Basic Books in Science
– a Series of books that start *at the beginning*

Book 2

Space – from Euclid to Einstein

Roy McWeeny
Professore Emerito di Chimica Teorica, Università di Pisa, Pisa (Italy)

Books in the Series will be available *–free of charge–*from the websites
<www.paricenter.com> (see 'Basic Books in Science')
<www.learndev.org> (see 'For the Love of Science')

(Last updated 10 June 2007)

BASIC BOOKS IN SCIENCE

About this Series

All human progress depends on **education**: to get it we need books and schools. Science Education is of key importance.

Unfortunately, books and schools are not always easy to find. But nowadays all the world's knowledge should be freely available to everyone – through the Internet that connects all the world's computers.

The aim of the Series is to bring basic knowledge in all areas of science within the reach of everyone. Every Book will cover in some depth a clearly defined area, starting from the very beginning and leading up to university level, and will be available on the Internet *at no cost to the reader*. To obtain a copy it should be enough to make a single visit to any library or public office with a personal computer and a telephone line. Each book will serve as one of the 'building blocks' out of which Science is built; and together they will form a 'give-away' science library.

About this book

This book, like the others in the Series, is written in simple English – the language most widely used in science and technology. It takes the next big step beyond "Number and symbols" (the subject of Book 1), starting from our first ideas about the measurement of distance and the relationships among objects in space. It goes back to the work of the philosophers and astronomers of two thousand years ago; and it extends to that of Einstein, whose work laid the foundations for our present-day ideas about the nature of space itself. This is only a small book; and it doesn't follow the historical route, starting from geometry the way Euclid did it (as we learnt it in our schooldays); but it aims to give an easier and quicker way of getting to the higher levels needed in Physics and related sciences.

Looking ahead –

Like the first book in the Series, Book 2 spans more than two thousand years of discovery. It is about the science of space – **geometry** – starting with the Greek philosophers, Euclid and many others, and leading to the present – when space and space travel is written about even in the newspapers and almost everyone has heard of Einstein and his discoveries.

Euclid and his school didn't trust the use of numbers in geometry (you saw why in Book 1): they used pictures instead. But now you've learnt things they didn't know about – and will find you can go further, and faster, by using numbers and algebra. And again, you'll pass many 'milestones':

- In Chapter 1 you start from distance, expressed as a number of **units**, and see how Euclid's ideas about straight lines, angles and triangles can be 'translated' into statements about distances and *numbers*.

- Most of Euclid's work was on geometry of the **plane**; but in Chapter 2 you'll see how any point in a plane is fixed by giving two numbers and how lines can be described by equations.

- The ideas of **area** and **angle** come straight out of plane geometry (in Chapter 3): you find how to get the area of a circle and how to measure angles.

- Chapter 4 is hard, because it ties together so many very different ideas, mostly from Book 1 – **operators**, **vectors**, **rotations**, **exponentials**, and **complex numbers** – they are all connected!

- Points which are *not* all in the same plane, lie in **3-dimensional space** – you need *three* numbers to say where each one is. In Chapter 5 you'll find the geometry of 3-space is just like that of 2-space; but it looks easier if you use vectors.

- Plane shapes, such as triangles, have properties like area, angle and side-lengths that don't change if you move them around in space: they belong to the shape itself and are called **invariants**. Euclid used such ideas all the time. Now you'll go from 2-space to 3-space, where objects also have **volume**; and you can still do everything without the pictures.

- After two thousand years people reached the last big milestone (Chapter 7): they found that Euclid's geometry was very nearly, *but not quite*, perfect. And you'll want to know how Einstein changed our ideas.

CONTENTS

Notes to the Reader. When Chapters have several Sections they are numbered so that "Section 2.3" will mean "Chapter 2, Section 3". Similarly, "equation (2.3)" will mean "Chapter 2, equation 3". Important 'key' words are printed in **boldface**: they are collected in the Index at the end of the book, along with the numbers of the pages where you can find them.

Chapter 1

Euclidean space

1.1 Distance

At the very beginning of Book 1 we talked about measuring the distance from home to school by counting how many strides, or paces, it took to get there: the pace was the *unit* of distance and the number of paces was the *measure* of that particular distance. Now we want to make the idea more precise.

The *standard* unit of distance is '1 metre' (or 1 m, for short) and is defined in a 'measuring-rod', with marks at its two ends, the distance between them fixing the unit. Any other pair of marks (e.g. on some other rod, or stick) are also 1 m apart if they can be put in contact, at the same time, with those on the standard rod; and in this way we can make as many copies of the unit as we like, all having the same length. In Book 1 we measured distances by putting such copies end-to-end (the 'law of combination' for distances) and if, say, three such copies just reached from one point to another then the two points were '3 m apart' – the 'distance between them was 3 m', or 'the length of the path from one to the other was 3 m' (three different ways of saying the same thing!).

Now the number of units needed to reach from one point 'A' to another point 'B' will depend on how they are put together: if they form a 'wiggly' line, like a snake, you will need more of them – the path will be longer. But the distance does not change: it is the **unique** (one and one only) *shortest* path length leading from A to B. (Of course the path length may not be exactly a whole number of units, but by setting up smaller and smaller 'mini-units' – as in Book 1, Chapter 4 – it can be measured as accurately as we please and represented by a decimal number.) The important thing is that the distance AB is the length of the *shortest path* between A and B. In practice, this can be obtained by marking the units (and mini-units) on a string, or tape, instead of a stiff measuring-rod: when the tape measure is pulled tight it can give a fairly accurate measure of the distance AB. The shortest path defines a **straight line** between the points A and B.

One thing we must remember about measuring distance (or any other quantity, like mass or time) – it's always a certain *number of units*, not the number itself. The distance from home to school may be 2000 m (the unit being the metre), but 2000 by itself is just a number: quantity = number × unit, where the number is the *measure* of the quantity in terms of a chosen unit. We can always change the unit: if a distance is large we can measure it in kilometres (km) and since 1 km means 1000 m the distance (d, say) from home to school will be $d = 2000$ m $= 2$ km. If we make the unit a thousand times bigger, the number that measures a certain quantity will become a thousand times *smaller*. Thus,

$$\begin{aligned} d &= \text{old measure} \times \text{old unit} \\ &= \text{new measure} \times \text{new unit} \\ &= \tfrac{\text{old measure}}{1000} \times (1000 \times \text{old unit}) \end{aligned}$$

and the same rule always holds. In some countries the unit of distance is the 'mile' and there are roughly 8 km to every 5 miles: 1 mile = (8/5) km. Thus, if I want a distance in miles instead of kilometers, (new unit)= (8/5)×(old unit) and (new measure)=(5/8)×(old measure). In this way we see the distance to the school is (5/8)×2 mile = 1.25 mile. Doing calculations with quantities is often called 'quantity calculus' – but there's nothing mysterious about it, it's just 'common sense'!

Euclidean **geometry** (the science of space) is based on the foundations laid by Euclid, the Greek philosopher, working around 300 BC) it starts from the concepts of points and straight lines; and it still gives a good description of the spatial relationships we deal with in everyday life. But more than 2000 years later Einstein showed that, in dealing with vast distances and objects travelling at enormous speeds, Euclidean geometry does not quite fit the experimental facts: the theory of **relativity** was born. One of the fundamental differences, in going from Euclid to Einstein, is that the shortest path between two points is *not quite* a 'straight line' – that space is 'curved'. There is nothing strange about this: a ship does not follow the shortest path between two points on the surface of the earth – because the earth is like a big ball, the surface is not flat, and what seems to be the shortest path (according to the compass) is in reality not a straight line but a curve. The strange thing is that *space itself* is very slightly 'bent', especially near very heavy things like the sun and the stars, so that Euclid's ideas are never *perfectly* correct – they are simply so close to the truth that, in everyday life, we can accept them without worrying.

In nearly all of Book 2 we'll be talking about Euclidean geometry. But instead of doing it as Euclid did – and as it's done even today in many schools all over the world – we'll make use of algebra (Book 1) from the start. So we won't follow history. Remember, the Greeks would not accept irrational numbers (Book 1, Chapter 4) so they couldn't express their ideas about space in terms of distances and had to base their arguments entirely on *pictures*, not on numbers. This was why algebra and geometry grew up separately, for two thousand years. By looking at mathematics as a whole (not as a subject with many branches, such as algebra, geometry, trigonometry) we can reach our goal much more easily.

1.2 Foundations of Euclidean geometry

The fact that the space we live in has a 'distance property' – that we can experimentally measure the distance between any two points, A and B say, and give it a *number* – will

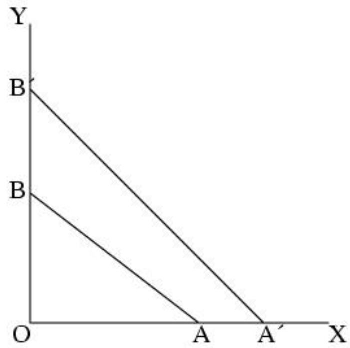

Figure 1

be our starting point. We make it into an 'axiom' (one of the basic principles, which we take as 'given'):

The distance axiom

There is a **unique** (one and one only) shortest path between two points, A and B, called the *straight line* AB; its length is the *distance* between A and B.

The first thing we have to do is talk about the properties of straight lines and the way they give us a foundation for the whole of Euclidean geometry. In fact, Euclid's geometry can be built up from the following 'construction', indicated in Fig.1, which can be checked by experiment. We take it as a second axiom:

The metric axiom

Given any two points, A and B, we can find a third, which we call the 'origin' O, such that the distances OA, OB, and AB are related by

$$AB^2 = OA^2 + OB^2 \tag{1.1}$$

and if the straight lines OA and OB are extended as far as we please (as in Fig.1) then the distance $A'B'$, between any two other points (A′, B′) is given by the same formula: $(A'B')^2 = (OA')^2 + (OB')^2$. (Note that AB, $A'B'$, etc denote single quantities, lengths, not products.)

Whenever this construction is possible mathematicians talk about **Euclidean space**; and say that (1.1) defines the 'metric' for such a space ('metric' meaning simply that distances can be measured). You can test (1.1) by taking special cases. For example, with OA = 3 cm ('cm' meaning 'centimetre', with 100 cm =1 m) and OB = 4 cm you will find AB = 5 cm; and $3^2 = 9$, $4^2 = 16$, so the sum of the squares is 9 + 16 = 25 – which is 5^2. The same formula is satisfied by OA = 5 cm, OB= 12 cm, and AB= 13 cm (25 + 144 = 169 = 13^2). If you take OA= 4 cm, OB = 5 cm you should find AB = 6.403 cm, because 6.403 is the square root of 41 (= 16 + 25). This construction gives us several new ideas and definitions:

- The lines OA and OB in Fig.1 are **perpendicular** or at **right angles**. The straight lines formed by moving A and B further and further away from the origin O, in either direction, are called **axes**. OX is the x-axis, OY is the y-axis.

- The points O, A, and B, define a 'right-angled' **triangle**, OAB, with the straight lines OA,OB,AB as its three sides. (The 'tri' means three and the 'angle' refers to the lines OA and OB and will measure how much we must turn one line around the origin O to make it point the same way as the other line – more about this later!)

- All straight lines, such as AB or A′B′, which **intersect** (i.e. cross at a single point) both axes, are said to 'lie in a **plane**' defined by the two axes.

From the axiom (1.1) and the definitions which follow it, the whole of geometry – the science we need in making maps, in dividing out the land, in designing buildings, and everything else connected with relationships in space – can be built up. Euclid started from different axioms and argued with pictures, obtaining key results called **theorems** and other results (called **corollaries**) that follow directly from them. He proved the theorems one by one, in a logical order where each depends on theorems already proved. He published them in the 13 books of his famous work called "The Elements", which set the pattern for the teaching of geometry throughout past centuries. Here we use instead the methods of algebra (Book 1) and find that the same chain of theorems can be proved more easily. Of course we won't try to do the whole of geometry; but we'll look at the first few links in the chain – finding that we don't need to argue with pictures, we can do it all with numbers! The pictures are useful because they remind us of what we are doing; but our arguments will be based on *distances* and these are measured by numbers.

This way of doing things is often called **analytical geometry**, but it's better not to think of it as something separate from the rest – it's just a part of a 'unified' ('made-into-one') mathematics.

Exercises

(1) Make a **tape measure** from a long piece of tape or string, using a metre rule to mark the centimetres, and use it to measure

- the distance (d) between opposite corners of this page of your book;

- the lengths of the different sides (x and y);

- the distance (AB) between two points (A and B) on the curved surface of a big drum (like the ones used for holding water), keeping the tape tightly stretched and always at the same height;

- the distance between A and B (call it L), when A and B come close together and the tape goes all the way round (this is called the **circumference** of the drum);

- the distance between two *opposite* points on the bottom edge of the drum (this – call it D – is the **diameter** of the drum).

(2) Check that the sum of x^2 and y^2 gives you d^2, as (1.1) says it should.

(3) Note that L is several times bigger than D: how many? (Your answer should give you roughly the number π (called "pi") that gave the Greeks so much trouble – as we know from Book 1)

(4) In some countries small distances are measured in "inches" instead of cm, 1 inch being roughly 2.5 cm (the length of the last bit of your thumb). Put into inches all the distances you measured in Exercise 1. Show that the answers you got in Exercises 2 and 3 are unchanged.

(5) Make a simple **set square** – a triangle like OAB in Fig.1, with sides of 9 cm, 12 cm and 15 cm, cut out from a piece of stiff card. Use it to mark out axes OX and OY on a big sheet of paper (e.g. newspaper or wrapping paper). Then choose several pairs of points, like A,B or A′,B′ in Fig.1, and verify that the distances AB,$A'B'$ etc. are always related to OA and OB (or OA' and OB' etc.) by equation (1.1).

(6) Take a big rectangular box and measure the lengths (a, b, c) of the three different edges and the distance (d) between opposite corners (the ones as far apart as you can get). Show, from your measurements, that $d^2 \approx a^2 + b^2 + c^2$, where the sign $\approx$ means 'approximately' or 'nearly' equal. Use the formula (1.1) to show that the 'exact' result should be

$$d^2 = a^2 + b^2 + c^2.$$

(Measurements are never quite perfect – so you can never use them to *prove* something.)

Chapter 2

Two-dimensional space

2.1 Parallel straight lines. Rectangles

A *plane* has been defined in the last Section: it is a region based on two intersecting straight lines of unlimited length, called axes. All straight lines which cut the two axes lie in the same plane and *any* pair with one point in common (to take as 'origin') can be used as alternative axes. Such a plane is a **two-dimensional** region called, for short, a **2-space**.

A special relationship between two intersecting straight lines is *perpendicularity*, defined in Section 1.2: two lines are perpendicular when they form a right angle. Thus, the lines AB and AP in Fig.2 are perpendicular and $BP^2 = AB^2 + AP^2$. (Note that the lines AQ and BP, shown as 'broken' lines in the Figure, are only put in to help us: they are "construction lines". Also AQ, for example, shown in Italic (sloping) type as AQ, is used to mean the *length* of the line AQ, which is measured by a single number.)

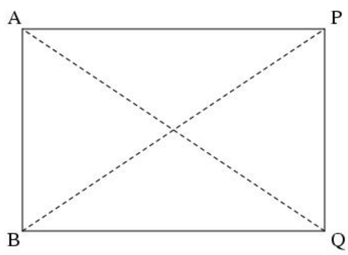

Figure 2

We now need another definition:

> *Definition.* If two straight lines in a plane are perpendicular to a third, they are said to be **parallel**.

Let's also note that in our 2-space *all* our straight lines lie in the same plane – so we won't always say it!

With this definition we can go to a first theorem:

> *Theorem.* Any straight line perpendicular to *one* of two parallel straight lines is also perpendicular to the *other*.

Proof (If you find a proof hard, skip it; you can come back to it any time.)

Suppose AB and PQ in Fig.2 are parallel, both being perpendicular to AP (as in the *Definition*), and that BQ is the other straight line perpendicular to AB. Then we must show that BQ is also perpendicular to PQ. In symbols, using (1.1),

$$\text{Given} \quad AP^2 + PQ^2 = AQ^2,$$

$$\text{show that} \quad BQ^2 + QP^2 = BP^2 = BA^2 + AP^2.$$

We must show that there *is* a point Q such that these relationships hold.

The lengths BQ and QP are unknown (they depend on where we put Q), but the possibilities are

$$\begin{array}{lll} (a) & BQ = AP, & PQ = AB, \\ (b) & BQ = AP, & PQ \neq AB, \\ (c) & BQ \neq AP, & PQ = AB, \\ (d) & BQ \neq AP, & PQ \neq AB. \end{array}$$

It is easy to see that (b) is not possible, because if $BQ = AP$ then $AQ^2 = AB^2 + BQ^2 = AB^2 + AP^2$; while $AQ^2 = AP^2 + PQ^2$. The two expressions for AQ^2 are only the same when $PQ = AB$, so possibility (b) is ruled out; and, by a similar argument, so is (c).

If we accept (a), it follows that $BQ^2 + QP^2 = BP^2 \quad (= BA^2 + AP^2)$ and this is the condition for the lines BQ and QP to be perpendicular: the theorem is then true. But when Q is fixed in this way possibility (d) is also ruled out – because it would mean there was *another* crossing point, Q′ say, with $BQ' \neq BQ$ and $PQ' \neq PQ$, whereas the perpendicular from B can intersect another line at only *one* point, already found. So (a) must apply and the theorem follows: BQ is perpendicular to PQ.

The proof of the theorem introduces other ideas:

(i) Plane ‘figures’ (or shapes), like the ‘box’ in Fig.2, are formed when two pairs of parallel lines intersect at right angles: they are called **rectangles** and their opposite sides are of equal length. When *all* sides have the same length the shape is a **square**.

(ii) There is only one shortest path from a point to a given straight line, this forming the line from the point to the given line and perpendicular to it.

(iii) The shortest path between two parallel straight lines, in a plane, is a straight line perpendicular to both; and all such paths have exactly the same length. This rules out the possibility of the parallel lines ever meeting (one of Euclid’s first axioms), since the shortest path would then have zero length for all pairs of points and the two lines would then coincide (i.e. there would be only one).

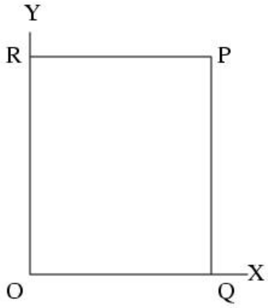

Figure 3

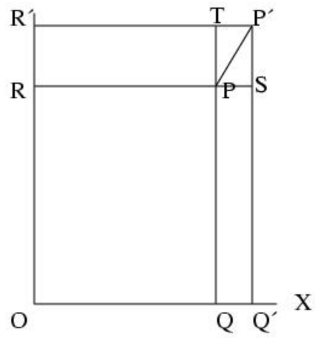

Figure 4

2.2 Points and straight lines in 2-space

We're now ready to describe any point in a plane by means of two numbers (more correctly they are *distances* but as in Book 1, Chapter 1 we'll often call them 'numbers', each distance being a number of units). Suppose the plane is defined by two axes, OX and OY in Fig.3, which we take to be perpendicular. From any point P we can drop perpendiculars onto OX and OY; and the position of the point P is then fixed by giving two distances, OQ ($= RP$) and OR ($=PQ$), the equalities following because ORPQ is a rectangle. These two distances, which we denote by x and y respectively, are called the **rectangular (or 'Cartesian') coordinates** of P with respect to the axes OX and OY. We'll always use axes that are perpendicular, for simplicity, and x, y are also called the **projections** of the line OP, from the origin to the point, on the axes. Any point in the plane is shown by giving its coordinates (x, y); and the whole of plane geometry can be developed *algebraically* in terms of the number-pairs (x, y) defining the points we want to talk about.

First let's think about straight lines. If P and P′ are any two points in the plane we can drop perpendiculars, as in Fig.4, to find their coordinates, namely (x, y) and (x', y'). From the results earlier in this Section, the line from P to P′ has projections $QQ' = x' - x$ and $RR' = y' - y$ on the two axes; and $QQ' = PS$, $RR' = PT$. The length of the line PP′, the separation of P and P′ (denoted by s) thus follows from

$$s^2 = (x' - x)^2 + (y' - y)^2 \tag{2.1}$$

and this is true no matter how close or far apart P and P′ may be. The starting point for Euclidean geometry (1.1) is now expressed in terms of coordinates in the form (2.1): it is usually written in the case where P and P′ are very close, so $x' - x$ and $y' - y$ are very small differences which we denote by $\mathrm{d}x$ and $\mathrm{d}y$, respectively, and call **differentials**. More about differentials in Book 3, Section 2.3. For now, just note that "d" in Roman type (written with a straight back) is used to mean that $\mathrm{d}x$ is "a little bit of x", not a product of two quantities d and x. (Remember that numbers and quantities are always shown in Italic, sloping, type.)

For points close enough together, then, (2.1) can be written

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2, \tag{2.2}$$

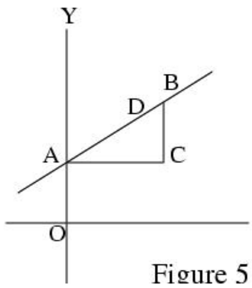

Figure 5

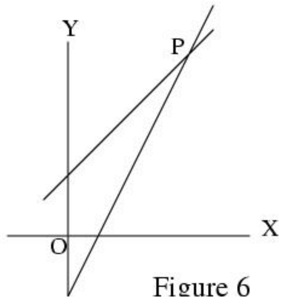

Figure 6

which is called the 'fundamental metric form'. In Euclidean space, the 'sum-of-squares' form is true whether the separation of two points is large or small: but if you are making a map you must remember that the surface of the earth is curved – so you can use (2.2) for small distances (e.g. your town) but not for large distances – your country. (Strictly speaking, (2.2) is only true 'in the limit' (see Book 1, Chapter 4) when the distances go to zero.) Space may be only *locally* Euclidean. Within the last hundred years our ideas about space have changed a lot, but in everyday life Euclidean geometry still serves perfectly well.

Now let's ask how to describe a straight line using rectangular Cartesian coordinates. Suppose the line intersects the y-axis at the point A with coordinates $(0, c)$ and that it is fixed by giving the coordinates (x_1, y_1) of just one other point, B, that lies on it (see Fig.5). The points A,B,C then define a right-angled triangle, whose sides AC and BC have lengths such that $(BC/AC) = m$: we say they stand in some definite *ratio* m, which is just a number – whatever units we use in measuring them. In terms of coordinates, this means $y_1 - c = mx_1$; and it then follows that the coordinates (x, y) of *any* point D, on the same line, are related in a similar way:

$$y = c + mx. \tag{2.3}$$

To test that the new point D, with coordinates related by (2.3), does lie on the same shortest path between A and B, we can use the length formula (2.1): thus $AB^2 = (y_1 - c)^2 + mx_1^2 = (1 + m^2)x_1^2$, $AD^2 = (1 + m^2)x^2$, and $DB^2 = (1 + m^2)(x_1 - x)^2$. On taking the square roots, $AB = \sqrt{1 + m^2}\, x_1, \quad AD = \sqrt{1 + m^2}\, x, \quad DB = \sqrt{1 + m^2}\, (x_1 - x)$.

From this it follows that $AD + DB = AB$; but this means that the two paths, AB and ADB (i.e. A to B, passing through the new point D), both have the same length – that of the unique shortest path between A and B. When the coordinates of any point D are related by (2.3) then that point lies on the straight line through A and B.

We say that (2.3) is the 'equation of a straight line', $m = BC/AC$ being called the **slope** of the line and $c = OA$ being its **intercept** on the y-axis.

Note that equation (2.3) will describe *any* straight line in the plane OXY and that the proof just given does not depend on point D being *between* A and B: if, for example, $x > x_1$, Fig.5 would show D on the line extended beyond B; and a similar argument would show that B must lie on the straight line AD. But we don't have to draw a different picture for every possible case: if x, y refer to points on the left of the y-axis, or beneath the x-axis, they will simply take negative values – and, as the laws of algebra hold for any numbers, our results will always hold good.

Sometimes two lines in a plane will cross, meeting at some point P, as in Fig.6. Whether they do or not is an important question – which was the starting point for all of Euclid's great work.

2.3 When and where will two straight lines meet?

Let's now look again at Euclid's 'parallel axiom' – that two parallel straight lines *never* meet. What does it mean in algebra?

Suppose the two lines have equations like (2.3) but with different values of slope (m) and intercept (c): let's say

$$y = c_1 + m_1 x, \qquad y = c_2 + m_2 x. \tag{2.4}$$

In Fig.6 two such lines cross at the point P. How can we find it? The first equation in (2.4) relates the x and y coordinates of any point on Line 1, while the second equation does the same for any point on Line 2. At a crossing point, the same values must satisfy both equations, which are then called **simultaneous equations** (both must hold at the same time). It is easy to find such a point in any given case: thus, if $m_1 = 1,\ m_2 = 2$ and $c_1 = 1,\ c_2 = -1$, the values of x and y must be such that

$$y = 1 + x \quad \text{and} \quad y = -1 + 2x,$$

which arise by putting the numerical values in the two equations. Thus, we ask that $1 + x = -1 + 2x$ at the crossing point; and this gives (see the Exercises in Book 1, Chapter 3) $x = 2$, with a related value of $y = 1 + x = 3$. This situation is shown in Fig.6, Point P being (2,3). If, instead, we took the second line to have the same slope ($m_2 = 2$) but with $c_2 = 3$, the result would be $x = -2,\ y = -1$. Try to get this result by yourself.

Finally, suppose the two lines have the same slope, $m_1 = m_2 = m$. In this case (x, y) at the crossing point must be such that

$$y = c_1 + mx = c_2 + mx,$$

which can be true only if $c_1 = c_2$, whatever the common slope of the two lines: but then the two lines would become the same (same slope and same intercept) – there would be only one! *All* points on either line would be 'crossing points'. As long as $m_1 \neq m_2$ we can find a true crossing point for $c_1 \neq c_2$; but as m_1 and m_2 become closer and closer the distance to the crossing point becomes larger and larger. This 'Point 3' can't be shown in Fig.6 – it is 'the point at infinity'!

This simple example is very important: it shows how an algebraic approach to geometry, based on the idea of *distance* and the metric (1.1), can lead to general solutions of geometrical problems, without the need to draw pictures for all possible situations; and it shows that Euclid's famous axiom, that parallel lines never meet, then falls out as a first result.

Before going on, let's look at one other simple shape in 2-space – the **circle** which the ancients thought was the most perfect of all shapes. It's easy enough to draw a perfect circle: you just hammer a peg into the ground and walk round it with some kind of marker, attached to the peg by a tightly stretched piece of string – the marker will mark out a circle! But how do you describe it in algebra?

Let's take the peg as origin O and the marker as point P, with coordinates x, y, say. Then if your string has length l, and you keep it tight, you know that the distance OP (the third side of a right-angled triangle, the other sides having lengths x and y) will always be the same – always l. But with the sum-of-squares metric this means

$$x^2 + y^2 = l^2 = \text{constant}, \tag{2.5}$$

however x and y may change. We say this is the "equation of a circle" with its centre at the origin O; just as (2.4) was the equation of a straight line, with a given slope (m) and crossing the y-axis at a certain point ($y = c$). The equation of the circle is of the 'second degree' (x and y being raised to the power 2); while that of the line is of the 'first degree' or *linear*. In the Exercises and in other Chapters you'll find many more examples.

Exercises

1) Suppose the corners of the rectangle in Fig.3 are at the points O(0,0), Q(3,0), P(3,4), R(0,4) and draw the straight line $y = \frac{1}{2}x$. At what point does it meet the side QP? (Any point on QP must have $x = 3$. So you only need to choose y.)

2) What happens if the line through the origin in Ex.1 is changed to $y = 2x$? (The point found in Ex.1 lies *between* Q and P: it is an *internal* point. The new point will lie on QP *extended* (beyond P): it is an *external* point, lying *outside* QP.)

3) Repeat Exercises 1 and 2, using in turn the lines

$$y = 3 - \tfrac{1}{2}x, \quad y = 3 - 2x, \quad y = -3 + \tfrac{1}{2}x, \quad y = -3 + 2x,$$

and describe your results.

4) Instead of using equation (2.3), take $y = 2 + \frac{1}{2}x^2$ and draw the curve of y against x. The new equation describes a **parabola**. Find values of x and y that fit the equation, using, in turn, the values $x = -3, -2, -1, 0, +1, +2, +3$ and 'plot' them (i.e. mark the points in a Figure and join them by a curved line.)

Find the points where the straight lines in Ex.3 cross the parabola (you need to know how to solve a *quadratic equation* – see Section 5.3 of Book 1) and show your results in a Figure.

Note In all the Exercises x, y, etc. are represented in the Figures as *distances*, so each stands for a number of *units*; but the size of the unit doesn't matter, so it is not shown.

Chapter 3

Area and angle

3.1 What is area?

We talked about rectangles in Section 2.1 and used them again in 2.2 in setting up the rectangular coordinates (x, y) of a point in a plane. One thing we all know about a rectangle is that it has an **area**: for example, if we are laying tiles to cover a rectangular shape as in Fig.7, we want to know how many will be needed – and this number measures the area. If our tiles are 20 cm square and we are covering a floor 3 m in one direction (the x direction, say) and 2 m in the y direction, then we shall need 3×5 tiles in each row and there will be 2×5 such rows; so we shall need 15×10 tiles and the area will be 150 units, the unit being '1 tile'. This is clear from Fig.7(a).

If the lengths of the two sides are instead L_1 m and L_2 m we shall need $L_1 \times L_2 \times 25$ tiles where L_1 and L_2 are numbers which *measure* the two lengths in metres. If we were using 'bigtiles', each being square with sides of length 1 m, then 1 bigtile would cover exactly the same area as 25 ordinary tiles; they would be equivalent in area and we could say this in the equation

$$1 \text{ bigtile} = 25 \text{ tiles}, \quad \text{or } 1 \text{ new unit} = 25 \text{ old units}.$$

Now we know already, from Section 1.1, that the measure of a quantity depends on the unit we are using: if we take a new unit k times as big as the old unit, then the number which measures the quantity will become k times smaller. So in this example the area of

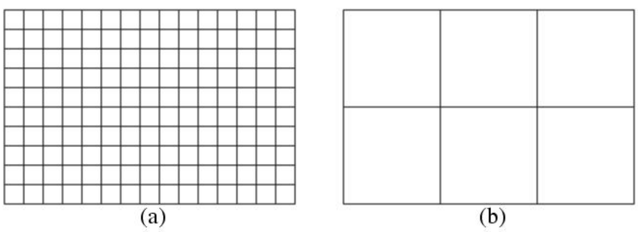

Figure 7

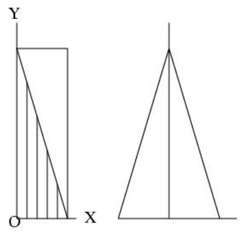

Figure 8

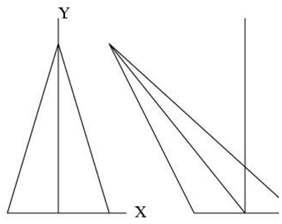

Figure 9

the room will be A = 150 tiles = (150/25) bigtiles, the 6 bigtiles corresponding to the area in 'square metres' of the 3 m × 2 m rectangle. This is shown in Fig.7(b): 6 bigtiles just fit.

With the metre as the standard unit of length we see that the unit of *area* is 1 m^2. So if the unit of *length* is multiplied by k, the measure of length will be divided by k; but the unit of *area* will be multiplied by k^2 and the measure of area will be divided by k^2. We usually say that area has the "**dimensions** of length squared" or, in symbols, [area] = L^2 (read as "the dimensions of area are el squared"). When we use symbols to stand for *quantities* we must always be careful to get the units right as soon as we use numbers to measure them!

The rectangle is a particular 'shape' with certain properties, like its area and the length of a side (i.e. the distance between two neighbouring corners). If we move it to another position in space, such properties do not change – they belong to the object. An important thing about the metric axiom (2.2) is that it means all distances will be left unchanged, or **invariant**, when we move an object without bending it or cutting it – an operation which is called a **transformation**. From this fact we can find the areas of other shapes. Two are specially important; the triangle △, which has only three sides, and the circle ◯, which has one continuous side (called its **perimeter**) at a fixed distance from its centre.

The area of a triangle follows easily from that of a rectangle: for a diagonal line divides the rectangle into two halves, each with the same area because each could be transformed into the other (as in Fig.8) without change of shape. To do this, think of the y-axis as a 'hinge' and turn the shaded half of the rectangle over (like a door); and then put the two halves together again, by sliding them in the plane until you get the 'equilateral triangle' (two sides equal). The base of the triangle is twice the bottom side of the rectangle; and its vertical height is the same as that of the rectangle. But the area of the triangle is still that of the original rectangle. So we get the simple formula

$$\text{Area of triangle} = \tfrac{1}{2}\ \text{base} \times \text{vertical height}. \qquad (3.1)$$

An interesting thing about this result is that it still holds good even when the top, or **vertex**, of the triangle is pushed over to one side as in Fig.9. This must be so because if you imagine each horizontal strip to be filled with tiny squares (elements of area), slide each one sideways, and then count the elements in all strips, the total number cannot have changed. So both the triangles shown in Fig.9 will have the same area, given by (3.1).

The area of a circle is not quite so easy to find, but the problem was solved by Archimedes

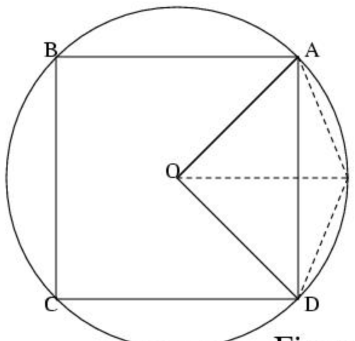

Figure 10

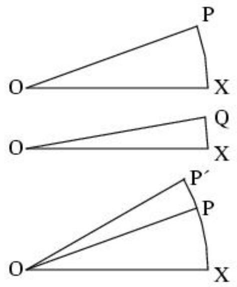

Figure 11

(another of the ancient Greeks), who used a very clever method. He noted that a circle could be filled, nearly but not quite, by putting inside it a shape (called a **polygon**) with N sides, as in Fig.10 for N=4, and that each side formed the base of a triangle with its vertex at the centre. Then, by making N bigger and bigger, he could find polygons whose areas would come closer and closer to the area of the circle.

For a circle of unit radius, $r = 1$, the first (very rough) approximation was the area of the square, with $A_4 = 4 \times (\frac{1}{2}r^2) = 2$, as follows from Fig.10. But Archimedes was then able to show that the polygon with $2N$ sides, instead of N, had an area A_{2N} given by the formula

$$A_{2N} = \frac{N}{2}\sqrt{2 - 2\sqrt{1 - \left(\frac{2A_N}{N}\right)^2}}. \qquad (3.2)$$

Using this formula (and given that $\sqrt{2} \approx 1.414214$) you can easily get the area A_8 of the 8-sided polygon (shown, in part, by the broken lines in Fig.10) in terms of A_4: it will be

$$A_8 = 2\sqrt{(2 - 2\sqrt{(1 - 1^2)})} = 2\sqrt{2} \approx 2.828427.$$

– compared with the first approximation $A_4 = 2$.

If you go on (you'll need a calculator!) you will find $A_{16} \approx 3.061468$ and if you go on long enough you will find something very close to 3.141593. This is a good approximation to the number, always denoted by the Greek letter π ('pi'), which is the **limit** of a series (Book 1, Section 5.1): it is the area of a circle of radius $r = 1$. If you want to go to a circle whose radius is measured by r instead of 1, it's enough to remember that $[A]=L^2$ – so that when a length is multiplied by r the area will be multiplied by r^2. This gives us the important formula

$$\text{Area of a circle of radius } r = \pi r^2 \qquad (\pi \approx 3.141593), \qquad (3.3)$$

which we'll need right away in defining **angle**.

3.2 How to measure angles

How can we measure the 'angle' between two intersecting straight lines when they are neither perpendicular nor parallel – when they simply 'point in different directions'. The *slope* m of a line is one such number, for it fixes the direction of the line AB in Fig.5 relative to AC, which is parallel to the x-axis. We say that AB 'makes an angle' with AC and call $m(= BC/AC)$ the **tangent** of the angle. This ratio is obtained easily for any pair of lines by dropping a perpendicular from a point on one of them to the other; and it also follows easily that it does not matter which line is taken first. Two other ratios, BC/AB and AC/AB, also give a simple arithmetic measure of the same angle: they are called, respectively, the **sine** and the **cosine** of the angle. There is, however, a *single* number which gives a more convenient measure of the angle – 'circular measure', since it relates directly to the circle. To get this we must think about *combining* angles.

Just as two *points* define a *linear displacement* which brings the first into coincidence with the second; two *straight lines*, with one point in common, define an *angular displacement*, or a *rotation*, which brings the first into coincidence with the second. The rotation angle is given a *sign*, positive (for anti-clockwise rotation) or negative (for clockwise) – for rotations in the two opposite senses are clearly different. Just as two linear displacements are called equal if their initial and final *points* can be put in coincidence (by sliding them about in the plane), we call two angular displacements equal if their initial and final *lines* can be brought into coincidence. And just as two linear displacements can be combined by making the end point of one the starting point of the next, we can combine two angular displacements by making the end line of one the starting line of the next. These ideas will be clear on looking at Fig.11. Angles are named by giving three letters: the first is the end point of the initial line; the last is the end point after rotation; and the middle letter is the point that stays fixed. The sum of the angles XOP and XOQ is the angle XOP′, obtained by taking OP as the starting line for the second angle, POP′, which is made equal to XOQ.

After saying what we mean by 'combination' and 'equality' of angles, we look for an 'identity' (in the algebra of rotations) and the 'inverse' of any angle, ideas which are old friends from Book 1. The 'identity' is now "don't do anything at all (or rotate the initial line through an angle zero)"; and the 'inverse' of an angle is obtained simply by changing the *sense* of the rotation – clockwise rotation followed by anti-clockwise rotation of the same amount is equal to no rotation at all! If we write R for a positive rotation and R^{-1} for its inverse (negative rotation). this means

$$\mathsf{R}\mathsf{R}^{-1} = \mathsf{R}^{-1}\mathsf{R} = \mathsf{I}.$$

Next we must agree on how to *measure* angles. There is a 'practical' method, which starts from the fact that rotation of the line OP through a complete circle around the point O, let's call it 1 'turn', is the same as doing nothing. The 'degree' is a small angle, such that 360 degrees = 1 turn; and the angle between two lines in a plane can therefore be measured by a number (of degrees) lying between 0 and 360. Angles, unlike distances (which can be as big as you like), are thus *bounded* – since we can't tell the difference between angles that differ only by 360 degrees (or any multiple of 360). This doesn't mean that angular *displacement* is bounded: we all know that, in turning a screw, for example, every turn (rotation through 360 degrees) is important; and it can be repeated again and again to get bigger and bigger rotation angles. It is only the angle between two lines in a plane that is bounded: in the case of a screw, rotation has an effect *outside* the plane and it is then useful to talk about rotations through angles greater than 360 degrees.

A more fundamental way of measuring angles follows from the equation (3.3) for the area of a circle. If we use θ to denote the angle XOP in Fig.11 (angles are usually named using Greek letters and θ is called 'theta'), then the 'circular measure' of θ is the ratio of two lengths: θ = arc/radius, where 'arc' is the length of the part of the circle between point P and the x axis. This is a pure number, not depending on the unit of length, and gets as close as we wish to $\tan\theta$ and $\sin\theta$ as the angle becomes smaller and smaller. To find this number we write (3.3) in another form. The area of the whole circle (A) is the sum of the areas of a huge number of tiny triangles, each one with a small area $a \approx \frac{1}{2}\text{arc} \times \text{radius}$: so we find $A = \frac{1}{2}(\text{whole arc}) \times \text{radius}$ where 'whole arc' means the sum of all the tiny arcs, one for each triangle, as we go round the perimeter of the whole circle. The length of this arc is the **circumference** of the circle. So what we have shown is that $A = \frac{1}{2}(\text{Circumference} \times \text{radius}$, and from (3.3) this gives the final result

$$\text{Circumference of a circle} = 2 \times \text{Area} \div \text{radius} = 2\pi r. \tag{3.4}$$

Since the circumference is the 'whole arc', which is $\Theta \times r$ (where Θ denotes the whole angle turned through in going all round the circle), we can write $\Theta = 2\pi$ radians. Here, the **radian** is the 'natural unit' of angle and since, in terms of 'degrees' 2π radians = 360 degrees, it follows from (3.3), that

$$1 \text{ radian} \approx 57.3 \text{ degrees}. \tag{3.5}$$

Usually, however, it is better to use radian measure: for example, two lines are perpendicular when the angle between them is $\pi/2$ and this does not depend on defining the 'degree'.

More on Euclid

Most of Euclid's work was on plane figures (shapes such as triangles and rectangles that lie in a plane). There's so much of it that it would fill a whole book, so we just give one or two definitions and key results to start things off:

- Two angles like A and B in Fig.12(a), whose sum is π, are called *complementary*; each is the *complement* of the other – together they complete the angle π. When the angles describe rotations of the arrow, about the fixed point O, the rotation A followed by B is the rotation $A + B = \pi$, which turns the arrow round and makes it point the other way.

- When two straight lines cross, as in Fig.12(b), they make two pairs of complementary angles A, B and A', B'. If we make a half-turn of the whole picture, around the crossing point, A goes into A' and B into B', but the angle A is *unchanged* by the operation: so $A' = A$ and similarly $B' = B$ – *'opposite' angles are equal.* So when

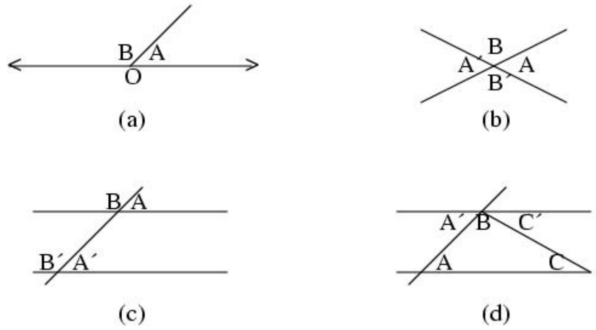

Figure 12

two lines cross they make two pairs of equal angles; and the different angles (A and B) are *complementary*.

- When a straight line crosses *two parallel* lines, as in Fig.12(c), it makes two other pairs of equal angles $A' = A$ and $B' = B$; for sliding the picture so as to send A into A' and B into B' is another transfomation (see Section 3.1) that does not change the angles. Such pairs of angles are called 'alternate'.

- By adding another straight line to the last picture (Fig.12(c)) we make a triangle (Fig.12(d)) with three *'internal' angles*, here called A, B, C. Now, from the last two results, A' (being opposite to the angle alternate to A) is equal to A and similarly $C' = C$. Also the sum of $A'(= A), B$, and $C'(= C)$ is the angle π in Fig.12(a). It follows that *the sum of the angles inside any plane triangle is* π *radians* (i.e. 180 degrees or two right angles).

Euclid and his school proved a great number of other results of this kind, each one following from those already obtained. All these theorems were numbered and collected and can still be found in any textbook of geometry.

Note: The next Chapter contains difficult things, usually done only at university, but also much that you will understand. Look at it just to see how many different ideas come together. Then come back to it when you're ready – perhaps a year from now!

Exercises

1) Look at Figs.9,10 and then calculate the area of the 8-sided polygon, part of which is shown by the broken lines in Fig.10. Check that your result agrees with equation (3.2) when you put $N = 4$. (The polygon holds $2N$ triangles, all with the same area. Find the base and the vertical height of each of them, taking the circle to have unit radius.)

2) Express all the angles in Fig.10 both in degrees and in radians.

Chapter 4

Rotations: bits and pieces

One of the great things about mathematics is that it contains so many 'bits and pieces' which, again and again, can be put together like bricks, in building up new ideas and theories. These small pieces are so useful that, once understood, they are never forgotten. In talking about angles and rotations we need to use vectors (Book 1, Section 3.2); the laws of indices (Book 1, Section 4.2); the exponential series (Book 1, Section 5.1); complex numbers (Book 1, Section 5.2); and the idea of rotation as an **operator** (as in Book 1, Section 6.1).

Let's start with a **vector** pointing from the origin O to any point P, as in Fig.13. In a rotation around O, any such vector is turned through some angle, let's call it θ, and, as in Book 1, Section 6.1, we can think of this operation as the result of applying a rotation **operator** R_θ. There is a **law of combination** for two such operators:

$$\mathsf{R}_{\theta'}\mathsf{R}_\theta = \mathsf{R}_{\theta+\theta'},$$

(don't forget we agreed in Section 6.1 that the one on the right acts first) and for every operator R_θ there is an **inverse** operator, denoted by R_θ^{-1}, with the property

$$\mathsf{R}_\theta\mathsf{R}_{-\theta} = \mathsf{R}_{-\theta}\mathsf{R}_\theta = \mathsf{I},$$

where I is the **Identity** operator (rotation through angle zero). These properties define a **group** (Book 1, Section 6.1) with an infinite number of elements – since θ can take any value between 0 and 2π (rotation through $\theta+2\pi$ not being counted as different from R_θ). We now want to put all this into symbols.

In 2-space any point P is found from its coordinates (x, y): to get there, starting from the origin (where $x = 0, y = 0$), you take x steps in the 'x-direction' (i.e. parallel to the x-axis) and y steps in the 'y-direction'. In Book 1, Section 3.2 there was only one axis and e was used to mean 1 step along that axis; but now there are *two* kinds of step (e_1 and e_2, say), so we write, for the vector describing the displacement from O to P,

$$\mathsf{r} = x\mathsf{e}_1 + y\mathsf{e}_2, \tag{4.1}$$

where e_1, e_2 are along the two directions and r is called the 'position vector' of P. From Book 1, Section 3.2, it's clear that the order in which the steps are made doesn't matter:

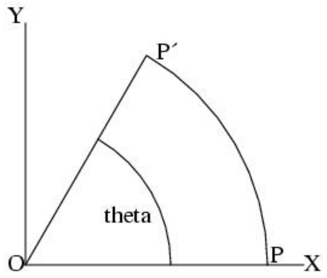

Figure 13

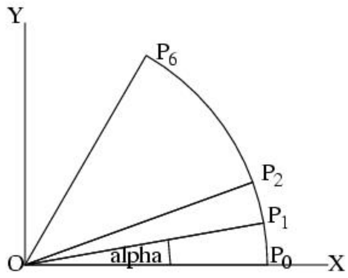

Figure 14

if $x = 2$ and $y = 3$ then $\mathsf{r} = \mathsf{e}_1 + \mathsf{e}_1 + \mathsf{e}_2 + \mathsf{e}_2 + \mathsf{e}_2$ or, just as well, $\mathsf{r} = \mathsf{e}_2 + \mathsf{e}_1 + \mathsf{e}_2 + \mathsf{e}_2 + \mathsf{e}_1$ – because you arrive at the same point in the end. The distance from O to P is the *length* of OP, or the *magnitude* of the vector r, and the coordinates x, y may be whole numbers or fractions, positive or negative, or even irrational, as we know from Book 1, Section 4.3. Now let's think about *rotating* a vector, turning it through an angle. A rotation of OP (Fig.13) through an angle θ around the origin can be described in symbols as

$$\mathsf{r} \rightarrow \mathsf{r}' = \mathsf{R}_\theta \mathsf{r}, \tag{4.2}$$

where $\rightarrow$ means "goes to" and r' is the position vector of point P′, after OP has been sent into OP′. The 'product' of two rotations, R_1 followed by R_2 through angles θ_1 and θ_2, respectively, is written

$$\mathsf{r} \rightarrow \mathsf{r}' = \mathsf{R}_2\mathsf{R}_1\mathsf{r} = \mathsf{R}_3\mathsf{r} \quad (\theta_3 = \theta_1 + \theta_2). \tag{4.3}$$

The fact that the product of two rotations is obtained by *adding* their rotation *angles*, reminds us of the laws of indices – where $a^m \times a^n = a^{m+n}$ – a result which is true even when the indices m, n are not only whole numbers. Let's now look for a connection.

In Book 1, Section 5.1 we met a number defined as the **limit** of a **series** (remember the shorthand used in Book 1, that $2! = 1 \times 2$, $3! = 1 \times 2 \times 3$, and so on, $n!$ being called "factorial n")

$$y = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \ = f(x), \tag{4.4}$$

when the number of terms becomes infinite. This number depends on the value we give to x and is denoted here by $f(x)$ (read as a "function of x" – or, in short, "eff of ex") to mean only that for every value of x we can find a related value of y: x is called the "independent variable" (we can give it any value we like), but y is the "dependent variable" whose value depends on that of x. The branch of mathematics that deals with functions is called Analysis, and we'll say more about it in other Books of the Seies. Here it's enough to think of a function as a *rule* – in this case the series (4.4) – by which we can calculate a value of y, given the value of x.

The function defined in (4.4) has amazing properties. Let's multiply two such series together, using two different values of x (call them $x = p$ in one series and $x = q$ in the other):

$$\begin{aligned} f(p)f(q) &= \left(1+p+\frac{p^2}{2!}+...\right)\left(1+q+\frac{q^2}{2!}+...\right) \\ &= 1+(p+q)+\left(\frac{p^2}{2!}+pq+\frac{q^2}{2!}\right)+... \\ &= 1+(p+q)+\frac{(p+q)^2}{2!}+...\ , \end{aligned} \tag{4.5}$$

– including terms only up to the 'second degree' (i.e. those with not more than *two* variables multiplied together). The result seems to be just the same function (4.4), but with the new variable $x = p + q$. And if you go on, always putting together products of the same degree, you'll find the next terms are

$$(p+q)^3/3! = (p^3 + 3p^2q + 3pq^2 + q^3)/3! \qquad \text{(third degree)}$$

and

$$(p+q)^4/4! = (p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4)/4! \qquad \text{(degree 4.)}$$

As you can guess, if we take more terms we're going to get the result

$$f(p)f(q) = 1+(p+q)+\frac{(p+q)^2}{2!}+\frac{(p+q)^3}{3!}+...\ = f(p+q). \tag{4.6}$$

To get a *proof* of this result is much harder: you have to look at all possible ways of getting products of the nth degree (n factors at a time) and then show that what you get can be put together in the form $(p+q)^n/n!$. So we'll just accept (4.6) as a basic property of the **exponential function**, defined in (4.4) and often written as "$\exp x$".

From (4.6) we find, by putting $p = q = x$, that $f(x)^2 = f(2x)$; and on doing the same again $f(x)^3 = f(x) \times f(2x) = f(3x)$. In fact

$$f(x)^n = f(nx). \tag{4.7}$$

This second basic property lets us *define* the nth power of a number even when n is *not an integer*; it depends only on the series (2.14) and holds good when n is any kind of number (irrational or even complex). Even more amazing, both (4.6) and (4.7) are true whatever the symbols (x, p, q) may stand for – as long as they satisfy the usual laws of combination, including $qp = pq$ (so that products can be re-arranged, as in getting the result (4.6)).

In Book 1, Section 1.7, the (irrational) number obtained from (4.4) with $x = 1$ was denoted by e:

$$e = 1+1+\frac{1}{2}+\frac{1}{6}+\frac{1}{24}+\ \dots \quad = 2.718281828\dots \tag{4.8}$$

and this gives us a 'natural' *base* for defining all real numbers. From (4.7), $e^n = f(n)$ is true for any n – not just for whole numbers but for *any* number. So changing n to x gives

$$e^x = 1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots\ , \tag{4.9}$$

and the 'laws of indices' can now be written in general form as

$$e^x e^y = e^{x+y}, \qquad (e^x)^y = e^{xy}. \tag{4.10}$$

We're now ready to go back to rotations in space! We know that rotations are combined according to (4.3) and that every rotation R_θ is labelled by its rotation angle θ, which is just a number. For some special values of θ, we also know what R_θ does to a vector in 2-space. For example, $\mathsf{R}_{2\pi}\mathsf{r} = \mathsf{r}$, but $\mathsf{R}_\pi\mathsf{r} = -\mathsf{r}$ because rotating a vector through *half* a turn makes it point in the opposite direction, which means giving it a negative sign. But how can we describe a general rotation?

Any rotation can be made in small steps, for example in steps of 1 degree at a time, so let's think of R_θ as the result of making n very small rotations through an angle α: so $\theta = n\alpha$ and what we mean is that $\mathsf{R}_\theta = (\mathsf{R}_\alpha)^n$, where we use the 'power' notation to mean the product $\mathsf{R}_\alpha\mathsf{R}_\alpha ... \mathsf{R}_\alpha$ with n factors. So n becomes a measure of the rotation angle θ in units of α; and if R_θ is followed by a rotation $\mathsf{R}_{\theta'}$, with $\theta' = m\alpha$, the result will be a rotation through $(m+n)\alpha$. Fig.14 gives a picture of the rotations which carry the position vector of a point P_0, on the x-axis, into P_1 (1 step), P_2 (2 steps), and so on – each step being through a very small angle α (magnified here, so you can see it).

The rotation R_α sends the point P_0, with position vector $\mathsf{r} = r\mathsf{e}_1 + 0\mathsf{e}_2$ (the components being $x = r$ and $y = 0$ when r points along the x-axis), into P_1 with $\mathsf{r}' = \mathsf{R}_\alpha = x'\mathsf{e}_1 + y'\mathsf{e}_2$. In general, the x- and y-components of any rotated vector (call them x, y for any rotation angle θ) are related to the *sine* and *cosine* of the angle turned through – as we know from earlier in this Section. The definitions are $\cos\alpha = x/r$ and $\sin\alpha = y/r$ and the rotation leading to P_1, with coordinates (x_1, y_1), thus gives

$$\mathsf{r}_1 = \mathsf{R}_\alpha\mathsf{r} = x_1\mathsf{e}_1 + y_1\mathsf{e}_2 = r\cos\alpha\ \mathsf{e}_1 + r\sin\alpha\ \mathsf{e}_2. \tag{4.11}$$

After repeating the operation n times we reach the vector ending on P_n: in short

$$\mathsf{r}_n = (\mathsf{R}_\alpha)^n\mathsf{r} = x_n\mathsf{e}_1 + y_n\mathsf{e}_2 = r\cos(n\alpha)\mathsf{e}_1 + r\sin(n\alpha)\mathsf{e}_2. \tag{4.12}$$

Of course, we know how to get the sine and cosine from the picture (by measuring the sides of a triangle) and we know their values for certain special angles like $\theta = 2\pi$, or π, or $\pi/2$, or even $\pi/4$; but what we really need is a way of calculating them for any angle θ $(= n\alpha)$.

To do this we start from the series (4.9), remembering that (4.10) gives us a way of finding its nth power just by writing nx in place of x (writing $y = n$ because it stands for *any* number). And since x is also *any* number let's experiment – putting $x = i\alpha$, where i is the 'imaginary unit' first introduced in Book 1, Section 5.2. The result is

$$e^{i\alpha} = 1 + i\alpha - \frac{\alpha^2}{2!} - i\frac{\alpha^3}{3!} + ... \tag{4.13}$$

where we're using the fact that $i^2 = -1$, $i^3 = i \times i^2 = -i$, and so on. On collecting together the real terms (no i factors) we discover a new series:

$$C_\alpha = 1 - \frac{\alpha^2}{2!} + \frac{\alpha^4}{4!} - ... \tag{4.14}$$

and, on doing the same with the imaginary terms, find another series

$$S_\alpha = \alpha - \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!} - \dots . \qquad (4.15)$$

Putting the two series together shows that

$$e^{i\alpha} = C_\alpha + iS_\alpha \qquad (4.16)$$

and from (4.10) there's a similar result when α is replaced by the large angle $n\alpha$; so

$$e^{in\alpha} = C_{n\alpha} + iS_{n\alpha}, \qquad (4.17)$$

where $C_{n\alpha}$ and $S_{n\alpha}$ are just like (4.14) and (4.15), but with $n\alpha$ instead of α.

Now look back at where we started: Equation (4.11) gives us the coordinates of P_1 after rotating OP_0 through a very small angle α and the (geometrically defined) values of $\sin\alpha$ and $\cos\alpha$ are, neglecting powers beyond α^2, $\sin\alpha \approx \alpha$ and $\cos\alpha \approx 1 - \frac{1}{2}\alpha^2$ – and these are the leading terms in the series (4.14) and (4.15)! For small angles, $C_\alpha \to \cos\alpha, \quad S_\alpha \to \sin\alpha$. From these starting values we continue by (i) making more rotations, in steps of α, getting (4.12) after n steps; and (ii) multiplying $e^{i\alpha}$ by the same factor, in every step, to get $e^{in\alpha}$ after n steps. The two things go hand in hand. We take a bold step and say that

$$\cos(n\alpha) = C_{n\alpha}, \qquad \sin(n\alpha) = S_{n\alpha}, \qquad (4.18)$$

are the algebraic expressions for the cosine and sine of *any* angle $n\alpha$.

So we write, for any angle θ, the general results

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \dots , \quad \sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \dots . \qquad (4.19)$$

And, from (4.17) with $n\alpha = \theta$,

$$e^{i\theta} = \exp i\theta = \cos(\theta) + i\sin(\theta) \qquad (4.20)$$

The above results lead to many others. Take an example: for any θ, we may square both sides of equation (4.20) to obtain

$$e^{2i\theta} = (\cos\theta + i\sin\theta)^2 = (\cos\theta)^2 - (\sin\theta)^2 + 2i\sin\theta\cos\theta.$$

But we also know that

$$e^{2i\theta} = \cos 2\theta + i\sin 2\theta$$

and (from Book 1, Section 5.2) that two complex numbers are equal only when their real and imaginary parts are separately equal; so comparing the last two equations shows that

$$\cos(2\theta) = (\cos\theta)^2 - (\sin\theta)^2, \quad \sin(2\theta) = 2\sin\theta\ \cos\theta \qquad (4.21)$$

– knowing the sine and cosine of any angle you can get them very easily for twice the angle. For example, we know that $\sin(\pi/4) = \cos(\pi/4) = \frac{1}{2}\sqrt{2}$ (from the right-angled triangle with sides of length 1, 1, $\sqrt{2}$); so doubling the angle gives $\sin(\pi/2) = 1$, $\cos(\pi/2) = 0$; doubling it again gives $\sin(\pi) = 0$, $\cos(\pi) = -1$; and yet again gives $\sin(2\pi) = 1$, $\cos(2\pi) = 0$. The last result shows that the angle 2π (or 360 degrees) looks no different from zero; and that every rotation through 2π gives us nothing new – the dependence of sine and cosine on the angle is said to be **periodic**, they take the same values whenever the angle increases by 2π, called the **period**. In other words

$$e^{2\pi i} = 1 \qquad (4.22)$$

– a connection between two irrational numbers (e, π) and the imaginary unit (i), almost beyond belief! This is one of the most remarkable results in the whole of Mathematics.

The sine and cosine of the *sum* of any two angles follows in the same way as for twice the angle. Taking

$$\exp i(\theta_1 + \theta_2) = \exp i\theta_1 \times \exp i\theta_2,$$

using (4.20) and expanding the right-hand side, we find (try it yourself!)

$$\begin{aligned} \cos(\theta_1 + \theta_2) &= \cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2, \\ \sin(\theta_1 + \theta_2) &= \sin\theta_1 \cos\theta_2 + \cos\theta_1 \sin\theta_2. \end{aligned} \qquad (4.23)$$

That's all you need to know about angles – the rest you can do for yourself! A long time ago, in school, when all of geometry was done the way Euclid did it, we had to learn all these results (and many more) by heart – chanting them over and over again – and all because the Pythagoreans threw away their great discovery of algebraic geometry, leaving it for the French mathematician René Descartes (1596-1650) to re-discover more than a thousand years later! Now you can get such results whenever you need them, remembering only the laws of indices and doing some simple algebra.

Exercises

1) Get the results labelled "(third degree)" and "(fourth degree)", just after equation (4.5), by multiplying together the results you already know.

2) Obtain the results (4.13) to (4.20) by starting from (4.9) and working through all the details.

3) Starting from (4.23), find expressions for $\cos(\theta_1 - \theta_2)$, $\sin(\theta_1 - \theta_2)$, $\cos 2\theta$, $\sin 2\theta$, $\cos 3\theta$, $\sin 3\theta$.

Chapter 5

Three-dimensional space

5.1 Planes and boxes in 3-space – coordinates

As we all know, from birth, the real 'physical' space we live in is *not* a 2-space, or plane, in which a point is specified by giving two numbers to define its position. There are points 'above' and 'below' any plane; and to define their positions we'll need a third number – to tell us how far up or how far down. Again, as in Section 1.2, we'll refer a point to a set of perpendicular axes, meeting at a point O – the origin – but now there will be three axes instead of two. Up to now, we've been talking about *plane* geometry; but now we turn to 3-space and to *solid* geometry. The basic ideas, however, are not much different: we start from an axiom, just like that we used in 2-space, referring to the shortest *distance* between points; then we set up a few theorems from which all of solid geometry can be derived by purely *algebraic* reasoning. Of course, we won't do all of it – just enough to make us feel sure that it *can* be done.

According to the first Axiom (Section 1.2) a straight line is the unique shortest path between two points. And from the definition of a plane (Section 1.2) it follows that if two planes intersect, then they cut each other *in a straight line* – for if any two points A and B are common to both planes then there is a unique straight line AB and all the points on AB lie at the same time in both planes (i.e. AB, which may be as long as we wish, is the line in which the planes intersect).

From this conclusion we can go to a first theorem:

> *Theorem.* If a straight line is perpendicular to *two* others, which it meets at a common point, then it is perpendicular to *all* others in the same plane and passing through the same point. It is then *perpendicular to the plane.*

The proof follows from Fig.15, where OP is taken perpendicular to both OA and OB and the angle OAB is taken to be a right angle. Let OC be any other straight line, through O, in the plane OAB. We must prove that COP is also a right angle.

This will be so only when $PC^2 = OP^2 + OC^2$ and this follows in two steps: First,

$$PB^2 = OP^2 + OB^2 = OP^2 + OA^2 + AB^2 = AP^2 + AB^2,$$

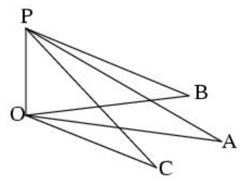

Figure 15

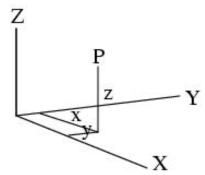

Figure 16

and therefore PAB is also a right angle. Then, second, we have

$$PC^2 = PA^2 + AC^2 = OA^2 + OP^2 + AC^2 = OC^2 + OP^2.$$

This proves the theorem.

Two other simple results follow:

- The perpendicular from a point to a plane is the shortest path from the point to any point in the plane.
- If a straight line is perpendicular to two others, which meet it at some point, then the two others lie in a plane.

These are 'corollaries' to the theorem, the second one being the **converse** of the theorem – saying it the other way round.

Cartesian coordinates in 3-space

We're now ready to set up the (rectangular) Cartesian coordinates of any point P in 3-space. First we take a plane OXY and the given point P, outside the plane as in Fig.16. If Q is the foot of the perpendicular from P onto the plane, then PQ is the unique shortest path from the point to the plane; let's call its length z. And point Q, lying in the plane, is uniquely defined (see Section 2.2) by giving its 2-space coordinates (x and y) relative to the axes OX and OY. The position of P is then completely defined by giving the *three* numbers (x, y, z), as in Fig.16. In the case of z, however, we must give the number a sign ($\pm$) to show whether P is *above* the plane or *below*: we agree that z will be counted positive (it will be on the 'positive side' of the plane) when a rotation carrying OX into OY would move a 'right-handed screw' (with its sharp end underneath point O) upwards, towards P.

Now the three axes OX, OY, and QP have not been freely chosen, for the third one must pass through the point P. We'd like to be able to talk about *all* points in space, not only those on one special line QP; we want one set of three perpendicular axes (OX, OY, OZ), all starting from a common origin (O), which can be used to describe *all* points. To do this, we need one more theorem

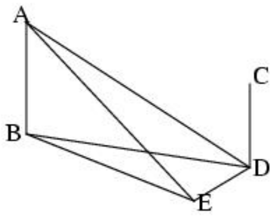

Figure 17

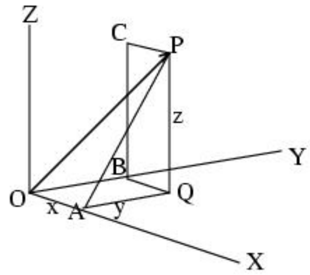

Figure 18

Theorem. Two straight lines, both perpendicular to a given plane, are parallel to each other.

The proof follows from Fig.17, where the two lines BA and DC are taken perpendicular to the plane BDE; and E is chosen so that DE is perpendicular to DB (i.e. BDE is a right angle).

We first show that EDA is also a right angle; and that CD, DA and DB must therefore lie in the same plane (by the previous theorem). This follows at once because $AE^2 = AB^2 + BE^2 = AB^2 + BD^2 + DE^2 = AD^2 + DE^2$, and so EDA is a right angle and the lines DB, DA, BA, and DC all lie in the same plane. Moreover, BA and DC, besides lying in the same plane, are perpendicular to the plane BDE; so they are perpendicular to the line BD which intersects them. Thus, by the *Definition* at the beginning of Section 2.1, BA and DC are *parallel* – proving the theorem.

Again, the theorem has a converse:

Converse. If two straight lines are parallel and one is perpendicular to a plane, then so is the other.

A whole chain of results follows from the theorem and its converse. We'll just say what they are when we need them (no proofs!), starting with a definition:

Definition. If two planes are perpendicular to the same straight line, then they are **parallel planes**.

It then follows that a perpendicular from any point on one plane, connecting it with a point on the other, will have the same length no matter what point we choose – this being the *shortest distance* between the two planes. If two pairs of points, A, B, and C, D, are connected in this way, then they lie at the corners of a *rectangle*, whose opposite sides have equal length.

5.2 Describing simple objects in 3-space

We can now go ahead exactly as we did in 2-space. But now we take any point O as **origin** and draw **rectangular axes** OX, OY, and OZ, as in Fig.18, each being perpendicular to the others. Any point P, anywhere in 3-space, can then be given rectangular (Cartesian) coordinates, x, y, z, which measure the shortest distances to the planes OYZ, OZX, and OXY, respectively. These distances are also the lengths of the **projections** of the line OP, shown in the Figure, along the three axes, OX, OY, OZ: the projection shown, OA, is the line from the origin O to the foot (A) of the perpendicular from P to the x-axis and the lengths of OA and QB are equal – being opposite sides of OAQB, which is a rectangle (as follows from the Theorems above, both lines being perpendicular to the plane OYZ).

The geometry of 2-space, in Section 2.2, was based on equation (2.1), which gave us the distance between any two points, P and P′; and on (2.2), which holds when they are close together. In 3-space, things look just the same, except that there are now three coordinates: the distance (r, say) from the origin O to any point P is given by

$$r^2 = x^2 + y^2 + z^2 \tag{5.1}$$

while for two infinitely close points the separation (dr) follows from

$$\mathrm{d}r^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \tag{5.2}$$

– $\mathrm{d}x, \mathrm{d}y, \mathrm{d}z$ being the **differentials**, such that a neighbouring point P′ has coordinates $x' = x + \mathrm{d}x, y' = y + \mathrm{d}y, z' = z + \mathrm{d}z$

Again (5.2) is the 'fundamental metric form' – but now in real three-dimensional space – and because it has sum-of-squares form at any point (and, according to (5.1), in any region, however large) the space is Euclidean, with all the properties first discovered by Euclid. Any *plane* is called a 2-dimensional **subspace** of 3-space and any straight line is a 1-dimensional subspace. Just as plane geometry, in the algebraic approach followed in Section 2.2, comes out of equations (2.1) and (2.2), the whole of *solid geometry* comes out of (5.1) and (5.2).

Again, in 3-space, the simplest geometrical object is a straight line; but now every point on the line will have three coordinates. In 2-space the coordinates x, y of a point on a straight line were *related* so the $y = mx+c$, where the numbers m and c fix the slope of the line and where it crosses the y-axis; we took x as the 'independent variable', which then determines y (the 'dependent variable'). But in 3-space things are a bit more complicated as the line doesn't have to lie in any one of the coordinate planes – it can point any way we please. The same is true for the next simplest object, the plane, which may have any **orientation** we please. We'll look at these things again in the next Section, after we've found a simpler way of dealing with them – namely, 'vector algebra'. But for the moment it's enough to note that lines and planes are described by *linear* equations, involving only *first* powers of the variables x, y, z, while circles (for example) require equations involving higher powers or products. The simplest examples are the coordinate planes themselves: thus $z = 0$ (constant) describes the plane containing the axes OX and OY, and similarly $z = p$ (constant) defines a plane parallel to OXY and at a perpendicular distance p from the origin. In both cases any point in the plane is determined by giving values, any we wish, of the other variables x, y.

The simplest solid object (after the **cube**, which has six plane faces) is the **sphere**, corresponding to the circle in 2-space. It has a single *curved* surface and the coordinates of any point on the surface are related by an equation of the *second* degree. The distance of a point $\mathrm{P}(x, y, z)$ from the origin is given by

$$r^2 = x^2 + y^2 + z^2 \tag{5.3}$$

and this distance (r) is the radius of the sphere, the same for all points on the surface. Thus (5.3) is the equation for the surface of a sphere centred on the origin. If you move the sphere (or the line or the plane) the equation will be more complicated. This is because our descriptions are based on choosing a set of axes that meet at the centre of the sphere and then using three distances (coordinates) to define every point; the set of axes is called a **reference frame**. If we decide to change the reference frame, so that the origin is no longer at the centre, then all the coordinates will have to be changed.

On the other hand, the objects we meet in 3-space have certain measurable *properties* (like length and area) which 'belong' to the object and do not depend in any way on how we choose the reference frame: as already noted (Chapter 3) they are **invariants**. We'd like to keep our equations as simple and as close as possible to what we're trying to describe: a line, for example, is a *vector* and could be denoted by a single symbol – instead of a set of numbers that will change whenever we change the reference frame. We'll see how to do this in the next Section.

5.3 Using vectors in 3-space

In ordinary algebraic number theory (Book 1, Chapter 4) we represented numbers by points on a straight line, or with the *displacements* which lead from an origin to these points. The displacements are in fact **vectors** in a 1-space, each being a numerical multiple of a unit 'step' which we called e; and any 1-vector a is written as $\mathsf{a} = a\mathsf{e}$, where a is just a number saying 'how many' steps we take in the direction of e. Of course, if a is an integer, the displacement will lead to a point labelled by that integer; but we know from Book 1 that this picture can be extended to the case where a is any real number and a is the vector leading to its associated point in the pictorial representation. The rules for combining vectors in 1-space are known from Book 1: we get the *sum* of two displacements, a and b, by making them one after the other (the end point of the first being the starting point for the second) and it doesn't matter which way round we take them. Thus

$$\mathsf{a} + \mathsf{b} = \mathsf{b} + \mathsf{a}, \tag{5.4}$$

and if there are three vectors it doesn't matter how we combine them,

$$(\mathsf{a} + \mathsf{b}) + \mathsf{c} = \mathsf{a} + (\mathsf{b} + \mathsf{c}). \tag{5.5}$$

We can also multiply a vector by any real number, as in writing a as a number a of units e: $\mathsf{a} = a\mathsf{e}$. Let's try to do the same things in 3-space. There will now be three different

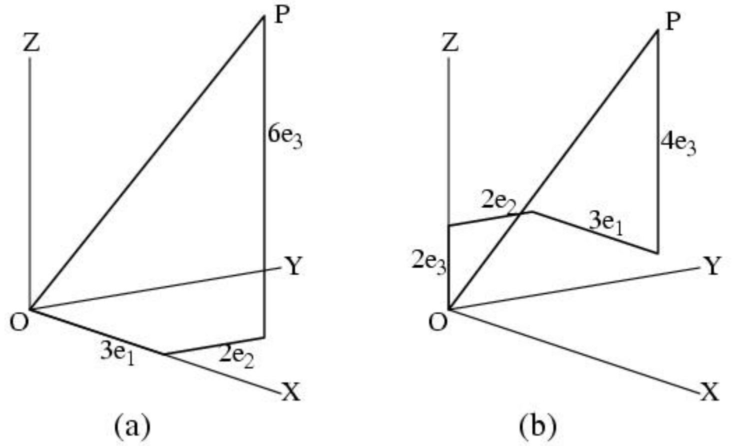

Figure 19

kinds of unit step – along the x-axis, the y-axis, and the z-axis – which we'll call $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$, respectively. They will be the **basis vectors** of our algebra and we take them to be of unit length (being 'unit steps') A vector pointing from the origin O to point P(x, y, z) (i.e. with Cartesian coordinates x, y, z) will be denoted by r and written

$$\mathsf{r} = x\mathsf{e}_1 + y\mathsf{e}_2 + z\mathsf{e}_3. \tag{5.6}$$

This is really just a *rule* for getting from O to P: If the coordinates are integers e.g. x=3, y=2, z=6, this reads "take 3 steps of type e_1, 2 of type e_2 and 6 of type e_3 – and you'll be there!" And the remarkable fact is that, even although the terms in (5.6) are in different directions, the order in which we put them together doesn't make any difference: you can take 2 steps parallel to the z-axis (type e_3), then 2 steps parallel to the y-axis (type e_2), 3 more steps of type e_1, and finally 4 steps of type e_3 – and you'll get to the same point. This is easy to see from Fig.19, remembering that (because the axes are *perpendicular*) space is being 'marked out' in rectangles, whose opposite sides are equal. In fact, the rules (5.4) and (5.5) apply generally for vector addition.

An important thing to note is that in combining the terms in (5.6) the vectors must be allowed to 'float', as long as they stay parallel to the axes: they are called 'free vectors' and are not tied to any special point in space. On the other hand, the **position vector** r is defined as a vector leading from the origin O to a particular point P: it is a 'bound vector'.

The numbers x, y, z in (5.6), besides being coordinates of the point P, are also **components** of its position vector. Any vector may be expressed in a similar form –

$$\begin{aligned} \mathsf{a} &= a_1\mathsf{e}_1 + a_2\mathsf{e}_2 + a_3\mathsf{e}_3 \\ \mathsf{b} &= b_1\mathsf{e}_1 + b_2\mathsf{e}_2 + b_3\mathsf{e}_3, \end{aligned}$$

etc. and addition of vectors leads to addition of corresponding components. Thus, rearranging the terms in the sum,

$$\mathsf{a} + \mathsf{b} = (a_1 + b_1)\mathsf{e}_1 + (a_2 + b_2)\mathsf{e}_2 + (a_3 + b_3)\mathsf{e}_3. \tag{5.7}$$

Similarly, multiplication of a vector by any real number c is expressed in component form by

$$c\mathsf{a} = ca_1\mathsf{e}_1 + ca_2\mathsf{e}_2 + ca_3\mathsf{e}_3. \tag{5.8}$$

Finally, note that the vector algebra of Euclidean 3-space is very similar to the ordinary algebra of real numbers (e.g. Book 1, Chapter 3). There is a 'unit under addition' which can be added to any vector without changing it, namely $\mathsf{0} = 0\mathsf{e}_1 + 0\mathsf{e}_2 + 0\mathsf{e}_3$; and every vector a has an 'inverse under addition', denoted by $-\mathsf{a} = -a_1\mathsf{e}_1 - a_2\mathsf{e}_2 - a_3\mathsf{e}_3$, such that $-\mathsf{a} + \mathsf{a} = \mathsf{0}$.

5.4 Scalar and vector products

From two vectors, a, b, it's useful to define special kinds of 'product', depending on their lengths (a, b) and the angle between them (θ). (The length of a vector a is often written as $a = |\mathsf{a}|$ and called the **modulus** of a.)

> *Definition.* The **scalar product**, written $\mathsf{a}\cdot\mathsf{b}$, is defined by $\mathsf{a}\cdot\mathsf{b} = ab\cos\theta$.
>
> *Definition.* The **vector product**, written $\mathsf{a}\times\mathsf{b}$, is defined by $\mathsf{a}\times\mathsf{b} = ab\sin\theta\ \mathsf{c}$,
>
> where c is a *new* unit vector, **normal** (i.e. perpendicular) to the plane of a, b and pointing so that rotating a towards b would send a right-handed screw in the direction of c.

The 'scalar' product is just a number (in Physics a 'scalar' is a quantity not associated with any particular direction); but the vector product is connected with the *area* of the piece of surface defined by the two vectors – and c points 'up' from the surface, so as to show which is its 'top' side (as when we first set up the z-axis). Both products have the usual 'distributive' property, that is

$$(\mathsf{a}+\mathsf{b})\cdot\mathsf{c} = \mathsf{a}\cdot\mathsf{c} + \mathsf{b}\cdot\mathsf{c}, \qquad (\mathsf{a}+\mathsf{b})\times\mathsf{c} = \mathsf{a}\times\mathsf{c} + \mathsf{b}\times\mathsf{c},$$

but, from its definition, the vector product changes sign if the order of the vectors is reversed ($\mathsf{b}\times\mathsf{a} = -\mathsf{a}\times\mathsf{b}$) – so whatever we do we must keep them in the right order.

The unit vectors $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$ each have unit modulus, $|\mathsf{e}_1| = |\mathsf{e}_2| = |\mathsf{e}_3| = 1$; and each is perpendicular to the other two, $\mathsf{e}_1\cdot\mathsf{e}_2 = \mathsf{e}_1\cdot\mathsf{e}_3 = \mathsf{e}_2\cdot\mathsf{e}_3 = 0$. It follows that the scalar product between any pair of vectors a, b is, in terms of their components,

$$\begin{aligned}\mathsf{a}\cdot\mathsf{b} &= (a_1\mathsf{e}_1 + a_2\mathsf{e}_2 + a_3\mathsf{e}_3)\cdot(b_1\mathsf{e}_1 + b_2\mathsf{e}_2 + b_3\mathsf{e}_3)\\ &= a_1b_1\mathsf{e}_1\cdot\mathsf{e}_1 + \dots\ + a_1b_2\mathsf{e}_1\cdot\mathsf{e}_2 + \dots\ ,\end{aligned}$$

where the dots mean 'similar terms'; and from the properties of the unit vectors (above) this becomes

$$\mathsf{a}\cdot\mathsf{b} = a_1b_1 + a_2b_2 + a_3b_3. \tag{5.9}$$

When $\mathsf{b} = \mathsf{a}$ we get $\mathsf{a} \cdot \mathsf{a} = a^2 = a_1^2 + a_2^2 + a_3^2$ (the original sum-of-squares form for a length); and for the position vector r of any point P we find

$$OP = r = \sqrt{x^2 + y^2 + z^2}. \tag{5.10}$$

Similarly, for two vectors r, r', the scalar product is

$$\mathsf{r} \cdot \mathsf{r}' = rr' \cos\theta = xx' + yy' + zz'$$

and this tells us how to find the angle between any two vectors. Remember that x, y, z are projections of r on the three coordinate axes, so $x/r = \cos\alpha$ (α being the angle between r and the x-axis); and similarly for the second vector, $x'/r' = \cos\alpha'$. The cosines of the angles between a vector and the three axes are usually called the **direction cosines** of the vector and are denoted by l, m, n. With this notation the equation above can be re-written as

$$\cos\theta = ll' + mm' + nn' \tag{5.11}$$

– a simple way of getting the angle θ, which applies for *any* two vectors in 3-space.

5.5 Some examples

To end this chapter it's useful to look at a few examples of how you can describe points, lines, planes, and simple 3-dimensional shapes in vector language. By using vectors you can often get the results you need much more easily than by drawing complicated diagrams and thinking of all the 'special cases' that can arise.

- *Angles in a triangle* In Section 1.2 we took the theorem of Pythagoras, for a right-angled triangle as the 'metric axiom'. There are many theorems concerned with triangles that we haven't even mentioned; and many of them refer to a general triangle, with no special angles. Let's take such a triangle, with vertices A,B,C, using the same letters to denote the corresponding angles A, B, C, and the small letters a, b, c to denote the lengths of the sides *opposite* to angles A, B, C. We can also use the special symbols $\mathsf{a}, \mathsf{b}, \mathsf{c}$ to mean the *vectors* pointing along the sides, following one another in the positive (anti-clockwise) direction. (Before going on, you should make a careful drawing of the triangle ABC, labelling the sides and angles. Then you'll have the picture in your head.)

  There are two basic 'laws' relating the sines and cosines of the angles. The first is very easy to get: if you drop a perpendicular from vertex C onto the line through A and B, calling its length h, then $\sin A = h/b, \sin B = h/a$; and so $h = b \sin A = a \sin B$. On dividing by ab we get $(\sin A/a) = (\sin B/b)$. Taking vertex A next, you find a similar result; and on putting them together you find

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}. \tag{5.12}$$

  This is the 'Law of Sines' for any plane triangle.

Now note that the sum of the vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$ (displacements following each other round the triangle and bringing you back to the starting point) is zero: $\mathsf{a}+\mathsf{b}+\mathsf{c} = \mathsf{0}$. So $\mathsf{a} = -(\mathsf{b}+\mathsf{c})$ and the squared length of a is

$$\begin{aligned} a^2 &= \mathsf{a}\cdot\mathsf{a} = (\mathsf{b}+\mathsf{c})\cdot(\mathsf{a}+\mathsf{c}) \\ &= \mathsf{a}\cdot\mathsf{a} + \mathsf{b}\cdot\mathsf{b} + 2\mathsf{a}\cdot\mathsf{b} = b^2 + c^2 + 2\mathsf{b}\cdot\mathsf{c}. \end{aligned}$$

From the definition of the scalar product in Section 5.2, $\mathsf{b}\cdot\mathsf{c} = bc\cos\theta$ when both vectors point *away* from the point of intersection: but that means turning c round, making it $-\mathsf{c}$. The result you get, along with two others like it (obtained by taking vertex B in place of A, and then vertex C) give us the 'Law of Cosines':

$$\begin{aligned} a^2 &= b^2 + c^2 - 2bc\cos A \\ b^2 &= c^2 + a^2 - 2ca\cos B \\ c^2 &= a^2 + b^2 - 2ca\cos C. \end{aligned} \tag{5.13}$$

- *Vector equation of a straight line* Suppose we want the line to pass through a point A, with position vector a, and to be parallel to a given vector b – which can be of unit length ($b^2 = \mathsf{b}\cdot\mathsf{b} = 1$). Then a general point on the line, P, with position vector r, will be given by

$$\mathsf{r} = \mathsf{a} + s\mathsf{b} \tag{5.14}$$

where s is any variable number – and that's the equation we need! If instead we want the equation for a line passing through two points, A and B (position vectors a, b), then we simply replace b in the last equation by the vector $\mathsf{b}-\mathsf{a}$, which points from A to B: the result is

$$\mathsf{r} = \mathsf{a} + s(\mathsf{b}-\mathsf{a}).$$

- *Vector equation of a plane* Suppose ON is a **normal** to the plane, drawn from the origin O to the foot of the perpendicular, N; and let n be a unit vector in the direction ON, so $\vec{ON} = p\mathsf{n}$ where p is the perpendicular distance from O to the plane. If r is the position vector of P, any other point in the plane, then its *projection* (Section 5.2) on the line ON must have the same value p. In other words,

$$\mathsf{r}\cdot\mathsf{n} = p \tag{5.15}$$

will be the equation defining a plane, with unit normal n, at perpendicular distance p from the origin.

- *Distance of a point from a plane* The perpendicular distance from the origin to a point P *in the plane*, given by (5.15), is $p = \mathsf{r}\cdot\mathsf{n}$. That from the origin to *any other* point, P′ with position vector r', will be $p' = \mathsf{r}'\cdot\mathsf{n}$ – where we're thinking of point P′ as being in some parallel plane (which will have the same normal n). The required distance is therefore

$$d = p' - p = \mathsf{r}'\cdot\mathsf{n} - p$$

and this will be positive when P′ is *above* the given plane, going out from the origin in the direction n.

- *Intersection of two planes* The angle θ between two planes means the angle between their normals; so it follows from

$$\cos\theta = \mathsf{n}\cdot\mathsf{n}',$$

n, n' being the two unit normals. If θ is zero the planes will be parallel; but otherwise they will intersect – somewhere, but where? A point (r) which lies on both planes must satisfy both equations, $\mathsf{r}\cdot\mathsf{n} = p, \quad \mathsf{r}\cdot\mathsf{n}' = p'$. It will then lie on the line of intersection; but if we multiply the two equations by any two numbers c and c' and add the results we'll get

$$\mathsf{r}\cdot(c\mathsf{n} - c'\mathsf{n}') = cp - c'p'.$$

And this is the equation of a plane with its normal in the direction $c\mathsf{n} - c'\mathsf{n}'$: it describes a plane through the line of intersection of the two given planes – *which* one depending on the values we give to c and c'.

Now a vector $d\mathsf{n} + d'\mathsf{n}'$ (the numbers d, d' to be chosen), starting from the origin, will contain the normals $(\mathsf{n}, \mathsf{n}')$ to both planes and will therefore cut the line of intersection: we take it as the vector a in equation(5.14), choosing d and d' so that the point will lie on both planes. Then we need only the direction, the unit vector b in (5.14), to fix the line. And since the line of intersection is perpendicular to both normals we can take b as the *vector product* $\mathsf{n}\times\mathsf{n}'$ defined in Section 5.4. Putting things together, the equation of the line of intersection is

$$\mathsf{r} = d\mathsf{n} + d'\mathsf{n}' + s\mathsf{n}\times\mathsf{n}', \tag{5.16}$$

where the value of s changes as you run along the line.

- *Equation of a sphere* We've already met the equation for a sphere centred on the origin, in Section 5.2, in terms of Cartesian coordinates. Let's now look at one centred on the point C (position vector c), with radius R. The distance from C to the surface is the length of the vector $\mathsf{r} - \mathsf{c}$ and the condition for point r to lie on the surface is thus $|\mathsf{r} - \mathsf{c}|^2 = R^2$. Thus, expanding,

$$r^2 - 2\mathsf{r}\cdot\mathsf{c} + (c^2 - R^2) = 0 \tag{5.17}$$

and this is the equation of the sphere centred on point c.

- *Intersection of a straight line and a sphere* Suppose the line is given by (5.14) and the sphere by (5.17): the point r must satisfy both these conditions. If we put the first in the second we get

$$(\mathsf{a} - s\mathsf{b})\cdot(\mathsf{a} - s\mathsf{b}) - 2(\mathsf{a} - s\mathsf{b})\cdot\mathsf{c} + (c^2 - R^2) = 0.$$

This contains the first and second powers of the variable number s and will therefore be a quadratic equation (Book 1, Section 5.3), which can be written as

$$As^2 + Bs + C = 0,$$

where

$$A = b^2 = 1,\ B = 2\mathsf{b}\cdot(\mathsf{a}-\mathsf{c}),\ C = a^2 + c^2 - R^2 - 2\mathsf{a}\cdot\mathsf{c}.$$

There will be two roots, both real numbers, when $B^2 > 4AC$; and these values of s fix the two points where the straight line meets the surface. If it happens that B^2 and $4AC$ are exactly equal, then the two points become *one* and the line just touches the surface in a single point. The line is then a **tangent** to the sphere.

Exercises

1) Find a unit vector perpendicular to each of the vectors $\mathsf{v}_1 = 2\mathsf{e}_1 - \mathsf{e}_2 + \mathsf{e}_3$ and $\mathsf{v}_2 = 3\mathsf{e}_1 + 4\mathsf{e}_2 - \mathsf{e}_3$. Calculate the angle between v_1 and v_2.

2) Find two vectors which make equal angles with e_1, are perpendicular to each other, and are perpendicular to $\mathsf{e}_1 + \mathsf{e}_2 + \mathsf{e}_3$.

3) What is the vector equation of a straight line through the points $\mathsf{e}_1 - 2\mathsf{e}_2 + \mathsf{e}_3$ and $3\mathsf{e}_3 - 2\mathsf{e}_2$? And where does this line meet the plane which contains the origin and the points $4\mathsf{e}_2$ and $2\mathsf{e}_1 + \mathsf{e}_2$?

4) Show that the line joining the mid points of two sides of a triangle is parallel to the third side and is of half its length.

5) Show that the three points whose position vectors are a, b, and $3\mathsf{a} - 2\mathsf{b}$ lie on the same straight line.

6) Find the equation of the straight line passing through the point with position vector d and making equal angles with the vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$.

7) Find the equation of the plane through the point $2\mathsf{e}_1 + 3\mathsf{e}_2 - \mathsf{e}_3$ which is perpendicular to the vector $3\mathsf{e}_1 - 4\mathsf{e}_2 + 7\mathsf{e}_3$.

8) Show that the points $\mathsf{e}_1 - \mathsf{e}_2 + 3\mathsf{e}_3$ and $3(\mathsf{e}_1 + \mathsf{e}_2 + \mathsf{e}_3)$ are each the same distance from the plane

$$\mathsf{r}\cdot(5\mathsf{e}_1 + 2\mathsf{e}_2 - 7\mathsf{e}_3) + 9 = 0,$$

but are on opposite sides of it.

Chapter 6

Area and volume: invariance

6.1 Invariance of lengths and angles

At the end of Section 5.2 we noted that the objects we meet in 3-space have properties 'of their own' which don't change if we move them around from one part of space to another – as long as we don't bend them or twist them or change their 'natural' shapes. The objects may be, for example, rods or sticks (with a *length* of their own; or plates (with an *area*); or bricks or buckets (with a *volume*). All such properties are *invariant* under the **transformations** that simply move an object from one place to another. And in the last Section we laid the foundations for describing invariance mathematically, by using single symbols (vectors) to stand for elements of space: the separation of two points in an object, for example, is described by a vector $\mathsf{d} = d_1\mathsf{e}_1 + d_2\mathsf{e}_2 + d_3\mathsf{e}_3$, say, whose length does not change when we move the object. In fact, such transformations have the fundamental property of *leaving invariant all distances and angles* – which define the shape of the object. This was the property used by the Greeks in their development of plane geometry – for example in comparing two triangles to see if they were exactly alike, meaning one could be placed on top of the other with all sides and angles matching. They used pictures, but here we're using *algebraic* methods and working in three dimensions (solid geometry) rather than two; and it's here that vectors are especially useful.

Let the position vectors of points P and Q, relative to an origin O and a set of unit vectors $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$, be

$$\mathsf{p} = p_1\mathsf{e}_1 + p_2\mathsf{e}_2 + p_3\mathsf{e}_3 \qquad \mathsf{q} = q_2\mathsf{e}_1 + q_2\mathsf{e}_2 + q_2\mathsf{e}_3,$$

where (so as not to be confused) we use p_1, p_2, p_3 for the components of p instead of x, y, z. The vector pointing from P to Q (often written $\vec{PQ}$) is the *difference*

$$\vec{PQ} = \mathsf{d}_{PQ} = \mathsf{q} - \mathsf{p} = (q_1 - p_1)\mathsf{e}_1 + (q_2 - p_2)\mathsf{e}_2 + (q_3 - p_3)\mathsf{e}_3.$$

The simplest transformation we can make is a **translation**, in which every point P is moved into its **image**, P′, with position vector $\mathsf{p}' = \mathsf{p} + \mathsf{t}$, where t is a constant vector. It is clear from Fig.20 that the vector from P′ to Q′ is just the same as that from P to Q,

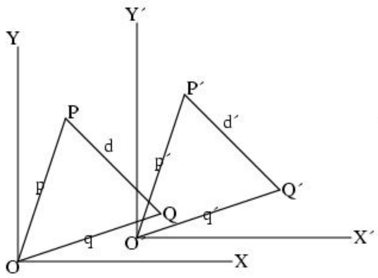

Figure 20

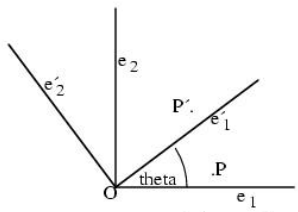

Figure 21

before moving the object: this idea can be expressed in the equation

$$\mathsf{d}_{P'Q'} = \mathsf{q}' - \mathsf{p}' = (\mathsf{q} + \mathsf{t}) - (\mathsf{p} + \mathsf{t}) = \mathsf{q} - \mathsf{p} = \mathsf{d}_{PQ}. \tag{6.1}$$

The vector separation of two points is *invariant* under the translation.

Let's think next of *rotating* the object into some new position: this is more difficult because an image point P′ now has a position vector p' related to p in a complicated way. But we can study a simple case – rotating the object around one axis, the z-axis with unit vector e_3. A rotation changes elements of space, not numbers, so we must ask what happens to the vectors $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$; and Fig.21 shows that a rotation through angle θ around e_3 (which points up out of the page) has the following effect –

$$\begin{aligned} \mathsf{e}_1 \to \mathsf{e}_1' &= \cos\theta \mathsf{e}_1 + \sin\theta \mathsf{e}_2, \\ \mathsf{e}_2 \to \mathsf{e}_2' &= -\sin\theta \mathsf{e}_1 + \cos\theta \mathsf{e}_2, \\ \mathsf{e}_3 \to \mathsf{e}_3' &= \mathsf{e}_3, \end{aligned} \tag{6.2}$$

where each unit vector turns into its image under the rotation, only e_3 (along the z-axis) staying as it was.

Now a point P, with position vector $\mathsf{p} = p_1\mathsf{e}_1 + p_2\mathsf{e}_2 + p_3\mathsf{e}_3$, is carried into P′, related in exactly the same way to the *new* unit vectors resulting from the rotation – nothing else has changed – and these are given in (6.2). The position vector of the image P′ is thus

$$\mathsf{p}' = p_1(\cos\theta \mathsf{e}_1 + \sin\theta \mathsf{e}_2) + p_2(-\sin\theta \mathsf{e}_1 + \cos\theta \mathsf{e}_2) + p_3\mathsf{e}_3,$$

when expressed in terms of the unit vectors before the rotation took place. This can be re-arranged to give

$$\mathsf{p}' = p_1'\mathsf{e}_1 + p_2'\mathsf{e}_2 + p_3'\mathsf{e}_3,$$

where

$$\begin{aligned} p_1' &= \cos\theta p_1 - \sin\theta p_2, \\ p_2' &= \sin\theta p_1 + \cos\theta p_2, \\ p_3' &= p_3. \end{aligned} \tag{6.3}$$

The new vector p' is clearly very different from p: but this is no surprise – what we are looking for is the invariance of *lengths* and *angles*. We'll just show that the length of the line OP is preserved in the rotation; then you can do the same for the angle between OP and OQ.

All we have to do is confirm that ${p'_1}^2 + {p'_2}^2 + {p'_3}^2$ (the square of the length OP′) is the same as before the rotation. The three terms are, from (6.3),

$$\begin{aligned}
{p'_1}^2 &= (\cos\theta)^2 p_1^2 + (\sin\theta)^2 p_2^2 - 2(\cos\theta\sin\theta)p_1p_2, \\
{p'_2}^2 &= (\sin\theta)^2 p_1^2 + (\cos\theta)^2 p_2^2 + 2(\cos\theta\sin\theta)p_1p_2, \\
{p'_3}^2 &= p_3^2,
\end{aligned}$$

and on adding these together, remembering that $(\cos\theta)^2 + (\sin\theta)^2 = 1$ for any angle θ, we get the expected result

$$ {p'_1}^2 + {p'_2}^2 + {p'_3}^2 = {p_1}^2 + {p_2}^2 + {p_3}^2. \tag{6.4}$$

The length of any vector is thus unchanged by rotation of the object.

After showing that the angles between any two vectors are also invariant, it follows that a transformation of this particular form (rotation around the z-axis) leaves unchanged the shape of an object, its surface area and its volume.

We must now think about area and volume in a bit more detail, but first let's note that what we've said about rotation around one special axis is true for *all* kinds of rotation. This is easy because, as we've just seen, an object is defined with reference to three unit vectors and its image (after rotation) is defined the same way in terms of the images of the unit vectors: so it's enough to know how $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$ are transformed. We also know that a rotated unit vector, pointing in *any* direction, can be found from the corresponding *direction cosines* (introduced just before (5.11)). If we use l_1, m_1, n_1 to fix the image e'_1 in terms of the original basis – and so on, we get as the most general transformation,

$$\begin{aligned}
\mathsf{e}_1 \to \mathsf{e}'_1 &= l_1\mathsf{e}_1 + m_1\mathsf{e}_2 + n_1\mathsf{e}_3, \\
\mathsf{e}_2 \to \mathsf{e}'_2 &= l_2\mathsf{e}_1 + m_2\mathsf{e}_2 + n_2\mathsf{e}_3, \\
\mathsf{e}_3 \to \mathsf{e}'_3 &= l_3\mathsf{e}_1 + m_3\mathsf{e}_2 + n_3\mathsf{e}_3.
\end{aligned} \tag{6.5}$$

These vectors will keep their original unit lengths provided

$$\mathsf{e}_1 \cdot \mathsf{e}_1 = l_1^2 + m_1^2 + n_1^2 = 1, \quad \text{etc.} \tag{6.6}$$

and will stay perpendicular to each other ($\cos\theta = 0$), provided

$$\mathsf{e}_1 \cdot \mathsf{e}_2 = l_1l_2 + m_1m_2 + n_1n_2 = 0, \quad \text{etc.} \tag{6.7}$$

according to (5.11). These are the general conditions that any rotation must satisfy in order that the image of an object will look exactly like the object before rotation. When all distances and angles are conserved in this way, the object and its image are said to be **congruent**. In fact, almost the whole of Euclid's geometry was based on the idea of congruence.

6.2 Area and volume

Starting from the idea of length, as the distance between the ends of a measuring rod, we have defined the surface area of a plane rectangular object (e.g. a plate) in Chapter 3: this quantity, a product of two lengths, was said to have "dimension L^2" and was measured by counting the number of 'units of area' (e.g. tiles) needed to cover it. In going from 2 to 3 dimensions similar ideas are used. The simplest definition of the *volume* of a box, whose sides are rectangles, is volume = product of the lengths of the 3 edges, a quantity with dimension L^3. The volume is measured by counting the number of 'units of volume' (e.g. bricks) needed to fill it. (See Book 1, Chapter 2, where we used this idea in setting up the laws for multiplying numbers: the number of bricks in a wall (Fig.7) was a product of three numbers – the numbers in the three directions, for length, thickness and height.)

To summarize the basic ideas, using vector language,

- Length (defined by one vector a) $= a = |\mathsf{a}|$
- Area (defined by two vectors, a, b) $= ab$
- Volume (defined by three vectors, $\mathsf{a}, \mathsf{b}, \mathsf{c}$) $= abc$

– the vectors being in the direction of the measurement and all being perpendicular to each other. Of course, we've taken for granted that the objects are rectangular (we've been working always with rectangular coordinates) and that a whole number of units will just fill the measured length, area, or volume. But when this is not so we know how to get round the difficulty by dividing the units into smaller and smaller 'sub-units'; or else, in the case of area, by breaking them into pieces (e.g. triangles, of known area) so as to fit more and more closely the area we're trying to measure. Finding the area of a circle (Section 3.1), by the method of Archimedes, is a beautiful example. In short, we can 'pin down' the quantity we're trying to measure as lying between 'this' and 'that' – where the 'this' and 'that' are **upper bounds** and **lower bounds**, respectively. And that means, in principle, that it can be measured by a real number (generally irrational, see Book 1) as accurately as we please!

So much for the simple definitions of length, area, and volume of simple shapes. More generally, we'll have to use ideas from another branch of mathematics – **calculus** – dealt with in other Books of the Series. But already things look a bit strange; because any length, in the definitions above, is measured by the vector distance between two points, which is taken positive only because we don't usually care whether it refers to 'going' or 'coming back' – and so decide to use the *modulus* of the vector. Similarly, the area may be defined in vector language as a vector product: the shape shown in Fig.22 (called a **parallelogram**), with two pairs of parallel sides, two of which are the vectors a, b, has a **vector area**

$$\mathsf{A} = \mathsf{a} \times \mathsf{b} = ab \sin\theta_{ab} \mathsf{n}, \qquad (6.8)$$

where n is a unit vector 'normal' (i.e. perpendicular) to the surface. (Notice that we're no longer talking only about rectangles, the vectors a, b being at any angle θ_{ab}.) The **normal** is determined (as in the definition following equation (5.8)) so as to point in the

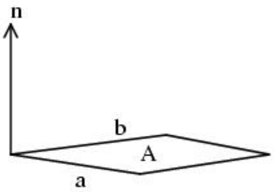

Figure 22

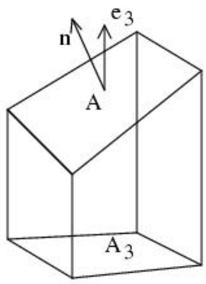

Figure 23

'right-hand screw' sense relative to a and b. When we talk about the area of the surface we're usually thinking of the *magnitude* of the vector area: $A = |\mathsf{A}|$. But if we need to know the difference between 'top' and 'bottom' we must always remember that the vector area A can carry a sign ($\pm$); and when we go on to look at *volume* we'll find similar problems. So we must deal with both things in a bit more detail.

Note: Skip the next Sections on first reading; but have a look at Chapter 7 (the last one!)

6.3 Area in vector form

Vector area is important when we think of something crossing or passing through a surface. If the surface is the open end of a water pipe the normal n can show the way the water flows (e.g. 'out', along n, when the vector in (6.8) is a positive number times n); and if we are thinking of the curved surface of an umbrella, then the resultant vector area will tell us how much cover it gives against the rain that falls on it.

Any kind of surface can be made out of very small elements (e.g. rectangles, with sides of lengths a and b), each with a vector area $\mathsf{A} = A\mathsf{n}$ (n chosen by the 'right-hand rule'). So we look at just one small element, writing its vector area as $\mathsf{A} = A_1\mathsf{e}_1 + A_2\mathsf{e}_2 + A_3\mathsf{e}_3$ where (taking a scalar product with e_1) $A_1 = \mathsf{A}\cdot\mathsf{e}_1$ and so on. The component A_3 is the *projection* of A in the direction e_3 (the z-axis in Fig.23) i.e. the projection on the xy-plane. Every element of the surface makes its own projection: so if we add the projections together we get the projection of the vector area of the *whole surface* on the xy-plane. If the xy-plane is the ground and the surface is a piece of board you're using to protect you against the rain, then

$$A_3 = \mathsf{A}\cdot\mathsf{e}_3 = A\mathsf{n}\cdot\mathsf{e}_3$$

and this projection will be the whole area of the board when you hold it horizontally, so that $\mathsf{n}\cdot\mathsf{e}_3 = 1$. But if you hold it sideways, so that n is parallel to the ground, then $\mathsf{n}\cdot\mathsf{e}_3 = 0$ and the projected area is zero – you get no cover at all!

Vector area is a very useful idea, as we'll find in other Books. For example, the vector area of any *closed* surface – like that of a rectangular box – is always *zero*: in this example opposite sides have the same area, but their normals (pointing out from the surface) are in opposite directions and the vector sum is zero. This is a general result: it means nothing can flow in or out through a surface that is closed – you'd have to make a hole in it.

Before starting on volume, it's useful to show how vector area can be written in terms of components. The vector area of a surface element defined in Fig.22 by the vectors a and b, with $\mathsf{a} = a_1\mathsf{e}_1 + a_2\mathsf{e}_2 + a_3\mathsf{e}_3,\ \mathsf{b} = b_1\mathsf{e}_1 + b_2\mathsf{e}_2 + b_3\mathsf{e}_3$, is $\mathsf{A} = \mathsf{a}\times\mathsf{b}$; and this becomes, remembering that $\mathsf{e}_1\times\mathsf{e}_2 = \mathsf{e}_3 \quad = -\mathsf{e}_2\times\mathsf{e}_1$, etc. and $\mathsf{e}_1\times\mathsf{e}_1 = 0$, etc.,

$$\begin{aligned}\mathsf{a}\times\mathsf{b} &= (a_1\mathsf{e}_1 + a_2\mathsf{e}_2 + a_3\mathsf{e}_3)\times(b_1\mathsf{e}_1 + b_2\mathsf{e}_2 + b_3\mathsf{e}_3)\\ &= (a_1b_2 - a_2b_1)\mathsf{e}_3 - (a_1b_3 - a_3b_1)\mathsf{e}_2\\ &\quad +(a_2b_3 - a_3b_2)\mathsf{e}_1.\end{aligned}$$

To remember things like this we first note that each component depends on two subscripts (e.g. the first on '1' and '2') and is multiplied by -1 if we change their order (e.g. '1,2' $\rightarrow$ '2,1') – it is **antisymmetric** under interchange of subscripts. There is a special notation for such quantities: we write

$$\begin{aligned}a_1b_2 - a_2b_1 &= \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}, \quad a_1b_3 - a_3b_1 = \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix},\\ a_2b_3 - a_3b_2 &= \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix},\end{aligned}$$

so that, from each array on the right, the corresponding component on the left is obtained as a product of the numbers on the 'leading diagonal' (e.g. a_1, b_2) *minus* the product of those on the 'second diagonal' (i.e. b_1, a_2). With this notation, the vector product above can be put in the (re-arranged) form

$$\mathsf{a}\times\mathsf{b} = \mathsf{e}_1\begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} - \mathsf{e}_2\begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + \mathsf{e}_3\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}. \qquad (6.9)$$

Each array, with the rule for 'multiplying it out' to get a single number, is called a **determinant**. We'll meet determinants in other Books, but for the moment we're just using the notation. Similar determinants can be set up, with any number of rows and columns, and any of them can be 'expanded' in terms of smaller determinants. To show how useful they can be in helping us to remember very complicated things, let's look at an expression for the vector product (6.9) as a *single* determinant with *three* rows and columns: it turns out to be

$$\mathsf{a}\times\mathsf{b} = \begin{vmatrix} \mathsf{e}_1 & \mathsf{e}_2 & \mathsf{e}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}. \qquad (6.10)$$

To expand this '3×3' determinant in the form (6.9) you take the element in the first row and the first column (it is e_1) and multiply it be the '2×2' determinant that's left when you strike out the first row and column; then you move to the next element in the first row (it is e_2) and do the same, multiplying it by the determinant that's left when you strike out the first row and *second* column; and then you move to the next element (e_3) and multiply it by the determinant that's left when you strike out the row and column that contain it. Finally, you add together the three contributions you have (one for e_1, one for e_2, and one for e_3) – but in working along the first row, in this way, you have to *multiply alternate contributions by* -1. If you use this simple recipe you will get (6.9).

We're now ready to find the volume of a 'box' (called a **parallelopiped**) defined by three vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$ as in Fig.24. This will be the 'volume element' in 3-space.

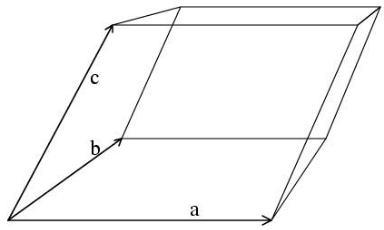

Figure 24

6.4 Volume in vector form

From Fig.24 we see that the whole object could be built up from thin slabs, each in the form of a parallelogram with area $ab\sin\theta_{ab}$ and thickness d i.e. with volume $abd\sin\theta_{ab}$. By stacking a number of such slabs, one on top of another, we get an approximation to the volume of any object with three sets of parallel faces (i.e. a parallelopiped). The top face is then at a vertical height $h = nd$ above the bottom face and the total volume (that of n slabs) is thus $abh\sin\theta_{ab}$. Now $h = c\cos\phi$, where c is the length of the vector c and ϕ is the angle it makes with the vertical (the normal to the plane of a and b). From this it follows that

$$V = abc\sin\theta_{ab}\cos\phi$$

and the formula will be exact in the limit where we take an enormous number of thinner and thinner slabs.

As in dealing with area, we can put this result in a convenient form even when all three vectors $(\mathsf{a}, \mathsf{b}, \mathsf{c})$ are expressed in terms of their components. The factor $ab\sin\theta_{ab}$ is the modulus of the vector area of the parallelogram, $\mathsf{A} = A\mathsf{n}$ (n being the upward-pointing normal in Fig.22), while $c\cos\phi = \mathsf{n}\cdot\mathsf{c}$ (ϕ being the Greek letter 'phi'); and the volume formula thus follows as a **triple product**

$$V = (\mathsf{a}\times\mathsf{b})\cdot\mathsf{c} = \mathsf{c}\cdot(\mathsf{a}\times\mathsf{b}). \tag{6.11}$$

Of course there's nothing special about the vector c: if we draw Fig.24 with vectors b and c along the edges of the bottom plane, instead of a and b, we'll get a different formula for the same volume. In this way we find

$$\begin{aligned} V &= \mathsf{a}\times\mathsf{b}\cdot\mathsf{c} = \mathsf{b}\times\mathsf{c}\cdot\mathsf{a} = \mathsf{c}\times\mathsf{a}\cdot\mathsf{b} \\ &= \mathsf{a}\cdot\mathsf{b}\times\mathsf{c} = \mathsf{b}\cdot\mathsf{c}\times\mathsf{a} = \mathsf{c}\cdot\mathsf{a}\times\mathsf{b} \end{aligned}$$

are all expressions for the same volume. The relative positions of the 'dot' and the 'cross' don't matter, so the triple product is often written as $[\mathsf{a\ b\ c}]$ and the last results then become

$$V = [\mathsf{a\ b\ c}] = [\mathsf{b\ c\ a}] = [\mathsf{c\ a\ b}],$$

where the different forms arise from a **cyclic interchange**, abc→bca→cab. Note that when the three vectors form a right-handed system, as in Fig.24, the volume V given in this way is always positive; but if you change this order the sign of the result is reversed. We needn't worry about this (we usually only want the *magnitude* of the volume) but we keep it in mind.

Finally, we express V in terms of the rectangular components of the vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$, as we did in the case of the vector area. Thus, writing $V = \mathsf{a} \cdot \mathsf{b} \times \mathsf{c}$ and using the formula (6.9), but with b, c in place of a, b, we see V can be written as the scalar product of

$$\mathsf{a} = a_1\mathsf{e}_1 + a_2\mathsf{e}_2 + a_3\mathsf{e}_3$$

and the vector product $\mathsf{b} \times \mathsf{c}$ in the form

$$\mathsf{e}_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - \mathsf{e}_2 \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + \mathsf{e}_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}.$$

From the properties of the Cartesian unit vectors ($\mathsf{e}_1 \cdot \mathsf{e}_1 = 1,\ \mathsf{e}_1 \cdot \mathsf{e}_2 = 0$, etc.) this product gives the volume V in the form

$$a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}.$$

But this is the expanded form of a single '3×3' determinant, as in (6.10); so we can write

$$V = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}. \tag{6.12}$$

This is a very general result: the vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$ can point in any directions and have any lengths we please – we only need to know their 3-space components and we can say at once what volume element they define.

Exercises

1) Use the transformation equation (6.3), which describes the rotation of all 3-space vectors around a common axis, to show that the angle between any two vectors, p and q, is unchanged by this rotation.

2) Show that the magnitude of the vector area defined by the two vectors a, b, and the volume of the parallelopiped defined by three vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$, are also invariant under the rotation (6.3).

3) Work out the volume of the parallelopiped in the last Exercise, and the vector areas of its six faces, when the vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$ are

$$\mathsf{a} = 3\mathsf{e}_1 + \mathsf{e}_2, \qquad \mathsf{b} = \mathsf{e}_1 + 2\mathsf{e}_2, \qquad \mathsf{c} = \mathsf{e}_1 + \mathsf{e}_2 + 2\mathsf{e}_3.$$

Make a drawing in which the vector areas are represented by arrows.

4) Besides the triple product in equation (6.11), which is a scalar quantity, there is also a *vector* triple product. For the three vectors $\mathsf{a}, \mathsf{b}, \mathsf{c}$ this is defined as the vector product of a with $\mathsf{b} \times \mathsf{c}$: $\mathsf{P}_{abc} = \mathsf{a} \times (\mathsf{b} \times \mathsf{c})$. Since P_{abc} is perpendicular to a and $\mathsf{b} \times \mathsf{c}$, while the latter is perpendicular to the plane containing b and c, the triple product must lie *in* the plane of b, c. Show that

$$\mathsf{P}_{abc} = (\mathsf{a} \cdot \mathsf{c})\mathsf{b} - (\mathsf{a} \cdot \mathsf{b})\mathsf{c}.$$

(This is quite hard! – and we don't use it unless we want to prove (7.19), near the end of the book. To get the result just given, you should introduce perpendicular unit vectors $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$, with e_2 parallel to b and e_3 in the plane of b, c. You can then put $\mathsf{b} = b\mathsf{e}_2$ and $\mathsf{c} = c_2\mathsf{e}_2 + c_3\mathsf{e}_3$ and also take $\mathsf{a} = a_1\mathsf{e}_1 + a_2\mathsf{e}_2 + a_3\mathsf{e}_3$. On expressing the vector products in $\mathsf{P}_{abc} = \mathsf{a} \times (\mathsf{b} \times \mathsf{c})$ in terms of the components of $\mathsf{a}, \mathsf{b}, \mathsf{c}$, you should find (noting that $\mathsf{b} \times \mathsf{c} = bc_3\mathsf{e}_2 \times \mathsf{e}_3 = bc_3\mathsf{e}_1$) $\mathsf{P}_{abc} = a_3bc_3\mathsf{e}_2 - a_2bc_3\mathsf{e}_3$.

This can be re-written - adding and subtracting a term $a_2bc_2\mathsf{e}_2$ -

$$\mathsf{P}_{abc} = (a_2c_2 + a_3c_3)b\mathsf{e}_2 - a_2b(c_2\mathsf{e}_2 + c_3\mathsf{e}_3).$$

The result we set out to prove is the same as this expression when we write the scalar products in terms of vector components.)

Chapter 7

Some other kinds of space

7.1 Many-dimensional space

So far we've been talking mainly about Euclidean spaces of 2 or 3 dimensions – 2-space and 3-space. They were *vector spaces*, containing all the vectors (v) that could be expressed in the form $\mathsf{v} = v_1\mathsf{e}_1 + v_2\mathsf{e}_2$ (2-space) or $\mathsf{v} = v_1\mathsf{e}_1 + v_2\mathsf{e}_2 + v_3\mathsf{e}_3$ (3-space), where $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$ are *basis vectors* and the coefficients v_1, v_2, v_3 are algebraic numbers called *vector components*. To include both cases we can write

$$\mathsf{v} = v_1\mathsf{e}_1 + v_2\mathsf{e}_2 + \dots\ v_n\mathsf{e}_n, \tag{7.1}$$

where $n = 2$ for 2-space and $n = 3$ for 3-space. Remember that every vector had a length (or magnitude) and a direction; and was often represented as an *arrow*, of given length and pointing in the given direction. (Mathematicians call the arrow a "directed line segment".)

Remember, too, that the components, $v_1, v_2, ...$, relate the vector to the basis and give us a way of labelling any point in space, P, as $\mathrm{P}(v_1, v_2, ...)$. The numbers $v_1, v_2, ...$ are components of a **position vector** (often denoted by r) corresponding to the line OP pointing from the origin O to the point P; and they are also called the *coordinates* of point P. So far, we have always chosen the basis vectors to be of *unit length* and *perpendicular* to one another. In the language of Chapter 6, any two basis vectors ($\mathsf{e}_i, \mathsf{e}_j$) have scalar products

$$\begin{aligned} \mathsf{e}_i \cdot \mathsf{e}_j &= 1 \text{ when } i = j\ ; \\ \mathsf{e}_i \cdot \mathsf{e}_j &= 0 \text{ when i} \neq \text{j}\ ; \end{aligned} \tag{7.2}$$

for all values of i, j in the range $1, 2, ..n$. This is the choice we started from in Chapter 1, taking it as the "metric axiom" for 2-space ($n = 2$). And the same choice, but with $n = 3$, leads to the 3-space considered in Chapter 6. In either case, the properties shown in (7.2) allow us to express the length of any vector in a 'sum-of-squares' form. In 3-space, for example, the square of a velocity vector, $|\mathsf{v}|^2 = \mathsf{v} \cdot \mathsf{v}$, is given by

$$|\mathsf{v}|^2 = (v_1\mathsf{e}_1 + v_2\mathsf{e}_2 + v_3\mathsf{e}_3) \cdot (v_1\mathsf{e}_1 + v_2\mathsf{e}_2 + v_3\mathsf{e}_3) = v_1^2 + v_2^2 + v_3^2, \tag{7.3}$$

where there are no terms such as $v_1 v_2$ because $\mathsf{e}_1 \cdot \mathsf{e}_2 = 0$.

The scalar products of the basis vectors are often set out in a square array, like this –

$$\begin{pmatrix} \mathsf{e}_1\cdot\mathsf{e}_1 & \mathsf{e}_1\cdot\mathsf{e}_2 & \mathsf{e}_1\cdot\mathsf{e}_3 \\ \mathsf{e}_2\cdot\mathsf{e}_1 & \mathsf{e}_2\cdot\mathsf{e}_2 & \mathsf{e}_2\cdot\mathsf{e}_3 \\ \mathsf{e}_3\cdot\mathsf{e}_1 & \mathsf{e}_3\cdot\mathsf{e}_2 & \mathsf{e}_3\cdot\mathsf{e}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{7.4}$$

An array of this kind is called the **metric matrix** of the space, and all such spaces – in which length can be defined as in (7.3) – are called "metric spaces".

Nothing we've said so far depends on n having the value 2 or 3: the simplest generalization of our ideas about geometry is just to keep everything, but allow n to become bigger than three. We then talk about "n-dimensional spaces". The fact that we can't imagine them, because we're so used to living in 3-space, is not important. If we can find a use for them, then we use them!

So let's put $n = 5$ and take it as an example of a 5-space. In Book 1, Chapter 6, we talked about a 'space' (though we didn't call it that) in which there were five *categories* of students in a class of 40. The categories were defined by putting the students into groups, according to the ranges into which their heights fall. Suppose we measure them and find the following results:

Heights of students	Numbers
Range (a): 1m 5cm to 1m 10cm	4 students
Range (b): 1m 10cm to 1m 15cm	8 students
Range (c): 1m 15cm to 1m 20cm	13 students
Range (d): 1m 20cm to 1m 25cm	12 students
Range (e): 1m 25cm to 1m 30cm	3 students

The numbers in these five categories show the 'state' of the class; and if we use a to stand for a student — no matter which one — in category (a), b for one in category (b), and so on, then we can describe the state of the class in symbols as

$$\mathsf{s} = 4\mathsf{a} + 8\mathsf{b} + 13\mathsf{c} + 12\mathsf{d} + 3\mathsf{e} \tag{7.5}$$

– which looks surprisingly like a vector! so we'll call it a *state vector*.

The students in the five categories can be 'sorted out' or *selected* by introducing *selection operators* (as we did in Book 1). Let's call them $\mathsf{A}, \mathsf{B}, \ldots\ \mathsf{E}$ so that A selects only students in group (a), and so on. These operators have (as we discovered) the algebraic properties

$$\mathsf{AA} = \mathsf{A}, \quad \mathsf{BB} = \mathsf{B}, \ldots \quad \mathsf{EE} = \mathsf{E} \tag{7.6}$$

and, for pairs of *different* operators,

$$\mathsf{AB} = \mathsf{BA} = 0,\ \mathsf{AC} = \mathsf{CA} = 0, \ldots\ \mathsf{DE} = \mathsf{ED} = 0. \tag{7.7}$$

And they work on the state vector s as follows:

$$\mathsf{As} = 4\mathsf{a}, \quad \mathsf{Bs} = 8\mathsf{b}, \quad \ldots \quad \mathsf{Es} = 3\mathsf{e},$$

This shows that each selects a part of the class and that putting the results together again we get the whole class:

$$(\mathsf{A}+\mathsf{B}+\mathsf{C}+\mathsf{D}+\mathsf{E})\mathsf{s} = 4\mathsf{a} + 8\mathsf{b} + ... + 3\mathsf{e} = \mathsf{s}.$$

In other words,

$$\mathsf{A}+\mathsf{B}+\mathsf{C}+\mathsf{D}+\mathsf{E} = \mathsf{1} \quad (7.8)$$

– the 'unit operator' which leaves any state vector unchanged. Operators with these properties form what mathematicians call a "spectral set": but here we've set them up using a very practical example, rather than snatching them out of the sky – as a real mathematician might do.

But let's get back to vector spaces. Algebra provides one way of dealing with selection, geometry provides another. When we use the vector (7.5) to stand for the 'state' of the school we're really thinking of $\mathsf{a}, \mathsf{b}, ... \mathsf{e}$ as 'basis vectors' or 'unit steps' along five different axes. And we can give them any properties we please – supposing, for example, that each of them is perpendicular to all the others, even though that would be impossible with 3-space thinking. The metric matrix will then no longer be (7.4): it will have five '1's along the diagonal and zeros everywhere else. It may all look strange – but who cares? We're only using a mathematical *language* and it's up to us to decide how the symbols should behave. Now that we've decided, we can think of s in (7.5) as the 5-dimensional vector formed by taking 4 steps of type a, 8 steps of type b, and so on, and combining them by addition (i.e. one after another, as in Fig.19). And the squared length of the vector, with this metric, will be the sum-of-squares of its components.

The selection operators can now be looked at geometrically: $\mathsf{As} = 4\mathsf{a}$ is simply the *projection* of the vector s on the axis defined by the unit vector a, while $\mathsf{Bs} = 8\mathsf{b}$ is its projection on the b axis. The property $\mathsf{AA} = \mathsf{A}$ then simply means that projecting *twice* on a given axis can produce nothing more than doing it only once; while $\mathsf{BA} = \mathsf{0}$ means that any projection on the a axis will have zero projection on the b axis – that's why we chose the unit vectors perpendicular (zero scalar products).

Sometimes it's useful to change this geometrical picture slightly. For example, if we want to compare two different classes, of different sizes, we'd be more interested in the *fractional* numbers of students in the various groups. In that case we might use a vector

$$\mathsf{s} = (4/40)\mathsf{a} + (8/40)\mathsf{b} + (13/40)\mathsf{c} + (12/40)\mathsf{d} + (3/40)\mathsf{e}$$

to show the state of the class, so that the projections along the five axes will represent these fractions directly. But then the 'pointer' s, which shows how the students are divided among the five groups, would not have very nice properties: if all the students belonged to the same group (a) we'd have $\mathsf{s} = (40/40)\mathsf{a} = \mathsf{a}$ and this would be a *unit* vector along the a axis – but that's a very special case. Is it possible to choose the vector components so that s will *always* be a unit vector, but will point in different directions according to the division of students into the five groups?

The components we've just tried, namely

$$(4/40),\ (8/40),\ (13/40),\ (12/40),\ (3/40),$$

won't do – because the sum of their squares doesn't give 1. But the sum of the numbers themselves *does* give 1. So why don't we try $\sqrt{4/40}$, $\sqrt{8/40}$, ...
$\sqrt{3/4}$? If we do this, the vector s showing the state of the class will become

$$\sqrt{4/40}\,\mathsf{a} + \sqrt{8/40}\,\mathsf{b} + \sqrt{13/40}\,\mathsf{c} + \sqrt{12/40}\,\mathsf{d} + \sqrt{3/40}\,\mathsf{e}$$

and the sum of the squares of the components will be exactly 1. So it *is* possible to represent the state of the class by a unit vector, pointing out from the origin in a 5-space, in a direction that will show the fractional number of students in each of the 5 categories.

If we want to compare two classes, to see if the heights of the students follow the same pattern, we just ask if the vectors s_1 and s_2 point in roughly the same direction. If they do, their scalar product $\mathsf{s}_1 \cdot \mathsf{s}_2$ will have a value close to 1; if the classes are very different (e.g. one of 5-year olds and one of 16-year olds) the scalar product of the vectors will be much closer to zero.

This example was about students, divided into groups according to height; but we might have been talking about potatoes of different sizes, or about objects produced in a factory and not all coming out quite right (some too big some too small), and we can use the same sort of vector description whenever we talk about categories. What's more, we can choose the metric in any way that seems useful for what we have in mind — as we'll see in the next two Sections.

7.2 Special Relativity: space-time

The starting point for this Book was the idea of distance and how it could be measured using a 'measuring-rod', whose **length** (the distance between its ends) was taken as the unit of distance. We also mentioned **time**, and how it could be measured using a 'clock' whose pendulum, swinging back and forth, marked out units of time; and also the **mass** of an object, which could be measured using a weighing machine. But so far mass and time haven't come into our picture of space: the idea of length alone has allowed us to build up the whole of Euclid's geometry.

Since about 1904, however, all that has changed. Space and time can't always be separated: it's no use giving my address (my 'coordinates' in space) if I don't live there any more — so perhaps my coordinates should really become x, y, z, t, the last one being the *time* at which I am (or was, or will be) there. The four coordinates together define a **space-time** point or an **event**; and when we talk about how things happen, or change, we need all four of them. This is especially true when two people (usually called the "observers") see the same event: one says it happens at the point x, y, z, t, the other says it happens at x', y', z', t'. But these numbers depend on the **reference frame** of the observer: from what origin in space (where $x = y = z = 0$) are the distances measured; and when was the clock started (by setting $t = 0$)? Einstein's theory of relativity is about how the numbers describing the same event, seen by different observers, are *related*.

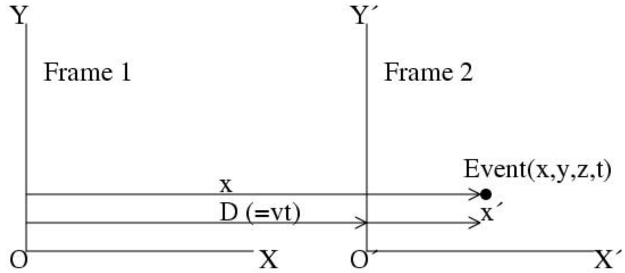

Figure 25

We've already looked at changes of reference frame in Chapter 6. Figure 20a showed how the distance between two points, P and Q, was left unchanged (invariant) when the frame was moved by a 'translation' in which $x_P \to x'_P = x_P + D$ etc. and $x_Q \to x'_Q = x_Q + D$ etc. – so the *differences* $x_P - x_Q$ stayed the same. But now we're going to move not the points but instead *the reference frame*, looking at the same points but seen by the different observers. And we'll take the simplest translation you can imagine (Figure 25), in which the frame is simply shifted along the x-axis. The *same point*, with coordinates x, y, z for the first observer, will then have coordinates x', y', z' for an observer in the shifted frame; and the relationship between the two sets of coordinates will be

$$x' = x - D, \quad y' = y, \quad z' = z.$$

If we want to include the time t, and suppose that the observers make their measurements at the same time, then the coordinates of the same *event* in 4-space will be related by

$$x' = x - D, \quad y' = y, \quad z' = z, \quad t' = t, \tag{7.9}$$

which is a very simple *linear transformation* (i.e. involving only first powers of the variables x, y, z, t and a 'constant' D).

When time is included, however, we have to think about *change* and *motion* – which we haven't done so far. If Frame 2 is moving relative to Frame 1, so that it goes a distance v to the right in every second (v not changing with time), then after t seconds it will have moved a distance $D = vt$. The constant v is called the **speed** of the motion. More generally, as in Fig.20a, D and v would become *vectors*, depending on *direction* and v would be the 'velocity vector'; so here v, the speed in the x-direction, is just the x-component of the velocity – and there's no harm in using the word "velocity" when we really mean speed.

After time t then, (7.1) will become

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t, \tag{7.10}$$

and this is called the "standard Galilean transformation". It goes back to the days of Galileo (1564–1642), who made some of the earliest experiments on motion. And it relates the coordinates of any given event, as measured by an observer in Frame 2, to those measured by one in Frame 1 — when Frame 2 moves with constant velocity v, relative to Frame 1, as in Fig.25. The science of **kinematics** (from the Greek word 'kinesis', meaning movement) deals with length, time, and movement; so now we're starting to think about kinematics. In this field the only 'tools' we need, in making experiments, are a measuring-rod and a clock; and very often we don't even need to actually *do* the experiments – it's enough to *think* about them, making a **thought-experiment**. We're going to make some amazing discoveries, just by thinking about things.

First of all, we'll suppose our clocks and measuring-rods are *perfect*. This means that if two lengths are found to be equal, then they will stay equal for all times (that's why we put in the word "perfect", because a real rod might get bent or broken); and similarly when two perfect clocks, both at the origin in some reference frame, show the same times, then they will do so even with a different choice of reference frame. As long as we're talking about kinematics (not about real objects, which have *mass* and are affected by 'gravity' – which we meet in Book 4) that's all we need.

Suppose you're in a train, waiting at a station for passengers to get on and off, and another train is passing. Each train is a reference frame, like the frames in Fig.25, and from your window you see people in the other train doing all the usual things – reading the newspaper, walking about, or even drinking tea: and perhaps you wonder for a moment which train is moving? Their train is moving with some velocity *v relative* to your train, but everything goes on as if it were'nt moving at all. In fact, *all movement is relative*: your train may not be moving relative to the station – but it is certainly moving (along with the whole station, the town, and the earth itself!) relative to the sun and the stars. You actually *feel* your relative motion only when it *changes*: if your train suddenly starts, you'll feel it; if you're standing you may even fall over. And the people in the other train will not notice they are moving with velocity v relative to you, unless *v changes*: if you see them falling over, or spilling their tea, you'll guess that the driver has put the brakes on and the train is slowing down. So there's something important about a relative velocity being *constant*: observers in two reference frames, moving with constant relative velocity, see things happening in exactly the same way. Albert Einstein (1879-1955) was the first to see just *how* important this was – for the whole of Physics. He took it as an axiom, which can be put in the following way:

> The laws of physics are exactly the same in any two reference frames in uniform relative motion (which means moving relative to each other with constant velocity in a straight line).

We'll call this *Einstein's Principle of Special Relativity* – "special" because objects with a *mass*, and subject to *gravity* (the force that makes things fall to the ground), are not yet included in the theory. The ideas of General Relativity, which takes account of mass and gravity, are much too difficult for this book, though we mention them briefly in the next Section. In Relativity Theory, frames "in uniform relative motion" are usually called **inertial frames** – but more about that in Book 4, where we begin to talk about mass.

Let's now go back to equation (7.10) which relates the coordinates of an *event*, as measured by observers in the reference frames of Fig.25. The observer in Frame 1, finds values

x, y, z, t, while the observer in Frame 2 finds values x', y', z', t' relative to his axes; both of them using the same standard unit of length and both having set their standard clocks to $t = t' = 0$ at the start of the experiment when, we suppose, the origin of Frame 2 is just on top of the origin of Frame 1. The distance in space to the event, call it s, is the same for both observers:

$$s^2 = x^2 + y^2 + z^2 = x'^2 + y'^2 + z'^2$$

and both believe $t = t'$, as they set their clocks to agree at the start (when O′ was passing O). The invariance of these quantities, in passing from one reference frame to the other is what leads to the 'transformation equations' (7.10), which now become

$$\begin{aligned} x' &= x - vt, \\ y' &= y, \\ z' &= z, \\ t' &= t. \end{aligned} \tag{7.11}$$

But this transformation is a bit too special: it keeps s^2 and t the same for both observers, but keeps them *separately* invariant – s^2 in 3-space, t in a 1-space. However, we agreed that time should be treated as just another coordinate. Is there a more general transformation, that will allow space and time coordinates to get mixed up? When this can happen, we'll be talking about a 4-space!

To see that such a transformation *can* be found, let's think of another simple event. We fire a gun, at the origin, at time $t = t' = 0$ just as O′ is passing O. The noise travels out from the gun, in all directions, with some constant speed which we can call c. After time t it will have reached all the points at a distance $r = ct$ from the origin O. These will lie on a surface of radius $r = ct$ (a sphere) such that

$$r^2 = c^2t^2 = x^2 + y^2 + z^2.$$

If we could assume that an observer in Frame 2 (along with his friends – all with standard clocks – placed at points where the noise arrived) all observed the same sphere of noise arrivals, then we'd suppose that

$$s^2 = c^2t^2 - x^2 - y^2 - z^2 = c^2t'^2 - x'^2 - y'^2 - z'^2 \tag{7.12}$$

was another invariant. We call it the squared **interval** (not just 'distance') and it depends on all four coordinates. Notice that (7.12) defines a 4-space metric that's a bit strange: it has a matrix like that in (7.4) but with three diagonal elements the same, the fourth having opposite sign (e.g. three -1s and one $+1$). But, after all, time (we've given it a 'time coordinate' ct) and space (with coordinates x, y, z) *are* different – and this shows up in the sign difference.

Of course, a 'thought experiment' like this would be difficult to do; and we don't know if it has any relationship to the real world. But it does suggest something we can try.

Let's *suppose* then, that in Einstein's 4-space the space and time coordinates of events observed from frames in uniform relative motion (Fig.25) are related so that (7.12) is satisfied. The big question is now: *What is this relationship?* And to get the answer we can argue as follows.

The new invariant contains a new constant (c), also a velocity, like the v in (7.11); and so v/c must be a pure number, which will go to zero if the constant c is big enough, or if v is small enough. Let's now define a number, usually called γ_v (Greek 'gamma', with a subscript to show it depends on the relative velocity v):

$$\gamma_v = \frac{1}{\sqrt{(1 - v^2/c^2)}}. \tag{7.13}$$

Notice that we've used the *squares* of the velocities in the denominator, because changing the direction of the x-axis will change the sign of a velocity – and we don't expect it will matter whether the axis points to the right or the left. Also, when v is small the denominator in (7.13) will go towards 1 – and so will γ_v. So if the new transformation equations depend on γ_v they will fall back into the Galilean transformation when the two reference frames are hardly moving– just as we'd expect.

Let's now try, instead of the first three equations in (7.11),

$$x' = \gamma_v(x - vt), \qquad y' = y, \qquad z' = z.$$

And instead of taking time to be universal, the same for both observers, let's try something a bit like the first equation above. If we put

$$t' = \gamma_v(t - ? \times x),$$

where '?' stands for something we don't yet know, then we can substitute the values of x', y', z', t' (given in the last four equations) into the right-hand side of (7.12); and comparing the two sides will tell us what to choose for the '?'. The only terms that contain t alone (not t^2) are $c^2\gamma_v^2 \times (-2xt\times?)$ and $2\gamma_v^2 xvt$. There's nothing to balance these terms on the left-hand side of (7.12), so the equality tells us that their sum must be zero and this fixes the '?' To get zero we must choose $?= v/c^2$ and so we must take

$$t' = \gamma_v\left(t - \frac{v}{c^2}x\right).$$

What we have shown is that the supposed invariance of the 'metric form' $c^2t^2 - x^2 - y^2 - z^2$ *requires that* the Galilean transformation equations be changed, becoming

$$\begin{aligned} x' &= \gamma_v(x - vt), \\ y' &= y, \\ z' &= z, \\ t' &= \gamma_v\left(t - \frac{v}{c^2}x\right). \end{aligned} \tag{7.14}$$

These are the equations of the **Lorentz transformation**, named after the Dutch mathematician and physicist Lorentz (1857–1928), who first got them, but never guessed how they would change the world! That was left to Einstein, who found them again and made them a cornerstone of his relativity theory.

Nowadays we're always hearing about mass and energy (who hasn't ever seen Einstein's famous equation $E = mc^2$?), atomic power, atomic bombs, space travel, and the strange things that happen in the universe. But let's stop for a minute! We haven't even got as far as physics: that will have to wait for other Books (beginning in Book 4). This Section is just a start, in which we're beginning to use some of the things we already know about number and space. Before this we didn't even include *time*, and we still haven't really thought about *mass*. So it's amazing that we can get so far just by thinking about things. Before stopping we'll connect briefly with what we call 'reality' – a few questions and a few conclusions.

The first question is What is the meaning of the constant c? and the second is How big is it? – and does it correspond to anything we can measure? In fact, there *is* something that travels through empty space with the velocity c: it is *light*, which we all know goes extremely fast – if you switch a light on it seems to fill the whole room in no time at all! Physics tells us what light is and gives us ways of finding how fast it travels: if the switched-on light starts from the origin, then it reaches a point with (space) coordinates x, y, z after a time t given by $t = (distance/velocity) = \sqrt{x^2 + y^2 + z^2}/c$, where c can be calculated in terms of quantities we can measure in the laboratory. And its value is almost exactly 300 million metres every second (3×10^8m s^{-1}), so in everyday life we needn't worry about using the Galilean equations (7.10). The other big question is How did we get so far without knowing any physics? The answer is not at all easy, but roughly speaking it's because we left out *mass* and *gravity*, and *electric charges*, and most of the things that go into physics – thinking only of kinematics (length, time, and motion) – except when we supposed that all the 'physics' was the same for "two observers in uniform relative motion". We didn't need all the details: the Lorentz transformation follows, as we saw, from *kinematical* principles. We're just lucky to find that physics supplies a 'natural' method of getting the *value* of the constant c.

What about conclusions? The first one is that there's a natural limit to the speed with which anything can move – even an observer in a spacecraft – and this limit is $v = c$. For then γ_v in (7.13) would become infinite; and for $v > c$ it would become imaginary. All the quantities we measure and relate must be *real*; and finite, so the only velocities we can consider must be less than c.

There are many more amazing conclusions. We'll just mention two: if an observer in Frame 1 looks at an object in Frame 2, he'll be surprised to find that it has shrunk in the direction of motion; and that a clock in Frame 2 is going slow!

The Lorentz contraction

Suppose we have a measuring-rod of length l_0, lying along the x-axis and not moving relative to Frame 2; and call its ends A and B. It will be moving *relative to us*, in Frame 1, with velocity v. But to an observer in Frame 2 it will be at rest and will have a **proper**

length, also called **rest length**,

$$l_0 = x'_B - x'_A, \tag{7.15}$$

not depending on what time his clock shows.

Looking at the rod from *our* reference frame (Frame 1), the length of the rod at time t on our clock will be

$$l = x_B(t) - x_A(t). \tag{7.16}$$

But we know from (7.14) how the coordinates measured in the two frames must be related:

$$x'_A = \gamma_v(x_A - vt), \qquad x'_B = \gamma_v(x_B - vt).$$

It follows that, using (7.15),

$$l_0 = x'_B - x'_A = \gamma_v(x_B - x_A) = \gamma_v l,$$

where l, given in (7.16), is the length of the rod *according to us*. Thus,

$$l = l_0/\gamma_v. \tag{7.17}$$

In other words, the measured length of the rod when it's moving away from us with velocity v, will be *less than* the *rest length* – as measured in a frame where it is not moving. This effect is called the **Lorentz contraction**. It is very small for speeds which are tiny compared with c ($\approx$ 300 thousand kilometres/second): so we never notice it in everyday life. But it is important in physics – and accurate measurements are in perfect agreement with the predictions.

Time dilation

Another remarkable conclusion follows just as easily. A clock moving away from us will register intervals of time different from those shown by a clock at rest in our reference frame: times get longer – an effect called **time dilation**.

Remember, we measured the times t, t' from the moment when the clock at the origin in Frame 2 passes that in Frame 1, setting $t' = t = 0$. The clock at the origin in Frame 2 will be at the point with $x' = 0$ but relative to Frame 2 its position at time t will be $x = vt$. Now according to the last equation in (7.14) the times shown, for the same event (as noted by two different observers), must be related by

$$t' = \gamma_v\left(t - \frac{v}{c^2}x\right) = \gamma_v t\left(1 - \frac{v^2}{c^2}\right) = \gamma_v t/\gamma_v^2 = t/\gamma_v,$$

where we've put in the value $x = vt$, for the moving clock, and used the definition of γ_v in (7.13). Thus,

$$t = \gamma_v t'. \tag{7.18}$$

In other words, all times measured in the moving system (Frame 2) must be multiplied by γ_v to get the times measured on our clock in Frame 1. Now the time taken for something to happen – the time between two events, A and B say, at a given position in space – will be $T_0 = t'_B - t'_A$ for an observer moving with his clock (Frame 2): he will call it his "proper time". And this leads to some very strange effects: for instance, if Frame 2 comes back to the origin O in Fig.25, after travelling all the way round the world, the Frame 1 observer (who stayed at home with his clock) will note that the journey took time $T = \gamma_v T_0$ – which is *longer* than the time (T_0) noted by the traveller. Who is right? Both are: each has his own 'proper time' and we shouldn't be surprised if they don't agree. The differences are normally almost too small to measure: but, by using extremely accurate ('atomic') clocks and taking them round the world on ordinary commercial aircraft, they *have* been measured and are in rough agreement with the formula. More accurate experiments really do confirm (7.18).

7.3 Curved spaces: General Relativity

In Section 1.1 we said that "space itself is very slightly 'bent', especially near very heavy things like the sun and the stars, so that Euclid's ideas are never perfectly correct ... " One of Einstein's most brilliant ideas, which he developed during the years 1905–1915, was that the mass of a heavy object produced a local 'curvature' in the space around it: this led him from the theory of Special Relativity to that of General Relativity, in which mass and its effects are included. As we haven't yet done any Physics we can't even begin to talk about General Relativity. But we *are* ready to think about 'curved space' and what it means.

In Special Relativity the 4-space metric (three space coordinates and one more for time) was very similar to that for ordinary Euclidean 3-space (Section 5.2): the square of the interval ('distance') between two events ('points' in space-time) still had a 'sum-of-squares' form, apart from the ± signs attached to the 4 terms; and it had the same form however big the interval. A space like that is called 'pseudo-Euclidean'.

In General Relativity, the metric form is no longer so simple; and it's no longer the same for all points in space – it can depend on where you are. To get an idea of what this means we'll use the example from Section 1.1: the surface of the earth is a curved space, though it's only a 2-space and it's a bit special because the curvature is the same at all points – how much it's bent depends only on the radius of the earth. Of course, the mathematics of curved surfaces is important for making maps. And it was important in the ancient world because the astronomers at that time believed the sun and the moon moved around the earth on spherical surfaces. The Hindus and Arabs invented many arithmetic rules for making calculations of their positions, but the rules were not turned into algebraic formulas until about the 13th century. The theory that followed tells us how to calculate lengths and angles for lines which are 'as straight as you can make them' on a spherical surface. Such a line follows the shortest path between two points, A and B, *on the surface* and is called a **geodesic** (from the Greek words for 'earth' and 'measurement'). If a ship sails from point A on the earth's surface, to point B, always keeping the same direction, and does the same in going from B to a third point C, then the three-sided path ABC is called a **spherical triangle**. The geometry of such paths was studied by mariners for hundreds of years and led to the branch of mathematics called *spherical trigonometry.*

What we want to get now is the form of the metric that determines the distance between points in a curved 2-space – points on a spherical surface 'embedded' in the 3-space world we live in. If we can do it for this case, then we'll get ideas about how to do it for a curved 3-space embedded in a 4-space – or for a curved 4-space embedded in a 5-space. Notice that if we want to 'bend' a space we always need (at least) one extra dimension to describe the bending: we can't describe the surface of a sphere, which is *two*-dimensional, without a third dimension to describe the sphere itself!

First of all we need to generalize the 'Law of Sines' and the 'Law of Cosines' (Section 5.5), which apply to a triangle with vertices, A,B,C, on a flat surface: we want corresponding results for the spherical surface shown in Fig.26.

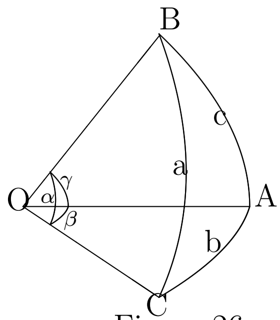

Figure 26

Suppose $\mathsf{A}, \mathsf{B}, \mathsf{C}$ are the position vectors of points A,B,C, relative to the centre of the sphere (the earth); and use A, B, C for the *angles* (on the surface) at the corners of the triangle. We'll also use a similar notation for the lengths of the sides, a for the side opposite to angle A, and so on.

The Law of Sines looks almost the same as for a flat surface, being

$$\frac{\sin A}{\alpha} = \frac{\sin B}{\beta} = \frac{\sin C}{\gamma}, \qquad (7.19)$$

but the denominators are *angles* – not side lengths. Remember, however, that the angles α, β, γ are at the centre of the sphere (Fig.26), not at the vertices of the triangle. At the same time, $\alpha = a/R$, where a is an *arc* length; so we can replace the angles in (7.19) by side lengths – as long as we remember the sides are bent! And then the formula looks exactly like that for a flat surface.

The Law of Cosines is the one we really need. It follows from what we know about the triple product (Section 6.4). The angle A is that between the planes AOB and AOC, the same as the angle between the *normals*: and a vector normal to AOB is $\mathsf{A} \times \mathsf{B}$, while one normal to AOC is $\mathsf{A} \times \mathsf{C}$. The angle A can thus be found from the scalar product of the two normals, which will give us $\cos A$. So let's look at the scalar product $(\mathsf{A} \times \mathsf{B}) \cdot (\mathsf{A} \times \mathsf{C})$, noting that choosing the radius $R = 1$ makes no difference to the angles.

The scalar product can be reduced using the result (see the Exercises on Chapter 6)

$$(\mathsf{A} \times \mathsf{B}) \cdot (\mathsf{A} \times \mathsf{C}) = (\mathsf{A} \cdot \mathsf{A})(\mathsf{B} \cdot \mathsf{C}) - (\mathsf{A} \cdot \mathsf{C})(\mathsf{A} \cdot \mathsf{B}).$$

For a sphere of unit radius,

$$\mathsf{B}\cdot\mathsf{C} = \cos\alpha, \qquad \mathsf{C}\cdot\mathsf{A} = \cos\beta, \qquad \mathsf{A}\cdot\mathsf{B} = \cos\gamma.$$

Also $\mathsf{A}\cdot\mathsf{B}$ is a vector of length $\sin\gamma$, normal to the plane AOB and pointing inwards (i.e.on the C-side); while $\mathsf{A}\cdot\mathsf{C}$ is of length $\sin\beta$, normal to plane AOC but pointing outwards.

On putting these values into the expression above, we find

$$(\mathsf{A}\times\mathsf{B})\cdot(\mathsf{A}\times\mathsf{C}) = \sin\beta\sin\gamma\cos A = \cos\alpha - \cos\beta\cos\gamma.$$

There are two other relations of similar form, obtained by starting from angle B and angle C (instead of A). They are all collected in the Law of Cosines for a spherical triangle:

$$\begin{aligned}\cos\alpha &= \cos\beta\cos\gamma + \sin\beta\sin\gamma\cos A,\\ \cos\beta &= \cos\gamma\cos\alpha + \sin\gamma\sin\alpha\cos B,\\ \cos\gamma &= \cos\alpha\cos\beta + \sin\alpha\sin\beta\cos C,\end{aligned} \tag{7.20}$$

for the cosines. The angles α, β, γ (radian measure) are related to the arc lengths BC, CA, AB on the spherical surface: for example, putting $BC = a$, the angle α is given by $\alpha = a/R$, where R is the radius of the sphere.

Now think of A as an 'origin of coordinates' on the surface and take the outgoing arcs, AB and AC, as axes, choosing the angle between them as $A = \pi/2$. On putting $\cos A = 0$, the first line in (7.20) tells us that

$$\cos\alpha = \cos\beta\cos\gamma \tag{7.21}$$

and this gives us all we need. For points near to A, it's enough to use the first few terms of the cosine series (Chapter 4) and to write the last equation as

$$1 - \frac{a^2}{2R^2} + \frac{a^4}{24R^4}.. = \left(1 - \frac{b^2}{2R^2} + \frac{b^4}{24R^4}..\right)\left(1 - \frac{c^2}{2R^2} + \frac{c^4}{24R^4}..\right).$$

If we multiply everything by $2R^2$ and compare the terms of second degree on the two sides of the = sign, the result is a first approximation:

$$a^2 \approx b^2 + c^2, \tag{7.22}$$

The squared length of the arc BC has Euclidean form: it is a sum of squares of distances along the other two arcs – in accordance with the metric axiom in Section 1.2 – just as it would be for a *flat* surface. But the metric is only *locally* Euclidean: more accurately, there are 'correction terms'

$$-(1/12R^2)a^2, \qquad \text{and} \qquad -(1/12R^2)(b^4 + c^4) - b^2c^2/R^2,$$

which must be added on the left and on the right, respectively, of equation (7.22).

Of course, when the *radius* of curvature R, is infinitely large the 2-space becomes flat (zero curvature); but in General Relativity even a very small curvature of 4-dimensional space-time is enough to account for many properties of the universe. Without Physics, which we'll start with in Book 4, it's not possible to go any further: but without the genius of Einstein and others like him it would never have been possible to get this far.

Exercises

1) When we use the vector (7.5) to stand for the 'state' of a class (how big are the students in it) we're using $\mathsf{a}, \mathsf{b}, ... \mathsf{e}$ as 'basis vectors'. The components we used, namely

$$\frac{4}{40}, \frac{8}{40}. \frac{13}{40}. \frac{12}{40}, \frac{3}{40}$$

(being the fractional numbers of students in the five height ranges) didn't give a *unit* vector – because the sum of their squares doesn't give 1.

Show that by using the *square roots* of the numbers as components you will always get a unit vector. So it *is* possible to represent the state of the class by a unit vector, pointing out from the origin in a 5-space in a direction that will show the fractional number of students in each of the 5 categories.

2) Suppose you want to compare two classes, to see if the heights of the students follow the same pattern. Prepare vectors s_1 and s_2, like that in Exercise 1 but for two different classes (e.g. 20 15-year old girls and 18 14-year old boys). Is the pattern of heights similar or not?

(You can either measure or just guess the heights. The patterns will be similar if the vectors point in roughly the same direction. If they do, their scalar product $\mathsf{s}_1 \cdot \mathsf{s}_2$ will have a value close to 1. For two very different classes (e.g. one of 5-year olds and one of 16-year olds) the scalar product of the vectors will be much closer to zero.)

Looking back –

You started this book knowing only about numbers and how to work with them, using the methods of algebra. Now you've learnt how to measure the **quantities** you meet in space (distances, area, volume), each one being a number of **units**. And you've seen that these ideas give you a new starting point for geometry, different from the one used by Euclid, and lead you directly to modern forms of geometry. Again, you've passed many milestones on the way:

- Euclid started from a set of **axioms**, the most famous being that two parallel straight lines never meet, and used them to build up the whole of geometry: in Chapter 1 you started from different axioms – a **distance axiom** and a **metric axiom** – which both follow from *experiment.*

- Two straight lines, with one point in common, define a **plane**; the metric axiom gave you a way of testing to see if the two lines are **perpendicular**; and then you were able to define two **parallel** straight lines – giving you a new way of looking at Euclid's axiom. Using sets of perpendicular and parallel straight lines you could find numbers (x, y), the coordinates, that define any point in the plane. Any straight line in the plane was then described by a simple **equation**; and so was a circle.

- In Chapter 3 you learnt how to calculate the **area** of a triangle and of a circle and to evaluate π ('pi') by the method of Archimedes. You studied angles and found some of the key results about the angles between straight lines that cross.

- Chapter 4 reminded you of some of the things you'd learnt in Book 1, all needed in the study of **rotations**. You learnt about the **exponential function**, e^x, defined as a **series**, and its properties; and found its connection with angle and the 'trigonometric' functions.

- In talking about 3-space, the first thing to do was to set up **axes** and decide how to label every point with *three* **coordinates**; after that everything looked much the same as in 2-space. But it's not easy to *picture* things in 3-space and it's better to use **vector algebra**. For any pair of vectors we found two new 'products' – a **scalar product** (just a number) and a **vector product** (a new vector), both depending on the lengths of the vectors and the angle between them. Examples and Exercises showed how useful they could be in 3-space geometry.

- Chapter 6 was quite hard! But the ideas underneath can be understood easily: lengths, areas and volumes are all *unchanged* if you move something through space – making a 'transformation'. This fact was often used by Euclid (usually in 2-space) in proving theorems about areas; but by the end of the Chapter you have all the 'tools' for doing things much more generally, as we do them today.

- To end the book (Chapter 7) you took a look at the next big generalization – to spaces of n dimensions, where n is *any* integer. Of course, you couldn't imagine them: but the *algebra* was the same, for any value of n. So you were able to invent new spaces, depending on what you wanted to use them for. One such space was invented by Einstein, just a hundred years ago, to bring *time* into the description of space – counting t as a fourth coordinate, similar to x, y, z. And you got a glimpse of some of the amazing things that came out as a result, things that could be checked by experiment and were found to be true.

Before closing this book, stop for a minute and think about what you've done. Perhaps you started studying science with Book 1 (two years ago? three or four years ago?) and now you're at the end of Book 2. You started from almost nothing; and after working through about 150 pages you can understand things that took people thousands of years to discover, some of the great creations of the human mind – of the Scientific Mind.

Index