{(p-1)!}a + 1.\end{aligned} \tag{2.27}$$

Apart from the first and last terms which are a^p and 1, the other terms are all of the form

$$\frac{p(p-1)(p-2)\cdots(p-(k-1))}{k!}a^{p-k}$$

[35] It does not make sense to use induction on p because if p is a prime it does not follow that $p+1$ is a prime.

[36] By multiplying out terms, you can check that

$$\begin{aligned}(a+b)^2 &= a^2 + 2ab + b^2 \\ (a+b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3 \\ (a+b)^4 &= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4 \\ (a+b)^5 &= a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5 \\ (a+b)^6 &= a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6 \\ &\vdots\end{aligned}$$

The general formula is called the "binomial formula" or "binomial theorem" and is

$$\begin{aligned}(a+b)^n =& a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}b^3 + \cdots \\ &+ \frac{n(n-1)(n-2)\cdots(2)}{(n-1)!}ab^{n-1} (= nab^{n-1}) + b^n.\end{aligned}$$

You will see it later in your other maths courses.

where $k = 1, 2, \ldots, p-1$.

Each coefficient $\dfrac{p(p-1)(p-2)\cdots(p-(k-1))}{k!}$ is an integer. This means that $k!$ divides $p(p-1)(p-2)\cdots(p-(k-1))$. But the p in the numerator cannot be cancelled by any term in the denominator $k!$ since all terms in $k!$ are less than p. It follows that $k!$ must divide $(p-1)(p-2)\cdots(p-(k-1))$. This implies that every term in (2.27) apart from the first and the last can be written in the form $p \times$ "natural number".

It follows that

$$(a+1)^p = a^p + p \times (\text{some natural number}) + 1.$$

Hence

$$\begin{aligned}(a+1)^p &\equiv a^p + 1 \bmod p \\ &\equiv a + 1 \bmod p, \text{ because we are assuming } P(a) \text{ is true.}\end{aligned}$$

In other words, $P(a+1)$ is true.

This means we have shown the inductive step.

It follows from the Principle of Mathematical Induction that $P(a)$ is true for every natural number a. That is, we have proved (2.24).

To prove (2.25) we first assume that $p \nmid a$. From what we have just proved, $p \mid (a^p - a)$ for every a, which means $p \mid a(a^{p-1} - 1)$. Because $p \nmid a$ it follows that $p \mid (a^{p-1} - 1)$. In other words, $a^{p-1} \equiv 1 \bmod p$.

If $p \mid a$ then $p \mid a^{p-1}$ and so $a^{p-1} \equiv 0 \bmod p$.

This completes the second proof of the Theorem. □

Questions

1 Find the number between 0 and 6 which is equivalent mod 7 to each of the following. The method you should use is to replace each number by its remainder after dividing by 7. And keep doing this.

1. $75 \times (37 \times 912 + 356)$
2. $96 \times 95 \times 94 \times 93 \times 92$

2 Use the method for Q2 on page 37 to find $14^2 \bmod 55$, $14^4 \bmod 55$, $14^8 \bmod 55$ and $14^{16} \bmod 55$.

Then use this information to find $14^{27} \bmod 55$.

The answer is in footnote[37]below.

3 Use the method for Q2 on page 37 to find $675^{307} \bmod 713$.

The answer is in footnote[38]below.

4 Work out the mod 9 addition and multiplication tables.

What patterns do you observe along the lines of the discussion on page 40?

5 Use the method on page 43 to:

a) Find which errors on an airline ticket are *not* detected if a single digit is changed.

[37]The answers are $14^2 = 31 \bmod 55$, $14^4 = 26 \bmod 55$, $14^8 = 16 \bmod 55$, $14^{16} = 36 \bmod 55$ and $14^{27} = 9 \bmod 55$.

[38]The answer is 3.

b) Find which errors on an airline ticket are *not* detected if 2 adjacent digits are transposed.

The answer is in footnote[39]below.

6 Use the method on page 43 to:

a) Show that *all* errors are detected on an ISBN number if a single digit is changed.

b) Show that *all* errors are detected on an ISBN number if any 2 (not necessarily adjacent) digits are switched.

[39]If a single digit is changed then the error is not detected if the digit is changed by 7, i.e. from 0 to 7 or 7 to 0, 1 to 8 or 8 to 1, 2 to 9 or 9 to 2. If two adjacent digits are transposed then the error is not detected if the digits differ by 7, i.e. if the adjacent digits are 07, 70, 18, 81, 29, or 92.

2.5 RSA Public Key Cryptography

Things that seem abstract and devoid of application today may be central in our daily lives tomorrow.

Overview

The theory and practice of RSA cryptography is developed. RSA cryptography was discovered in the 1970s. You use RSA cryptography whenever you encrypt material on your hard disk, buy something from eBay or Amazon.com, do any other secure or secret transaction over the internet, or use EFTPOS or an ATM.

RSA cryptography is at first almost unbelievable. You publish certain information on your website (see page 61) explaining to anyone in the world how they can take a secret message they want to send you, encode this secret message as a "coded" public number, and publish this coded number on their website.

Anyone in the world can see this coded number, but only you can decode it and get back the original mesage. Everyone in the world can send you secret messages, they all use exactly the same coding method, yet only you can decode messages meant for you. Not the CIA, not the FBI, not Mossad, not the KGB, not ASIO, not anyone.

Security agencies are not happy about this. For many years the U.S. government tried to prevent the export and distribution of RSA cryptographic methods. However, there was a legal loophole that meant one could export a hardcopy of the source code but not an electronic copy. This was used to legally circumvent the ban. In any case, the mathematical result had already been presented at conferences, and if it had not been then someone else would soon have discovered the method. See also the Addendum at the end of this Chapter.

- In the section "Simple Coding and Decoding" we first discuss some elementary things about codes.
- In "Working with BIG numbers" we get a feeling for how "big" are the big numbers which we use.
- In "Background and Overview of RSA Cryptography" we discuss in outline the method of RSA cryptography.
- In "A Real Example of RSA encryption" we work through a real life example using Maple.
- In "Summary of the Method" we go back and summarise the method used in the previous Example.
- In "A Toy Example" we work through a very small example (which is totally insecure and useless) in order to understand better how and why the method works.
- In "Mathematical Theory of RSA Cryptography" we prove the Theorem which shows why it all works.

- In the "Addendum" I make a few remarks about matters such as history, competitions and quantum computing.

The last six Sections are largely independent and the same ideas are repeated a number of times. If you get stuck, just move on. Don't be discouraged if you find the mathematical explanation too confusing at this stage. It definitely is not examinable material!

Simple Coding and Decoding

[HM, 95,96]

Simple Coding Methods One of the simplest ways to encode a message is for the sender to replace every letter by another letter in an agreed manner. If the receiver of the message knows how this is done then it is usually easy to reverse the process and obtain the original message.

For example, the coding method might be to replace each letter by the one below it according to the following table:

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

A highly secret message such as "the key is under the mat by the front door" would be coded as"ofn znv hl tcqnb ofn ipo uv ofn mbaco qaab".

But someone who intercepts the message might figure that "the" is a very common word. So probably "ofn" means "the" and this means that "o, f, n" are decoded as "t, h, e" respectively.

To make it a little harder to crack it would be good to arrange the original message in blocks of 5 (say) letters, i.e. "theke yisun derth ematb ythef rontd oor" and then encode it as "ofnzv vhltc qnbof nipou vofnm bacoq aab"

If the receiver knows how messages are encoded then he/she can reverse the process and decode the message, obtaining "theke yisun derth ematb ythef rontd oor" and reading it as "the key is under the mat by the front door".

Frequency Analysis Any method of replacing one letter by another is not very intelligent. Certain letters such as "e" occur more frequently than others. By doing a frequency count of letters in a sufficiently long message, and a bit of trial and error, it is fairly easy to work out the original message.

Improved Coding Methods An alternative would be to have a single code *number* for each block of 5 letters.

To code and decode messages in this way would require a table of $26 \times 26 \times 26 \times 26 \times 26 = 26^5 = 11,881,376$ entries, consisting of blocks of 5 letters and matching numbers from 1 to 26^5. But a computer program could crack such a code with a built in dictionary and some trial and error without much trouble. One might want to use blocks of 80 letters, but this requires a table with 26^{80} numbers, and that is more than the number of atoms in the universe (currently estimated to be about 10^{80}).

Problems with these Coding Methods There are two major problems with the coding methods just discussed.

1. Very soon someone will lose their copy of the code book, or sell it to some unscrupulous third party, or have it stolen. If the coding method is sent over the internet it is likely to be intercepted.
2. If a malicious person knows how to encode a message they can reverse the process and so decode any messages they intercept.

But, and this is truly surprising and unexpected, the second problem *is* avoidable by using RSA Coding. That is, you can tell anyone how to encode messages they want to send you, but only you can decode these messages!

Moreover, with RSA cryptography, because there is no secret codebook that people need in order to send you a secret message, the first problem does not arise at all.

⋆*Working with BIG numbers*

Since we will work with incredibly big numbers, let's get a feel for them.

Examples of Big Numbers

- A *million* is 10^6.
- A *billion*[40] is one thousand million or 10^9.
- A *trillion*[41] is one thousand billion or 10^{12}.
- The world's wealthiest individual is worth about 90 billion dollars. This is more than the total financial worth of the 100 million people in the worlds poorest countries.
- There are 10^{21} atoms in a pinhead.
- There are about 10^{22} grains of sand on the earth and at least this many stars in the universe.
- There are about 10^{80} atoms in the universe.
- The number of ways I can arrange 60 books in a row on my bookshelf[42] is 60!, which is approximately 8.32×10^{81} and so more than the number of atoms in the universe.
- A *googol* [43] is defined to be 10^{100}. Notice that it has 101 digits.

  A googol equals $10^{20} \times 10^{80} = 100 \times 10^9 \times 10^9 \times 10^{80}$ and so is about 100 billion billion times the number of atoms in the universe.
- A *googolplex* is 10 raised to the power of a googol, or 1 followed by a googol of 0's. A googolplex equals $10^{\left(10^{100}\right)}$ and we write it as $10^{10^{100}}$.

  A googolplex has far more *digits* than atoms in the universe. This is a much stronger statement than merely saying it is larger than the number of atoms in the universe — you only need about 80 digits for that to be true.

---

[40] Unfortunately, we have to live with the American convention on this one. An older, and scientifically more logical, British definition is that a billion is a million million, or 10^{12}.

[41] The older definition of a trillion is a million million million, i.e. 10^{18}.

[42] Think of 60 slots in a row that are to be filled by the 60 books. There are 60 possibilities for the first slot and for each of these there are 59 possibilities for the second slot. Thus there are 60×59 possibilities for the first 2 slots and for each of these possibilities there are 58 possibilities for the third slot. Thus there are $60 \times 59 \times 58$ possibiliteis for the first 3 slots. Etc., etc. So there are 60! possibilities altogether.

✍ *By writing out the various possibilities in a systematic way, convince yourself there are 3! ways of arranging 3 books and 4! ways of arranging 4 books.*

[43] Said to have been coined in the 1940s by the 9 year old nephew of the American mathematician E. Kasner, at Kasner's request.

- A natural number with 2 digits is at least 10 and is less than 100, a natural number with 3 digits is at least 100 and is less than 1000, a natural number with 4 digits is at least 1000 and is less than 10,000, and a natural number with 150 digits is at least 10^{149} and is less than 10^{150}.

  More generally:

  $$\textit{A natural number } x \textit{ has } M \textit{ digits iff } 10^{M-1} \le x < 10^M. \qquad (2.28)$$

  It follows that there are $10^M - 10^{M-1} = 9 \times 10^{M-1}$ natural numbers with exactly M digits.
- If you multiply two 150 digit numbers you will get a number with 299 or 300 digits.

  The reason is that both numbers will be at least 10^{149} and both will be less than 10^{150}. So their product will be at least 10^{298} and will be less than 10^{300}.

  If the product is at least 10^{298} but less than 10^{299} it will have 299 digits.

  If the product is at least 10^{299} but less than 10^{300} it will have 300 digits.
- The previous idea holds in general, so we will call it a Theorem and write out the proof.

  Theorem 2.5.1. *If a natural number x has M digits and another natural number y has N digits then their product will have $M+N-1$ or $M+N$ digits.*

  Proof. We know $10^{M-1} \le x < 10^M$ and $10^{N-1} \le y < 10^N$.
  Multiplying, we get $10^{M+N-2} \le xy < 10^{M+N}$.
  We write this as

  $$10^{M+N-2} \le xy < 10^{M+N-1} \text{ or } 10^{M+N-1} \le xy < 10^{M+N}.$$

  From (2.28) it follows that in the first case xy has $M+N-1$ digits and that in the second case xy has $M+N$ digits. □

Big Numbers in Cryptography

- For cryptography we need to work with prime numbers of about 150 digits (much much more than a trillion trillion trillion trillion trillion times the number of atoms in the universe).
- To find primes of this size by some kind of random process will take a few seconds using Maple. See the calculation of p and q on page 58.

  Because there are so many such primes, the probability that we or anyone else will ever again find the same 2 primes is much much much less than the probability that 2 people would choose at random the same 2 atoms in the universe.
- We will also need the product of 2 such prime numbers, which as we saw before has about 300 digits, and is almost incomprehensibly large. Maple does the multiplication in a flash. See the calculation of n on page 59.
- If I give someone the product of 2 prime numbers each about 150 digits long, but not the primes themselves, it would take this person much much more than 100 years using all the world's current computers in parallel to find the 2 prime factors. *No one will ever find the 2 factors!*

- If I raise one 300 digit number, or even a number as small as 2, to a power given by another 300 digit number, the answer N will have far more digits than atoms in the universe. In fact, because $2^{10} = 1024$ and a 300 digit number is $\geq 10^{299}$,

$$N \geq 2^{(10^{299})} = 2^{10\times 10^{298}} = (2^{10})^{10^{298}} > 1000^{10^{298}} > 10^{10^{298}}.$$

  Notice that the answer is much larger than a googolplex, which is $10^{10^{100}}$. It is even far far larger than a googolplex multiplied by itself a googol times.[44]

  And yet if I ask Maple for the remainder after dividing the answer N by a third 300 digit number it will give me the answer in a fraction of a second! (It uses the method on page 37.) See the calculation of $D = C^d$ on page 62, where C is on page 62 and d is on page 60.

Summary

- Finding large primes is very quick (there are special tests for being a prime), i.e. takes seconds.
- Finding remainders after dividing one large number raised to another large number by a third large number is very quick, i.e. takes seconds.
- Factorising large numbers except in special cases takes an impossibly long time.

Background and Overview of RSA Cryptography

[HM, 97–99]

Representing Messages as Numbers We can easily represent a message as a (natural) number. Just replace letters by 1 through 26 using the following table:

$$A, B, C, D, E, F, G, H,\ I,\ J, K,\ L, M, N, O, P, Q,\ R,\ S,\ T, U, V, W, X, Y,\ Z$$
$$01,02,03,04,05,06,07,08,09,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26$$

Replace spaces by 27, commas by 28 and full stops by 29.

For example, the word "the" becomes the number "200805". The secret message "the key is under the mat by the front door" with 41 letters and spaces is translated into the number 20080527110525270919272114040518272008052713012027022527200805270618151420270415151 8. This is a big number, it has 84 digits.

This is not the coding method. It is nothing more than a simple way of representing messages as numbers. Conversely, given the number we can easily use the previous table to go back to the message. There are more efficient ways of translating messages into numbers, but that is not an essential feature and so we won't digress to discuss it.

In this way in future *we think of messages as numbers and numbers as messages.*

We split messages with 150 or more letters and spaces, which give numbers with 300 or more digits, into blocks, and treat each block separately.

---

[44] $10^{10^{298}} >> 10^{10^{200}} = 10^{(10^{100}\times 10^{100})} = \left(10^{10^{100}}\right)^{10^{100}}$, which is a googolplex multiplied by itself a googol times. By ">>" is meant "is far greater than".

To be on the super safe side any secret message with less than 100 letters (and so less than 200 digits in the corresponding number) should have its corresponding number padded out with garbage digits at the beginning to bring it up to 200 digits, before encoding.

To repeat: in future *a secret message is a secret number* with less than 300 digits.

Coding Secret Numbers Instead of coding a secret message we will talk about coding a secret number, which we have just seen is in effect the same thing.

The Very Basic Idea of RSA Cryptography Making some inessential simplifications, RSA cryptography works like this.

1. You (more precisely your computer) generate 2 very large prime numbers p and q, each 150 digits or more. You (or your computer) then multiply these 2 numbers to give a number n about 300 digits or more long.
2. You announce the product $n = pq$ to the world (or to any other computer that is interested) by posting it on your website, for example. See page 61. But you do *not* tell anyone else what p and q are.
3. There are 3 important computational points.
   a) *There are many such large primes* p and q, more than 10^{147} of them. This is over a trillion trillion trillion trillion trillion times more than all the atoms in the universe (there are about 10^{80} such atoms).
   b) It is *incredibly easy for your computer to find such primes* and to do it in such a way they will be different from primes ever found in the future or the past by you or any other computer. Just randomly churn out any 150 digit number and ask the computer to find the next prime after that. If the computer uses the package Maple (which you will yourself use) it will give you the next prime following this random number in a few seconds.[45]
   c) It is *essentially impossible to factorise a large number* n which is a product of two primes p and q each about 150 digits long unless you already know one of these primes.
4. Suppose someone has a secret number W (i.e. message) to send you. (It needs to be less than 300 digits, otherwise it must be split into smaller numbers.) They *encode* W in a certain way to give a coded number (i.e. coded message) C *which is made public.* The encoding method uses just the number n and a certain other publicly known number e called the $e$ncoding exponent, but does not involve knowing p or q.
5. The number C cannot be *decoded* to give back W, even in the lifetime of the universe,[46] by just reversing the encoding process. You need to also know p or q or some similar information.[47]
6. However, there *is* a fast way of decoding coded numbers if you know p or some similar information. Since only you know p, and no one will have

[45] Strictly speaking Maple uses a probabilistic method. But the chance of an error is much less than the probability that if two people each pick an atom somewhere in the universe at random then they both end up picking the same atom!

[46] OK, to be safe, let's say 100 years.

[47] More precisely, you need a secret number d calculated from p and q and which is called the $d$ecoding exponent. This will all be explained.

the resources to find p by factorising n, *only* you can decode messages that were coded up by others using n.

Any message sent to you by RSA encoding is totally safe.[48]

It is far from clear how this scheme works, and in fact almost unbelievable that it could work. But it does, and we will see it all in the following.

⋆A Real Example of RSA encryption

[HM, 99–101], but only small "toy" examples are discussed there. Here we give a "real" example.

At this point you might like to jump ahead to the Section "Summary of the Method" beginning on page 63 in order to gain an overview.

Generating the Public and Private Keys By randomly running your fingers over the keyboard enter a natural number *pp* with 150 digits. With the command `pp:=` you are telling Maple to set `pp` equal to this number *and* to store this number `pp` in its memory. Maple confirms the request and displays the value of *pp*.

```
> pp:= 836530018326471739718456981244510066629365454667349982776558
       965767174365745391729836879548878795879098374560987666398764
       099376398998764589702527386362;
```

$$pp := 836530018326471739718456981244510066629365454667349982776558965767174365745391729836879548878795879098374560987666398764099376398998764589702527386362$$

Do this again to store another randomly generated natural number *qq* with 150 digits in Maple's memory.

```
> qq:= 943987627897384756895479800200710098812730100111111110586
       162605018856757411563456705614547510406538564747747846501
       601040501476671045767176585018950214;
```

$$qq := 943987627897384756895479800200710098812730100111111110586162605018856757411563456705614547510406538564747747846501601040501476671045767176585018950214$$

Ask Maple to find the next *prime* after *pp* and store it in its memory as *p*.

```
> p:= nextprime(pp);
```

$$p := 836530018326471739718456981244510066629365454667349982776558965767174365745391729836879548878795879098374560987666398764099376398998764589702527386399$$

[48] Well, not quite! Maybe someone is secretly filming or logging your key strokes when you are creating your public and private keys, or before you encode a secret message.

Ask Maple to find the next prime after qq and store it in its memory as q.

```
> q:= nextprime(qq);
```

$q :=$ 943987627897384756895479800200710098812730100111111110586162
60501885675741156345670561454751040653856474774784650160104 0
501476671045767176585018950251

Ask Maple to multiply p and q and store the result in its memory as n.

```
> n:= p*q;
```

$n :=$ 789673987664961855991059957987616862212524804835927253312317
07000328412583200865081291547042010431512879185125984625531
87927258900664171743292196973884356836180150868084489142979
56501441003571997494034363955077785997244186969293793361609
41148782602255798417805912821812316634398078633517672163503
6149

Ask Maple to multiply $p - 1$ and $q - 1$ and store the result in its memory as m.

```
> m:= (p-1)*(q-1);
```

$m :=$ 789673987664961855991059957987616862212524804835927253312317
07000328412583200865081291547042010431512879185125984625531
87927258900664171743292196973866551659717912303118349775165
11281275561476442715573270592356215211213063812338606819115
31509862360489486186922496021831856549091074180341043408869
9500

Ask Maple to find the remainder after dividing m by 3. Since the remainder is 0, you see that 3 is a factor. Similarly 5 is a factor. But for 7 the remainder is 6 and so 7 is not a factor. In this way you quickly find a prime which is *not* a factor of m.

This will be the "encoding exponent" e which you set equal to 7 in Maple's memory.

```
> irem(m,3);
                    0
> irem(m,5);
                    0
> irem(m,7);
                    6
> e:= 7;
                  e := 7
```

In practice you will quickly find e in a few steps.[49]

Because e is a prime less than m and e does not divide m, it follows $\gcd(e, m) = 1$. This means by Theorem 2.4.1 that there is a natural number d less than m such that $ed = 1 \bmod m$. Ask Maple to find this d.

```
> d:= 1/e mod m;
```

$d :=$ 1128105696664231222844371368553738374589321149765610361874738671433263036902869501161307814885863307326845501799780364741132465557237738820470313853409507379959701757588335682166444687536516394918165104672274794593158875805446198009741593307283748006992659813178514597597950701301059719157762983856 43

None of the previous calculations take Maple more than a second or so.

The Information You Put on Your Website In this way you now have your public key (n, e) and your private key d. On your website you should publish your public key and the other information in Figure 2.2 on page 61.

The reason for the padding in instruction 1 on your home page is that if the secret number is too small then there are other ways of decoding it, although they will still take a very long time.

The "99" and "00" are so that when you decode the secret number you will know if there is junk padding and where the junk ends.

The Information You Keep Secret Keep the private key d to yourself. No one other than you will know what d is, and they do not need to know d in order to send you coded messages.

---

[49] One way to see this is as follows. There is one chance in 3 that 3 *will* divide m, since there are exactly 3 possible remainders after dividing m by 3 and only a remainder of 0 means that 3 divides m. There is similarly one chance in 5 that 5 will divide m and so there is only one chance in 15 that *both* 3 and 5 will divide m. There is one chance in 7 that 7 will divide m, and so only one chance in $3 \times 5 \times 7 = 105$ that 3, 5 and 7 will *all* divide m, etc. In other words there is 104 chances in 105 that at least one of 3, 5 or 7 will *not* divide m.

As we test more and more primes the chance is miniscule that all of them will divide m. Remember that we only require one prime that does *not* divide m.

Here is another way to see that you will eventually find a prime which does *not* divide m. If for example $3, 5, 7, 11, 13$ all divide m then m must be a multiple of $3 \times 5 \times 7 \times 11 \times 13 =$ 152,460 and so m must certainly be larger than 152,460. In the same way if all of the first 130 primes divide m then m must be larger than the product of these 130 primes. However, this product contains 304 digits and so is larger than m.

So one of the first 130 primes *must* divide m. By the way, the 130th prime is 733.

Remark: The Maple code to use is

```
> prod:= 1:
        for i from 1 to 130 do
    prod:= prod * ithprime(i)
                        end do:
                          prod;
```

My public key is:

n :789673987664961855991059957987616862212524804835927253312317
07000328412583200865081291547042010431512879185125984625531
87927258900664171743292196973884356836180150868084489142979
56501441003571997494034363955077785997244186969293793361609
41148782602255798417805912821812316634398078633517672163503
6149

e :7

If you have a secret message for me, do the following:

1. In your message replace A by 01, B by 02, ..., Z by 26, blank space by 27, comma by 28 and period by 29.
   Let W be the number you get in this manner.
   If W is 300 or more digits long (i.e. corresponds to 150 or more characters) first break it into blocks of less than 300 digits and treat each block separately.
   To be super cautious, if the number for any block is less than 200 digits (i.e. if the block corresponds to less than 100 characters), first pad it out with junk digits at the beginning to make it 200 or more digits long. The junk should begin with 99 and end with 00, and contain no other 99 or 00.
2. Compute the remainder after dividing W^e by n and call it C (for coded message). (This is a quick computation for the computer provided it does it the right way.)
3. Send me C or put it up on your own website and let me know there is a coded secret message for me. No one other than me will be able to decode it.

Figure 2.2: Your Webpage

At this stage it would also be a very smart idea to completely destroy pp, p, qq, q and m. They are no longer needed. Put them in the computer trash and secure erase your trash!

Coding a Message Only You Can Decode Following instructions your friend translates his/her secret message "the key is under the mat by the front door" into the following secret number. We did this on page 56 and obtained

20080527110525270919272114040518272008052713012027022527200
8052706181514202704151518

Since this is 84 digits long your friend pads it out with two lines of junk digits at the beginning, as instructed on your homepage, and enters this number into

his/her computer:

```
>W:= 998218676572478561365412136354137654965482743651724766354656
     827366545476623654826235487765626456757639972725417847002008
     052711052527091927211404051827200805271301202702252720080527
     061815142027041515 18;
```

W :=998218676572478561365412136354137654965482743651724766354656
827366545476623654826235487765626456757639972725417847002008
052711052527091927211404051827200805271301202702252720080527
061815142027041515 18

Your friend next uses his/her computer to encode W into a public number C by again following the instructions on your homepage and computing the remainder after dividing W^e by n.

It would be impossible to do this before the universe ends if the computer tried to first calculate W^e. Instead it has to proceed like we did on page 37 when we computed 2136^{1035} and 507^{107} mod 14. This is the reason for the "`&`" before "`^e`" in the instructions to Maple.

```
> C:= W&^e mod n;
```

C :=412807365828314445505340698766576568603950314867328252560244834087708501914239407341441267427025327400872387162386718427867243400044687355743392075312560215812710738869957475814897584219029802174103799320014864255767115240878601597547120258297781625130345722715692192237689056771898356139943705573335

Your friend now sends you the number C or just publishes C on his/her website. No one apart from you will ever be able to decode it! Of course your friend does not publish W since anyone who reads your instructions will know the original message if they know W.

How You Decode the Coded Message When you receive the coded message C you ask your computer to decode C by computing D as follows. Note that this requires your secret decoding exponent d as well as the public number n.

```
> D:= C&^d mod n;
```

D :=998218676572478561365412136354137654965482743651724766354656
827366545476623654826235487765626456757639972725417847002008
052711052527091927211404051827200805271301202702252720080527
061815142027041515 18

Note that D is the same as the original secret number W!!

Finally you strip off the junk padding which begins with 99 and ends with the first 00. This gives the number

200805271105252709192721140405182720080527130120270225272008052706181514202704151518.

Translate this back into an English sentence by replacing 20 by T, 08 by H, 05 by E, etc. This gives me the message "the key is under the mat by the front door".

Summary of the Method

[HM, 101–103]

Generating the Public and Private Keys

1. Choose two different prime numbers p and q (in practice each about 150 digits long in a random way that no one else could ever emulate).
2. Multiply p and q together and call the result n, i.e. $n = pq$ (in practice n will have about 300 digits).
3. Multiply $p-1$ and $q-1$ together and call the result m, i.e. $m = (p-1)(q-1)$ (in practice m will also have about 300 digits).
4. Choose a natural number $e < m$ ("e" is for encoding) which has no common factors with m (i.e. is relatively prime to m).[50]
5. Find the unique natural number $d < m$ ("d" is for decoding) such that $ed \equiv 1 \bmod m$. This is justified by Theorem 2.4.1. The Extended Euclidean Algorithm is used (in practice used by Maple) to find d.
6. Publish your public key numbers n and e on your website.
   Your private key is the number d. This is to be kept secret.

Coding a Message Only You Can Decode Suppose that someone has a secret number W that they want you to know.

They simply compute the remainder C (for coded message) after dividing W^e by n. They then publish C on their website for anyone, including you, to see. Of course, they do *not* put W on their website.

How You Decode the Coded Message Simply compute the remainder after dividing C^d by n. This will equal W.

We prove this in Theorem 2.5.2

A Toy Example

[HM, 101–103]

Now we will go through the whole process again, following the description under the previous "Summary of the Method".

[50] This is easy. Start with 3 and see if it divides m. If 3 does *not* divide m let $e = 3$. Otherwise try the next prime 5, and then 7, etc. In practice you will quickly find a prime that does not divide m and so has no common factor with m. See Footnote 49.

Notice by the way that the number 2 divides m. The reason is that p and q are primes larger than 2 so they must be odd. This means $p-1$ and $q-1$ must be even. But this means that their product m is also even.

Note: on page 59 we looked for an e which was a prime, but in fact all we really need is that e is relatively prime to m.

Generating the Public and Private Keys

1. Choose the primes $p = 3$ and $q = 11$. This is completely stupid for security purposes. But it will allow us to get a good idea of why RSA encryption works.

   ✍ *In order to increase your understanding you should simultaneously work the example $p = 5$ and $q = 7$.*
2. Multiply p and q together and call the result n, i.e. $n = pq = 3 \times 11 = 33$.

   ✍ *What is n in your example?*
3. Multiply $p - 1$ and $q - 1$ together and call the result m, i.e. $m = (p - 1)(q - 1) = 2 \times 10 = 20$.

   ✍ *What is m in your example?*
4. Choose a natural number e ("e" is for **e**ncoding) which has no common factors with $m = 20$. So $e = 1$, 3, 7, 9, 11, 13, 17, or 19. *Don't* use $e = 1$. We will choose $e = 13$

   ✍ *What are the possible values for e in your example?*
5. Find the natural number d such that $ed \equiv 1 \bmod m$, i.e. $ed \equiv 1 \bmod 20$. The answer is $d = 17$.

   ✍ *Why does this follow from Table 2.1? What are the values of d for the other possible values of e?* One allowable value of e in your example is 5. *What is the corresponding value of d?*

$\otimes$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	0	2	4	6	8	10	12	14	16	18	0	2	4	6	8	10	12	14	16	18
3	0	3	6	9	12	15	18	1	4	7	10	13	16	19	2	5	8	11	14	17
4	0	4	8	12	16	0	4	8	12	16	0	4	8	12	16	0	4	8	12	16
5	0	5	10	15	0	5	10	15	0	5	10	15	0	5	10	15	0	5	10	15
6	0	6	12	18	4	10	16	2	8	14	0	6	12	18	4	10	16	2	8	14
7	0	7	14	1	8	15	2	9	16	3	10	17	4	11	18	5	12	19	6	13
8	0	8	16	4	12	0	8	16	4	12	0	8	16	4	12	0	8	16	4	12
9	0	9	18	7	16	5	14	3	12	1	10	19	8	17	6	15	4	13	2	11
10	0	10	0	10	0	10	0	10	0	10	0	10	0	10	0	10	0	10	0	10
11	0	11	2	13	4	15	6	17	8	19	10	1	12	3	14	5	16	7	18	9
12	0	12	4	16	8	0	12	4	16	8	0	12	4	16	8	0	12	4	16	8
13	0	13	6	19	12	5	18	11	4	17	10	3	16	9	2	15	8	1	14	7
14	0	14	8	2	16	10	4	18	12	6	0	14	8	2	16	10	4	18	12	6
15	0	15	10	5	0	15	10	5	0	15	10	5	0	15	10	5	0	15	10	5
16	0	16	12	8	4	0	16	12	8	4	0	16	12	8	4	0	16	12	8	4
17	0	17	14	11	8	5	2	19	16	13	10	7	4	1	18	15	12	9	6	3
18	0	18	16	14	12	10	8	6	4	2	0	18	16	14	12	10	8	6	4	2
19	0	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Table 2.1: Multiplication mod 20

Coding a Message Only You Can Decode The secret number is a natural number W less than $n = 33$. For example suppose $W = 15$.

This is coded up by the natural number C less than n such that $C \equiv W^e \bmod n$, i.e. C is the remainder after dividing W^e by n.

In this case C is the natural number less than 33 such that $C \equiv 15^{13} \bmod 33$, i.e. C is the remainder after dividing 15^{13} by 33. It follows from Table 2.2 that $C = 9$.

Find C directly by the methods on page 37 without using Table 2.2.
Find C from Table 2.2 if $W = 21$ and if $W = 9$.
Find C in your example if $W = 15$.

a	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
a^1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
a^2	0	1	4	9	16	25	3	16	31	15	1	22	12	4	31	27	25	25	27	31	4	12	22	1	15	31	16	3	25	16	9	4	1
$\boldsymbol{a^3}$	0	1	**8**	**27**	**31**	**26**	**18**	**13**	**17**	**3**	**10**	**11**	**12**	**19**	**5**	**9**	**4**	**29**	**24**	**28**	**14**	**21**	**22**	**23**	**30**	**16**	**20**	**15**	**7**	**2**	**6**	**25**	**32**
a^4	0	1	16	15	25	31	9	25	4	27	1	22	12	16	4	3	31	31	3	4	16	12	22	1	27	4	25	9	31	25	15	16	1
a^5	0	1	32	12	1	23	21	10	32	12	10	11	12	10	23	12	1	32	21	10	23	21	22	23	21	1	23	12	10	32	21	1	32
a^6	0	1	31	3	4	16	27	4	25	9	1	22	12	31	25	15	16	16	15	25	31	12	22	1	9	25	4	27	16	4	3	31	1
$\boldsymbol{a^7}$	0	1	**29**	**9**	**16**	**14**	**30**	**28**	**2**	**15**	**10**	**11**	**12**	**7**	**20**	**27**	**25**	**8**	**6**	**13**	**26**	**21**	**22**	**23**	**18**	**31**	**5**	**3**	**19**	**17**	**24**	**4**	**32**
a^8	0	1	25	27	31	4	15	31	16	3	1	22	12	25	16	9	4	4	9	16	25	12	22	1	3	16	31	15	4	31	27	25	1
$\boldsymbol{a^9}$	0	1	**17**	**15**	**25**	**20**	**24**	**19**	**29**	**27**	**10**	**11**	**12**	**28**	**26**	**3**	**31**	**2**	**30**	**7**	**5**	**21**	**22**	**23**	**6**	**4**	**14**	**9**	**13**	**8**	**18**	**16**	**32**
a^{10}	0	1	1	12	1	1	12	1	1	12	1	22	12	1	1	12	1	1	12	1	1	12	22	1	12	1	1	12	1	1	12	1	1
$\boldsymbol{a^{11}}$	0	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
a^{12}	0	1	4	9	16	25	3	16	31	15	1	22	12	4	31	27	25	25	27	31	4	12	22	1	15	31	16	3	25	16	9	4	1
$\boldsymbol{a^{13}}$	0	1	**8**	**27**	**31**	**26**	**18**	**13**	**17**	**3**	**10**	**11**	**12**	**19**	**5**	**9**	**4**	**29**	**24**	**28**	**14**	**21**	**22**	**23**	**30**	**16**	**20**	**15**	**7**	**2**	**6**	**25**	**32**
a^{14}	0	1	16	15	25	31	9	25	4	27	1	22	12	16	4	3	31	31	3	4	16	12	22	1	27	4	25	9	31	25	15	16	1
a^{15}	0	1	32	12	1	23	21	10	32	12	10	11	12	10	23	12	1	32	21	10	23	21	22	23	21	1	23	12	10	32	21	1	32
a^{16}	0	1	31	3	4	16	27	4	25	9	1	22	12	31	25	15	16	16	15	25	31	12	22	1	9	25	4	27	16	4	3	31	1
$\boldsymbol{a^{17}}$	0	1	**29**	**9**	**16**	**14**	**30**	**28**	**2**	**15**	**10**	**11**	**12**	**7**	**20**	**27**	**25**	**8**	**6**	**13**	**26**	**21**	**22**	**23**	**18**	**31**	**5**	**3**	**19**	**17**	**24**	**4**	**32**
a^{18}	0	1	25	27	31	4	15	31	16	3	1	22	12	25	16	9	4	4	9	16	25	12	22	1	3	16	31	15	4	31	27	25	1
$\boldsymbol{a^{19}}$	0	1	**17**	**15**	**25**	**20**	**24**	**19**	**29**	**27**	**10**	**11**	**12**	**28**	**26**	**3**	**31**	**2**	**30**	**7**	**5**	**21**	**22**	**23**	**6**	**4**	**14**	**9**	**13**	**8**	**18**	**16**	**32**
a^{20}	0	1	1	12	1	1	12	1	1	12	1	22	12	1	1	12	1	1	12	1	1	12	22	1	12	1	1	12	1	1	12	1	1
a^{21}	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
a^{22}	0	1	4	9	16	25	3	16	31	15	1	22	12	4	31	27	25	25	27	31	4	12	22	1	15	31	16	3	25	16	9	4	1
a^{23}	0	1	8	27	31	26	18	13	17	3	10	11	12	19	5	9	4	29	24	28	14	21	22	23	30	16	20	15	7	2	6	25	32
a^{24}	0	1	16	15	25	31	9	25	4	27	1	22	12	16	4	3	31	31	3	4	16	12	22	1	27	4	25	9	31	25	15	16	1
a^{25}	0	1	32	12	1	23	21	10	32	12	10	11	12	10	23	12	1	32	21	10	23	21	22	23	21	1	23	12	10	32	21	1	32
a^{26}	0	1	31	3	4	16	27	4	25	9	1	22	12	31	25	15	16	16	15	25	31	12	22	1	9	25	4	27	16	4	3	31	1
a^{27}	0	1	29	9	16	14	30	28	2	15	10	11	12	7	20	27	25	8	6	13	26	21	22	23	18	31	5	3	19	17	24	4	32
a^{28}	0	1	25	27	31	4	15	31	16	3	1	22	12	25	16	9	4	4	9	16	25	12	22	1	3	16	31	15	4	31	27	25	1
a^{29}	0	1	17	15	25	20	24	19	29	27	10	11	12	28	26	3	31	2	30	7	5	21	22	23	6	4	14	9	13	8	18	16	32
a^{30}	0	1	1	12	1	1	12	1	1	12	1	22	12	1	1	12	1	1	12	1	1	12	22	1	12	1	1	12	1	1	12	1	1
a^{31}	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
a^{32}	0	1	4	9	16	25	3	16	31	15	1	22	12	4	31	27	25	25	27	31	4	12	22	1	15	31	16	3	25	16	9	4	1

Table 2.2: Power Table Mod 33

How You Decode the Coded Message Find the natural number D less than n such that $D \equiv C^d \bmod n$.

In this case D is the natural number less than 33 such that $D \equiv 9^{17} \bmod 33$. It follows from Table 2.2 that $D = 15$. This is the same as W, as it should be.

✍

Check that the two coded messages C which you obtained from $W = 21$ and $W = 9$ decode to $D = 21$ and $D = 9$ respectively, i.e. back to the original W's.

Find D in your example when $W = 15$ and confirm that the answer is indeed 15.

Card Shuffling In [HM, 97,98] there is a discussion of card shuffling as an analogy for coding messages.

If you look at the rows in black in Table 2.2 you will see that they correspond to powers 3, 7, 9, 13, 17, 19. In each of these rows the numbers $2, 3, 4, \ldots, 32$ are shuffled around and each appears exactly once. These also are the rows corresponding to the powers e we found and for which $\gcd(e, m) = 1$, i.e. $\gcd(e, 20) = 1$.

In each of the other rows some numbers from $2, 3, 4, \ldots, 32$ do not appear at all while other numbers appear more than once.

The idea is that raising secret numbers $W = 2, 3, 4, 5, \ldots, 32$ to the power e and taking the remainder mod 33 to give encoded numbers C, corresponds to shuffling the pack of numbers $2, 3, 4, \ldots, 32$. (We will prove this in general in Question 2.) Raising numbers C to the power d mod 33 to give decoded numbers D corresponds to unshuffling the pack back to its original order.

In general your public key (n, e) allows anyone to shuffle the pack of numbers $2, 3, 4, \ldots, n-1$. But with numbers n of about 300 digits, only you with the private key d will ever be able to unshuffle the pack.

It is completely unclear at this stage why it works, but we prove it does work in Theorem 2.5.2. *We will show that if W is a natural number less than n, if C is the remainder after dividing W^e by n, and D is the remainder after dividing C^d by n, then $D = W$.*

The other way to get W back from C would be to raise each of the n numbers $2, 3, 4, \ldots, n-1$ to the power e until we found the number which gives C. This then is W.

But in real life n will have about 300 digits and this means we will have to check incredibly more cases than there are atoms in the universe. No one can do this. There may just possibly be a quick way to do it without knowing d, but no one knows how and the experts in cryptography and number theory believe it is not possible. However, no one has actually proved it is not possible!!

⋆*Mathematical Theory of RSA Cryptography*

In [HM, 103–106] a numerical example is examined. Here we give the complete theory.

We saw on page 63 that p and q are different prime numbers. We then let $n = pq$ and $m = (p-1)(q-1)$.

The next step was to find a natural number $e < m$ having no common factor with m. This is the same as requiring $\gcd(e, m) = 1$. It follows from Theorem 2.4.1 there is a unique natural number $d < m$ such that $ed \equiv 1 \bmod m$.

If someone has a "secret" (natural) number $W < n$ they want to send you they obtain the coded number C from W by finding the remainder after dividing W^e by n. So $C \equiv W^e \bmod n$.

From the coded number C *you* find the decoded number D by finding the remainder after dividing C^d by n. So $D \equiv C^d \bmod n$.

The point to all this is that $D = W$. This is what we will now prove.

Theorem 2.5.2 (The Main RSA Theorem)**.** *Suppose p and q are distinct prime numbers. Let $n = pq$ and let $m = (p-1)(q-1)$.*

Suppose e and d satisfy $ed \equiv 1 \bmod m$.

Suppose $W < n$ is a natural number, C is the remainder after dividing W^e by n, and D is the remainder after dividing C^d by n.

Then $D = W$.

Understanding this will req
aminable material.

Proof. Because of the way we defined the numbers C and D we have

$$C \equiv W^e \bmod n, \quad D \equiv C^d \bmod n, \quad 1 \leq W, C, D < n.$$

Working mod n this means

$$D \equiv C^d \equiv (W^e)^d = W^{ed} \bmod n. \tag{2.29}$$

Suppose we can show

$$W^{ed} \equiv W \bmod n. \tag{2.30}$$

Then it will follow from (2.29) and (2.30) that $D \equiv W \bmod n$. This implies $D = W$ since both D and W are natural numbers less than n.

So now our goal is to prove (2.30). This means we want to show that $n \mid W^{ed} - W$. But because $n = pq$ is a product of 2 different prime numbers this is equivalent[51] to proving that $p \mid W^{ed} - W$ and $q \mid W^{ed} - W$. So now we have the new and equivalent goal of proving that

$$W^{ed} \equiv W \bmod p \quad \text{and} \quad W^{ed} \equiv W \bmod q. \tag{2.31}$$

If we can do this we will have finished the proof of the Theorem!

First notice that since

$$ed \equiv 1 \bmod m,$$

the remainder after dividing ed by m is 1, and so for some integer k,

$$ed = km + 1. \tag{2.32}$$

If $p \mid W$ then the first equivalence in (2.31) is true since both W and W^{ed} are equivalent to 0 mod p.

If $p \nmid W$ then

$$\begin{aligned} W^{ed} &= W^{km+1} \quad \text{(from (2.32))} \\ &= WW^{km} = WW^{k(p-1)(q-1)} = W(W^{p-1})^{k(q-1)} \\ &\equiv W1^{k(q-1)} \bmod p \quad \text{(using Fermat's Little Theorem, Theorem 2.4.2)} \\ &= W \bmod p. \end{aligned}$$

[51] This is where we use in the proof the fact p and q are different. If $p = q$ then $n = p^2$, but it does not follow that $p^2 \mid W^{ed} - W$ if $p \mid W^{ed} - W$.

In any case we would not allow $p = q$. Because if this were true then we could find p from n by taking the square root of n, which is a trivial thing for Maple to do.

This proves the first equivalence in (2.31).
(Prove the second equivalence in the same way.)
This means we have proved (2.31) and so we have proved the Theorem. □

Addendum

The True History of RSA The official discoverers of RSA cryptography are three mathematicians Ron Rivest, Adi Shamir and Len Adleman, who published the method in 1977. A couple of years earlier, W. Diffie and M. Hellman first discovered the general concept but did not have a secure way to use it.

However, another mathematician Clifford Cocks, working for the British security agency, actually discovered the full method in 1973. For bureaucratic security reasons it was classified and not published, even though the British thought it did not have any use!

Factoring Competitions and Prizes The RSA company has a list of numbers of varying sizes. It pays out money to anyone who can factor these numbers, the larger the number the larger the prize. This competition is used to decide how big the primes p and q in RSA encryption should be for security reasons. See `http://www.rsasecurity.com/rsalabs/node.asp?id=2094`

Quantum Computing and Factorisation It was shown in 1995 by a mathematician Peter Shor that a computer using the principles of quantum mechanics could factor large numbers so quickly that RSA cryptography would no longer be viable. However, although there is an enormous amount of research and investement in this field, no one has yet been able to build a quantum computer.

Questions

1
- Suppose you know that the sum of two numbers p and q is a and you also know that their product is b. What is a quadratic equation, with coefficients expressed *just* in terms of a and b, whose roots are p and q ?

  Use this to find a formula for p and q in terms of a and b. Check your answer.
- Suppose someone rummages through your trash and discovers the number m (which equals $(p-1)(q-1)$) that you thought you had destroyed and which you used in finding your private decoding number d.
  - Do they have enough information to find your private decoding number d?
  - Use the first part of this question to show how they can use the value of m together with your public number n to find p and q.

2 On page 66 we said that the numbers $2^e, 3^e, 4^e, 5^e, \ldots, 32^e$ after taking the remainder mod 33 are a "shuffle" of the numbers $2, 3, 4, 5, \ldots, 32$.

More general the following is true:

Theorem 2.5.3. *Suppose p and q are distinct prime numbers. Let $n = pq$ and let $m = (p-1)(q-1)$. Suppose e and m are relatively prime.*

Then the numbers $2^e, 3^e, 4^e, 5^e, \ldots, (n-1)^e$ after taking the remainder mod n are a permutation (i.e. "shuffle") of the numbers $2, 3, 4, 5, \ldots, n-1$.

In order to prove this Theorem it is sufficient to show that if W_1 and W_2 are two natural numbers less than n then

$$W_1^e \text{ and } W_2^e \text{ give the same remainder mod } n \implies W_1 = W_2. \quad (2.33)$$

1. Explain why the Theorem follows from (2.33).
2. Use the RSA Theorem to show (2.33). (It is just a few lines.)

2.6 Irrational Numbers

Following assumptions to their logical conclusions can yield powerful results.

Overview

The assumption that all numbers are rational, i.e. can be written as fractions m/n where m and n are integers and $n \neq 0$, at first seems to be a reasonable one. But as we will see it leads to an impossibility. From this it follows some numbers are in fact not rational — they are "irrational".

A number of different methods will be given that can be used to show certain numbers are irrational.

Finally, examples of numbers that everyone believes to be irrational, but no one yet can prove this to be the case, will be given.

Rational and Irrational Numbers

In Section 2.1 we briefly discussed the natural numbers, the integers and the real numbers. In the following we divide the real numbers into the *rational numbers* (those numbers which can be written as fractions, which includes the integers) and the irrational numbers (those numbers which cannot be written as fractions).

Definition. A real number is *rational* if it can be written in the form m/n where m and n are integers and $n \neq 0$. If a real number cannot be written this way it is called *irrational*.

It is often convenient to assume that m and n have been cancelled down, so that they have no common factors. For example,

$$\frac{450}{165} = \frac{90}{33} = \frac{20}{11}.$$

However, we do not necessarily assume this is the case.

It is easy to construct rational numbers. They are just the fractions. But are there any numbers which are not rational, i.e. which are irrational? We will soon see that in fact many numbers are not rational.

There are Lots of Rational Numbers

Between any two rational numbers, no matter how close, there is always another rational number. All we have to do is take the average of the two numbers.

In other words, suppose a and b are different rational numbers, and let's name them so that $a < b$. The number $\dfrac{a+b}{2}$ is also rational, *why?* And we

have that

$$a < \frac{a+b}{2} < b.$$

We can repeat the argument and in this way get another rational number midway between a and $\frac{a+b}{2}$, namely $\frac{1}{2}\left(a + \frac{a+b}{2}\right) = \frac{3a}{4} + \frac{b}{4}$. Next we get $\frac{7a}{8} + \frac{b}{8}$. In fact we can get an infinite decreasing sequence of rational numbers

$$a < \cdots < \frac{15a}{16} + \frac{b}{16} < \frac{7a}{8} + \frac{b}{8} < \frac{3a}{4} + \frac{b}{4} < \frac{a}{2} + \frac{b}{2}.$$

which is eventually as close as we wish to a.

In a similar way we can also get an infinite *increasing* sequence of rational numbers which is eventually as close as we wish to a.

This was one of the reasons that the ancient Greeks initially thought that all numbers were rational.

The Ancient Greeks

[HM, 110–112]

Greece in the period 600–300 BC was home to a remarkable flowering of human endeavour and was the foundational culture of Western Civilization. Philosophy, Literature, Mathematics, Science, Music, Theatre, Architecture, Sculpture, Pottery, Political Science and Democracy, were developed to an extent which, at least in the West, was in most cases not surpassed until the Renaissance beginning around 1400 AD.

The Greeks initially thought of natural numbers and positive fractions as the only types of positive numbers.[52] This was not unreasonable for the following reasons:

- Certainly the natural numbers are "natural". We use them to count.
- Fractions are natural as they correspond to ways of dividing quantities into smaller equal parts. For example, 1/3 corresponds to the result of dividing a given line segment into 3 segments of equal length. Likewise 2/5 corresponds to dividing the given line segment into 5 segments of equal length and then placing 2 of the segments end to end.
- The Greeks thought of any two quantities as being commensurable one with the other, i.e. "capable of being measured by a common standard", i.e. integer multiples of some small common quantity.

  Suppose we take a line segment of length L and a reference ruler which is one unit long. (The Greeks used the "pous" unit which is about 316mm.) Then the Greeks thought that if they divided their unit ruler into a sufficiently large number n of small parts of equal length $1/n$ pous, then the length L of the original line segment should be equal to some integer multiple m of $1/n$. This idea was reinforced by the fact that in principle one could choose $1/n$ as small as one wished.

But the Greeks then discovered a major problem with their world view.

They knew that if you draw a right angled triangle for which the two shorter sides are of length 1, then the length L of the hypotenuse is given by $L =$

[52] The Greeks did not have a useful concept of negative number, however.

$\sqrt{2}$. (After all, they knew Pythagoras' Theorem, since Pythagoras lived in Greece approximately 580–500 BC.) More precisely they thought of this fact geometrically. Namely, in Fig. 2.3 where we construct a square on each of the three sides of the right angled triangle, the area L^2 of the square on the hypotenuse equals the sum of the areas on the other two sides, i.e. $L^2 = 1+1 = 2$, or as we would write it, $\sqrt{L} = 2$.

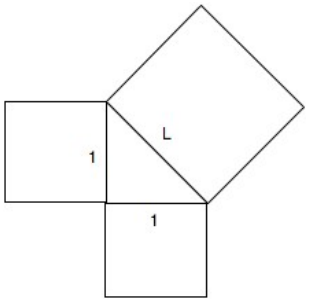


Figure 2.3: Larger square has area $L^2 = 1^2 + 1^2 = 2$

From their assumption that $L = m/n$ for positive integers m and n, the Greeks were led to an impossible conclusion, as we will soon see.

Examples of Irrational Numbers

[HM, 123–124]

The Irrationality of $\sqrt{2}$. As we discussed previously, the ancient Greeks initially thought that all numbers are rational. But we will now prove[53] that $\sqrt{2}$ is in fact irrational. Remember that $\sqrt{2}$ has a very simple geometric interpretation — it is the length of the hypotenuse of the right angled triangle whose two smaller sides are both one unit long.

The proof uses the *Method of Contradiction.* That is, we will *assume* that $\sqrt{2}$ is rational and from this assumption we will obtain a conclusion which is clearly false. It follows that the assumption was incorrect, in other words $\sqrt{2}$ is irrational.

Theorem 2.6.1. *$\sqrt{2}$ is irrational.*

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

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

where m and $n \neq 0$ 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.*

[53] The proof is essentially that used in Euclid's Elements, written about 290 BC. See `http://aleph0.clarku.edu/ djoyce/java/elements/toc.html`. This website gives a complete online version of Euclid's elements. First click on "A quick trip through the Elements". Then go to Proposition 8 of Book 8, and then read the "Guide" at the bottom of the webpage to see what this has got to do with $\sqrt{2}$! (Not that it is all that clear, to say the least.)

Using these *new* m and n which *do not have any common factor*, and squaring both sides, we get

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

and so

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

It follows that m^2 is even since 2 must be a factor. But this means 2 divides $m \times m$ and so 2 must divide m alone (see Theorem 2.3.10 on page 33). In other words, m is also even

Because m is even, we can write

$$m = 2k$$

for some integer k, and hence

$$m^2 = 4k^2.$$

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

$$4k^2 = 2n^2,$$

and hence

$$2k^2 = n^2.$$

But now we can argue as we did before for m, and deduce that n^2 is even and hence n is even.

Thus (the new) m and n both have the common factor 2, which contradicts the fact they have *no* common factors.

This contradiction implies that our original assumption in the first two lines was wrong, and so $\sqrt{2}$ is not rational. In other words, $\sqrt{2}$ is irrational □

The Irrationality of $\sqrt{3}$

Theorem 2.6.2. *$\sqrt{3}$ is irrational.*

The proof is very similar to that for $\sqrt{2}$. *Write out the proof for yourself* by making the appropriate changes. The main point is just to replace 2 by 3 at various places. Instead of certain numbers being even, i.e. divisible by 2, they now will be divisible by 3.

You can check your proof against what is written in [HM, p 115].

More Irrational Numbers

- In [HM, p119, Q11] you are asked to use similar arguments to show that, for example, $\sqrt{6}$ is irrational. *Try this yourself.*
- In fact, the square root of any integer which is not itself a perfect square is irrational. The proof is similar, but messier to write out neatly.
- In [HM, p116] it is shown that $\sqrt{2} + \sqrt{3}$ is irrational. *Look at the proof.*

- If $2^x = 32$ then $x = 5$.

  Is there a rational number m/n such that $2^{m/n} = 33$?

  If there were, then raising both sides to the power n we would get $2^m = 33^n$.

  But this means that 2 is a factor of 33^n and so 2 is a factor of 33. But this is not true and so we have a contradiction.

  It follows that if $2^x = 33$ then x is irrational. Later in your other maths courses you will write this x as $\log_2 33$ and call it the log of 33 to the base 2. It is not surprising that x is a little bigger than 5, and your calculator will show $x = 5.044394118\ldots$.

  Similarly, if $10^x = 33$ then we write $x = \log_{10} 33$, or sometimes just $x = \log_{10} 33$

- Here is a trickier question. *Show that if* $2^x = 30$ *then* x *is irrational.*
- The number π and the number e (see page 7) are both irrational, but this is more difficult to prove.
- Mathematicians believe that numbers like 2^π and π^e are irrational, but no one so far has been able to prove this.

Questions

There are lots of Questions in [HM, pp118–120].

2.7 The Real Number System

Sometimes things that seem commonplace and ordinary are actually exotic, and things that at first seem rare and exceptional are in fact the norm.

Overview

The decimal expansion of a number x can be interpreted as an infinite series. It can also be thought of as giving the "address" of x on the real number line.

Some numbers have more than one infinite decimal expansion. These numbers are precisely those which also have a finite decimal expansion. One example is $1 = 1.\overline{0} = .\overline{9}$.

Instead of decimal expansions, one can consider binary expansions. These correspond to continually subdividing intervals into 2, rather than 10, smaller parts.

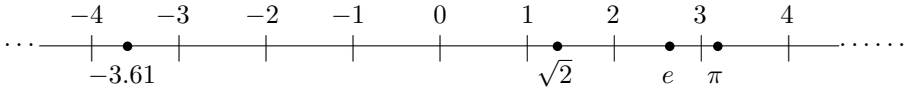
We will see that between any two numbers, no matter how close, there are infinitely many rational and infinitely many irrational numbers. However, if a number is chosen "at random", then it is "certain" to be irrational.

Although these ideas are discussed in [HM], we discuss them here in more depth.

The Real Number Line

[HM, 121–122]

On page 7 we briefly discussed how the real numbers correspond to the points on a line.



To do this we first need to decide which points corresponding to the numbers 0 and to 1. After this, every number then corresponds to a specific point on the line, and every point on the line corresponds to a specific number.

For example, the number 2 corresponds to the point twice the distance from 0 as 1 is from 0, and on the same side of 0 as is 1. The distance from 0 to $\sqrt{2}$ is the length of the hypotenuse of the right angled triangle whose two small sides are each 1 unit long.

We will soon examine decimal expansions of real numbers by studying properties of the real number line.

⋆*Decimal Expansions as Infinite Series*

Your calculator will tell you that

$$\sqrt{2} = 1.414213\ldots, \quad \text{i.e. } \sqrt{2} = 1+\frac{4}{10}+\frac{1}{10^2}+\frac{4}{10^3}+\frac{2}{10^4}+\frac{1}{10^5}+\frac{3}{10^6}+\ldots \quad (2.34)$$

On the right side we have an *infinite series* (or *infinite sum*). We can think of this as an infinite *sequence* of numbers:

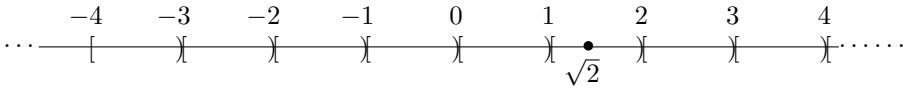
$$1, \quad 1.4 = 1+\frac{4}{10}, \quad 1.41 = 1+\frac{4}{10}+\frac{1}{10^2}, \quad 1.414 = 1+\frac{4}{10}+\frac{1}{10^2}+\frac{4}{10^3},$$
$$1.4142 = 1+\frac{4}{10}+\frac{1}{10^2}+\frac{4}{10^3}+\frac{2}{10^4},$$
$$1.41421 = 1+\frac{4}{10}+\frac{1}{10^2}+\frac{4}{10^3}+\frac{2}{10^4}+\frac{1}{10^5},$$
$$1.414213 = 1+\frac{4}{10}+\frac{1}{10^2}+\frac{4}{10^3}+\frac{2}{10^4}+\frac{1}{10^5}+\frac{3}{10^6}, \quad \ldots .$$

The idea is that this infinite sequence eventually gets as close as we wish to $\sqrt{2}$. We say that the sequence *converges* to $\sqrt{2}$. One can make this idea quite rigorous, although we will not do so here.

Geometric Interpretation of Decimal Expansions

[HM, 123–124]

Suppose that the real line is first partitioned into intervals of width one, beginning at 0.



Notation. We include the first endpoint in each interval but not the second. For example, one interval consists of all real numbers x such that $0 \le x < 1$. We read this as "$0 \le x$ *and* $x < 1$". We write the interval as $[0, 1)$, where "[" indicates we include 0 and ")" indicates we do not include 1.[54]

The interval to the right of $[0, 1)$ consists of all real numbers x such that $1 \le x < 2$, and we write this as $[1, 2)$. The interval to the left of $[0, 1)$ consists of all x such that $-1 \le x < 0$, and we write this as $[-1, 0)$. The next interval to the right of these three intervals consists of all x such that $2 \le x < 3$, and we write this as $[2, 3)$. Etc.

Dividing and Redividing the Real Number Line. Recall the decimal expansion of $\sqrt{2}$ in (2.34).

The number $\sqrt{2}$ is in the interval $[1, 2)$, see Figure 2.4. The 1 in $[1, 2)$ corresponds to the fact that the digit before the decimal point in the expansion of $\sqrt{2}$ is also 1.

[54] *More Notation*: The interval $(3, 7)$ consists of all numbers x such that $3 < x < 7$, the interval $[3, 7]$ consists of all numbers x such that $3 \le x \le 7$, and the interval $(3, 7]$ consists of all numbers x such that $3 < x \le 7$.

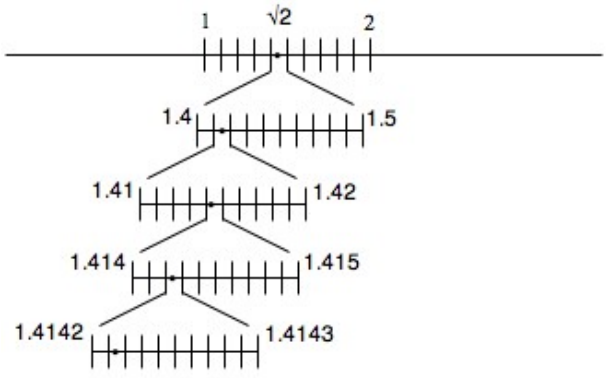


Figure 2.4: Decimal Expansion of $\sqrt{2}$

Next divide the interval $[1, 2)$ into 10 intervals of equal width $1/10$,

$$[1, 1.1), [1.1, 1.2), [1.2, 1.3), \ldots, [1.9, 2),$$

and number these intervals $0, 1, 2, \ldots, 9$. Then the first digit after the point in the decimal expansion for $\sqrt{2}$ tells us which of these intervals contains $\sqrt{2}$. This digit is 4 and the corresponding interval is $[1.4, 1.5)$.

Now divide the interval $[1.4, 1.5)$ into 10 intervals of equal width $1/10^2$,

$$[1.4, 1.41), [1.41, 1.42), [1.42, 1.43), \ldots, [1.49, 1.5),$$

and number these intervals $0, 1, 2, \ldots, 9$. Then the second digit after the decimal point in the decimal expansion for $\sqrt{2}$ tells us which of these intervals contains $\sqrt{2}$. This digit is 1 and the corresponding interval is $[1.41, 1.42)$.

Now divide the interval $[1.41, 1.42)$ into 10 intervals of equal width $1/10^3$,

$$[1.41, 1.411), [1.411, 1.412), [1.412, 1.413), \ldots, [1.419, 1.42),$$

and number these intervals $0, 1, 2, \ldots, 9$. Then the third digit after the decimal point in the decimal expansion for $\sqrt{2}$ tells us which of these intervals contains $\sqrt{2}$. This digit is 4 and the corresponding interval is $[1.414, 1.415)$.

Etc.

Addresses You can think of the decimal expansion $\sqrt{2} = 1.41421356\ldots$ as giving an infinite *address* for $\sqrt{2}$. The 1 before the decimal point gives the country, the first digit 4 after the decimal point gives the state, the second digit 1 gives the city, the third digit 4 gives the suburb, the fourth digit 2 gives the street, the fifth digit 1 gives the street number, the sixth digit 3 gives the apartment number, the seventh digit 5 gives the room number, etc.

Finding Addresses Notice that we can find the intervals containing $\sqrt{2}$, and hence the decimal expansion of $\sqrt{2}$, even if we did not know the decimal expansion before we started.

For example, $1^2 \le 2 < 2^2$, and so taking square roots $1 \le \sqrt{2} < 2$, and so the integer part of $\sqrt{2} = 1$.

Similarly $1.4^2 \le 2 < 1.5^2$, and so $1.4 \le \sqrt{2} < 1.5$, and so the decimal expansion of $\sqrt{2}$ up to one decimal point is 1.4.

Similarly, $1.41^2 \le 2 < 1.42^2$, and so $1.41 \le \sqrt{2} < 1.42$, and so the decimal expansion of $\sqrt{2}$ up to two decimal points is 1.41.

Similarly, $1.414^2 \le 2 < 1.415^2$, and so $1.414 \le \sqrt{2} < 1.415$, and so the decimal expansion of $\sqrt{2}$ up to three decimal points is 1.414. Etc.

✍ *Now try Question 1.*

Types of Decimal Expansions

[HM, 124–129]

There are three types of decimal expansions.

Definition 2.7.1. A decimal expansion is called *finite* (or *terminating*) if it stops after a finite number of places. Examples are

$$1.7, \quad 2.37, \quad 3.24658\,.$$

An infinite decimal expansion is called *periodic* if beginning from some position there is a finite pattern of digits which repeats itself forever. Examples are

$$.3333\cdots = .\overline{3}, \quad 2.34687168716871\cdots = 2.34\overline{6871}\,.$$

An infinite decimal expansion is called *non periodic* if it is *not* periodic.

In Theorem 2.7.3 we will see that the rational numbers correspond to finite or periodic decimal expansions while the irrational numbers correspond to non periodic decimal expansions.

Finite Decimal Expansions. A finite decimal expansion represents the same number as the infinite periodic decimal expansion obtained by adding an infinite string of 0's at the end. For example,

$$1.7 = 1.7\overline{0},\ 2.37 = 2.37\overline{0},\ 3.24658 = 3.24658\overline{0}.$$

We use the infinite decimal expansion with 0's when we want to think of the decimal expansion as giving an "address" of the number.

The Decimal Expansion $.\overline{9}$. *A result which often seems surprising at first is that* $1 = 1.0 = .\overline{9}$. It is *not* the case that "$.\overline{9}$ is just a little bit less than 1".

One way to see this is as follows. First notice that multiplication by 10 moves the decimal point one digit to the right. For example, if

$$x = \frac{3}{10} + \frac{1}{10^2} + \frac{5}{10^3} + \frac{7}{10^4} + \frac{6}{10^5} + \cdots = .31576\ldots\,,$$

then

$$10x = 3 + \frac{1}{10} + \frac{5}{10^2} + \frac{7}{10^3} + \frac{6}{10^4} + \cdots = 3.1576\ldots\,.$$

So in particular, if $x = .\overline{9}$ then $10x = 9.\overline{9}$. If we subtract, this gives $9x = 9$ and so $x = 1$.

Each of the *approximations* .9, .99, .999, , 999, ... to 1 is less than 1. But the difference from 1 can be made as small as we wish by taking an approximation with sufficiently many 9s. The number *represented* by the infinite decimal expansion $.\overline{9}$ is *exactly* 1.

More than One Infinite Decimal Expansion. Any number which has a finite decimal expansions can also be written *both* as a periodic expansion with an infinite string of 0's at the end *and* as a periodic decimal expansions with an infinite string of 9's at the end. For example,

$$\begin{gathered} 1.7 = 1.7\overline{0} = 1.6\overline{9},\ 2.37 = 2.37\overline{0} = 2.36\overline{9}, \\ 3.248 = 3.248\overline{0} = 3.247\overline{9},\ 6 = 6.\overline{0} = 5.\overline{9},\ 300 = 300.\overline{0} = 299.\overline{9} \end{gathered} \tag{2.35}$$

You can show these with a method similar to that used for $.\overline{9}$. *Write out the proof for the above numbers.* We will also write out the method in Step **1** in the proof of Theorem 2.7.3.

Here is an important Theorem. Look at the examples in (2.35) to understand what it is saying. Unfortunately the proof tends to obscure the main ideas, which are essentially contained in (2.35).

Theorem 2.7.2. *If a real number $x > 0$ has a finite decimal expansion then it also has exactly two infinite decimal expansions. One of these infinite decimal expansions is the same as the finite decimal expansion followed by an infinite string of 0's. The second infinite decimal expansion is obtained by decreasing the last non zero digit in the finite expansion by 1 and then following this by an infinite string of 9's.*

Real numbers x which do not *have a finite decimal expansion have exactly one infinite decimal expansion.*

Proof. First suppose x has two or more decimal expansions, two of which agree up to the third place (say) and then disagree in the fourth place, such as $6.3274\ldots$ and $6.3275\ldots$. Even though we will deal with a particular example for simplicity, we will see that our argument is general.

The largest number with a decimal expansion of the form $6.3274\ldots$ is $6.3274\overline{9}$. Any other number with a decimal expansion of the form $6.3274\ldots$ is strictly smaller than $6.3274\overline{9}$. For example, $6.327499989999\ldots$ is smaller by .00000001.

The smallest number with a decimal expansion of the form $6.3275\ldots$ is $6.3275\overline{0}$. Any other number with a decimal expansion of the form $6.3275\ldots$ is strictly larger than $6.3275\overline{0}$.

Thus the *only* way x can have decimal expansions of the form $6.3274\ldots$ and $6.3275\ldots$ is if the first expansion is $6.3274\overline{9}$, the second is $6.3275\overline{0}$, and x then has the finite decimal expansion 6.3275.

Similarly, for any x which has two infinite decimal expansions agreeing in the first three places (say) and disagreeing in the fourth place, let us write these two decimal expansions as $.a_1a_2a_3b\ldots$ and $.a_1a_2a_3c\ldots$, where we take b less than c. In order for the two decimal expansions to be *equal* we must have,

by a similar argument to before, $x = .a_1a_2a_3b\,\overline{9} = .a_1a_2a_3c\,\overline{0} = .a_1a_2a_3$, and $b + 1 = c$.

A similar argument also applies more generally to any x with two or more infinite decimal expansions by looking at the first place where two of the decimal expansions differ. It follows that such an x has a finite decimal expansion and has *exactly* two infinite decimal expansions.

Summarising: A number x has two or more infinite decimal expansions if and only if it has a finite decimal expansion, and it then has *exactly* two infinite decimal expansions.

We have just seen that if a number has more than one infinite decimal expansion then it has a finite decimal expansion. This means that if it does *not* have a finite decimal expansion then it does *not* have more than one infinite decimal expansion.[55]

This proves the statement in the second paragraph of the theorem. □

Decimal Expansions of Rational and Irrationals. The following Theorem tells us which infinite decimal expansions represent rational numbers and which represent irrational numbers.

Note that finite decimal expansions always represent rational numbers. But not all rational numbers have finite decimal expansions. *Why?*

Theorem 2.7.3. *An infinite decimal expansion is periodic if and only if it represents a rational number. It is non periodic if and only if it represents an irrational number.*

Proof. **1.** If a decimal expansion is periodic then we can write it as a fraction by using the following method.

The basic idea is to move the decimal point to the right so that the repeating parts again line up. This requires multiplication by a suitable power of 10. For example, if the length of the repeating part is 4 then multiply by $10^4 = 10000$ to move the decimal point along 4 digits. After subtracting the original decimal expansion we then get a *terminating* expansion which can easily be written as a fraction.

For example,

$$\begin{aligned} x &= .333333333\ldots = .\overline{3}\,, \qquad y = .\overline{9}\,, \\ 10x &= 3.333333333\ldots = 3.\overline{3}\,, \quad 10y = 9.\overline{9}\,. \end{aligned}$$

Subtracting gives $9x = 3$ and so $x = \dfrac{1}{3}$. Similarly, $9y = 9$ so $y = 1$.

Here is another example. Since the repeating part has length 5 we multiply by $10^5 = 100,000$.

$$\begin{aligned} x &= 2.34\overline{68715} \\ 100\,000x &= 234\,687.15\overline{68715} \end{aligned}$$

[55] There is a logical principle involved here. Namely, if $P \Longrightarrow Q$ then not $Q \Longrightarrow$ not P. Think about it!

Subtraction gives $99\,999x = 234\,684.81 = \dfrac{23\,468\,481}{100}$ and so $x = \dfrac{23\,468\,481}{9\,999\,900}$.

We have shown that if an infinite decimal expansion is periodic then it gives a rational number.

2. We will now show that if a number is rational then it either has a terminating decimal expansion or otherwise its infinite decimal expansion is periodic.[56]

In order to find the decimal expansion of a rational number[57] we need to do a long division. For example, to find the decimal expansion of 491/165 we do the following long division.

$$
\begin{array}{rl}
 & 2.975\ldots \\
165)\overline{491.000\ldots} & \\
 & \underline{330} \quad \text{(multiplication of 165 by 2)} \\
 & \mathbf{161}0 \\
 & \underline{1485} \quad \text{(multiplication of 165 by 9)} \\
 & \mathbf{125}0 \\
 & \underline{1155} \quad \text{(multiplication of 165 by 7)} \\
 & \mathbf{95}0 \\
 & \underline{825} \quad \text{(multiplication of 165 by 5)} \\
 & \mathbf{125}0
\end{array}
$$

The remainder at each stage is indicated in bold. It must be between 0 and 164. Here we see the remainders are $161, 125, 95$ and then 125 again. From here on the long division repeats itself and so we see that $491/165 = 2.9\overline{75}$.

The number of possible remainders is always the same as the number we are dividing by. So eventually we must come again to a remainder that has already occurred. From then on the long division repeats itself, which means that the decimal expansion is periodic.

3. We have now shown that a number is rational if and only if its decimal expansion is terminating or is periodic. It follows that a number is irrational if and only if its decimal expansion is non periodic. □

Some Curious Irrational Numbers. The following numbers

$$
\begin{aligned}
a &= .10100100010000100000100000010000000001\ \ldots \\
b &= .01011011101111011111011111011111110\ldots
\end{aligned}
\tag{2.36}
$$

have nice patterns. But are they rational? NO.

(We are using base 10 here, not base 2. If you wish, you can for example replace 1 everywhere by 7, and the same result will be true)

Theorem 2.7.4. *The numbers a and b in* (2.36) *are irrational.*

[56] This does not follow from what we have already shown. For example, cats are mammals, but mammals are not always cats.

[57] We will see in the next Section that some rational numbers have two decimal expansions, but then *both* expansions will be periodic.

Proof. We saw in Theorem 2.7.3 that the number given by a decimal expansion is rational if, and only if, the decimal expansion is periodic.

But the decimal expansion for a is not periodic. This looks pretty clear, and here is a careful argument.

If the decimal expansion given for a is periodic then there is a repeating pattern of length N (say) first starting at the kth digit (say) and then again at $k+N, k+2N, k+3N, \ldots, k+pN, \ldots$. On the other hand, there are runs of 0 which are as long as we want, and certainly of length much greater than N. For example there will be (infinitely many) sequences of 0s of length at least googol $\times N$.

Eventually there will be a *first* p such that $k+pN$ is in one of these very long sequences of 0's, and in particular there will only be 0's up to $k+(p+1)N$. This means the repeating block consists only of 0's, and so the decimal expansion for a will consist only of 0's beyond a certain point. But this contradicts the fact that there are infinitely many 1's in a.

So a is *not* rational, i.e. a is irrational.

In a similar way, b is irrational. □

Notice by the way that $a + b = .\overline{1} = 1/9$, which is rational.

In [HM] a similar proof is given that

$$0.12345678910111213141516171819202122232425\ldots$$

is irrational.

Binary Expansions

[HM, 124]

In Footnote 15 we briefly discussed different bases other than the base 10 (decimal) system for natural numbers. In particular we mentioned the bases 2, 16 and 60.

In fact we can find the expansion of arbitrary real numbers, not just natural numbers, to any base. For example, the base 2 (or *binary*) expansion of $\sqrt{2}$ is

$$\sqrt{2} = 1.0110101000001001111\cdots = 1 + \frac{0}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{0}{2^4} + \frac{1}{2^5} + \ldots.$$

✍ To find this binary expansion divide the interval $[1, 2)$ into two parts $[1, 3/2)$ and $[3/2, 1)$. Then $\sqrt{2}$ is in the left interval (*why?*), see Fig. 2.5, and so the first digit after the point is 0. Dividing this interval into two parts $[1, 5/4)$ and $[5/4, 6/4)$ we can check that $\sqrt{2}$ is in the right interval and so the second digit after the point is 1. If we divide again then $\sqrt{2}$ is again in the right interval and so the third digit after the point is 1. Etc.

✍ *Now try Question 2.*

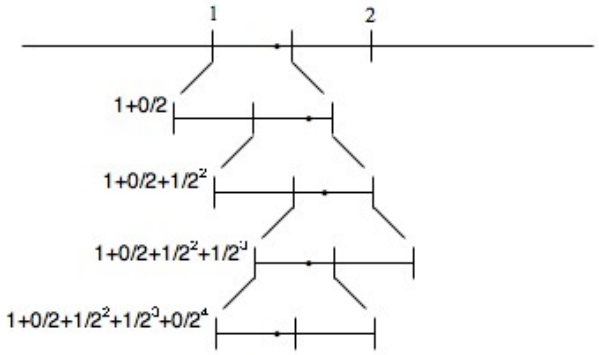


Figure 2.5: Binary Expansion of $\sqrt{2}$.

⋆*Density of the Rationals and the Irrationals*

We saw on page 70 that there are lots of rational numbers. For any two rational numbers, no matter how close, there is a rational number between them, just take their average.

More generally, we have the following Theorem.

Theorem 2.7.5. *Between any two distinct numbers a and b, no matter how close, there is both a rational number and an irrational number.*

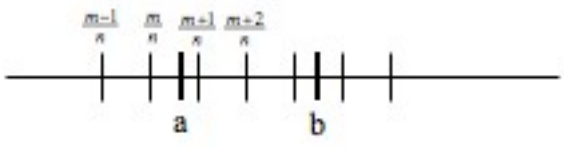
In fact, there are infinitely many rational numbers, and infinitely many irrational numbers, between a and b.

Proof. If a and b are distinct numbers, let a be the smaller. Choose n sufficiently large that $\dfrac{1}{n} < b - a$. Notice that $b - a$ is just the distance from a to b.

Now consider the numbers

$$\ldots, -\frac{4}{n}, -\frac{3}{n}, -\frac{2}{n}, -\frac{1}{n}, 0, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \frac{4}{n}, \ldots.$$

If m is the *largest* integer such that $\dfrac{m}{n} \leq a$ then we see from the following diagram that $a < \dfrac{m}{n} + \dfrac{1}{n} < b$.[58]



Thus we have found a rational number $\dfrac{m+1}{n}$ between a and b.

[58]The reason is that $1/n$ is too small to "bridge the gap" from a and b. The careful argument is that $(m+1)/n > a$ by the definition of m as the largest integer such that $m/n \leq a$. But if $(m+1)/n \geq b$ then $(m+1)/n - m/n \geq b - a$ by properties of inequalities. This implies $1/n \geq b - a$, which contradicts the fact we chose n so that $1/n < b - a$.

One way to find an irrational number between a and b is to choose n even larger if necessary to ensure that $\frac{\sqrt{2}}{n} < b - a$. Again choose the largest integer m such that $m/n \leq a$. It follows as before that the irrational number $\frac{m}{n} + \frac{\sqrt{2}}{n} = \frac{m + \sqrt{2}}{n}$ is between a and b. *Why?*

✍

In order to obtain an infinite set of rational numbers between a and b, we just reuse the result already proved.

More precisely, we first obtain a rational number r_1 where $a < r_1 < b$. Then by what we have already shown, there is another rational number r_2 where $a < r_2 < r_1$. And then there is another rational number r_3 where $a < r_3 < r_2$. Etc.

In a similar way we obtain an infinite number of irrational numbers between a and b. □

It follows from the Theorem that:
For any real number there is no next real number immediately before it or immediately after it.

Because of the Theorem and the above comment we say that the set of rational numbers, and the set of irrational numbers, are each *dense* in the set of all real numbers.

No Holes, Nothing Missing

[HM, 127]

Even though the rational numbers are dense in the set of real numbers, we still managed to further squeeze in the irrational numbers. Do you think there are even more numbers we could squeeze in if we were suitably ingenious? The answer is NO, but we first need to formulate the question more precisely.

We know that $\sqrt{2}$ can be approximated by an increasing sequence of *rational* numbers obtained from its decimal expansion:

$$1.4, 1.41, 1.414, 1.4142, 1.41421, 1.414213, \ldots .$$

In other words, we have a sequence of *rational* numbers which is increasing and which gets closer and closer to $\sqrt{2}$, which is *not* rational.

In fact, if we take *any* sequence of real (not necessarily rational) numbers which is increasing, provided it is not getting arbitrarily large (such as the sequence $1, 2, 3, 4, \ldots$), it will always get closer and closer to one, and exactly one, real number. This property of the real numbers is known as *Sequential Completeness*. We can think of it as a very strong statement about there being no holes in the set of real numbers!

Sequential completeness follows from the usual properties of decimal expansions. But we won't go any further in trying to write this out carefully. In most rigorous developments of the real number system sequential completeness is taken to be one of the axioms, or part of one of the axioms, and the properties of decimal expansions are then derived from the axioms.

Random Reals

[HM, 129-130]

We saw in Theorem 2.7.5 that between any two distinct real numbers, no matter how close, there are infinitely many rational numbers and infinitely many irrational numbers. However, in a probabilistic sense, we will now see that irrational numbers are far more common than rational numbers.

A Thought Experiment. We will pick a random number between 0 and 1 as follows.

Spin a roulette wheel divided into 10 equal sectors and marked $\{0, 1, 2, \ldots, 9\}$. Suppose the result is 3. Write down the number .3.

Now spin the wheel a second time and suppose the result is 7. Write down the number .37.

Now spin the wheel a third time and suppose the result is 0. Write down .370.

Keep going in this manner and we might, for example after 10 steps, have .3708854216.

Now imagine that we do this infinitely often, thus generating a decimal expansion and hence a number x where $0 \leq x \leq 1$.

The idea is that the first digit is chosen from $\{0, 1, 2, \ldots, 9\}$ where each digit has probability 1/10 of being chosen. The second digit is chosen randomly in a similar manner, independently of the first digit. And so on.

You can ask Maple to do this with its inbuilt random number generator.

Is the number x rational or irrational?

If x were rational then after some point we would repeat the same pattern again and again forever. How likely is this? It could happen in principle, but the probability of it happening is a number less than any positive number, and greater than or equal to 0. The only such number is 0 itself.

This may seem paradoxical. You will probably agree that it is extremely unlikely that we would repeat the same pattern again and again forever, and probably agree that the probability of this happening should be smaller than any positive real number. Yet you may initially be unhappy about the idea of assigning 0 to the probability of an event which could happen in principle even though it is "certain" it will never happen.

This takes us into the area of modern probability theory, which we will not have time to discuss in this course.

Questions

1 Find the first few digits in the decimal expansion of $\sqrt[3]{3}$. Use your calculator but only to multiply numbers. Do not use the cube root function, or anything fancy like that.

2 Find the first few digits to base 2 of $\sqrt{3}$ and $\sqrt[3]{2}$ (note that here when we write $\sqrt{3}$ and $\sqrt[3]{2}$ we are using base 10).

DON'T LOOK YET but here[59] are the answers.

See [HM, pp132–134] for other Questions.

[59] The answers are $1.1011101101100111101\ldots$ and $1.0100001010001010001\ldots$.

Chapter 3

Infinity

Infinity has long been a source of philosophical speculation. With mathematics we can make sense of it. We will replace vague ideas with precise notions with which we can work.

For some history, see page 129.

Is infinity interesting? Yes. We will encounter some amazing and initially counterintuitive ideas.

Is infinity useful for anything? The answer is a resounding yes. Almost all mathematical analysis in science and technology, economics and engineering, uses ideas involving infinity. The notion of an infinite limit is fundamental to calculus, and to the extensions of calculus to more general notions.

We will later show that some infinities are larger than others. There is an infinity of infinities. We will discover and discuss many other amazing and profound results

We begin with an analysis of when two collections are the same size.

Contents

3.1 Comparing Sets

A key to understanding the complex unknown can be a deep understanding of the simple and familiar.

Overview

We discuss what it means for two collections (i.e. sets) to be the same size via the idea of a one-to-one correspondence between collections.

We also introduce some notation about sets.

One-to-One Correspondences

[HM, 138–140]

Sets of the Same Size How can we show two collections have the same size?

Suppose we have a collection of basket balls and another collection of large boxes. We could count each collection, and if we get the same number in each case we would agree that the two collections have the same size.

But suppose the number of basket balls and of boxes was very large and we could not count that high, or that we did not trust ourselves to not make a mistake.

In any case, another way would be to pair up balls and boxes by putting each ball into a box so that every ball is in exactly one box and every box contains exactly one ball. If we could do this, with no balls or boxes left over, we would agree that the collection of balls and the collection of boxes were each of the same size.

We say there is a *one-to-one pairing* or *one-to-one correspondence* between the collection of balls and the collection of boxes.

Definition of One-to-One Correspondence We now make this idea more precise. In the following definition you should first think of A as the collection of basket balls and B as the collection of boxes in the previous discussion.

Definition 3.1.1. A *one-to-one correspondence*[1] between two collections A and B is a pairing of the members of A with the members of B so that under this pairing

- every member of A is paired with exactly one member of B, and
- every member of B is paired with exactly one member of A.

[1] For those of you who have seen the idea of a function and related notions, here is an equivalent approach:

A *one-to-one correspondence* between two sets A and B is given by a function $f : A \to B$ i.e. a function f from A to B) which is one-to-one and onto.

The actual one-to-one correspondence is given by $a \leftrightarrow f(a)$ for all $x \in A$.

Such a function f has an inverse $f^{-1} : B \to A$, which is defined by $f^{-1}(b) = a$ if and only if $f(a) = b$. The inverse function f^{-1} is one-to-one and onto. The one-to-one correspondence is also given by $f^{-1}(b) \leftrightarrow b$ for all $b \in B$.

Examples

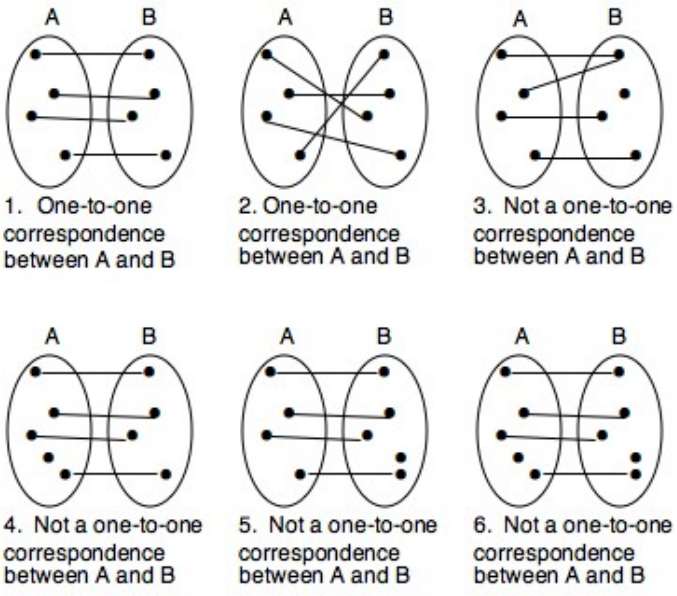


Figure 3.1: One-to-One Correspondence

1. Examples 1 and 2 in Figure 3.1 are one-to-one correspondences between A and B.
2. Example 3 is not a one-to-one correspondence since the top two elements in A are paired with the same element in B. It is also not a one-to-one correspondence since one element in B is not paired with any element in A.

   However, it is possible to give a one-to-one correspondence, see Example 1 or 2.
3. Examples 4 and 5 are not one-to-one correspondences since in the first case there is an element in A which is not paired with any element in B, and in the second case there is an element in B which is not paired with any element in A.
4. Example 6 is not a one-to-one correspondence since there is an element in A which is not paired with any element in B. It is also not a one-to-one correspondence since there is an element in B which is not paired with any element in A.

   However, it is possible to give a one-to-one correspondence. *Find one. Now find a second.*

✍

Definition of Same Cardinality The following Definition is fundamental to the rest of our wok on infinity. It gives a precise description on when two sets have the same size, or as we usually say, have the same cardinality.

Definition 3.1.2. Two sets A and B have the same size (or same *cardinality*) if[2] there exists a one-to-one correspondence between A and B.

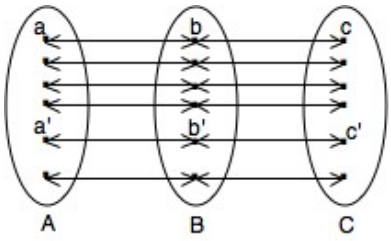
This material is not in [HM

⋆Comparing Three Sets The following Theorem is not surprising.

Theorem 3.1.3. *Suppose there is a one-to-one correspondence between A and B, and another one-to-one correspondence between B and C.*

Then there is a one-to-one correspondence between A and C.

Proof.



Consider any element a in A (see the above diagram). It will be paired with exactly one element b in B. The element b is paired with exactly one element c in C. The one-to-one correspondence between A and C is obtained by pairing a with c.

It is fairly clear that this gives a one-to-one correspondence between A and C according to Definition 3.1.1.

Here is a brief explanation of why the two dot point requirements in Definition 3.1.1 are satisfied for A and C.

- Every element a in A is paired with exactly one element c in C, namely the element c we defined as above.
- On the other hand, consider any element c in C. By taking the unique b in B which is paired with c and then the unique a in A which is paired with this b , we certainly obtain *some* element a in A which is paired with c by our method.

  This c cannot be paired with any other a' in A. The reason is that a' will be paired with some b' different from b, and this b' will in turn be paired with some c' different from c. But this means that a' is paired with c' which is *different* from c. That is, if a is different from a' then c is different from c'.

Thus we have shown that our pairing is indeed a one-to-one correspondence between A and C according to Definition 3.1.1. □

Notation for Sets

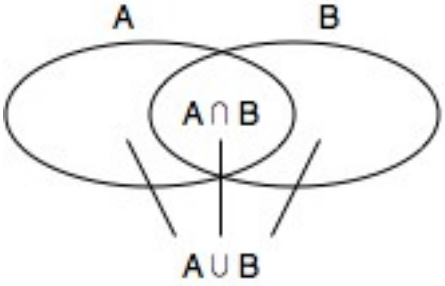
This material is not in [HM

1. A *collection* of objects is usually called a *set* in mathematics. Occasionally we also use the word *class*.

[2]It is standard practice in Mathematics to use the word "if" in a definition to mean "if and only if". This is logical since it would not be a complete definition of a concept if we did not mean "if and only if".

2. A member of a set is also called an *element* of the set. If a is an element of the set A we write $a \in A$.
3. We say that two sets are *equal* (or are the same) if they have the same elements.
4. If the elements of a set A are a, b and c then we write $A = \{a, b, c\}$ (and similarly for other examples). The sets $\{a, b, c\}$, $\{b, a, c\}$, $\{c, b, a\}$ etc. are all the same.

   For example, the set of even integers between 0 and 10 inclusive is the set $\{0, 2, 4, 6, 8, 10\}$.
5. If A and B are sets, then the *union* $A \cup B$ of A and B is the set of all elements that are in A *or* B (or both).[3] The *intersection* $A \cap B$ of A and B is the set of all elements that are in both A *and* B.



   For example, if $A = \{1, 2, 3\}$ and $B = \{2, 3, 4, 5\}$ then $A \cup B = \{1, 2, 3, 4, 5\}$ and $A \cap B = \{2, 3\}$.

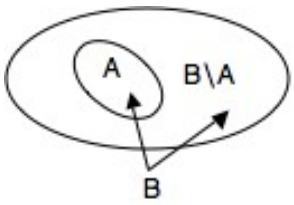
   Notice that we do not normally write $A \cup B = \{1, 2, 2, 3, 3, 4, 5\}$. For example, 3 is either in, or is not in, a set. It cannot be in a set "twice".
6. If A, B and C are sets, then the *union* $A \cup B \cup C$ is the set of all elements that are in A or B or C. (Remember that "or" includes the possibility that more than one of these alternatives is true.) The *intersection* $A \cap B \cap C$ is the set of elements that are in A and B and C.

✍ *Draw a diagram analogous to that for $A \cup B$ and $A \cap B$.*
7. It is convenient to have a set with no members, which we call the *empty set* and denote by $\emptyset$. This is useful because then the intersection of two sets is always a set. If the two sets have no elements in common then their intersection is the empty set.

   There is only one empty set, since any two empty sets have the same members, and so by item 3 are the same!
8. If A is a set then we say B is a *subset* of A, and we write $B \subset A$, if every element in B is also in A. In particular, $\emptyset \subset A$ and $A \subset A$.[4]
9. If A and B are sets and $A \subset B$ then the set $B \setminus A$ is the set of all elements in B that are not in A. $B \setminus A$ is called "B minus A" or "B take A".

[3] In Mathematics, when we say "a statement P is true *or* a statement Q is true" we *always* allow the possibility that both are true. So when we say that $x \in A \cup B$ if and only if $x \in A$ or $x \in B$, we allow the possibility that *both* $x \in A$ and $x \in B$ are true.

[4] This may be slightly different from the notation you will see elsewhere. Some books write $B \subseteq A$ where we write $B \subset A$. Some books write $B \subset A$ only if B is a subset *and* there is at least one element in A that is not in B.



Remark It follows from Theorem 3.1.3 that if A has the same cardinality as B and B has the same cardinality as C then A has the same cardinality as C.

Further discussion of the Examples

1. In Examples 1 and 2 on page 90 we saw that A and B have the same cardinality.
2. Example 3 does not give a one-to-one correspondence. But A and B still have the same cardinality as it *is* possible to give a one-to-one correspondence between A and B in this case, as we have already noted.
3. In Examples 4 and 5 it is not possible to find *any* one-to-one correspondence between A and B. In the first case, no matter how we try to pair up the elements in A with those in B, there there will always be one element in A left unpaired. A similar situation applies in Example 5.
4. Example 6 does not give a one-to-one correspondence between A and B. But A and B still have the same cardinality as it *is* possible to give a one-to-one correspondence between A and B in this case, as we have noted previously.

Questions

See [HM, pp142–144] for Questions. Try Questions 8, 9, 21, 22.

3.2 Countably Infinite Sets

Using a precise definition carefully can lead to counterintuitive discoveries that liberate our thinking and our view of the world.

Overview

We discuss what it means for two infinite sets to have the same cardinality (size) and come up with some rather surprising results.

We will show that the set of natural numbers, the set of integers, and the set of rational numbers all have the same cardinality in a certain precise sense.

Sets with Equal Cardinality

From now on we will take Definition 3.1.2 as fundamental! But we will use it with infinite sets.

It follows (*why?*) from Definition 3.1.2 that:

> *(i) If we can find* ***some*** *one-to-one correspondence between two sets then they have the same (or equal) cardinality (size).*
>
> *(ii) If we can show there is* ***no*** *one-to-one correspondence between two sets then they do* ***not*** *have the same (or equal) cardinality (size).*

At this stage you will need to commit yourself to Definition 3.1.2. Perhaps you could write down on a piece of paper:

> I hereby declare that two sets have the same size if they can be paired up via a one-to-one correspondence.
>
> Moreover, if it can be established by some means that there is no one-to-one correspondence between these two sets (by using an argument by contradiction or otherwise) then the two sets do not have the same size.
>
> ---
>
> *Signature*

Finding one-to-one correspondences in certain cases can be quite tricky, and the results are often surprising.

Showing there is *no* one-to-one correspondence in certain cases (not just that we ourselves cannot find one!) is even more tricky and subtle, and we will do this in Section 3.3.

Comparing Some Sets of Natural Numbers

[HM, 145–148]

Even and Odd Natural Numbers We will use $\mathbb{N}$ for the set of natural numbers, E for the set of even natural numbers, and O for the set of odd natural numbers. That is

$$\begin{aligned}\mathbb{N} &= \{1, 2, 3, 4, 5, 6, \dots\},\\ E &= \{2, 4, 6, 8, 10, 12, \dots\},\\ O &= \{1, 3, 5, 7, 9, 11, \dots\}.\end{aligned}$$

It is clear that there is a one-to-one correspondence between E and O. For example

$$2 \leftrightarrow 1, 4 \leftrightarrow 3, 6 \leftrightarrow 5, 8 \leftrightarrow 7, 10 \leftrightarrow 9, 12 \leftrightarrow 11, \dots.$$

We can even write down a formula:

$$2n \leftrightarrow 2n - 1 \quad \text{for} \quad n \in \mathbb{N}.$$

There are also many other one-to-one correspondences. For example we might change just the first two pairings:

$$2 \leftrightarrow 3, 4 \leftrightarrow 1, 6 \leftrightarrow 5, 8 \leftrightarrow 7, 10 \leftrightarrow 9, 12 \leftrightarrow 11, \dots.$$

Even v. All Natural Numbers Somewhat more surprising is that there is a one-to-one correspondence between $\mathbb{N}$ and E. The simplest one-to-one correspondence between $\mathbb{N}$ and E is

$$1 \leftrightarrow 2, 2 \leftrightarrow 4, 3 \leftrightarrow 6, 4 \leftrightarrow 8, 5 \leftrightarrow 10, 6 \leftrightarrow 12, \dots, n \leftrightarrow 2n, \dots.$$

(Here we have also written the general formula.) Every natural number corresponds to exactly one even natural number, and under this pairing every even natural number corresponds to exactly one natural number.

This is surprising because we have shown a one-to-one correspondence between the set $\mathbb{N}$ and another set E, where E is obtained by omitting certain elements (the odd integers) in $\mathbb{N}$. Thus we have an example where a set $\mathbb{N}$ is the same size as a second set E obtained after discarding certain elements, in fact infinitely many, from $\mathbb{N}$.

Another Set of Natural Numbers Another example is a one-to-one correspondence between $\mathbb{N}$ and the set $\mathbb{N}^* = \{2, 3, 4, 5, 6, 7, \dots\}$ consisting of all the natural numbers ≥ 2. So $\mathbb{N}^*$ is obtained from $\mathbb{N}$ by discarding the number 1.

One example of a one-to-one correspondence between $\mathbb{N}$ and $\mathbb{N}^*$ is:

$$1 \leftrightarrow 2, 2 \leftrightarrow 3, 3 \leftrightarrow 4, 4 \leftrightarrow 5, 5 \leftrightarrow 6, \dots, n \leftrightarrow n + 1, \dots.$$

Finite and Countably Infinite Sets

Most of this material is no

At this stage it is convenient to make the following Definition.

Definition 3.2.1.

- A set S is *finite* if it is the empty set $\emptyset$ or if it has the same cardinality as $\{1, 2, \ldots, k\}$ for some natural number k. We say S has *cardinality* 0 in the first case and *cardinality* k in the second case.
- A set is *infinite* if it is not finite.[5]
- A set S is *countably infinite* if it has the same cardinality as $\mathbb{N}$. We say S has *cardinality* d or S has *cardinality* $\aleph_0$.
- A set is *countable* if it is finite or countably infinite.

We say $0, 1, 2, \ldots, n, \ldots$ and d (or $\aleph_0$) are *cardinals* or *cardinal numbers.*

The d in the previous definition stands for "denumerable". The symbol $\aleph$ is the first letter "aleph" of the Hebrew alphabet, and we say "aleph zero" for $\aleph_0$.

We saw E, O and $\mathbb{N}^*$ have cardinality d by writing the elements in each set as the elements of an infinite sequence (or "list") $s_1, s_2, s_3, s_4, \ldots$. The elements in each set occurred exactly once in the corresponding sequence:

Sequence:	s_1	s_2	s_3	$\ldots$	s_n	$\ldots$
All natural numbers:	1	2	3	$\ldots$	n	$\ldots$
Even natural numbers:	2	4	6	$\ldots$	$2n$	$\ldots$
Odd natural numbers:	1	3	5	$\ldots$	$2n-1$	$\ldots$
Integers ≥ 2:	2	3	4	$\ldots$	$n+1$	$\ldots$

The above are examples of a general fact which we state as the following Theorem, which is essentially a rewording of Definition 3.2.1. A little more informally the Theorem says:

> *A set is finite iff it is empty or can be listed as a finite sequence.*
>
> *A set is countably infinite iff it can be listed as an infinite sequence.*

Theorem 3.2.2. *A set S is finite iff it is the empty set or for some natural number k there is a finite sequence*[6]

$$s_1, \ldots, s_k$$

such that each element of S occurs exactly once in the sequence.

A set S is countably infinite iff there is an infinite sequence

$$s_1, s_2, \ldots, s_n, \ldots$$

such that each element of S occurs exactly once in the sequence.

Proof. From Definition 3.2.1:

- S is finite iff it is the empty set or there is a one-to-one correspondence between it and the set $\{1, 2, \ldots, k\}$ for some natural number k.

[5] Another equivalent definition is that a set is infinite if it can be put in one-to-one correspondence with a proper subset of itself. See Theorem 3.3.4.

[6] *It is important to realise that the order matters when we list a sequence. Changing the order changes the sequence. On the other hand the order does not matter when we describe a set. Changing the order does not change the set.*

- S is countably infinite iff there is a one-to-one correspondence between it and the set $\mathbb{N} = \{1, 2, \ldots, n, \ldots\}$.

The proof now is just a matter of noticing that:

- There is a one-to-one correspondence between S and $\{1, 2, \ldots, k\}$ is just another way of saying that S can be written as a "finite" sequence

$$s_1, \ldots, s_k.$$

The one-to-one correspondence gives us the finite sequence and the finite sequence gives us the one-to-one correspondence. *Why?* ✍

- There is a one-to-one correspondence between S and $\{1, 2, \ldots, n, \ldots\}$ is just another way of saying that S can be written in an infinite sequence

$$s_1, \ldots, s_n, \ldots$$

The one-to-one correspondence gives us the infinite sequence and the infinite sequence gives us the one-to-one correspondence. *Why?* ✍

This completes the proof. □

At this stage you might think that *every* infinite set can be arranged in a sequence and so every infinite set has the same cardinality as $\mathbb{N}$. In Section 3.3 we will show that in fact the set of all real numbers *cannot* be written in an infinite sequence! In the rest of this section we will give some more sets which *can* be written as an infinite sequence and so are countably infinite.

Countability of $\mathbb{Z}$

[HM, 151,152]

The set of all integers is denoted by $\mathbb{Z}$:

$$\mathbb{Z} = \{\ldots, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, \ldots\}. \tag{3.1}$$

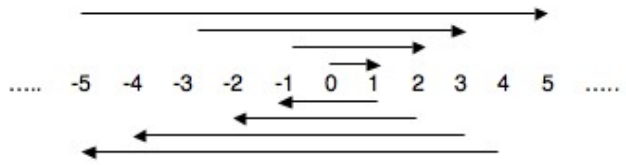
Is $\mathbb{Z}$ countably infinite? In other words, does $\mathbb{Z}$ have the same cardinality as $\mathbb{N}$?

Someone might answer that $\mathbb{Z}$ is countably infinite *because* it is already arranged in an infinite sequence. But their logic would be wrong, since (3.1) is not a sequence in our sense. A sequence as used in Theorem 3.2.2 is something that has a first element, then a second element, then a third element, ..., then an nth element, ... (We do not require that the first element be the smallest element, by the way.)[7]

This person might then answer that $\mathbb{Z}$ is not countably infinite *because* it is not arranged in an infinite sequence in our sense. But again their logic would be wrong — Theorem 3.2.2 only requires that the set in question *can* be arranged in a sequence.

In fact it is possible to arrange $\mathbb{Z}$ in a sequence. Beginning with 0 move one step right to 1, then 2 steps left to -1, then 3 steps right to 2, then 4 steps left to -2, etc. This is perhaps best seen in the following diagram.

[7] Occasionally you might see the arrangement in (3.1) referred to as a *bi-infinite* sequence or *two-sided* sequence.



Theorem 3.2.3. *The set $\mathbb{Z}$ of all integers is countably infinite.*

Proof. The sequence

$$0, 1, -1, 2, -2, 3, -3, 4, -4, 5, -5, \ldots \tag{3.2}$$

includes every element in $\mathbb{Z}$ exactly once. It follows from Theorem 3.2.2 that $\mathbb{Z}$ is countably infinite. □

Although we do not really need a formula for the proof of Theorem 3.2.3, you can check that the nth term in the sequence is $-(n-1)/2$ if n is odd and is $n/2$ if n is even.

Countability of $\mathbb{Q}$

[HM, 152–155]

The set $\mathbb{Q}$ of rational numbers is also countably infinite. This is very surprising at first.

In the case of $\mathbb{Z}$ we were able to rearrange the natural ordering (3.1) into the sequence (3.2). But it is very far from clear how we might do this for $\mathbb{Q}$. One problem is that between any two numbers there are infinitely many rational numbers. (We say $\mathbb{Q}$ is *dense* in the set $\mathbb{R}$ of all real numbers, see page 83.)

The solution lies in returning to the original definition of a rational number. Recall that a number is rational if it can be written in the form m/n where m and n are integers and $n \neq 0$.

The proof of the following theorem is not at all obvious, and indeed is quite remarkable.

Theorem 3.2.4. *The set $\mathbb{Q}$ of rational numbers is countably infinite.*

Proof. We will first prove that the set $\mathbb{Q}^+$ of *positive* rational numbers, i.e. those rational numbers of the form m/n where m and n are both > 0, is countably infinite.

So that each element of $\mathbb{Q}^+$ will only occur once, we represent each rational number by m/n where m and n have been cancelled down as much as possible. That is, m and n have no common factors. Each element of $\mathbb{Q}^+$ has *exactly one* representation of this form.

In a row list all positive rational numbers which are of the form $m/1$, i.e. list all positive integers.
In the next row list all positive rational numbers which are of the form $m/2$ after cancelling down.
In the next row list all positive rational numbers which are of the form $m/3$ after cancelling down.

In the next row list all positive rational numbers which are of the form $m/4$ after cancelling down.
Etc.

See the following diagram and at this stage ignore the arrows.

$$\begin{array}{ccccccccccc}
1/1 & & 2/1 & \to & 3/1 & & 4/1 & \to & 5/1 & & \dots \\
\downarrow & & \uparrow & & \downarrow & & \uparrow & & \downarrow & & \\
1/2 & \to & 3/2 & & 5/2 & & 7/2 & & 9/2 & & \dots \\
 & & & & \downarrow & & \uparrow & & \downarrow & & \\
1/3 & \leftarrow & 2/3 & \leftarrow & 4/3 & & 5/3 & & 7/3 & & \dots \\
\downarrow & & & & & & \uparrow & & \downarrow & & \\
1/4 & \to & 3/4 & \to & 5/4 & \to & 7/4 & & 9/4 & & \dots \\
 & & & & & & & & \downarrow & & \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \ddots
\end{array}$$

Each element of $\mathbb{Q}^+$ occurs exactly once in the above infinite array, but so far $\mathbb{Q}^+$ is not listed in a single sequence. However, by following the arrows as indicated we do obtain a sequence in which each number in $\mathbb{Q}^+$ occurs exactly once. The first few terms are:

$$1, \frac{1}{2}, \frac{3}{2}, 2, 3, \frac{5}{2}, \frac{4}{3}, \frac{2}{3}, \frac{1}{3}, \frac{1}{4}, \frac{3}{4}, \frac{5}{4}, \frac{7}{4}, \frac{5}{3}, \frac{7}{2}, 4, 5, \frac{9}{2}, \frac{7}{3}, \frac{9}{4}, \dots .$$

So $\mathbb{Q}^+$ is countably infinite.

We abbreviate this sequence for $\mathbb{Q}^+$ to:[8]

$$r_1, r_2, r_3, r_4, \dots, r_n, \dots .$$

The set of negative rational numbers can then be written as

$$-r_1, -r_2, -r_3, -r_4, \dots, -r_n, \dots .$$

Finally, the set $\mathbb{Q}$ of all rational numbers can be written as a sequence by following the arrows in the following:

$$\begin{array}{ccccccc}
0 & & & & & & \\
\downarrow & & & & & & \\
r_1 & & r_2 & & r_3 & & r_4 \quad \cdots \\
\downarrow \nearrow & & \downarrow \nearrow & & \downarrow \nearrow & & \downarrow \nearrow \cdots \\
-r_1 & & -r_2 & & -r_3 & & -r_4 \quad \cdots
\end{array}$$

This gives

$$0, r_1, -r_1, r_2, -r_2, r_3, -r_3, r_4, -r_4, \dots ,$$

[8] You may object that there is no obvious formula for r_n. In fact one could write a complicated formula, but this does not really matter. The point is that we have demonstrated that there is a sequence that contains each positive rational number exactly once.

In this example it would not be too hard to write a computer program which would actually generate as many terms of the above sequence as we wished.

In some even more complicated situations it may be possible to prove the *existence* of a sequence which lists every member of a certain set exactly once. But the proof may be an argument by contradiction, and we may not actually have a nice way of generating the sequence.

i.e.

$$0, 1, -1, \frac{1}{2}, -\frac{1}{2}, \frac{3}{2}, -\frac{3}{2}, 2, -2, 3, -3, \frac{5}{2}, -\frac{5}{2}, \frac{4}{3}, -\frac{4}{3}, \dots .$$

(One could even write a computer program to do this.) □

New Infinite Sets from Old

This material is not in [HM].

⋆Subsets of Countably Infinite Sets Suppose we have a set which is countably infinite. What can we say about the possible cardinalities of its subsets?

For example, if

$$A = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20, \dots\}$$

then one subset is

$$B = \{4, 22, 24, 36, 52, 54, 58, 102, \dots\}$$

where the sequence listing B continues forever, and another is

$$C = \{4, 16, 22, 484\}$$

where the sequence listing C is finite.

The following theorem says that this is essentially all that can happen.

Theorem 3.2.5. *Suppose a set is countably infinite. Then any subset is either finite or countably infinite.*

Suppose a set is finite with cardinality n. Then any subset is finite with cardinality $k \leq n$.

Proof. Let A be a countably infinite set. Suppose B is a subset of A.

We can write

$$A = \{a_1, a_2, a_3, \dots, a_n, \dots\}. \tag{3.3}$$

Use the enumeration in (3.3) to consider each element of A in turn. If the element is in B then write it down in a new list (i.e. sequence). Three things can happen. Either

1. the process will never start because $B = \emptyset$, or
2. the process will stop after a finite number of steps (say k), or
3. the process will never stop.

In the first case the set B is certainly finite.

In the second case we obtain a finite listing of the elements of B, which we write as

$$B = \{b_1, b_2, \dots, b_k\}.$$

This implies that B is finite with cardinality k.

In the third case we obtain an infinite sequence listing the elements of B, which we can write as

$$B = \{b_1, b_2, \dots, b_n, \dots\}.$$

This implies B is countably infinite.

If A is finite with cardinality n, then B is either the same as A or is obtained by removing some elements from A. It follows that B will have cardinality k where $k \leq n$. □

⋆Unions of Two Countably Infinite Sets In the proofs that $\mathbb{Z}$ and $\mathbb{Q}$ are countably infinite, one of the steps was to show that if certain sets A and B are countably infinite then so is $A \cup B$. The argument we used works more generally.

Theorem 3.2.6. *Suppose at least one of the sets A and B is countably infinite and the other is either finite or countably infinite. Then $A \cup B$ is countably infinite.*

Proof. Because A is countably infinite we can write

$$A = \{a_1, a_2, \dots, a_n, \dots\}.$$

Because B is finite or countably infinite we can write

$$B = \{b_1, b_2, \dots, b_k\}$$

for some k, or

$$B = \{b_1, b_2, \dots, b_n, \dots\}.$$

A sequence which lists all members of $A \cup B$ exactly once is obtained by first considering the path which alternates between elements of A to B as in the following diagram.

$$\begin{array}{ccccccccccc} a_1 & & a_2 & & a_k & & a_{k+1} \to & a_{k+2} \to & a_{k+3} & \cdots \\ \downarrow & \nearrow & \downarrow & \cdots & \downarrow & \nearrow & & & & \\ b_1 & & b_2 & & b_k & & & & & \end{array} \quad \text{or} \quad \begin{array}{ccccccccc} a_1 & & a_2 & & a_3 & & a_4 & & \\ \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \nearrow & \downarrow & \cdots & . \\ b_1 & & b_2 & & b_3 & & b_4 & & \end{array}$$

Thus in case B has k elements the path is

$$a_1 \to b_1 \to a_2 \to b_2 \to \cdots \to a_k \to b_k \to a_{k+1} \to a_{k+2} \to a_{k+3} \dots,$$

and in case B is infinite the path is

$$a_1 \to b_1 \to a_2 \to b_2 \to a_3 \to b_3 \to a_4 \to b_4 \to \dots.$$

If we next strike out any elements that have already occurred, we obtain an infinite sequence listing each element of $A \cup B$ exactly once.

This shows $A \cup B$ is countably infinite. □

Suppose that the first few elements of A and B are given by

$$A = \{2, 13, 3, 5, 8, 18, \dots\}, \quad B = \{1, 5, 14, 3, 2, 6, \dots\}.$$

What is the path and what are the first few elements in the enumeration of $A \cup B$ obtained in the previous proof? ✍

⋆Comments and Further Results

Union of More Than Two Countable Sets Suppose we have three countable sets and at least one of them is infinite. Then it follows from applying Theorem 3.2.6 twice that the union of these three sets is countably infinite. *See Question 2.* ✍

Subsets of $\mathbb{Q}$ From Theorem 3.2.4 the set $\mathbb{Q}$ of rational numbers is countably infinite. We already knew that E, O and $\mathbb{Z}$ are countably infinite, but this also follows from Theorem 3.2.5.

Questions

1 Questions 6–44 on pp 156–161 of [HM] are all relevant. Questions 16–18 are an interesting way to explore some of the paradoxes of infinity.

2 Suppose A, B and C are three countable sets and A is countably infinite. Use the result of Theorem 3.2.6 twice to prove that $A \cup B \cup C$ is countably infinite.

Now prove that if fours sets are countable and at least one is countably infinite, then their union is countably infinite.

What happens for more than four sets?

3 Suppose that for each natural number n we have a set A_n which is countably infinite. It will be convenient to write

$$\begin{aligned} A_1 &= \{a_1^1, a_2^1, a_3^1, \ldots, a_k^1, \ldots\} \\ A_2 &= \{a_1^2, a_2^2, a_3^2, \ldots, a_k^2, \ldots\} \\ &\vdots \\ A_n &= \{a_1^n, a_2^n, a_3^n, \ldots, a_k^n, \ldots\} \\ &\vdots \end{aligned}$$

By tracing out a path as in the proof that the set of rationals is countably infinite, prove that the set A of all elements from $A_1, A_2, \ldots, A_n, \ldots$ is countably infinite.

To simplify matters, first do the case where no two of the A_n's have any element in common.

Then explain how to deal with the situation where some elements may occur in more than one of the A_n.

We write

$$A = A_1 \cup A_2 \cup \cdots \cup A_n \cup \ldots, \quad \text{or} \quad A = \bigcup_{n \geq 1} A_n,$$

and say "A is the union of the A_n for $n \geq 1$". The result can be stated as: *the union of a countably infinite collection of countably infinite sets is countably infinite.*

4 Why can the previous result NOT be proved by induction on n? What similar result can be proved by induction on n?

5 a Suppose one number is removed from $\mathbb{N}$. What is the cardinality of the remaining set? Explain briefly.

b Suppose a finite set of numbers is removed from $\mathbb{N}$. What is the cardinality of the remaining set? Explain briefly.

c Suppose an infinite set of numbers is removed from $\mathbb{N}$. Give examples where the remaining set has one element, has 23 elements, and has infinitely many elements respectively.

d Explain why similar results apply to *any* set of cardinality d.

3.3 Different Sizes of Infinity

Extending a new idea to its logical conclusion can lead to surprising and counterintuitive outcomes as well as a more accurate view of reality.

Overview

So far we have seen that the set $\mathbb{N}$ of natural numbers, the set $\mathbb{Z}$ of all integers, the set E of even numbers, and the set $\mathbb{Q}$ of rational numbers, all have the same cardinality d. We called these sets *countably infinite*.

We also saw that if two sets are countably infinite then so is their union. In Question 3 in the previous section we even saw that the union of a countably infinite collection of countably infinite sets is countably infinite.

It would be *reasonable* to suspect that "all infinite sets are the same size". In other words, it would be reasonable to suspect that all infinite sets have the same cardinality as the set $\mathbb{N}$ of natural numbers. Another way of expressing this would be to claim that every infinite set can be written as an infinite sequence, see Theorem 3.2.2.

However, we will see this is not true! We will see that the set $\mathbb{R}$ of all real numbers gives a larger infinity than the set $\mathbb{N}$ of natural numbers. Perhaps a more surprising way of expressing this is that: the set $\mathbb{R}$ of all real numbers gives a larger infinity than the set $\mathbb{Q}$ of rational numbers. More precisely, we will prove in Theorem 3.3.1 that $\mathbb{R}$ is not countably infinite. We say that $\mathbb{R}$ has cardinality c.

We will also show that the set $\mathbb{I}$ of irrational numbers is not countably infinite, and in fact has the same cardinality c as $\mathbb{R}$.

Reread the comments and commitment you made under **Sets with Equal Cardinality** on page 94. In this Section we want to show there is no one-to-one correspondence between $\mathbb{N}$ and $\mathbb{R}$. In other words we want to show there is *no* sequence which includes every real number.

The Dodge Ball Game

First look at [HM, Story 5, pp 8,16,21].

Diagonalising Out of a Sequence Suppose a friend writes down a list of 7 natural numbers, each 7 digits long. For example, the 7 numbers might be 5699785, 3407467, 4679068, 4235675, 3915609, 1377465, 8769689. We write

these as follows (at this stage ignore the boxes):

[5]	6	9	9	7	8	5
3	[4]	0	7	4	6	7
4	6	[7]	9	0	6	8
4	2	3	[5]	6	7	5
3	9	1	5	[6]	0	9
1	3	7	7	4	[6]	5
8	7	6	9	6	8	[9]

It is very easy to find another 7 digit number not in the list. We will denote this number by n. One systematic way to find such an n is to use the boxed digits as follows:

1. Take the first digit in n to be different from the first digit in the first number in the list. We can choose any digit other than 5. Let's choose 4.
2. Take the second digit in n to be different from the second digit in the second number in the list. We can choose any digit other than 4. Let's choose 3.
3. Take the third digit in n to be different from the third digit in the third number in the list. We can choose any digit other than 7. Let's choose 2.
4. Take the fourth digit in n to be different from the fourth digit in the fourth number in the list. We can choose any digit other than 5. Let's choose 4.
5. Take the fifth digit in n to be different from the fifth digit in the fifth number in the list. We can choose any digit other than 6. Let's choose 0.
6. Etc.

Proceeding in this way, we might end with $n = 4324012$, for example.

To summarise, we start with the top left digit and proceeding down the diagonal, construct an integer n as follows:

1. The first digit in n is selected so as to be different from the first digit in the diagonal.
2. The second digit in n is selected so as to be different from the second digit in the diagonal.
3. The third digit in n is selected so as to be different from the third digit in the diagonal.
4. The fourth digit in n is selected so as to be different from the fourth digit in the diagonal.
5. The fifth digit in n is selected so as to be different from the fifth digit in the diagonal.
6. Etc.

We could do the above in many ways. We might do the following. First decide that the only two digits we will use in constructing n are 4 and 6. Then follow the rules:

- If a diagonal digit is not 4, the corresponding digit in n will be 4.

- If a diagonal digit is 4, the corresponding digit in n will be 6.

Notice that:

1. After the first step we already know that n will be different from the first number in the list. This is because n differs in the first digit place from the first number in the list. It does not matter how we select the later digits in n.
2. After the second step we know that n will be different from the second number in the list, no matter how we select the other digits in n.
3. After the third step we know that n will be different from the third number in the list, no matter how we select the other digits in n.
4. Etc.

Extensions We do not have to use a sequence of 7 numbers with 7 digits. For example, the method would work on a sequence of one google natural numbers each containing one googol digits. I do not suggest you verify this by means of an example, but it is clear that by following the same method of working down the diagonal, we could show that there is a natural number with one googol digits which is not in the original sequence.

There are other ways of showing there is a natural number with one googol digits which is not in the original sequence. In fact, as we saw in (2.28) on page 55, there are $10^{\text{googol}-1}$ natural numbers with one googol digits. This is far bigger than one googol, which is how many numbers there were in the original sequence. So there will always be plenty of numbers with a googol digits which are not in the original sequence.

However, the "Dodge Ball" idea is what we will need to use in the following.

$\mathbb{R}$ *is not Countably Infinite*

[HM, 162–168]

Cantor's Diagonalisation Method We want to show there is *no* sequence which lists every real number. We will do this by using a variation of the Dodge Ball Game called *Cantor's Diagonalisation Method*.

It is very easy to write down an infinite sequence of real numbers that does not include every real number. For example,

$$0, .1, .2, \ldots, 1, -.1, -.2, \ldots, -1, 1.1, 1.2, \ldots, 2, -1.1, -1.2, \ldots, -2, \ldots.$$

Any sequence of real numbers we think of will always miss some real numbers, but perhaps that just means we did not think about it hard enough. After all, it was quite hard work to list all the rational numbers in a single sequence. Maybe there is also an ingenious way of listing all the real numbers, not just the rationals, in an infinite sequence.

In fact there is no such sequence. But how could we prove this? It is no good just showing that certain sequences of real numbers do not contain all real numbers. What we want is to show that for *every* sequence of real numbers, no matter how such a sequence might have been produced, there will always be some real number *not* in the sequence.

It is important to understand that, when we produce a real number not in a given sequence, this real number will depend on the sequence. The real number produced will vary from one sequence to another. Similarly, in the

game of Dodge Ball we produced a 7 digit number not in the original finite sequence. But if we had started with a different sequence then the Dodge Ball method would have usually produced a different number which is not in the finite sequence.

Theorem 3.3.1. *The set $\mathbb{R}$ is not countably infinite.*

Proof. We will show there is no sequence which includes all real numbers. By Theorem 3.2.2 this will imply that $\mathbb{R}$ is not countably infinite.

More precisely, we will show:

> If $s_1, s_2, \ldots, s_n, \ldots$ is a sequence of real numbers then there is another real number r which is not in the sequence.[9]

Once we have done this, the Theorem is proved. *Why?*

Using decimal expansions we first write the given sequence $s_1, s_2, \ldots, s_n, \ldots$ in the form:

$$\begin{aligned}
s_1 &= a_1.\boxed{a_{11}}a_{12}a_{13}a_{14}\ldots a_{1n}\ldots \\
s_2 &= a_2.a_{21}\boxed{a_{22}}a_{23}a_{24}\ldots a_{2n}\ldots \\
s_3 &= a_3.a_{31}a_{32}\boxed{a_{33}}a_{34}\ldots a_{3n}\ldots \\
s_4 &= a_4.a_{41}a_{42}a_{43}\boxed{a_{44}}\ldots a_{3n}\ldots \\
&\vdots \\
s_n &= a_n.a_{n1}\,a_{n2}\,a_{n3}a_{n4}\ldots\boxed{a_{nn}}\ldots \\
&\vdots
\end{aligned} \tag{3.4}$$

For example, if $s_1 = 17.325168432\ldots$ and $s_2 = -0.298461705\ldots$ then[10]

$$\begin{aligned}
&a_1 = 17, a_{11} = 3, a_{12} = 2, a_{13} = 5, a_{14} = 1, a_{15} = 6, \ldots, \\
&a_2 = -0, a_{21} = 2, a_{22} = 9, a_{23} = 8, a_{24} = 4, a_{25} = 6, \ldots\ .
\end{aligned}$$

We now show there is a real number r not in this sequence by extending the Dodge Ball method to infinite sequences. In fact the number r we obtain will always be between 0 and 1.

To do this we will use the digits 4 and 6 as on page 104. (You can use your own different choice of digits, but do not use 9 or 0 for a reason we will soon discuss.) Then we define

$$r = .r_1r_2r_3r_4\ldots r_n\ldots, \tag{3.5}$$

where:

1. If $a_{11} \neq 4$ then $r_1 = 4$ and if $a_{11} = 4$ then $r_1 = 6$.
2. If $a_{22} \neq 4$ then $r_2 = 4$ and if $a_{22} = 4$ then $r_2 = 6$.
3. If $a_{33} \neq 4$ then $r_3 = 4$ and if $a_{33} = 4$ then $r_3 = 6$.
4. If $a_{44} \neq 4$ then $r_4 = 4$ and if $a_{44} = 4$ then $r_4 = 6$.

[9] The method is discussed for a particular example in [HM, p165]. But the method there is quite general and here we write out the same method using a more general notation.

[10] Taking $a_2 = -0$ might look a bit odd. It probably helps to think of a_2 as part of the description of s_2 rather than as a number.

$\vdots$

n. If $a_{nn} \neq 4$ then $r_n = 4$ and if $a_n = 4$ then $r_n = 6$.

$\vdots$.

Some numbers have two decimal expansions, but this only happens when the decimal expansions end in an infinite sequence of 0s or an infinite sequence of 9's (see Theorem 2.7.2). This does not happen with the number r since its decimal expansion uses only 4s and 6s. So the decimal expansion (3.5) is the *only* decimal expansion for r.

It follows that:

1. Since the decimal expansion for r is different from the decimal expansion for s_1 in the first decimal place, $r \neq s_1$.
2. Since the decimal expansion for r is different from the decimal expansion for s_2 in the second decimal place, $r \neq s_2$.
3. Since the decimal expansion for r is different from the decimal expansion for s_3 in the third decimal place, $r \neq s_3$.
4. Since the decimal expansion for r is different from the decimal expansion for s_4 in the fourth decimal place, $r \neq s_4$.

$\vdots$

n. Since the decimal expansion for r is different from the decimal expansion for s_n in the nth decimal place, $r \neq s_n$.

$\vdots$

We have now shown that the number r is different from every real number in the sequence $s_1, s_2, s_3, \ldots, s_n, \ldots$.

This proves the statement in the second paragraph of the proof and so completes the proof of the Theorem. □

Comment on the Proof One student might say that the proof is flawed because we can always include the number r in the sequence $s_1, s_2, \ldots, s_n, \ldots$ by, for example, putting it at the beginning of the sequence and moving all other terms one place to the right. A second student might reply that the proof then applies to the new sequence and there will be another number r not in the new sequence.

But the objection of the first student is not valid. And while the second student has made a correct statement, the statement is not necessary in order to justify the proof, and perhaps even slightly misses the point of the proof.

The point of the proof is that the proof really does show that for *any* sequence of real numbers there is a real number r not in that sequence. It is similar to the Dodge Ball Method in this respect.

The fact that we could make a new sequence which *does* include r, and obtain from this new sequence yet another real number not in the new sequence, is true. And it is also helpful to our understanding. But this information about how we would deal with the new sequence is not necessary in order to justify the validity of the proof! The objection of the first student is not valid because we only need to show that r is not in the original sequence. Remember that the original sequence was any sequence of real numbers, it was completely arbitrary.

Uncountable Sets

This material is not in [HM].

Definition 3.3.2. A set is *uncountable* if it is not finite or countably infinite.

If a set can be put in one-to-one correspondence with $\mathbb{R}$ we say it has *cardinality c*.[11]

We say that c is a *cardinal* or a *cardinal number*.

So now we have the cardinalities n where n is 0 or any natural number, the cardinality d and the cardinality c. We will see in Section 3.4 that this is just the beginning of the story!

A Common Error Suppose A is an infinite set. Then it is not necessarily correct to say "let $A = \{a_1, a_2, \dots\}$". The reason is that this implicitly assumes that A is countably infinite!

Countable Subsets of Uncountable Sets Suppose we have a set A which is infinite. Think of A as being uncountable. One possible example of A is the set $\mathbb{R}$, but there are many others. Then A will always have a subset S (in fact many) which is countably infinite, in other words which has cardinality d, as we see in the next Theorem.

For this reason we say that *d is the smallest infinite cardinal number.*

Theorem 3.3.3. *Suppose A is an infinite set. Then there is a subset S of A which is countably infinite.*

Proof. Because A is not the empty set there is certainly an element $a_1 \in A$.

✍ Because $A \setminus \{a_1\}$ is also not the empty set (*why?*) there is an element $a_2 \in A \setminus \{a_1\}$.

✍ Because $A \setminus \{a_1, a_2\}$ is also not the empty set (*why?*) there is an element $a_3 \in A \setminus \{a_1, a_2\}$.

Etc.

In this way we choose[12] a countably infinite set $S = \{a_1, a_2, a_3, \dots, a_n, \dots\}$ which is a subset of A. □

⋆*Removing Part of an Infinite Set*

This material is not in [HM].

If we remove one or more elements from a finite set then we decrease the cardinality (size) of the set.

In Question 5 on page 102 we saw that if we remove a finite set from a countably infinite set the new set is still countably infinite. If we remove a countably infinite set from a countably infinite set the new set may be empty, finite, or countably infinite.

[11] The letter c comes from *continuum*, an old way of referring to the set $\mathbb{R}$.

[12] Since in general there is no rule or "constructive" way to do this, we are using what is called in mathematics the "Axiom of Choice". In fact, we are using the "countable" Axiom of Choice. Further discussion of this takes us deep into the Foundations of Mathematics. See "Foundations of Set Theory" and "A Cardinal between d and c?" on page 120.

The next Theorem gives particular examples, which are sufficient for our purposes, of the following more general result: *If from an infinite set B a subset A of smaller cardinality is removed, then the remaining set has the same cardinality as B.*

We prove that if one removes a finite subset A from an infinite set B then the remaining set $B \setminus A$ is still infinite, and its cardinality is the same as the cardinality of B. We also prove that if we remove a *countably* infinite subset A from an *uncountably* infinite set B then the remaining set $B \setminus A$ is uncountably infinite, and its cardinality is the same as the cardinality of B.

The reason the proof of the Theorem is tricky is that we *cannot* write the set B in an infinite sequence

$$b_1, b_2, \ldots, b_n, \ldots,$$

unless B is *countably* infinite. See "A Common Error" on page 108.

For example, if B is $\mathbb{R}$ then we cannot write B as a sequence. So instead, we choose an appropriate *countably* infinite subset $r_1, r_2, \ldots, r_n, \ldots$ from B and work with this subset.

In the Theorem, think of B as the set $\mathbb{R}$ of real numbers. The set A might be a finite set of numbers, or a countably infinite set such as $\mathbb{N}$ or $\mathbb{Q}$.

Theorem 3.3.4.

1. *Suppose B is an infinite set and A is a finite subset. Then the set $B \setminus A$ has the same cardinality as B.*
2. *Suppose B is an uncountably infinite set and A is a countably infinite subset. Then the set $B \setminus A$ has the same cardinality as B.*

Proof. We begin with the proof of Part 1.

To understand the ideas, first suppose A contains just the single element r_1. We write $A = \{r_1\}$.

The trick is to choose an infinite sequence $r_1, r_2, \ldots, r_n, \ldots$ of distinct elements from B which begins with r_1.

We then write

$$B = \{r_1, r_2, \ldots, r_n, \ldots \quad\} \cup \big(B \setminus \{r_1, r_2, \ldots, r_n, \ldots\}\big). \quad (3.6)$$

It follows

$$B \setminus A = B \setminus \{r_1\} = \{r_2, r_3, \ldots, r_{n+1}, \ldots\} \cup \big(B \setminus \{r_1, r_2, \ldots, r_n, \ldots\}\big). \quad (3.7)$$

We can now use (3.6) and (3.7) to define a one-to-one correspondence between B and $B \setminus A$ as follows:

$$\begin{array}{ll} r_1 \leftrightarrow r_2 & \\ r_2 \leftrightarrow r_3 & \\ \quad\vdots & \& \quad \text{every element in } B \setminus \{r_1, r_2, \ldots, r_n, \ldots\} \\ r_n \leftrightarrow r_{n+1} & \quad\ \text{corresponds to itself.} \\ \quad\vdots & \end{array}$$

This proves that B and $B \setminus A$ have the same cardinality.

Let us next suppose we remove a set $A = \{r_1, \ldots, r_{23}\}$, for example, from B. The proof that B and $B \setminus A$ have the same cardinality is similar to before.

Beginning with $r_1, \ldots, r_{23}$ choose an infinite sequence $r_1, \ldots, r_{23}, \ldots, r_n, \ldots$ of distinct elements from B.

We then write

$$B = \{r_1, \ldots, r_n, \ldots \quad\} \cup \big(B \setminus \{r_1, r_2, \ldots\}\big). \quad (3.8)$$

It follows

$$B \setminus A = B \setminus \{r_1, \ldots, r_{23}\} = \{r_{24}, \ldots, r_{n+23}, \ldots\} \cup \big(B \setminus \{r_1, r_2, \ldots\}\big). \quad (3.9)$$

Using (3.8) and (3.9) the one-to-one correspondence between B and $B \setminus A$ is

$$\begin{array}{l} r_1 \leftrightarrow r_{24} \\ r_2 \leftrightarrow r_{25} \\ \quad \vdots \\ r_n \leftrightarrow r_{n+23} \\ \quad \vdots \end{array} \quad \& \quad \begin{array}{l} \text{every element in } B \setminus \{r_1, r_2, \ldots\} \\ \text{corresponds to itself.} \end{array}$$

This proves that B and $B \setminus A$ have the same cardinality. A similar argument works if we remove any finite set of elements from B.

In order to prove part 2, suppose B is uncountable. Suppose a countable set $A = \{r_1, r_2, \ldots, r_n, \ldots\}$ of distinct elements is removed from B.

So as to give ourselves room to manoeuvre we choose another sequence $s_1, s_2, \ldots, s_n, \ldots$ of elements from B, distinct from each other and distinct from the r_n's. (Notice that if we had to stop choosing at some stage because there were no more elements left, then that would imply B is countably infinite, which is not the case.)

Now combine these two sequences into the single sequence:

$$r_1, s_1, r_2, s_2, \ldots, r_n, s_n, \ldots \ .$$

We write

$$B = \{r_1, s_1, r_2, s_2, \ldots\} \cup (B \setminus \{r_1, s_1, r_2, s_2, \ldots\}).$$

It follows

$$B \setminus A = B \setminus \{r_1, r_2, \ldots\} = \{s_1, s_2, s_3, s_4, \ldots\} \cup (B \setminus \{r_1, s_1, r_2, s_2, \ldots\}).$$

Now define a a one-to-one correspondence between B and $B \setminus A$ as follows:

$$\begin{array}{l} r_1 \leftrightarrow s_1 \\ s_1 \leftrightarrow s_2 \\ r_2 \leftrightarrow s_3 \\ s_2 \leftrightarrow s_4 \\ \quad \vdots \\ r_n \leftrightarrow s_{2n-1} \\ s_n \leftrightarrow s_{2n} \\ \quad \vdots \end{array} \quad \& \quad \begin{array}{l} \text{every element in } B \setminus \{r_1, s_1, r_2, s_2, \dots\} \\ \text{corresponds to itself.} \end{array}$$

This shows B and $B \setminus A$ have the same cardinality. □

⋆*The Set* $\mathbb{I}$ *of Irrationals.*

This material is not in [HM

Theorem 3.3.5. *The set* $\mathbb{I}$ *of irrational numbers is uncountable.*

Proof. Assume $\mathbb{I}$ is countable. (We will obtain a contradiction.)

Because $\mathbb{Q}$ is also countably infinite and $\mathbb{R} = \mathbb{Q} \cup \mathbb{I}$, it then follows from Theorem 3.2.6 that $\mathbb{R}$ is countably infinite.

But we know from Theorem 3.3.1 that $\mathbb{R}$ is not countably infinite.

Thus we have a contradiction and so our *assumption* is wrong. That is, $\mathbb{I}$ is uncountable. □

We can actually show something more than this. We will show that the set $\mathbb{I}$ has the same cardinality c as $\mathbb{R}$.

The trick is to apply Theorem 3.3.4. Since $\mathbb{I}$ is obtained by removing the countably infinite set $\mathbb{Q}$ from $\mathbb{R}$, it follows that $\mathbb{I}$ has cardinality c.

Theorem 3.3.6. *The set* $\mathbb{I}$ *and the set* $\mathbb{R}$ *have the same cardinality* c.

Proof. We know that $\mathbb{I} = \mathbb{R} \setminus \mathbb{Q}$. Since $\mathbb{R}$ is uncountable and $\mathbb{Q}$ is countable, it follows from Theorem 3.3.4 that $\mathbb{I}$ and $\mathbb{R}$ have the same cardinality, namely c. □

Questions

1 Questions 7, 8, 10, 13–24 on pp 169–172 of [HM] are good.
Notice that Questions 13, 14, 16 are variations on a similar idea.
How?

2 Show that if A is an infinite set and b is not in A then $A \cup \{b\}$, the set obtained by "adding" b to A, has the same cardinality as A.
HINT: This is easy by applying Theorem 3.3.4 to $A \cup \{b\}$.
Note that we do not really need to assume b is not in A. Why?

3 Show that if A is an infinite set and B is any countable set, then $A \cup B$ has the same cardinality as A.

HINT: First assume that B has no elements in common with A and apply Theorem 3.3.4.

How can you get the general result from this case?

3.4 An Infinite Hierarchy of Infinities

Use a powerful idea or technique repeatedly to discover more and more fascinating and surprising results.

Overview

[HM, 173,174]

You should read [HM, §3.4]. Although the material in Section 3.4 here is self contained and a little more advanced, [HM] has discussion and motivation for the ideas.

We now have a precise notion of what it means for two sets to have the same cardinality. We have used this to discover some very surprising results. For example, the sets $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ all have the same cardinal d. The sets $\mathbb{R}$ and $\mathbb{I}$ have the same cardinal c. The cardinal d is smaller than the cardinal c.

Here are more Questions, which we can now phrase in a precise manner.

1. We have seen that the cardinal d is smaller than the cardinal c.[13]
   Is there an infinity (i.e. a cardinal) between d and c? The assumption that there is no such cardinal is called the *Continuum Hypothesis*.
2. Is there an infinity greater than the cardinal c of the set $\mathbb{R}$?
3. Are there infinitely many different cardinals, i.e. sizes of infinity?
4. Is there a largest infinity?
5. Is there a set containing all sets?

We begin with Question 2 and show the answer is YES by extending Cantor's Diagonalisation Method used in the last Section.

These methods will allow us also to show that the answer to Question 3 is YES.

This helps us show that the answers to Questions 4 and 5 are NO.

The most profound is Question 1. The astounding answer is neither YES nor NO!

There is some further discussion and history concerning these Questions and related matters beginning on page 129.

The Power Set

[HM, 175]

So far we have usually discussed sets whose elements are numbers. Examples are $E, \mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{I}, \mathbb{R}$. But there is also the set of books in this room, the set of

[13] From Theorem 3.3.3 any set of cardinality c always contains a set of cardinality d. See also Definition 3.4.3.

people in this class with green hair (frequently, but perhaps not always, it is the empty set), the set of infinite sequences of integers which start with 1 and do not contain 7, etc., etc.

The point is that the elements of a set need not be numbers. The elements might be books, people, infinite sequences, etc. The elements of a set may even themselves be sets.

If A is a set, we will be particularly interested in the set whose elements are all the subsets of A. We call this the *power set* of A and denote it by $\mathcal{P}(A)$.

Definition 3.4.1. The *power set* $\mathcal{P}(A)$ of a set A is the set consisting of all the subsets of A.

Example Suppose $A = \{a, b, c, d\}$. (The elements of A might be $1, 2, 3, 4$, but they could be any four distinct objects.) Then the subsets of A are:

$$\begin{gathered} \emptyset \\ \{a\}, \{b\}, \{c\}, \{d\}, \\ \{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}, \\ \{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}, \\ \{a, b, c, d\} \end{gathered} \tag{3.10}$$

Remarks

1. We include both the empty set $\emptyset$ and the original set A as elements of $\mathcal{P}(A)$, i.e. as subsets of A.
2. Because the order of elements does not matter, we do not write $\{b, a\}$ as well as $\{a, b\}$. Both represent the *same* set. Similarly for other cases.
3. The set $\{a\}$ is not the same as the element a. For example, the set $\{a\}$ has exactly one element, namely a, and the cardinality of $\{a\}$ is one. But if a is itself a set then a may have cardinality 23, say. However, $\{a\}$ still has cardinality one!

Describing the Power Set One way to think of subsets of A is to think of the elements of A lined up in a row.

$$a \quad b \quad c \quad d$$

A subset is determined by those elements in the row that we "push forward" to the next row. For example

$$\begin{array}{cccc} a & & c & \\ & b & & d \end{array}$$

determines the set $S = \{b, d\}$,

$$\begin{array}{cccc} a & & & \\ & b & c & d \end{array}$$

determines the set $S = \{b, c, d\}$ and

$$\begin{array}{cccc} a & & c & d \\ & b & & \end{array}$$

determines the set $S = \{b\}$.

In describing an element $S \in \mathcal{P}(A)$, i.e. in describing a set $S \subset A$, we need to decide for each element of A whether or not to push it forward and so include it in S.

The Power Set of a Finite Set

[HM, 176–178]

The Number of Subsets Suppose A is the finite set $\{a, b, c, d\}$ as before.

In describing an element $S \in \mathcal{P}(A)$, i.e. in describing a set $S \subset A$, we have 2 choices (or possibilities) for a ($a \in S$ or $a \notin S$), 2 choices for b ($b \in S$ or $b \notin S$), 2 choices for c ($c \in S$ or $c \notin S$) and 2 choices for d ($d \in S$ or $d \notin S$).

For each of the 2 choices for a we have 2 choices for b, making a total of $2 \times 2 = 4$ choices for a and b. For each of these 4 choices for dealing with a and b, there are 2 choices for c, making $4 \times 2 = 8$ choices for what to do with a, b and c. Finally , for each of these 8 choices there are 2 choices for d, leading to $8 \times 2 = 16$ choices as to what to do with a, b, c and d.

If you go back and count the number of subsets in (3.10) you will indeed get 16.

This leads to a Theorem.

Theorem 3.4.2. *If A is a finite set with n elements, then $\mathcal{P}(A)$ has 2^n elements. In other words, there are 2^n subsets of A.*

Proof. We can write A in the form

$$A = \{a_1, a_2, a_3, \ldots, a_n\}.$$

A subset S of A can be described by assigning to each element of A either Y (for YES, the element is in S) or N (for NO, the element is not in S).

There are 2 possibilities for a_1, 2 possibilities for a_2, 2 possibilities for a_3, $\ldots$, 2 possibilities for a_n.

The total number of possibilities is obtained by multiplying (*not* by adding, *why?*), and this gives $2 \times 2 \times 2 \times \cdots \times 2 = 2^n$, since there are n factors. □

More Subsets than Elements Suppose A is a finite set with n elements. Since $2^n > n$, the cardinality of the power set of A is certainly larger than the cardinality of A.

But what we need later is a method *which will extend to infinite sets* for showing the cardinality of the power set of A is larger than the cardinality of A.

For this, suppose we have two columns, each a list of the same length. The first column is a listing of all the elements of A. In the second column there are certain subsets of A, one corresponding to each element of A in the first column. For example, we might have $A = \{a_1, a_2, \ldots, a_7\}$ and the following two lists:

All elements of A	Certain Subsets of A
a_1	$\{a_1, a_4\}$
a_2	$\{a_3, a_5, a_6\}$
a_3	$\{a_2, a_4, a_6, a_7\}$
a_4	$\emptyset$
a_5	$\{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$
a_6	$\{a_6\}$
a_7	$\{a_4\}$

The goal is to obtain, in a systematic manner, a subset of A which is not included in the list on the right side.

We will call this set M as in [HM]. (The M stands for "mysterious".)

We want M to be different from the set corresponding to a_1, different from the set corresponding to a_2, different from the set corresponding to a_3, ..., and finally different from the set corresponding to a_n.

The Dodge Ball Game Revisited We want M to be a subset of A which is different from every set in the second column.

First, M should be different from the set paired with a_1, which is the set $\{a_1, a_4\}$. Moreover, it would be nice if we could do this just on the basis of whether or not to put a_1 in M. But this is easy. Since a_1 is in $\{a_1, a_4\}$ we decide to *not* put a_1 in M. So M is different from $\{a_1, a_4\}$, no matter what we do with the other elements $a_2, a_3, \ldots, a_7$ from A.

Similarly, M should be different form the set paired with a_2, which is the set $\{a_3, a_5, a_6\}$. Moreover, it would be nice if we could do this just on the basis of whether or not to put a_2 is in M. But this is again easy. Since a_2 is not in $\{a_3, a_5, a_6\}$ we decide to put a_2 in M.

Similarly, since a_3 is not in $\{a_2, a_4, a_6, a_7\}$ we decide to put a_3 in M.

Since a_4 is not in $\emptyset$ we decide to put a_4 in M.

Since a_5 is in $\{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$ we decide to *not* put a_5 in M.

Since a_6 is in $\{a_6\}$ we decide to *not* put a_6 in M.

Since a_7 is not in $\{a_4\}$ we decide to put a_7 in M.

So the final result is that we take $M = \{a_2, a_3, a_4, a_7\}$.

Observations on the Method

1. For the element a_3 in A (and similarly for every other element in A), the decision whether or not to put a_3 in M is based solely on whether or not a_3 is in the set $\{a_2, a_4, a_6, a_7\}$ with which it is paired.

   Once we have made the decision to put or not put a_3 in M, no matter what else we decide for the other elements in A, the set M will be different from the set in the list corresponding to a_3.
2. The decision we make for a_3 (and similarly for every other element in A) is independent of the decision we make for the other elements in A.
3. We can make our decisions in any order. We could make all our decisions simultaneously, at least in principle.
4. We have shown there is no one-to-one correspondence between A and $\mathcal{P}(A)$. We have shown that for any way of pairing up every element in A with a subset of A, there will always be some set M which is not paired with any element in A.

The method here will work also when A is infinite. In fact, it will enable us to *prove* the cardinality of $\mathcal{P}(A)$ is larger than the cardinality of A even when A is infinite.

The Power Set of an Infinite Set

[HM, 179–182]

What We Will Prove We know the cardinality of the power set of a finite set A is larger than the cardinality of A. But if A is infinite, particularly if A has a large infinite cardinality such as c, there is much more room to "manoeuvre". Maybe there *is* a one-to-one correspondence between A and $\mathcal{P}(A)$. But we will see this is not so.

We will show *there is no one-to-one correspondence between A and $\mathcal{P}(A)$*. An equivalent way of expressing this is that:

$$\begin{gathered}\text{if every element of } A \text{ is paired with some subset} \\ \text{of } A, \text{ then another subset of } A \text{ will be left unpaired.}\end{gathered} \tag{3.11}$$

Parallels with the Finite Case We will prove (3.11) by using the Dodge Ball Method, or more precisely Cantor's Diagonal Argument, as in the finite case with the Table on page 115.

That is, we will show (3.11) by showing that for any pairing in which every element of A is paired with some subset of A, there is a subset M of A which is "left over" in the pairing process.

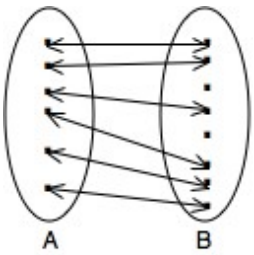
The only difference in the argument from the finite case is that the set A cannot be listed in a column, not even an infinite column if A is uncountable (see Theorem 3.3). But this does not make any *essential* difference to the proof.

We will also need the following Definition.

Definition 3.4.3. The cardinality of a set B is *larger* than the cardinality of a set A if:

- There is a way of pairing every element of A with a different element in B. (This is the same as saying that there is a one-to-one correspondence between A and a subset of B.)
- There is no way of pairing every element of A with a different element of B so that there are no elements of B left unpaired. That is, there is no one-to-one correspondence between A and B.

We also say that the cardinality of A is *smaller* than the cardinality of B.



While reading the proof of the next Theorem you should refer back to the Table on page 115 and also the discussion of the Dodge Ball Game on page 116.

In the Theorem we are particularly interested in the case A is infinite.

Theorem 3.4.4. *Suppose A is set. If every element of A is paired up with some subset of A, there will always be some subset M of A which is not paired up with any element from A.*

It follows that there is no one-to-one correspondence between A and $\mathcal{P}(A)$, and that the cardinality of $\mathcal{P}(A)$ is larger than the cardinality of A.

Proof. First note that there are many ways of pairing every element of A with a different element of $\mathcal{P}(A)$. For example, pair each element a with the subset $\{a\}$.

Now consider *any* pairing in which every element of A is paired with some subset of A. We will show there is always a "left over" subset of A (which we call M) which is not paired with any element from A.

For each a in A denote the corresponding subset paired with a by A_a.[14]

We now construct the "mysterious" set M, which is a subset of A, as follows:

For each a in A, if a is in A_a then do *not* put a in M. If a is not in A_a then *do* put a in M.

This ensures M is not the same as the set A_a for every a in A.

So we have proved the claim in the first paragraph of the Theorem.

It follows that there is no one-to-one correspondence between A and $\mathcal{P}(A)$.

It now follows from Definition 3.4.3 that the cardinality of $\mathcal{P}(A)$ is larger than the cardinality of A. □

An Infinity of Infinities

There is some discussion of this in [HM, 179].

⋆Aleph and All That Now things are going to get really weird.

Let's begin with the set $\mathbb{N}$. It has cardinality $d = \aleph_0$.

From Theorem 3.4.4 the set $\mathcal{P}(\mathbb{N})$ has a larger cardinality than $\mathbb{N}$. In fact the cardinality of $\mathcal{P}(\mathbb{N})$ is c, as we will see in Question 2 on page 121.

The Continuum Hypothesis says that there is no cardinal between d and c. So c is the next cardinal after $\aleph_0$ and is referred to as $\aleph_1$.

But now if we apply Theorem 3.4.4 to $\mathcal{P}(\mathbb{N})$ we get a set $\mathcal{P}(\mathcal{P}(\mathbb{N}))$ with an even larger cardinal number. The Generalised Continuum Hypothesis[15] says there is no cardinal between this cardinal and $\aleph_1$. We then write this cardinal as $\aleph_2$.

If we then apply Theorem 3.4.4 to $\mathcal{P}(\mathcal{P}(\mathbb{N}))$ we get a set $\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))$ with an even larger cardinality again. Again assuming the generalised Continuum Hypothesis we write this cardinal as $\aleph_3$.

Etc., etc.

[14] For example, in the case on page 115 where A is finite, $A_{a_1} = \{a_1, a_4\}$, $A_{a_2} = \{a_3, a_5, a_6\}$, $A_{a_3} = \{a_2, a_4, a_6, a_7\}$, etc.

[15] The basic idea here of an infinite hierarchy of cardinals still holds, even if we do not assume the Continuum Hypothesis and the Generalised Continuum Hypothesis. But the notation is a bit simpler if we do.

In this way we get a sequence of sets

$$\mathbb{N},\ \mathcal{P}(\mathbb{N}),\ \mathcal{P}(\mathcal{P}(\mathbb{N})),\ \mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N}))),\ \ldots, \tag{3.12}$$

with larger and larger cardinals

$$\aleph_0 < \aleph_1 < \aleph_2 < \aleph_3 < \cdots < \aleph_n < \ldots.$$

But we can even go further than this. Imagine the set which contains all the elements in $\mathbb{N}$, all the elements in $\mathcal{P}(\mathbb{N})$, all the elements in $\mathcal{P}(\mathcal{P}(\mathbb{N}))$, all the elements in $\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))$ etc.

What is the cardinality of this monster set? It is in fact larger than $\aleph_n$ for every natural number n. This is because if we take any n then $\aleph_n < \aleph_{n+1}$ and the cardinality of the monster set is at least as large as $\aleph_{n+1}$! It is standard to denote the cardinality of the monster set by $\aleph_\omega$. So now we have

$$\aleph_0 < \aleph_1 < \aleph_2 < \aleph_3 < \cdots < \aleph_n < \cdots < \aleph_\omega.$$

The letter ω is called "omega" and is the last letter of the Greek alphabet.

We are on a roll here, so let's keep going. Taking the power set of the monster set, and the power set of that, and so on, we get more and more cardinals usually written as follows:

$$\begin{aligned}&\aleph_0 < \aleph_1 < \aleph_2 < \aleph_3 < \cdots < \aleph_n < \ldots\\ &< \aleph_\omega < \aleph_{\omega+1} < \aleph_{\omega+2} < \aleph_{\omega+3} < \cdots < \aleph_{\omega+n} < \ldots.\end{aligned}$$

Why stop here? Consider the double monster set which consists of all elements in all sets obtained so far. It has a cardinality larger than all sets so far, and its cardinal is usually written $\aleph_{\omega\cdot 2}$. And onwards and onwards. And we have barely begun.

Do We Need This? From a philosophical and psychological point of view it is amazing how the human mind can begin to grasp such extraordinary concepts through a precise mathematical analysis.

From the practical and applications point of view it is important. Although the sets we deal with in mathematics will usually belong to at most the fifth or so level in (3.12), the fact that we can continue is essential to developing the theory required for applications. To develop the theory we need to know we can take the power set without any prior restriction on how often, and so we need to know we can continue up through the monster and the double monster and so on.

Set Theory Paradoxes

HM, 183–185]

A Set of All Sets? Is there a set which contains all sets? The answer is NO.

One way to see this is to suppose there were such a set, which we will call the universal set U. By Theorem 3.4.4 $\mathcal{P}(U)$ has a larger cardinality than U and so must contain some sets not in U. This contradicts the fact U contains all sets.

This is at first odd. It seems reasonable that we should at least be able to talk about the collection of all sets. But if we do so, this collection cannot be treated like an ordinary set.

A Largest Infinity? Is there a largest cardinal, i.e. a largest infinity? Again the answer is NO.

For suppose there were such a cardinal. Let us denote it by κ. (It is traditional to denote large cardinals by κ, which is pronounced "kappa" and is another letter of the Greek alphabet.) Then if A is a set whose cardinal is κ, we get a larger cardinal by taking the cardinality of the set $\mathcal{P}(A)$.

⋆Russell's Paradox Does every property define a set?

Suppose "$P(x)$" is an abbreviation for the statement "x has a certain property P". For example: "x is a real number", "x is a pink elephant", "x is a set of sets of sets", etc. It seems reasonable that there should always be a set S consisting precisely of those objects x with the property P, even though S might be the empty set.

In fact, we have seen that this is not the case. For example, if $P(x)$ says "x is a set" then we have seen there is no set of all sets and so there is no set whose elements are precisely those objects x satisfying the property $P(x)$.

The fact that for some properties there is not a corresponding set was first realised and understood by Bertrand Russell in 1902. Russell's example is quite simple and does not use the idea of cardinality. He considered the property P given by[16] [17]

$$P(A) \text{ iff } (A \text{ is a set and } A \notin A).$$

Assume there is a set S consisting precisely of those sets A such that $P(A)$ is true. In other words, assume that S is the set of all sets A which are not members of themselves.

We ask ourselves if $S \in S$ or $S \notin S$? Either way we get a contradiction:

- If $S \in S$ then $P(S)$ is true, i.e. $S \notin S$.
- If $S \notin S$ then $P(S)$ is not true, and since S is a set this implies $S \in S$.

So the conclusion we are forced to draw is that the assumption there is such a set S is in fact not correct![18]

⋆Foundations of Set Theory In the years 1900–1930 there were many attempts to put the foundations of set theory on a firm basis.

Russell and Whitehead wrote "Principia Mathematica" and developed what is known as the *Theory of Types*. This work attempts to reduce the foundations of mathematics to logic and was extremely influential. However, it is very unwieldy. It took 500 pages to establish $1 + 1 = 2$, but went considerably further than this!

The approach now used most frequently is due to Zermelo and Fraenkel and is called Zermelo-Fraenkel set theory. Another approach is due to Gödel, Bernays and Von Neumann. It allows for both sets and "classes". There is a class of all sets but not a class of all classes.

[16] By "$x \notin S$" we mean "x is *not* an element of S", and we read it as "x is not in S".

[17] It is in fact not easy to find an example of a set that *is* a member of itself. One example is the set of strange ideas. The set of strange ideas is a strange idea, and so is a member of itself!

For a more "mathematical" example of a set which *is* a member of itself one could try the set S of all sets with more than two elements. In fact, S is not a set for reasons similar to those which showed that there is no set of all sets.

[18] See the History concerning Russell's Paradox and the impact on Frege's work mentioned on page 130.

A Cardinal Between d and c? The statement that there is no cardinal between d and c is called the "Continuum Hypothesis". Gödel showed in 1940 that the Axiom of Choice[19] and the Continuum Hypothesis cannot be proved to be false using the other axioms of set theory. In 1963 Paul Cohen proved that the Axiom of Choice and the Continuum Hypothesis cannot be proved to be true using the other axioms of set theory, for which he received a Fields medal. These results together show the Axiom of Choice and the Continuum Hypothesis are independent of the other axioms of set theory.

There is no agreement on whether or not we should accept the Continuum Hypothesis. Either way leads to counterintuitive conclusions. While most mathematicians accept the Axiom of Choice, it also has some very surprising consequences.

Gödel also showed in 1934 that there is no set of axioms which will capture all mathematical truths.

This is all quite mind boggling. It means that we can prove that we cannot prove certain things. It implies that in mathematics there is no absolute truth. The consequences have been far reaching — in the philosophy and foundations of mathematics. The ideas involved have major implications to fields such as computational science, automata theory and artificial intelligence.

Questions

1 Questions 1, 3, 8, 9, 10, 13–22 on pp 185–189 of [HM] are good.

2 Prove that the cardinality of $\mathcal{P}(\mathbb{N})$ is c as follows:

1. Show there is a one-to-one correspondence between $\mathcal{P}(\mathbb{N})$ and the set of all infinite sequences of 0's and 1's.
2. Show that every point in the interval $[0, 1]$ has a binary expansion $\cdot a_1 a_2 a_3 \ldots$, where each a_n is either 0 or 1.
3. Explain why, in a manner analogous to that for decimal expansions, every number in $[0, 1]$ has either one or two binary expansions.
   What are the numbers with two binary expansions?
   Why is this set countable?
4. Deduce that the set of (different) binary expansions has cardinality c.
   Deduce that $\mathcal{P}(\mathbb{N})$ has cardinality c.

[19] See Footnote 12 on page 108.

3.5 Geometry and Infinity

Geometry can illuminate and geometry can mislead.

Overview

We will use geometric arguments to show that all line segments, both bounded and unbounded, have the same cardinality c.

We will also show that the plane $\mathbb{R}^2$ has the same cardinality as the set $\mathbb{R}$, despite the fact that it has an extra dimension. In fact the set $\mathbb{R}^3$ of points in space also has cardinality c, see Question 8 page 131.

All Line Segments are the Same Size

[HM, 190–195]

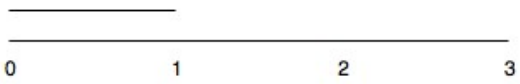
A line segment is just an interval where the first endpoint is strictly less than the second. So we do not allow the interval $[2, 2]$, which is not very interesting since it just contains the number 2, as a line segment. A line segment may or may not contain endpoints. It may be unbounded in one or both directions.

See the Notation for intervals on page 76. The words "line segment" are used here to emphasise the geometric aspect.

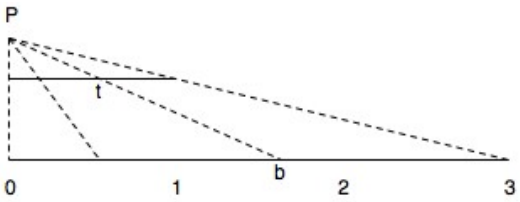
Closed Bounded Line Segments We will start with the two line segments below. The first consists of all real numbers between 0 and 1, including both 0 and 1, and is written $[0, 1]$. The second interval is $[0, 3]$.

Such intervals are said to be *closed* and *bounded*. They are"closed" because they include their endpoints and they are "bounded" because they do not go on forever in either direction.

Because the second interval is three times as long as the first we might think it has a larger cardinality.



However, the following diagram shows that in fact there is a one-to-one correspondence between $[0, 1]$ and $[0, 3]$.



Align P and the two line segments as shown. Then each line through P which intersects the top segment will also intersect the bottom segment. In this way, each point t on the top line segment $[0,1]$ is paired with exactly one point b on the bottom line segment $[0,3]$, there are no points left unpaired, and so there is a one-to-one correspondence between $[0,1]$ and $[0,3]$.

In a similar way any two line segments of the form $[a,b]$ and $[c,d]$ will have the same size, no matter how short the first and how long the second.

We can also give a formula for the one-to-one correspondence between $[0,1]$ and $[0,3]$. It is given by the correspondence $x \leftrightarrow 3x$.

What is the formula for a one-to-one correspondence between $[a,b]$ and $[c,d]$?

Theorem 3.5.1. *Any two closed bounded line segments $[a,b]$ and $[c,d]$ have the same size.*

Proof. This is proved either by a geometric argument as discussed above, or by giving a formula as discussed above. □

⋆Other Bounded Line Segments Consider the line segment $(0,1]$. It is obtained by removing the single point (or number) 0 from $[0,1]$. Since $[0,1]$ is an infinite set (*why?*), it follows from Theorem 3.3.4 that $(0,1]$ has the same cardinality as $[0,1]$.

However, the one-to-one correspondence between $(0,1]$ and $[0,1]$ we obtain from the proof of Theorem 3.3.4 is *not* a nice geometric one.

Theorem 3.5.2. *The line segments $[a,b]$, $(a,b]$, $[a,b)$, (a,b) all have the same cardinality.*

Proof. The last three intervals are obtained by removing one or two points from the interval $[a,b]$. It follows from Theorem 3.3.4 that they all have the same cardinality as $[a,b]$. □

The next theorem includes Theorem 3.5.2

Theorem 3.5.3. *Any two bounded line segments, whether or not they contain one or both of their endpoints, have the same cardinality.*

Proof. The line segments $[a,b]$, $(a,b]$, $[a,b)$, (a,b) all have the same cardinality by the previous Theorem. Similarly, the line segments $[c,d]$, $(c,d]$, $[c,d)$, (c,d) all have the same cardinality.

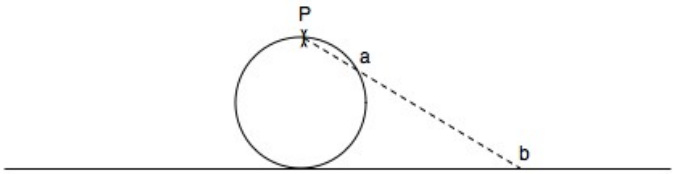
But the line segments $[a,b]$ and the line segments $[c,d]$ have the same cardinality by Theorem 3.5.1.

It follows from Theorem 3.1.3 that any bounded line segments have the same cardinality. *Why?*

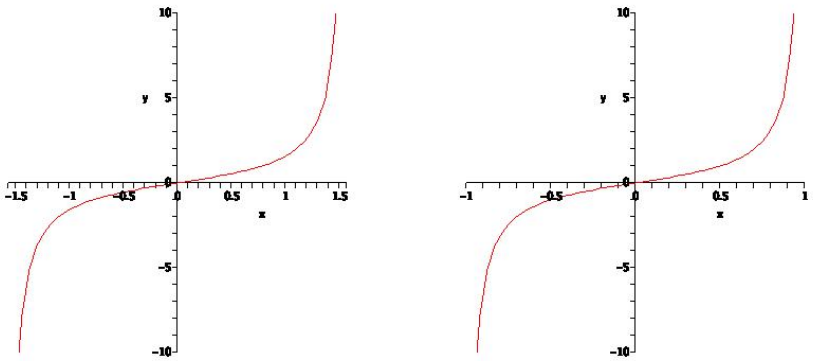
□

Unbounded Line Segments The most important of these is $\mathbb{R}$ itself.

There is a nice geometric way, stereographic projection, to get a one-to-one correspondence between the line segment $(-1, 1)$ (or any line segment of the type (a, b)) and $\mathbb{R}$. The idea is to wrap $(-1, 1)$ up into a circle with one point P missing. Then use P to stereographically project each point a on $(-1, 1)$ onto a point b on $\mathbb{R}$. See the following diagram.



Another way to get a one-to-one correspondence between the line segment $(-1, 1)$ and $\mathbb{R}$ is as follows. The graph of $\tan x$ gives a one-to-one correspondence between $(-\pi/2, \pi/2)$ and $\mathbb{R}$. So the graph of $\tan \pi x/2$ gives a one-to-one correspondence between $(-1, 1)$ and $\mathbb{R}$. *Why?*



Graphs of $y = \tan x$ and $y = \tan(\pi x/2)$.

The following Theorem generalises Theorems 3.5.2 and 3.5.3.

Theorem 3.5.4. *The cardinality of any bounded line segment is c.*

Proof. There is a one-to-one correspondence between $(-1, 1)$ and $\mathbb{R}$ by using sterographic projection as above, so $(-1, 1)$ has cardinality c.

But all bounded line segments have the same cardinality as $(-1, 1)$ by Theorem 3.5.3.

So any bounded line segment has cardinality c. □

We have not discussed unbounded line segments which go arbitrarily far in just one direction. These are line segments of the form (a, ∞), $[a, \infty)$, $(-\infty, b)$ and $(-\infty, b]$.

For example, (a, ∞) is the set of all numbers greater than or equal to a. *What are the other three?*

Note that the symbols "$-\infty$" and "∞" *DO NOT* denote numbers. They are used here in the context "(a, ∞)" etc. as a convenient shorthand way to represent certain line segments.

We will see in Question 6 that all unbounded line segments have cardinality c.

This together with Theorem 3.5.4 gives the following important result.

All line segments have cardinality c. (3.13)

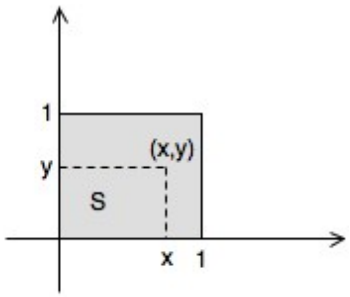
Sets in the Plane

[HM, 196–200]

The set of points in the plane is denoted by $\mathbb{R}^2$ and consists of all pairs of real numbers (x, y). It would seem reasonable to think that $\mathbb{R}^2$ has a larger cardinality than $\mathbb{R}$. After all, $\mathbb{R}^2$ has two dimensions, one more than $\mathbb{R}$. And we know there are certainly sets of larger cardinality than $\mathbb{R}$, such as the set $\mathcal{P}(\mathbb{R})$ of all subsets of $\mathbb{R}$.

However, it turns out that $\mathbb{R}$, $\mathbb{R}^2$ and even $\mathbb{R}^3$ (the set of points in space) all have cardinality c. In fact for any natural number n, the set $\mathbb{R}^n$ of all n-tuples $(x_1, \ldots, x_n)$ of real numbers, has cardinality c.

The Unit Square We will first see that the set of points in the "unit square" S, i.e. the set of points with coordinates (x, y) where $0 < x < 1$ and $0 < y < 1$, has cardinality c.



Theorem 3.5.5. *The cardinality of the unit square S, i.e. of the set of points* (x, y) *such that* $0 < x < 1$ *and* $0 < y < 1$*, has cardinality c.*

"Proof". [20]

For each point (x, y) in S consider the infinite decimal expansions:

$$x = \cdot x_1 x_2 x_3 x_4 x_5 x_6 \ldots, \quad y = \cdot y_1 y_2 y_3 y_4 y_5 y_6 \ldots .$$

[20] *Caveat*: We will cheat a little (hence the inverted commas around "Proof") and use the Cantor-Schroeder-Bernstein theorem on page 128), which we do not actually prove! This theorem is not surprising, although the proof is subtle. You do not need to read ahead — we will discuss it in the following.

Recall that some numbers have two decimal expansions. This happens if the number can be expressed with an infinite tail of 0's, in which case it can also be expressed with an infinite tail of 9's. For example, $.3700000\cdots = .3699999\ldots$. *We will always use the expansion with an infinite tail of 9's in such cases.*

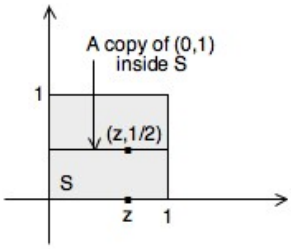
Next combine the decimal expansions of these two numbers x and y to obtain a third number

$$z = .x_1y_1x_2y_2x_3y_3x_4y_4\ldots\ .$$

Because x and y lie strictly between 0 and 1 it follows that z also lies strictly between 0 and 1. *Try a few examples.*

Different points (x, y) and (x', y') will give different *decimal expansions* for z and z'. Moreover, the decimal expansions for z and z' will never end in an infinite string of 0's because none of the decimal expansion for x, y, x', y' end in an infinite string of 0's. It follows that the *numbers* z and z' are also different.[21]

In other words, every point (x, y) in S corresponds to exactly one point z in the line segment $(0, 1)$, and different points in S correspond to different points in $(0, 1)$. For this reason we say that *the cardinality of S is less than or equal to the cardinality of* $(0, 1)$.



On the other hand, we can put a copy of the line segment $(0, 1)$ inside S, by letting $z \in (0, 1)$ correspond to $(z, 1/2) \in S$ for example. In this way, every point z in $(0, 1)$ corresponds to exactly one point in S, and different points in $(0, 1)$ correspond to different points in S. For this reason we say that *the cardinality of* $(0, 1)$ *is less than or equal to the cardinality of S.*

It follows from the two previous italicized facts that *the cardinality of* $(0, 1)$ *is equal to the cardinality of S.* But to prove this rigorously and actually construct a one-to-one correspondence between S and $(0, 1)$ requires the Cantor-Schroeder-Bernstein Theorem (see page 128.)

We know that the cardinality of $(0, 1)$ is c from Theorem 3.5.4, and so this completes the "proof". □

[21] *Aside*: Certain points z in the line segment $(0, 1)$ do not correspond to any point (x, y). These are the points z which have an infinite expansion ending in an infinite sequence of 0's in every even place *or* ending in an infinite sequence of 0's in every odd place. For example, we do not get any z with a decimal expansion $z = \cdot 168403070606050408000303050900000 2010408\ldots$, because this would require $x = \cdot 180000000000000000\ldots$ and $y = \cdot 64376654803359002148\ldots$ and we have ruled out decimal expansions for x or y ending in an infinite string of 0's.

Remark The map we constructed in the proof is very "ungeometric". For example

$$(x, y) = (\cdot 39999\ldots, \cdot 39999\ldots) \leftrightarrow z = \cdot 3399999999\ldots,$$
$$(x', y') = (\cdot 40001\ldots, \cdot 40001\ldots) \leftrightarrow z' = \cdot 4000140001\ldots\ .$$

The difference between x and x' is only $\cdot 00002\ldots$, and similarly for y and y'. But the difference between z and z' is $\cdot 06\ldots$, which is much larger. In fact we can get x and x' as close as we like, and similarly for y and y', while z and z' will still be about $\cdot 06\ldots$ apart. *How?*

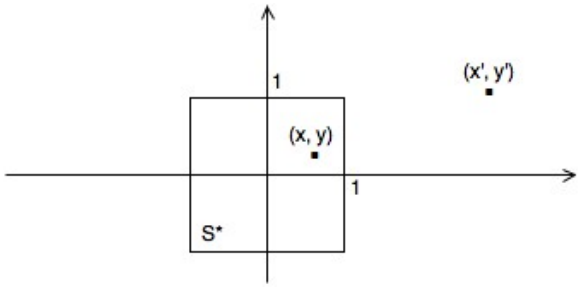
To summarise: Points that are close in S did not always go to points that are close in $(0, 1)$.

⋆The Plane $\mathbb{R}^2$ The plane is often denoted by $\mathbb{R}^2$. It can be represented as the set of pairs of numbers (x, y) where x and y are any real numbers.

Theorem 3.5.6. *The cardinality of the plane $\mathbb{R}^2$ is c.*

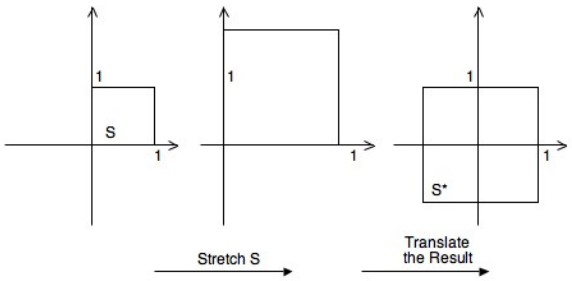
Proof. In Theorem 3.5.4 we used stereographic projection to obtain a one-to-one correspondence $x \leftrightarrow x'$ between points x in $(-1, 1)$ and points x' in $\mathbb{R}$.

Instead of the unit square S consider the square S^* consisting of all points (x, y) such that x is in the line segment $(-1, 1)$ and y is also in the line segment $(-1, 1)$. Then $(x, y) \leftrightarrow (x', y')$ is a one-to-one correspondence between S^* and $\mathbb{R}^2$. *Why?* See the following Diagram.



There is also a one-to-one correspondence between S and S^* which is obtained by stretching S by a factor 2 in both the x and y directions and then translating the result. The formula is

$$(x^*, y^*) = (2x - 1, 2y - 1).$$



Because there is a one-to-one correspondence between S and S^* and another between S^* and $\mathbb{R}^2$, it follows from Theorem 3.1.3 that there is a one-to-one correspondence between S and $\mathbb{R}^2$. Since S has cardinality c it follows that $\mathbb{R}^2$ has cardinality c. □

⋆*More Advanced Topics*

This material is not in [HM].

Notation It is convenient to denote the cardinality of a set A by $|A|$. So $|\mathbb{N}| = |\mathbb{Q}| = d$ and $|\mathbb{R}| = |\mathbb{R}^2| = |\mathbb{I}| = c$.

If the cardinality of a set A is less than the cardinality of B as in Definition 3.4.3 we write $|A| < |B|$. If the cardinality of A is equal to the cardinality of B we write $|A| = |B|$.

In particular,

$$1 < 2 < 3 < \cdots < n < \cdots < d < c.$$

If there is a one-to-one correspondence between a set A and some subset of a set B (which may be all of B) we say that the cardinality of A is less than or equal to the cardinality of B and we write $|A| \leq |B|$.

It follows from Definition 3.4.3 that

$$|A| \leq |B| \quad \text{iff} \quad \big(|A| < |B| \text{ or } |A| = |B|\big).$$

In terms of cardinal numbers α and β [22] we can write this as

$$\alpha \leq \beta \quad \text{iff} \quad \big(\alpha < \beta \text{ or } \alpha = \beta\big).$$

There are two major results that we did not prove.

Cantor-Schroeder-Bernstein Theorem This says that for any two sets A and B,

$$\big(|A| \leq |B| \ \&\ |B| \leq |A|\big) \text{ implies } |A| = |B|.$$

This may seem obvious when we write it this way. But what it is saying is that if there is a one-to-one correspondence between A and some subset of B, and another one-to-one correspondence between B and some subset of A, then there is a third one-to-one correspondence between A and B.

Another way of writing the Cantor-Schroeder-Bernstein Theorem is that for any two cardinal numbers α and β,

$$\big(\alpha \leq \beta \ \&\ \beta \leq \alpha\big) \text{ implies } \alpha = \beta.$$

Comparing Cardinals Theorem This says that for any two sets A and B,

$$|A| \leq |B| \text{ or } |B| \leq |A|.$$

(As always in mathematics, "or" allows for both statements to be true.) Although this theorem may seem unsurprising, it is far from obvious. It is equivalent to saying that given any two sets there is always a one-to-one correspondence from one of them into a subset of the other. The proof of this requires the Axiom of Choice.

Another way of expressing this is that for any two cardinals α and β,

$$\alpha \leq \beta \text{ or } \beta \leq \alpha.$$

[22] These are the first two letters alpha and beta of the Greek alphabet.

Ordering Cardinals It follows from the previous results that for any two cardinals α and β, *exactly one* of the following is true:

$$\alpha < \beta \text{ or } \alpha = \beta \text{ or } \beta < \alpha.$$

A Brief History of Set Theory

This material is not in [HM

Abbreviated and slightly modified from `http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/Beginnings_of_set_theory.html`.

The idea of infinity has been the subject of deep thought from the time of the ancient Greeks. By the Middle Ages discussion of the infinite had led to comparison of infinite sets.

In 1847 Bolzano defended the concept of an infinite set at a time when many believed that infinite sets could not exist. He gave examples to show that, unlike the case for finite sets, the elements of an infinite set could be put in one-to-one correspondence with elements of one of its proper subsets.

In 1874 Cantor published an article in Crelle's Journal which marks the birth of set theory. In his paper Cantor considered at least two different kinds of infinity. Before this, orders of infinity did not exist and all infinite collections were considered "the same size". However Cantor showed that the rational numbers are in one-to-one correspondence with the natural numbers. In the same paper he shows that the real numbers cannot be put into one-to-one correspondence with the natural numbers using an argument which is more complex than that used today (which is also due to Cantor in a later paper of 1891).

However, set theory was now becoming the centre of controversy. Kronecker, who was on the editorial staff of Crelle's Journal, was unhappy about the revolutionary new ideas contained in Cantor's paper.

In his next paper in 1878 Cantor introduced the idea of equivalence of sets and said two sets are equivalent or have the same power (cardinality) if they can be put in one-to-one correspondence. He proved that the natural numbers have the smallest infinite cardinality and showed that the set of points in the plane or in space has the same cardinality as $\mathbb{R}$. He showed further that countably many copies of $\mathbb{R}$ still have the same cardinality as $\mathbb{R}$.

Cantor published a six part treatise on set theory from the years 1879 to 1884. But there was growing opposition to his ideas. The leading figure in the opposition was Kronecker, whose criticism was built on the fact that he only accepted mathematical objects that could be constructed finitely from the intuitively given set of natural numbers. Cantor's array of different infinities were impossible under this way of thinking.

The year 1884 was one of mental crisis for Cantor. He seemed to lose confidence in his own work and applied to lecture on philosophy rather than on mathematics. The crisis did not last long, by early 1885 he recovered and his faith in his own work had returned. In 1897 the first International Congress of Mathematicians was held in Zurich and at that conference Cantor's work was held in the highest esteem.

In 1899 Cantor discovered the paradox which arises from the set of all sets. Clearly it must have the greatest possible cardinal, yet the cardinal of the set

of all subsets of a set always has a greater cardinal than the set itself. It began to look as if the criticism of Kronecker might be at least partially right since extension of the set concept too far seemed to be producing the paradoxes.

Bertrand Russell (mathematician, philosopher and peace activist) discovered Russell's paradox in 1902. Russell wrote to Frege telling him about the paradox as Frege had nearly completed his major treatise on the foundations of arithmetic. Frege added an acknowledgement to his book: "A scientist can hardly meet with anything more undesirable than to have the foundation give way just as the work is finished. In this position I was put by a letter from Mr Bertrand Russell as the work was nearly through the press."

By this stage, however, set theory was beginning to have a major impact on other areas of mathematics. Rather than dismiss set theory because of the paradoxes, ways were sought to keep the main features of set theory while eliminating the paradoxes.

Gödel showed in 1939 that the Axiom of Choice and the Continuum Hypothesis cannot be disproved using the other axioms of set theory. This, and other work earlier work of Gödel, is some of the most influential and profound work in mathematics in the 20th century. In 1963 Paul Cohen proved that the Axiom of Choice and the Continuum Hypothesis cannot be proved from the other axioms of set theory, for which he received a Fields medal.

Russell's paradox had undermined the whole of mathematics according to Frege. Russell, trying to repair the damage, made an attempt to put mathematics back onto a logical basis in his major work "Principia Mathematica" written with Whitehead. However their methods did not seem a very satisfactory way around the problems and others sought different ways.

Zermelo in 1908 was the first to attempt an axiomatisation of set theory. Many other mathematicians attempted to axiomatise set theory. Fraenkel, von Neumann, Bernays and Gödel are all important figures in this development. Gödel showed the limitations of any axiomatic theory and that the aims of many mathematicians such as Frege and Hilbert could never be achieved.

Questions

1 Questions 6–22 on pp 202–205 of [HM].

2 Show that $(0, \infty)$ has the same cardinality c as $\mathbb{R}$.
HINT: Think of the graph of the function $y = \log x$.

3 Show that $(0, \infty)$ and (a, ∞) have the same cardinality for any a.
HINT: What is a nice geometric correspondence between these two intervals.

4 Show that (a, ∞) and $[a, \infty)$ have the same cardinality.
HINT: Use Theorem 3.3.4.

5 Show that $[a, \infty)$ and $(-\infty, -a]$ have the same cardinality. Similarly for (a, ∞) and $(-\infty, -a)$.

6 Use the previous Questions to show that (a, ∞), $[a, \infty)$, $(-\infty, b)$ and $(-\infty, b]$ all have the same cardinality c.

7 (We saw in Theorem 3.2.6 that the union of two sets of cardinality d has cardinality d. It follows from Question 3 on page 112 that the union of a set of cardinality c and a set of cardinality d has cardinality c. In this Question you will see that the union of two sets of cardinality c is again of cardinality c.)

Suppose A and B have no elements in common and each have cardinality c. Prove that $A \cup B$ has cardinality c.

HINT: There is a one-to-one correspondence between A and the interval $(-\infty, 0)$, and a one-to-one correspondence between B and the interval $[0, \infty)$.

(If A and B have some elements in common then the result is still true and not surprising. But this is trickier to prove, unless we use the Cantor-Schroder-Bernstein Theorem, see page 128.)

8 Using two coordinate axes we represented each point in the plane by a pair (x, y) of real numbers. In a similar way by using three coordinate axes we can represent each point in space by a triple (x, y, z) of real numbers. For this reason we often write $\mathbb{R}^3$ for the set of points in space.

Prove $\mathbb{R}^3$ has cardinality c.

HINT: Use Theorem 3.5.6 to show there is a one-to-one correspondence between $\mathbb{R}^3$ and $\mathbb{R}^2$.

Chapter 4

Chaos and Fractals

This Chapter corresponds to Chapter 6 in [HM].

"Chaos" and "Fractals" are now almost household words. They represent areas of mathematics that have been investigated and applied extensively in the last 30 years or so. Their theoretical investigation and their applications are closely connected with computer simulations, and as we will see many "user friendly" software packages are now freely available.

There is no precise definition of "chaos" or "fractal". But informally, chaos or chaotic behaviour corresponds to a process which eventually behaves in an erratic and essentially non-predictable manner.[1] This in fact happens very frequently in nature and even in economics. Examples are long term growth of various animal and other populations, fluctuations in the money markets, weather patterns, and many more.

Fractal sets or images have the property that if we look at them under a microscope, using larger and larger magnifications, we always continue to see more and more structure. (On the other hand if you look at a smooth curve or surface under a microscope with increasing magnification, the result is rather boring — either a straight line or a flat surface.) In the following Section and in Section 6.1 there are many examples of fractals — have a look.

Many objects in nature can best be modelled by fractals. Examples include

- *Biology*: breast tissue patterns, structure and development of plants, blood vessel patterns, morphology of fern leaves;
- *Chemistry*: pattern-forming alloy solidification, diffusion processes;
- *Physics*: transport in porous media, statistical mechanics, dynamical systems, turbulence, wave propagation;
- *Geology*: particle-size distributions in soil, landscape habitat diversity;
- *Computer Science*: image compression, adding machines and wild attractors, compression of audio signals, compression for multimedia, evolutionary algorithms, neural networks, cellular automata;
- *Engineering*: rough surfaces, image encoding, antennae, stochastic optimal control, signal processing, fragmentation analysis of thin plates.

Some aspects of fractal sets have been studied by mathematicians since at least 1872, when Wierstrass discovered a function which is continuous but not

[1] But chaotic processes *do* usually behave in a manner about which we can have good *statistical* information. For example, even though a fair dice and an unfair dice will both be unpredictable, their statistical behaviour is different but may well be predictable in each case.

differentiable at any point.[2] But such examples were then considered to be only of interest to pure mathematicians.

Benoit Mandelbrot was the real founder of the subject of fractals. In the 1960's he realised that fractal behaviour occurred in many natural phenomena including turbulence, noise and errors in electronic transmissions, geographical features such as coastlines. In 1982, he published his seminal book "The Fractal Geometry of Nature". This created an enormous amount of interest: mathematical, applied, and in the public imagination.

Mandelbrot also coined the word "fractal". He took it from the Latin "fractus" ("broken"), which is the past participle of the Latin "frangere" ("to break" or create irregular fragments).

Chaotic behaviour has been observed in mathematical and physical systems at least as far back as the 1850's. It was discussed by Maxwell in 1860 that arbitrarily small changes in the initial position of atoms in a gas would lead to wildly different positions at later times and that this could be used to predict the properties of gases. In 1890 Poincare explained how arbitrarily small changes in the initial position of three interacting gravitational bodies (e.g. planets or stars) could lead to completely different later behaviour.

In 1962 the meteorologist Lorenz discovered that miniscule changes in the equations used to predict the weather would soon lead to completely different weather predictions for a few days later, known as the *butterfly effect*. Lorenz also discovered "order" in this chaos, and showed that even a very much simplified version of his equations led to similar effects. More precisely, the solution to these simplified equations is a point moving around in three dimensional space and following a path that becomes arbitrarily close to the lines on what is called the *Lorenz Attractor*, see Figures 4.1 and 4.2.

In 1976 the Australian Robert May, a theoretical physicist and later an ecologist, wrote a famous paper[3] in the journal *Nature* explaining how simple mathematical models exhibiting chaotic behaviour arise in the biological, economic and social sciences.

The underlying theme in this Chapter is that of a repeated or *iterative* process. An example is repeatedly applying a simple process such as the quadratic function $f(x) = x^2 + c$ where c is a fixed number. This will lead us both to chaotic behaviour and to incredibly complicated and beautiful fractals (Julia sets and the Mandelbrot set).

Contents

[2] See Figure 4.7 for an example of the graph of such a function, although pixellation effects might make the function there appear differentiable at some points. Wierstrass's example itself was not generated by a random process, as is the case in Figure 4.7.

[3] *Simple Mathematical Models with very complicated dynamics*, Nature 1976, **261**, 459–467.

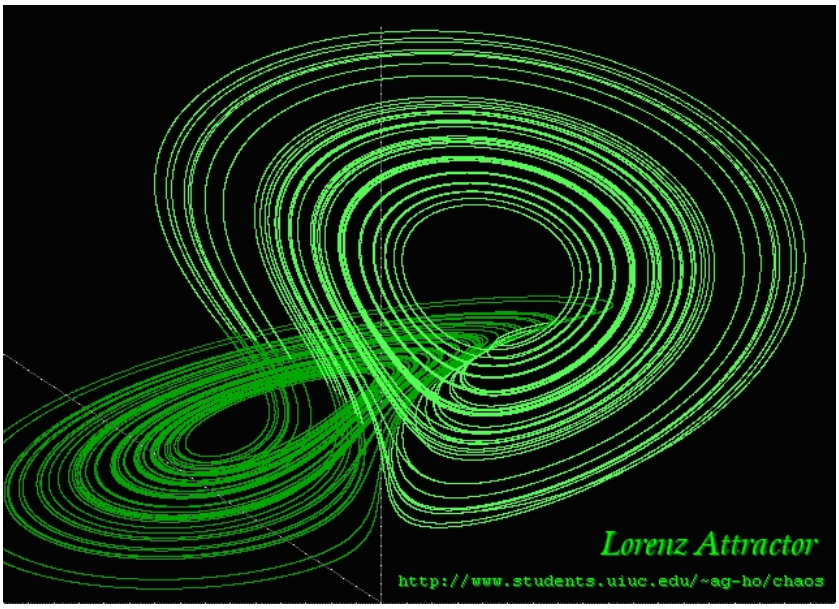


Figure 4.1: The Lorenz Attractor

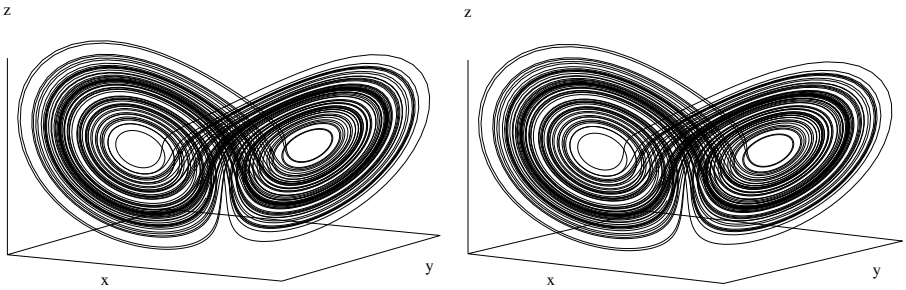


Figure 4.2: The Lorenz Attractor in Stereo. To view, stare at the centre of the two images and cross your eyes until the two images merge. Allow your eyes to relax so they can refocus.

[4]Once you are out you never get back in. Even if you stay in, the situation is very unstable. There are points arbitrarily close by which will eventually be thrown out.

4.1 A Gallery of Fractals

Great fleas have little fleas,
Upon their backs to bite 'em,
And little fleas have lesser fleas,
And so ad infinitum.

Augustus De Morgan,
A Budget of Paradoxes, 1872

Overview

[HM, 404–411]

Look closely at the images shown here. I will make a few comments but we will discuss these examples in more detail in the following sections.

Sierpinski Triangle

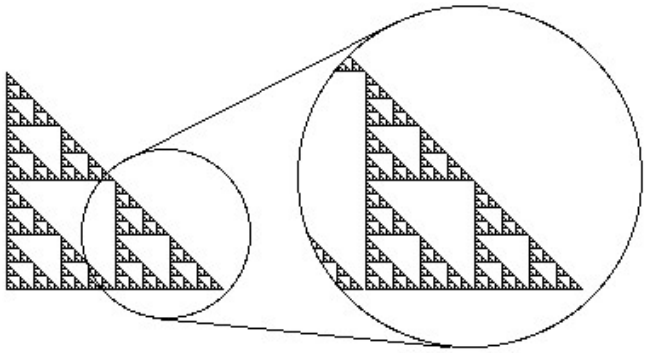


Figure 4.3: Sierpinski Triangle

The large "triangle" on the left of Figure 4.3 is composed of (i.e. is the union of) three smaller "triangles", each of which is a scaled model of the original. The right side shows what you see under a microscope, with magnification $\times 2$, applied to the small triangle on the bottom right side of the large triangle on the left.

Each of the three small triangles on the left of Figure 4.3 is in turn the union of three even smaller triangles, and so on. But of course, we can only draw the "real" Sierpinski triangle up to a certain resolution.

We say the Sierpinski Triangle is *self-similar*.

The Sierpinski triangle here is right-angled, more often it is equilateral. The Sierpinski triangle is an important example of a fractal set. We will see a number of ways to construct it, and analyse its properties, in Section 4.3.

Dendrites

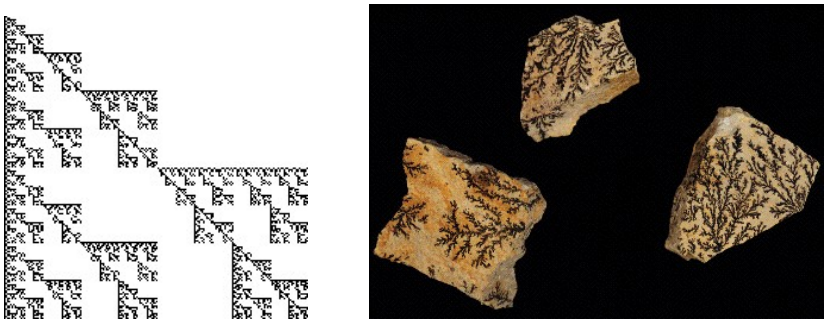


Figure 4.4: Some Dendrites

The mathematical dendrite on the left of Figure 4.4 is composed of three small dendrites. Each small dendrite is obtained by scaling the original by 1/2, translating, and in one case by also performing a reflection. *Which one? What is the reflection?* The small dendrites are composed of smaller dendrites, and so on.

On the right of Figure 4.4 are three naturally occurring dendrites. Each is a realisation of a *random self-similar* fractal since in a certain *statistical* sense smaller parts of the dendrite look like the original. It is possible to make these ideas mathematically precise, although we will not do that here.

In naturally occurring phenomena self-similarity only occurs over a range of scales. For example, when we go down to the atomic level and in fact well before, the self-similarity will certainly break down. Nonetheless, mathematical fractals where self-similarity occurs over arbitrarily small scales are extremely useful models in analysing natural fractals.

More Fractals

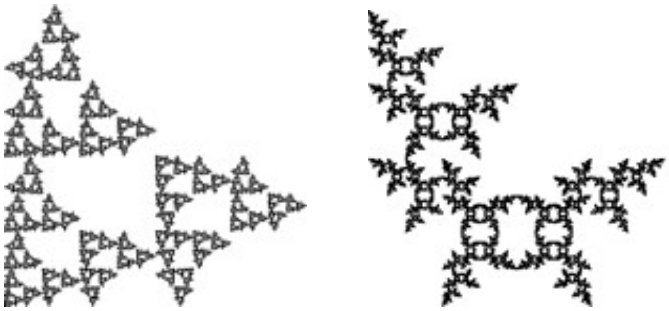


Figure 4.5: More Fractals

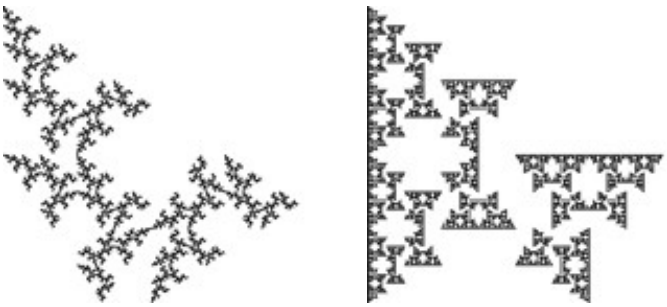


Figure 4.6: And More Fractals

✍ Each of the fractals in Figures 4.5 and 4.6 is the union of three copies of itself scaled by 1/2. Some cases are a bit tricky to see. *Can you see the three copies in each case and how they are obtained?*

Random Fractals

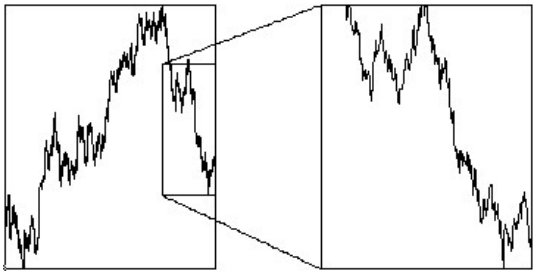


Figure 4.7: A Random Graph

The first graph in Figure 4.7 is the type of graph obtained from plotting short term fluctuations in the money markets. The second is a rescaling of a portion of the first; it is "statistically self-similar" to the full (first) graph. They are examples of random fractals.

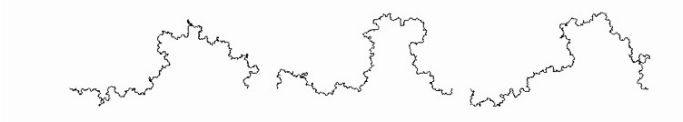


Figure 4.8: Random Koch Curves

The three curves in Figure 4.8 are examples of a "random Koch curve". Each is the union of two pieces which when suitably rescaled are themselves further examples of a random Koch curve. All are examples of random fractals.

We will not be treating random fractals in any detail in this course.

Julia Sets and the Mandelbrot Set

We will discuss these incredibly bizarre yet beautiful fractal sets in Section 4.6. They arise from repeatedly applying (iterating) the simple quadratic function $f(z) = z^2 + c$, where c is a fixed *complex* number and z is any *complex* number. They have spawned an enormous amount of computer graphics. They are typical of the behaviour obtained by iterating functions in two dimensions and are studied mathematically for this reason.

There is just one Mandelbrot set, it is the centre set (cardioid with discs and filaments) in Figure 4.9.

The other 7 sets in Figure 4.9 are particular examples of Julia sets, translated in the diagram so they do not overlap one another. There is one Julia set for every point c in the plane. The point c is indicated in each of these 7 examples by the dot at the other end of the line pointing to the corresponding Julia set.

The Mandelbrot set is the set of values of c for which the corresponding Julia set is connected. Otherwise the corresponding Julia set is a totally disconnected "dust" of points, as in the bottom example in Figure 4.9.

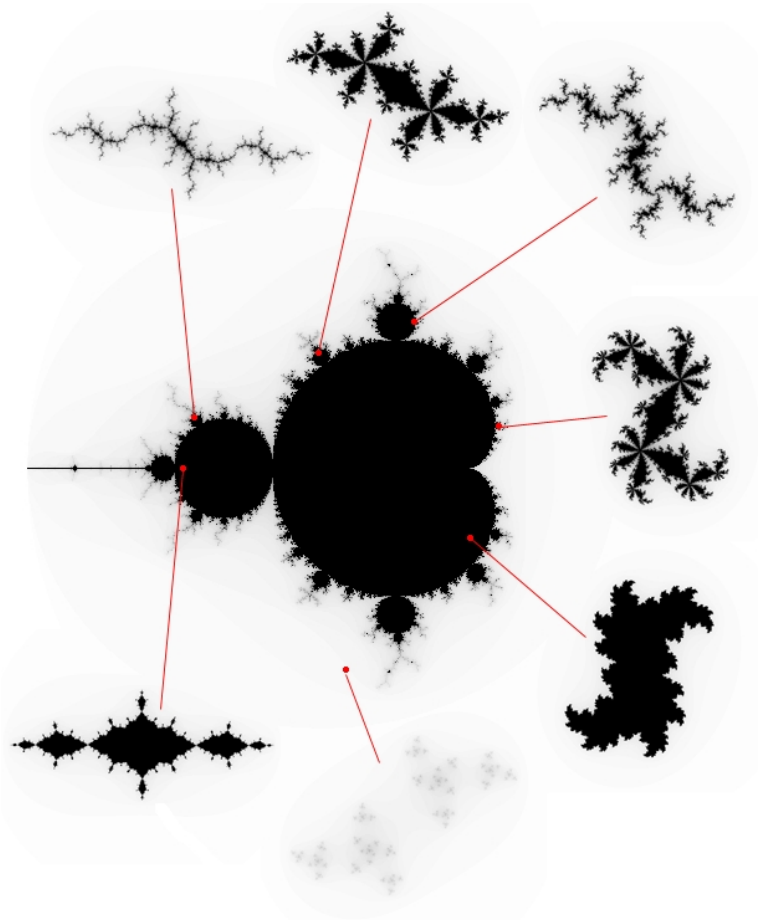


Figure 4.9: The Mandelbrot Set and some Julia Sets

Also look at the Mandelbrot set and Julia sets and other fractals in [HM, pp 404–411]. (The colouring used for the Mandelbrot and Julia set has a mathematical explanation, which we will discuss in Section 4.6.)

Questions

1 Questions 2–10 on pp 410, 411 of [HM].

4.2 Iterative Dynamical Systems

Using simple models for complicated systems can help understand important features.

Overview

An iterative dynamical system is a process where the output from each time step is used as the input for the next time step.

- Many natural processes can be understood and analysed this way.
- Iterative dynamical systems are the key to understanding chaos and fractals.

In this section we will look at a number of interesting examples of dynamical systems. We will study the connections with chaos and fractals in the following sections.

Bank Interest

[HM, 413]

Suppose we have \$100 invested in the bank, and the interest is fixed at 5% per annum

- After 1 year, we will $\$100 \times 1.05$.
- After 2 years we will have $\$100 \times 1.05 \times 1.05 = \$100 \times 1.05^2$.
- After 3 years we will have $\$100 \times 1.05^3$.

$\vdots$

- After n years we will have $\$100 \times 1.05^n$.

$\vdots$

The key point is that if we know the amount in the bank (or "output") for any particular year, then to find the amount in the bank ("output") one year later we just the known amount by 1.05.

In this case we can write down a simple formula for what the amount is after n years, namely $\$100 \times 1.05^n$. Another way of expressing this is that the number of dollars in the bank in successive years is given by the terms of the *geometric progression*

$$100, 100 \times 1.05, 100 \times 1.05^2, 100 \times 1.05^3, \ldots, 100 \times 1.05^n, \ldots.$$

In the more complicated systems we study in connection with chaos and fractals the situation will be much more complicated. There will not be any simple or useful formula for the situation after n time steps. None-the-less, we will still be able to say a good deal about what happens. And it will often be very surprising.

Simple Population Growth Model

[HM, 414]. But

Description of the Model The simplest way to model population growth of an animal, plant or bacterial population is to assume that the population grows (increases) after one time step by a certain fixed proportion of the current population. This proportion is called the *growth rate*. The time steps might be a year, or a month, or a minute, or perhaps a generation. This is called the *Simple Population Growth Model.*

The previous model for bank interest is one example of simple population growth. The population is the number of dollars in the bank, each time step is one year, and the *growth rate* is $s/100$ if the interest rate is $s\%$ per annum.

Definition 4.2.1. Let P_n be the *population at time* n. The *change in population* from time n to time $n+1$ is $P_{n+1} - P_n$. It is positive if the population is increasing and negative if it is decreasing.

The *growth rate* from time n to time $n+1$ is

$$\frac{P_{n+1} - P_n}{P_n}. \tag{4.1}$$

Fish Example If there are 1000 fish at time n and 1200 at time $n+1$ then the change in population from time n to time $n+1$ is 200 and the corresponding growth rate is

$$\frac{1200 - 1000}{1000} = \frac{1}{5}.$$

Theorem 4.2.2. *Suppose r is the growth rate for the Simple Population Growth Model. Then the population P_{n+1} at time $n+1$ is determined from the population P_n at time n by*

$$P_{n+1} = (1+r)P_n. \tag{4.2}$$

The population at any time n is determined from the initial population P_0 by

$$P_n = (1+r)^n P_0. \tag{4.3}$$

Proof. From (4.1),

$$\frac{P_{n+1} - P_n}{P_n} = r$$

for all n. From this equation we see that (4.2) follows. *Why?*

In particular, we see the population after one time step is

$$P_1 = (1+r)P_0.$$

After two time steps it is

$$P_2 = (1+r)P_1 = (1+r)^2 P_0.$$

After three time steps it is

$$P_3 = (1+r)P_2 \times r = (1+r)^3 P_0.$$

And so after n time steps the population using the Simple Growth model is

$$P_n = (1+r)^n P_0.$$

(We could use induction to prove this carefully, but we would not usually bother to do so.) □

Another way of expressing this is that the population at successive time steps is given by the terms of the *geometric progression*

$$P_0, (1+r)P_0, (1+r)^2 P_0, \ldots, (1+r)^n P_0, \ldots.$$

Fish Example Again In the case of the fish, if the growth rate r is $1/5$ then we get

$$P_{n+1} = \frac{6}{5} P_n.$$

Suppose the initial population $P_0 = 100$. Then after n time steps we see from (4.3) that the population is

$$P_n = \left(\frac{6}{5}\right)^n \times 100.$$

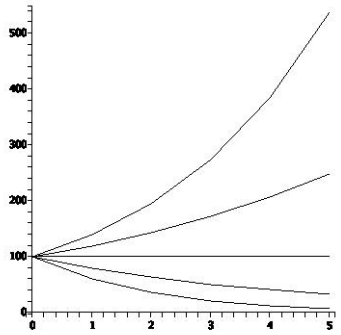


Figure 4.10: Initial population 100 fish. Growth rates $2/5, 1/5, 0, -1/5, -2/5$ respectively from top to bottom. Populations after 1, ..., 5 timesteps. Points connected by straight line segments.

In Figure 4.10 we show what happens to an initial population of 100 fish under different growth rates.

Problems with the Model The Simple Population Growth Model is not a very good model if there is a large number of time steps. If $r > 0$ then the population grows without bound. But this cannot happen due, for example, to lack of resources. We will later see that the *Logistic model* (also called the *Verhulst model*) is much better in this respect.

Notice also that the population should normally be given by an integer. But this may not happen in this model even for $P_1 = (1+r)P_0$. However, this is *not* usually a serious problem, as we are only looking at approximate models.

Iterative Processes The simple population growth model is an example of an *iterative process* or an *iterative dynamical system.*

The key point is that if we know the population P_n at some time n, then we can use this as input to find the population P_{n+1} at time $n+1$. In (4.2) we just multiplied the input P_n by $1+r$. That is,

$$P_{n+1} = (1+r)P_n.$$

From this, we showed in (4.3) that $P_n = (1+r)^n P_0$ for any n.

In the more complicated models we study later we will have a method as in (4.2) which allows us to go from knowing P_n to finding P_{n+1}. But there will *not* usually be a simple formula as in (4.3) which enables us to go directly from knowing the initial data P_0 to finding P_n.

The Verhulst Model

[HM, 417–420]. Here the material is developed further.

The Simple Growth Model is not very realistic if there is a large number of time steps, as we discussed before. A more realistic model is the *Verhulst Model*[5], also called the *Logistic Model*[6].

Population Density We assume that there is a *maximum sustainable population* which we denote by S. The actual value of S will be determined by the amount of resources available, the number of predators, and various other things. For simplicity we will assume that S is fixed and does not change with time.

Let P_n be the population after n time steps. It is convenient to let p_n be the *population density* after n time steps, given by the fraction

$$p_n = \frac{P_n}{S}. \tag{4.4}$$

So if $P_n < S$ then $p_n < 1$, if $P_n = S$ then $p_n = 1$ and if $P_n > S$ then $p_n > 1$. (It is traditional, and a little easier, to work with p_n rather than P_n. We will usually do this.)

If $P_n < S$ then we expect the growth rate (defined in (4.1)) from time n to time $n+1$ to be positive. If $P_n > S$ then we expect the growth rate to be negative. *Why?*

The Main Assumption In the Verhulst Model we make the assumption that the growth rate is proportional to $1 - p_n$. From (4.1) this implies

$$\frac{P_{n+1} - P_n}{P_n} = a(1 - p_n) \tag{4.5}$$

for some fixed number $a > 0$. By dividing numerator and denominator on the left side by S we get

$$\frac{p_{n+1} - p_n}{p_n} = a(1 - p_n). \tag{4.6}$$

[5] Pierre Verhulst was a Belgian mathematician who published his work in *Mémoires de l'Académie Royale de Belgique* in 1844 and 1847.

[6] *Logistics* [noun, plural or singular]: the detailed consideration of a complex operation involving many people, facilities or supplies.

We call a the *parameter* of the Verhulst Model.

Theorem 4.2.3. *Suppose the Verhulst Model has growth rate as in* (4.5)*. Then the population density p_{n+1} at time $n+1$ is given in terms of the population density p_n at time n by*

$$p_{n+1} = p_n + ap_n(1 - p_n) = (1 + a)p_n - ap_n^2. \tag{4.7}$$

Proof. Simplify (4.6). □

Unlike the situation for the Simple Growth Model there is no simple or useful formula which gives p_n (or P_n) in terms of p_0 (or P_0).

Use (4.7) to write p_1 in terms of p_0, then use it again to write p_2 in terms of p_1 and hence in terms of p_0. Now find p_3 in terms of p_0. As you will see, it rapidly becomes a mess.

But in this case messy things can be important, and the analysis can be very interesting and deep. See the following Examples and Section 4.5.

Examples Suppose the initial population density $p_0 = 0.1$. If we take $a = 1$ then you can check on your calculator that

$$p_0 = 0.9, p_1 = 0.19, p_2 = 0.3439, p_3 = 0.56953279, p_4 = 0.8146979811,$$
$$p_5 = 0.9656631616, p_6 = 0.9988209813, p_7 = 0..9999986103, \ldots$$

At least numerically it appears that the population density converges to 1, and this is indeed the case. See the first graph in Figure 4.11. This is not particularly surprising.

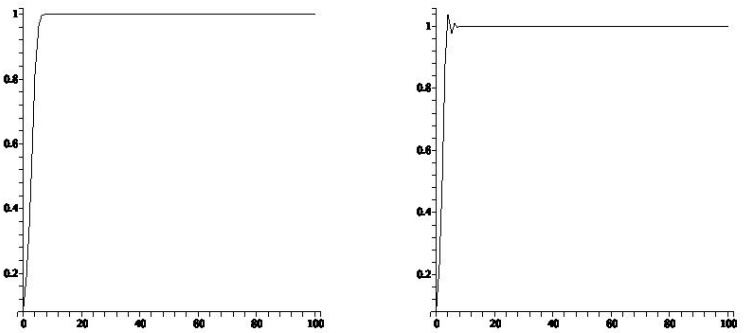


Figure 4.11: Logistic Model for initial population density 0.1 and $a = 1, 1.5$.

As a takes the values $a = 1.5, 1.9, 2.1, 2.5, 3$ we get widely differing behaviour. See Figures 4.11–4.13. For $a = 1.5, 1.9$ the population density converges to 1, for $a = 2.1$ it oscillates between two values (although pixelation effects obscure this), for $a = 2.1$ it oscillates between four values, and for $a = 3$ the behaviour appears to be chaotic with no discernible pattern.

The "problem" is that the population density p_1 significantly overshoots the maximum sustainable density which is 1. As a increases towards 3 this overshooting and undershooting becomes quite wild.

The initial population density usually makes little difference to the overall features. We will discuss the chaotic properties for a problem essentially the same as the Logistic Model in Section 4.5.

If $a > 3$ the model is no longer physically reasonable. See Question 3.

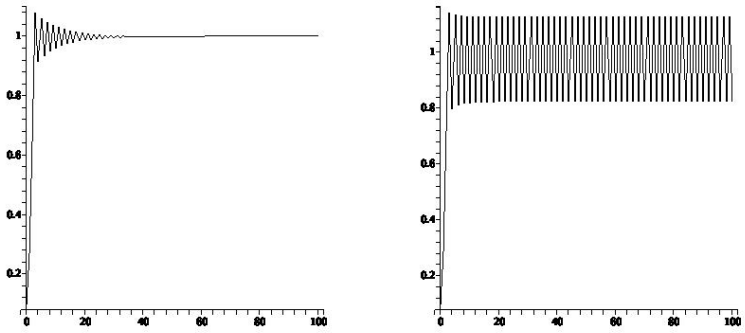


Figure 4.12: Logistic Model for initial population density 0.1 and $a = 1.9, 2.1$ respectively.

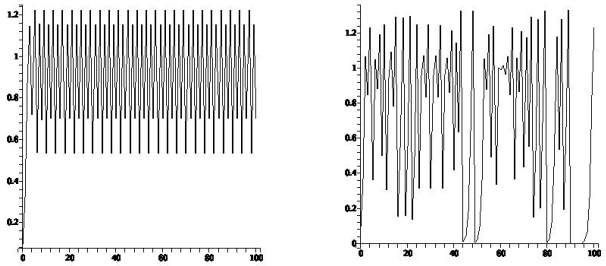


Figure 4.13: Logistic Model for initial population density 0.1 and $a = 2.5, 3$ respectively.

Game of Life

[414–416]

History The *Game of Life* was invented by the Cambridge mathematician John Conway and discussed in Martin Gardner's column in *Scientific American* 1970, **223**, 120–123.

It has analogies with the life cycles and patterns of societies and various living organisms. It developed out of ideas involving automata theory and self-

replicating machines. It shows how complex patterns can evolve from very simple rules. For these reasons it is of interest to physicists, biologists, economists, computer scientists and mathematicians.

Rules The Game of Life is played on an infinite square grid. Notice that each square S in such a grid has exactly 8 neighbours: 4 squares that each have a side in common with S and 4 squares that each have a corner in common with S. See Figure 4.14, where of course the grids however are finite.

At each time interval (or generation) a square is either alive or dead. We begin with a finite number of live squares and then proceed as follows:

- If a square is alive at one generation and it has exactly 2 or 3 live neighbours, then it remains alive in the next generation. Otherwise it dies at the next generation.
- If a square is dead at one generation and it has exactly 3 live neighbours, then it comes to life at the next generation. Otherwise it remains dead.

You might think of the rules as follows:

- If a square is alive and it has less than 2 live neighbours then it dies from loneliness. It it has more than 3 live neighbours then it dies from overcrowding. If it has exactly 2 or 3 live neighbours then conditions are optimal and it survives to the next generation.
- If a square is dead and it has exactly 3 live neighbours then conditions are just right for the square to come to life at the next generation. (Perhaps 2 squares are required for reproduction and a third to assist in child rearing!) If a dead square has less than 3 live neighbours it remains dead from a lack of genetic material and assistance. If it has more than 3 live neighbours then it remains dead because of overcrowding and a lack of resources.

Examples[7] A single live square will die at the next generation. *Why?* ✍

If there are exactly two live squares, no matter where placed, they will die at the next generation. *Why?* ✍

With three live squares the situation is more interesting.

Examples a, b and c in Figure 4.14 will die after 2 generations. *Check!* ✍

Example d forms a block after one generation and this does not change in succeeding generations.

Example e forms a vertical line of 3 live squares after one generation, then again a horizontal line, etc. It is called a *blinker*

You should now run all the examples from "Conway's Game of Life" under "Chaos and Fractals" from the CD in the book [HM]. (The program on the CD may be confusing, since if a pattern moves off one side of the screen it will reappear at the diagonally opposite position on the opposite side.)

You might also look at the java applet at `bitstorm.org/gameoflife/`. (It also has a small bug when patterns move off the screen.)

[7] These are taken from the original Scientific American article noted in the previous historical comments.

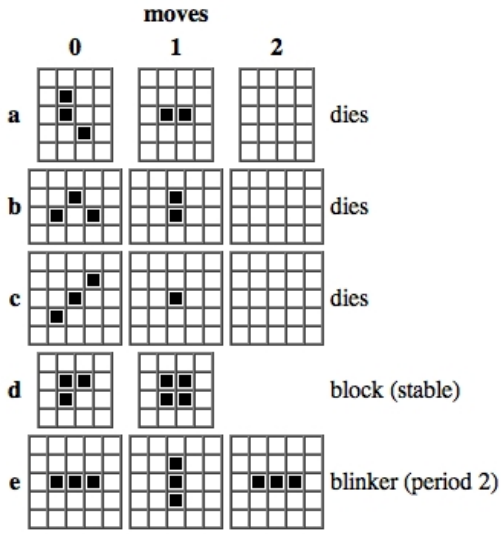


Figure 4.14: Game of Life

In general, some patterns die out, some evolve into a static pattern, some evolve into oscillating patterns, some translate across the board, some "oscillate" except that every cycle creates additional material which then moves away from the main group. See the "glider guns" on the CD or java applet for the last type of bevaviour.

Questions

1 Suppose a population of 1000 is slowly becoming extinct and that the growth rate is -0.1 per annum. We sometimes say the decay rate is 0.1 per annum.

Write down a formula using the Simple Growth Model which gives the population after n years.

How many years until the population is 500, 400, 300, 200, 100?

2 Suppose an initial population is given by the number P_0 and the growth rate for each time step is r.

If $r > 0$ find a formula for the minimum number of time steps until the population is at least double. If $r < 0$ find a formula for the number of time steps until the population is at least halved.

3 Take an initial population density $p_0 = 2/3$ in the Logistic Model. SDhow that

$$p_1 = \frac{2}{3} + \frac{2}{9}a, \quad p_2 = \frac{2}{81}(3-a)(3+a)(2a+3).$$

For which a is $P_2 < 0$? (For such a the Logistic Model is not physically reasonable.)

4 Questions 10–11, 21–23 from p423 of [HM].

Remark You could also do Questions 24–40 on pp 423–428 of [HM] at this stage. They are a "lead in" to later Sections in this Chapter. We will return to these Questions at the appropriate time.

4.3 Fractals By Repeated Replacement

Multiple repetitions of simple processes can lead to complex structures.

This Section corresponds to
The material is developed h

Overview

We discuss how to generate fractals by a process of repeated replacement.

We also look at some of the surprising properties of fractals. For example, fractal sets have the following self-similarity property: arbitrarily small parts of a fractal set will, after scaling, look similar to the original set. We saw examples of this in Section 4.1 and we will see more examples here.

We proceed by studying four typical examples: the Koch Curve, the Sierpinski Triangle, the Menger Sponge, and the Cantor Set.

Koch Curve

Construction Begin with a line segment of length one, and replace it by a polygonal line consisting of 4 straight segments each of length 1/3 as in Figure 4.15. (The middle segment "removed" from the initial segment has length 1/3. It and the two oblique segments form an equilateral triangle.)

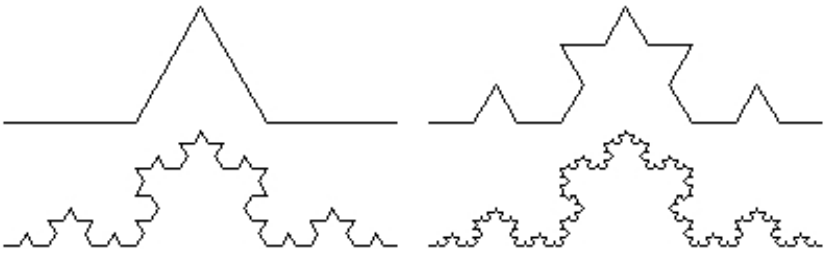


Figure 4.15: Approximations to the Koch Curve

Replace each of these 4 new segments by another four segments each 1/3 the length of the segment being replaced. This leads to the third polygonal line consisting of 16 ($= 4 \times 4 = 4^2$) segments.

Next replace each of these new segments by another four segments each 1/3 the length of the segment being replaced. This leads to the fourth polygonal line consisting of 64 ($= 4^3$) segments.

Next replace each of these new segments by another four segments each 1/3 the length of the segment being replaced. This leads to the fifth polygonal line consisting of 256 ($= 4^4$) segments.

Etc.

The limiting set is the *Koch Curve*. It is a continuous curve but it does not have a tangent at any point.[8]

To make the construction mathematically precise we need to take the limit in a certain sense. Computationally or in nature the iterative process is only done finitely often. But you should think of it as being done a few billion times! Or at least 100 times.

The Koch curve, or even the approximation after a large number of iterations, is very complicated. The protruberances become spirals, and in the Koch curve itself become infinite spirals.

Self Similarity The Koch curve is the union of 4 copies of itself, where each copy is scaled by 1/3. *What are the 4 copies?* We say the Koch Curve is *self-similar*.

Each of these 4 copies is in turn the union of 4 smaller scaled copies of itself, where each copy is scaled by 1/3. So it follows that the Koch curve is the union of 16 copies of itself, each scaled by $(1/3)^2$. *What are the 16 copies in Figure 4.15?*.

Repeating this argument, we see the Koch curve is the union of $64{=}4^3$ copies of itself, where each copy is scaled by $(1/3)^3$.

In fact for any natural number n, the Koch curve is the union of 4^n copies of itself, where each copy is scaled by $(1/3)^n$.

In fact if we zoom in on the Koch curve to *any* scale of magnification, we keep seeing more and more copies of itself.

In fact, the Koch curve is also the union of 2 copies of itself, each obtained by scaling, rotating and then reflecting in a certain line. *What are the 2 copies? What is the scaling factor? What is the line of reflection in each case?*

Length The length of the first approximation to the Koch curve is 4/3, of the second is $(4/3)^2$, of the third is $(4/3)^3$, and of the nth is $(4/3)^n$. So we expect that the length of the Koch curve is infinite.

It is possible to give a precise definition of the length of the Koch curve, and then its length is indeed infinite. See Question 1.

It is important to realise however, that just because the length of the approximations approaches infinity, it does not automatically follow that the length of the Koch curve is infinite. See Question 2.

Sierpinski Triangle

Construction Begin with a solid equilateral triangle.

[8] To make these statements precise we would need to give a precise definition of a continuous curve and of a tangent line. This can be done, although we will not be doing it in this course.

Figure 4.16: Approximations to the Sierpinski Triangle

Replace it by three equilateral triangles each 1/2 the side length of the original triangle, as in Figure 4.16.

Replace each of these triangles by three equilateral triangles each 1/2 the side length of the triangle being replaced. There are now $3^2 = 9$ triangles.

Replace each of these new triangles by three equilateral triangles each 1/2 the side length of the triangle being replaced. There are now $3^3 = 27$ triangles.

Replace each of these new triangles by three equilateral triangles each 1/2 the side length of the triangle being replaced. There are now $3^4 = 81$ triangles.

Etc.

The *Sierpinski Triangle* is the limit, in a sense that can be made precise, of this process. In practice, the process can only be done a large finite number of times.

Self-similarity The Sierpinski triangle is the union of 3 copies of itself, where each copy is scaled by 1/2. *What are the 3 copies?* We say the Sierpinski triangle is *self-similar*.

Each of these 3 copies is in turn the union of 3 smaller scaled copies of itself, where each copy is scaled by 1/2. So it follows that the Sierpinski triangle is the union of 9 copies of itself, each scaled by $(1/2)^2$. *What are the 9 copies in Figure 4.16?*.

Repeating this argument, we see the Sierpinski triangle is the union of $27{=}3^3$ copies of itself, where each copy is scaled by $(1/2)^3$.

In fact for any natural number n, the Sierpinski triangle is the union of 3^n copies of itself, where each copy is scaled by $(1/2)^n$.

If we zoom in on the Sierpinski triangle to *any* scale of magnification, we keep seeing more and more copies of itself.

Length and Area Suppose that the length of each of the sides of the original triangle in Figure 4.16 is 1. Then the length of the boundary, i.e. perimeter, is 3. The area of the triangle is $\sqrt{3}/4$. *Why?*

We saw above that the nth approximation consists of 3^n triangles each of which is obtained by scaling the original triangle by $(1/2)^n$ in all directions.

A very important general fact is the following:

When we scale in all directions by r, lengths are multiplied by r, areas by r^2 and volumes by r^3.

We say the scaling factor for length is r, for area is r^2, and for volume is r^3.

Explain why this is true for rectangles, triangles, circles and cubes. ✍

It follows that for the nth approximation to the Sierpinski triangle:

boundary length L_n

$$= \text{ no. of triangles} \times \text{boundary length of initial triangle} \times \text{length scaling factor}$$

$$= 3^n \times 3 \times \left(\frac{1}{2}\right)^n = 3\left(\frac{3}{2}\right)^n$$

area A_n $= \text{ no. of triangles} \times \text{area of initial triangle} \times \text{area scaling factor}$

$$= 3^n \times \frac{\sqrt{3}}{4} \times \left(\left(\frac{1}{2}\right)^n\right)^2 = 3^n \times \frac{\sqrt{3}}{4} \times \left(\frac{1}{4}\right)^n = \frac{\sqrt{3}}{4}\left(\frac{3}{4}\right)^n.$$

Notice that L_n gets arbitrarily large as n increases. We say L_n approaches infinity as n approaches infinity. On the other hand, the area A_n approaches zero. We write[9]

$$L_n \to \infty,\ A_n \to 0 \text{ as } n \to \infty.$$

It is possible to give a precise definition of length and area, and of boundary. Then it turns out that the area of the Sierpinski triangle is zero but the length of its boundary is infinite. Moreover, the boundary of the Sierpinski triangle is in fact the entire Sierpinski triangle.

However, all this does not follow automatically, just as we noted before in the case of the Koch curve.

Menger Sponge

Construction Begin with a cube as in Figure 4.17.

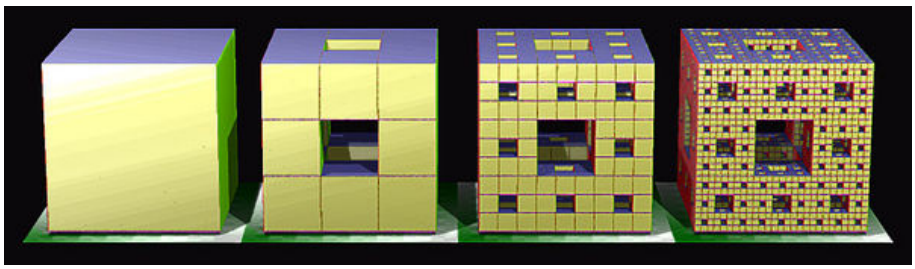


Figure 4.17: Approximations to the Menger Sponge

Replace this cube by 20 cubes each of 1/3 the side length. (The 7 "missing" cubes are the one from the top middle, the one from the bottom middle, one from the middle of each of the four sides, and the one from the centre.)

Replace each of these 20 cubes by another 20 each of 1/3 their side length. There are now $20^2 = 400$ cubes.

Replace each of these new cubes by another 20 each of 1/3 their side length. There are now $20^3 = 8000$ cubes.

[9] Of course, ∞ is *not* a number. It is just a convenient way to help express the fact that L_n and n can be arbitrarily large.

Etc.

The *Menger Sponge* is the limit, again in a sense that can be made precise, of this process.

Properties The Menger sponge is the union of 20 copies of itself, each scaled by 1/3. So we say that the Menger sponge is *self-similar*. Remember that we can only draw approximations to the actual Menger sponge, as in Figure 4.17.

The nth approximation to the Menger Sponge consists of 20^n cubes each of which is obtained by scaling the original cube by $(1/3)^n$.

Suppose the initial cube has side length equal to one. It follows that for the nth approximation to the Menger Sponge:

surface area A_n

$$= \text{ no. of cubes} \times \text{surface area of initial cube} \times \text{area scaling factor}$$

$$= 20^n \times 6 \times \left(\left(\frac{1}{3}\right)^n\right)^2 = 20^n \times 6 \times \left(\frac{1}{9}\right)^n = 6\left(\frac{20}{9}\right)^n$$

volume V_n

$$= \text{ no. of cubes} \times \text{volume of initial cube} \times \text{volume scaling factor}$$

$$= 20^n \times 1 \times \left(\left(\frac{1}{3}\right)^n\right)^3 = 20^n \times 1 \times \left(\frac{1}{27}\right)^n = \left(\frac{20}{27}\right)^n.$$

Notice that the surface area A_n approaches infinity as n approaches infinity. On the other hand, the volume V_n approaches zero. We write

$$A_n \to \infty,\ V_n \to 0 \text{ as } n \to \infty.$$

It is possible to give a precise definition of volume and surface area. Then it turns out that the surface area of the Menger Sponge is infinite but the volume is zero. Moreover, the surface of the Menger Sponge is in fact the same as the Menger Sponge.

Once again, none of this follows automatically, but requires careful definition and proof.

Cantor Set

The Cantor set is the simplest fractal set. For this reason it is useful to study it in more detail.

Figure 4.18: Seven approximations to the Cantor Set

Construction Begin with a closed line segment of length one, and replace it by 2 line segments each of length 1/3 as in Figure 4.15.[10] (The middle open segment "removed" from the initial segment has length 1/3.)

Replace each of these 2 new segments by another 2 segments each 1/3 the length of the segment being replaced (i.e. remove the middle open third of each segment). This gives $4 = 2^2$ line segments.

Next replace each of these new segments by another 2 segments each 1/3 the length of the segment being replaced (i.e. remove the middle open third of each segment). This gives $8 = 2^3$ line segments.

Etc.

Let $C_0, C_1, C_2, C_3, \ldots$ denote the sets in Figure 4.15. That is:

$$C_0 = [0, 1]$$

$$C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right] \tag{4.8}$$

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{3}{9}\right] \cup \left[\frac{6}{9}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right] \tag{4.9}$$

$$C_3 = \left[0, \frac{1}{27}\right] \cup \left[\frac{2}{27}, \frac{3}{27}\right] \cup \left[\frac{6}{27}, \frac{7}{27}\right] \cup \left[\frac{8}{27}, \frac{9}{27}\right] \cup \left[\frac{18}{27}, \frac{19}{27}\right] \cup \left[\frac{20}{27}, \frac{21}{27}\right] \cup \left[\frac{24}{27}, \frac{25}{27}\right] \cup \left[\frac{26}{27}, 1\right]$$

etc.

The *Cantor Set* C is the set of points which are in *every* C_n. We write

$$C = \bigcap_{n \geq 0} C_n. \tag{4.10}$$

Notice that C_n consists of 2^n intervals each of length $(1/3)^n$ and that

$$C_0 \supset C_1 \supset C_2 \supset C_3 \supset \cdots \supset C_n \supset \cdots \supset C.$$

The Cantor set is the union of 2 copies of itself, each obtained by scaling by 1/3. For this reason the cantor set is *self-similar*.

Describing the Points in the Cantor Set

Endpoints of Intervals Every endpoint of every interval used in the construction, is itself in the Cantor Set. For example, $0 \in C$ since it is clear that $0 \in C_n$ for every n.

Similarly, $1/3 \in C$ since $1/3 \in C_n$ for every n. Likewise for 8/27 and so on.

One way to see this is to notice that at each stage in the construction we are throwing away the middle open third of each interval. So we always keep any endpoints.

It might seem that the only points in the Cantor set are the endpoints of such intervals. *This is wrong!*

[10] We have drawn a very "fat" line for visual purposes. Of course, a line really has not "thickness".

Addresses of Points To better understand what points are in the Cantor set, let us go back and look again at the construction.

Every point x in the Cantor set C is either in the left interval $[0, 1/3]$ or the right interval $[2/3, 1]$.

For example, suppose $x \in C$ is in the right interval $[2/3, 1]$. Then either x is in the left interval $[6/7, 7/9]$ or the right interval $[8/9, 1]$.

Suppose $x \in C$ is in the left interval $[6/7, 7/9]$. Then either x is in the left interval $[18/27, 19/27]$ or the right interval $[20/27, 21/27]$.

Suppose $x \in C$ is in the left interval $[18/27, 19/27]$. Etc.

If we write L for left and R for right, then such a point x will be described by an infinite sequence of the form

$$RLL\ldots.$$

Every point in the Cantor set C can be described by an infinite sequence of L's and R's in this way. Moreover, *every* such infinite sequence describes a point in C.

For example, the infinite sequence

$$LRRLRLLRLRRLLLRLRLRLLLLRRRLLRLLRLLLRLRRRRRL\ldots$$

describes a point x which is in the interval $[0, 1]$, also in $[0, 1/3]$ (go Left), also in $[2/9, 3/9]$ (go Right), also in $[8/27, 9/27]$ (go Right), also in $[24/81, 25/81]$ (go Left), also in $[74/243, 75/243]$ (go Right), also in $[222/729, 223/729]$ (go Left), etc.

With this notation, left endpoints of intervals obtained in the construction of the C_n correspond to an infinite sequence ending in L's. Right endpoints of intervals obtained in the construction of the C_n correspond to an infinite sequence ending in R's.

For example, the point $8/9$ corresponds to $RRLLLLL\ldots$, since we went right twice and then forever stay left. In a similar way, $2/27$ corresponds to $LLRLLLLL\ldots$, since we went left twice and then right and then forever stay left. *Mark* $8/9$ *and* $2/27$ *in Figure 4.18.* ✍

Tree Representation A convenient way to represent infinite sequences consisting of the terms L and R is by means of an infinite tree which branches twice at each node. See Figure 4.19, where the first few sub branches are shown.

Each *infinite* branch corresponds to a point in the Cantor set and each point in the Cantor set corresponds to exactly one *infinite* branch.

Base 3 Representation of Points in C We know that every point in the interval $[0, 1]$ has a decimal expansion. We saw on page 82 that it also has a binary expansion.

Because of the following Theorem, we will be interested here in the base 3, or *ternary*, expansion of numbers $x \in [0, 1]$. For this we need just the numerals 0, 1 and 2. We then have the following result.

Theorem 4.3.1. *A number x in the interval $[0, 1]$ is in the Cantor set C if and only if it has a* ***ternary*** *expansion consisting just of 0s and 2s. Moreover, x will have* ***exactly one*** *ternary expansion which consists only of 0s and 2s.*

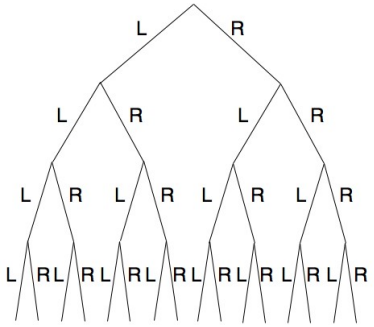


Figure 4.19: First few branches of a two branching tree

Proof. **Step A.** Suppose for example the ternary expansion of some $x \in [0,1]$ is

$$x = .0220202\ldots.$$

We will show that $x \in C$.

If we consider the intervals $[0,1/3]$, $[1/3,2/3]$ and $[2/3,1]$, because the first digit in the ternary expansion of x is 0, it follows that x is in the first of these three intervals, i.e. $[0,1/3]$. *Draw a diagram.* In particular, $x \in C_1$, see (4.8).

Subdividing $[0,1/3]$, consider the intervals $[0,1/9]$, $[1/9,2/9]$ and $[2/9,3/9]$. It follows from the ternary expansion for x, because the second digit is 2, that x is in the third of these, i.e. $[2/9,3/9]$. In particular, $x \in C_2$, see (4.9).

Subdividing $[2/9,3/9]$ we consider the intervals $[6/27,7/27]$, $[7/27,8/27]$ and $[8/27,9/27]$. It follows from the ternary expansion for x, because the third digit is 2, that x is in the third of these, i.e. $[8/27,9/27]$. In particular, $x \in C_3$.

Etc.

In this way we see that $x \in C_n$ for every n, and so $x \in C$. See (4.10)

The same argument shows for *any* $x \in [0,1]$, that if x has a ternary expansion using just the numerals 0 and 2 but not 1, then $x \in C$.

Step B. On the other hand, suppose $x \in C$. *We will now show that* x *has some ternary expansion which uses the numerals* 0 *and* 2 *but not* 1.

Since $x \in C$ it follows in particular that $x \in C_1$. This implies x is in either the interval $[0,1/3]$ or the interval $[2/3,1]$. (It might also be in the interval $[1/3,2/3]$, but then it must be one of the two endpoints of $[1/3,2/3]$, and so also is in either $[0,1/3]$ or $[2/3,1]$.) So the first digit in the ternary expansion of x can be taken to be 0 or 2.

We subdivide the corresponding interval [0, 1/3] or [2/3, 1] into three subintervals. Because $x \in C$, x is in either the first or third of these subintervals. This means the second digit of its ternary expansion can be taken to be 0 or 2.

We subdivide the relevant subinterval into three subsubintervals. Because $x \in C$, x is in either the first or third of these subsubintervals. This means the third digit of its ternary expansion can also be taken to be 0 or 2.

Etc.

In this way we see that if $x \in C$ then x has a ternary expansion with only 0s and 2s. I

Because we have just one choice, namely left or right interval, at each stage, there is moreover *exactly one* ternary expansion for $x \in C$ which consists just of the numerals 0 or 2. □

How Large is the Cantor Set?

Length Since the approximation C_n consists of 2^n intervals of length $(1/3)^n$, its total length is $(2/3)^n$.

We will not give a precise definition of the length of complicated sets like the Cantor set. But it is possible to do this. All we need here is the property that if $A \subset B$ then the length of A is $\leq$ the length of B.

Since $C \subset C_n$ (*why?*) it follows that the length of C is $\leq (2/3)^n$ for *every* n. So in the sense of "length", C is small.

Cardinality The Cantor set is certainly infinite. This is clear because we have an infinite number of choices to make using the L, R tree representation.

But is the Cantor set countable or uncountable? (The proof of the following Theorem is incomplete, in that it uses a theorem we have not actually proved. See the discussion within the proof itself.)

Theorem 4.3.2. *The Cantor set has cardinality c.*

"Proof". Let C denote the Cantor set. We will prove:

1. There is a one-to-one correspondence between C and some subset of $[0, 1]$.
2. There is a one-to-one correspondence between $[0, 1]$ and some subset of C.

It seems reasonable that from these two facts there should be a one-to-one correspondence between (all of) C and (all of) $[0, 1]$. This is indeed the case, but to show it one needs the Cantor-Schroeder-Bernstein Theorem on page 128, which we have not proved. For this reason I have written "*Proof*".

From Theorem 3.5.4 on page 124, the cardinality of $[0, 1]$ is c. So once we know there is a one-to-one correspondence between C and $[0, 1]$ it follows that C also has cardinality c.

The first fact (1.) is easy to prove, since C is a subset of $[0, 1]$. The one-to-one correspondence just sends each $x \in C$ to the same $x \in [0, 1]$.

For (2.) we use the fact from page 82 that every $x \in [0, 1]$ has at least one binary expansions of the form $.a_1a_2a_3 \ldots a_n \ldots$, where each a_i is either 0 or 1. For each $x \in [0, 1]$ choose one such binary expansion, replace each 0 by L and each 1 by R, and so get an infinite sequence of L's and R's which we call $s(x)$. (Notice that if $x \neq y$ then $s(x) \neq s(y)$, *why?*)

Using addresses of points in the Cantor set as on page 157 we can identify each $s(x)$ with a member of the Cantor set C, and in this way we get a one-to-one correspondence between $[0, 1]$ and a subset of C. *Why is it only a subset of C?*

This completes the "Proof" because of the comments above. □

Questions

1 In Figure 4.3 we begin with the line segment from $x = 0$ to $x = 1$ on the x-axis. We then sketch the first four approximations to the Koch curve (each "begins" at the point $x = 0$ on the x-axis and "ends" at the point $x = 1$).

What is the length of the first approximation to the Koch curve? How about the second, third, fourth, nth approximations?

2 For n a positive integer consider the "n-tooth" curve whose graph is a straight line from $(0, 0)$ to $(\frac{1}{2n}, \frac{1}{\sqrt{n}})$, a straight line from $(\frac{1}{2n}, \frac{1}{\sqrt{n}})$ to $(\frac{2}{2n}, 0)$, a straight line from $(\frac{2}{2n}, 0)$ to $(\frac{3}{2n}, \frac{1}{\sqrt{n}})$, a straight line from $(\frac{3}{2n}, \frac{1}{\sqrt{n}})$ to $(\frac{4}{2n}, 0)$, a straight line from $(\frac{4}{2n}, 0)$ to $(\frac{5}{2n}, \frac{1}{\sqrt{n}})$, a straight line from $(\frac{5}{2n}, \frac{1}{\sqrt{n}})$ to $(\frac{6}{2n}, 0)$, ..., finally ending up at $(\frac{2n}{2n}, 0) = (1, 0)$. There are thus n teeth. Why?

Draw a diagram for $n = 3$.

What is the length of the n-tooth curve? What happens to the length of the n-tooth curve as n approaches infinity?

What curve does the the n-tooth curve approach as n approaches infinity?

What is the length of this limit curve?

3 For the Koch curve constructed as in Question 1, what is the area between the x-axis and the first, second, third and fourth approximations? What about the nth approximation? What do you expect the area to be between the x-axis and the Koch curve?

4.4 Iterated Function Systems

New ways of looking at something can have surprising and useful applications.

Overview

This Section corresponds to
The material is developed h

We discuss the idea of an *Iterated Function System* or *IFS*, a very useful way of examining a large class of fractals.

In particular, the idea of an IFS leads to two different ways of generating fractals, the *Deterministic Algorithm* and the *Random Algorithm* or *Chaos Game*.

We discuss the IFS corresponding to the Sierpinski Triangle and the IFS corresponding to the Koch curve. But the ideas in these two cases can be generalised in a more or less straightforward way to any IFS.

A Little History The idea that many fractals can be characterised by an IFS and that such fractals can be generated by the deterministic algorithm was introduced and developed in a 1981 paper[11] of the author.

The idea of the chaos game to generate fractals was first developed by Barnsley and Demko[12] in 1985. A few years later Barnsley applied these ideas to image compression and was a founder of the company "Iterated Systems", at one stage valued at $200,000,000(US), later known as "Media Bin" and then acquired by "Interwoven".

What is an IFS?

[HM, 437–446]

Three Maps and the Sierpinski Triangle S If you look at the Sierpinski Triangle S in Figure 4.20 you can see that S is made up of three copies of itself each scaled by $1/2$.[13] We will denote these three copies by S_1, S_2 and S_3, where the vertex $P_1 = (0,0) \in S_1$, the vertex $P_2 = (1,0) \in S_2$ and the top vertex $P_3 = (1/2, \sqrt{3}/2) \in S_3$. *Indicate all this on Figure 4.20.*

Suppose that each point (x, y) in the plane $\mathbb{R}^2$ is moved closer to the point P_1 by the factor $1/2$. For example, $(3,0)$ is mapped to $(3/2, 0)$, $(1,1)$ is mapped to $(1/2, 1/2)$, $(0,3)$ is mapped to $(0, 3/2)$.

In this way, points in S are mapped to points in S. For example, $(1,0)$ is mapped to $(0, 0.5)$, $(3/4, \sqrt{3}/4)$ which is the midpoint of the "right edge" of S is mapped to $(3/8, \sqrt{3}/8)$ (*where is this on S?*) and $P_3 = (1/2, \sqrt{3}/2)$ is

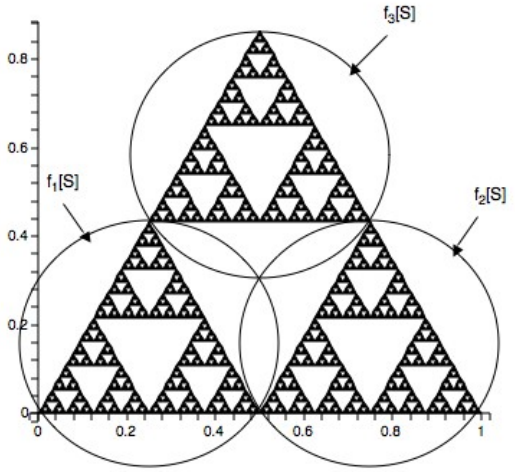


Figure 4.20: Sierpinski Triangle S

mapped to $(1/4, \sqrt{3}/4)$. (*where is this on S?*).

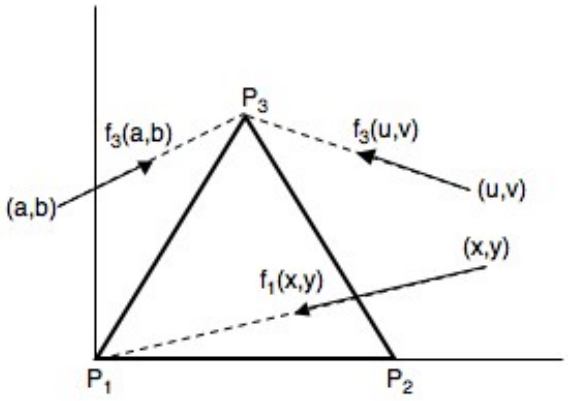


Figure 4.21: The maps f_1, f_2 and f_3.

The map (or function) we just described is the function $f_1 : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$f_1(x, y) = \left(\frac{x}{2}, \frac{y}{2}\right), \tag{4.11}$$

[11]Hutchinson, John E., *Fractals and self-similarity.* Indiana Univ. Math. J. **30** (1981), 713–747.

[12]Barnsley, M. F.; Demko, S. *Iterated function systems and the global construction of fractals.* Proc. Roy. Soc. London Ser. A **399** (1985), 243–275.

[13]Of course, as usual, we can only sketch an approximation to S.

see Figure 4.21.

Notice from Figure 4.20 that

$$S_1 = f_1[S],$$

where by the right side of the equality we mean the set of all points of the form $f_1(x, y)$ for $(x, y) \in S$. That is

$$f_1[S] = \{f_1(x, y) : (x, y) \in S\}.$$

We read this as "$f_1[S]$ is the set of points of the form $f_1(x, y)$ for some $(x, y) \in S$".

In a similar way,

$$S_2 = f_2[S] \quad \text{and} \quad S_3 = f_3[S],$$

where f_2 maps every point $(x, y) \in \mathbb{R}^2$ to the midpoint between (x, y) and P_2, and f_3 maps every point $(x, y) \in \mathbb{R}^2$ to the midpoint between (x, y) and P_3.

Why are the following formulae true? ✍

$$\begin{aligned} f_1(x, y) &= \left(\frac{x}{2}, \frac{y}{2}\right) \\ f_2(x, y) &= (1, 0) + \frac{1}{2}\Big((x, y) - (1, 0)\Big) = \left(\frac{x}{2} + \frac{1}{2}, \frac{y}{2}\right) \\ f_3(x, y) &= \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) + \frac{1}{2}\left((x, y) - \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)\right) = \left(\frac{x}{2} + \frac{1}{4}, \frac{y}{2} + \frac{\sqrt{3}}{4}\right). \end{aligned} \tag{4.12}$$

If you know about matrices and column vectors, then f_1, f_2 and f_3 can also be conveniently described that way.[14]

The IFS for S We have seen that for the Sierpinski triangle S,

$$S = S_1 \cup S_2 \cup S_3$$

where

$$S_1 = f_1[S],\ S_2 = f_2[S],\ S_3 = f_3[S].$$

So we have the important relation

$$\boxed{S = f_1[S] \cup f_2[S] \cup f_3[S]} \tag{4.13}$$

Definition 4.4.1. The set of maps $\mathcal{F} = \{f_1, f_2, f_3\}$ with f_1, f_2 and f_3 are as in (4.12) is called the *Iterated Function System* or *IFS* corresponding to the Sierpinski Triangle S.

Because of (4.13) we say that S is *invariant* under the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$.

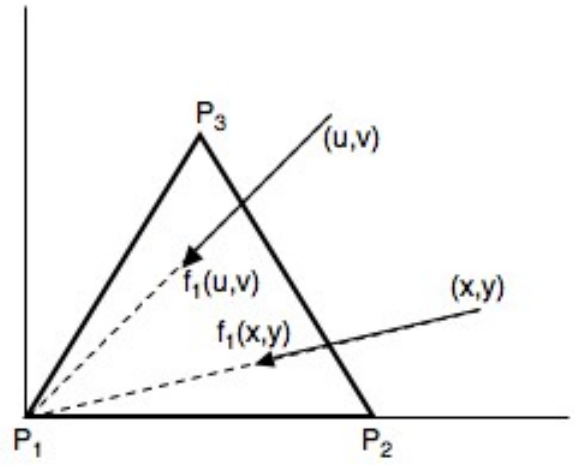


Figure 4.22: f_1 is contractive

Contractive Maps Take two initial points $\mathbf{x} = (x, y)$ and $\mathbf{y} = (u, v)$ and apply the map f_1 to each. The image points are $f_1(x, y) = (x/2, y/2)$ and $f_1(u, v) = (u/2, v/2)$ respectively, see (4.11).

The distance[15] between the two image points is exactly $1/2$ the distance between the two initial points. This is essentially because of the factor $1/2$ in the definition of f_1, see (4.11).

If you know about vectors, then you will see that the vector from (x, y) to (u, v) can be written as $(u-x, v-y)$. The vector from $(x/2, y/2)$ to $(u/2, v/2)$ is $(u/2 - x/2, v/2 - y/2) = \frac{1}{2}(u - x, v - y)$, which is exactly half the length of the vector $(u - x, v - y)$ from (x, y) to (u, v).

You could also use the formula for computing *the distance* d between points. For example,

$$\begin{aligned} d\big(f_1(x,y), f_1(u,v)\big) &= d\big((x/2, y/2), (u/2, v/2)\big) \\ &= \sqrt{(x/2 - u/2)^2 + (y/2 - v/2)^2} \\ &= \frac{1}{2}\sqrt{(x-u)^2 + (y-v)^2} \\ &= \frac{1}{2} d\big((x,y),(u,v)\big) \end{aligned} \tag{4.14}$$

In the same way we see that the maps f_2 and f_3 also reduce the distance between points by the factor $1/2$.

[14] If we represent points in the plane by column vectors then it follows from (4.12) that

$$f_1\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}, \quad f_2\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix}, \quad f_3\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \frac{1}{4} \\ \frac{\sqrt{3}}{4} \end{bmatrix}.$$

[15] The distance between $x, y \in \mathbb{R}$ is $d(x, y) = |x - y|$.
The distance between $\mathbf{x} = (x_1, x_2)$, $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$ is $d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$.
The distance between $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{y} = (y_1, y_2, y_3) \in \mathbb{R}^3$ is $d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$.

Definition 4.4.2. A function $f : \mathbb{R} \to \mathbb{R}$, $f : \mathbb{R}^2 \to \mathbb{R}^2$ or $f : \mathbb{R}^3 \to \mathbb{R}^3$, is *contractive* if there is a number r with $0 \le r < 1$ such that

$$d\big(f(\boldsymbol{x}), f(\boldsymbol{u})\big) \le r d(\boldsymbol{x}, \boldsymbol{u})$$

for all points $\boldsymbol{x}$ and $\boldsymbol{u}$ in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$ respectively.

The number r is called a *contractivity factor*[16] for f.

For example, it follows from (4.14) that f_1 is contractive with contractivity factor $1/2$. Similarly, f_2 and f_3 are also contractive with contractivity factor $1/2$.

It will generally be the case, at least in this course, that the maps in an IFS are all contractive.

The Deterministic Algorithm

[HM, 437–446]

The IFS Determines S *A surprising and very important fact is that from just knowing the IFS* $\mathcal{F} = \{f_1, f_2, f_3\}$ *in Definition 4.4.1 we can find* S.

To see this, begin with *any* set (picture) E, such as the face in Figure 4.23, and apply the IFS $\mathcal{F}$ to get a new set (picture)

$$E_1 = \mathcal{F}(E) = f_1[E] \cup f_2[E] \cup f_3[E].$$

So E_1 consists of 3 little faces.

Next apply $\mathcal{F}$ to E_1 to get a new set (picture) E_2 consisting of 9 smaller faces.

$$E_2 = \mathcal{F}(E_1) = f_1[E_1] \cup f_2[E_1] \cup f_3[E_1].$$

Next apply $\mathcal{F}$ to E_2 to get a new set (picture) E_3 consisting of 27 smaller faces.

$$E_3 = \mathcal{F}(E_2) = f_1[E_2] \cup f_2[E_2] \cup f_3[E_2].$$

And so on.

In the limit we obtain the Sierpinski Triangle S, no matter what set E we start from. See Theorem 4.4.3.

For any set E we defined the set

$$\mathcal{F}(E) = f_1[E] \cup f_2[E] \cup f_3[E]. \tag{4.15}$$

We have the following result.

Theorem 4.4.3. *Consider the IFS* $\mathcal{F} = \{f_1, f_2, f_3\}$ *in Definition 4.4.1. Suppose* E *is any closed*[17] *and bounded*[18] *set.*

[16] Notice that if r is a contractivity factor then so is any number larger than r. One can show there is always a smallest contractivity factor, and then this is usually called *the* contractivity factor for f.

[17] A set E is *closed* if it contains all its boundary points. This is discussed in more detail on page 171.

[18] A set $E \subset \mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$ is *bounded* if there is some number M such that the distance from every point in E to the origin is at most M. For example, the Sierpinski triangle and the Koch curve and the set of points inside any disc, are all bounded. But the entire plane $\mathbb{R}^2$ is not bounded, nor is the set of all points (x, y) in the plane for which $x \ge 0$ and $y \ge 0$.

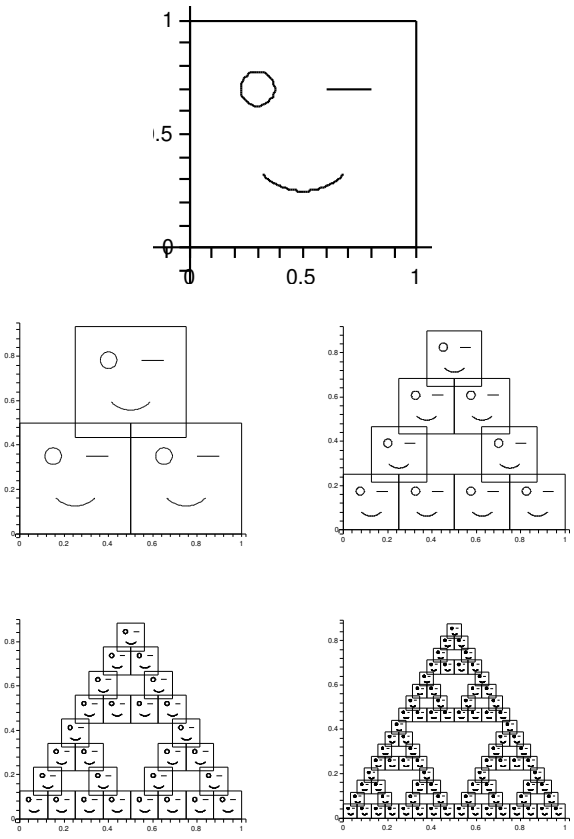


Figure 4.23: A sequence of sets $E, E_1, E_2, E_3, E_4, \ldots$, beginning with a face and obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$, converging to the Sierpinski Triangle S.

Then the sequence of sets

$$E,\ E_1 = \mathcal{F}(E),\ E_2 = \mathcal{F}(E_1),\ E_3 = \mathcal{F}(E_2),\ \ldots,\ E_n = \mathcal{F}(E_{n-1}), \ldots \quad (4.16)$$

converges to S, independent of the starting set E.

Note: If we take $E = \mathbb{R}^2$ in the statement of the Theorem, then every set in the sequence (4.16) is $\mathbb{R}^2$. *Why?* So the sequence of sets in this case will not converge to S. *Why does the theorem not apply in this case?*

"Proof". We cannot give a complete and rigorous proof, as we have not defined what we mean by the limit of a sequence of sets.

Also, to make the following rigorous requires the filling in of quite a few details about converging sequences of sets.

But I will describe the essential parts of the argument.

Somewhat informally, by saying the set S is the limit of the sequence of sets (4.16), i.e. the sequence of sets converges to S, we mean that for *every*

positive number ϵ,[19] which we can think of as a very small "tolerance", the following is true:

There is an integer N depending on ϵ such that if $n \geq N$ then

1. for every $x \in E_n$ there is some $y \in S$ within distance ϵ of x,[20] and
2. for every $y \in S$ there some $x \in E_n$ within distance ϵ of y.[21]

From the way we originally defined the Sierpinski Triangle S beginning on page 152, we know S is the limit of the sequence

$$T,\ T_1 = \mathcal{F}(T),\ T_2 = \mathcal{F}(T_1),\ T_3 = \mathcal{F}(T_2),\ \ldots,\ T_n = \mathcal{F}(T_{n-1}), \ldots\ , \quad (4.17)$$

where T is the black equilateral triangle as in Figure 4.24.

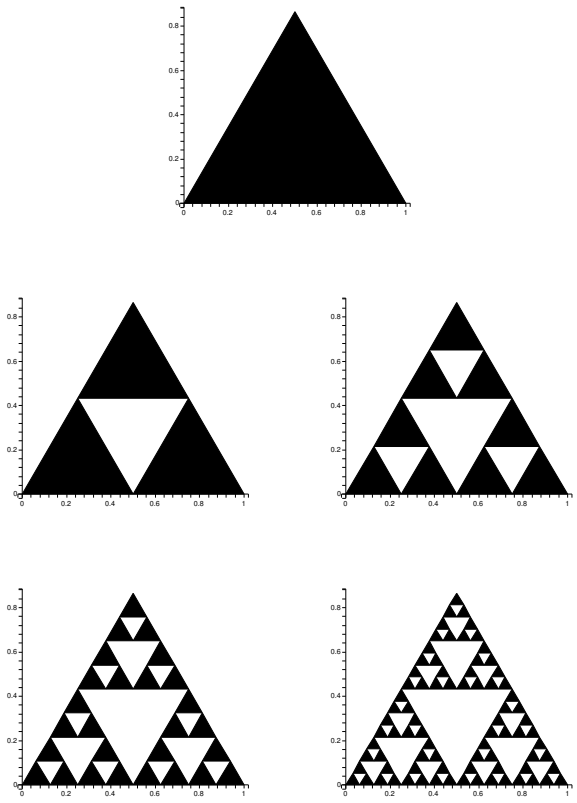


Figure 4.24: A sequence of sets $T, T_1, T_2, T_3, T_4, \ldots$, beginning with a triangle and obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3\}$, converging to the Sierpinski Triangle S.

There is certainly some positive real number, let us call it α, such that

[19] In Mathematics we commonly use the Greek letter ϵ, called "epsilon", to represent a number which is small and positive.

[20] The point y depends on x, on ϵ and on n.

[21] The point x depends on y, on ϵ and on n.

1. every point in the triangle T is within distance α of some point in the set E, and
2. every point in the set E is within distance α of some point in the triangle T.

For example, $\alpha = 1$ would do if E is the face in Figure 4.23. In fact smaller α will also work, but that does not make a difference to the following proof. *What is a smaller α that works?*

Because the functions f_1, f_2 and f_3 contract distances by 1/2,
1. every point in T_1 is within distance $\alpha/2$ of some point in E_1, and
2. every point in E_1 is within distance $\alpha/2$ of some point in T_1.

Again because the functions f_1, f_2 and f_3 contract distances by 1/2,
1. every point in T_2 is within distance $\alpha/4$ of some point in E_2, and
2. every point in E_2 is within distance $\alpha/4$ of some point in T_2.

Again because the functions f_1, f_2 and f_3 contract distances by 1/2,
1. every point in T_3 is within distance $\alpha/8$ of some point in E_3, and
2. every point in E_3 is within distance $\alpha/8$ of some point in T_3.

Etc.

Beginning on page 152 we essentially saw that the sequence (4.17) converges to S, in fact this was essentially how we defined S. We also have just seen that the sets in the sequence (4.16) are getting closer and closer to the sets in the sequence (4.17). It follows that the sets in the sequence (4.16) also converge to S.

This argument did not depend on the initial set E. For different E we will get a different α, but nothing else changes. □

Where in the proof did we use the fact that E was bounded?

Deterministic Algorithm for Generating S This is what we have just discussed. Begin with any set E and take the sequence (4.16). This will give better and better approximations to S.

The terminology "deterministic algorithm" is used to distinguish this algorithm from the "random algorithm" or "chaos game" discussed on page 178.

There is a nice java applet at
`www.geom.uiuc.edu/java/IFSoft/IFSs/welcome.html#findingattractors`
Scroll down to the blue window, draw your own face, and use the Iterate button to step through successive iterations. Note that the Sierpinski Triangle there, and hence the three functions used, are a little different from the example we have just been discussing.

The Koch Curve K and its IFS The Koch Curve K and its first approximation is shown in Figure 4.25. See also page 151.

The four line segments in the first approximation each have length 1/3. The second and third line segments form an equilateral triangle with the x-axis. From this is is easy to check that the vertices of the five corners are

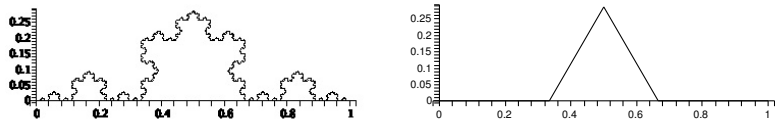


Figure 4.25: Koch Curve and its first approximation

$P_1 = (0,0)$, $P_2 = (1/3,0)$, $Q = (1/2,\sqrt{3}/6)$, $P_3 = (2/3,0)$ and $P_4 = (1,0)$. *Check it!*

The Koch Curve K can be written as the union of 4 scaled copies of itself each scaled by $1/3$.

$$K = K_1 \cup K_2 \cup K_3 \cup K_4 = f_1[K] \cup f_2[K] \cup f_3[K] \cup f_4[K]. \tag{4.18}$$

Here K_1 is the left "quarter" joining the points $(0,0)$ and $(0,1/3)$, K_2 joins $(0,1/3)$ and $(1/2,\sqrt{3}/6)$, K_3 joins $(1/2,\sqrt{3}/6)$ and $(2/3,0)$, K_4 joins $(2/3,0)$ and $(1,0)$.

Geometrically:

1. f_1 contracts points towards $(0,0)$ with contraction ratio $1/3$;
2. f_2 contracts points towards $(0,0)$ with contraction ratio $1/3$, then rotates anti clockwise, i.e. in the "positive" direction, by $60°$ or equivalently $\pi/3$ radians, and finally translates in the x-direction by $1/3$;
3. f_3 contracts points towards $(0,0)$ with contraction ratio $1/3$, then rotates by $-60°$ or equivalently $-\pi/3$ radians, and finally translates the origin $(0,0)$ to $Q = (1/2,\sqrt{3}/6)$;
4. f_4 contracts points towards $(0,0)$ with contraction ratio $1/3$ and then translates in the x-direction by $2/3$.

The formulae for $f_1, \ldots f_4$ are:

$$\begin{aligned} f_1(x,y) &= (x/3, y/3), \\ f_2(x,y) &= (x/6 - \sqrt{3}y/6 + 1/3, \sqrt{3}x/6 + y/6), \\ f_3(x,y) &= (x/6 + \sqrt{3}y/6 + 1/2, -\sqrt{3}x/6 + y/6 + \sqrt{3}/6), \\ f_4(x,y) &= (x/3 + 2/3, y/3). \end{aligned} \tag{4.19}$$

If you know a little about matrices and how to represent rotations by matrices, the geometric descriptions will allow you to compute the functions $f_1, \ldots, f_4$.[22]

[22] For example, from the description of f_2, the point (x,y) is first sent to $(x/3, y/3)$. Since a rotation by θ radians is represented by the matrix $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, it follows that $(x/3, y/3)$ is then sent to

$$\begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{bmatrix} \begin{bmatrix} x/3 \\ y/3 \end{bmatrix} = \begin{bmatrix} x/6 - \sqrt{3}y/6 \\ \sqrt{3}x/6 + y/6 \end{bmatrix}.$$

Finally, translation by $1/3$ in the x-direction adds $1/3$ to the first coordinate. This gives the formula for $f_2(x,y)$.

Use a similar argument to find the formulae for $f_4(x,y)$ and $f_3(x,y)$.

The IFS corresponding to K is

$$\mathcal{F} = \{f_1, f_2, f_3, f_4\}, \tag{4.20}$$

where $f_1, \ldots, f_4$ are as in (4.19).

The contractivity factor for the maps $f_1, \ldots, f_4$ is $1/3$. *Why?*

If E is *any* set then we define

$$\mathcal{F}(E) = f_1[E] \cup f_2[E] \cup f_3[E] \cup f_4[E].$$

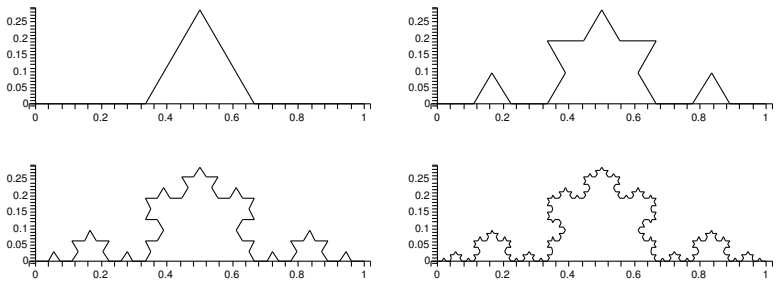


Figure 4.26: A sequence of sets which began with a line segment, obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$ in (4.20), converging to the Koch Curve K.

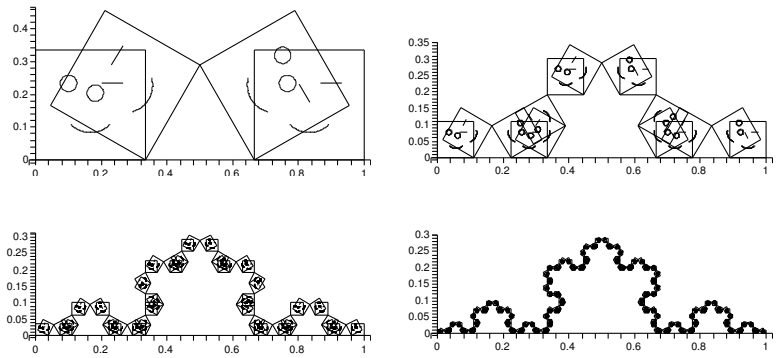


Figure 4.27: A sequence of sets which began with a face, obtained by repeatedly applying the IFS $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$, converging to the Koch Curve K.

Just as in Theorem 4.4.3 for the Sierpinski Triangle, there is a similar theorem for the Koch Curve.

Theorem 4.4.4. *Consider the IFS $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$ in (4.20). Then the sequence of sets*

$$E,\ E_1 = \mathcal{F}(E),\ E_2 = \mathcal{F}(E_1),\ E_3 = \mathcal{F}(E_2),\ \ldots,\ E_n = \mathcal{F}(E_{n-1}), \ldots \tag{4.21}$$

converges to K, independent of the starting set E, provided E is bounded.

"Proof". The proof is very similar to that for the Sierpinski Triangle. The only significant difference is that the contractivity factor here for the maps $f_1, \ldots, f_4$ is $1/3$ instead of $1/2$, as was the case for the Sierpinski Triangle IFS. □

The General IFS Theorem We have the following very general result, Theorem 4.4.5. It extends Theorems 4.4.3 and 4.4.4.

In order to give a precise statement we need the idea of a *closed* and *bounded* set. We have already seen in Footnote 18 what it means for a set to be bounded.

A set E is said to be *closed* if it contains all its boundary points.

For example, the interval $(2, 3) \subset \mathbb{R}$ is not closed because it does not contain its boundary points 2 and 3. However, the interval $[2, 3]$ *is* closed because it contains its boundary points. The interval $[2, 3)$ is not closed. *Why?* ✍

The set A of points $(x, y) \in \mathbb{R}^2$ such that $x^2 + y^2 < 1$, is not closed. It does not contain its boundary points, which are the points (x, y) on the circle given by $x^2 + y^2 = 1$. On the other hand, the set B of points $(x, y) \in \mathbb{R}^2$ such that $x^2 + y^2 \leq 1$ *is* closed, because it *does* contain its boundary points, which as for A are the points (x, y) on the circle $x^2 + y^2 = 1$.

Here[23] is a precise definition of "boundary point".

All the fractal sets we discuss are closed and bounded.

Theorem 4.4.5. *Suppose $\mathcal{F}$ is any IFS consisting of maps $f_1, \ldots, f_n$ such that each $f_i : \mathbb{R}^2 \to \mathbb{R}^2$ has a contractivity factor less than 1.*

Then there is exactly one closed bounded set A, depending on $\mathcal{F}$, such that

$$A = \mathcal{F}(A), \quad \textit{i.e. } A = f_1[A] \cup \cdots \cup f_n[A].$$

Moreover, beginning from any closed bounded set E, the sequence of sets

$$E,\ E_1 = \mathcal{F}(E),\ E_2 = \mathcal{F}(E_1),\ E_3 = \mathcal{F}(E_2),\ \ldots,\ E_n = \mathcal{F}(E_{n-1}), \ldots$$

converges to the set A.

"Proof". In Theorems 4.4.3 and 4.4.4 we knew the set A to which the sequence converged, namely the Sierpinski Triangle or the Koch Curve respectively.

Here the problem is that we may not know beforehand what A is. However, this problem can be overcome. Just as in the previous two Theorems, essentially because all the maps f_i are contractions, if we start from different sets E and F and repeatedly apply $\mathcal{F}$ to get

$$E,\ E_1 = \mathcal{F}(E),\ E_2 = \mathcal{F}(E_1),\ E_3 = \mathcal{F}(E_2),\ \ldots,\ E_n = \mathcal{F}(E_{n-1}), \ldots$$
$$F,\ F_1 = \mathcal{F}(F),\ F_2 = \mathcal{F}(F_1),\ F_3 = \mathcal{F}(F_2),\ \ldots,\ F_n = \mathcal{F}(F_{n-1}), \ldots$$

[23] Suppose $A \subset \mathbb{R}^2$ (or $\mathbb{R}$ or $\mathbb{R}^3$). A point $x \in \mathbb{R}^2$ is a *boundary point* of A if for every $\epsilon > 0$ there is at least one point in A within distance ϵ of x, and at least one point *not* in A within distance ϵ of x. *A sketch will help you understand this definition.*

✍ *Why does this give the same boundary points for the sets $(2, 3), [2, 3], [2, 3) \subset \mathbb{R}$ and for the sets $A, B \subset \mathbb{R}^2$ as discussed in the previous few paragraphs?*

then the sets E_n and F_n will get closer and closer together as $n \to \infty$.

Using this information, one can show the two sequences of sets converge to some set A and that A does not depend of whether we begin with E or F or any indeed other closed bounded set.

We saw in the second paragraph of Footnote 18 why we work with bounded sets. One reason we also want A to be closed is in order to have a *unique* limit for the previous sequences of sets. Here[24] is an example of the sort of thing that can happen otherwise.

Filling in the details of the proof requires some more background than we have developed in this course. But hopefully the above discussion makes the result plausible. □

The Collage Method In [HM, 437–439] there is a discussion of a "collage process" for generating fractals.[25] This is, as here, what is normally called the "deterministic algorithm".

There is something else usually called the "collage method". It is the basis for a method of compressing images based on iterated function systems. The key idea is that an image or picture is split my means of a fixed grid into perhaps thousands of small squares. For each small square S in the picture, the square of twice the size which most looks like a scaled up copy of S (perhaps after rotations or reflections) is found. Then the functions that actually do the scaling are stored in the computer as a type of very large IFS. When needed, the image is rapidly reconstructed from this IFS. For many years in the 1990's the images on Microsoft's online encyclopedia Encarta where stored in this manner.

More Examples The site `www.math.utah.edu/~korevaar/fractals/` has many examples of fractals given by IFS's. Figures (4.28), (4.29) and (4.30) are from this site.

In each case the functions $f_1, f_2, f_3, \ldots$ in the corresponding IFS $\mathcal{F}$ are the unique affine[26] functions of the form

$$f_i(x, y) = (a_i x + b_i y + c_i,\ d_i x + e_i y + f_i) \tag{4.22}$$

which map the *standard unit square* with vertices $(0, 1), (0, 0), (1, 0), (1, 1)$ into the various parallelograms shown. The $\mathbb{L}$ in each parallelogram indicates the

[24] Consider the sequence of sets

$$[1/3, 2/3],\ [1/4, 3/4],\ [1/5, 4/5], [1/6, 5/6],\ \ldots.$$

According to the Definition of limit of of a sequence of sets used in the "Proof" of Theorem 4.4.3, this sequence converges to the set $(0, 1)$ as well as to the set $[0, 1]$. But it is possible to prove that there is only one limit which is closed.

[25] On p438 of [HM] the image next to the one captioned "3rd stage" should be captioned "first stage", and the image at the bottom right of p439 should be captioned "2nd stage".

[26] An *affine function* is the same as a linear transformation followed by a translation. In (4.22) we can write

$$f_i\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} a_i & b_i \\ d_i & e_i \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} c_i \\ f_i \end{bmatrix}.$$

So f_i is the same as first applying $\begin{bmatrix} a_i & b_i \\ d_i & e_i \end{bmatrix}$ and then translating by $\begin{bmatrix} c_i \\ f_i \end{bmatrix}$.

manner in which an $\mathbb{L}$ in the standard unit square would need to be transformed in order to move that square into the given parallelogram.

For each function f_i there are 6 numbers $a_i, b_i, c_i, d_i, e_i, f_i$ to be found. But a parallelogram is determined by any three of its vertices, and because of the $\mathbb{L}$ we know which of the 3 square vertices $(0,1)$, $(0,0)$, $(1,0)$ from the standard unit square map to which of the parallelogram vertices. There are two equations to be satisfied for each vertex, and so there are a total of 6 equations to solve.[27] Since there are 6 unknowns $a_i, b_i, c_i, d_i, e_i, f_i$ we expect there is a unique solution for $a_i, b_i, c_i, d_i, e_i, f_i$, and in this setting that is indeed the case.

Another site with material you might like to look at is
`http://classes.yale.edu/fractals/index.html` .

In [HM, 440–441] see the Barnsley fern, which is an aesthetically pleasing and somewhat realistic example of a fractal set in nature.

Chaos Game

Sierpinski Addresses We saw on page 157 that every point on the Cantor set has a unique address given by an infinite sequence of L's and R's, such as

$$RLL\ldots\ .$$

Instead of R and L we might use the symbols 0 and 2 (see Theorem 4.3.1), or 1 and 2, etc., depending upon the application.

In a similar way, every point on the Sierpinski Triangle S has an address given by an infinite sequence of numbers from the set $\{1,2,3\}$. But the address is not always unique as we will soon see.

Look at Figure 4.31. Suppose a point $x \in S$ has an address of the form $3132122311\ldots$. At each level, 1 corresponds to the bottom left third, 2 to the bottom right third and 3 to the top third. You might think of the first digit as giving the continent to which x belongs, the second digit as giving the country, the third digit giving the state, the fourth giving the city, the fifth giving the

[27] For examples, suppose that the $\mathbb{L}$ in the standard unit square and the image of $\mathbb{L}$ in some parallelogram tell us that the vertices $(0,1),(0,0),(1,0)$ are mapped by f_1 in (4.22) to $(1,3),(2,5),(4,6)$ respectively.

✍ *Draw a diagram.*

Then the 6 equations are

$$\begin{aligned} 1 &= b_1 + c_1, & 3 &= e_1 + f_1, \\ 2 &= c_1, & 5 &= f_1, \\ 4 &= a_1 + c_1, & 6 &= d_1 + f_1. \end{aligned}$$

This gives $a_1 = 2$, $b_1 = -1$, $c_1 = 2$, $d_1 = 1$, $e_1 = -2$ and $f_1 = 5$. So

$$f_i(x,y) = (2x - y + 2,\ x - 2y + 5).$$

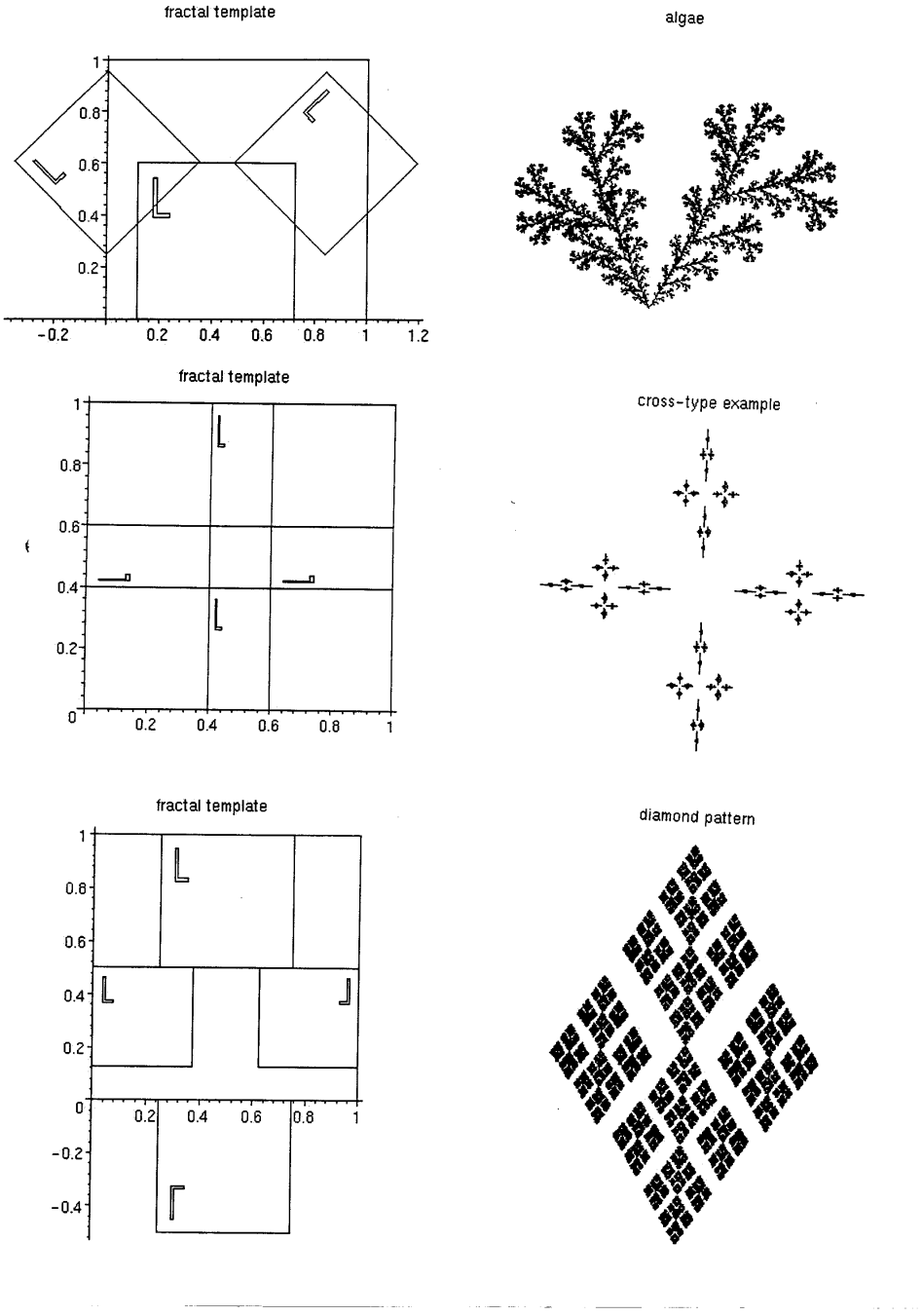


Figure 4.28: IFS Fractals from `www.math.utah.edu/ korevaar/fractals/`.

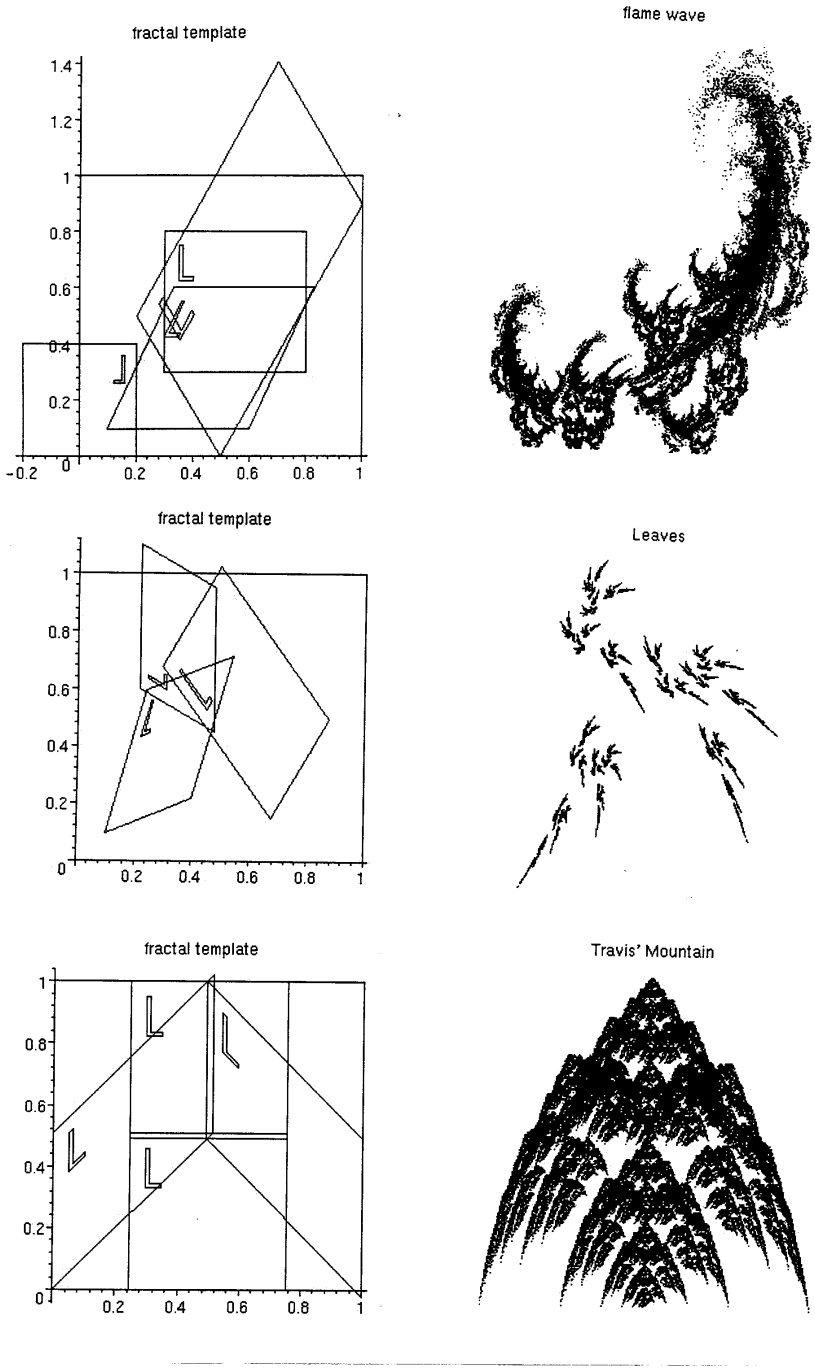


Figure 4.29: More IFS Fractals.

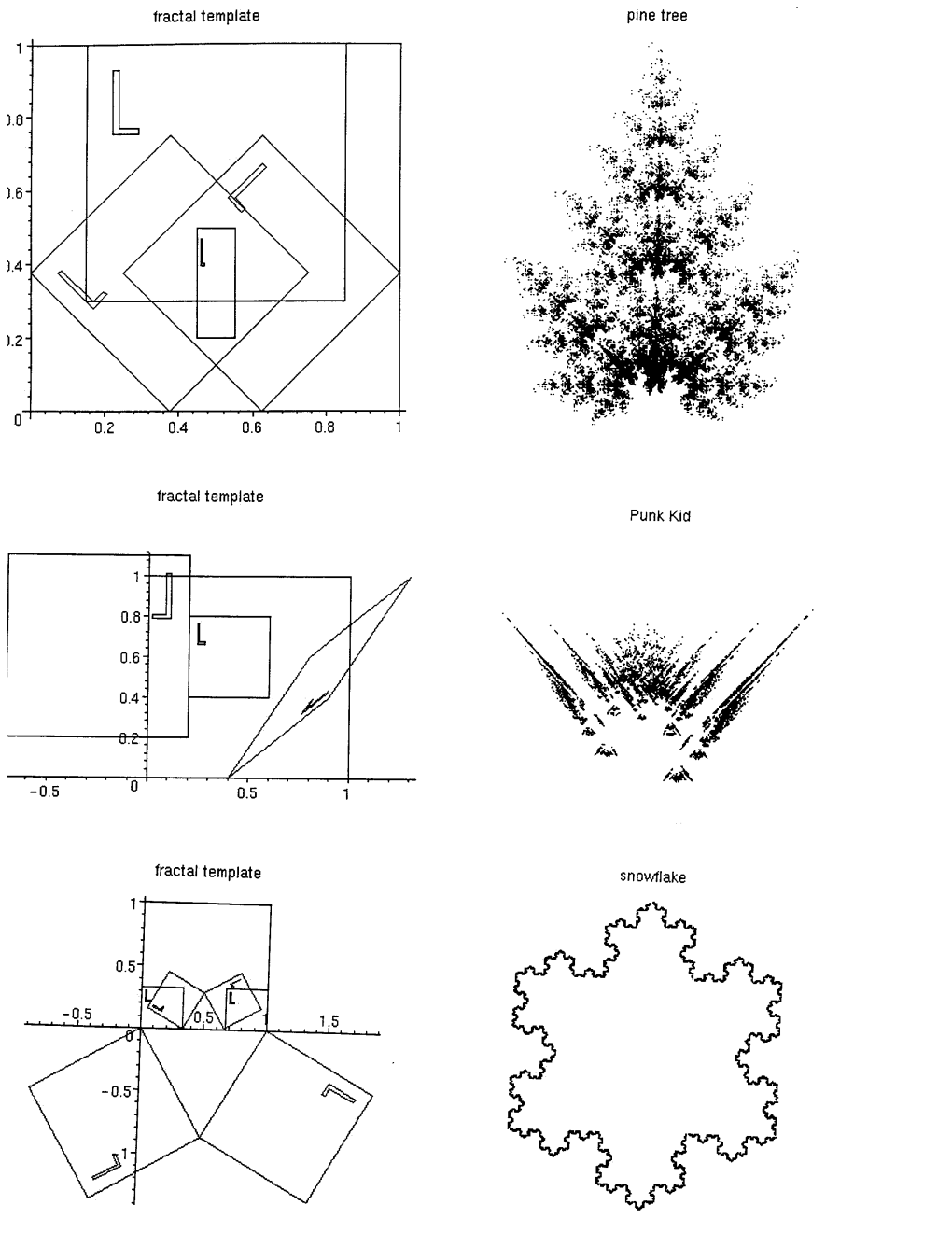


Figure 4.30: And More.

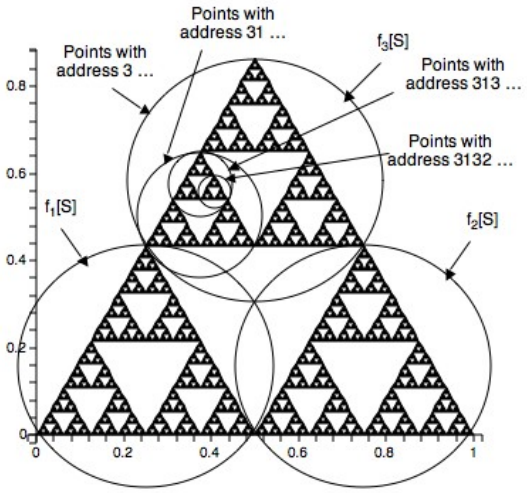


Figure 4.31: Addresses on the Sierpinski triangle S

suburb, the sixth giving the street, the seventh giving the house, the eighth giving the room, etc.

Since x has address $3132122311\ldots$ it is in the top continent, then within that continent it is in the bottom left country, then within that country it is in the top state, then within that state it is in the bottom right city, etc.

The point $(1/2, 0) \in S$ has the address $1222\ldots$, which we write as $1\overline{2}$ and read as "1 then 2 repeating". It also has the address $2\overline{1}$. *Where is the point with address* $31\overline{3}$ *and what is its other address?*[28]

Addresses and Maps Suppose we have a point $x \in S$ with an address $213131221\ldots$, for example. What is the address of $f_1(x)$?

Since f_1 sends every point in S into the bottom left third of S, the address of $f_1(x)$ will start with 1. Moreover, since x was previously in the bottom right third of S, it will move to the bottom right third *of* the bottom left third of S. So the address of $f(x)$ will start with 12. Similarly, the address of $f(x)$ will start with 121. Etc.

In fact, a similar argument shows that the address of $f_1(x)$ is precisely the address as for x but with 1 placed in front and every digit of x moved one place to the right.

Similarly for applying f_2 and f_3.

[28] In fact points with two addresses are those of the form "blah blah"$a\overline{b}$ where $a \neq b$. The other address is then "same blah blah"$b\overline{a}$.

For example, in terms of addresses

$$f_1(332113\ldots) = 1332113\ldots,$$
$$f_2(332113\ldots) = 2332113\ldots,$$
$$f_3(332113\ldots) = 3332113\ldots\ .$$

The Chaos Game Method This is a very useful, and initially surprising, way to generate S.

Begin with any point $x_0 \in \mathbb{R}^2$, sometimes called *the seed*.

1. With probability 1/3 in each case, apply either f_1, f_2 or f_3 to x_0 and denote the result by x_1.
2. Independently of what has already happened and with probability 1/3 in each case, apply either f_1, f_2 or f_3 to x_1 and denote the result by x_2.
3. Independently of what has already happened and with probability 1/3 in each case, apply either f_1, f_2 or f_3 to x_2 and denote the result by x_3.
4. Independently of what has already happened and with probability 1/3 in each case, apply either f_1, f_2 or f_3 to x_3 and denote the result by x_4.
5. Etc.

In this way we obtain a potentially infinite sequence of points

$$x_0, x_1, x_2, x_3, x_4, \ldots, x_n, \ldots\ . \tag{4.23}$$

We could do this by throwing a standard dice at each stage and agree that if 1 or 4 is thrown then one applies the function f_1, if 2 or 5 is thrown then one applies the function f_2, and if 3 or 6 is thrown then one applies the function f_3.

Geometrically: $f_1(x)$ is exactly half way from the point x to the point $P_1 = (0,0)$, $f_2(x)$ is exactly half way from the point x to the point $P_2 = (1,0)$, $f_3(x)$ is exactly half way from the point x to the point $P_3 = (1/2, \sqrt{3}/2)$.

Each time we run a sequence of trials we will get a new sequence of points in (4.23).

For each n let A_n be the first n points in the sequence (4.23). For example,

$$A_{1057} = \{x_0, x_2, x_2, \ldots, x_n, \ldots, x_{1056}\}$$

The amazing result is that if n is large, such as $n = 1057$, then A_n is always an extremely good approximation to the Sierpinski Triangle.

We will make this more precise in Theorem 4.4.6. But meanwhile, here are some computer experiment results.

Now check out "Barnsley Fern and Fractal Gardens" under "6. Chaos and Fractals" from the CD in *The Heart of Mathematics*.

There is a nice java applet at
`http://math.bu.edu/DYSYS/applets/fractalina.html`.
Initially I suggest you try the 6 preset IFS's. Just select one and press the "Start" button. An explanation "More about Fractalina" is at the top of the web page.

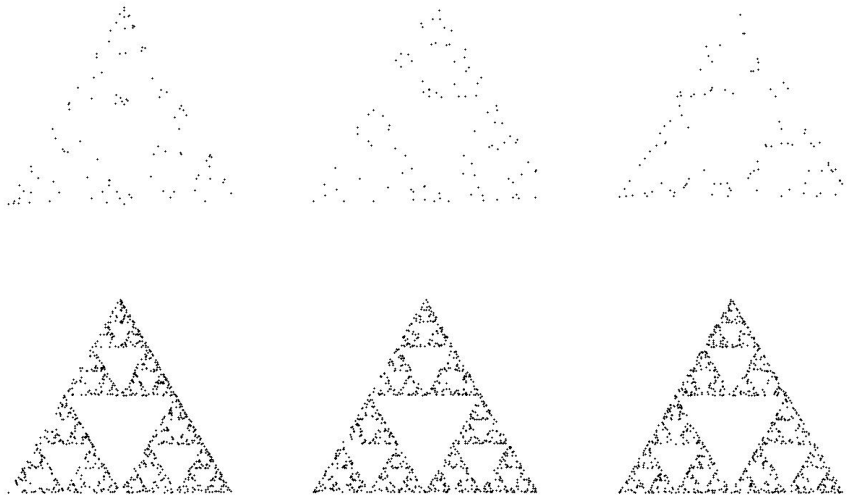


Figure 4.32: Three examples of 100 points obtained from the Chaos Game and three examples of 1000 points.

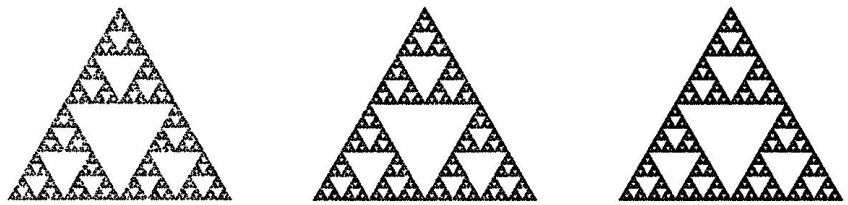


Figure 4.33: The Chaos Game with 5000 points, 10 000 points and 20 000 points.

The following Theorem is (at first) rather amazing. In Figures 4.32 and 4.33 there are examples of $A_{100}, A_{1000}, A_{5000}, A_{10\,000}, A_{20\,000}$ from different runs of the Chaos Game.

We do not have the tools to really state the Theorem precisely, but the main ideas are hopefully reasonably clear. So do not worry too much if it is a bit vague/confusing!

Theorem 4.4.6. *Let x_0 be the initial point in the Chaos Game for the Sierpinski Triangle S. Consider any run of the Chaos Game and for each n let A_n be the set consisting of the first n points obtained.*

If $x_0 \in S$ then with probability one the sequence of sets

$$A_0, A_1, A_2, \ldots, A_n, \ldots,$$

converges to S.

Even if $x_0 \notin S$, for any given "tolerance", with probability one and by ignoring sufficiently many initial points (in practice, say 100), the sets A_n will all eventually be closer to S than the given tolerance.

"Proof". First suppose that $x_0 \in S$. Then x_0 will have some address

$$a_1 a_2 a_3 \dots a_k \dots .$$

Suppose we run the Chaos Game and the first functions chosen were, for example,

$$f_3, f_1, f_2, f_3, f_2, f_2, f_1, f_3, \dots ,$$

in that order.

By the discussion "Addresses and Maps" on page 177, the points obtained from the Chaos Game will have addresses as follows:

$$\begin{aligned}
x_0 &= a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 \dots \\
x_1 = f_3(x_0) &= 3 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 \dots \\
x_2 = f_1(x_1) &= 13 a_1 a_2 a_3 a_4 a_5 a_6 a_7 \dots \\
x_3 = f_2(x_2) &= 213 a_1 a_2 a_3 a_4 a_5 a_6 \dots \\
x_4 = f_3(x_3) &= 3213 a_1 a_2 a_3 a_4 a_5 \dots \\
x_5 = f_2(x_2) &= 23213 a_1 a_2 a_3 a_4 \dots \\
x_6 = f_2(x_2) &= 223213 a_1 a_2 a_3 \dots \\
x_7 = f_1(x_2) &= 1223213 a_1 a_2 \dots \\
x_8 = f_3(x_2) &= 31223213 a_1 \dots \\
&\vdots
\end{aligned}$$

For the set of points $A_n = \{x_0, x_1, x_2, \dots, x_{n-1}\}$ with n large, approximately 1/3 of the addresses will begin with 1, approximately 1/3 will begin with 2, and approximately 1/3 will begin with 2. Equivalently, approximately 1/3 of the points in A_n will be in each of the three Sierpinski "subtriangles" S_1, S_2, S_3, see Figure 4.20.

The larger n is, the closer we will get to 1/3.

The 9 "level 2" Sierpinski subtriangles for S correspond to addresses beginning with 11, 12, 13, 21, 22, 23, 31, 32 and 33 respectively. So for large n, approximately 1/9 of the points in A_n will be in each of these 9 subtriangles. The larger is n, the closer we get to 1/9.

Similarly, for n large, approximately 1/27 of the points in A_n will be in each of the 27 "level 3" Sierpinski subtriangles for S.

Etc.

The main point is that for large n the points in A_n will be "evenly spread" over the small subtriangles of S. The larger is n the smaller the subtriangles over which A_n is "evenly spread". In this way, one can show that the sets A_n converge to the Sierpinski Triangle S.

To make all this more precise requires the theory of probability.

Next suppose the seed x_0 is *not* in S and the distance from x_0 to the *closest* point in S is α. Let us call the closest point y. We know that for $i = 1, 2, 3$,

$$d(f_i(x_0), f_i(y)) = \frac{1}{2} d(x_0, y) = \frac{\alpha}{2},$$

since each of f_1, f_2, f_3 has contraction ratio $1/2$.

Because $x_1 = f_i(x_0)$ for some i and because $f_i(y) \in S$, the distance from x_1 to the closest point in S is at most $\alpha/2$.

Similarly, the distance from x_2 to the closest point in S is at most $\alpha/4$, the distance from x_3 to the closest point in S is at most $\alpha/8$, etc.

Eventually x_n will be within any given tolerance of some point in S, say x^*. Take x^* as the seed and apply the same functions as those applied to x_n to get $x_{n+1}, x_{n+2}, x_{n+3}, \ldots$. Then the iterates of x^* and the corresponding iterates of x_n will be even closer to each other than x_n is to x^*. *Why?* ✍

Since $x^* \in S$ we already know that the first part of the Theorem applies to the seed x^*. So a large number of iterates of x^* is a good approximation to S, and it follows that a similarly large number of iterates of x_n is also a a good approximation to S.

This gives the main ideas, but to make it more precise would take us too far afield. □

Generalisation to any IFS For any IFS consisting of contractive maps, the Chaos Game using those maps will generate the unique fractal given by the IFS in Theorem 4.4.5. In applications, this is in fact the most efficient and effective way to generate the fractal.

The proof is essentially the same as for the Sierpinski triangle case.

Questions

1 Can you think of an IFS with two maps which gives the Koch Curve K?
HINT: Look at the two "halves" of K.
How about an IFS with 3 maps? And one with 8 maps?

2 Define $f(x) = x + 1/x$ for $x \in [0, \infty)$. Draw the graph.
Prove that $|f(x_1) - f(x_2)| < |x_1 - x_2|$ whenever $x_1 \neq x_2$.
Explain why the smallest contractivity factor for f is 1.

4.5 Simple Processes Can Lead to Chaos

Overview

The diagrams here were done using java applets developed by Bob Devaney and others, available at
`http://math.bu.edu/DYSYS/applets/`,
or by using the Maple "Chaos" package from Ken Monks, available at
`http://math.scranton.edu/monks/software/chaos/ChaosDemo.html`.

Review

The Logistic Model In Section 4.2 we discussed the Verhulst model, also called the logistic model. The population density after n time steps is written as p_n. We saw in Theorem 4.2.3 that if we make the "Main Assumption" on page 146 then the population density p_{n+1} after $n+1$ time steps can be calculated from the population density p_n after n time steps by the formula

$$p_{n+1} = p_n + ap_n(1-p_n) = (1+a)p_n - ap_n^2. \tag{4.24}$$

The number a is a parameter[29] which depends on the natural reproductive rate, the food supply, the prevalence of predators, etc. The parameter a is in the range $0 \le a \le 3$ if the model is to be physically reasonable. See the Examples on page 147.

We also saw for $a = 3$, and in fact also for values of a less than 3 but near 3, that the population density fluctuates wildly and is chaotic and essentially unpredictable.

Initial Seeds and Orbits Because of (4.24) we are interested in the function f given by

$$f(x) = (1+a)x - ax^2, \tag{4.25}$$

where a is a number in the range $0 \le a \le 3$.

We want to analyse the long term behaviour of the sequence $x_0, x_1, x_2, \ldots$ which begins with some *initial seed* x_0 in the range $0 \le x_0 \le 1$ and is obtained as follows:

$$x_0,\ x_1 = f(x_0),\ x_2 = f(x_1), \ldots,\ x_n = f(x_{n-1}), \ldots . \tag{4.26}$$

The sequence (4.26) is called the *orbit* or *trajectory* of x_0. It will turn out that the initial seed x_0 is usually irrelevant to the important aspects of the long term behaviour of the orbit.

Other Functions Instead of working with the function (4.25) it is a little easier to work with the function

$$f(x) = cx(1-x), \tag{4.27}$$

[29] The word "parameter" is used to indicate a number a on which the particular model depends. We will usually be interested in studying how important features of the model change as a changes.

where $0 \leq x \leq 1$ and c is a parameter such that $0 < c \leq 4$. See "Staying in the Box" on page 185. (The case $c = 0$ is not very interesting and so we omit it since we often want to divide by c.) This equation is also called the "logistic equation", and in future we will use the words "logistic equation" in this sense.

All the important features we study are the same whether we use (4.25) or (4.27). In fact, all the features are the same with essentially any one parameter family of quadratic maps. An example we will find convenient to use later when we discuss the Mandelbrot set is given by $f(x) = x^2 + c$ where $-2 \leq c \leq 0.25$.

Cobweb Diagrams

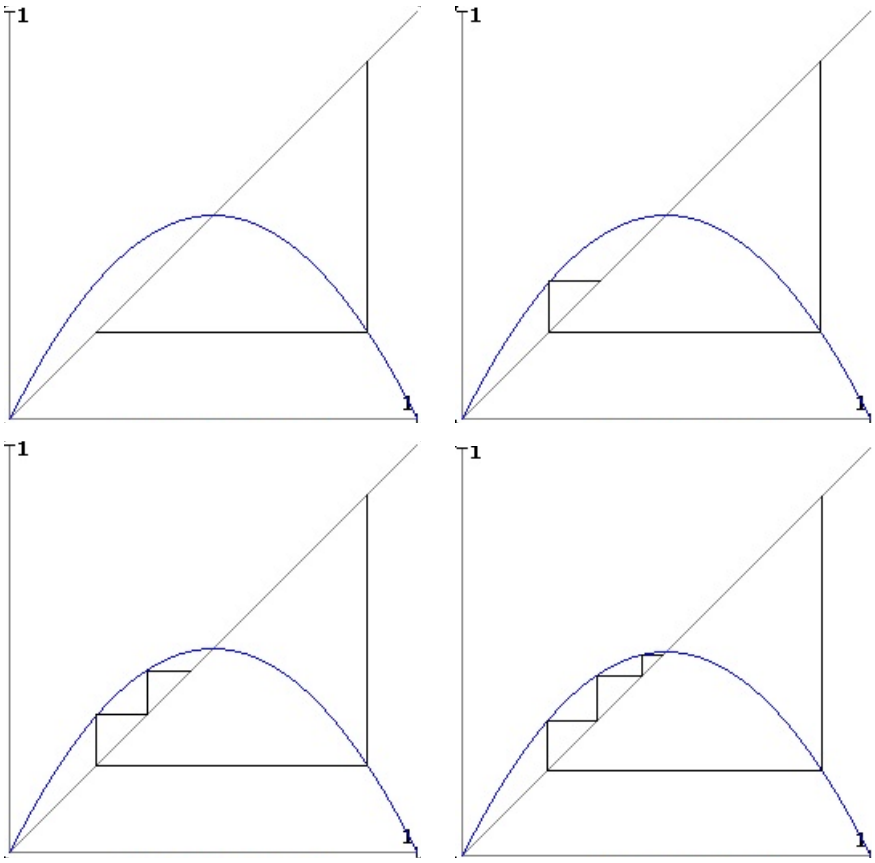


Figure 4.34: One, two three and four iterations of $f(x) = 2x(1 - x)$. Always start with a vertical line from the diagonal.

Finding Orbits In Figure 4.34 we begin with a seed $x_0 \approx 0.9$ and find the orbit points

$$x_1 = f(x_0),\ x_2 = f(f(x_0)),\ x_3 = f(f(f(x_0))) \text{ and } x_4 = f(f(f(f(x_0))))$$

respectively. We usually write

$$x_0,\ x_1 = f(x_0),\ x_2 = f^2(x_0),\ x_3 = f^3(x_0),\ x_4 = f^4(x_0), \ldots .$$

You can tell in the diagrams at which end we start because the first line is vertical, *not* horizontal. I will explain what is happening in the section "the Cobweb Process" below. But you should see if you can first figure it out yourself.

Composition of Functions Notice that by f^2 we mean the composition $f \circ f$ of f with itself, and not the square of f.

For example, if $f(x) = \sin x$ then

$$f^2(x) = (f \circ f)(x) = \sin(\sin x), \ \textit{not} \ (\sin x)^2$$

If $f(x) = 3x(1-x)$ then

$$\begin{aligned} f^2(x) = (f \circ f)(x) &= f\big(3x(1-x)\big) \\ &= 3\big(3x(1-x)\big)\big(1-3x(1-x)\big) \\ &= 9x(1-x)(1-3x+3x^2). \end{aligned}$$

On the other hand

$$\big(f(x)\big)^2 = 9x^2(1-x)^2.$$

Notice that in the example $f^2(x)$ is a quadric, $f^3(x)$ is a sixth order polynomial, and so on.

The Cobweb Process Here is how it works. See Figure 4.34.

Think of numbers as being represented by points on the main diagonal, i.e. on the line described by $y = x$. For example, the number 0.9 is represented by the point on the diagonal with coordinates $(0.9, 0.9)$. In general, the number a is represented by the point on the diagonal with coordinates (a, a).

Start with the number a represented by the point (a, a). You can find the point representing the number $f(a)$, i.e. find the point $(f(a), f(a))$, by the following geometric procedure:

> *Move vertically from (a, a) until you meet the graph. Then move horizontally until you meet the diagonal. This will give the point $(f(a), f(a))$ representing $f(a)$.*

When you start at (a, a) and move vertically to the graph, the point you meet on the graph is $(a, f(a))$, and the subsequent point on the diagonal is $(f(a), f(a))$. *Why?*.

To find the point representing $f(f(a))$ just repeat the process, starting at $(f(a), f(a))$. And so on.

By the way, it is easy to remember that one first goes vertical and then goes horizontal. If one tries to first go horizontal there might not be a corresponding point on the graph, or there may be more than one point. Look at what happens with the first diagram in Figure 4.34.

Examples of Cobwebs Diagrams show why we use the terminology "cobweb".

In Figure 4.35, Examples 1, 2, 3 and 5 show iterations from some initial seed. In Examples 4 and 6 we show the approximately "steady state" situation

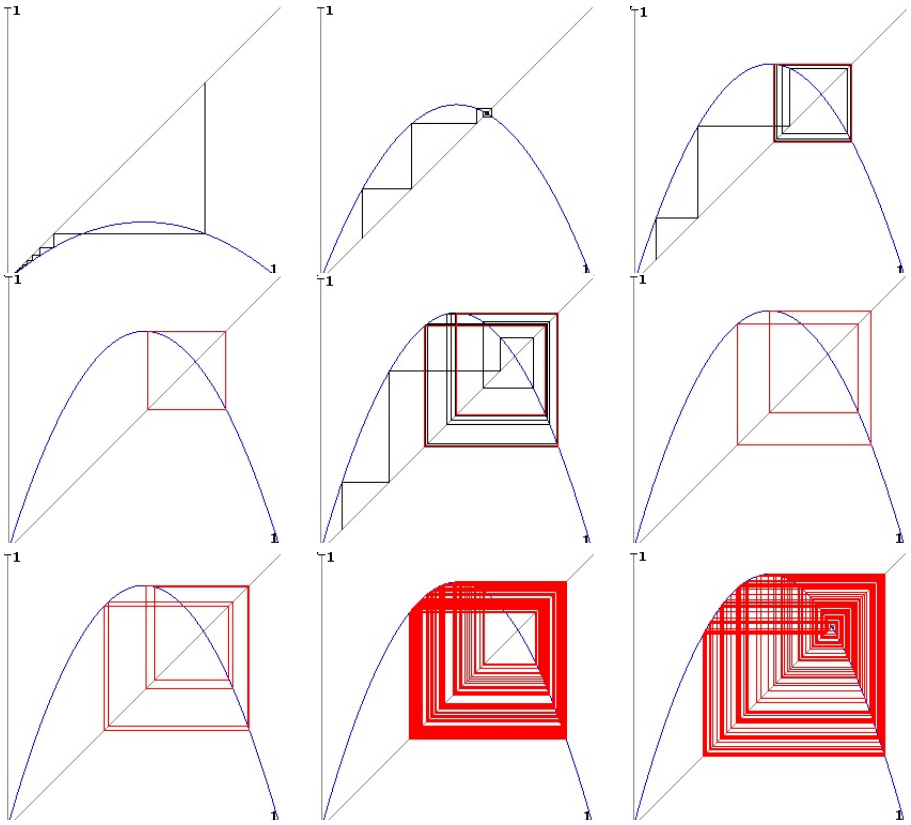


Figure 4.35: Cobwebs for $f(x) = cx(1-x)$ with $c = 0.85$, 2.57, 3.20, 3.20, 3.50, 3.50, 3.55, 3.60, 3.70, respectively. To get the correct direction, remember to move vertically from points on the diagonal.

obtained from Examples 3 and 5 by omitting the first 15 iterates. In Example 7 and onwards, and for all examples in Figure 4.36, we omit the first 15 iterates.

We will explain what happens in these diagrams in the remainder of this section.

Staying in the Box Notice that the parabola $y = cx(1-x)$ crosses the x-axis at $x = 0$ and $x = 1$, no matter what is the value of c.

Also, the maximum of the function f given by $f(x) = cx(1-x)$ is taken at $x = 1/2$. This is clear geometrically by symmetry. It also follows by setting the derivative $f'(x) = 0$, which implies $c - 2cx = 0$ and so $x = 1/2$.

At the maximum point $x = 1/2$ we have $f(1/2) = c/4$. So it follows that the graph of f lies in the box bounded by the points $(0,0)$, $(1,0)$, $(1,1)$ and $(0,1)$, provided $0 \leq c \leq 4$. So for this range of c, the cobweb process remains in the box. For other c there will always be initial seeds for which the cobweb process moves outside the box. *Draw a few diagrams showing why this is so.* ✍

We always only consider c such that $0 < c \leq 4$.

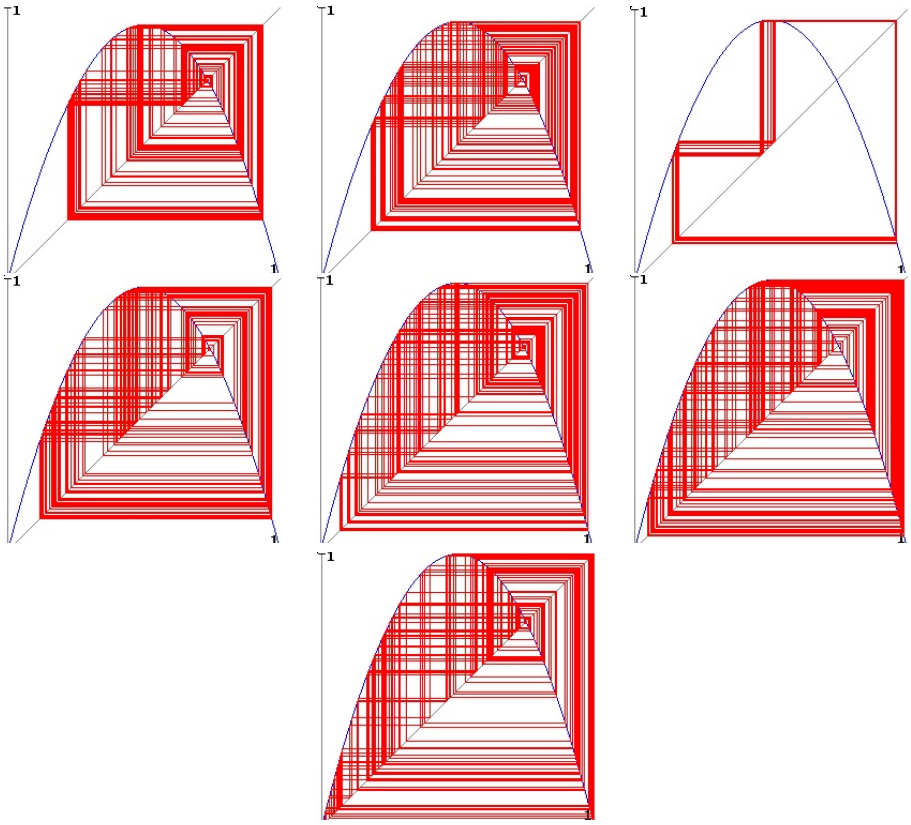


Figure 4.36: Cobwebs for $f(x) = cx(1-x)$ with $c = 3.75$, 3.80, 3.85, 3.87, 3.92, 3.95, 4.00, respectively.

Fixed Points

Finding Fixed Points We want to find *fixed points* of the function f. These are numbers x such that $f(x) = x$. They occur exactly where the diagonals in Figures 4.34, 4.35 and 4.36 cross the parabolas. Algebraically, we want to solve

$$cx(1-x) = x.$$

The solutions are

$$a = 0,\ a = 1 - \frac{1}{c}.$$

✍ *Why?* Since we are only interested in fixed points $0 \le x \le 1$ we will only consider the second solution if $c \ge 1$. And remember that we always want $c \le 4$, see "Staying in the Box" on page 185.

✍ **Cobwebs near Fixed Points** If we start the cobweb diagram at a fixed point a, it stays there. *Why?*

A more interesting question is "what happens if we start the cobweb diagram *sufficiently near* a fixed point a?" In this case there are four major cases to consider. They depend on the derivative $f'(a)$ at a.

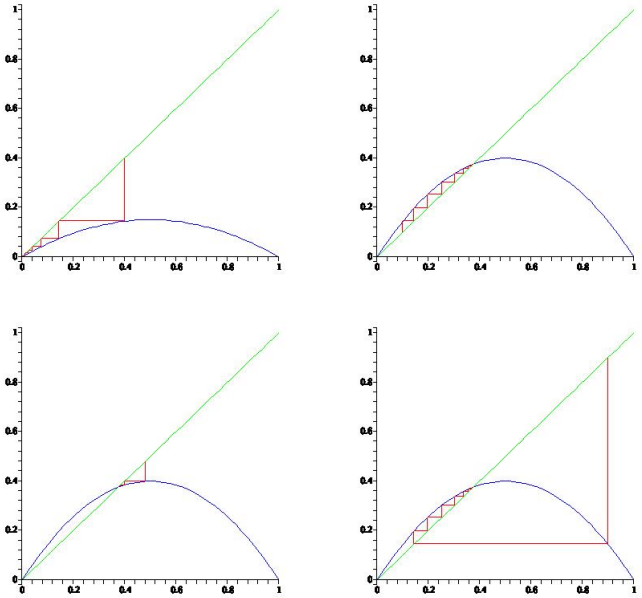


Figure 4.37: Cobwebs for $f(x) = cx(1-x)$ with $c = 0.6, 1.6, 1.6, 1.6$ respectively. In the first example $0 < f'(0) < 1$ and the cobweb steps down to the stable fixed point 0. In the next 3 examples the cobwebs eventually step up or down to the stable fixed point $a = 1 - 1/1.6 = 3/8$. Note $-1 < f'(a) < 1$.

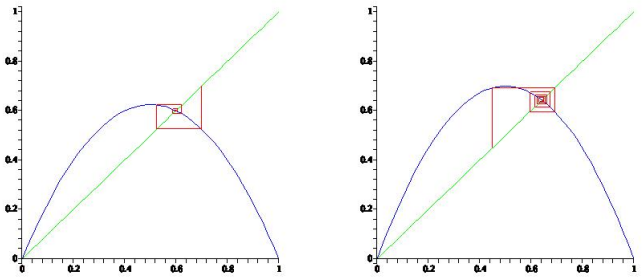


Figure 4.38: Cobwebs for $f(x) = cx(1-x)$ with $c = 2.5, 2.8$ respectively. The cobwebs spiral towards the stable fixed points $a = 1 - 1/2.5 = 3/5$ and $a = 1 - 1/2.8 = 9/14$, respectively. Note $-1 < f'(a) < 0$ in both cases.

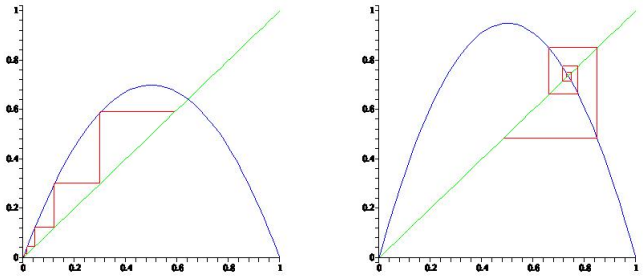


Figure 4.39: Cobwebs for $f(x) = cx(1-x)$ with $c = 2.8, 3.8$ respectively. The first cobweb spirals away from the unstable fixed point 0. To see what happens to it, look at the similar example 2 in Figure 4.35. The second cobweb spirals away from the unstable fixed point $a = 1-1/3.8 = 14/19$. To see what happens to it, look at example 2 in Figure 4.36.

1. $(0 < f'(a) < 1.)$ In this case the cobweb converges to the fixed point a on the diagonal by eventually "stepping up" or "stepping down" towards a. See Figure 4.37.
2. $(-1 < f'(a) < 0.)$ In this case the cobweb converges to the fixed point a by eventually "spiralling in" towards a. See Figure 4.38.
3. $(f'(a) > 1.)$ In this case the cobweb diverges from the fixed point by "stepping away" from a. See the first diagram in Figure 4.39.
4. $(f'(a) < -1.)$ In this case the cobweb diverges from the fixed point by "spiralling away" from a. See the second diagram in Figure 4.39.

Stable Fixed Points

Definition 4.5.1. A fixed point a is *stable* for f if there is some open interval I containing a [30] such that whenever the seed x_0 belongs to I, the orbit $x_0, f(x_0), f^2(x_0), f^3(x_0), f^4(x_0), \ldots$ converges to a.

A fixed point is *unstable* if it is not stable.

Sometimes we call a stable fixed point an *attractor* or *attractive* fixed point. An unstable fixed point is then called a *repellor* or a *repelling* fixed point. However, terminology differs a little from some books to others.

When is a Fixed Point Stable?

Theorem 4.5.2. *A fixed point a for the function f is stable if $|f'(a)| < 1$. The fixed point is unstable if $|f'(a)| > 1$*

Comments. We will not give a rigorous proof. But in the discussion and diagrams in the previous section we saw what is happening.

We discussed the cases $-1 < f'(a) < 0$ and $0 < f'(a) < 1$ on page 188 and saw by means of a geometric analysis that a is then stable.

[30] This means that $I = (b, c)$ where $b < a < c$.

In the case $f'(a) = 0$, if we start from a seed x_0 sufficiently close to a then the iterates of x_0 will in fact converge *very* fast to a, and so a is stable. Look at Figure 4.34 and take $a = 1/2$, corresponding to the point where the diagonal crosses the parabola. Whether the iterates step or spiral will depend on the second and perhaps higher derivatives. *Draw a couple of diagrams in this case to see what is happening.*

For $|f'(a)| > 1$ we noted on page 188 that the iterates of x_0 for x_0 near a will move away from a. So a is unstable.

If $f'(a) = \pm 1$ then the iterates from x_0 near a may slowly move towards a or may slowly move away from a. It depends on the second and perhaps higher derivatives at a. *Draw a couple of diagrams in this case to see what is happening.* □

Note. In the cases we will consider, if $|f'(a)| = 1$ then a will be stable.

Stability Analysis for Different c We consider the cases $0 < c \leq 4$, as discussed previously. We have

$$f(x) = cx(1-x) = cx - cx^2, \quad f'(x) = c - 2cx.$$

The fixed points a are given by $f(a) = a$, which gives $a = 0$ and $a = 1 - 1/c$. Moreover

1. For $0 < c \leq 1$, the fixed point 0 is stable and all orbits are attracted to it. There is no other fixed point. See Diagram 1 in Figures 4.35 and 4.37. *Find $f'(0)$ in these cases and use the previous Theorem and the subsequent "Note".*
2. For $1 < c \leq 3$ the fixed point 0 is unstable and the fixed point $1 - 1/c$ is stable. All orbits with initial seed different from 0 or 1 converge to $1-1/c$. See Diagram 2 in Figure 4.35 and Diagrams 2, 3, 4 in Figure 4.37. *Find $f'(0)$ and $f'(1 - 1/c)$ in these cases and use the previous Theorem and the subsequent "Note".*
3. For $3 < c \leq 4$ both of the fixed points 0 and $1 - 1/c$ are unstable. Things really get interesting in this range, as we will discuss. Look at Figures 4.35 and 4.36. *Find $f'(0)$ and $f'(1-1/c)$ in these cases and use the previous Theorem and the subsequent "Note".*

Periodic Cycles

If $3 < c \leq 4$ then the two fixed points 0 and $1 - 1/c$ are both unstable. So orbits are going to "bounce around" quite a lot.

Numerical Experiments If we look at Example 3 and particularly Example 4 in Figure 4.35 it appears that there are 2 points a and b such that $f(a) = b$ and $f(b) = a$. They are the two points on the diagonal which are also corners of the square in Example 4.

If we look at Example 5 and particularly Example 6 in Figure 4.35 it appears that there are 4 points p, q, r and s such that $f(p) = q$, $f(q) = r$, $f(r) = s$ and $f(s) = p$. *Describe these 4 points in Example 6.*

Our observations are indeed correct.

Definition 4.5.3. We say that two distinct points a and b form a *period two cycle* for f if $f(a) = b$ and $f(b) = a$.

We say that three distinct points p, q and r form a *period three cycle* for f if $f(p) = q$, $f(q) = r$ and $f(r) = p$.

Etc.

Finding Period Two Cycles If $f(a) = b$ and $f(b) = a$ then $f^2(a) = a$. *Why?* (By f^2 we mean $f \circ f$, *not* the square of f.)

On the other hand, if $f^2(a) = a$ and $a \neq f(a)$, then a and $f(a)$ form a period two cycle. *Why?*

So a very nice fact is that *instead of looking for period two cycles for f we can look for fixed points of the function f^2*. In other words, we want to solve

$$f(f(x)) = x. \tag{4.28}$$

Using this idea we can prove the following Theorem.

Theorem 4.5.4. *Let $f(x) = cx(1-x)$ where $0 < c \leq 4$ is fixed. Suppose $0 \leq x \leq 4$.*

If $0 < c \leq 3$ then f has no period two cycle.

If $3 < c \leq 4$ then there is exactly one periodic two cycle. For c in this range, the quadratic equation $c^2x^2 - (c^2 + c)x + (c + 1) = 0$ has two distinct real solutions for x,

$$x = \frac{(c+1) \pm \sqrt{(c+1)(c-3)}}{2c},$$

and these two solutions give a period two cycle for f.

Proof. Since $f(x) = cx(1-x)$ we can write (4.28) as

$$\begin{aligned} 0 &= x - f(f(x)) \\ &= x - c\,f(x)\,(1 - f(x)) \\ &= x - c\,\big(cx(1-x)\big)\,\big(1 - cx(1-x)\big) \\ &= x - c^2x(1-x)(1 - cx + cx^2) \\ &= x + c^2x(x-1)(cx^2 - cx + 1) \\ &= x\big(1 + c^2(x-1)(cx^2 - cx + 1)\big) \\ &= x\big(c^3x^3 - 2c^3x^2 + c^2(1+c)x + (1 - c^2)\big) \end{aligned} \tag{4.29}$$

After division by x we are left with a cubic equation. Cubics are usually messy to solve.[31] However, in this case there is a way around the problem.

[31] MAPLE can solve cubics exactly and also quadrics. But there is no formula for solving quintics exactly in terms radicals (square roots, cube roots, fourth roots, etc.). The theorem that there is no formula was proved by Galois. Unfortunately he was killed in 1832 in a duel at the age of 20.

Remember that $a = 1 - 1/c$ is a fixed point of f. This implies a is also a fixed point of $f \circ f$ since

$$f(f(a)) = f(a) = a.$$

And this implies that $x = 1 - 1/c$ is a solution of (4.28) and so $x - (1 - 1/c)$ is a factor of (4.29). Nice!

If you are now brave enough to divide through by $x - (1 - 1/c)$ (*do it!*) you will see that the expression in (4.29) factorises to

$$x\left(x - (1 - 1/c)\right) c\left(c^2x^2 - (c^2 + c)x + (c + 1)\right).$$

Doing the $b^2 - 4ac$ thing it follows the quadratic part equals zero, i.e.

$$c^2x^2 - (c^2 + c)x + (c + 1) = 0, \tag{4.30}$$

if

$$x = \frac{(c^2 + c) \pm \sqrt{(c^2 + c)^2 - 4c^2(c + 1)}}{2c^2} = \frac{(c + 1) \pm \sqrt{(c + 1)(c - 3)}}{2c}. \tag{4.31}$$

This gives two distinct real roots if and only if

$$(c + 1)(c - 3) > 0.$$

For c in the range we are considering, namely $0 < c \leq 4$, this is true precisely when $3 < c \leq 4$.

We should check for these c that the solutions for x of (4.30) do lie in the range $0 \leq x \leq 1$ and that they are different from the other solutions $x = 0$ and $x = 1 - 1/c$. *Check it*, but it is also clear from Figure 4.41.

To summarise, we have seen that there are two distinct real solutions of (4.29) if $0 < c \leq 3$ and there are 4 solutions if $3 < c \leq 4$. In the second case, the two solutions other than $x = 0$ and $x = 1 - 1/c$ form a two cycle.

To see this let the 4 solutions of (4.29) be

$$0, 1 - 1/c, a \text{ and } b,$$

where a and b are given by (4.31).

Since a is a fixed point of f^2 it follows that $f(a)$ is also a fixed point. *Why?*

So $f(a)$ must be one of 0, $1 - 1/c$, a or b. For $c > 3$, none of these are equal, see Figure (4.41).

If $f(a) = 0$ then $f^2(a) = f(0) = 0$, and so $a = 0$. *Why?*

If $f(a) = 1 - 1/c$ then $a = 1 - 1/c$ by a similar argument. *Do it!*

If $f(a) = a$ then a must be one of the two fixed points of f and so $a = 0$ or $a = 1 - 1/c$.

So the only possibility is that $f(a) = b$. In a similar way, $f(b) = a$, and so $\{a, b\}$

□

Examples We now go back to Examples 3 and Example 4 in Figure 4.35. The equation was

$$f(x) = 3.2x(1-x).$$

Putting $c = 3.2$ in the quadratic equation (4.30) and solving gives

$$x = 0.79945\ldots,\ x = 0.51304\ldots$$

according to MAPLE.[32] Setting $a = 0.79945\ldots$ and $b = 0.51304\ldots$, MAPLE will also confirm that $f(a) = b$ and $f(b) = a$ up to the order of accuracy you choose.

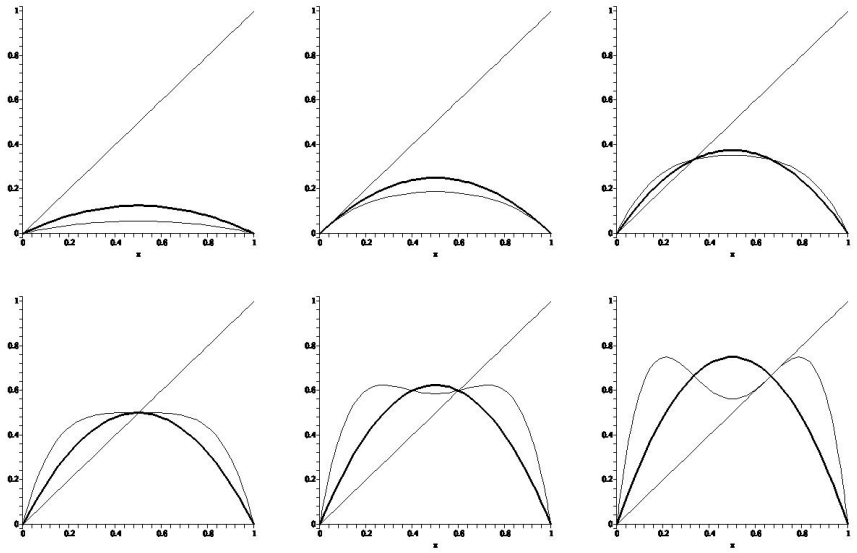


Figure 4.40: Graphs of f (heavy) and f^2 (i.e. $f \circ f$) for $c = .5, 1, 1.5, 2, 2.5, 3$. The equation $f^2(x) = x$ has the same real solutions as $f(x) = x$, for such c.

In Figure 4.40 we show graphs of f and f^2 for c in the range $0 < c \leq 3$. Notice that the real solutions of $f(x) = x$ and $f^2(x) = x$ are the same. In Figure 4.40 we show graphs of f and f^2 for c in the range $3 < c \leq 4$. Notice that the real solutions of $f^2(x) = x$ are the two solutions of $f(x) = x$ plus two more solutions.

Stable Two Cycles

When is a Two Cycle Stable? I will be somewhat informal in this discussion.

[32] I used the commands:

```
> c := 3.2 ;,
> eq := c^3*x^2 -(c^3+c^2)*x + (c^2+c) = 0 ;
> solve(eq,x) ;
```

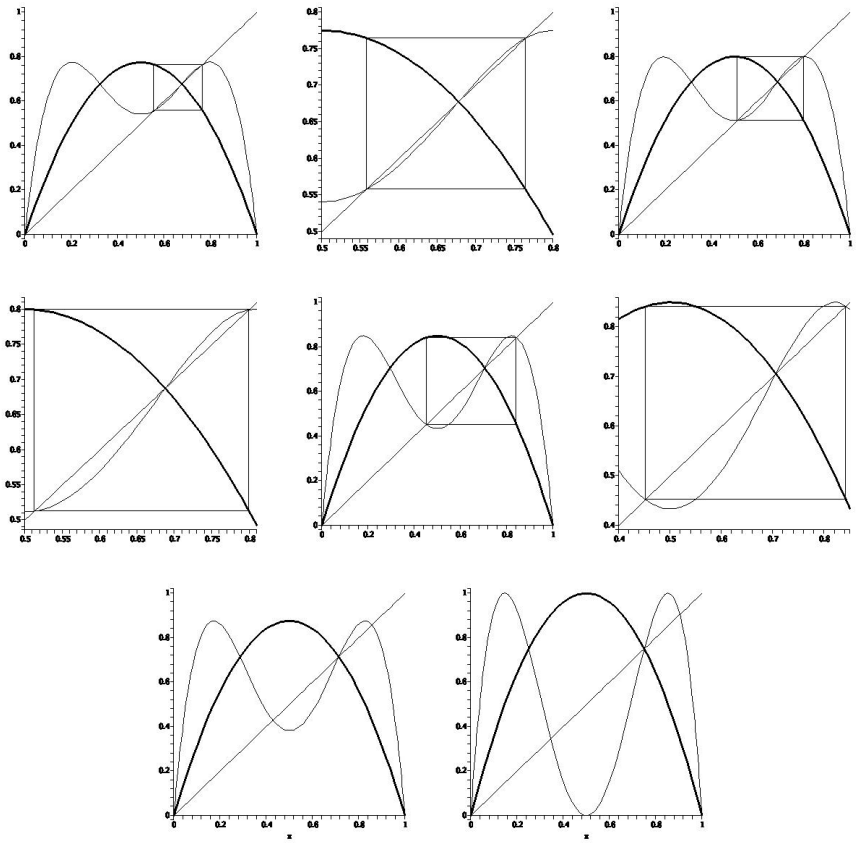


Figure 4.41: Graphs of f (heavy) and f^2 (i.e. $f \circ f$) for $c = 3.1$ (including a zoom), 3.2 (including a zoom), 3.4 (including a zoom), 3.5 and 4.0. In the first three cases the two cycle is "stable" and is shown. In the last two cases it is "unstable" and is not shown. But you should be able to draw it. We will discuss stability for two cycles beginning on page 192.

We have seen that for $3 < c \leq 4$ there is a two cycle $\{a, b\}$, with $f(a) = b$ and $f(b) = a$.

In the case of a fixed point a, i.e. $f(a) = a$, it was important to look at what happens to points near a. Are they attracted to a or are they repelled from a? Similarly, if $\{a, b\}$ is a two cycle we can ask if a point x_0 near a has the property that the orbit

$$x_0, f(x_0), f^2(x_0), f^3(x_0), f^4(x_0), f^5(x_0), f^6(x_0), \ldots$$

moves closer and closer, i.e. is "attracted to" the orbit

$$a, b, a, b, a, b, \ldots,$$

or not. This is equivalent to asking if the sequence

$$x_0, f^2(x_0), f^4(x_0), f^6(x_0), \ldots$$

converges to a for x_0 near a, and the sequence

$$y_0, f^2(y_0), f^4(y_0), f^6(y_0), \ldots$$

converges to b for y_0 near b. (Think of y_0 as $f(x_0)$).

Why is Stability Important? When we run a numerical experiment we will observe stable fixed points because lots of orbits will converge to them. In fact in our case there will be at most one stable fixed point, and if this is the case then essentially all orbits will converge to it. We will not observe unstable fixed points except in incredibly rare circumstances. In practice, if there is a little bit of "noise", an orbit will eventually be driven away from any unstable fixed point even if it happens to land on it.

Similarly, we will observe stable periodic cycles, but we will not observe unstable periodic cycles.

Definition 4.5.5. A two cycle $\{a, b\}$ for f is *attractive* or *stable* if there is an open interval I containing a, and an open interval J containing b, such that for any $x_0 \in I$ the orbit starting from x_0 converges to the orbit $a, b, a, b, \ldots$, and for any $x_0 \in J$ the orbit starting from x_0 converges to the orbit $b, a, b, a, b, a, \ldots$.

If this does not happen we say the two cycle $\{a, b\}$ is *repelling* or *unstable*.

Theorem 4.5.6. *If $\{a, b\}$ is a two cycle for f then*

$$(f^2)'(a) = (f^2)'(b) = f'(a)\, f'(b).$$

The cycle $\{a, b\}$ is stable if $|(f^2)'(a)| = |(f^2)'(b)| < 1$. It is unstable if $|(f^2)'(a)| = |(f^2)'(b)| > 1$.

Comments and Proof. Notice that the Theorem does not give any information in the case $|(f^2)'(a)| = |(f^2)'(b)| = 1$. But for the problems we consider, the two cycle will be stable in this case.

The equality follows from the chain rule in calculus:

$$\begin{aligned}
(f^2)'(a) &= \frac{d}{dx} f(f(x))\Big|_{x=a} \text{ just a change in terminology} \\
&= f'(f(x))\, f'(x)\Big|_{x=a} \text{ the chain rule} \\
&= f'(f(a))\, f'(a) \text{ setting } x = a \\
&= f'(b)\, f'(a) \text{ since } f(a) = b
\end{aligned}$$

The result for the derivative at b is obtained by switching a and b.

Suppose $|(f^2)'(a)| = |(f^2)'(b)| < 1$. It follows from Theorem 4.5.2 that a and b are stable fixed points for f^2. In the discussion before this Theorem we indicated why it then follows that the cycle $\{a, b\}$ for f is stable (we are not being very rigorous here!).

If $|(f^2)'(a)| = |(f^2)'(b)| > 1$ then from Theorem 4.5.2 it follows that a and b are unstable stable fixed points for f^2. From the discussion before this Theorem it follows that the cycle $\{a, b\}$ for f is unstable. □

Stable Two Cycles for Different c

Theorem 4.5.7. *The two cycle corresponding to $f(x) = cx(1-x)$ for $3 < c \leq 4$ is stable if $3 < c < 1 + \sqrt{6} = 3.449499\ldots$. It is unstable if $c > 1 + \sqrt{6}$.*

Proof. We want to use Theorem 4.5.6. So we calculate

$$\begin{aligned}(f^2)'(a) = (f^2)'(b) &= f'(a)\, f'(b) \\ &= c^2(1-2a)(1-2b) \\ &= c^2\big(1 - 2(a+b) + 4ab\big).\end{aligned}$$

Since a and b are the solutions of (4.30) ,

$$a + b = \frac{c^2 + c}{c^2}, \quad ab = \frac{c+1}{c}.$$

Why? It follows that

$$\begin{aligned}f'(a)\, f'(b) &= c^2\left(1 - 2\frac{c^2+c}{c^2} + 4\frac{c+1}{c}\right) \\ &= c^2 - 2c^2 - 2c + 4c + 4 \\ &= 4 + 2c - c^2.\end{aligned}$$

But for c in the range we are considering,

$$\begin{aligned}1 > 4 + 2c - c^2 > -1 \quad &\text{if} \quad 3 < c < 1 + \sqrt{6}, \\ 4 + 2c - c^2 < -1 \quad &\text{if} \quad 1 + \sqrt{6} < c \leq 4,\end{aligned}$$

Why?

The Theorem now follows from Theorem 4.5.6. □

Note: Even though we do not prove it, the two cycle is also stable if $c = 1 + \sqrt{6}$.

Don't Panic! The Story so Far

OK, this is important. You may well be a bit lost by the previous details. Do not worry! Here are the key points, presented three slightly different ways.

In computer experiments we will observe the following. In a few cases we might need to have a miniscule amount of noise to bump us off any unstable fixed points or cycles.

1. If $0 < c \leq 1$ then every orbit will converge to 0.
2. If $1 < c \leq 3$ then every orbit will converge to $1 - 1/c$.
3. If $3 < c \leq 1 + \sqrt{6} = 3.449499\ldots$ then every orbit will converge to a cycle bouncing back and forth between two points which depend on c.

In more precise language:

1. If $0 < c \leq 1$ then f has exactly one stable fixed point at 0. It has no stable cycles (and in fact no cycles at all).

2. If $1 < c \leq 3$ then f has exactly stable fixed point at $1 - 1/c$. It has no stable cycles (and in fact no cycles at all).
3. If $3 < c \leq 1 + \sqrt{6} = 3.449499\ldots$ then f has exactly one stable 2 cycle. It has no other stable fixed points or stable cycles. The stable 2 cycle contains the points

$$x = \frac{(c+1) \pm \sqrt{(c+1)(c-3)}}{2c}.$$

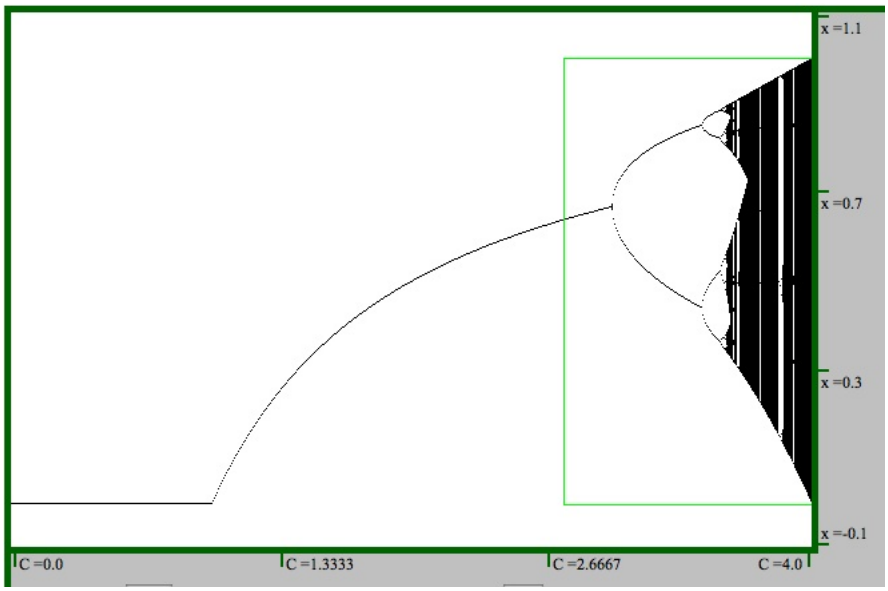


Figure 4.42: Bifurcation diagram for $f(x) = cx(1-x)$.

And there is a third way of expressing this information by the bifurcation diagram in Figure 4.42, also called the Feigenbaum diagram.

The horizontal axis shows c in the range $[0, 4]$. (Unfortunately the axis is divided in a silly way, and goes a little past 0. Read no significance into this.) The vertical axis shows x in the range $[0, 1]$. (It goes a little past this, but nothing happens for x outside $[0, 1]$.)

(The only point to the green vertical rectangle is that we will blow it up later, see Figure 4.44.)

1. The horizontal part of the diagram above $0 < c \leq 1$ shows the stable fixed point $x = 0$ for c in this range.
2. The arc above $1 < c \leq 3$ shows the stable fixed point $x = 1 - 1/c$ for c in this range.
3. The two arcs above $3 < c \leq 1 + \sqrt{6} = 3.449499\ldots$ show the two points $x = \big((c+1) \pm \sqrt{(c+1)(c-3)}\,\big)/2c$ forming the stable two cycle, for c in this range.

The Rest of the Story

The Stable Four Cycle From Figures 4.40 and 4.41, with $c = .5, 1, 1.5, 2, 2.5, 3, 3.1, 3.2, 3.4, 3.5, 4$ you can see the following:

1. For $0 < c < 1$;
   a) f crosses the diagonal at 0,
   b) $0 < f'(0) < 1$ and so 0 is stable.
2. As c passes through 1;
   a) a second crossing point $1 - 1/c$ is introduced,
   b) $f'(0)$ increases above 1 and so 0 becomes unstable,
   c) $f'(1 - 1/c)$ decreases down from 1 and so $1 - 1/c$ is stable.
3. For $1 < c < 3$;
   a) $f'(1 - 1/c)$ decreases from 1 to -1 and so $1 - 1/c$ remains stable.
4. As c passes through 3;
   a) $f'(1 - 1/c)$ decreases below -1 and so $1 - 1/c$ becomes unstable,
   b) the crossing point $1 - 1/c$ for f^2 splits into the three crossing points $1 - 1/c$ and a, $b = \big((c+1) \pm \sqrt{(c+1)(c-3)}\,\big)/2c$,
   c) $f^2(a) = a$, $f^2(b) = b$, $f(a) = b$ and $f(b) = a$.
   d) $(f^2)'(a) = (f^2)'(b)$ decrease down from 1 and so the two cycle $\{a, b\}$ is stable for f (see Theorem 4.5.6).
5. For $3 < c \le 1 + \sqrt{6} = 3.449499\ldots$;
   a) $(f^2)'(a) = (f^2)'(b)$ decreases down from 1 to -1 and so the two cycle $\{a, b\}$ is stable.

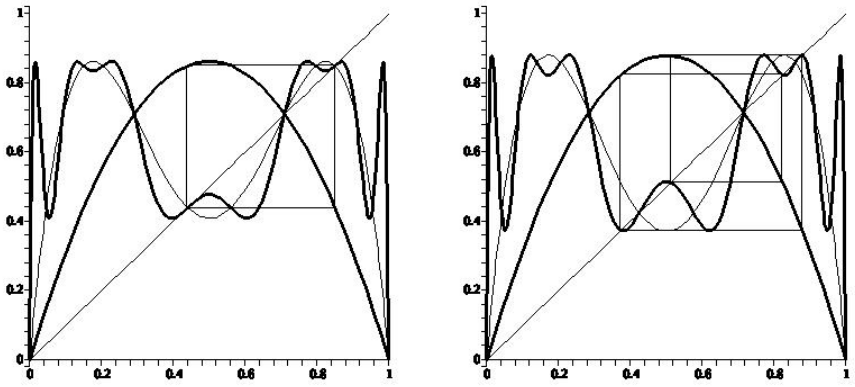


Figure 4.43: Graphs of f (heavy), f^2 and f^4 (heavy), for $c = 1 + \sqrt{6} = 3.34\ldots$ and 3.5. In the first case the 2 cycle is stable, but for $c > 1 + \sqrt{6}$ the 2 cycle becomes unstable and bifurcates into a stable 4 cycle.

I will now try to explain informally what next happens as c passes through $1 + \sqrt{6}$. You should look at Figure 4.43.

1. $(f^2)'(a) = (f^2)'(b)$ both simultaneously decrease below -1 and the 2 cycle $\{a, b\}$ becomes unstable,
2. $(f^4)'(a) = (f^4)'(b)$ both simultaneously increase above 1,
3. the crossing point a for f^4 splits into 3 crossing points a, a_1 and a_2,
4. the crossing point b for f^4 splits into 3 crossing points b, b_1 and b_2,
5. $f^4(a_1) = a_1$, $f^4(a_2) = a_2$, $f^4(b_1) = b_1$ and $f^4(b_2) = b_2$,
6. $f(a_1) = b_2$, $f(b_2) = a_2$, $f(a_2) = b_1$ and $f(b_1) = a_1$,
7. $\{a_1, b_2, a_2, b_1\}$ is a stable 4 cycle for c up to approximately 3.544090.

Period Doubling With $c > 1 + \sqrt{6} = 3.449499\ldots$ we are in the region where the so-called *period doubling* takes place. Look again at Figure 4.42. Now look at Figures 4.44–4.48 and relate them to steps 1–6 below.

1. After 3.544090... the 4 cycle becomes unstable and splits into a stable 8 cycle for c up to approximately 3.564407,
2. After 3.564407... the 8 cycle becomes unstable and splits into a stable 16 cycle for c up to approximately 3.568759,
3. After 3.568759... the 16 cycle becomes unstable and splits into a stable 32 cycle for c up to approximately 3.569692,
4. After 3.569692... the 32 cycle becomes unstable and splits into a stable 64 cycle for c up to approximately 3.569891,
5. After 3.569891... the 64 cycle becomes unstable and splits into a stable 128 cycle for c up to approximately 3.569934,
6. After 3.569934... the 64 cycle becomes unstable and splits into a stable 128 cycle for c up to approximately 3.569946,
7. etc.

The numbers above approach a limit value $\alpha = 3.569945671205296863\ldots$.

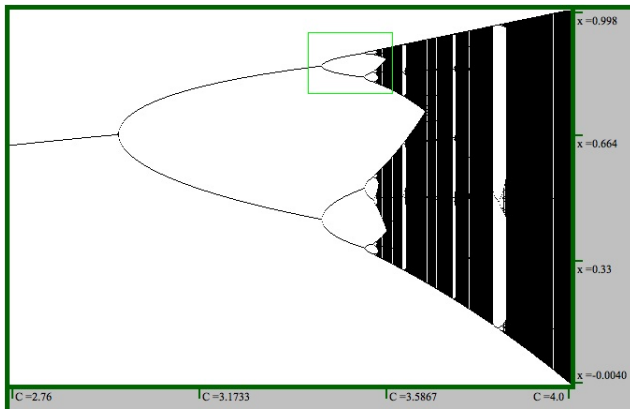


Figure 4.44: Blow up of part of Figure 4.42.

The Chaotic Regime For $c > \alpha$ there is a sea of chaos interrupted by regions of tranquil waters. See Figures 4.49–4.52.

But within the waters of tranquility there are seas of chaos. See Figure 4.51.

Yet within even the darkest seas there are still regions of light. See Figures 4.53 and 4.54.

And there are copies of the Feigenbaum diagram within copies of the Feigenbaum diagram within copies of the Feigenbaum diagram within copies of the Feigenbaum diagram ....

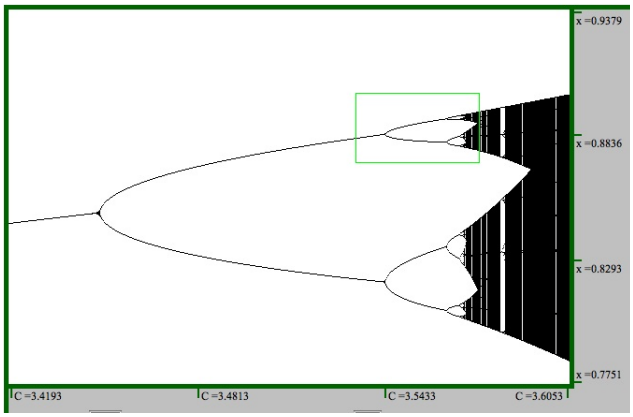


Figure 4.45: Blow up of part of Figure 4.44.

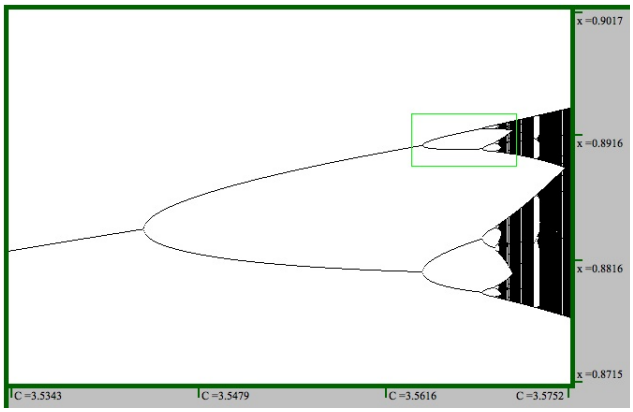


Figure 4.46: Blow up of part of Figure 4.45.

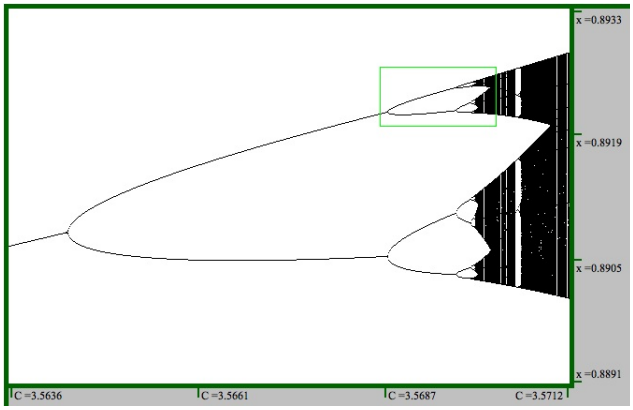


Figure 4.47: Blow up of part of Figure 4.46.

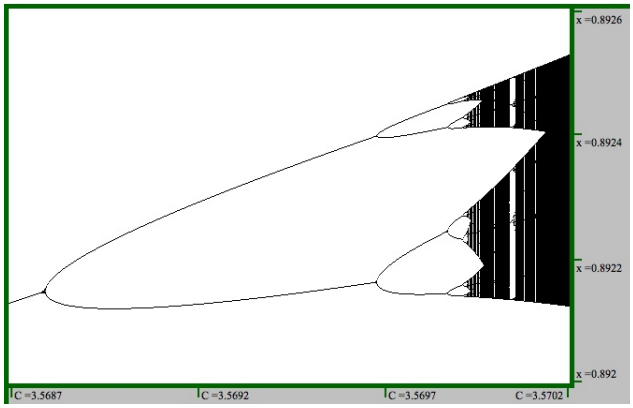


Figure 4.48: Blow up of part of Figure 4.47.

If you look carefully at the largest white band to the right of Figure 4.44 you will see a stable period 3 cycle. In fact, there are cycles of every period in this band, but they are unstable and so you do not see them!

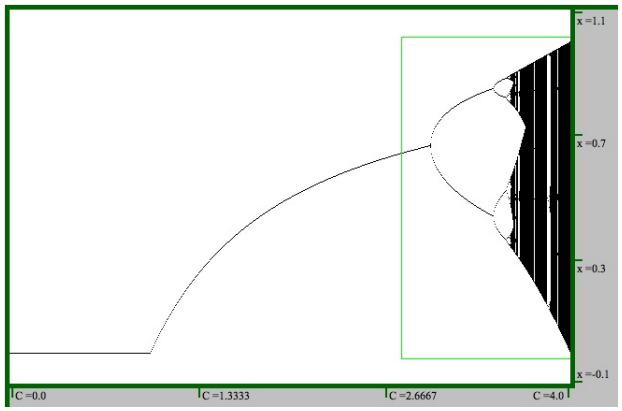


Figure 4.49: The Feigenbaum diagram again.

Questions

1 Find the fixed points of the function f given by $f(x) = x^2$. Which is stable and which is unstable? Why?

2 Consider the function given by $f(x) = \sin x$. What is its fixed point? What is f' at the fixed point? Is the fixed point stable? Why?

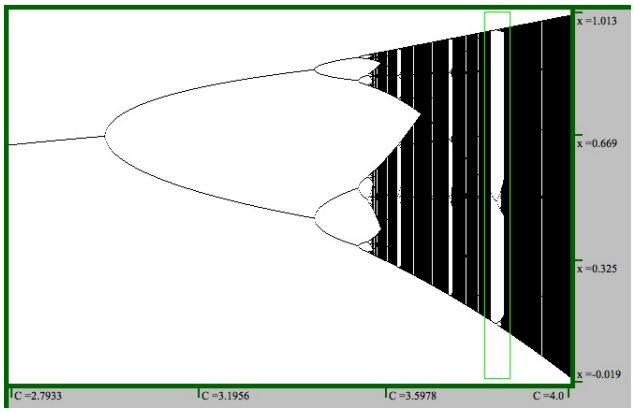


Figure 4.50: Blow up of part of Figure 4.49

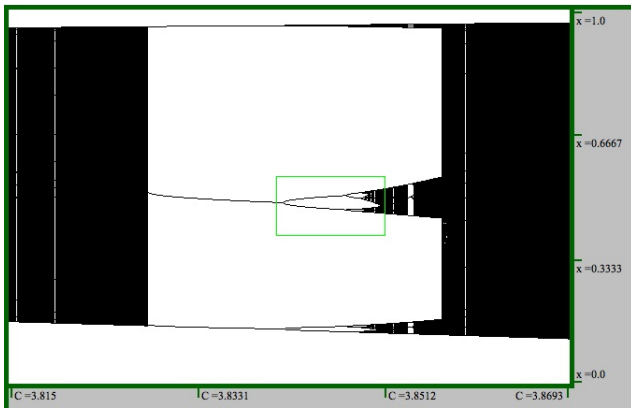


Figure 4.51: Blow up of part of Figure 4.50

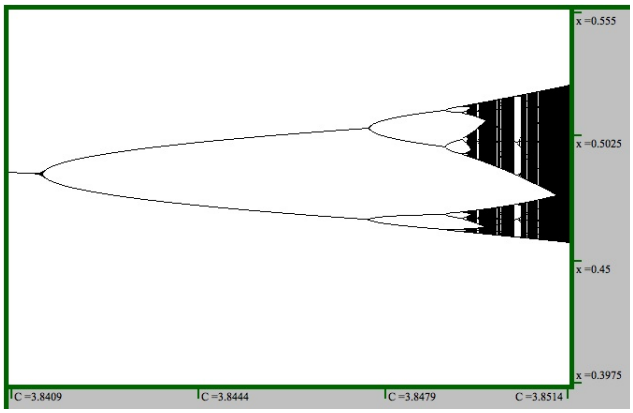


Figure 4.52: Blow up of part of Figure 4.51

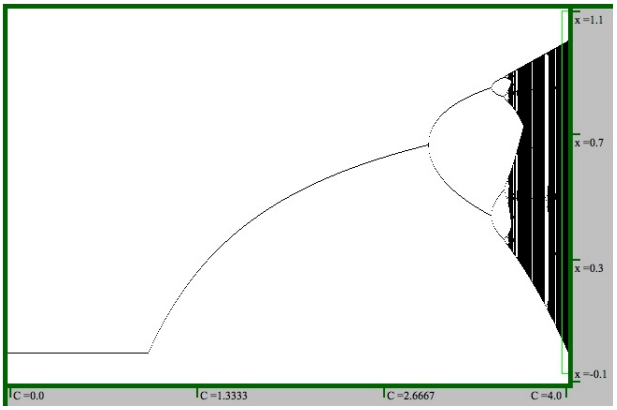


Figure 4.53: The Feigenbaum diagram again.

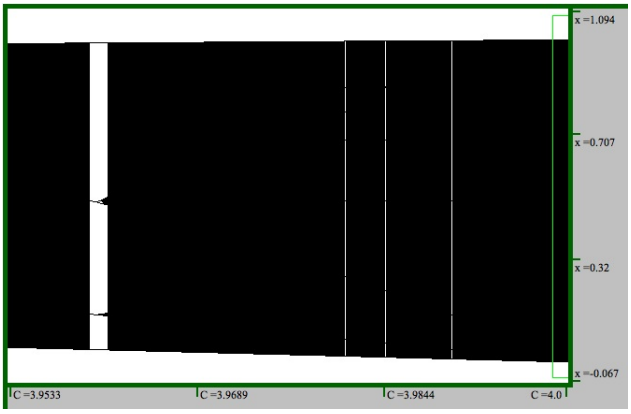


Figure 4.54: Blow up of part of Figure 4.53

Explain why the orbit starting at any point x_0 satisfies $-1 \leq x_1 \leq 1$. What happens after this?

3 In (4.25) we considered a function of the form $f(x) = (1+a)x - ax^2$. We will take $0 < a \leq 3$. Notice that its graph crosses the x-axis at $x = 0$ and $x = (1+a)/a$, and it meets the diagonal at $(1, 1)$. Draw a diagram.

It will be convenient to change notation a little. We use u and v to denote coordinates. The function will be denoted by g, so that

$$g(u) = (1+a)u - au^2 \quad \text{with graph given by} \quad v = (1+a)u - au^2.$$

In order for the orbits of g to look the same as those of the function

$$f(x) = cx(1-x) \quad \text{with graph given by} \quad y = cx(1-x),$$

it looks like we should scale the horizontal and vertical units by the same amount λ, say. This will leave both the origin and diagonal unchanged.

What value of λ, in terms of a, transforms the first graph into the second? What is the corresponding value of c in terms of a? As a varies over the interval $(0, 3]$, what happens to c.

4 Use the java applet at
`http://math.bu.edu/DYSYS/applets/Iteration.html` to view the logistic map as a time series for different values of c (λ in the applet). Look also at the histogram. Watch what happens as you pass through the values of b_n.

5 Compute the ratio $\dfrac{b_n - b_{n-1}}{b_{n+1} - b_n}$ for $n = 2, \ldots, 7$. It appears to be converging to 4.699 .... This is a special constant called δ, and it is the same in essentially all bifurcation situations.

6 Period 3 cycle.

1. Use the Devaney/Enchev java applet "Nonlinear web" to find to 2 decimal places the first value of c such that $f^3(x) = x$ has 3 real roots.
2. Find this value of c on Figure 4.51. Note that a stable period 3 cycle suddenly emerges out of a sea of chaos at this value of c. Let us call this value μ.
3. Use the Nonlinear Web applet to explain why a stable period 3 cycle, and an unstable period 3 cycle, occur as c passes above μ.
4. Prove that if $\{\alpha, \beta, \gamma\}$ is a 3 cycle then

$$(f^3)'(\alpha) = (f^3)'(\beta) = (f^3)'(\gamma) = f'(\alpha)\, f'(\beta)\, f'(\gamma).$$

7 The *decimal shift map* is defined

$$f(x) = 10x \bmod 1$$

for $x \in [0, 1]$. In other words, multiply by 10 and take the decimal part. For example:

$$f(0.3275\ldots) = 0.275\ldots,\ f(0.6918\ldots) = 0.918\ldots,\ f(1/2) = 0,$$
$$f(0.020202\ldots) = 0.20202\ldots,\ f(0) = 0,\ f(1) = 0,\ f(1/3) = 1/3.$$

1. Sketch the graph.
2. Find all fixed points. HINT: Write x in decimal form.
3. Find some period 2 cycles. Describe all period 2 cycles.
4. Find a period 3 cycle. Describe all period 3 cycles.
5. Explain why for every positive integer n there is a cycle of period n.
6. For which initial starting points is the corresponding orbit a cycle?
7. For which initial starting points is the corresponding orbit not a cycle?
8. Take one of the 2 cycles you described. Explain why it is not stable.
9. Take one of the 3 cycles you described. Explain why it is not stable.
10. Explain why there are no stable cycles.
11. Find an initial starting point whose orbit comes arbitrarily close to every number $x \in [0, 1]$. HINT: It is sufficient that the orbit includes every number x with a finite decimal expansion.

4.6 Julia Sets and Mandelbrot Sets

The following are preliminary notes for this section. Use the book [HoM] as the main source and these notes as a supplement.

Overview

Our goal here is to understand some of the properties of Julia sets and the Mandelbrot set. First look again at the discussion and the diagram beginning on page 141.

The Julia sets J_c (there is one of them for every complex number c) and the Mandelbrot set M are subsets of the plane $\mathbb{R}^2$. But the only way we can really understand these sets is by using complex numbers. Each point (a, b) in $\mathbb{R}^2$ corresponds to a complex $a + ib$. We will discuss complex numbers beginning on page 209.

We will be guided in our study by first looking at a much simpler situation. The baby Julia sets bJ_c (there is one of them for every *real* number c) and the baby Mandelbrot set bM are subsets of the *line* $\mathbb{R}$. Although these sets are not particularly interesting themselves, they do provide some motivation for the standard Julia sets J_c and Mandelbrot set M.

In Section 4.5 we discussed in detail the behaviour of iterations of the logistic map $f(x) = ax(1-x)$, for different values of the parameter a. There is nothing particularly special about this map. Similar behaviour occurs for iterates of essentially *any* function f provided it is not a affine map, i.e. is not of the form $f(x) = ax + b$. *What are the first two iterates in this case?*

Instead of the logistic map, here we will look at what happens if we iterate the *quadratic map* $f(x) = x^2 + c$. This will lead to the sets bJ_c and bM. Then we will look at what happens if we replace x by a complex number z and the parameter c also by a complex number and iterate the map $f(z) = z^2 + c$. This will lead to the sets J_c and M.

Baby Julia Sets and Baby Mandelbrot Set

The Real Quadratic Map In the case of the logistic map $f(x) = ax(1-x)$, we saw on page 185 that if $0 \le a \le 4$ and if we begin an orbit at any $x_0 \in [0, 1]$ then the orbit is bounded and in fact stays in the range $[0, 1]$. Geometrically, the orbit "stays in the box". If we begin an orbit at any x_0 *not* in the interval $[0, 1]$ then the orbit is unbounded and in fact diverges to $+\infty$.[33] *Explain graphically.*

We will now see that a similar graphical analysis works for the (real) quadratic map $f(x) = x^2 + c$. In Figure 4.55 you can see the graphs of these functions for $c = 0.5, -1.5, -2.3$.

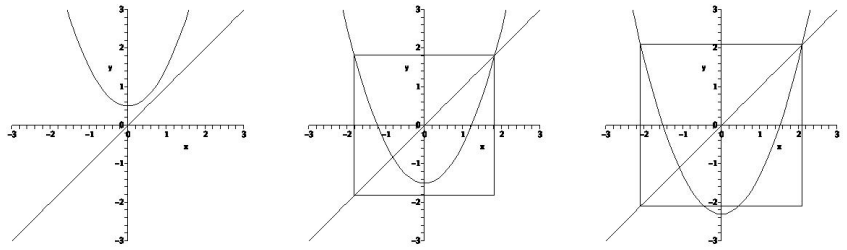


Figure 4.55: The maps $f(x) = x^2 + c$ for $c = 0.5, -1.5, -2.3$.

The Case $c > 0.25$ In this case there are no real solutions of the equation $f(x) = x$. *Check this.* See the first diagram in Figure 4.55.

If an orbit starts at *any* value of x_0 then it will eventually diverge to $+\infty$.[34] *Explain graphically.*

The Case $-2 \leq c \leq 0.25$ For $c \leq 0.25$ there are two real solutions of the equation $f(x) = x$.

See the second diagram in Figure 4.55. The square box analogous to the one for the logistic map is obtained here by starting from the top right point where the graph crosses the diagonal. The top of the box is the horizontal line beginning at this point and ending at the other point on the graph with the same y-coordinate. The left vertical side is obtained by starting from this point and ending at the point on the line $y = x$ with the same x-coordinate. Now complete the box in the only possible way. *Why is it a square box?*

The main difference from the logistic map case is that now the box depends on c.

A calculation shows that if $c \leq 0.25$ then the vertical sides of the box cross the x-axis at $\pm(1 + \sqrt{1 - 4c})/2$. A further calculation shows that if $-2 \leq c \leq 1/4$ then the vertex of the graph is in the box (we always include the edges of the box as part of the box). *Do the calculations.*

Suppose $-2 \leq c \leq 0.25$ and an orbit starts at some x_0 in the box, i.e. x_0 is in the interval $[-(1+\sqrt{1-4c})/2, (1+\sqrt{1-4c})/2]$. Then the orbit is bounded and in fact stays in the same interval. If an orbit begins at some point x_0 *not* in the interval $[-(1+\sqrt{1-4c})/2, (1+\sqrt{1-4c})/2]$ then the orbit is unbounded. *Explain graphically.*

For example:

1. if $c = .25$ then the orbit starting from x_0 is bounded if and only if $x_0 \in [-1/2, 1/2]$;

[33] There is no point $+\infty$. What is meant is that for any positive real number K, after a certain "time" all points on the orbit are eventually greater than K.

[34] Of course, $+\infty$ is not a number or a point. What is meant more precisely is that for any positive real number M, no matter how large, all points in the orbit will eventually be greater than M.

2. if $c = 0$ then the orbit starting from x_0 is bounded if and only if $x_0 \in [-1, 1]$;
3. if $c = -2$ then the orbit starting from x_0 is bounded if and only if $x_0 \in [-2, 2]$.

The case $c < -2$ (the Big Brother Syndrome[35]) If $c < -2$ then part of the graph of f is below the box.

If an orbit starts at some point x_0 on the diagonal outside the box then it stays outside the box and eventually diverge to infinity. *See Figure 4.56 and explain graphically.*

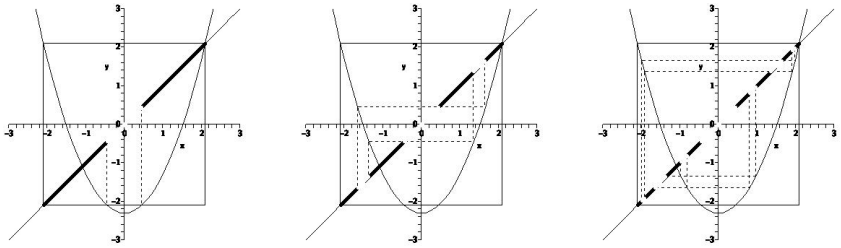


Figure 4.56: Being ejected from the box.

If an orbit starts at x_0 inside the blanked out portion of the diagonal in the first diagram in Figure 4.56, then after one step it is ejected outside the box, and then stays outside forever more. *Why?*

If an orbit starts at x_0 inside the two smaller blanked out portions of the diagonal in the second diagram in Figure 4.56, then after one step it is in the larger blanked out portion, after two steps it is ejected outside the box, and then it stays outside forever more. *Why?*

If an orbit starts at x_0 inside the four smallest blanked out portions of the diagonal in the third diagram in Figure 4.56, then after one step it is in the two smaller blanked out portions of the diagonal, after two steps it is in the larger blanked out portion, after three steps it is ejected outside the box, and then it stays outside forever more. *Why?*

Etc.

In this way we see there is a totally disconnected Cantor style[36] set of points S with the property that if we start an orbit from any point $x_0 \in S$ then the orbit is bounded and in particular stays in the box. The 8 black segments

[35] Once you are out you never get back in. Even if you stay in, the situation is very unstable. There are points arbitrarily close by which will eventually be thrown out.

[36] A *Cantor style set* S is a set with the property that if we join any two distinct points in S by an arc, then there will always be at least one point on the arc which is not in S. In fact there will be infinitely many such points on the arc. A more common terminology is that S is *totally disconnected*.

in the third diagram in Figure 4.56 are just the third approximation to this Cantor type set.

Baby Julia Sets Much of the previous discussion is summarised in Figure 4.57 .

Definition 4.6.1. For each real number c the *baby Julia set* bJ_c is the set of all real numbers x_0 such the orbit for $f(x) = x^2 + c$ starting from x_0 remains bounded.

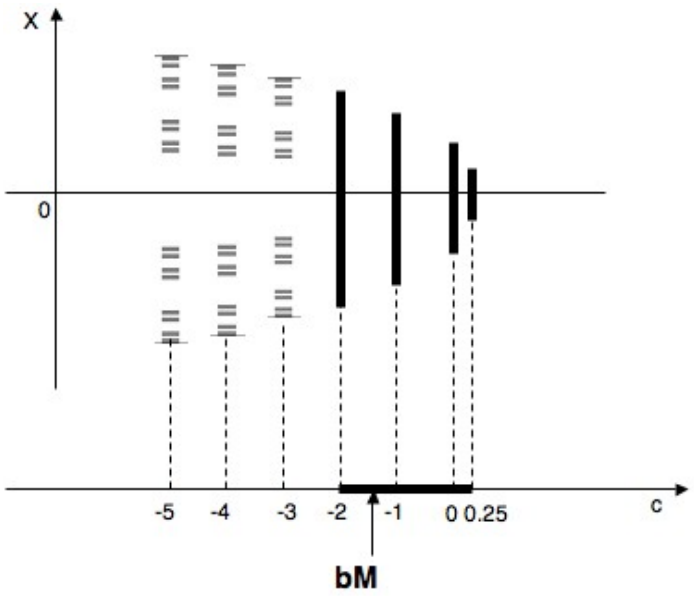


Figure 4.57: The baby Julia set bJ_c corresponding to various values of c is indicated by the "fibre" above c. If $c < -2$ then bJ_c is totally disconnected and is a Cantor style set — of course we can only draw an approximation. If $-2 \leq c \leq .25$ then bJ_c is connected and is an interval. If $c > .25$ then bJ_c is the empty set $\emptyset$. The baby Mandelbrot set $bM = [-2, 1/4]$ is the set of c such that bJ_c is connected and nonempty.

- For $c > 1/4$,
$$bJ_c = \emptyset.$$
- For $-2 \leq c \leq 1/4$, bJ_c is the closed interval
$$bJ_c = \left[-\frac{1+\sqrt{1-4c}}{2},\ \frac{1+\sqrt{1-4c}}{2}\right].$$
For example,
$$bJ_{-2} = [-2, 2],\ bJ_0 = [-1, 1],\ bJ_{\frac{1}{4}} = [-\frac{1}{2}, \frac{1}{2}].$$

- For $c < -2$, bJ_c is a totally disconnected Cantor type subset of

$$\left[-\frac{1+\sqrt{1-4c}}{2}, \frac{1+\sqrt{1-4c}}{2}\right].$$

Theorem 4.6.2. *A baby Julia set is*
- *either a connected subset of* $\mathbb{R}$*, in fact a closed bounded interval, or*
- *it is a totally disconnected set (Cantor type set) or the empty set.*

Proof. We have proved this in so far as we have discussed the various cases in terms of the values of c. □

There is an equivalent way of describing those real numbers c such that the baby Julia set bJ_c is totally disconnected or empty. We only need to examine the orbit starting at 0.

The significance of the point 0 is that $f'(0) = 0$. We say that 0 is a *critical point* for f. Notice that for a quadratic function there is exactly one critical point. *Why? What is the critical point for the logistic map?*

Theorem 4.6.3. *The baby Julia set* bJ_c *is totally disconnected or empty if and only if the orbit for* $f(x) = x^2 + c$ *starting at* 0 *is unbounded. It is connected and non empty if and only if the orbit starting at* 0 *is bounded.*

"Proof" and Discussion. Look at Figure 4.57. Remember that if $x_0 \in bJ_c$ then the orbit starting at x_0 is bounded, and if $x_0 \notin bJ_c$ then the orbit starting at x_0 is unbounded.

For $c < -2$ the baby Julia set bJ_c is totally disconnected.

In this case the orbit starting from 0 is unbounded because $0 \notin bJ_c$. You can also see from the third diagram in Figure 4.56 that the orbit starting from 0 moves outside the box at the next step, and then diverges to $+\infty$.

For $-2 \leq c \leq .25$ the baby Julia set bJ_c is an interval. In this case $0 \in bJ_c$ and so the orbit starting at 0 is bounded. You can also see from the second diagram in Figure 4.55 that the orbit starting from 0 is bounded.

For $c > .25$ the baby Julia set bJ_c is empty. So certainly $0 \notin bJ_c$ and this means the orbit starting at 0 is unbounded. You can also see from the first diagram in Figure 4.56 that the orbit starting from 0 (in fact starting from anywhere) is unbounded. □

The Baby Mandelbrot Set We are now in position to define the baby Mandelbrot set.

Definition 4.6.4. The *Baby Mandelbrot set* bM is:

$$\begin{aligned} bM &= \{c : bJ_c \text{ is connected (and not empty)}\} \\ &= \{c : \text{Orbit starting from the seed 0 is bounded}\} \end{aligned}$$

From Theorem 4.6.3 we see that these two definitions are equivalent. From Figure 4.57 we see that

$$bM = [-2, 1/4].$$

This is not a very interesting set. But the analogous definitions of the Julia sets J_c and the Mandelbrot set M in the complex plane will lead to some *very* interesting sets!

Complex Numbers

I will assume you have looked already at the information on complex numbers in [HoM, pp 462–469].

Complex numbers as points in the plane You can think of complex numbers as a way of representing points in the plane. The complex number $z = x + iy$ corresponds to the point (x, y). So there is nothing "imaginary" about them.

Complex addition Addition of two complex numbers corresponds to vector addition, see the diagram [HoM, p464].

Polar coordinates Another representation of complex numbers is as follows.

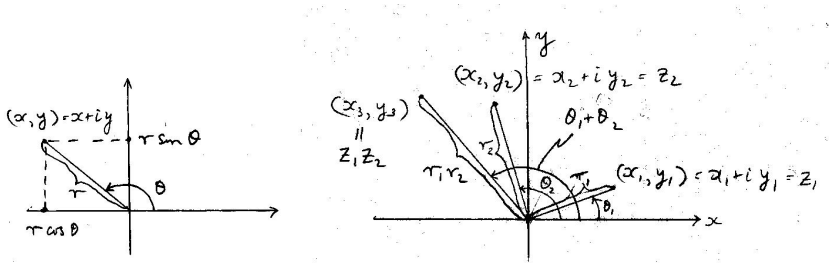


Figure 4.58: Polar coordinates and complex multiplication

Let $z = x + iy = (x, y)$. Suppose the distance of z from the origin is r and the angle in the counterclockwise direction measured from the positive x -axis is θ. See Figure 4.58.

Then by basic trigonometry

$$x = r\cos\theta,\ \ y = r\sin\theta,\ \text{ and so } z = r\cos\theta + ir\sin\theta.$$

We say that z is represented in polar coordinates by r and θ.

Complex multiplication Suppose the two complex numbers z_1 and z_2 are

$$\begin{aligned} z_1 &= x_1 + iy_1 = r_1\cos\theta_1 + ir_1\sin\theta_1, \\ z_2 &= x_2 + iy_2 = r_2\cos\theta_2 + ir_2\sin\theta_2. \end{aligned}$$

See Figure 4.58.

Then by using basic properties of complex multiplication including $i^2 = -1$, and some basic trigonometric formulae,

$$\begin{aligned} z_1 z_2 &= (r_1\cos\theta_1 + ir_1\sin\theta_1)\,(r_2\cos\theta_2 + ir_2\sin\theta_2) \\ &= r_1 r_2\big(\cos\theta_1\cos\theta_2 + i\cos\theta_1\sin\theta_2 + i\sin\theta_1\cos\theta_2 + i^2\sin\theta_1\sin\theta_2\big) \\ &= r_1 r_2\big((\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2)\big) \\ &= r_1 r_2\big(\cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2)\big). \end{aligned}$$

In other words, *multiplying two complex numbers is the same as multiplying their distances from the origin and adding their angles from the positive x-axis.*

As an exercise, *if $z_1 = z_2 = i$, what are $r_1, r_2, \theta_1, \theta_2, x_1, x_2, y_1, y_2$?*

In fact, we could even think of representing points (x, y) in the plane by the "complex" number $x + iy$ and then making the rule $i^2 = -1$, as just being a very convenient way of allowing us to "multiply" points in the plane according to the previous rule of multiplying distances and adding angles!

Julia Sets

The notes here are quite brief. You should look at [HoM, pp 469–471].

For each complex number c we will be interested in iterates of the map $f(z) = z^2 + c$. Just as in Definition 4.6.1 for the baby Julia set we now define the Julia set J_c.

Important Note: What we defined before would more accurately be called the *filled in baby Julia set.* We now define the usual *filled in Julia set* and the *Julia set.*

Definition 4.6.5. For each complex number c the *filled in Julia set* fJ_c is the set of all initial seeds z_0 such the orbit for $f(z) = z^2 + c$ starting from z_0 is bounded.

The *Julia set* J_c is the boundary of the filled in Julia set fJ_c.

The Julia sets are very beautiful.

There are some pictures of filled in Julia sets fJ_c for various values of c in [HoM, pp470, 471]. The filled in Julia sets are shown in black. The Julia sets J_c are the boundary of the filled in Julia set. They can also be thought of as the *interface* between the filled in Julia set and the complement of the filled in Julia set.

The different colours on page 470 in the complement of fJ_c correspond to how long it takes the orbit of a point to move a certain fixed distance from the origin. Remember that the orbit from any point in the complement of fJ_c is unbounded and in fact keeps moving further and further from the origin.

The colours are those of the rainbow: red, orange, yellow, green, blue, indigo and violet (Roy G. Biv). Points marked red are the fastest to move away, those marked violet take much longer and are closer to fJ_c. (Except for the two pictures at the top of p470 where the colouring scheme is reversed.)

For the examples shown the sets fJ_c and J_c are connected except for $c = -0.194 + 0.6557i$, $c = -0.74543 + 0.11301$ and $c = -0.15652 - 1.03225i$. *In these three cases the sets fJ_c and J_c are the same and are totally disconnected.* This is not completely clear due to pixelation effects.

You should use the excellent applet at
`http://math.bu.edu/DYSYS/applets/Quadr.html`
to further examine these and other examples. Notice how changing the number of iterations affects the diagrams. *Why is this?* Also, *follow the link* there to the Mandelbrot Set Explorer.

Properties of the Julia sets Similarly to Theorem 4.6.2 for the (filled in) baby Julia set we have

Theorem 4.6.6. *A filled in Julia set fJ_c is*
- *either connected, or*
- *is a totally disconnected Cantor (dust) style of set.*

Proof. The proof is too long to give here. □

Here are some more properties:

1. The sets fJ_c and J_c are never empty. This is different from the "baby" case.

   In fact, the two solutions of $f(z) = z$ must belong to fJ_c. *Why?* Remember that a quadratic equation always has solutions if we allow complex numbers.

   Moreover, the solutions of $f^2(z) = z$ must also belong to fJ_c. And in fact the solutions of $f^n(z) = z$ must belong to fJ_c for every n. *Why?* .
2. If fJ_c is connected then so is its boundary J_c. If fJ_c is totally disconnected then fJ_c and J_c are the same. (This is a general fact about totally disconnected Cantor style sets.)

Just as in Theorem 4.6.3 for the baby case, there is another way of describing those complex numbers c such that fJ_c is totally disconnected and those complex numbers c such that fJ_c is connected.

Theorem 4.6.7. *The filled in Julia set fJ_c is totally disconnected if and only if the orbit for $f(x) = x^2 + c$ starting at 0 is unbounded. It is connected if and only if the orbit starting at 0 is bounded.*

Proof. Again, it is too long to give here. □

The Mandelbrot Set

Once again, by analogy with Definition 4.6.4 for the baby Mandelbrot set, the Mandelbrot set is defined as follows. The two definitions are equivalent from Theorem 4.6.7.

Definition 4.6.8. The *Mandelbrot set* M consists of certain values of the parameter c and is defined by:

$$\begin{aligned} M &= \{c : J_c \text{ is connected } \} \\ &= \{c : \text{The orbit for } z^2 + c \text{ starting from the seed 0 is bounded}\} \end{aligned}$$

The Mandelbrot set has an incredibly rich structure. If you blow it up then you get more and more amazing patterns. Look at[37]
`http://math.bu.edu/DYSYS/applets/Quadr.html`

[37]This applet will easily allow you to blow up by a factor of 10^{14} for example, which is better than blowing up a postage stamp to a sheet of paper whose width and height are each the distance to the sun and back. And better than blowing up something the size of an atom to a ball with a 10km diameter.

For Julia sets J_c corresponding to points c on or near the boundary of the Mandelbrot set M, it is possible to read off a lot of information about J_c by looking at the Mandelbrot set near c. You will find a lot of information at
`http://math.bu.edu/DYSYS/explorer/page1.html`

There are some great movies at
`http://math.bu.edu/DYSYS/movies.html` . The one called
A Little Trip Through the Mandelbulbs is a tour of the Julia sets as you move around the Mandelbrot set. You can also make your own movies at the previous site
`http://math.bu.edu/DYSYS/applets/Quadr.html`

4.7 Dimensions Which Are Not Integers

The following are preliminary notes only for this section. You should use the book [HoM] as the main source and these notes as a supplement.

Overview

Dimension is a convenient way of measuring the "size" of a set. On the other hand it is also a rather "crude" measure. *Every* "nice" line has dimension one and *every* "nice" surface has dimension two.

On the other hand, fractal sets will usually have a dimension, called the *similarity dimension*, which is not an integer. The similarity dimension is only defined for sets which are *self-similar* in the sense we discussed in Section 4.1 for the Sierpinski triangle, page 152 for the Koch curve, page 153 for the Sierpinski triangle, page 155 for the Menger sponge and page 156 for the Cantor set.

There are also other notions of dimension. The *Hausdorff dimension* is defined for all sets and agrees with the similarity dimension if a set is self-similar. But this is rather technical and we will not have time to explore it.

The *box counting dimension* is very useful for experimentally computing dimensions. If you do a Google you will find lots of information. Box counting dimension is often, but not always, agrees with the Hausforff dimension.

And finally there is the standard *topological dimension*, which is always a integer, and agrees with our usual idea of dimension. But even this is very difficult to define rigorously.

Similarity Dimension

Motivation See the diagram on page 506 of [HoM].

Let

$$N = \text{ no. of copies}, \quad S = \text{ scaling factor}$$

Then we get:

$$\begin{aligned} \text{Line: } & S = 3,\ N = 3\ (= 3^1) \\ \text{Square: } & S = 3,\ N = 9\ (= 3^2) \\ \text{Cube: } & S = 3,\ N = 27\ (= 3^3) \end{aligned}$$

In all cases,

$$N = S^d,$$

where d is the dimension in the common every day sense of "dimension".

Definition of Similarity Dimension Motivated by the previous examples we now make the following Definition:

Definition 4.7.1. Suppose the set E is the union of N copies of itself, and E is obtained from each copy by scaling it up by the factor S. Then the *similarity*

dimension of E is

$$\boxed{d = \frac{\log N}{\log S}}. \tag{4.32}$$

We also need to assume that the copies do not intersect each other as is the case for the Cantor set, or have "minimal overlap" as in the cases of the Sierpinski triangle, Koch curve and Menger sponge. However, we will not make the notion of "minimal overlap" precise here.

In this way we get:

$$\begin{aligned}
\text{Koch Curve: } & S = 3,\ N = 4,\ d = \log 4/\log 3 = 1.261859507\ldots \\
\text{Sierpinski Triangle: } & S = 2,\ N = 3,\ d = \log 3/\log 2 = 1.584962501\ldots \\
\text{Menger Sponge: } & S = 3,\ N = 20,\ d = \log 20/\log 3 = 2.726833027\ldots \\
\text{Cantor Set: } & S = 3,\ N = 2,\ d = \log 2/\log 3 = .6309297534\ldots
\end{aligned}$$

Explain why this is so in each case.

Applications

The dimension of a set is an important way of analysing it, as I mentioned previously.

Dimension of the Universe Experimental observations indicates that over a very large range of scales, the amount of matter in the universe scales something like $(\text{distance})^{1.5}$. This is somewhat paradoxical, but it indicates that in a sense the universe has "dimension" around 1.5.

Computer simulations of 10,000 or more point masses moving under Newtonian gravity and certain other assumptions also gives a similar value.

This is not to be confused with "string theory" models in physics which indicate that our universe is, in the sense of topological dimension, 10 or perhaps more. See `www.columbia.edu/cu/record/23/18/14.html`.

Dimension of Attractors The Lorenz attractor in Figures 4.1 and 4.2 has been estimated to have Hausdorff dimension $2.06 \pm .01$.

Dimensions of Physical Objects In the 1999 paper, *Fractal analysis of surface roughness by using spatial data*[38] Peter Hall and S. Davies at the ANU and CSIRO respectively, wrote in their abstract:

> We develop fractal methodology for data taking the form of surfaces. ... Our results and techniques are applied to analyse data on the surfaces of soil and plastic food wrapping. For the soil data, interest centres on the effect of surface roughness on retention of rain-water, and data are recorded as a series of digital images over time. Our analysis captures the way in which both the fractal dimension and the scale change with rainfall, or equivalently with

[38] Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (1), 337.

time. The food wrapping data are on a much finer scale than the soil data and are particularly anisotropic. The analysis allows us to determine the manufacturing process which produces the smoothest wrapping, with least tendency for micro-organisms to adhere.

Questions

1 Give an example of a Sierpinski type triangle T with a different scaling factor such that the similarity dimension of T is exactly one.

Show how to find a Sierpinski type triangle T_α such that the similarity dimension of T_α is α, for any number $0 < \alpha \leq \log 3/\log 2$.

What do you speculate happens for $\log 3/\log 2 < \alpha \leq 2$? I don't think anyone knows a complete answer.

Chapter 5

Geometry and Topology

This chapter corresponds to some of the sections in Chapters 4 and 5 of [HM]. Although we do not cover all the sections there, those we do are done here in more depth.

Geometry originated from the study of shape and size, such as occurred in measuring farm land or building the pyramids. Geometry now also deals with curved spaces and spaces of any number of dimensions. These profound extensions of geometry arose first in mathematics but are now the basis of the theory of relativity and contemporary theories of the universe, where space has 10 or 11 dimensions. This is yet another example of the fact that the general patterns and structures which arise naturally in mathematics are also the patterns and structures which are essential to model and analyse the universe in which we live.

Topology is the study of the classification of geometric objects, where two objects are considered to be the same if they can be deformed continuously one into the other (actually, this is a simplification of what really happens, but will do for now). In this way, a coffee cup and a doughnut are the same, but there are many more profound applications, some of which we will discuss.

Contents

5.1 Euclidean Geometry and Pythagoras's Theorem

Abstract ideas can be made tangible, and manipulating simple shapes can lead to profound results.

Not in [HM]

Euclidean Geometry

Euclid's Elements The subject of Euclidean Geometry was first written down by Euclid in about 300 B.C. The 13 volumes can be seen in the original Greek and in translation at
`http://aleph0.clarku.edu/ djoyce/java/elements/elements.html`
and
`http://farside.ph.utexas.edu/euclid.html` .

Euclid's Elements has appeared in over 1000 editions and is second only to the Bible in this respect. It is the most used textbook of all times and was used regularly in schools up to early last century. It is largely a collection of theorems due to earlier mathematicians, but Euclid was the first to organise the material in a systematic manner.

The work is profound in a number of ways. It attempted to derive all of plane and three dimensional geometry, and the basic number theory known at that time, including the infinite number of primes, geometric series, irrational numbers, the irrationality of $\sqrt{2}$, and the computation of areas under various curves by an approach related to integration. The method in Euclid's Elements was to begin from a small number (five) of "indisputable" facts or postulates, or what we now call *axioms*, and proceed by the rules of logic to deduce increasingly complicated, and far from obvious, results. It is the first example of the axiomatic approach to mathematics.

Actually, as we now know, the work is flawed. There are quite a few "hidden assumptions" besides the 5 axioms. But these other assumptions are so "obvious" when we think of their geometric interpretation that it is not surprising that it took a long time to appreciate the problem. In fact, it was really only after the discovery of other "curved" or non Euclidean geometries, that the flaws were realised.

Hilbert's Axioms There have been a number of approaches which give a complete set of axioms for Euclidean geometry, and perhaps the best approach is due to Hilbert in the early part of the 20th century. You can download a translation of Hilbert's book at
`http://www.gutenberg.org/etext/17384` .

The significance of Hilbert's axioms is not that Euclidean geometry is best done in this manner (it is not) but that his work considered fundamental matters like the independence, consistency and completeness of the axioms. Math-

ematics now relies on the axiomatic approach, and Hilbert's axiomatisation of geometry was a very important early example of this approach. The contemporary approach to Euclidean geometry is a mixture of using Cartesian coordinates and defining Euclidean geometry to be the study of those properties (such as angle, perpendicularity, length, area, parallelism) which are invariant under rotations and translations.

Euclid's Method Euclid begins with 23 definitions such as point, line, surface, angle, right angle, triangle, circle, parallel.

He then states five postulates (or "axioms")

1. A straight line segment can be drawn by joining any two points.
2. A straight line segment can be extended indefinitely in a straight line.
3. Given a straight line segment, a circle can be drawn using the segment as radius and one endpoint as center.
4. All right angles are congruent. (can be translated and rotated one into another)
5. If two lines are drawn which intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended far enough.

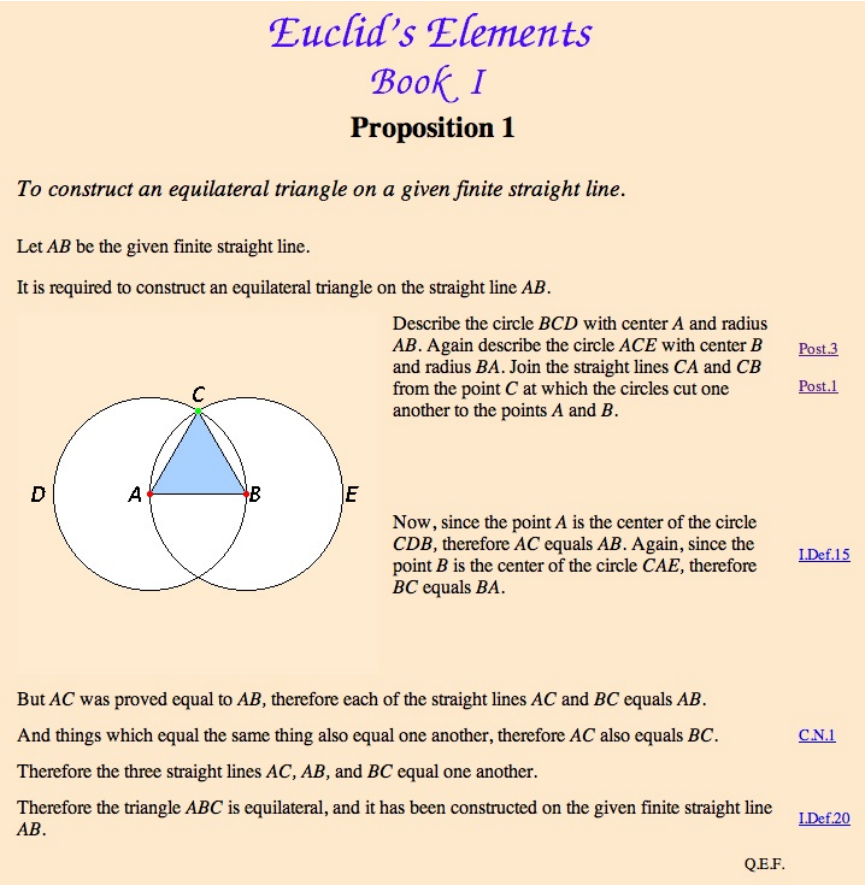
Finally there are five "common notions" (which are today called logical and arithmetical axioms).

1. Things which equal the same thing are equal to one another. (Transitivity of equality)
2. If equals are added to equals, then the sums are equal. (Addition property of equality)
3. If equals are subtracted from equals, then the remainders are equal. (Subtraction Property of Equality)
4. Things which coincide with one another are equal to one another. (Reflexive property of equality)
5. The whole is greater than the part.

We will not develop the subject of Euclidean geometry, with the following exceptions. In Figure 5.1 we show the translation of the First Proposition and its proof. In fact there are already some "problems" with this proof, in that we use more than is in the axioms. For example, it does not follow from the axioms that the two circles meet. There are also other things that do not follow just from Euclid's axioms, but the remarkable fact is that all the results in Euclidean geometry *do* follow from Hilbert's axioms.

Beginning on page 222 we state and prove Pythagoras's theorem. In Section 5.2 we prove the main results on Platonic solids from the final volume of Euclid's Elements.

Remarks on "Proofs" . In this chapter our "geometric 'proofs", like Euclid's, will rely on various facts which are geometrically clear but which we will not establish rigorously from a set of axioms. It is possible to derive everything from either Hilbert's axioms or by using Cartesian coordinates, but this is not very interesting and life is too short anyway.



Euclid's Elements

Book I

Proposition 1

To construct an equilateral triangle on a given finite straight line.

Let *AB* be the given finite straight line.

It is required to construct an equilateral triangle on the straight line *AB*.

Describe the circle *BCD* with center *A* and radius *AB*. Again describe the circle *ACE* with center *B* and radius *BA*. Join the straight lines *CA* and *CB* from the point *C* at which the circles cut one another to the points *A* and *B*. Post.3 Post.1

Now, since the point *A* is the center of the circle *CDB*, therefore *AC* equals *AB*. Again, since the point *B* is the center of the circle *CAE*, therefore *BC* equals *BA*. I.Def.15

But *AC* was proved equal to *AB*, therefore each of the straight lines *AC* and *BC* equals *AB*.

And things which equal the same thing also equal one another, therefore *AC* also equals *BC*. C.N.1

Therefore the three straight lines *AC*, *AB*, and *BC* equal one another.

Therefore the triangle *ABC* is equilateral, and it has been constructed on the given finite straight line *AB*. I.Def.20

Q.E.F.

Figure 5.1: The first Proposition in Euclid's Elements. The marginal comments refer to the relevant Postulate, Definition or Common Notion used to justify the adjacent step.

Pythagoras's Theorem

[HM, 208–211]

See Figure 5.2. This theorem and its proof was known to Pythagoras about 600B.C. But in fact the result, if not the proof, was known to the Babylonians about 1900B.C. The proof we give is due to the Hindi mathematician Bhaskara in the second century A.D. You are asked to find other proofs in the Questions at the end of this section.

Theorem 5.1.1. *The area of a square whose side is the hypotenuse (the longest side) of a right angled triangle is the some of the areas of the two squares whose sides are given by the other two sides of the triangle.*

Proof. Consider any right angled triangle T as in Figure 5.3 where c is the length of the hypotenuse and a and b are the lengths of the other two sides.

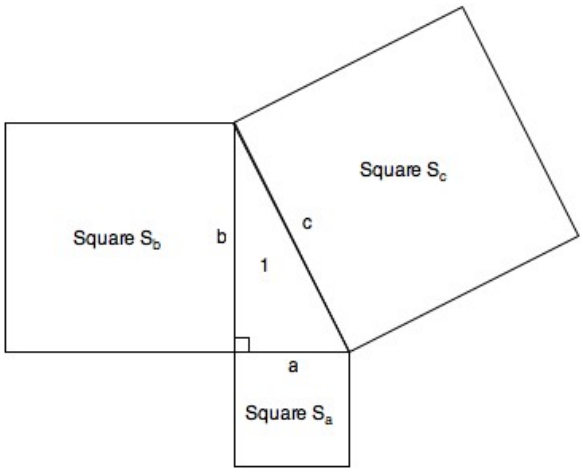


Figure 5.2: Pythagoras's Theorem. The area of S_c is the sum of the areas of S_a and S_b. Algebraically, $c^2 = a^2 + b^2$.

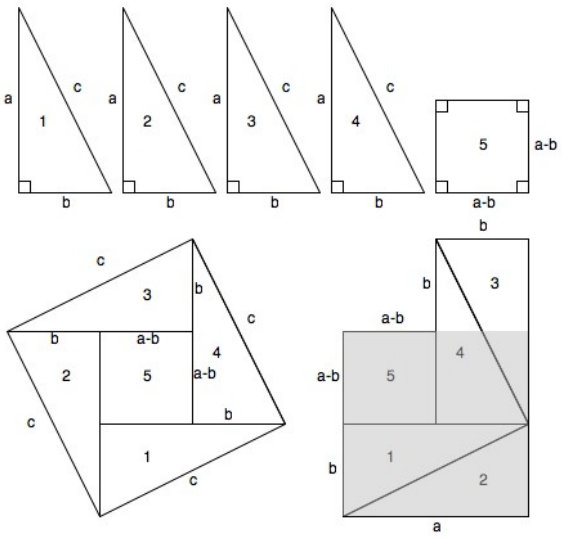


Figure 5.3: Proof of Pythagoras's Theorem by "rearrangement".

The four copies of T and the small square in Figure 5.3 are first rearranged to form a square of side c; see the large square in Figure 5.3.

By translating and rotating triangles 2 and 3 we next obtain an L shaped figure which can be decomposed into the shaded square of side a and the remaining unshaded square of side b.

It follows that the area of a square of side c is the sum of the areas of a square of side a and a square of side b. □

If we really wanted to make the previous proof a rigorous one from some axioms, we would have to prove or have an axiom which says that areas are not changed by translations and rotations. We would also need to prove that the sum of the angles of a triangle is 180° and in particular that the two acute angles of a right angled triangle add to give a right angle.

Questions

1 Using only the first rearrangement in Figure 5.3, prove Pythagoras's Theorem by computing the area of the four triangles and the area of the small square.

2 Use Figure 5.4 to give another proof of Pythagoras's Theorem.

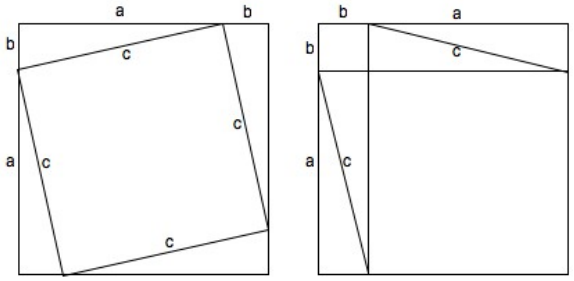


Figure 5.4: Another proof of Pythagoras's Theorem.

3 Do [HM, p216, Q20].

4 Do [HM, p216, Q21].

5 Do [HM, p216, Q21].

5.2 Platonic Solids and Euler's Formula

Overview

This section corresponds to

What are They? The Platonic solids are the most symmetric, regular (and aesthetic) solid objects, apart from the solid ball. There are exactly 5 Platonic solids — the tetrahedron, cube or hexahedron, octahedron, dodecahedron and icosahedron,[1] see Figure 5.5.

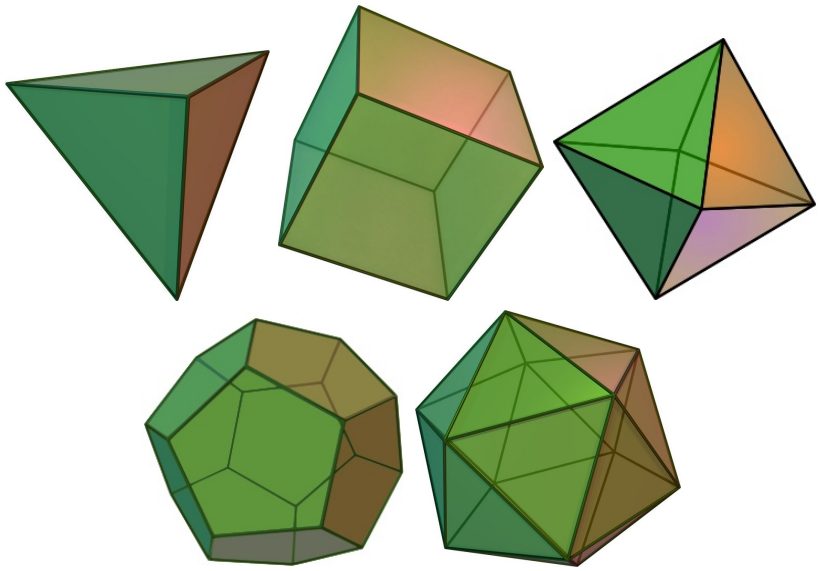


Figure 5.5: The five Platonic solids: Tetrahedron, Hexahedron (Cube), Octahedron, Dodecahedron, Icosahedron.

History The Platonic solids were constructed in the thirteenth and final volume of Euclid's Elements, and the fact that these are the only possible Platonic solids was also proved there. These results were probably due to Theaetetus (ca. 417 – 369 B.C.) although Pythagoras (ca. 575–495 B.C.) probably knew of the existence of the first four.

The philosopher Plato in about 360 B.C. associated four of the Platonic solids with the elements earth, air, fire and water. Fire is the tetrahedron as it is sharp and jagged, earth is the cube as it is not very stable and crumbles easily, air is the octahedron since it slides easily (O.K., that does not seem so well justified), and water is the icosahedron since it is almost spherical as are

[1] The names indicate the number of faces. Here are some Greek numerical prefixes: mono (1), di (2), tri (3), tetra (4), penta (5), hexa (6), hepta (7), octa (8), ennea (9), deca (10), hendeca (11), dodeca (12), icosa (20).

beads of water. The dodecahedron is left out so Plato suggested the gods used it to arrange the constellations in the heavens!

In the sixteenth century Kepler postulated that if the 5 Platonic solids were arranged with the octahedron inside the icosahedron inside the dodecahedron inside the tetrahedron inside the cube, so that the circumscribed sphere of one is the inscribed sphere of the next, then the orbits of the 6 planets Mercury, Venus, Earth, Mars, Jupiter, and Saturn lay on the 6 spheres so obtained. This was not a highly successful theory, particularly since Uranus was later discovered but there were no more Platonic solids. Kepler's laws of planetary motion have been somewhat more enduring.

Examples in Nature and Applications Three platonic solids occur naturally as crystals — the cube (e.g. halite), the octahedron (e.g. spinels) and the dodecahedron (e.g. garnet). Certain species of Radiolaria which occur in plankton produce skeletons corresponding to the Platonic solids and are named accordingly; for example Ircoporus octahedrus, Circogonia icosahedra, Lithocubus geometricus and Circorrhegma dodecahedra. Numerical models of the earth used in meteorological simulations are sometimes based on an icosahedron instead of using latitude and longitude coordinates, as this gives a more uniform grid near the poles.

The Mathematics We will construct the 5 Platonic solids and give the proof from Euclid that there are no other Platonic solids.

We will also give a proof due to Cauchy in 1809 of the formula of Euler (1707–1783), which leads to another proof that there are only 5 Platonic solids. Euler's formula and its generalisations are very important in the subject of topology, and we will discuss this here and in a later section.

Polygons

Properties of Polygons A *polygon* is a planar figure bound by straight line segments.

A *convex* polygon is a polygon with the property that if a straight line segment connects any two points in the polygon (including its boundary) then the segment lies entirely within the polygon.

✍ *Draw a couple of examples of polygons that are not convex.*

The *external angle* at a vertex of a polygon is the angle through which, when traversing the perimeter of the polygon in a counterclockwise direction, the edge leading into the vertex would rotate if it were to point in the same direction as the edge leading out of the vertex. See Figure 5.6.

The *internal angle* at a vertex of a polygon is 180° minus the external angle.

✍ *For any polygon, the sum of the external angles is* 360°. *Why?*

✍ The external angle could be negative for a polygon, but not for a convex polygon. *Draw an example.* In this case the internal angle is larger than 180°.

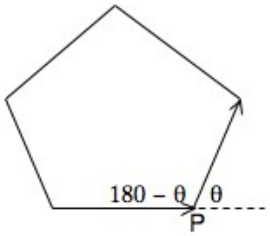


Figure 5.6: A polygon with external angle θ, and internal angle $180° - \theta$, at the vertex P.

Regular Polygons

Definition 5.2.1. A *regular polygon* is a polygon for which all sides are equal and all internal angles are equal.

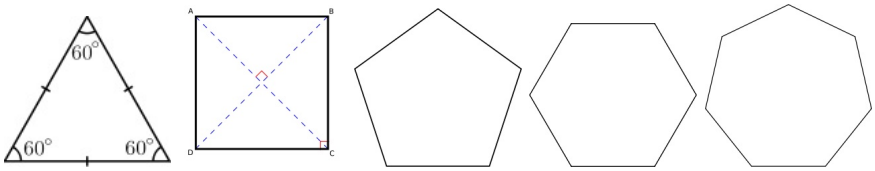


Figure 5.7: Regular Polygons: Equilateral Triangle, Square, Regular Pentagon, Regular Hexagon, Regular Heptagon, etc.

There is one regular polygon with n sides for each integer $n \geq 3$. So there are infinitely many regular polygons. See Figure 5.7.

Notice that the regular polygons are all convex. I am not asking for a rigorous proof, but *can you give an explanation of why this is so?* ✍

Theorem 5.2.2. *If a regular polygon has n sides then the internal angle at each vertex is* $180° - \dfrac{360°}{n}$.

Proof. Let ϕ be the internal angle at each vertex. Then the external angles are $180° - \phi$.

Since the sum of the n external angles is $360°$,

$$n(180° - \phi) = 360°,$$

and so

$$\phi = 180° - \frac{360°}{n}.$$

□

What is the internal angle at each vertex of an equilateral triangle, square, regular pentagon, regular hexagon, regular heptagon? ✍

Platonic Solids and Their Construction

Polyhedra Just as there is a notion of a polygon in two dimensions, so there is an analogous notion of a polyhedron in three dimensions.

A *polyhedron* is a solid figure bounded by polygonal faces and straight edges.

A *convex* polyhedron is a polyhedron with the property that if a straight line segment connects any two points in the polyhedron (including its boundary) then the segment lies entirely within the polyhedron.

What are Platonic Solids? Platonic solids are the analogue in three dimensions of regular polygons in two dimensions.

Definition 5.2.3. A *Platonic solid* is a convex[2] polyhedron such that all the faces are congruent[3] and the solid internal angles at each vertex are all equal.[4]

Are there any Platonic Solids? What are they?

It turns out that there are exactly 5 Platonic solids, see Figure 5.5.[5] This contrasts with we saw before that there are infinitely many regular polygons. We will give two different proofs of this important fact.

Coordinates The 4 vertices of the tetrahedron can be taken to be

$$(1,1,1),\ (-1,-1,1),\ (-1,1,-1),\ (1,-1,-1).$$

The 8 vertices of the cube can be taken to be[6]

$$(\pm 1, \pm 1, \pm 1).$$

The 6 vertices of the tetrahedron can be taken to be

$$(\pm 1,0,0),\ (0,\pm 1,0),\ (0,0,\pm 1).$$

The 20 vertices of the dodecahedron can be taken to be

$$(\pm 1,\pm 1,\pm 1),\ (0,\pm 1/\phi,\pm\phi),\ (\pm 1/\phi,\pm\phi,0),\ (\pm\phi,0,\pm 1/\phi),$$

where $\phi = \dfrac{1+\sqrt{5}}{2}$ is the golden ratio.

The 12 vertices of the icosahedron can be taken to be

$$(0,\pm 1,\pm\phi),\ (\pm 1,\pm\phi,0),\ (\pm\phi,0,\pm 1).$$

Foldouts It is fairly clear that we could construct the tetrahedron, cube and octahedron from the "foldouts" in Figure 5.8, and that the faces would indeed fit together exactly. *Think about this.*

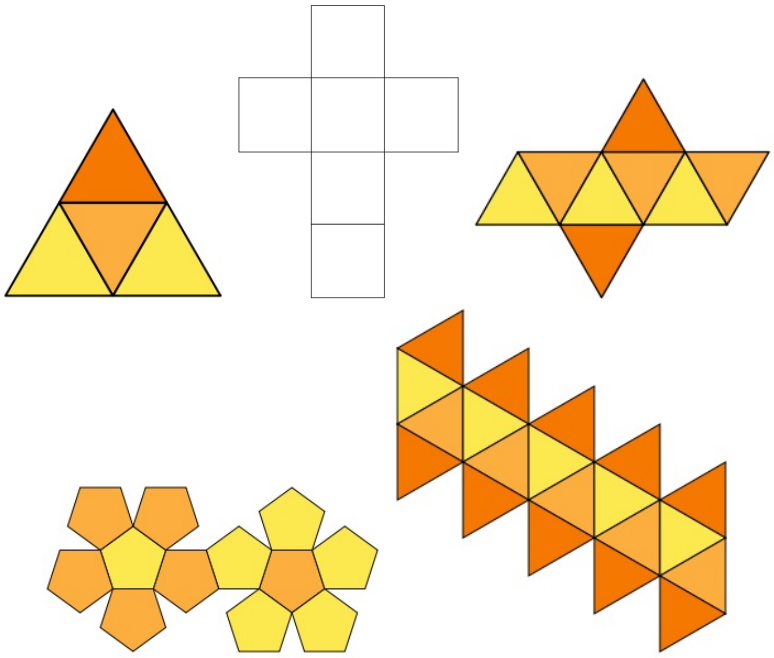


Figure 5.8: Foldouts for the tetrahedron, cube, octahedron, dodecahedron and icosahedron.

However, it is less clear that the other two foldouts in Figure 5.8 will fit together exactly to give the dodecahedron and the icosahedron. One needs to do some calculations to show this. But we will not stop to do it.

There are cutouts with tabs which you can download and assemble at `http://www.worksheetworks.com/math/geometry/polyhedra.html` .

Duality If you connect the centres of the six faces of the cube as in the second diagram of Figure 5.9,[7] you will obtain an octahedron such that each vertex of the octahedron is a face of the cube. Conversely, if you connect the centres of the faces of an octahedron you obtain a cube. We say that the cube and the octahedron are *dual*.

[2]In fact convexity follows from the fact that the solid internal angles at each vertex are all equal. To see this imagine a plane from a long way out which is moved parallel to itself until it first touches the Platonic solid at some vertex. The internal angle at this vertex must "point outwards", and since all internal angles are the same they must all point outwards. It can be shown from this that the Platonic solid is convex.

[3]Two figures are *conguent* if they are essentially identical copies of each other. More precisely, if each can be obtained from the other by a composition of translations and rotations.

[4]Two solid angles are equal if each can be obtained from the other by a composition of translations and rotations.

[5]This and many of the other diagrams in this section are taken from Wikipedia.

[6]By $(\pm 1, \pm 1, \pm 1)$ is meant the 8 possibilities $(1,1,1)$, $(1,1,-1)$, $(1,-1,1)$, $(1,-1,-1)$, etc. Similar comments apply in other cases.

[7]See `www.math.uiowa.edu` .

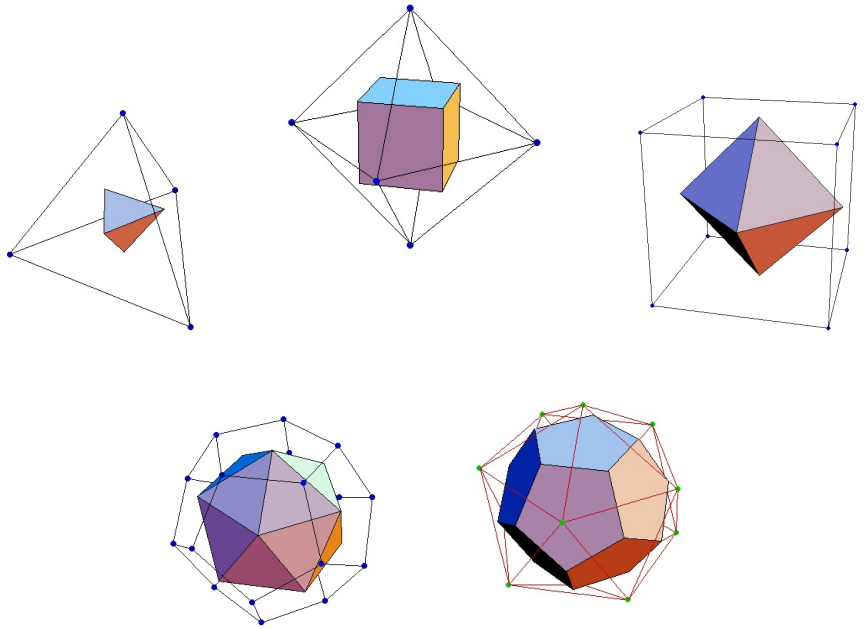


Figure 5.9: The tetrahedron is self dual, the octahedron and the cube are dual, the dodecahedron and the icosahedron are dual.

Similarly, the tetrahedron is *self dual*. The dodecahedron and icosahedron are dual.

Note that on passing from one Platonic solid to its dual, the number of vertices and the number of faces are switched. See Table 5.2. *Explain why this is so without referring to the Table.*

✍

Counting Vertices, Edges and Faces

Doing the Count If you count the number of vertices, edges and faces of the Platonic solids you should obtain the first 4 columns in Table 5.2. The q and p columns should be clear.

It is easy to count the number of faces by using the cutouts in Figure 5.8. You will probably go a little cross-eyed trying to count the number of vertices or edges in the case of the dodecahedron or icosahedron. However, the following formulae (5.2) allow us to compute the number of edges and the number of vertices from the number of faces and some other simple information.

A Combinatorial Relation

Theorem 5.2.4. *For any Platonic solid let q denote the number of edges (and of faces) at each vertex and let p denote the number of edges (and of vertices)*

Name	V	E	F	q	p	$V-E+F$
Tetrahedron	4	6	4	3	3	2
Cube	8	12	6	3	4	2
Octahedron	6	12	8	4	3	2
Dodecahedron	20	30	12	3	5	2
Icosahedron	12	30	20	5	3	2

Table 5.1: Vertices (V), edges (E), faces (F), edges or faces at a vertex (q), edges or vertices of a face (p) and Euler Characteristic ($V-E+F$), for the Platonic solids.

of each face. Then

$$E=\frac{pF}{2},\quad V=\frac{pF}{q}, \tag{5.1}$$

where V is the number of vertices, E is the number of edges and F is the number of faces.

Proof. (*Follow this proof through in the case of one of the three simplest Platonic solids to help your understanding.*)

If we multiply the number F of faces by the number p of *edges* per face, then the number pF which we obtain is *not* the number of edges. Each edge is counted *twice*, since each edge occurs in exactly two faces. So the total number of edges is $pF/2$. This proves the first formula in (5.1).

If we multiply the number F of faces by the number p of *vertices* per face, then the number pF which we obtain is *not* the number of vertices. Each vertex is counted q times since each vertex occurs in exactly q faces. So the total number of vertices is pF/q. This proves the second formula in (5.1). □

In (5.1) we saw that if we know p and q, then we can compute V and E from F. From this it follows by a little algebra that if we know p and q then we can also compute E and F from V, and also F and V from E. Namely

$$\begin{aligned} V&=\frac{pF}{q},\quad E=\frac{pF}{2},\\ E&=\frac{qV}{2},\quad F=\frac{qV}{p},\\ F&=\frac{2E}{p},\quad V=\frac{2E}{q}. \end{aligned} \tag{5.2}$$

Check this. It is just a line or two. ✍

Exactly Five Platonic Solids; Euclid's Proof

We have already seen that there are at least 5 Platonic solids and have described them. We want to show there are no more.

As usual we will give a "geometric proof" relying on some (hopefully) geometrically clear facts. This is essentially the proof in Euclid's Elements.

Theorem 5.2.5. *There are exactly 5 Platonic Solids.*

Proof. Suppose P is any Platonic solid.

Let q be the number of faces or edges of P which meet at each vertex.

Let the internal angle made by the two edges at each vertex of each face be ϕ. For example, if the faces are all equilateral triangles then $\phi = 60°$, $\phi = 90°$ for squares, $\phi = 108°$ for pentagons, $\phi = 120°$ for hexagons, etc. See Theorem 5.2.2.

The first important fact is that

$$q\phi < 360°. \tag{5.3}$$

✍ *Check that this is true for the 5 Platonic solids we already know by writing down the value of $q\phi$ in each case.*

To see this would be true for *any* Platonic solid P let X be a vertex of P and let S be a plane through X such that P lies on one side of S. See Figure 5.10 where $q = 5$. (Such an S exists since P is convex.) If we cut along

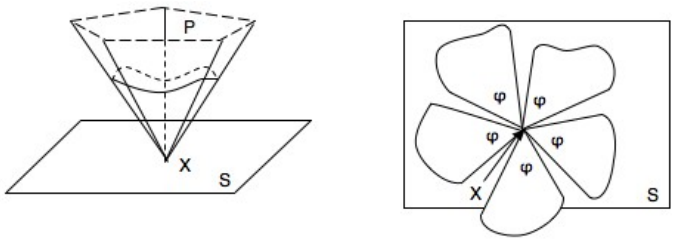


Figure 5.10: Each angle at P equals ϕ. Cut out the angles along the curved line and the edges meeting at P. Lay the angles down flat to see their sum is $< 360°$.

the edges meeting at X and lay them flat on S, we see they do not cover a full 360° around X, and so the sum of the q angles ϕ is *less* than 360°. This proves (5.3).

We now use (5.3) to investigate the various possible values for q and ϕ.

✍ First note that $q \geq 3$. *Why?*

✍ If the faces are equilateral triangles then $\phi = 60°$ and so the only possible values of q in (5.3) are 3, 4 or 5. *Why?* These cases occur for the tetrahedron, octahedron and icosahedron respectively. See Table 5.2.

✍ If the faces are squares then $\phi = 90°$ and the only possible value of q in (5.3) is 3. *Why?* This case occurs for the cube.

✍ If the faces are pentagons then $\phi = 108°$ and the only possible value of q in (5.3) is 3. *Why?* This case occurs for the dodecahedron.

To summarise: we have shown that there are exactly 5 possibilities for the pair (q, ϕ) and that each of these possibilities actually occurs.

But could two *different* Platonic solids P_1 and P_2 have the same values of q and ϕ? The answer is *NO* because of the following informal geometric argument.

First note that since the angle ϕ is the same for both P_1 and P_2, the polygonal faces of P_1 and P_2 must be the same shape — either all are equilateral triangles, all are squares or all are pentagons. After rescaling, the faces of P_1 and P_2 will also have the same edge length. Since both q and ϕ are the same in each case, if X_1 and X_2 are vertices of P_1 and P_2 it follows that P_1 and P_2 will be congruent near X_1 and X_2. Since the distance from X_1 and X_2 to the neighbouring vertices is the same in both cases, we can extend the congruence out past these neighbouring vertices. Repeating the argument a few times we eventually see that P_1 and P_2 are congruent.

So to summarise: there are exactly 5 possibilities for q and ϕ and each of these 5 possibilities leads to *exactly one* Platonic solid. Thus there are exactly 5 Platonic solids. □

Exactly Five Platonic Solids; Euler's Proof

The following theorem is very important. It has many extensions and applications as we will discuss later. Just by counting as in Table 5.2 we can confirm directly that the theorem is true for the 5 Platonic solids.

Theorem 5.2.6 (Euler's Formula)**.** *Let P be* any *convex polyhedron (not necessarily a Platonic solid) and let V be the number of vertices, E the number of edges and F the number of faces. Then*[8]

$$V - E + F = 2. \tag{5.4}$$

Proof. First Step: Remove one of the faces of P and deform the result to obtain a "planar network" N of vertices, edges and faces. The edges of the planar network N may be curved. In Figure 5.11 this is shown for the tetrahedron, cube and octahedron.

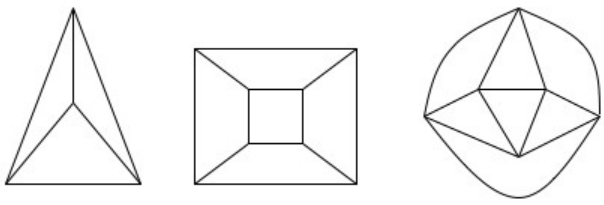


Figure 5.11: Remove one face from the tetrahedron, hexahedron and octahedron respectively, and deform the result to make a planar network. The resulting edges may be curved.

[8]To remember the order in Euler's Formula, note that the vertices are 0-dimensional and come first, the edges are 1-dimensional and come next, while the faces are 2-dimensional and come last.

In the first diagram in Figure 5.12 there is another example of a planar network that could arise in this manner.

We will often use the word "region" instead of "face" when we are discussing a planar network. When we speak of the number of regions (or faces) of the planar network we will include the unbounded region which lies outside all the edges.

We will again use the symbols V, E and F for the number of vertices, edges and regions of the planar network.

It is clear that the number of vertices and edges is unchanged when we pass from P to N. Moreover the number of faces for P is the same as the number of regions for N. The face removed from P corresponds to the unbounded region for N.

So the quantity $V - E + F$ is unchanged in passing from the convex polyhedron to the planar network.

We will show that

$$V - E + F = 2 \tag{5.5}$$

for any planar network obtained in this manner. From this it follows that $V - E + F = 2$ for any convex polyhedron P.

Second Step: Connect vertices of the planar network by new edges so that all regions, except the unbounded region, are *triangular*, possibly with curved edges. This is done in passing from the first diagram in Figure 5.12 to the second. Each time a new edge is introduced in this manner the number V

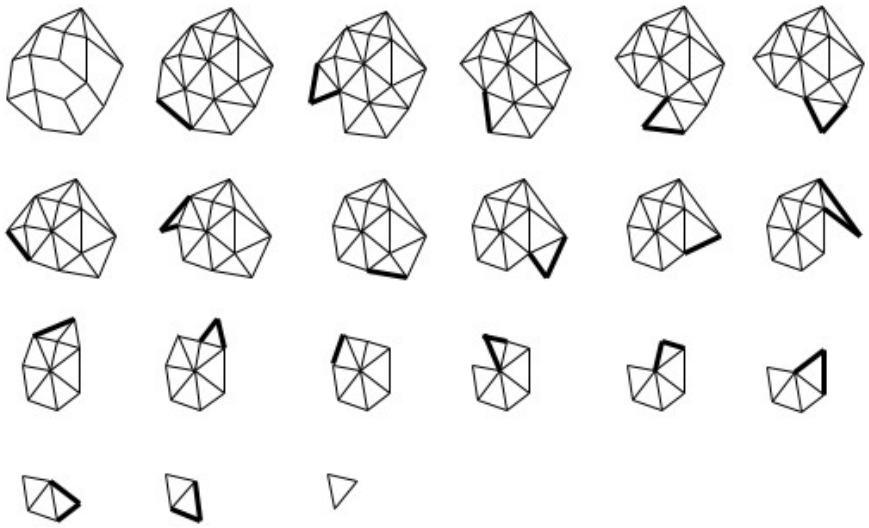


Figure 5.12: The planar network is first changed into a planar network of triangular faces. Triangles with one or two consecutive outside edges are then progressively removed. The quantity $V - E + F$ is unchanged by this process.

is unchanged, the number E is increased by one, and the number F is also increased by one. *Why?* So the number $V - E + F$ is unchanged.

Third Step: Repeatedly remove triangles in the planar network with one or two consecutive "outside" edges. In Figure 5.12 this is done by starting at the second diagram, moving from left to right and then down the rows.

1. Suppose the removed outer triangle has one outside edge, as is the case in moving from the second to the third diagram, the fourth to the fifth and the seventh to the eighth. Then V is unchanged while E and F are each reduced by one, so $V - E + F$ is unchanged.
2. Suppose the removed outer triangle has two consecutive outside edges, as is the case in moving from the third to the fourth diagram, the fifth to the sixth and the sixth to the seventh. Then V is reduced by one, E is reduced by two, and F is reduced by one, so $V - E + F$ is unchanged.

We keep doing this until finally one triangle is left, in which case

$$V - E + F = 3 - 3 + 2 = 2.$$

Since $V - E + F$ was unchanged at each step in changing the planar network, and since after the last step $V - E + F = 2$, this proves (5.5) and so proves the Theorem. □

Remark: In [HM: Section 5.3] the Euler Formula (5.5) is proved for slightly more general planar networks than we do in Steps 2 and 3 of the previous Theorem. In fact all that is required of the network is that it be "connected". This means that any two vertices can be connected by a sequence of edges. We will return to this in a later section.

Theorem 5.2.7. *There are exactly 5 Platonic Solids.*

Proof. From Euler's Formula (5.4) we have

$$V - E + F = 2.$$

From the third line in (5.2),

$$V = \frac{2E}{q} \text{ and } F = \frac{2E}{p}, \tag{5.6}$$

and so

$$E\left(\frac{2}{q} - 1 + \frac{2}{p}\right) = 2. \tag{5.7}$$

For this to happen we must have

$$\frac{2}{p} + \frac{2}{q} - 1 > 0,$$

why? That is,

$$\frac{2}{p} + \frac{2}{q} > 1. \tag{5.8}$$

Because p is the number of edges of each face and q is the number of edges at each vertex, both $p \geq 3$ and $q \geq 3$.

It follows from this that there are not many possible values for p and q such that (5.8) is true.

In fact:

1. if $p = 3$ then $q = 3$, 4 or 5;
2. if $p = 4$ then $q = 5$;
3. if $p = 5$ then $q = 3$;
4. finally, it is not possible that $p \geq 6$.

✍ *Explain why 1–4 are true.*

There are thus only 5 possible cases:

1. if $p = 3$ and $q = 3$ then $E = 6$ from (5.7) and then $V = F = 4$ from (5.6) — this gives the tetrahedron;
2. if $p = 3$ and $q = 4$ then $E = 12$ from (5.7) and then $V = 6$ and $F = 8$ from (5.6) — this gives the octahedron;
3. if $p = 3$ and $q = 5$ then $E = 30$ from (5.7) and then $V = 12$ and $F = 20$ from (5.6) — this gives the icosahedron;
4. if $p = 4$ and $q = 3$ then $E = 12$ from (5.7) and then $V = 8$ and $F = 6$ from (5.6) — this gives the cube;
5. if $p = 5$ and $q = 3$ then $E = 30$ from (5.7) and then $V = 20$ and $F = 12$ from (5.6) — this gives the dodecahedron.

The fact none of these 5 cases can give more than one Platonic solid is even more direct than in the proof of Theorem 5.2.5. In each of the 5 cases we know the exact number of faces, we know their shape and the angles at each vertex (it is given by p), and we know the number of edges and vertices. This is enough to completely determine the polyhedron up to scaling. □

⋆*Groups of Rotations*

(This will not make much sense unless you have done a course on Groups! And even then, the following comments are briefly explained and intended just to give you the flavour of what is happening.)

There are often important connections between geometry and algebra. In the case of the Platonic solids one connection is via the *Rotation Group* of transformations of each Platonic solid. The rotation group is the set of rotations which send the solid into itself. The reason that the rotations form a group is that the composition of two rotations is a rotation and the inverse of a rotation is also a rotation.

Slightly more general is the *Symmetry Group* of transformations of each Platonic solid. This group consists of transformations obtained from composing rotations from the rotation group with reflections in any plane which divides the Platonic solid into two symmetric halves. It can be shown that every such transformation is the same as a rotation or a rotation followed by reflection in a single fixed plane. For this reason the *order* of (i.e. number of elements in) the symmetry group for any Platonic solid is twice the order of the corresponding rotation group.

One important fact is that the rotation group and the symmetry group will be the same for each Platonic solid as for its dual. Rotating or reflecting the cube into itself gives a way of rotating or reflecting the octahedron into itself and conversely, as is probably clear from looking at Figure 5.9. A similar comment applies to the other dual pair consisting of the dodecahedron and the icosahedron.

The Tetrahedron. What are the different ways of rotating the tetrahedron into itself?

Label the vertices of the tetrahedron by A, B, C and D. The vertex A can be rotated into 4 vertices, including itself. After this there are three choices for vertex B. But the locations of vertices C and D are then determined.

So the rotation group for the tetrahedron has order $4 \times 3 = 12$.

The symmetry group for the tetrahedron has order $2 \times 12 = 24$.

The Cube and the Octahedron. What are the different ways of rotating the cube into itself?

There is the identity transformation which leaves everything fixed, giving 1 possibility.

There are 3 pairs of opposite faces and it is possible to rotate about the axis through the centres of each pair by $90°$, $180°$ or $270°$, giving $3 \times 3 = 9$ possibilities.

There are 6 pairs of opposite edges and it is possible to rotate by $180°$ about the axis through the centres of each pair, giving 6 possibilities.

There are 4 pairs of opposite vertices (corresponding to the 4 diagonals) and it is possible to rotate by $120°$ or $240°$ about these diagonals, giving $4 \times 2 = 8$ possibilities.

Adding all this gives 24 rotations. This gives all rotations of the cube, and hence also of the octahedron. So the rotation group of the cube and also of the octahedron has order 24.

The symmetry group for the cube or the octahedron has order $2 \times 24 = 48$.

The Dodecahedron and the Icosahedron. What are the different ways of rotating the dodecahedron into itself?

There is the identity transformation which leaves everything fixed, giving 1 possibility.

There are 6 pairs of opposite faces and for each pair there are 4 rotations about the axis through their centres by multiples of $360°/5 = 72°$, giving $6 \times 4 = 24$ possibilities.

There are 15 pairs of opposite edges and for each pair there is one rotation of $180°$ about the axis through their centres, giving 15 possibilities.

There are 10 pairs of opposite vertices and for each there are two rotations of $120°$ and $240°$ about the diagonal connecting them, giving $10 \times 2 = 20$ possibilities.

Adding all this gives 60 different rotations. So the rotation group for the dodecahedron and the icosahedron has order 60.

The symmetry group for the dodecahedron or the icosahedron has order $2 \times 60 = 120$.

Questions

1 A *regular n-gon* is a regular polygon with n sides, where $n \geq 3$.

Find formulae for the following quantities for a regular n-gon if the distance from the centre to any vertex is r:

1. The angle subtended at the centre by each edge;
2. Each interval and external angle;

3. The length of each edge and the circumference;
4. The area.

2 Let p and q be as usual, the number of edges (and of vertices) of each face and the number of edges (and of faces) at each vertex respectively. Prove from Euler's formula (5.4) and Theorem 5.2.4 that, for any Platonic solid, the number of vertices, edges and faces are given by the formulae

$$V = \frac{4p}{4-(p-2)(q-2)},\ E = \frac{2pq}{4-(p-2)(q-2)},\ F = \frac{4q}{4-(p-2)(q-2)}.$$

3 Question 17 p286 of [HM].

4 Question 18 p286 of [HM].

5 Question 19 p286 of [HM].

6 Question 20 p287 of [HM].

7 Prove that it is not possible to have a convex polyhedron built out of 60 triangles with the property that all vertices have the same number of triangles coming out of them.

HINT: Let n be the number of edges coming out of each vertex. How many edges and how many vertices are there? Now apply Euler's formula.

8 Question 35 p372 of [HM]. This is tricky! Notice that from the soccer ball, by joining the vertices by straight lines, we can construct a convex polyhedron made up of pentagons and hexagons. (These pentagons and hexagons need not be regular.) Moreover each pentagon is surrounded by 5 hexagons.

Let P be the number of pentagons and H be the number of hexagons. What information can you get by carefully applying Euler's formula?

5.3 Visualising the Fourth Dimension

[HM, pp307–320]

What is Dimension?

There are a number of different but related notions of "dimension" within mathematics.

Informally, a line should be one dimensional, a surface two dimensional and a solid object three dimensional.

In Section 4.7 we discussed the idea of "similarity dimension" and we saw that it need not be an integer.

In this section we will be discussing a notion of dimension which is much closer to our usual informal idea of dimension.

Dimension One Think of a straight line which continues indefinitely in both directions. See Figure 5.13. Fix a base point 0 on the line and another point

Figure 5.13: A line is one dimensional. The point P has coordinate 2.3.

1 to give a "unit of length". We can think of the line as a "numbered axis". Every point P on the line will then have a unique real number a associated with it and every real number will correspond to a unique point. We can think of the real number a as the *address* or *coordinate* of the point P. We often just write $P = a$ as in Figure 5.13.

Since any point P on the line is given by *one* piece of information, namely the real number a, we say the line is *one dimensional*.

Dimension Two Next think of a plane which continues indefinitely in all directions. See Figure 5.14. Fix a base point in the plane and two perpendicular

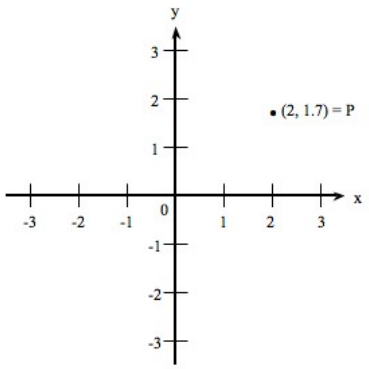


Figure 5.14: A plane is two dimensional. The point P has coordinates $(2, 1.7)$.

numbered axes passing through this point. We often call these axes the x and y axes respectively. The unit of length should be the same on both axes.

Each point P on the plane will have a unique pair of real numbers (a, b) associated with it and every pair of real numbers will correspond to a unique point. We can think of the numbers (a, b) as the *address* or *coordinates* for P — go distance a along the x-axis (right if $a > 0$ and left if $a < 0$) and then distance b vertically (up if $b > 0$ and down if $b < 0$).

Since a point P on the plane is given by *two* pieces of information, namely the pair of real numbers (a, b), we say the line is *two dimensional*.

Dimension Three Finally think of the "3-dimensional space" in which we live. See Figure 5.15. Fix a base point and three perpendicular numbered

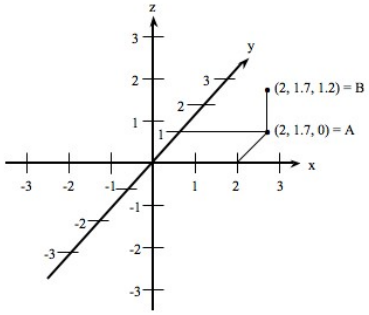


Figure 5.15: Space is three dimensional. The point P has coordinates $(2, 1.7)$.

axes passing through this point. We often call these axes the x, y and z axes respectively. The unit of length should be the same on all axes.

Each point P in space will have a unique triple of real numbers (a, b, c) associated with it and every triple of real numbers will correspond to a unique point. We can think of the numbers (a, b, c) as the *address* or *coordinates* for P — go distance a along the x-axis (direction of the arrow if $a > 0$ and opposite direction if $a < 0$), then distance b parallel to the y-axis (direction of the arrow

if $b > 0$ and opposite direction if $b < 0$) and finally distance c parallel to the z-axis (direction of the arrow if $c > 0$ and opposite direction if $c < 0$).

Since a point P in space is given by *three* pieces of information, namely the triple of real numbers (a, b, c), we say space is *three dimensional*.

Geometric and Analytic Representations

What are Lines, Planes and Space? We can think of a straight line, a plane, or 3D space itself as representations of idealised physical entities[9] from the world in which we live.

From a mathematical perspective we can think of a straight line, a plane, or 3D space as geometric objects given by Euclid's axioms, or more precisely by Hilbert's axioms, see page 220. Alternatively we can think of them as the set of real numbers, the set of pairs of real numbers, or the set of triples of real numbers.[10]

Coordinate Formulation The identification between a line and the set $\mathbb{R}$, between a plane and the set $\mathbb{R}^2$ of pairs of real numbers, and between space and the set $\mathbb{R}^3$ of triples of real numbers, allows us to discuss geometric concepts and ideas in terms of real numbers. This is what we do when we are doing *coordinate geometry*, also called *analytic geometry* or *Cartesian geometry*.

For example, the line through the points $(0, 0)$ and $(1, 2)$ in a plane can be described as the set of points with coordinates (x, y) for which $y = 2x$.

The *square* in Figure 5.16 can be described as the set of points (x, y) such that $0 \leq x \leq 1$ and $0 \leq y \leq 1$. That is, it is the set given by $\{(x, y) : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$.

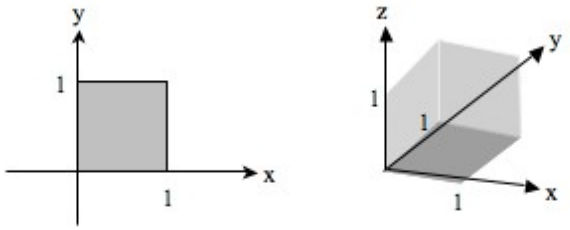


Figure 5.16: The shaded square is $\{(x, y) : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$. The cube is $\{(x, y, z) : 0 \leq x \leq 1,\ 0 \leq y \leq,\ 0 \leq z \leq 1\}$

The *cube* can be described as the set of points with coordinates (x, y, z) such that $0 \leq x \leq 1$, $0 \leq y \leq 1$ and $0 \leq z \leq 1$. That is, the cube is $\{(x, y, z) : 0 \leq x \leq 1,\ 0 \leq y \leq,\ 0 \leq z \leq 1\}$.

[9] They are idealisations because a line or a plane for us has no thickness. Another idealisation is that we are using classical ideas of space — not the more sophisticated ideas involved in the theory of relativity.

[10] To be more precise, we need to fix a base point, various axes, units of lengths, etc before we can make the correspondence between points on a plane and pairs of real numbers. Similar comments apply to points in space and triples of real numbers.

It now makes sense to define *four dimensional space* as $\mathbb{R}^4$, where

$$\mathbb{R}^4 = \{(x, y, z, u) : x, y, z, u \in \mathbb{R}\},$$

i.e. the set of 4-tuples of real numbers.

The *4D hypercube* in four dimensional space is then defined to be the following set:

$$\{(x, y, z, u) : 0 \leq x, y, z, u \leq 1.\} \tag{5.9}$$

We want to gain a "geometric" understanding of what this means!

Flatland

Let us now do the following thought experiment.[11] Imagine that there are certain beings which live in a two dimensional world. What would be the consequences of this for them? How could we explain our three dimensional world to them?

Living in a 2-Dimensional World There is a classical book *Flatland* written by Edwin Abbott over a century ago, available online at
`http://www.geom.uiuc.edu/ banchoff/Flatland/` .
There is some background history and commentary at
`http://www.geom.uiuc.edu/ banchoff/ISR/ISR.html` .
In Flatland the consequences of living in a 2D world are explored in depth. The book is both a satire on Victorian English social mores and an introduction to the understanding of higher dimensions.

In a 2D world, everyone is either a right facing individual when standing up, like Blah and Blip in Figure 5.17, or a leftie like Blog. Blip needs to stand on his head if he wants to look at Blog's back.

We, from our 3D world, can see inside the head of Blah, Blip and Blog. But they cannot see inside each other's head.

Describing a Cube to Someone in a 2D Universe When we draw a cube, we are really projecting it down into two dimensional space. See the top right diagram in Figure 5.18. This diagram is how we might start to explain a 3D cube to someone (e.g. Blog) living in a two dimensional world.

It would be better to explain the 3D cube as an infinite stack of 2D squares, as in the top left diagram in Figure 5.18, although I have only drawn 5 squares. But notice that the squares do not really look like squares, so you would need to explain that the right angles are changed because of the projection. You would also need to explain that there really are five "squares" in the picture although four of them are only partially to be seen.

You could use the bottom left diagram in Figure 5.18 and explain that the original layered squares do *not* overlap; they only appear to do so because of the fact they have been projected down to two dimensions. (You should also

[11] In a thought experiment we work through in our mind the consequences of a hypothetical scenario.

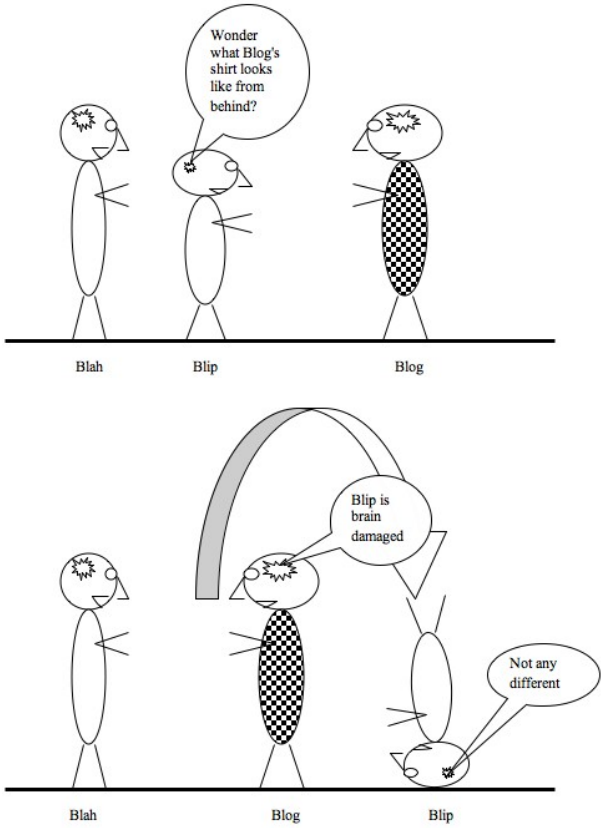


Figure 5.17: Blah and Blip are right facing individuals, Blog is a leftie. We can see inside their heads but they cannot see inside each others head.

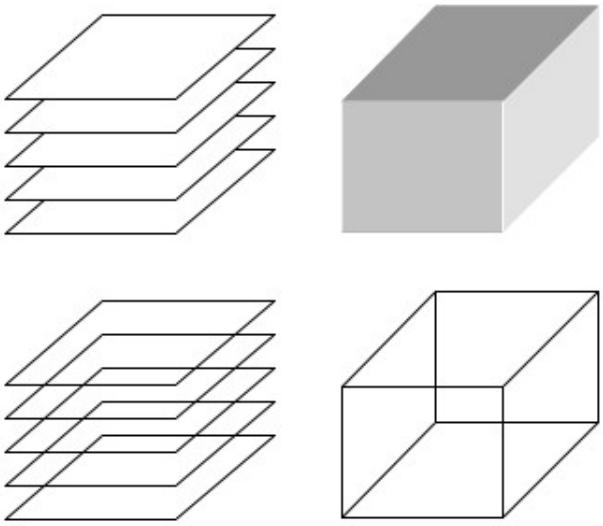


Figure 5.18: Projecting a cube, or square slices of a cube, onto a plane.

draw the lines in a translucent way so the 2D creatures can see all the lines you have drawn.)

An alternative would be to draw the frame in the bottom right diagram. (The lines should again be translucent.) You would have to explain:

1. The 12 lines only meet at common vertices and otherwise they do not intersect up in 3D space.
2. The 6 squares only meet at common edges and otherwise do not intersect up in 3D space.
3. The 4 edges of a square bound the square, as Blog ought to know. In an analogous manner it is true that the 6 squares in the bottom right diagram of Figure 5.18 bound a 3D cube up in 3D space. This is very difficult for Blog to understand, or indeed anyone else living in 2D space!

You could also draw a number of copies of the plane with x and y coordinate axes, make them translucent, and stack them up, as in Figure 5.19. Then

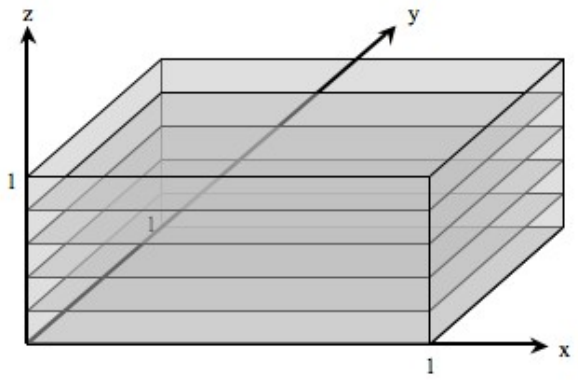


Figure 5.19: Projecting a cube as many copies of a square.

describe a point P in a 3D cube to Blog as follows:

1. Ask Blog to imagine an infinite collection of squares: although they overlap in his/her world the corresponding squares in the 3D world do not overlap.
2. Tell Blog to think of each square as having a number z associated with it, that number z lies in the range $[0, 1]$, it corresponds to the z coordinate of the square, and it represents how far that square lies above the $x - y$ plane in 3D space.
3. Point out that each point P in the cube will be on a certain level, and the particular level will be given by a number z such that $0 \leq z \leq 1$.
4. Once we have the level z for P, the x and y coordinates of P are found in the usual manner by using the x and y axes for that particular level. You will find that $0 \leq x \leq 1$ and $0 \leq y \leq 1$.

You could also ask Blog to imagine a single square whose colour changes continuously from red to orange to green to yellow to blue to indigo to violet, which corresponds to the z coordinate of the corresponding square slice in the cube changing from 0 to 1.

Finally, you could ask Blog to think of just 2 squares, as in the "top" and "bottom" squares in the bottom right diagram from Figure 5.18. Then ask Blog to imagine joining the 4 corresponding pairs of vertices by (vertical) lines, one vertex of each pair from the top square and one from the bottom square.

Then point out that corresponding edges from the top square and the bottom square have been "joined" and in this way 4 new squares have been formed. There are now 6 squares, and tell Blog that because of all your previous explanations, it should be possible to imagine how in a 3D world the corresponding squares actually bound a cube!

Describing a 4D Universe

4D Universe as $\mathbb{R}^4$ We saw that standard space can be described by $\mathbb{R}^3$, the set of triples of real numbers.

We defined four dimensional space to be $\mathbb{R}^4$, the set of 4-tuples (x, y, x, u) fo real numbers. We then defined the 4D cube, the hypercube, by

$$\text{4D cube} = \{(x, y, z, u) : 0 \leq x \leq 1,\ 0 \leq y \leq 1,\ 0 \leq z \leq 1,\ 0 \leq u \leq 1\}. \quad (5.10)$$

This is probably the best way to do things mathematically, but it is not very helpful for our intuition. So we will try and gain some geometric insight by thinking of a 4D universe from our 3D perspective in a manner analogous to they way we saw how to explain 3D matters to Blog living in a 2D world.

Understanding the Hypercube Geometrically In a manner parallel to our previous discussions, imagine two unit cubes in our 3D space. They may overlap, but we think of them as being the projections of two non overlapping cubes in 4D space, corresponding to $u = 0$ and $u = 1$ respectively in (5.10). See Figure 5.20. Now join corresponding vertices of the top and bottom cubes.

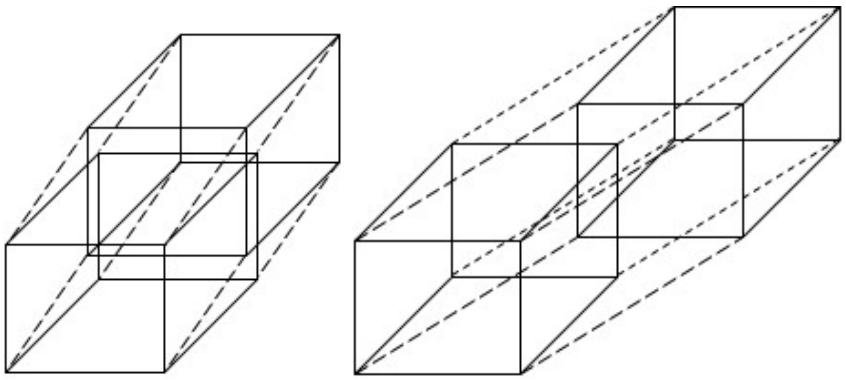


Figure 5.20: A 4D hypercube projected onto three dimensional space.

In Figures (5.18) and (5.19) we joined the vertices of the top square to the corresponding vertices of the bottom square. The result was that the 4 edges of the top square were connected to the corresponding 4 edges of the bottom square to give 4 new squares and hence a total of 6 squares. Up in 3D space

these 6 squares only meet along any common edges and were the boundary of a 3D cube.

In Figure 5.20 the result of joining vertices is that in each case the 6 faces of the bottom cube are connected to the corresponding face in the upper cube to give 6 new cubes and hence a total of 8 cubes. Up in 4D space these 8 cubes only meet along any common faces *and they form the boundary of the 4D hypercube*. This latter is difficult to understand!

In Figure 5.20 you can also think of the bottom cube at time $u = 0$ being translated slowly into the top cube at time $u = 1$. Any two cubes in 4D space corresponding to two cubes in 3D space at different times do *not* overlap. As u passes from 0 to 1, the entire 4D hypercube is traced out.

Alternatively, you might think of a single cube coloured red (think of "colour" $u = 0$) which slowly changes its colour through the colours of the rainbow to colour violet (think of "colour" $u = 1$). Every point in the hypercube is described by coordinates x, y and z in the range $[0, 1]$ and a "colour" coordinate u also in the range $[0, 1]$.

In Figure 5.21 there is a stereo version of a 4D cube projected into 3D space.

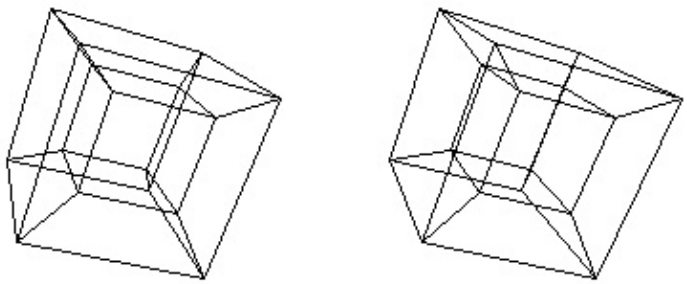


Figure 5.21: A 4D hypercube projected onto three dimensional space, in stereo. To view, stare at the centre of the two images and cross your eyes until the two images merge. Allow your eyes to relax so they can refocus.

An excellent site on the 4D cube is
`http://dogfeathers.com/java/hyprcube.html` .

Use the red-blue glasses from the back of [HM]. Put the red lens on the left (important). Click through the stereo button on the above site until the red blue version appears (red green is not bad either). Click on the detach button and enlarge to full screen size. The projection of the 4D cube will sit in front of the screen out over the keyboard. (I found slowing speed to about 5 and giving a projection of about .3 was useful.)

Be warned that some people have difficulty seeing the stereo effect.

By way of comparison, in
http://mathforum.org/alejandre/applet.polyhedra.html
you can see how a 3D cube projects into 2 dimensions. Each of the six 2D squares (i.e. faces) bounding the 3D cube projects onto a distorted square (i.e. quadrilateral) in 2D space. Although these six distorted squares intersect one another in the 2 dimensional projection, the original six squares only intersect along any common edges.

In the first site the 4D cube projects in an analogous manner into 3 dimensions. There are now eight 3D cubes bounding the 4D cube and each 3D cube projects onto a distorted cube in 3D space. Although these eight distorted cubes intersect one another in the 3 dimensional projection, the original eight cubes in 4D space only intersect long any common faces.

Does the Fourth Dimension "Really" Exist?

It is often important to represent information by four or more real numbers. For example, we might represent the weather by the temperature, barometric pressure, wind speed and precipitation rate. This would give the weather as a point in $\mathbb{R}^4$, altough using only 4 numbers is a gross simplification. An economy or a physical system might be represented by n parameters, where n is very large, and so be represented by a point in $\mathbb{R}^n$.

In the theory of relativity it is natural to take time as the fourth dimension. This is different from what we are doing here because "time" is not a "spatial" direction, the notion of distance needs to be modified, and it is not possible to go "backwards" in time. None the less, there are also important analogues between the two notions of four dimensions.

In contemporary theories in physics, the universe is modelled by a 10 or 11 dimensional curved space. The additional dimensions are sort of analogous to incredibly small 7 or 8 dimensional spheres, of the order of 10^{-35} metres in size, and much to small to observe. See
http://en.wikipedia.org/wiki/Why_10_dimensions%3F

Four dimensional geometries are particularly complicated, particularly when we allow "curved geometries"! But in the last year or two an amazing breakthrough was achieved. See
http://www.insidescience.org/reports/2006/021.html
for the story.

We will look briefly at two dimensional curved geometries in a later section.

5.4 Topology, Isotopy and Homeomorphisms

[HM, pp328–338]

Changing, modifying or tweaking some aspects of reality can reveal hidden structure in the world.

Overview

This section is a commentary and in some cases an extension of the material in [HM]. You will definitely need to read the relevant material from there.

Topology can, very loosely, be thought of as "rubber sheet" geometry. That is, topology studies those aspects of shape and structure which are preserved when we stretch, twist or bend an object, but are not necessarily preserved when we glue or cut an object. We will see many examples in the following. But in topology one does not necessarily restrict to objects sitting in three dimensional space, as mostly we do here.

Topology is a major unifying concept in almost all areas of contemporary mathematics. *General topology* is the study of properties like continuity and connectedness. *Algebraic topology* uses ideas from algebra, and particularly from group theory, to classify geometric objects. *Differential topology* studies geometric objects where there is a notion of smoothness via what is called a differentiable structure. *Low dimensional topology* is the study of topology of three and four dimensional objects. But as with all of mathematics there are no clear boundaries between these subjects, and each draws on the others.

Topology has applications to network theory, image modelling, relativity theory, mathematical economics, optimisation theory, study of vision, DNA and protein structures, ...

The Main Definitions

Isotopy The first six curves in Figure 5.22 have the property that each can be continuously distorted into the other without any cutting (breaking, tearing) or gluing. We say that they are *isotopic*[12] to each other.[13]

[12] *isotopic* is derived from the greek words *iso* meaning 'equal' and *topos* meaning 'place'.

[13] Here are the main definitions in a slightly more, but still not completely, precise manner.

If A and B are subsets of R^3 then a function $f : A \to B$ is *continuous* if f sends nearby points in A to nearby points in B. More precisely, $f : A \to B$ is continuous if whenever $a_n \to a$ where the a_n and a are points in A, then $f(a_n) \to f(a)$.

Two subsets A and B of $\mathbb{R}^3$ are *homeomorphic* if there exists a one-to-one and onto function $f : A \to B$ such that f is continuous and its inverse $f^{-1} : B \to A$ is also continuous. We call f a *homeomorphism*.

We need to require that the inverse is continuous since there are one-to-one, onto and continuous functions $f : A \to B$ whose inverse f^{-1} is not continuous. For example, if A is the half open interval and B is the circle in the following diagram

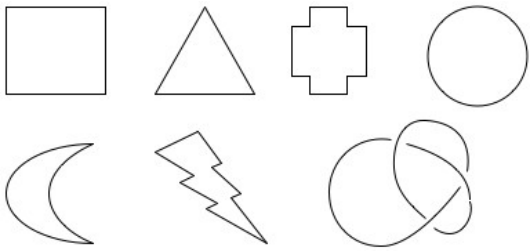


Figure 5.22: The first six curves are isotopic. All seven curves are homeomorphic. The last curve is a "knotted loop" in $\mathbb{R}^3$.

More generally, two subsets A and B of $\mathbb{R}^3$ are *isotopic* if they can be distorted one into the other within $\mathbb{R}^3$ without any cutting (breaking, tearing) or gluing. We call the distortion an *isotopy*.

In Figure 5.23 we cannot distort A into B if we are only allowed to do our distortions in $\mathbb{R}^2$. This is clear informally, since we would have to pass one

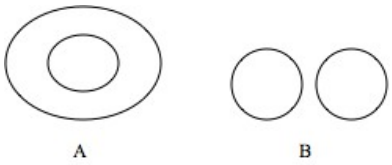
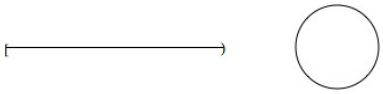


Figure 5.23: The sets A and B each consist of two closed curves. They are not isotopic in $\mathbb{R}^2$ but they are isotopic in $\mathbb{R}^3$.

of the loops in A through the other. (However, it is not easy to prove this rigorously.)

On the other hand, if we work in $\mathbb{R}^3$ it is easy to distort A into B. Simply lift the inner circle a little out of the plane, slide it across, and move it back down into the plane. Then deform it a little if necessary so it ends up as the second circle in B. Next move and distort the outer ellipse in A until it ends up as the first circle in B.



then there is a continuous $f : A \to B$, but its inverse f^{-1} is *not* continuous. *Explain.*

We say that two subsets A and B of $\mathbb{R}^3$ are *isotopic* if there is a continuous one parameter family of functions $f_t : A \to \mathbb{R}^3$ such that f_t is a homeomorphism for every $t \in [0,1]$, $f_0 : A \to A$ is the identity function given by $f_0(a) = a$, and the range of f_1 is B. More precisely, subsets A and B of $\mathbb{R}^3$ are *isotopic* if there is a continuous function $F(t,a)$ defined for all $t \in [0,1]$ and all $a \in A$, such that for each fixed t the function $F(t,a)$ is a homeomorphism onto some subset of $\mathbb{R}^3$, for $t = 0$ the function $F(0,a)$ is the identity function and for $t = 1$ the range of $F(t,a)$ is B. We call F an *isotopy*. We say that A and B are *isotopic in* $\mathbb{R}^3$.

Similarly, we can define what it means to be isotopic in $\mathbb{R}^2$, or in $\mathbb{R}^k$ for any positive integer k.

If we think of the seven curves in Figure 5.22 as sitting in $\mathbb{R}^4$ rather than in $\mathbb{R}^3$, then in fact they *are* isotopic. This is because we can unknot the knot in $\mathbb{R}^4$ by the same sort of trick as is done in the diagram in [HM p314], and by an idea analogous to what we just did with the sets A and B in $\mathbb{R}^2$. See Question 2.

It is usually clear from the context that we are only allowing isotopies in $\mathbb{R}^2$ or in $\mathbb{R}^3$ (or perhaps in some higher dimensional space). But if there is any ambiguity, we will clarify the situation.

Homeomorphism The first six curves in Figure 5.22 are not isotopic to the knot, since the knot cannot be unravelled without cutting through itself.

There is a however a one-to-one way of matching up points on the circle (say) and points on the knot.

To see this fix a point on the circle and fix another point on the knot. Move continuously around the circle and at the same time move continuously around the knot, until after one unit of time we return to the initial point in each case. Match up the two points which corresponding at each time and define the function f from the circle to the knot in this manner. The function f is continuous in that nearby points are mapped to nearby points, and the same is true for the inverse function f^{-1}. For this reason we say that the circle and the knot are *homeomorphic* and say that f is a *homeomorphism*.

In a similar manner, we can define a homeomorphism from any of the first six curves to the knot.

Summary

- Two sets are *isotopic* if they can be distorted into each other without any cutting or tearing.
- Two sets are *homeomorphic* if there is a one-to-one continuous correspondence from one to the other which is continuous and its inverse is continuous.
- If two sets are isotopic then they are homeomorphic. If two sets are homeomorphic they are not necessarily isotopic.

Topology This is the study of those properties of sets which are preserved under homeomorphiisms.

In this chapter we will mainly be looking at isotopies. Because two isotopic sets are also homeomorphic, isotopic sets will have the same topological properties.

Three Surprising Isotopies

It is important to realise that we are discussing distortations that can actually be done with real rubber objects provided these objects are sufficiently flexible.

[HM, 329–331]

Removing Your Vest The problem is to remove your vest (or jumper) from under your buttoned jacket without tearing it. You have to imagine that your vest is made of very stretchy material!

329–332]

Turning a Punctured Tyre Inside Out It is important to realise here, and elsewhere that it is allowable to do things like

- make the puncture as large as you like,
- slide the puncture around the tube,
- compress parts of the tyre down to small strips

[HM, 330–332]

The Ring Challenge The problem is to distort the blue stretchable rubber with two holes, and a ring through both holes, in such a way that the ring passes through just one holes. See the figure half way down page 330. That this can be done is at first truly surprising.

Showing Some Sets are Not Isotopic

[HM, 332–333]

An Example of Non Isotopic Sets A rubber band, and a rubber band that has been cut, are *not* isotopic. The ends of the second would need to be glued together to get the first, and the first would need to be cut to get the second.

Show that these two examples are not isotopic by using some of the following ideas.

[HM, 333]

Removing Points See the diagram at the bottom of [HM, P333]. The circle has the property that when we remove *any* point we still have one piece left. This property is preserved if we distort the circle. From this fact we can show that a circle cannot be distorted into a circle with a line segment attached to the circle at one end.

If we remove two points from a circle then the circle breaks into two pieces. This property is preserved if we distort the circle. From this fact we can show that a circle cannot be distorted into a θ shaped curve.

[HM, 334]

Local Properties The immediate neighbourhood of any point on a circle looks like a small line segment. This property is preserved if we distort the circle, although the line segment may become very wiggly or bent.

However, there are two points on a theta curve which have different neighbourhoods. At these two points there are three directions one can move, not two.

So this is another reason that the circle is not isotopic to a theta curve.

The immediate neighbourhood of any point on a sphere looks like the immediate neighbourhood of a point on the plane. The same is also true for any point on the torus. So we cannot distinguish a sphere and a torus in this manner.

Any object with the property that around every point there is a small neighbourhood equivalent to a small neighbourhood of a point in the plane is called a *surface*. We will look more closely at surfaces in a later section.

[HM, 334–336]

Removing Circles If we remove any circle from a sphere we get two pieces. This property is preserved under distortion.

If we remove certain circles from a torus then it is *not* cut into two pieces. So a torus is *not* isotopic to to a sphere.

More Surprising Isotopies

[HM, 336,337]

Two Holed Tori Half way down page 336 of [HM] there are 5 figures which are isotopic to a two holed torus. Remember that a two holed torus is the surface of a two holed donut!

The fourth (box like) figure might be a bit surprising. It is meant to be a shoe box with two holes cut out the top and two out the bottom, and toilet rolls then glued in.

(To see the isotopy it is often psychologically convenient to think of the objects in question as being big, not small!)

Much more surprising is that a standard two holed torus can be distorted into the linked two holed torus shown in the margin of p336. This is shown in the green diagram on p337.

An important key to understanding this diagram is the idea that the "holes" where the four green tubes fit into the green balloon blob in the fifth green diagram, can slide around the green balloon blob.

[HM, 336,337]

Jello Blobs Through the first jello cube at the top of p337, there are drilled out two vertical tubes. This leaves us with a three dimensional lump of jello.

In the second jello cube the first tube is drilled out in a knotted manner and the second tube is drilled out as before.

In the third jello cube the first tube is drilled out vertically and the second is drilled out in a way which intertwines it with the first tube.

The first two cubes are not isotopic. This is perhaps not surprising, although we will not prove it here.

The first and the third cubes are isotopic, and this is quite surprising at first.

The key point is that we can slide the holes around the outside of the tube. We can then slide the hole of one tube down the side of the second tube, beginning from one end of the second tube and ending at its other end.

Questions

1 Do a selection of Questions 4, 6, 7, 9–40 on pp 339–344.

2 *The point to this Question is to see more carefully how to unknot a knot in four dimensions, as is shown informally on pp314,315 of [HM].*

In Figure 5.24 we begin with the string APB. The (x, y, z) coordinates are $A = (-1, 0, 0)$, $P = (0, 0, 1)$ and $B = (1, 0, 0)$.

The goal is to distort the string APB into the position AQB where $Q = (0, 0, -1)$, by keeping A and B fixed, by not cutting the string and

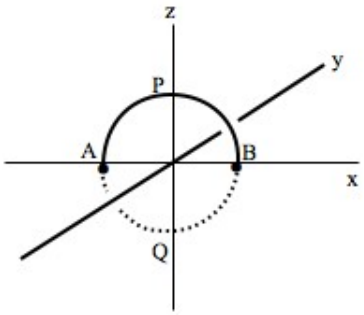


Figure 5.24: Distort APB to AQB without passing through the y axis.

by not passing through the y axis (the y axis stretches arbitrarily far in both directions, so we cannot just stretch the string over the end of the y axis).

It is not surprising that this is not possible in $\mathbb{R}^3$, although we will not prove this fact. But we will show it *is* possible to do this in four dimensional space.

The idea is to first lift the string APB up into 4D space, except for A and B which remain fixed. Second, without changing the fourth coordinate of any point on the string, press the string down in the z direction until it agrees with the curve AQB except in the fourth coordinate. Finally, drop the curve back down into 3D space so it lines up with AQB. See the discussion on pp 313–314 of [HM] of a similar situation.

We now make this precise.

Suppose that in $\mathbb{R}^3$ the string APB is given by $(x, y, z) = (s, 0, \sqrt{1 - s^2})$ where $-1 \leq s \leq 1$. Notice that A corresponds to $s = -1$, P corresponds to $s = 0$ and B corresponds to $s = 1$. Similarly, the curve AQB is given by $(x, y, z) = (s, 0, -\sqrt{1 - s^2})$ where $-1 \leq s \leq 1$.

One way to distort the string APB into AQB within $\mathbb{R}^3$ is to send each point $(s, 0, \sqrt{1 - s^2})$ vertically "down" to the point $(s, 0, -\sqrt{1 - s^2})$. If we imagine doing this uniformly over the time interval $0 \leq t \leq 1$ then at time t the point $(s, 0, \sqrt{1 - s^2})$ will move to $(s, 0, (1 - 2t)\sqrt{1 - s^2})$.

The problem is that at $t = 1/2$ the curve has moved to $(s, 0, 0)$ where $-1 \leq s \leq 1$, which for $s = 0$ corresponds to the origin. But the origin lies on the y axis and so the string cuts through the y-axis.

Suppose next we are in $\mathbb{R}^4$ and the coordinates of points are given by (x, y, z, w). We assign $w = 0$ to all the points in $\mathbb{R}^3$. That is, let $A = (-1, 0, 0, 0)$, $P = (0, 0, 1, 0)$, $B = (1, 0, 0, 0)$ and the string APB is given by $(s, 0, \sqrt{1 - s^2}, 0)$ where $-1 \leq s \leq 1$.

The y-axis is given by the set of points of the form $(0, y, 0, 0)$ where y is any real number.

We now do the "unknotting".

1. Write down a formula which, over the time interval $0 \leq t \leq 1$, sends each point $(s, 0, \sqrt{1 - s^2}, 0)$ on the string at time $t = 0$, in a continuous manner to the point $(s, 0, \sqrt{1 - s^2}, \sqrt{1 - s^2})$ at time $t = 1$. All points on the string, except A and B, should "lift" into

$\mathbb{R}^4$. The maximum lift will occur for the point P. The x, y and z coordinates of each point on the string should be unchanged in this step. *HINT*: We just did something similar in $\mathbb{R}^3$.

2. Write down a formula which over the time interval $1 \leq t \leq 2$ sends each point $(s, 0, \sqrt{1-s^2}, \sqrt{1-s^2})$ on the string (now in $\mathbb{R}^4$) in a continuous manner down to the point $(s, 0, -\sqrt{1-s^2}, \sqrt{1-s^2})$. The x, y and w coordinates of each point on the string should be unchanged in this step.
3. Write down a formula which over the time interval $2 \leq t \leq 3$ sends each point $(s, 0, -\sqrt{1-s^2}, \sqrt{1-s^2})$ in its new position in $\mathbb{R}^4$ in a continuous manner to the point $(s, 0, -\sqrt{1-s^2}, 0)$ on AQB. The x, y and z coordinates of each point on the string should be unchanged in this step.

Show that at no time $0 \leq t \leq 3$ did any point on the string lie on the y axis. That is, show that at no time $0 \leq t \leq 3$ did any point on the string have coordinates of the form $(0, y, 0, 0)$ for any number y.

Putting all this together, over the time interval $0 \leq t \leq 3$ we have continuously moved the string APB into the position AQB, keeping A and B fixed throughout and not at any point passing through the y axis.

Remark Here is a way of thinking about the three steps without doing any calculations. Imagine that the string APB and the y axis are at temperature $0°$, which you may think of as colour blue.

1. Heat the string so that the temperature at A and B remains $0°$ but the temperature along the rest of the string slowly rises until at time 1 the temperature is $100°$ or colour red at P, and otherwise changes continuously from $0°$ to $100°$ along the string. Think of the temperature, or the colours across the rainbow from blue to red, as giving the w coordinate.
2. Keep the temperature fixed but move the string down to position AQB.
3. Lower the temperature back to $0°$ along the string.

The string does not really pass through the y axis since the 4th coordinate, i.e. the colour or the temperature of the string, is different at the point on the string where the "crossing" would otherwise takes place.

5.5 One Sided Surfaces and Non Orientable Surfaces

[HM, pp346–353]

Looking at concepts or objects in new ways can often lead to surprising discoveries.

Overview

In this section we discuss *surfaces*, and in particular what it means for a surface to be *one sided*, and what it means for a surface to be *non orientable*. In [HM] the emphasis is on the first of these ideas and there is little discussion of the second idea.

The important distinction is that the notion of the "side" of a surface depends on the fact the surface under consideration it sitting in 3D space. For this reason we say that the notion of side of a surface is an *extrinsic* notion.

On the other hand, the notion of "orientability" of a surface is *not* defined in terms of how the surface sits in 3D space, and is defined by reference just to the surface itself.[14] For this reason the notion of orientability of a surface is called an *intrinsic* notion.

The difference between the two ideas is confusing because for surfaces in 3D space a surface is one sided *if and only if* it is non orientable! See Question 3.

Another major idea we discuss is the idea of an "identification diagram" used to describe and analyse a surface. This will play a very important role when we discuss the classification of surfaces in the following chapter.

What is a Surface?

We can loosely define a *surface* S to be a geometric object such that every point in S has a neighbourhood which is "equivalent" to a disc in the plane. More precisely, by "equivalent" we mean "homeomorphic".

A sphere is a surface, and so is a torus.

In many cases we talk about the *surface of* a solid object which is in $\mathbb{R}^3$. For example, the surface of a ball is a sphere and the surface of a donut is a torus!

We will also want to look at surfaces with one or more *edges*, as in Figure 5.25. Another example is the Mobius band in Figure 5.26. We won't give a precise definition, however.

[14] In fact, in the next section we will discuss surfaces without having them necessarily sitting in 3D space or in any other space.

What is a Side, Locally?

Definition Imagine a point P on a surface S in 3D space, see Figure 5.25. We have a good intuitive idea of what we mean by the two sides of the surface

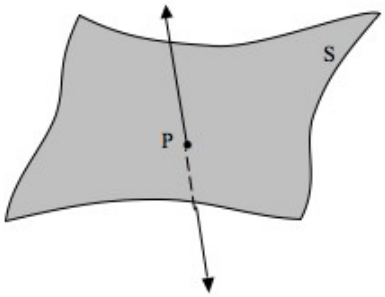


Figure 5.25: A surface S in 3D space. A point P on S, and normal vectors at P pointing in the two possible directions.

near P. A bug at P must be on one side or the other. The two (local) sides of a surface near P correspond to the two essentially different directions in which a bug can look into space from P.[15]

More precisely, *the two (local) sides of a surface S in 3D space, near a point P on S, correspond to the two possible directions of a vector normal to S with base at P.*

Every surface in 3D space is *locally* two sided.

Examples Recall from Section 5.4 the sphere, the torus (surface of a doughnut), the torus with two holes, the torus with three holes, and so on. These surfaces have no edges, i.e. no boundary.[16]

Surfaces with boundary also locally have two sides near any point P which is on the surface but is not on the boundary. Simple examples are a spherical cap or the surface in Figure 5.25.

Later we will discuss the Mobius band and the Klein Bottle. These also are *locally* two sided.

Sides of a Curve Imagine a curve in 2D space. Near any point P on the curve other than at endpoints, there are locally two different sides.

But if we draw a curve in 3D space then the notion of a side does not make sense any more! *Draw diagrams of a curve in 2D space and a curve in 3D space.*

The moral is that the idea of "side of a curve" or "side of a surface" depends on the space in which the curve or surface is sitting.

[15] There is both a local and a global notion of side for a surface sitting in 3D space. It will usually be clear from the context which we mean, but we often use the words "local" and "global " for emphasis.

[16] We are using the word "edge" here in a different sense from the way we used it in the discussion of edges for a Platonic solid on page 230, or in the discussion around Euler's formula on page 233. We can here instead use the word "boundary".

Sides of 3D space Our 3D space does not have "sides". But if it were sitting in 4D space then it would locally have two sides at every point, although we could not see them from our 3D point of view. If you think of the fourth dimension as time then near any point in 3D space one "side" would correspond to looking into the future and the other would correspond to looking into the past!

What is a Side, Globally?

[HM, top of p347], [HM, 3

Definition Suppose a bug starts out on one of the two sides at a point P on the surface S in 3D space, goes for a long walk, never crosses the boundary of S if there is any, and finally returns to P.

If the bug always ends up at P on the same side from which it started then we say S is *(globally) two sided*. If it is possible for the bug to finish on the opposite side at P from which it started then we say that S is *(globally) one sided*.

More precisely, suppose S is a surface sitting in 3D space. Begin with a vector whose base point is at some point P on S and which is normal to S. The base of the vector is moved around the surface so that the vector changes its direction in a *continuous* way and is always normal to S. The path of the base is not allowed to cross the boundary of S if there is any.

> ** *If the direction of the vector after it returns to P is always the same as when it began, then we say that S is* (globally) two sided. *If for some path the direction of the vector after it returns to P is the opposite from when it began, then we say that S is* (globally) one sided. **

We usually write "one sided" for "globally one sided", and "two sided" for "globally two sided".

Examples The sphere and a torus with one or more holes are (globally) two sided surfaces, as is the example in Figure 5.25.

The Mobius Band and the Klein Bottle, which we soon discuss in detail, are (globally) one sided surfaces.

Sides are Extrinsic It is important to realise that the definition of a "side" of a surface (or a curve) depends not just on the surface (or curve) but also on the larger space in which the surface (or curve) is sitting. See Question 2.

The Mobius Band

[HM, 347–350]

Construction Twist and tape a strip of paper as in Figure 5.26. The two edges marked "a" are glued together and the arrows indicate how the edges are "matched up". We usually say that the two edges are *identified*.

Applications in "Real Life" Recycling Logo. One sided conveyor belt (it wears uniformly on both "sides'). See "Mobius Bands Abound" in [HM, 350].

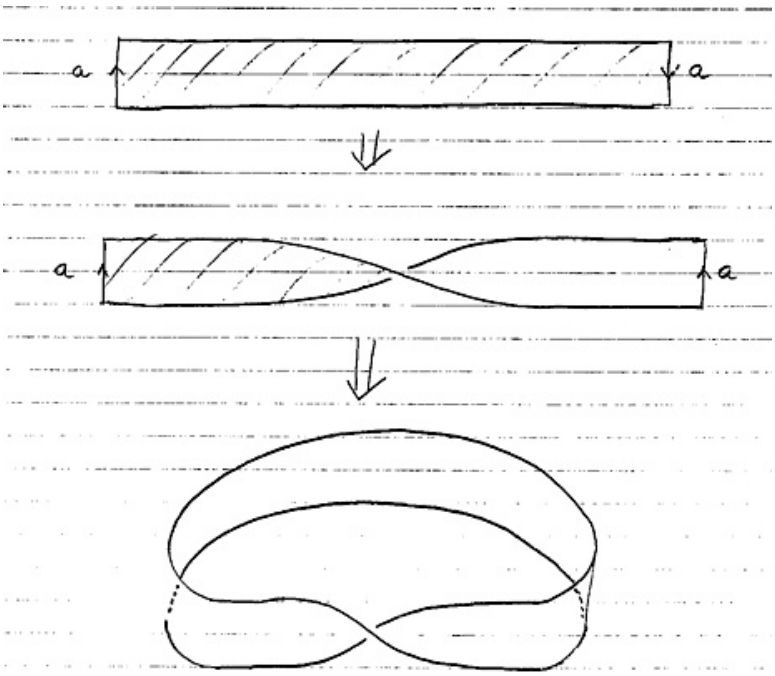


Figure 5.26: The Mobius Band.

The Identification Diagram This provides a very powerful way of understanding the properties of the Mobius band.

When constructing the Mobius band we glued, i.e. identified, the two ends of the strip of paper after doing a half twist. We represent this by the identification diagram in Figure 5.27, see also Figure 5.26. The direction of the arrows show

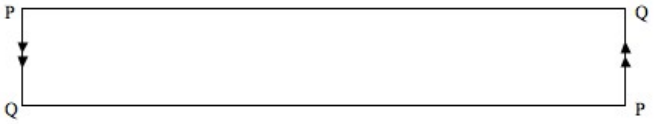


Figure 5.27: Mobius band identification diagram.

in what direction the edges of the original strip should be lined up and glued together, i.e. identified.

Notice that the two points marked P are identified and so are the two points marked Q.

✍ *What is the surface you get if the arrows in Figure 5.27 are in the same direction at each end?*

Number of Edges and Sides Begin by marking an edge of the Mobius band with a red pen, and continue until you return to the starting point. There is

just one edge!

Begin by marking a line along the centre of one side and continue until you return to the starting point. There is just one side (globally).

See Experiments 1 and 2 in [HM, 347,348].

Cutting Down the Centre Use a pair of scissors to cut down the middle. How many pieces? Just one!

How many edges and sides for this new band? Two edges, two sides. And one full twist.

We can best understand the centre cut from the identification diagram.

See Experiment 3 in [HM, 348,349].

Cutting One Third In From the Edge Use a pair of scissors to cut one third in from one side. Continue until you return to the start. How many pieces? Two!

We can also best understand this from the identification diagram.

See Experiment 4 in [HM, 348,350].

One Sided Suppose a bug starts at some point P on the Mobius band and travels once around the band back to P. The bug will then be on the opposite side at P from which it started. For this reason the Mobious band is *(globally) one sided*.

Explain why the Mobius band is one sided by discussing of a vector whose base moves around the Mobius band and which is always normal to the band. Use the definition of *one sided* marked with ** on page 257.

Non Orientability

This is not in *[HM]*

Mobius Band Imagine the Mobius band to be transparent and consider a small circle drawn on the Mobius band with an arrow to indicate orientation. You might like to think of a watch. See Figure 5.28. Move the circle *once* around the band until it comes back to the starting place. The orientation of the circle will be reversed.[17] We say the Mobius band is *non orientable*.

Notice in Figure 5.28 what happens as the oriented circle crosses the line "a". The leading arc points from Q to P and the trailing arc points from P back to Q. You should observe this first in the 3D diagram and then in the 2D identification diagram.

Terminology Why use the terminology "orientable"?

If we draw any circle in 2D space then we can orient it, i.e. put an arrow on it, in two different directions. See the first two circles in Figure 5.29. The first orientation is the mirror image of the second orientation, and conversely. If we are only allowed to isotope the first circle together with its arrow (i.e. to

[17] Remember that we are not on one "side" of the Mobius band. Think of being "in" the Mobius band, analogous to the way a point might be "in" 2D space.

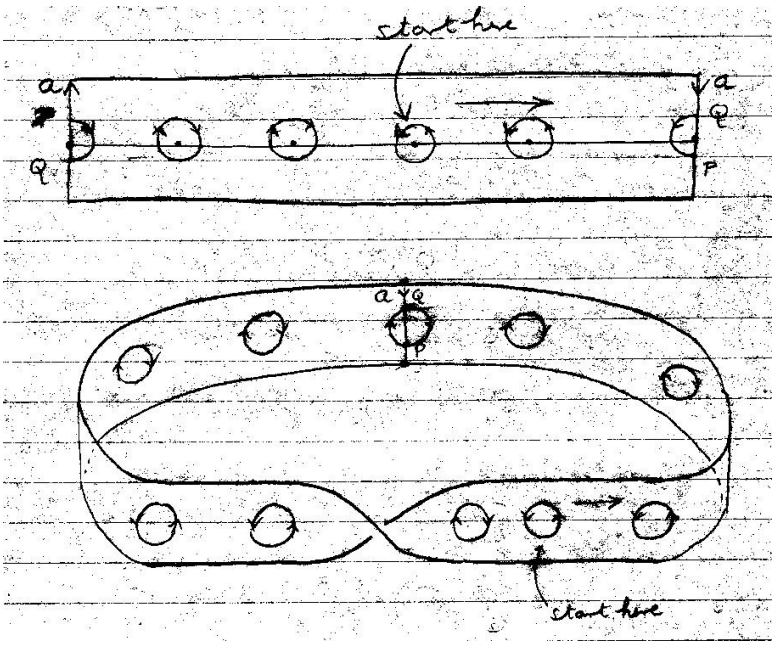


Figure 5.28: An oriented circle moved once around the Mobius Band returns with the opposite orientation.

Figure 5.29: On the left are two circles with opposite orientations. In the centre is the *Universal Standard* orientation. On the right are three crcles with "good" orientations according to the Universal Standard.

translate, rotate and distort, but *not* reflect or pass it through itself), then we will never be able to obtain the second circle together with its arrow.

In order to have consistency in these matters, we will select some circle with an orientating arrow, keep it safe under presidential type bodyguards, and call it the *Universal Standard*. By translating, rotating and distorting the Universal Standard we can assign a "good" rotation to any circle. The "bad" orientation of a circle will be the other of the two possible orientations which does not agree with the Universal Standard.

This method works in 2D space and on a torus, but it does not work on the Mobius band. The problem is that if we move the Universal Standard around the Mobius band to a circle in order to determine which orientation of the circle is the good one and which is the bad one, we will get a different answer depending on which transportation route we use!

For this reason we say 2D space and the torus are *orientable*, but the Mobius

band is *non orientable.*

One Sidedness and Non Orientability There is a subtle difference between being the notion of being one sided and the notion of being non orientable. The first notion requires placing the Mobius band in a larger space. In this sense the notion of being one sided is an *extrinsic* notion — to describe it we need to have a larger space containing the Mobius band. The second notion is an *intrinsic* notion. We can describe it without referring to any larger space containing the Mobius band.

However a surface in 3D space is one sided if and only if it is non orientable. See Question 3.

The Klein Bottle

[HM, 351,352]

Construction and Identification Diagram See the diagram at the top of p352 of [HM] and see Figure 5.30.

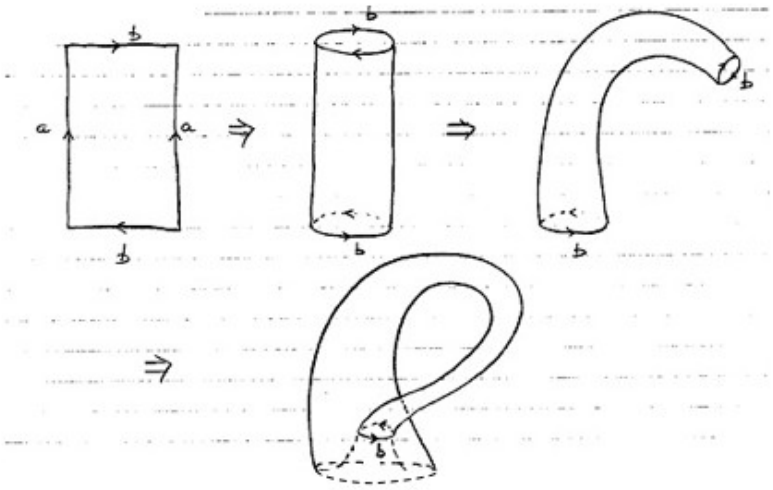


Figure 5.30: The Klein bttle.

The identification diagram is the first rectangle in Figure 5.30. It tells us which sides are identified and in which way. Notice that there are no unmatched edges in this case.

The Klein bottle can be constructed in 3D space only if we allow it to self-intersect. However, it can be put in 4D space without self intersection. *Explain, using the idea of the Remark on page 254.*

One Sided Suppose a bug were to start at some point P on the outside of the thin neck in Figure 5.30, walk down the neck and (mysteriously) pass through the surface of the bottle but stay on the same side of the neck, keep

going down to the bottom of the bottle, then up the bottle (it would now be inside the bottle) to the top, and then back down to its starting point P. The bug would have returned to P but now be on the inside of the thin neck, the opposite side of P from which it started.

For this reason we say that the Klein bottle in 3D space is one sided — the inside and the outside are the same. See the crystal Klein bottle at the bottom of p352 of [HM].

Non Orientable In Figure 5.31 we begin with a small oriented circle on the Klein bottle as shown and move it along the curve in the direction shown. ✍ *What does this curve look like when drawn on the Klein bottle at the bottom of Figure 5.30?*

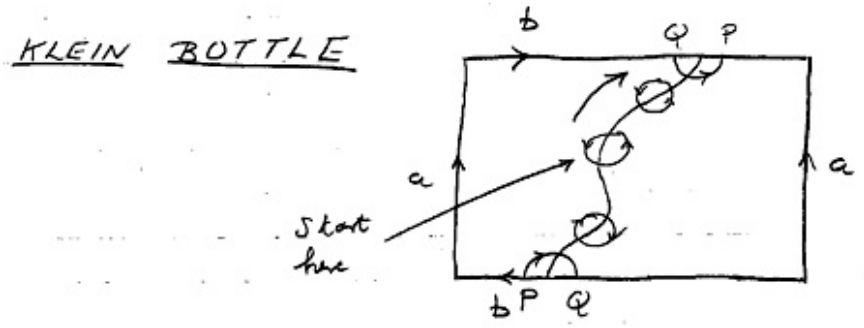


Figure 5.31: Moving an oriented circle around the Klein bottle.

We see that if the oriented circle crosses b just once then its orientation is reversed when it comes back to its starting position.

For this reason *the Klein bottle is non orientable.*

In fact, the circle's orientation is reversed when it comes back to its starting position after crossing b an odd number of times, but its orientation is unchanged after crossing b an even number of times.[18] Crossing a does not affect the orientation of the small circle when it comes back to its starting position.

Questions

1 Do a selection of Questions 8–40 on pp 354–358 of [HM].

2 Supose C is the curve we used on page 259 to cut down the centre of the Mobius band. Notice that it is just a circle.

Suppose P is a point on C. How many sides does C have locally near P (where C is sitting in — i.e. C is regarded as a subset of — the Mobius band)? Explain.

[18] *A possible misconception*: The orientation of the small circle does *not* suddenly "reverse" when the circle crossed b. Nothing dramatic happens when the small oriented circle crosses either b or a! After all, b is just a simple closed loop on the Klein bottle, as is a.

How many sides does C have globally? Explain.

Without changing C, explain how C can be taken as sitting in another surface D in such a way that C now has a different number of sides globally.

3 Explain why a surface S in 3D space is one sided if and only if it is non orientable.

HINT: The two sides near a point P on S correspond to the two possible directions of an arrow at P which is normal to S. A bg travelling around S corresponds to the base of a normal arrow moving around S.

Next notice that by the "right hand screw rule" in 3D space, each orientation of any small circle in S centred at P, gives a unique direction normal to S at P. Conversely, each direction normal to S at P, gives a unique orientation of any small circle in S centred at P.

5.6 Classifying Surfaces

This material is not in [HM]

Overview

In this chapter we discuss the classification of surfaces.

You will not be required to know all this material. It will probably be sufficient to understand the material up to the "Classification Theorem" on page 278, to understand the statement but not the proof of this theorem, and to be able to do some "cut and paste" type arguments. Ask your teachers.

The following summary will not make complete sense until you have looked more carefully at the rest of the material in this chapter.

The surfaces we classify are called *closed* surfaces.[19] It is possible to classify surfaces which are not closed, but essentially no new ideas are involved.

Closed surfaces are surfaces which are *connected* (have only one component or "piece"), *compact* (i.e. can be built up from a *finite* number of polygons — this excludes the plane and other surfaces of "infinite extent") and have *no boundary edge* in the usual informal sense of the words "boundary edge" (this rules out the Mobius band but includes the torus and the Klein bottle).

The *main theorem* is that every closed surface is equivalent to a sphere, a sphere with p handles sewn in for some integer $p \geq 1$ (i.e. a p-holed torus), or a sphere with q cross caps (Bishops hats) sewn in for some integer $q \geq 1$. See (5.11).

Spheres, and spheres with handles, arise as the boundaries of solid objects in 3D space. But spheres with cross caps and Klein bottles do not arise in this way.

We first review how every surface can be built up from an *identification diagram*, i.e. a set of polygon patches with rules for sewing pairs of edges together. We will see that for each compact connected surface only *one* polygon is needed, sometimes called a *fundamental polygon* for the surface. See the section beginning on page 268.

The edges of a fundamental polygon can be given a *symbolic representation*. For example, $aba^{-1}b^{-1}$ gives the torus in Figure 5.38 and $aba^{-1}b$ gives the Klein bottle in Figure 5.30. For a surface without edges, every edge in the fundamental polygon appears exactly twice. *Can you see why, for the examples in these two Figures, we might use these particular symbolic representations? What others might we use?*

We show that a surface is *orientable* if and only if in its fundamental polygon *every* edge appears once in the form x and once in the form x^{-1}. A surface is *non orientable* if and only if at least one pair of edges appears both times as x or both times as x^{-1}. Using this, we see from (5.11) that the sphere and the tori are orientable but a sphere with cross caps is non orientable.

[19] The word "closed" here has quite a different and only very loosely related meaning to that which we used previously for closed intervals or closed subsets of the plane, etc.

From the *main theorem*, by collapsing, cutting and pasting, we can change a fundamental polygon of any closed surface to be in exactly one of the following forms:

$$\begin{aligned}
\text{sphere:} &\quad aa^{-1},\\
\text{sphere with one handle (torus):} &\quad aba^{-1}b^{-1},\\
\text{sphere with 2 handles (2-torus):} &\quad aba^{-1}b^{-1}cdc^{-1}d^{-1},\\
\text{sphere with 3 handles (3-torus):} &\quad aba^{-1}b^{-1}cdc^{-1}d^{-1}efe^{-1}f^{-1},\\
&\quad \vdots\\
\text{sphere with cross cap (projective plane):} &\quad aa,\\
\text{sphere with 2 cross caps (Klein bottle):} &\quad aabb,\\
\text{sphere with 3 cross caps:} &\quad aabbcc,\\
&\quad \vdots
\end{aligned} \tag{5.11}$$

Every closed surface is equivalent (more precisely "homeomorphic") to exactly one of the surfaces in (5.11). For example, we mentioned before that as in Figure 5.30 the Klein bottle can be represented by a fundamental polygon of the form $aba^{-1}b$. But it can also be represented by a fundamental polygon of the form $aabb$, as in (5.11).

Finally, we will see that the Euler number $E - V + F$ is 2 for a sphere (we essentially already know this from Theorem 5.2.6), $2 - 2p$ for a sphere with p handles, and $2 - q$ for a sphere with q cross caps. It will follow that the surfaces we classify are completely determined by their Euler number and their orientability.

Surfaces via Identification Diagrams

Identification Diagram for Two Holed Torus Imagine any of the examples we have so far of a closed surface. Cover it by a a "quilt" of patches consisting of triangles, rectangles, pentagons, etc. with curved sides. See Figure 5.32.

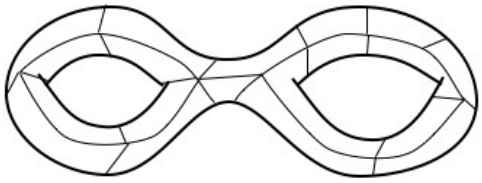


Figure 5.32: The two holed torus as a "quilt" of triangles, rectangles, pentagons, etc.

We can also describe the surface in Figure 5.32 by:
1. providing a copy of each polygonal patch;
2. listing all pairs of matching edges and indicating by arrows how each pair is matched

We call this way of describing a surface an *identification diagram.*[20]

By further subdividing each polygon if necessary, we can make sure all the polygons are triangles. Sometimes this is convenient. We say in this case that we have a *triangulation* of the surface.

Simple Example of an Identification Diagram In Figure 5.33 we have two patches with matching edges a, b, c and d and an unmatched edge e.

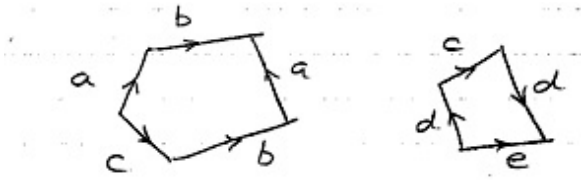


Figure 5.33: Two patches with instructions to form a "quilt".

The arrows show how the two edges marked a are matched, and similarly for b, c and d. Here e is not matched and represents the boundary of the corresponding surface. The patches are considered to be stretchable or compressible as much as is required.[21] The first step in "sewing" this quilt together might be to match up the edges c as in Figure 5.34.

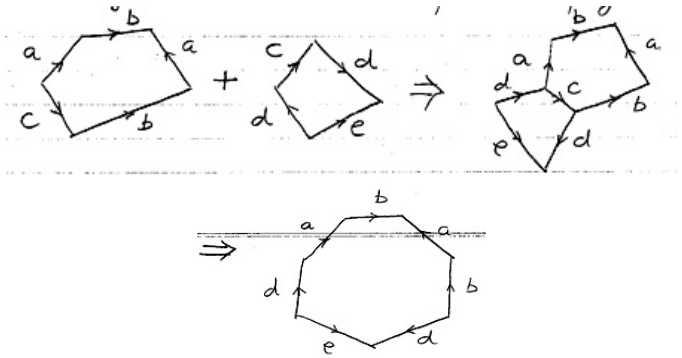


Figure 5.34: Sewing patches together.

We still have to sew together the edges marked a, b and d. The quilt need not actually be constructible in 3D space, unless we are allowed to pass one patch through another!

Since there is one unmatched edge e, the resulting surface is not closed in this example.

[20] To do this precisely requires that certain natural "compatibility conditions" be specified. In particular, vertices match vertices, not some other point elsewhere on an edge. Edges match in pairs, but many vertices may be matched together. *Why?*

[21] But a patch cannot be compressed to an edge nor an edge to a vertex.

Describing Surfaces to Citizens in 2D Space The identification diagrams in Figures 5.33 and 5.34 are, in principle, understandable to someone living in 2D space, although the actual patching can only be done in 3D or 4D space.[22]

In Figure 5.33 if a citizen from 2D space starts in the first patch and crosses the left edge a then he/she will reappear through the right edge a back into the first patch. If one starts in the first patch and crosses the edge c then one will reappear through the edge c into the *second* patch.

Paired edges such as a, b, c and d in Figure 5.35 are not observable to a 2D citizen "living" in the surface corresponding to the identification diagram. Similarly the edges in the identification diagram for a torus in Figure 5.38 are not observable to a 2D citizen living in the torus.

Identification Diagrams in General In future we will usually describe a surface by means of an *identification diagram*. This is a finite collection of triangles, rectangles, pentagons, etc., together with letters and arrows to provide instructions for matching certain edges in pairs. See Figures 5.30, 5.33, 5.34,

This approach via identification diagrams is very natural because we manage to avoid using a higher dimensional space in order to describe the surface.

Types of Surfaces

Connected Surfaces A surface S is *connected* if any two points in S can be connected by a line which lies in S.

Informally, a connected surface is a surface consisting of just one part or component. The surface consisting of both a torus *and* a sphere is not a connected surface.

The surface given by the identification diagram in Figure 5.35, consisting of three polygons, is not connected. The first two polygons are joined along c but the third polygon is not joined to either of the first two. The surface given by the identification diagram in Figure 5.34 *is* connected.

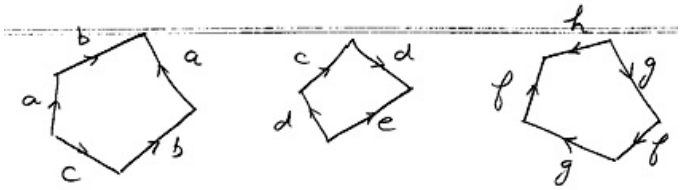


Figure 5.35: The first two patches are connected but the third patch is not connected to the first two.

[22] We could do the patching in 2D space if we were allowed to superimpose patches one on another. But this might not be very enlightening.

Compact Surfaces A surface is *compact* if it can be built out of a *finite* number of polygons by matching edges. This excludes the plane and other "unbounded" surfaces.

Closed Surfaces A *closed* surface is a connected compact surface without boundary edges. Examples are the sphere, torus with one or more holes and the Klein bottle, see Figures 5.30 and 5.32. Mobius bands and spherical caps have a boundary and so they are *not* closed surfaces.

If a closed surface is described by an identification diagram then each edge is matched with exactly one other edge. See Figures 5.30, 5.37, 5.38, 5.39 and 5.40.

Usually, but not always, we will deal with closed surfaces.

Fundamental Polygons

Connected Surfaces Suppose a connected compact surface S (e.g. a torus or a Mobius band) is given by an identification diagram. Then as in Figure 5.34 we can continue sewing patches together until just one polygon remains. We can then deform the polygon into a convex polygon. This is called a *fundamental polygon* for the surface.

Each fundamental polygon for a *closed* surface has an *even* number of edges occurring in matching pairs. Edges in each pair will be indicated by the same letter, with arrows to indicate in which of the two possible ways the two edges are matched.[23] See the identification diagrams in Figures 5.37, 5.38, 5.39 and 5.40.

Symbolic Representation When a surface is given by a fundamental polygon we can describe the way edges are matched up as follows.

Begin at any vertex and travel in a clockwise direction. Write each side as x or x^{-1} according as it points in the direction of travel (i.e. clockwise) or not (i.e. anticlockwise).

Thus in Figure 5.37 we can represent the sphere by aa^{-1} or by $a^{-1}a$, depending on the starting vertex. Either will do.

From Figure 5.38 one representation of the torus is $aba^{-1}b^{-1}$.

From Figure 5.39[24] one representation of the double torus is $aba^{-1}b^{-1}cdc^{-1}d^{-1}$.

From Figure 5.40 one representation of the 3-torus is $aba^{-1}b^{-1}cdc^{-1}d^{-1}efe^{-1}f^{-1}$.

There are many others representations in each case, obtained by starting at the other vertices. We can also reverse the arrows on *both* edges corresponding to any matching pair, and still obtain the same surface. *Why?*

Testing Orientability There is a simple way to see from a fundamental polygon for a connected surface if the surface is orientable or not.

[23] When two edges are matched we only need to know which end corresponds to which end. Other than this, different matchings will give essentially the same surface. More precisely, they will give homeomorphic surfaces. *Why?*

[24] The P's are to indicate that *all* vertices are identified. *Why is that true for this example?*

First look at the example of a Klein bottle in Figure 5.31. One of the possible symbolic representations is $aba^{-1}b$. The two edges marked b both point in the *same* direction, namely clockwise. If an oriented circle is transported around the Klein bottle and back to its starting position, in such a manner that it crosses the edge b exactly one, then its orientation will be reversed.[25] The Klein bottle is non orientable.

Next look at the example of the torus in Figure 5.38. One of the possible symbolic representations is $aba^{-1}b^{-1}$. The two edges marked a point in opposite directions, as do the two edges marked b. An oriented circle transported around the torus and back to its starting position will not have its orientation changed by crossing either a or b. The torus is orientable.

To summarise:

- If the two members of *every* pair of edges point in *opposite* directions then the surface is *orientable*;
- If the two members of *one or more* pairs of edges points in the *same* direction then the surface is *non orientable*.

Representations of the Sphere and Tori

Spheres with Handles Later on it will be useful to think of a torus as a sphere with two discs removed and a handle handle sewn in. See Figure 5.36.

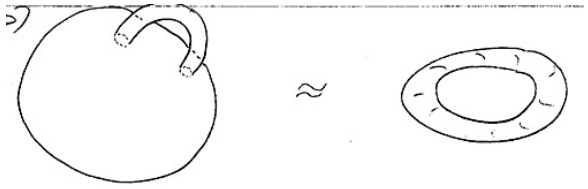


Figure 5.36: A sphere with a handle is a torus.

Similarly, a 2-holed torus is a sphere with two handles sewn in (after removing 4 disks), and more generally a p-holed torus is a sphere with p handles sewn in after removing $2p$ disks. *Why?*

Fundamental Polygons The sphere, torus, double torus and triple torus can be represented by fundamental polygons as shown in Figures 5.37, 5.38, 5.39 and 5.40 respectively.

I will not try to draw a p-holed torus, but by an analogous argument it can be similarly represented.

[25]The orientation is not changed suddenly as the edge b is crossed. If a two dimensional explorer walks around the Klein bottle and crosses edge b exactly once, then when she returns to her starting place she will be "reversed" — she will be the mirror image of her original

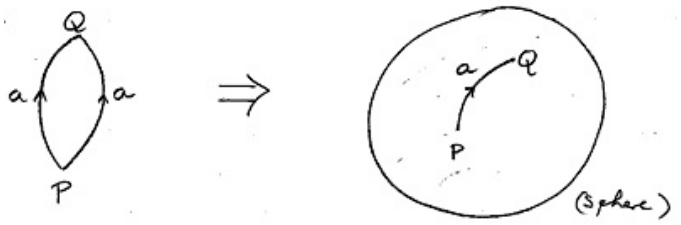


Figure 5.37: Identification Diagram aa^{-1} for the Sphere and how it is sewn together.

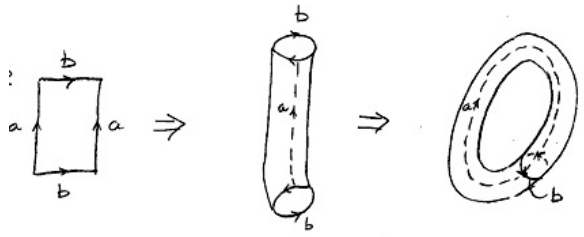


Figure 5.38: Identification Diagram $aba^{-1}b^{-1}$ for the Torus and how it is sewn together.

Vertices The four vertices of the fundamental polygon in Figure 5.38 each map to the same point on the torus. You can see this by following the construction.

But an easier way to proceed, particularly in more complicated cases, is as follows.

1. Mark the top left vertex of the fundamental polygon at the left of Figure 5.38 by P.
2. Since P is at the "end" of a it is identified with the vertex at the end of the other edge marked a. This is the top right corner of the fundamental polygon. Mark this vertex also as P.
3. Since the original P is at the "beginning" of b it is identified with the vertex at the beginning of the other edge marked b. This is the bottom left vertex. Mark this vertex also as P.
4. Finally, the bottom left vertex is identified with the bottom right vertex, as both are at the "beginning" of a. Mark the latter vertex also as P.

In this way we see all four vertices are identified on the actual torus.

A similar procedure for the 2-torus and the 3-torus in Figures 5.39 and 5.40 respectively shows that in each case all vertices of the fundamental polygon map to the same vertex on the corresponding surface. *Explain.*

self. Her left hand will now be on the right side, together with any ring originally on the left hand. Any book taken along for the journey will have its writing reversed. But the book will not look any different to the explorer, only to those who remained behind. On the other hand, all books in the local library will be reversed from the explorer's perspective.

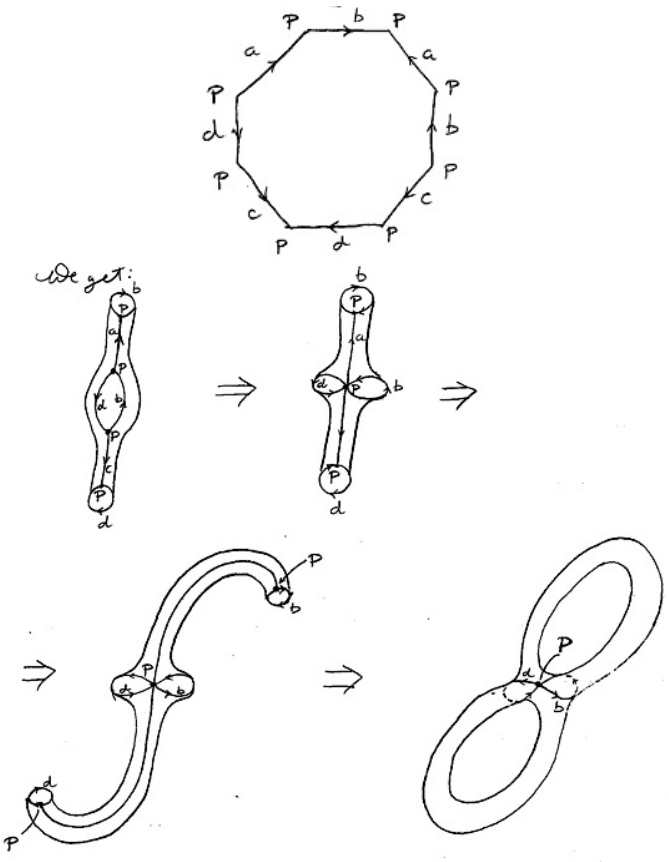


Figure 5.39: Identification Diagram $aba^{-1}b^{-1}cdc^{-1}d^{-1}$ for the Double Torus and how it is sewn together.

A similar procedure also shows that all vertices in the fundamental polygon for the p-torus will map to the same point on the p-torus. *Explain.*

On the other hand, the two vertices in the fundamental polygon in Figure 5.37 are mapped to distinct points on the sphere. *Explain.*

Summary A sphere, and a sphere with handles, are usually represented by the following fundamental polygons. For the sphere there are two distinct vertices. For a sphere with one or more handles, all vertices of each of the

The way for the explorer to undo this undesirable situation is to repeat the journey and double reverse back to the original orientation!

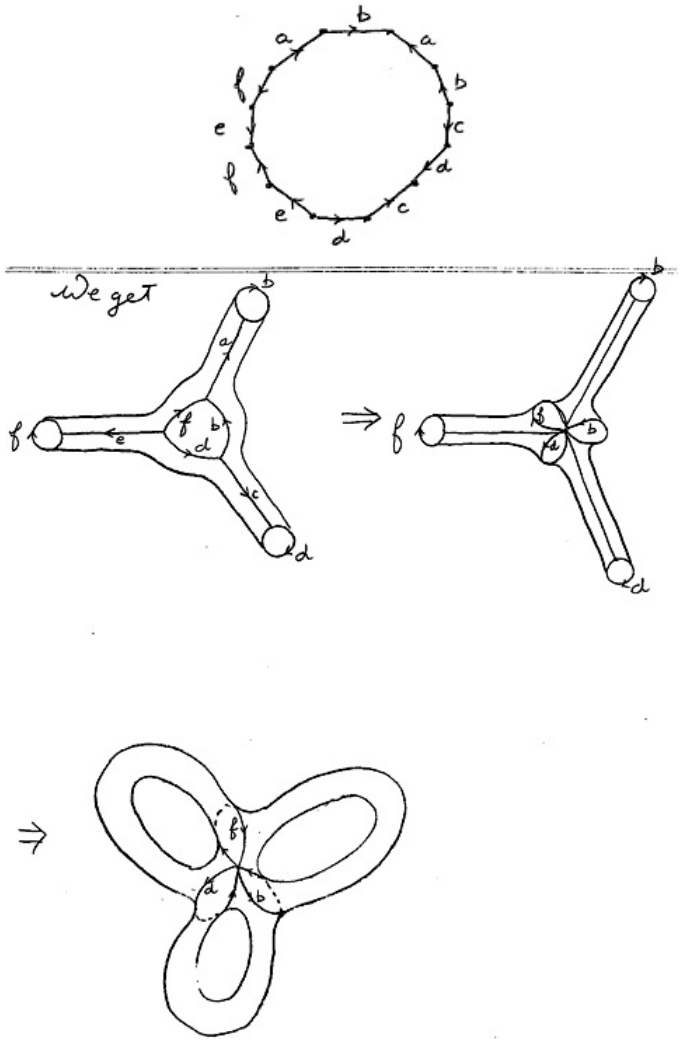


Figure 5.40: Identification Diagram $aba^{-1}b^{-1}cdc^{-1}d^{-1}efe^{-1}f^{-1}$ for the 3-Torus and how it is sewn together.

following fundamental polygons are identified.

$$\begin{aligned}
\text{sphere:} &\quad aa^{-1},\\
\text{sphere with handle = torus:} &\quad aba^{-1}b^{-1},\\
\text{sphere with 2 handles = 2-torus:} &\quad aba^{-1}b^{-1}cdc^{-1}d^{-1},\\
\text{sphere with 3 handles = 3-torus:} &\quad aba^{-1}b^{-1}cdc^{-1}d^{-1}efe^{-1}f^{-1},\\
&\vdots\\
\text{sphere with p handles = p-torus:} &\quad a_1b_1{a_1}^{-1}{b_1}^{-1}a_2b_2{a_2}^{-1}{b_2}^{-1}\dots a_pb_p{a_p}^{-1}{b_p}^{-1},\\
&\vdots\ .
\end{aligned} \tag{5.12}$$

Representations of some Non Orientable Surfaces

Klein Bottle In Figure 5.30 we gave the fundamental polygon $aba^{-1}b$ for the Klein bottle.

Projective Plane A very important example of a non orientable surface is the *projective plane*. It is represented by the simple fundamental "polygon" cc. In order to have some feeling for what this surface looks like it is convenient to replace each c by ab^{-1}. See Figure 5.41. When we match up edges we obtain the surface shown.

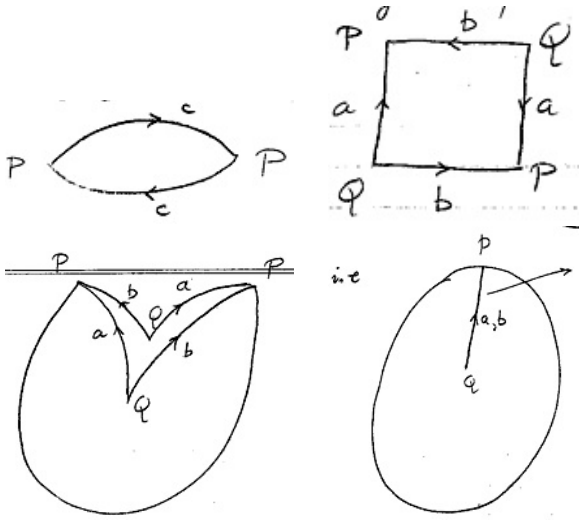


Figure 5.41: Identification diagrams for the projective plane and the projective plane in 3D space. The line PQ in the last diagram, with the arrow pointing away from it, is a line of self-intersection and is counted twice.

Cross Cap We can think of the projective plane as a *cross cap* (sometimes called a *Bishops hat*) sewn onto a sphere after removing a disc, see Figures 5.41 and 5.42.

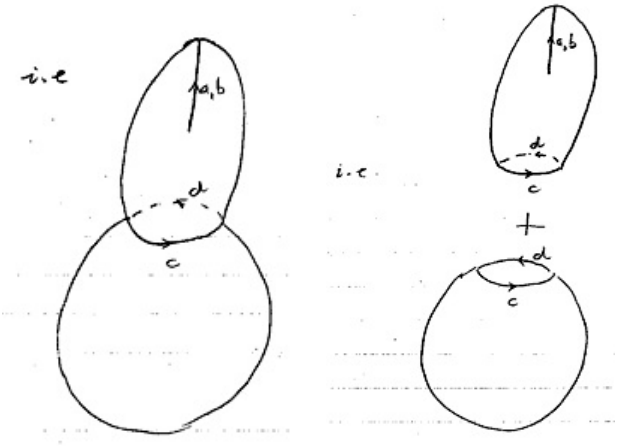


Figure 5.42: The projective plane is a cross cap sewn onto a sphere.

The cross cap is the same as the Mobius band. To see this study Figure 5.43. Renaming edges in the last diagram in Figure 5.43 it follows that they can both be represented as *abac*. *Why?*

✍

Another representation of the Mobius band is given by Figure 5.44. Renaming edges in the last diagram in Figure 5.44 it follows that the Mobius band and the cross cap can be represented by *aab*. *Why?*

✍

The Klein Bottle Again The Klein bottle is the same as two cross caps sewn together along their boundaries (i.e. is the same as two Mobius bands sewn together along their boundaries). This is shown in Figure 5.45. Renaming edges in the last diagram in Figure 5.45 it follows that the Klein bottle can be represented by *aabb*. *Why?*

✍

Since the Klein bottle is the same as two cross caps sewn along their boundaries, it follows from Figure 5.46 that the Klein bottle is also the same as two cross caps sewn to a sphere after two discs have been removed from the sphere.

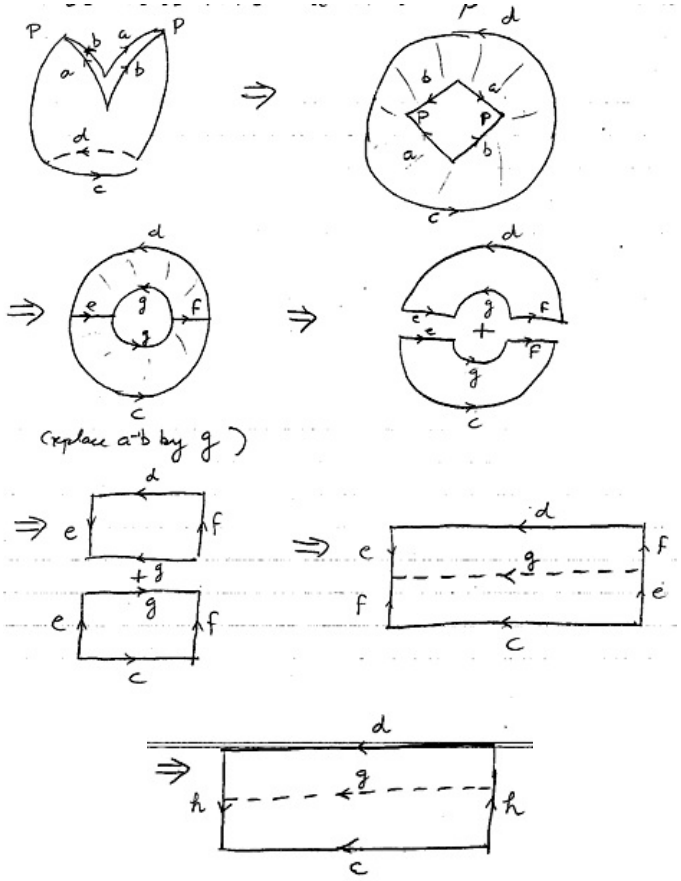


Figure 5.43: The cross cap is the Mobius band.

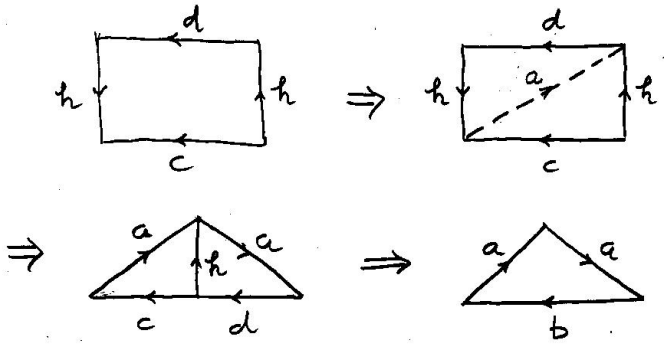


Figure 5.44: Another representation of the Mobius band, i.e. the cross cap. Note that b is the boundary and since all vertices in the final triangle are identified, b is the same as a circle.

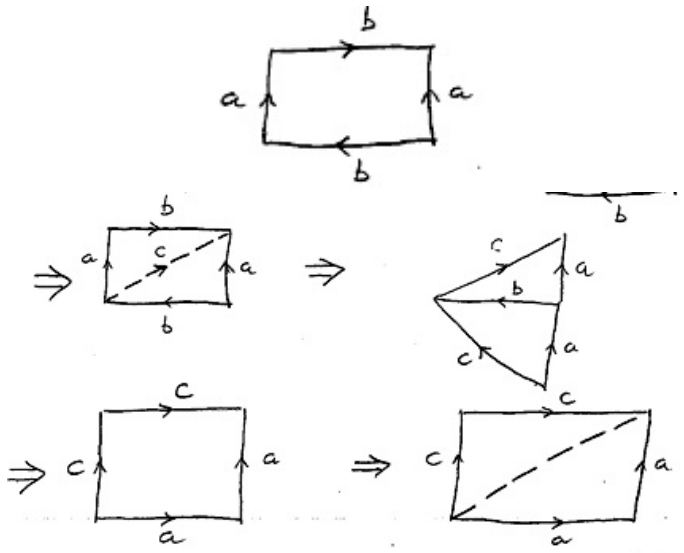


Figure 5.45: The Klein bottle is the same as two Mobius bands, i.e. cross caps, sewn together along their boundaries.

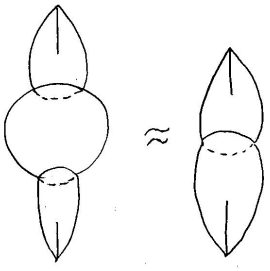


Figure 5.46: The Klein bottle is the same as two cross caps bands sewn to the sphere after removing two disks.

Summary We have seen the following fundamental polygon representations and topological equivalences (i.e. homeomorphisms):

$$\begin{aligned}
\text{projective plane}: &\quad aa \\
&\sim \text{ one cross cap sewn to sphere (after removing a disc from sphere),} \\
\text{Klein bottle}: &\quad aba^{-1}b^{-1} \text{ or } aabb \\
&\sim \text{ 2 cross caps sewn to sphere (after removing 2 discs from sphere)} \\
&\sim \text{ 2 cross caps sewn together along their boundaries,} \\
\text{Mobius band}: &\quad abac \text{ or } aab \\
&\sim \text{ cross cap.}
\end{aligned} \tag{5.13}$$

The projective plane and the Klein bottle are closed surfaces, while the Mobius band (i.e. cross cap) has a boundary.

All vertices in each fundamental polygon representation given for the pro-

jective plane and for the Klein bottle are identified. *Check this.*

In the second representation above for the Mobius band all three vertices are identified. In the first representation the four vertices of the fundamental polygon correspond to two distinct vertices on the Mobius band. *Check these facts.*

Representations of Other Non Orientable Surfaces

We saw in Figures 5.41 and 5.42 that the projective plane, represented by aa, is the same as a sphere with a cross cap. In Figures 5.44, 5.45 and 5.46 we saw that the Klein bottle, usually represented by $aba^{-1}b$, can also be represented by $aabb$, and is the same as the sphere with two cross caps.

What do the surfaces represented by $aabbcc$, $aabbccdd$ and more generally by $a_1a_1a_2a_2 \dots a_qa_q$ look like? Not surprisingly they are spheres with 3, 4 and q cross caps sewn in after removing the same number of disks.

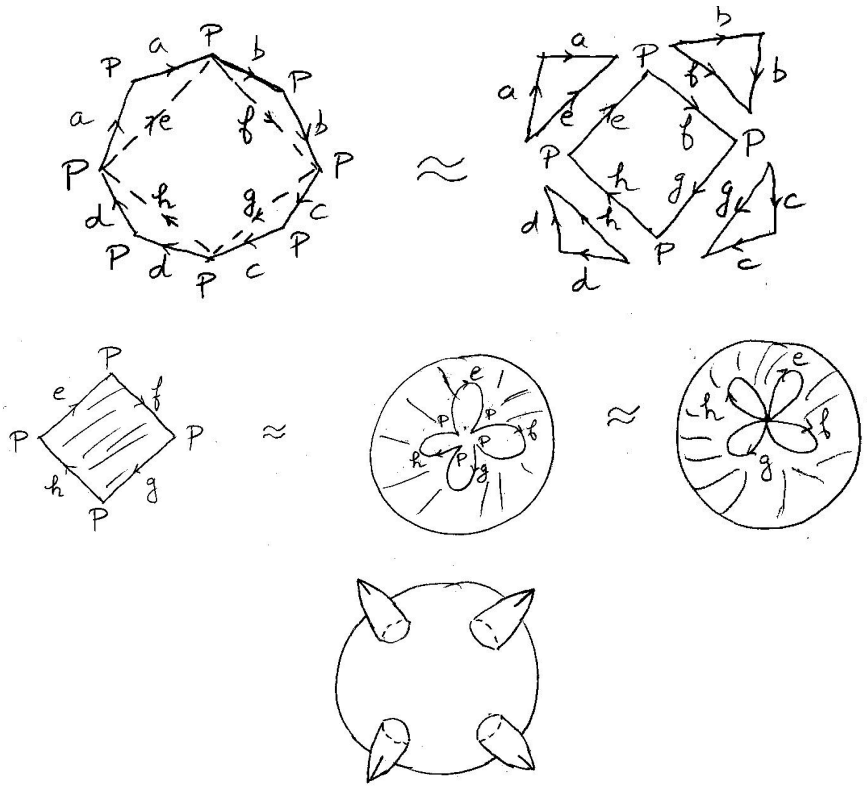


Figure 5.47: The fundamental polygon $aabbccdd$ is first disassembled into 4 cross caps and an inner rectangle. The inner rectangle (with all vertices identified) is equivalent to a sphere with 4 vertices removed as indicated in the second row. Sewing back in the 4 cross caps and moving them apart gives a sphere with 4 cross caps.

For the fundamental polygon *aabbccdd* see Figure 5.47. This polygon is equivalent to a rectangle $efgh$ plus four cross caps aae, bbf, ccg and ddh. Because all vertices are identified, $efgh$ is equivalent to a sphere with 4 discs removed and one point in common to all 4 boundaries. Sew in the cross caps. By first flattening the cross caps near P to be tangential to the sphere one can then slide the holes around the sphere to obtain 4 cross caps as in the last diagram in Figure 5.47.

The Classification Theorem

It turns out that we have now described all possible closed surfaces. More precisely we have the following theorem. (We will discuss Euler numbers in the section "Euler Numbers" on page 284.)

Theorem 5.6.1.

- *Every orientable closed surface is either*
  - *a sphere and has fundamental polygon* aa^{-1}*, or*
  - *is a sphere with* $2p$ *disks removed and* p *handles sewn in, and has fundamental polygon* $a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}\dots a_pb_pa_p^{-1}b_p^{-1}$*, for some* $p \geq 1$*.*
- *Every non orientable closed surface is a sphere with* q *disks removed and* q *cross caps sewn in, and has fundamental polygon* $a_1a_1a_2a_2\dots a_qa_q$*, for some* $q \geq 1$*.*

None of these surfaces are homeomorphic to any other. The Euler number for a sphere is 2, for a sphere with p *handles is* $2-2p$*, and for a sphere with* q *cross caps is* $2-q$*.*

A closed surface is completely determined by its orientability and its Euler number.

⋆*Proof.* (We will leave the part concerning Euler numbers for the section beginning on page 284.)

We first deal with the case that the surface is *orientable.*

1. *Represent the surface by a* single *fundamental polygon.*
   See the discussion under "Connected Surfaces" on page 268 for this part.
   Since the surface has no boundary and is orientable, each edge will occur twice in the fundamental polygon and with opposite directions.
   If there are two sides we have the sphere as in Figure 5.37. So we now assume 4 or more edges in the fundamental polygon.
2. *Remove any adjacent edges of the type* aa^{-1}*.* See Figure 5.48.
3. *Make all vertices equivalent.*
   For example, if there are two types of vertices P and Q then the number of Q vertices can be reduced to 0 by systematically cutting and pasting, and cancelling any new adjacent edges, as in Figure 5.49.

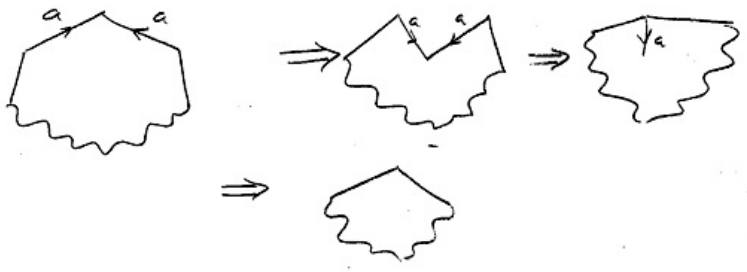


Figure 5.48: Removing two adjacent edges identified in opposite directions.

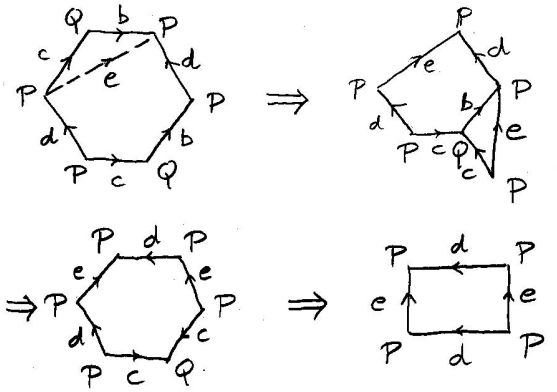


Figure 5.49: Reducing the number of Q vertices from two to one by cutting along a line e from a P vertex which is attached to a Q vertex. (The e line should have a Q vertex on both sides.) Then reduce the number of Q vertices from one to zero by collapsing two adjacent edges cc^{-1}.

4. *Put edges in the $xyx^{-1}y^{-1}$ form.*

   Suppose not every edge of the fundamental polygon is part of expressions of the form $xyx^{-1}y^{-1}$, even after renaming and switching both arrows of some pairs. See Figure 5.50. Choose one of these "bad" edges and call it a.

   Join the base points of the two a's by a line x, cut along x, and rejoin the two pieces along another common edge. (This can be shown to be always possible using the fact that all vertices are identified.) See Figure 5.50.

   The two a's will be separated by just an x, but the two x's need not be separated by just an a. In this case join the base of the two x's by a line y, cut along y and rejoin along a. See Figure 5.50. The original a's will now cancel out but, perhaps after changing arrow directions in pairs, we will have the x and y edges in the form $xyx^{-1}y^{-1}$ form and similarly for the c and d edges.

This completes the main ideas involved when the surface is orientable.

In case the surface is *non orientable* we proceed as follows.

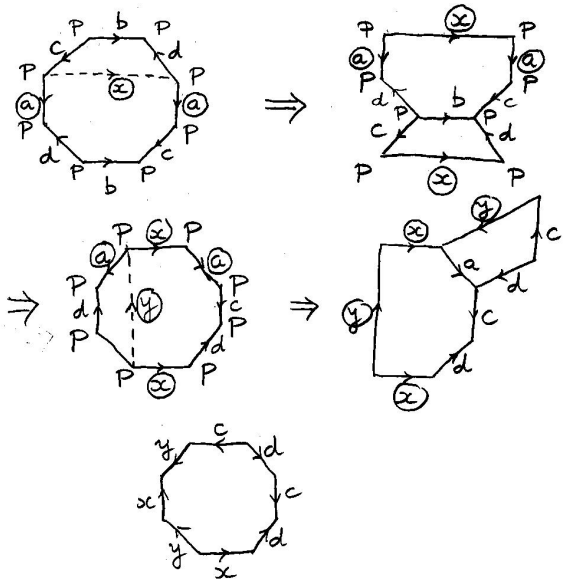


Figure 5.50: The a edges are not in the desired form as part of some $uvu^{-1}v^{-1}$ so cut along a line x joining their bases and repatch. Put in convex form. The x edges are not in the desired form so cut along a line y joining their bases and repatch along the a. After reversing arrows in pairs this gives two sequences of the form $uvu^{-1}v^{-1}$.

1. *Represent the surface by a* single *fundamental polygon.* As in Step 1 for the orientable case.
2. *Cancel any adjacent edges of the type* xx^{-1}. As in Step 2 for the orientable case.
3. *Make all vertices equivalent.* As in Step 3 for the orientable case.
4. *Replace all pairs in the* $a * a$ *form by* bb. To do this cut from the base of the first a to the base of the second a and paste along a. See Figure 5.51.

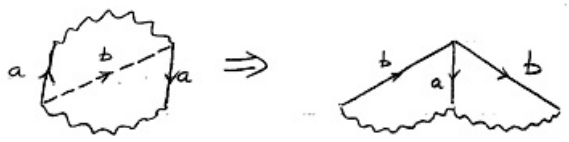


Figure 5.51: Replacing $a * a$ by bb.

5. *Remove pairs* $a*a^{-1}$ *and obtain something in the* $xyx^{-1}y^{-1}$ *form.* Similar to Step 4 in the orientable case.

   We will now have a polygon with all edges occurring as something of the form aa or $cdc^{-1}d^{-1}$. There is at least one of the former since the surface is non orientable.

6. *Remove sequences of the type* $xyx^{-1}y^{-1}$. We first replace any aa and $cdc^{-1}d^{-1}$ by 3 pairs pointing in the *same* direction, but they will not necessarily adjacent pairs. See Figure 5.52.

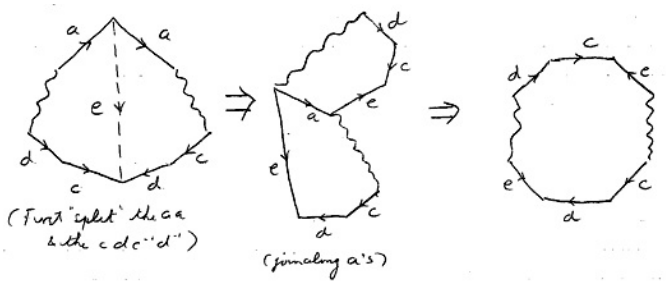


Figure 5.52: Replacing aa and $cdc^{-1}d^{-1}$ by 3 pairs of not necessarily adjacent edges, each pointing in the same direction.

Then we use the method of Step 4 to replace non adjacent pairs in the same direction by adjacent pairs in the same direction. See Figure 5.53.

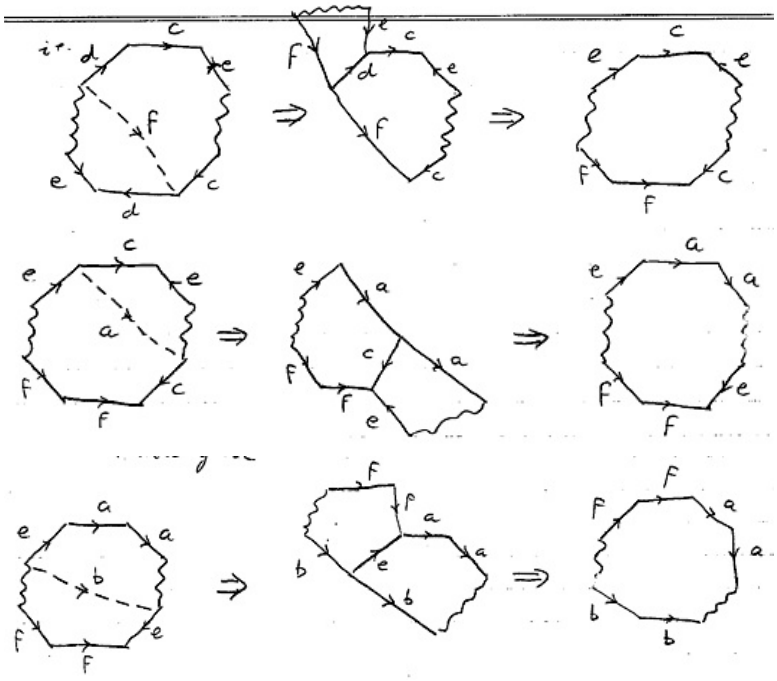


Figure 5.53: Begin with the last diagram in Figure 5.52. First replace $d * d$ by an adjacent pair ff. The f's are adjacent and in the same direction, the e's are in opposite directions and the c's are in the same direction. So in the next row we work on the c's and replace them by an adjacent pair aa. Finally in the last row we replace $e * e$, now pointing in the same direction, by an adjacent pair bb.

In this way, assuming the surface is non orientable and so has at least one pair initially pointing in the same direction, we end up with only adjacent pairs and the edges of each pair point in the same direction. By renaming edges such as $e^{-1}e^{-1}$ to ee, we finally get the for $aabbccddee\ldots$. This gives the required form.

□

Cut and Paste Examples

Example 1 We want to find the standard form for the surface corresponding to the first diagram in Figure 5.54. The surface is non orientable because of

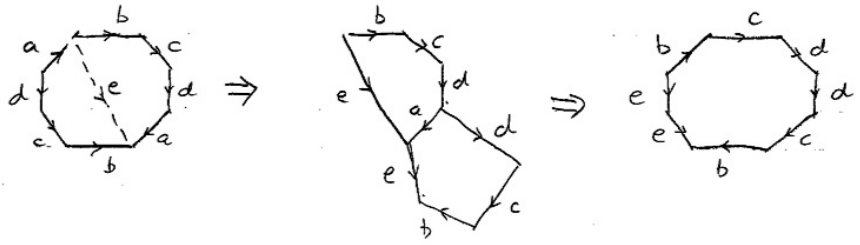


Figure 5.54: Beginning with the fundamental polygon on the left, work on the non adjacent pair of a's.

the a's. We can check that all vertices are identified. *Do it.*

Proceeding to Step 4 for non orientable surfaces we work on the pair of non adjacent a's. This gives adjacent d's and c's.

Next work on the non adjacent b's as in Figure 5.55.

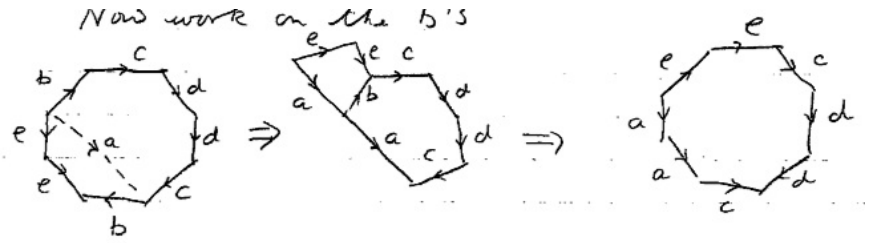


Figure 5.55: Work on the non adjacent pair of b's.

Finally we work on the c's and get after renaming $aabbccdd$, a sphere with 4 crosscaps. *Do it.*

Example 2 We want to find the standard form for the surface corresponding to the first diagram in Figure 5.56. The surface is non orientable because of

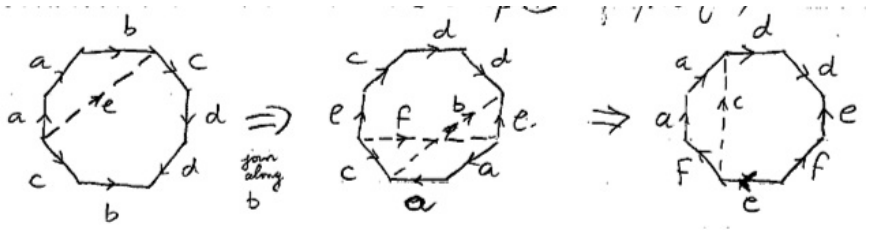


Figure 5.56: Remove the non adjacent pair of c's pointing in opposite directions.

the a's (also because if the d's).

Because pairs pointing in the same direction are already adjacent we go to Step 5 for non orientable surfaces and work on the two c edges pointing in opposite directions.

By Step 6 for non orientable surfaces we know we can replace $e^{-1}f^{-1}ef$ in the last diagram in Figure 5.56. We do it for practice in Figure 5.57. After

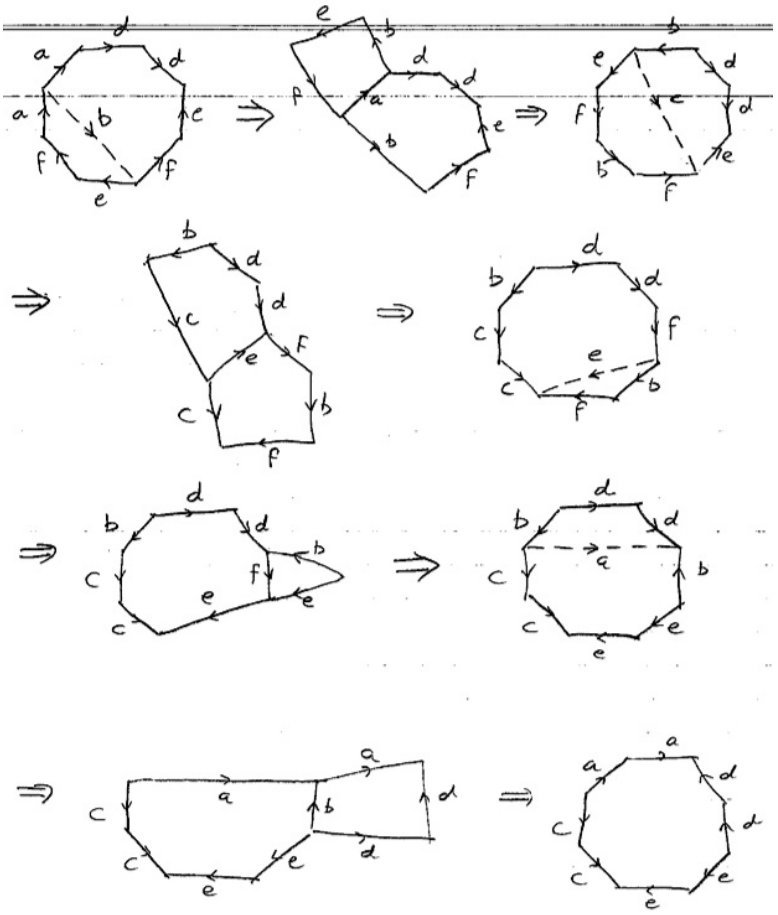


Figure 5.57: Replace $e^{-1}f^{-1}ef$ by pairs of edges in the same directions.

renaming we get *aabbccdd*, a sphere with 4 crosscaps.

Euler Numbers

I will be fairly brief in this section.[26]

Properties Suppose S is a closed surface covered by a finite set of polygons, having common edges in pairs, as in Figure 5.32. The Euler Number or *Euler Characteristic* of S is given by

$$V - E + F$$

where V is the number of vertices, E is the number of edges and F is the number of faces.

We have the following important facts:

1. *the Euler number of a surface S depends only on S and not on the covering used;*
2. *if two surfaces are homeomorphic then they will have the same Euler number, so if they have different Euler numbers then they are not homeomorphic;*
3. *the Euler number can be computed from any identification diagram for S, and in particular from the fundamental polygon of S put in standard form as in Theorem 5.6.1.* (For counting purposes we need to take account of the fact that each pair of edges in the identification diagram corresponds to one edge in S, and that many vertices in the identification diagram will correspond to one vertex in S.)

The reason for 1. is that we can change from one covering to another by adding or subtracting vertices from the middle of edges and by adding or subtracting edges joining a pair of vertices. None of these operations changes the Euler number, by an argument similar to that in the *Third Step* on page 235 or in [HM: Section 5.3].

The reason for 2. is that we can use the homeomorphism to pass from a covering by polygons of the first surface to a covering of the second surface, and this does not change the number of vertices, edges or faces. *Why?*

The reason for 3. is that the Euler number for a surface is clearly the same as the Euler number for the identification diagram corresponding to the covering of the surface, *provided* we take account of the fact that each edge in the surface covering is represented twice in the identification diagram and that each vertex in the surface covering will also be represented a number of times in the identification diagram. Moreover, when we do cut and paste operations on identification diagrams the Euler number is unchanged. This also follows by an argument similar to that in the *Third Step* on page 235 or in [HM: Section 5.3].

Computing the Euler Number The Euler number of the fundamental polygon for the *sphere*, see Figure 5.37, is

$$V - E + F = 2 - 1 + 1 = 2.$$

[26] The details will be filled out in a later version.

Notice that the two vertices are not identified, there is only one edge after identification, and there is one face.

The Euler number for the sphere with p handles, i.e. the torus with p holes, is

$$V - E + F = 1 - 2p + 1 = 2 - 2p.$$

This comes from considering the fundamental polygon

$$a_1 b_1 {a_1}^{-1} {b_1}^{-1} a_2 b_2 {a_2}^{-1} {b_2}^{-1} \dots a_p b_p {a_p}^{-1} {b_p}^{-1}.$$

All vertices are identified so there is really only one vertex, there are 2p distinct edges and there is one face. See Figures 5.38, 5.39 and 5.40 for the cases $p = 1, 2, 3$ respectively.

The Euler number for the sphere with q crosscaps is

$$V - E + F = 1 - q + 1 = 2 - q.$$

This comes from considering the fundamental polygon

$$a_1 a_1 a_2 a_2 \dots a_p a_p.$$

All vertices are identified, there are q distinct edges and there is one face. See Figures 5.41, 5.45 (last diagram) and 5.47 for the cases $q = 1, 2, 4$ respectively.

The Classification Theorem Again The previous discussion completes the proof of the assertions in the last two paragraphs of Theorem 5.6.1. The sphere or a sphere with handles is orientable, and so cannot be homeomorphic to a sphere with crosscaps which is nonorientable.[27] Moreover, the sphere, and spheres with different numbers of handles, have different Euler numbers and so cannot be homeomorphic to one another by fact 2. on page 284. Similarly, spheres with different numbers of crosscaps have different Euler numbers and so cannot be homeomorphic to one another, again by fact 2. on page 284.

Questions

1 Use Theorem 5.6.1 to describe the surfaces with Euler number 2, 1, 0, -1, -2, -3, -4, -5, -6, -7, -8.

[27] An orientable surface cannot be homeomorphic to a non orientable surface. *Why?*