

## PART IV. PROBLEMS, HINTS, SOLUTIONS

---

This last part offers the reader additional opportunity for practice.

The problems require no more preliminary knowledge than the reader could have acquired from a good high-school curriculum. Yet they are not too easy and not mere routine problems; some of them demand originality and ingenuity.<sup>12</sup>

The hints offer indications leading to the result, mostly by quoting an appropriate sentence from the list; to a very attentive reader ready to pick up suggestions they may reveal the key idea of the solution.

The solutions bring not only the answer but also the procedure leading to the answer, although, of course, the reader has to supply some of the details. Some solutions try to open up some further outlook by a few words placed at the end.

The reader who has earnestly tried to solve the problem has the best chance to profit by the hint and the solution. If he obtains the result by his own means, he may learn something by comparing his method with the method given in print. If, after a serious effort, he is inclined to give up, the hint may supply him with the

<sup>12</sup> Except Problem 1 (widely known, but too amusing to miss) all the problems are taken from the Stanford University Competitive Examinations in Mathematics (there are a few minor changes). Some of the problems were formerly published in *The American Mathematical Monthly* and/or *The California Mathematics Council Bulletin*. In the latter periodical also some solutions were published by the author; they appear appropriately rearranged in the sequel.

missing idea. If even the hint does not help, he may look at the solution, try to isolate the key idea, put the book aside, and then try to work out the solution.

## PROBLEMS

1. A bear, starting from the point *P*, walked one mile due south. Then he changed direction and walked one mile due east. Then he turned again to the left and walked one mile due north, and arrived exactly at the point *P* he started from. What was the color of the bear?

2. Bob wants a piece of land, exactly level, which has four boundary lines. Two boundary lines run exactly north-south, the two others exactly east-west, and each boundary line measures exactly 100 feet. Can Bob buy such a piece of land in the U.S.?

3. Bob has 10 pockets and 44 silver dollars. He wants to put his dollars into his pockets so distributed that each pocket contains a different number of dollars. Can he do so?

4. To number the pages of a bulky volume, the printer used 2989 digits. How many pages has the volume?

5. Among Grandfather's papers a bill was found:

72 turkeys \$—67.9—

The first and last digit of the number that obviously represented the total price of those fowls are replaced here by blanks, for they have faded and are now illegible.

What are the two faded digits and what was the price of one turkey?

6. Given a regular hexagon and a point in its plane. Draw a straight line through the given point that divides the given hexagon into two parts of equal area.

7. Given a square. Find the locus of the points from

which the square is seen under an angle (a) of  $90^\circ$  (b) of  $45^\circ$ . (Let  $P$  be a point outside the square, but in the same plane. The smallest angle with vertex  $P$  containing the square is the "angle under which the square is seen" from  $P$ .) Sketch clearly both loci and give a full description.

8. Call "axis" of a solid a straight line joining two points of the surface of the solid and such that the solid, rotated about this line through an angle which is greater than  $0^\circ$  and less than  $360^\circ$  coincides with itself.

Find the axes of a cube. Describe clearly the location of the axes, find the angle of rotation associated with each. Assuming that the edge of the cube is of unit length, compute the arithmetic mean of the lengths of the axes.

9. In a tetrahedron (which is not necessarily regular) two opposite edges have the same length  $a$  and they are perpendicular to each other. Moreover they are each perpendicular to a line of length  $b$  which joins their midpoints. Express the volume of the tetrahedron in terms of  $a$  and  $b$ , and prove your answer.

10. The vertex of a pyramid opposite the base is called the *apex*. (a) Let us call a pyramid "isosceles" if its apex is at the same distance from all *vertices* of the base. Adopting this definition, prove that the base of an isosceles pyramid is *inscribed* in a circle the center of which is the foot of the pyramid's altitude.

(b) Now let us call a pyramid "isosceles" if its apex is at the same (perpendicular) distance from all sides of the base. Adopting this definition (different from the foregoing) prove that the base of an isosceles pyramid is *circumscribed* about a circle the center of which is the foot of the pyramid's altitude.

11. Find  $x$ ,  $y$ ,  $u$ , and  $v$ , satisfying the system of four equations

$$\begin{aligned}x + 7y + 3v + 5u &= 16 \\8x + 4y + 6v + 2u &= -16 \\2x + 6y + 4v + 8u &= 16 \\5x + 3y + 7v + u &= -16\end{aligned}$$

(This may look long and boring: look for a short cut.)

12. Bob, Peter, and Paul travel together. Peter and Paul are good hikers; each walk  $p$  miles per hour. Bob has a bad foot and drives a small car in which two people can ride, but not three; the car covers  $c$  miles per hour. The three friends adopted the following scheme: They start together, Paul rides in the car with Bob, Peter walks. After a while, Bob drops Paul, who walks on; Bob returns to pick up Peter, and then Bob and Peter ride in the car till they overtake Paul. At this point they change: Paul rides and Peter walks just as they started and the whole procedure is repeated as often as necessary.

- (a) How much progress (how many miles) does the company make per hour?
- (b) Through which fraction of the travel time does the car carry just one man?
- (c) Check the extreme cases  $p = 0$  and  $p = c$ .

13. Three numbers are in arithmetic progression, three other numbers in geometric progression. Adding the corresponding terms of these two progressions successively, we obtain

$$85, \quad 76, \quad \text{and} \quad 84$$

respectively, and, adding all three terms of the arithmetic progression, we obtain 126. Find the terms of both progressions.

14. Determine  $m$  so that the equation in  $x$

$$x^4 - (3m + 2)x^2 + m^2 = 0$$

has four real roots in arithmetic progression.

15. The length of the perimeter of a right triangle is 60 inches and the length of the altitude perpendicular to the hypotenuse is 12 inches. Find the sides.

16. From the peak of a mountain you see two points,  $A$  and  $B$ , in the plain. The lines of vision, directed to these points, include the angle  $\gamma$ . The inclination of the first line of vision to a horizontal plane is  $\alpha$ , that of the second line  $\beta$ . It is known that the points  $A$  and  $B$  are on the same level and that the distance between them is  $c$ .

Express the elevation  $x$  of the peak above the common level of  $A$  and  $B$  in terms of the angles  $\alpha$ ,  $\beta$ ,  $\gamma$ , and the distance  $c$ .

17. Observe that the value of

$$\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots + \frac{n}{(n+1)!}$$

is  $1/2$ ,  $5/6$ ,  $23/24$  for  $n = 1, 2, 3$ , respectively, guess the general law (by observing more values if necessary) and prove your guess.

18. Consider the table

$$\begin{array}{rcl} 1 & = & 1 \\ 3 + 5 & = & 8 \\ 7 + 9 + 11 & = & 27 \\ 13 + 15 + 17 + 19 & = & 64 \\ 21 + 23 + 25 + 27 + 29 & = & 125 \end{array}$$

Guess the general law suggested by these examples, express it in suitable mathematical notation, and prove it.

19. The side of a regular hexagon is of length  $n$  ( $n$  is an integer). By equidistant parallels to its sides the hexagon is divided into  $T$  equilateral triangles each of which has sides of length 1. Let  $V$  denote the number of vertices appearing in this division, and  $L$  the number of boundary lines of length 1. (A boundary line belongs to one or two triangles, a vertex to two or more triangles.) When

$n = 1$ , which is the simplest case,  $T = 6$ ,  $V = 7$ ,  $L = 12$ . Consider the general case and express  $T$ ,  $V$ , and  $L$  in terms of  $n$ . (Guessing is good, proving is better.)

20. In how many ways can you change one dollar? (The "way of changing" is determined if it is known how many coins of each kind—cents, nickels, dimes, quarters, half dollars—are used.)

## HINTS

1. *What is the unknown?* The color of a bear—but how could we find the color of a bear from mathematical data? *What is given?* A geometrical situation—but it seems self-contradictory: how could the bear, after walking three miles in the manner described, return to his *starting point*?

2. *Do you know a related problem?*

3. If Bob had very many dollars, he would have obviously no difficulty in filling each of his pockets differently. *Could you restate the problem?* What is the minimum number of dollars that can be put in 10 pockets so that no two different pockets contain the same amount?

4. *Here is a problem related to yours:* If the book has exactly 9 numbered pages, how many digits uses the printer? (9, of course.) *Here is another problem related to yours:* If the book has exactly 99 numbered pages, how many digits does the printer use?

5. *Could you restate the problem?* What can the two faded digits be if the total price, expressed in cents, is divisible by 72?

6. *Could you imagine a more accessible related problem? A more general problem? An analogous problem?* (GENERALIZATION, 2.)

7. *Do you know a related problem?* The locus of the points from which a given segment of a straight line is

seen under a given angle consists of two circular arcs, ending in the extreme points of the segment, and symmetric to each other with respect to the segment.

8. I assume that the reader is familiar with the shape of the cube and has found certain axes just by inspection—but are they *all* the axes? *Can you prove* that your list of axes is exhaustive? Has your list a clear principle of classification?

9. *Look at the unknown!* The unknown is the volume of a tetrahedron—yes, I know, the volume of any pyramid can be computed when the base and the height are given (product of both, divided by 3) but in the present case neither the base nor the height is given. *Could you imagine a more accessible related problem?* (Don't you see a more accessible tetrahedron which is an aliquot part of the given one?)

10. *Do you know a related theorem?* Do you know a related . . . simpler . . . analogous theorem? Yes: the foot of the altitude is the mid-point of the base in an isosceles triangle. Here is a *theorem related to yours and proved before*. *Could you use . . . its method?* The theorem on the isosceles triangle is proved from congruent right triangles of which the altitude is a common side.

11. It is assumed that the reader is somewhat familiar with systems of linear equations. To solve such a system, we have to combine its equations in some way—look out for relations between the equations which could indicate a particularly advantageous combination.

12. *Separate the various parts of the condition. Can you write them down?* Between the start and the point where the three friends meet again there are three different phases:

- (1) Bob rides with Paul
- (2) Bob rides alone
- (3) Bob rides with Peter.

Call  $t_1$ ,  $t_2$ , and  $t_3$  the durations of these phases, respectively. How could you split the condition into appropriate parts?

13. *Separate the various parts of the condition. Can you write them down? Let*

$$a - d, \quad a, \quad a + d$$

be the terms of the arithmetic progression, and

$$bg^{-1}, \quad b, \quad bg$$

be the terms of the geometric progression.

14. *What is the condition?* The four roots must form an arithmetic progression. Yet the equation has a particular feature: it contains only even powers of the unknown  $x$ . Therefore, if  $a$  is a root,  $-a$  is also a root.

15. *Separate the various parts of the condition. Can you write them down?* We may distinguish three parts in the condition, concerning

- (1) perimeter
- (2) right triangle
- (3) height to hypotenuse.

16. *Separate the various parts of the condition. Can you write them down?* Let  $a$  and  $b$  stand for the lengths of the (unknown) lines of vision,  $\alpha$  and  $\beta$  for their inclinations to the horizontal plane, respectively. We may distinguish three parts in the condition, concerning

- (1) the inclination of  $a$
- (2) the inclination of  $b$
- (3) the triangle with sides  $a$ ,  $b$ , and  $c$ .

17. *Do you recognize the denominators 2, 6, 24? Do you know a related problem? An analogous problem? (INDUCTION AND MATHEMATICAL INDUCTION.)*

18. Discovery by induction needs observation. Observe the right-hand sides, the initial terms of the left-hand sides, and the final terms. What is the general law?

19. *Draw a figure.* Its observation may help you to discover the law inductively, or it may lead you to relations between  $T$ ,  $V$ ,  $L$ , and  $n$ .

20. *What is the unknown?* What are we supposed to seek? Even the aim of the problem may need some clarification. *Could you imagine a more accessible related problem? A more general problem? An analogous problem?* Here is a very simple analogous problem: In how many ways can you pay one cent? (There is just one way.) Here is a more general problem: In how many ways can you pay the amount of  $n$  cents using these five kinds of coins: cents, nickels, dimes, quarters, and half dollars. We are especially concerned with the particular case  $n = 100$ .

In the simplest particular cases, for small  $n$ , we can figure out the answer without any high-brow method, just by trying, by inspection. Here is a short table (which the reader should check).

|       |   |   |   |    |    |    |    |    |    |    |
|-------|---|---|---|----|----|----|----|----|----|----|
| $n$   | 4 | 5 | 9 | 10 | 14 | 15 | 19 | 20 | 24 | 25 |
| $E_n$ | 1 | 2 | 2 | 4  | 4  | 6  | 6  | 9  | 9  | 13 |

The first line lists the amounts to be paid, generally called  $n$ . The second line lists the corresponding numbers of "ways of paying," generally called  $E_n$ . (Why I have chosen this notation is a secret of mine which I am not willing to give away at this stage.)

We are especially concerned with  $E_{100}$ , but there is little hope that we can compute  $E_{100}$  without some clear method. In fact the present problem requires a little more from the reader than the foregoing ones; he should *create a little theory*.

Our question is general (to compute  $E_n$  for general  $n$ ),

but it is "isolated." Could you imagine a more accessible related problem? An analogous problem? Here is a very simple analogous problem: Find  $A_n$ , the number of ways to pay the amount of  $n$  cents, using only cents. ( $A_n = 1$ .)

## SOLUTIONS

1. You think that the bear was white and the point  $P$  is the North Pole? Can you prove that this is correct? As it was more or less understood, we idealize the question. We regard the globe as exactly spherical and the bear as a moving material point. This point, moving due south or due north, describes an arc of a *meridian* and it describes an arc of a *parallel circle* (parallel to the equator) when it moves due east. We have to distinguish two cases.

(1) If the bear returns to the point  $P$  along a meridian different from the one along which he left  $P$ ,  $P$  is necessarily the North Pole. In fact the only other point of the globe in which two meridians meet is the South Pole, but the bear could leave this pole only in moving northward.

(2) The bear could return to the point  $P$  along the same meridian he left  $P$  if, when walking one mile due east, he describes a parallel circle exactly  $n$  times, where  $n$  may be 1, 2, 3 . . . In this case  $P$  is not the North Pole, but a point on a parallel circle very close to the South Pole (the perimeter of which, expressed in miles, is slightly inferior to  $2\pi + 1/n$ ).

2. We represent the globe as in the solution of Problem 1. The land that Bob wants is bounded by two meridians and two parallel circles. Imagine two fixed meridians, and a parallel circle moving away from the equator: the arc on the moving parallel intercepted by the two fixed meridians is steadily shortened. The center of the land that Bob wants should be on the equator: he can not get it in the U.S.

3. The least possible number of dollars in a pocket is obviously 0. The next greater number is at least 1, the next greater at least 2 . . . and the number in the last (tenth) pocket is at least 9. Therefore, the number of dollars required is at least

$$0 + 1 + 2 + 3 + \dots + 9 = 45$$

Bob cannot make it: he has only 44 dollars.

4. A volume of 999 numbered pages needs

$$9 + 2 \times 90 + 3 \times 900 = 2889$$

digits. If the bulky volume in question has  $x$  pages

$$2889 + 4(x - 999) = 2989$$

$$x = 1024$$

This problem may teach us that a preliminary estimate of the unknown may be useful (or even necessary, as in the present case).

5. If \_679\_ is divisible by 72, it is divisible both by 8 and by 9. If it is divisible by 8, the number 79\_ must be divisible by 8 (since 1000 is divisible by 8) and so 79\_ must be 792: the last faded digit is 2. If \_6792 is divisible by 9, the sum of its digits must be divisible by 9 (the rule about "casting out nines") and so the first faded digit must be 3. The price of one turkey was (in grandfather's time)  $\$367.92 \div 72 = \$5.11$ .

6. "*A point and a figure with a center of symmetry* (in the same plane) are given in position. Find a straight line that passes through the given point and bisects the area of the given figure." The required line passes, of course, through the center of symmetry. See INVENTOR'S PARADOX.

7. In any position the two sides of the angle must pass through two vertices of the square. As long as they pass through the same pair of vertices, the angle's vertex

moves along the same arc of circle (by the theorem underlying the hint). Hence each of the two loci required consists of several arcs of circle: of 4 semicircles in the case (a) and of 8 quarter circles in the case (b); see Fig. 31.

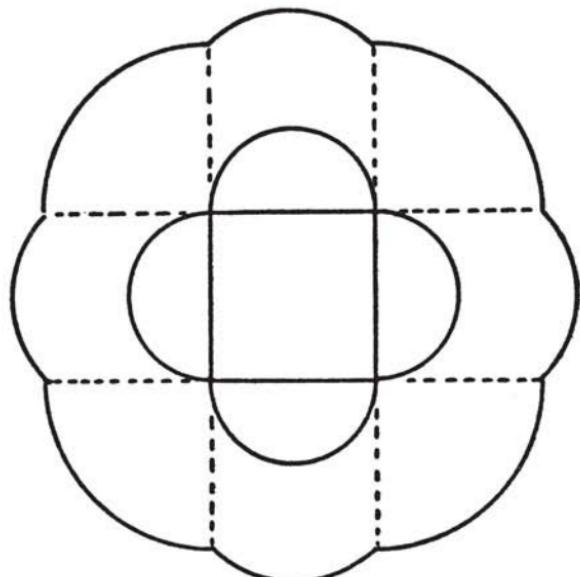


FIG. 31

8. The axis pierces the surface of the cube in some point which is either a vertex of the cube or lies on an edge or in the interior of a face. If the axis passes through a point of an edge (but not through one of its endpoints) this point must be the midpoint: otherwise the edge could not coincide with itself after the rotation. Similarly, an axis piercing the interior of a face must pass through its center. Any axis must, of course, pass through the center of the cube. And so there are three kinds of axes:

- (1) 4 axes, each through two opposite vertices; angles  $120^\circ, 240^\circ$

- (2) 6 axes, each through the mid-points of two opposite edges; angle  $180^\circ$   
 (3) 3 axes, each through the center of two opposite faces; angles  $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ .

For the length of an axis of the first kind see section 12; the others are still easier to compute. The desired average is

$$\frac{4\sqrt{3} + 6\sqrt{2} + 3}{13} = 1.416.$$

(This problem may be useful in preparing the reader for the study of crystallography. For the reader sufficiently advanced in the integral calculus it may be observed that the average computed is a fairly good approximation to the "average width" of the cube, which is, in fact,  $3/2 = 1.5$ .)

9. The plane passing through one edge of length  $a$  and the perpendicular of length  $b$  divides the tetrahedron into two *more accessible* congruent tetrahedra, each with base  $ab/2$  and height  $a/2$ . Hence the required volume

$$= 2 \cdot \frac{1}{3} \cdot \frac{ab}{2} \cdot \frac{a}{2} = \frac{a^2 b}{6}.$$

10. The base of the pyramid is a polygon with  $n$  sides. In the case (a) the  $n$  lateral edges of the pyramid are equal; in the case (b) the altitudes (drawn from the apex) of its  $n$  lateral faces are equal. If we draw the altitude of the pyramid and join its foot to the  $n$  vertices of the base in the case (a), but to the feet of the altitudes of the  $n$  lateral faces in the case (b), we obtain, in both cases, *n right triangles of which the altitude (of the pyramid) is a common side*: I say that these  $n$  right triangles are congruent. In fact the hypotenuse [a lateral edge in the case (a), a lateral altitude in the case (b)] is of the same length in each, according to the definitions

laid down in the proposed problem; we have just mentioned that another side (the altitude of the pyramid) and an angle (the right angle) are common to all. In the  $n$  congruent triangles the third sides must also be equal; they are drawn from the same point (the foot of the altitude) in the same plane (the base): they form  $n$  radii of a circle which is circumscribed about, or inscribed into, the base of the pyramid, in the cases (a) and (b), respectively. [In the case (b) it remains to show, however, that the  $n$  radii mentioned are perpendicular to the respective sides of the base; this follows from a well-known theorem of solid geometry on projections.]

It is most remarkable that a plane figure, the isosceles triangle, may have *two different analogues* in solid geometry.

11. Observe that the first equation is so related to the last as the second is to the third: the coefficients on the left-hand sides are the same, but in opposite order, whereas the right-hand sides are opposite. Add the first equation to the last and the second to the third:

$$\begin{aligned} 6(x + u) + 10(y + v) &= 0, \\ 10(x + u) + 10(y + v) &= 0. \end{aligned}$$

This can be regarded as a system of two linear equations for two unknowns, namely for  $x + u$  and  $y + v$ , and easily yields

$$x + u = 0, \quad y + v = 0.$$

Substituting  $-x$  for  $u$  and  $-y$  for  $v$  in the first two equations of the original system, we find

$$\begin{aligned} -4x + 4y &= 16 \\ 6x - 2y &= -16. \end{aligned}$$

This is a simple system which yields

$$x = -2, \quad y = 2, \quad u = 2, \quad v = -2$$

12. Between the start and the meeting point each of the friends traveled the same distance. (Remember, distance = velocity  $\times$  time.) We distinguish two parts in the condition:

Bob traveled as much as Paul:

$$ct_1 - ct_2 + ct_3 = ct_1 + pt_2 + pt_3,$$

Paul traveled as much as Peter:

$$ct_1 + pt_2 + pt_3 = pt_1 + pt_2 + ct_3.$$

The second equation yields

$$(c - p)t_1 = (c - p)t_3.$$

We assume, of course, that the car travels faster than a pedestrian,  $c > p$ . It follows

$$t_1 = t_3;$$

that is, Peter walks just as much as Paul. From the first equation, we find that

$$\frac{t_3}{t_2} = \frac{c + p}{c - p}$$

which is, of course, also the value for  $t_1/t_2$ . Hence we obtain the answers:

$$(a) \quad \frac{c(t_1 - t_2 + t_3)}{t_1 + t_2 + t_3} = \frac{c(c + 3p)}{3c + p}$$

$$(b) \quad \frac{t_2}{t_1 + t_2 + t_3} = \frac{c - p}{3c + p}$$

(c) In fact,  $0 < p < c$ . There are two extreme cases:

If  $p = 0$  (a) yields  $c/3$  and (b) yields  $1/3$

If  $p = c$  (a) yields  $c$  and (b) yields  $0$ .

These results are easy to see without computation.

13. The condition is easily split into four parts expressed by the four equations

$$\begin{aligned}a - d + bg^{-1} &= 85 \\a + b &= 76 \\a + d + bg &= 84 \\3a &= 126.\end{aligned}$$

The last equation yields  $a = 42$ , then the second  $b = 34$ . Adding the remaining two equations (to eliminate  $d$ ), we obtain

$$2a + b(g^{-1} + g) = 169.$$

Since  $a$  and  $b$  are already known, we have here a quadratic equation for  $g$ . It yields

$$g = 2, \quad d = -26 \quad \text{or} \quad g = 1/2, \quad d = 25.$$

The progressions are

$$68, 42, 16 \qquad \qquad \qquad 17, 42, 67$$

or

$$17, 34, 68 \qquad \qquad \qquad 68, 34, 17$$

14. If  $a$  and  $-a$  are the roots having the least absolute value, they will stand next to each other in the progression which will, therefore, be of the form

$$-3a, -a, a, 3a.$$

Hence the left-hand side of the proposed equation must have the form

$$(x^2 - a^2)(x^2 - 9a^2).$$

Carrying out the multiplication and comparing coefficients of like powers, we obtain the system

$$\begin{aligned}10a^2 &= 3m + 2, \\9a^4 &= m^2.\end{aligned}$$

Elimination of  $a$  yields

$$19m^2 - 108m - 36 = 0.$$

Hence  $m = 6$  or  $-6/19$ .

15. Let  $a$ ,  $b$ , and  $c$  denote the sides, the last being the hypotenuse. The three parts of the condition are expressed by

$$\begin{aligned} a + b + c &= 60 \\ a^2 + b^2 &= c^2 \\ ab &= 12c. \end{aligned}$$

Observing that

$$(a + b)^2 = a^2 + b^2 + 2ab$$

we obtain

$$(60 - c)^2 = c^2 + 24c.$$

Hence  $c = 25$  and either  $a = 15$ ,  $b = 20$  or  $a = 20$ ,  $b = 15$  (no difference for the triangle).

16. The three parts of the condition are expressed by

$$\sin \alpha = \frac{x}{a},$$

$$\sin \beta = \frac{x}{b},$$

$$c^2 = a^2 + b^2 - 2ab \cos \gamma$$

The elimination of  $a$  and  $b$  yields

$$x^2 = \frac{c^2 \sin^2 \alpha \sin^2 \beta}{\sin^2 \alpha + \sin^2 \beta - 2 \sin \alpha \sin \beta \cos \gamma}.$$

17. We conjecture that

$$\frac{1}{2!} + \frac{2}{3!} + \cdots + \frac{n}{(n+1)!} = 1 - \frac{1}{(n+1)!}.$$

Following the pattern of INDUCTION AND MATHEMATICAL INDUCTION, we ask: Does the conjectured formula remain

true when we pass from the value  $n$  to the next value  $n + 1$ ? Along with the formula above we should have

$$\frac{1}{2!} + \frac{2}{3!} + \cdots + \frac{n}{(n+1)!} + \frac{n+1}{(n+2)!} = 1 - \frac{1}{(n+2)!}$$

Check this by subtracting from it the former:

$$\frac{n+1}{(n+2)!} = -\frac{1}{(n+2)!} + \frac{1}{(n+1)!}$$

which boils down to

$$\frac{n+2}{(n+2)!} = \frac{1}{(n+1)!}$$

and this last equation is obviously true for  $n = 1, 2, 3, \dots$  hence, by following the pattern referred to above, we can prove our conjecture.

18. In the  $n$ th line the right-hand side seems to be  $n^3$  and the left-hand side a sum of  $n$  terms. The final term of this sum is the  $m$ th odd number, or  $2m - 1$ , where

$$m = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2};$$

see INDUCTION AND MATHEMATICAL INDUCTION, 4. Hence the final term of the sum on the left-hand side should be

$$2m - 1 = n^2 + n - 1.$$

We can derive hence the initial term of the sum considered in two ways: going back  $n - 1$  steps from the final term, we find

$$(n^2 + n - 1) - 2(n - 1) = n^2 - n + 1$$

whereas, advancing one step from the final term of the foregoing line, we find

$$[(n - 1)^2 + (n - 1) - 1] + 2$$

which, after routine simplification, boils down to the same: good! We assert therefore that

$$(n^2 - n + 1) + (n^2 - n + 3) + \cdots + (n^2 + n - 1) = n^3$$

where the left-hand side indicates the sum of  $n$  successive terms of an arithmetic progression the difference of which is 2. If the reader knows the rule for the sum of such a progression (arithmetic mean of the initial term and the final term, multiplied by the number of terms), he can verify that

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

and so prove the assertion.

(The rule quoted can be easily proved by a picture little different from Fig. 18.)

19. The length of the perimeter of the regular hexagon with side  $n$  is  $6n$ . Therefore, this perimeter consists of  $6n$  boundary lines of length 1 and contains  $6n$  vertices. Therefore, in the transition from  $n - 1$  to  $n$ ,  $V$  increases by  $6n$  units, and so

$$V = 1 + 6(1 + 2 + 3 + \cdots + n) = 3n^2 + 3n + 1;$$

see INDUCTION AND MATHEMATICAL INDUCTION, 4. By 3 diagonals through its center the hexagon is divided into 6 (large) equilateral triangles. By inspection of one of these

$$T = 6(1 + 3 + 5 + \cdots + 2n - 1) = 6n^2$$

(rule for the sum of an arithmetic progression, quoted in the solution of Problem 18). The  $T$  triangles have jointly  $3T$  sides. In this total  $3T$  each internal line of division of length 1 is counted twice, whereas the  $6n$  lines along the perimeter of the hexagon are counted but once. Hence

$$2L = 3T + 6n, \quad L = 9n^2 + 3n.$$

(For the more advanced reader: it follows from Euler's theorem on polyhedra that  $T + V = L + 1$ . Verify this relation!)

20. Here is a well-ordered array of analogous problems: Compute  $A_n$ ,  $B_n$ ,  $C_n$ ,  $D_n$  and  $E_n$ . Each of these quantities represents the number of ways to pay the amount of  $n$  cents; the difference is in the coins used:

$A_n$  only cents

$B_n$  cents and nickels

$C_n$  cents, nickels, and dimes

$D_n$  cents, nickels, dimes, and quarters

$E_n$  cents, nickels, dimes, quarters, and half dollars.

The symbols  $E_n$  (reason now clear) and  $A_n$  were used before.

All ways and manners to pay the amount of  $n$  cents with the five kinds of coin are enumerated by  $E_n$ . We may, however, distinguish two possibilities:

*First.* No half dollar is used. The number of such ways to pay is  $D_n$ , by definition.

*Second.* A half dollar (possibly more) is used. After the first half dollar is laid on the counter, there remains the amount of  $n - 50$  cents to pay, which can be done in exactly  $E_{n-50}$  ways.

We infer that

$$E_n = D_n + E_{n-50}.$$

Similarly

$$D_n = C_n + D_{n-25},$$

$$C_n = B_n + C_{n-10},$$

$$B_n = A_n + B_{n-5}.$$

A little attention shows that these formulas remain valid if we set

$$A_0 = B_0 = C_0 = D_0 = E_0 = 1$$

(which obviously makes sense) and regard any one of the quantities  $A_n, B_n \dots E_n$  as equal to 0 when its subscript happens to be negative. (For example,  $E_{25} = D_{25}$ , as can be seen immediately, and this agrees with our first formula since  $E_{25-50} = E_{-25} = 0$ .)

Our formulas allow us to compute the quantities considered *recursively*, that is, by going back to lower values of  $n$  or to former letters of the alphabet. For example, we can compute  $C_{30}$  by simple addition if  $C_{20}$  and  $B_{30}$  are already known. In the table below the initial row, headed by  $A_n$ , and the initial column, headed by 0, contain only numbers equal to 1. (Why?) Starting from these initial numbers, we compute the others recursively, by simple additions: any other number of the table is equal either to the number above it or to the sum of two numbers: the number above it and another at the proper distance to the left. For example,

$$C_{30} = B_{30} + C_{20} = 7 + 9 = 16$$

The computation is carried through till  $E_{50} = 50$ : you can pay 50 cents in exactly 50 different ways. Carrying it further, the reader can convince himself that  $E_{100} = 292$ : you can change a dollar in 292 different ways.

| $n$   | 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|-------|---|---|----|----|----|----|----|----|----|----|----|
| $A_n$ | 1 | 1 | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  |
| $B_n$ | 1 | 2 | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 |
| $C_n$ | 1 | 2 | 4  | 6  | 9  | 12 | 16 | 20 | 25 | 30 | 36 |
| $D_n$ | 1 | 2 | 4  | 6  | 9  | 13 | 18 | 24 | 31 | 39 | 49 |
| $E_n$ | 1 | 2 | 4  | 6  | 9  | 13 | 18 | 24 | 31 | 39 | 50 |

# An Introduction to Mathematical Reasoning

numbers, sets and functions

PETER J. ECCLES



An Introduction to

# **Mathematical Reasoning**



# An Introduction to **Mathematical Reasoning**

## numbers, sets and functions

Peter J. Eccles  
*Department of Mathematics  
University of Manchester*



CAMBRIDGE UNIVERSITY PRESS  
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo,  
Delhi, Dubai, Tokyo

Cambridge University Press  
The Edinburgh Building, Cambridge, CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

[www.cambridge.org](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9780521597180](http://www.cambridge.org/9780521597180)

© Cambridge University Press 2007

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without  
the written permission of Cambridge University Press.

First published 1997  
Thirteenth printing 2010

Printed in the United Kingdom at the University Press, Cambridge

*A catalogue record of this publication is available from the British Library*

*Library of Congress Cataloging in Publication data*  
Eccles, Peter J., 1945-

An introduction to mathematical reasoning: lectures on numbers, sets,  
and functions / Peter J. Eccles.

p. cm.

Includes bibliographical references and index.

ISBN 0 521 59269 0 hardback – ISBN 0 521 59718 8 paperback

1. Proof theory. I. Title.

QA9.54.E23 1997

511.3—dc21 97-11977 CIP

ISBN 978-0-521-59269-7 hardback

ISBN 978-0-521-59718-0 paperback

Cambridge University Press has no responsibility for the persistence or accuracy  
of URLs for external or third-party internet websites referred to in this publication,  
and does not guarantee that any content on such websites is, or will remain,  
accurate or appropriate. Information regarding prices, travel timetables and other  
factual information given in this work are correct at the time of first printing but

Cambridge University Press does not guarantee the accuracy of such  
information thereafter.

I too will something make  
And joy in the making;  
Altho' tomorrow it seem  
Like the empty words of a dream  
Remembered on waking.

Robert Bridges, *Shorter poems*.

In loving memory of

Elizabeth Baron and John Baron

(Auntie Lizzie and Uncle Jack)

and those Woodside Bank summers

# Contents

|   |   |     |
|---|---|-----|
| <i>Preface</i>                                    | <i>page</i>                             | ix  |
| <b>Part I: Mathematical statements and proofs</b> |   |     |
| 1   | The language of mathematics             | 3   |
| 2   | Implications                            | 10  |
| 3   | Proofs                                  | 21  |
| 4   | Proof by contradiction                  | 30  |
| 5   | The induction principle                 | 39  |
|   | Problems I                              | 53  |
| <b>Part II: Sets and functions</b>                |   |     |
| 6   | The language of set theory              | 61  |
| 7   | Quantifiers                             | 74  |
| 8   | Functions                               | 89  |
| 9   | Injections, surjections and bijections  | 101 |
|   | Problems II                             | 115 |
| <b>Part III: Numbers and counting</b>             |   |     |
| 10  | Counting                                | 123 |
| 11  | Properties of finite sets               | 133 |
| 12  | Counting functions and subsets          | 144 |
| 13  | Number systems                          | 157 |
| 14  | Counting infinite sets                  | 170 |
|   | Problems III                            | 182 |
| <b>Part IV: Arithmetic</b>                        |   |     |
| 15  | The division theorem                    | 191 |
| 16  | The Euclidean algorithm                 | 199 |
| 17  | Consequences of the Euclidean algorithm | 207 |
| 18  | Linear diophantine equations            | 216 |
|   | Problems IV                             | 225 |

|    |   |     |
|----|---|-----|
|    | <b>Part V: Modular arithmetic</b>                   | 229 |
| 19 | Congruence of integers                              | 231 |
| 20 | Linear congruences                                  | 240 |
| 21 | Congruence classes and the arithmetic of remainders | 250 |
| 22 | Partitions and equivalence relations                | 262 |
|    | Problems V  | 271 |
|    | <b>Part VI: Prime numbers</b>                       | 275 |
| 23 | The sequence of prime numbers                       | 277 |
| 24 | Congruence modulo a prime                           | 289 |
|    | Problems VI   | 295 |
|    | <i>Solutions to exercises</i>                       | 299 |
|    | <i>Bibliography</i>                                 | 345 |
|    | <i>List of symbols</i>                              | 346 |
|    | <i>Index</i>  | 347 |

# Preface

This book is based on lectures given at the Victoria University of Manchester to first year Honours Mathematics students including those taking joint or combined degrees. In common with most other British mathematics departments, the Manchester mathematics department has thoroughly reviewed its curriculum in recent years in the attempt to meet more adequately the needs of students who have experienced the effects of the great changes in the teaching of mathematics in schools, as well as the increased numbers of ‘mature students’ and students from non-standard backgrounds.

It was clear to us at the University of Manchester that we should completely rethink and broaden our curriculum, including material which we had previously expected students to know on entry to the course but also including introductory material on combinatorics, computer skills and numerical mathematics, as well as encouraging the development of problem solving skills.

A key ingredient of this new University of Manchester curriculum is a module on Mathematical Reasoning whose purpose is to introduce the basic ideas of mathematical proof and to develop skills in writing mathematics, helping to bridge the gap between school and university mathematics. This book is based on this course module. The ability to write correct and clear mathematics is a skill which has to be acquired by observing experienced practitioners at work in lectures and tutorials, by learning to appreciate the details of mathematical exposition in books, and by a great deal of practice. It is a skill readily transferable to many other areas and its acquisition is likely to be one of the main benefits of a mathematics degree course for most students who may well make no further use of most of the specific mathematical content of the course!

There is no absolutely *correct* way of writing out a given proof. For

example, it is necessary to take into account who the intended reader is. But, whoever the reader is, the proof should be expressed as clearly as possible and to achieve this the writer needs to understand the logic of the proof. Writing a proof is not separate from discovering the proof in the way that writing up a scientific experiment is separate from carrying out the experiment or performing a piece of music is separate from composing it. Attempts to write out a proof are an important part of the discovery process. Alison Leonard has written in a totally different context:

Not only are human ideas conveyed by language, they are actually formed by the language available to us.<sup>†</sup>

So it is in mathematics, where we find a parallel development of mathematical ideas and mathematical language.

This is reflected in this book. The emphasis is on helping the reader to understand and construct proofs and to learn to write clear and concise mathematics. This can only be achieved by exploring some particular mathematical topics and the contexts chosen are set theory, combinatorics and number theory. These topics

- provide good examples to illustrate a range of basic methods of proof, in particular proof by induction and proof by contradiction,
- include some fundamental ideas which are part of the standard tool kit of any mathematician, such as functions and inverses, the binomial theorem, the Euclidean algorithm, the pigeonhole principle, the fundamental theorem of arithmetic, and congruence,
- build on ideas met in early schooling illustrating ways in which familiar ideas can be formulated rigorously, for example counting or the greatest common divisor,
- include some of the all time great classic proofs, for example the Euclidean proofs of the irrationality of root 2 and the infinity of primes, but also ideas from throughout mathematical history so that there is an opportunity to present mathematics as a continually developing subject.

Roughly speaking, the first three parts of the book are about the basic language of mathematics and the final three parts are about number theory, illustrating how the ideas of the earlier parts are applied to some significant mathematics. The reader may find that the later parts

<sup>†</sup> See Alison Leonard, *Telling our stories: wrestling with a fresh language for the spiritual journey*, Dorton, Longman and Todd, 1995.

contain some more straightforward material than the early parts simply because there is material on problem solving techniques which can then be practised on specific numerical examples. The topics selected for these later parts, the Euclidean algorithm, modular arithmetic and prime numbers, include material from the whole of mathematical history from classical Greek times to the present day.

I would encourage the reader in working through the first three parts not to expect to understand everything at first. Part I introduces various forms of mathematical statements and the standard methods of proof. Proof by contradiction and proof by induction are explained in detail and these methods are then used again and again throughout the remainder of the book: so many great theorems are proved by contradiction and some are included in this book. Part II covers the basic material on sets and functions which provides the language in which much mathematics is best expressed. It includes a leisurely discussion of universal and existential statements which are so important in university mathematics, particularly analysis or advanced calculus. This is not a book on mathematical logic but inevitably some ideas from the beginnings of that subject are included in these first two parts. The reader may find some of the material on counting in Part III to be more difficult than the rest of the book. It is best not to become discouraged by this but if necessary to move on to Part IV which is probably the easiest part but also includes some very striking and attractive results, returning to Part III when more familiarity with the language of sets and functions has been acquired. This material on counting provides good practice in using the language of sets and functions in a very familiar context and also illustrates how a familiar process may be made mathematically precise and how this then enables the process to be extended to a less familiar context, counting infinite sets.

The book is divided into twenty four chapters and grew out of a series of twenty four lectures. However, there is far more material in most chapters than could reasonably be covered in a single lecture. A lecture course based on this book would need to be selective covering either a subset of the chapters or, more likely, a subset of the material in most or all of the chapters as the author's lecture courses have done.

There is a great range of ability, experience and knowledge amongst students embarking on university mathematics courses and a real attempt has been made here to provide material meeting the needs of weaker or ill-prepared students whilst at the same time providing something which will interest and challenge the most able students. This is

achieved by including material and, in particular, problems of varying difficulty; many of the over 250 problems are routine and computational but others are quite demanding. In the later stages of the book some significant developments of the ideas in the book are approached in the problems.

Each chapter concludes with a number of exercises many of which are very closely related to material in the chapter and intended to be relatively straightforward and routine. Occasionally something a little harder is included. The reader is encouraged to work through all of these and full solutions for them are given at the back of the book. There are also six sets of problems, some similar to the exercises in order to consolidate the techniques involved but others more wide ranging and challenging. I hope that every reader will find some of the exercises and problems to be fun.

Mathematicians are not good at encouraging students to read around the subject but the intention here is that material in this book not covered in a lecture course will provide additional reading particularly for stronger students and, through the references given, lead to wider study in some areas. I have tried to write the book which I would have welcomed for my own students.

Finally, I am only too conscious that anyone who writes a book about how to write mathematics well lays themselves open to ridicule when the proofs in the book are found to be confused or inadequate. If the reader does find failings in the proofs in this book then I hope that acquiring the ability to see these failings will be seen as a useful step in developing that self-criticism which is necessary for the writing of clear and beautiful mathematics.

**Acknowledgements.** The author is indebted to very many people who have knowingly or unknowingly influenced the material in this book or who have provided specific advice. I would in particular like to acknowledge the contributions of the following: Pyotr Akhmet'ev, Michael Barratt, Francis Coghlan, Mark Eccles, Michael Eccles, Pamela Eccles, Douglas Gregory, Brian Hartley, Martin Huxley, John King-Hele, E. Makin, Mick McCrudden, Jeff Paris, Mike Prest, Nigel Ray, John Reade, András Szűcs, Grant Walker, George Wilmers, J.R. Winn and Reg Wood. In addition I wish to thank the staff of Cambridge University Press for their careful editorial work and an anonymous referee whose report on a preliminary version was extremely helpful.

Peter J. Eccles, Manchester, June 1997.

# **Part I**

## Mathematical statements and proofs



# 1

## The language of mathematics

Pure mathematics is concerned with the exploration of mathematical concepts arising initially from the study of space and number. In order to capture and communicate mathematical ideas we must make statements about mathematical objects and much mathematical activity can be described as the formulation of mathematical statements and then the determination of whether or not such statements are true or false. It is important to be clear about what constitutes a mathematical statement and this is considered in this first chapter. We begin with simple statements and then examine ways of building up more complicated statements.

### 1.1 Mathematical statements

It is quite difficult to give a precise formulation of what a mathematical statement is and this will not be attempted in this book. The aim here is to enable the reader to recognize simple mathematical statements. First of all let us consider the idea of a *proposition*. A good working criterion is that *a proposition is a sentence which is either true or false (but not both)*. For the moment we are not so concerned about whether or not propositions are in fact true. Consider the following list.

- (i)  $1 + 1 = 2$ .
- (ii)  $\pi = 3$ .
- (iii) 12 may be written as the sum of two prime numbers.
- (iv) Every even integer greater than 2 may be written as the sum of two prime numbers.
- (v) The square of every even integer is even.
- (vi)  $n$  is a prime number.

- (vii)  $n^2 - 2n > 0$ .
- (viii)  $m < n$ .
- (ix)  $12 = 11$ .
- (x)  $\pi$  is a special number.

Of these the first five are propositions. This says nothing about whether or not they are true. In fact (i) is true and (ii) is false. Proposition (iii) is true since  $12 = 7 + 5$  and 5 and 7 are prime numbers.<sup>†</sup> Propositions (iv) and (v) are general statements which cannot be proved by this sort of simple arithmetic. In fact it is easy to give a general proof of (v) (see Exercise 3.3) showing that it is true. However, at the time of writing, it is unknown whether (iv) is true or false; this statement is called the Goldbach conjecture after Christian Goldbach who suggested that it might be true in a letter to Leonhard Euler written in 1742.

On the other hand the others on the above list are not propositions. The next two, (vi) and (vii), become propositions once a numerical *value* is assigned to  $n$ . For example if  $n = 2$  then (vi) is true and (vii) is false, whereas if  $n = 3$  then they are both true. The next, (viii), becomes a proposition when values are assigned to both  $m$  and  $n$ , for example it is true for  $m = 2, n = 3$  and false for  $m = 3, n = 2$ . Sentences of this type are called *predicates*. The symbols which need to be given values in order to obtain a proposition are called *free variables*.

The word *statement* will be used to denote either a proposition or a predicate. So in the above list the first eight items are statements. We will use a single capital letter  $P$  or  $Q$  to indicate a statement, or sometimes an expression like  $P(m, n)$  to indicate a predicate, with the free variables listed in brackets.

The last two entries in the above list are not statements: (ix) is not even a sentence and (x) doesn't mean anything until we know what 'special' means. Very often mathematicians do give technical meanings to everyday words (as in 'prime' number used in (iii) and (iv) above and 'even' number used in (v)) and so if 'special' had been given such a meaning as a possible property of a number then (x) would be a proposition and so a statement. Of course the fact that (i) to (viii) are statements relies on a number of assumptions about the meanings of the symbols and words which have been taken for granted.

<sup>†</sup> A positive integer is a *prime number* if it is greater than 1 and is divisible only by 1 and itself. Thus the first few prime numbers are 2, 3, 5, 7, 11, .... This definition is discussed in Chapter 23.

In particular it is quite complicated to give a precise definition of the number  $\pi$  in statement (ii). The number  $\pi$  was originally defined geometrically as the ratio between the length of the circumference of a circle and the length of the diameter of the circle. In order for this definition to be justified it is necessary to define the length of the circumference, a curved line, and having done that to prove that the ratio is independent of the size of the circle. This was achieved by the Greeks in classical times. A regular hexagon inscribed in a circle of unit diameter has circumference 3 and so it is clear that  $\pi > 3$ .† Archimedes showed by geometrical means in his book *On the measurement of the circle* that  $3\frac{10}{71} < \pi < 3\frac{1}{7}$  and also provided a beautiful proof of the formula  $\pi r^2$  for the area of a circle of radius  $r$ . The modern definition of  $\pi$  is usually as twice the least positive number for which the cosine function vanishes, and the details may be found in any text on analysis or advanced calculus.‡

## 1.2 Logical connectives

In mathematics we are often faced with deciding whether some given proposition is true or whether it is false. Many statements are really quite complicated and are built up out of simpler statements using various ‘logical connectives’. For the moment we restrict ourselves to the simplest of these: ‘or’, ‘and’ and ‘not’. The truth or falsehood of a complicated statement is determined by the truth or falsehood of its component statements. It is important to be clear how this is done.

### *The connective ‘or’*

Suppose that we say

For integers  $a$  and  $b$ ,  $ab = 0$  if  $a = 0$  or  $b = 0$ .

The statement ‘ $a = 0$  or  $b = 0$ ’ is true if  $a = 0$  (regardless of the value of  $b$ ) and is also true if  $b = 0$  (regardless of the value of  $a$ ). Notice that the statement is true if both  $a = 0$  and  $b = 0$  are true. This is called

† Some fundamentalist Christians have claimed that 1 Kings vii 23, in which a round object of diameter 10 cubits is stated to have a circumference of 30 cubits, ‘proves’ that  $\pi = 3$  but mathematical truth can never be a matter simply of faith (nor can religious truth either in the author’s experience) and in any case there is no suggestion in the biblical passage that the measurements were highly accurate!

‡ See for example R. Haggerty, *Fundamentals of mathematical analysis*, Addison-Wesley, Second edition 1993.

the ‘inclusive’ use of ‘or’. Its meaning is best made precise by means of a *truth table*. Any statement may be either true or false: we say that ‘true’ and ‘false’ are the two possible *truth values* for the statement. So given two statements  $P$  and  $Q$  each has two possible truth values giving four possible combinations in all. The truth table for ‘or’ which follows specifies the truth value for ‘ $P$  or  $Q$ ’ corresponding to each possible combination of truth values for  $P$  and for  $Q$ , one line for each. In the table  $T$  indicates ‘true’ and  $F$  indicates ‘false’.

**Table 1.2.1**

| $P$ | $Q$ | $P$ or $Q$ |
|-----|-----|------------|
| $T$ | $T$ | $T$        |
| $T$ | $F$ | $T$        |
| $F$ | $T$ | $T$        |
| $F$ | $F$ | $F$        |

In this table, the truth values in the second line for example indicate that if  $P$  is true and  $Q$  is false then the statement ‘ $P$  or  $Q$ ’ is true.

‘ $P$  or  $Q$ ’ is called the *disjunction* of the two statements  $P$  and  $Q$ .

In everyday speech ‘or’ is often used in the exclusive sense as in the first sentence of this section: ‘we are faced with deciding whether some given proposition is true or whether it is false’, in which it is implicitly understood (and emphasized by the uses of the word ‘whether’) that it is not possible for a proposition to be both true and false.

To give another everyday example of the ‘exclusive or’, when we say ‘everyone will travel there by bus or by train’ this would normally be taken to mean that everyone uses one or other form of transport but not both. If we wanted to allow for someone using both we would probably say ‘everyone will travel there by bus or by train or by both’. In practice the precise meaning of ‘or’ is made clear by the context or there is some ambiguity (but not both!). The meaning of mathematical statements must be precise and so we avoid these ambiguities by always using ‘or’ in the inclusive way determined by the above truth table.

The connective ‘or’ is sometimes hidden in other notation. For example when we write ‘ $a \leq b$ ’ where  $a$  and  $b$  are real numbers this is a shorthand for ‘ $a < b$  or  $a = b$ ’. Thus both the statements ‘ $1 \leq 2$ ’ and ‘ $2 \leq 2$ ’ are true.

Similarly ‘ $a = \pm b$ ’ is a shorthand for ‘ $a = b$  or  $a = -b$ ’. There is some possibility of confusion here for the true statement that  $1 = \pm 1$  does not mean that 1 and  $\pm 1$  are interchangeable: ‘if  $x^2 = 1$  then  $x = \pm 1$ ’ is a

true statement whereas ‘if  $x^2 = 1$  then  $x = 1$ ’ is a false statement! It is necessary to be clear that  $a = \pm b$  is not asserting that a single equality is true but is asserting that one (or both) of two equalities  $a = b$ ,  $a = -b$  is true. The following truth table illustrating this may help.

| $a$ | $b$ | $a = b$ | $a = -b$ | $a = \pm b$ |
|-----|-----|---------|----------|-------------|
| 1   | 2   | F       | F        | F           |
| 1   | 1   | T       | F        | T           |
| 2   | -2  | F       | T        | T           |
| 0   | 0   | T       | T        | T           |

Observe how the truth values in the final column may be read off from the truth values in the previous two columns using the truth table for ‘or’ (Table 1.2.1).

In fact the statement  $a = \pm b$  can be written as a single equation, namely  $|a| = |b|$  where  $|a|$  denotes the *absolute value* of  $a$  (i.e.  $|a| = a$  if  $a \geq 0$  and  $|a| = -a$  if  $a \leq 0$ ).

### *The connective ‘and’*

We use this when we wish to assert that two things are both true. Again this word can be hidden in other words or notation. Thus if we assert that

$\pi$  lies between 3 and 4

or in symbols

$3 < \pi < 4$

then this means that

$\pi > 3$  and  $\pi < 4$

which is called the *conjunction* of the two statements ‘ $\pi > 3$ ’ and ‘ $\pi < 4$ ’.

You should construct a truth table for ‘and’.

### *The connective ‘not’*

Finally we have the idea of the *negation* of a statement. The negation of a statement is true when the original statement is false and it is false when the original statement is true. This is described by the following truth table.

**Table 1.2.2**

| $P$ | not $P$ |
|-----|---------|
| T   | F       |
| F   | T       |

The negation of a statement is usually obtained by including the word ‘not’ in the statement but does require a little care since this is often used very loosely in everyday speech.

**Example 1.2.3** Consider the following statement about a polynomial  $f(x)$  with real coefficients, such as  $x^2 + 3$  or  $x^3 - x^2 - x$ .

- (i) For real numbers  $a$ , if  $f(a) = 0$  then  $a$  is positive (i.e.  $a > 0$ ).

What is the negation of this statement? The following possibilities spring to mind from the everyday use of ‘negation’.

- (ii) For real numbers  $a$ , if  $f(a) = 0$  then  $a$  is negative.
- (iii) For real numbers  $a$ , if  $f(a) = 0$  then  $a$  is non-positive.
- (iv) For real numbers  $a$ , if  $f(a) \neq 0$  then  $a$  is positive.
- (v) For real numbers  $a$ , if  $a$  is positive then  $f(a) = 0$ .
- (vi) For real numbers  $a$ , if  $a$  is positive then  $f(a) \neq 0$ .
- (vii) For real numbers  $a$ , if  $a$  is non-positive then  $f(a) \neq 0$ .
- (viii) For some non-positive real number  $a$ ,  $f(a) = 0$ .

If you think about it you will see that none but the last is the negation. For example, if  $f(x) = x^3 - x = x(x+1)(x-1)$  then statement (i) is certainly false since, for the real number 0,  $f(0) = 0$  but 0 is not positive. On the other hand so are all the other statements apart from (viii) which is true since 0 is a non-positive number such that  $f(0) = 0$ . We will consider the negation of statements of this form more formally in the next chapter.

### Exercises

**1.1** Construct a truth table for ‘and’ as follows.

| $P$ | $Q$ | $P$ and $Q$ |
|-----|-----|-------------|
| T   | T   |             |
| T   | F   |             |
| F   | T   |             |
| F   | F   |             |

**1.2** Construct truth tables for the statements

- (i) not ( $P$  and  $Q$ );
- (ii) (not  $P$ ) or (not  $Q$ );
- (iii)  $P$  and (not  $Q$ );
- (iv) (not  $P$ ) or  $Q$ .

**1.3** Using the truth table for ‘or’ complete the following truth table for the statement  $a \leq b$ .

| $a$ | $b$ | $a < b$ | $a = b$ | $a \leq b$ |
|-----|-----|---------|---------|------------|
| 1   | 1   |         |         |            |
| 2   | 1   |         |         |            |
| 1   | 2   |         |         |            |
| -2  | 1   |         |         |            |

**1.4** Consider the following statement.

- (i) All girls are good at mathematics.

Which of the following statements is the negation of the above statement?

- (ii) All girls are bad at mathematics.
- (iii) All girls are not good at mathematics.
- (iv) Some girl is bad at mathematics.
- (v) Some girl is not good at mathematics.
- (vi) All children who are good at mathematics are girls.
- (vii) All children who are not good at mathematics are boys.

Can you find any statements in this list which have the same meaning as the original statement (i)?

**1.5** Prove that  $|a|^2 = a^2$  for every real number  $a$ .

# 2

## Implications

In the first chapter we were mainly interested in the meaning of mathematical statements. However, mathematics is primarily concerned with establishing the truth of statements. This is achieved by giving a proof of the statement. The key idea in most proofs is that of implication and this idea is discussed in this chapter.

### 2.1 Implications

A proof is essentially a sequence of statements starting from statements we know† to be true and finishing with the statement to be proved. Each statement is true because the earlier statements are true. The justification for such steps usually makes use of the idea of ‘implication’; an implication is the assertion that if one particular statement is true then another particular statement is true.

The symbol usually used to denote implication in pure mathematics‡ is  $\Rightarrow$  although there are a variety of forms of words which convey the same meaning. For the moment we can think of ' $P \Rightarrow Q$ ' as asserting that if statement  $P$  is true then so is statement  $Q$ , which is often read as ' $P$  implies  $Q$ '. The meaning will be made precise by means of a truth table. Before doing this it is necessary to clarify what this meaning should be and to do this we consider an example concerning an integer  $n$ .

Suppose that  $P(n)$  is the statement ' $n > 3$ ' and  $Q(n)$  is the statement

† See the section on ‘mathematical truth’ at the end of this chapter.

‡ In mathematical logic the symbol  $\rightarrow$  is usually used instead of  $\Rightarrow$ . Other symbols used in mathematical logic are ' $P \vee Q$ ' for ' $P$  or  $Q$ ', ' $P \wedge Q$ ' for ' $P$  and  $Q$ ', and ' $\neg P$ ' or ' $\sim P$ ' for 'not  $P$ '.

' $n > 0$ ', where  $n$  is an integer. Then the statement  $P(n) \Rightarrow Q(n)$ , i.e.

$$n > 3 \Rightarrow n > 0,$$

is the assertion that if an integer  $n$  is greater than 3 then it is greater than 0, which is certainly a true statement. Notice that this statement is true for each integer  $n$  even though the statement ' $n > 3$ ' is true for some values of  $n$  and false for others as is the statement ' $n > 0$ '. Let us consider the possibilities.

|         | $P(n)$ | $Q(n)$ |
|---------|--------|--------|
| $n < 0$ | F      | F      |
| $n = 0$ | F      | F      |
| $n = 1$ | F      | T      |
| $n = 2$ | F      | T      |
| $n = 3$ | F      | T      |
| $n > 3$ | T      | T      |

This table demonstrates that if it is true that  $n > 3$  then it is indeed true that  $n > 0$ . But if  $n \not> 3$  (i.e.  $P(n)$  is false) then we may have  $n > 0$  (so  $Q(n)$  is true) for example when  $n = 2$ , and we may have  $n \not> 0$  (so  $Q(n)$  is false) for example when  $n = -2$ . However, there is no value of  $n$  for which  $P(n)$  is true and  $Q(n)$  is false; and this is precisely what we mean when we say that if  $P(n)$  is true then  $Q(n)$  is true. We can summarize this in the following truth table.

**Table 2.1.1**

| $P$ | $Q$ | $P \Rightarrow Q$ |
|-----|-----|-------------------|
| T   | T   | T                 |
| T   | F   | F                 |
| F   | T   | T                 |
| F   | F   | T                 |

In an implication  $P \Rightarrow Q$ , the statement  $P$  is called the *hypothesis* or *antecedent* and the statement  $Q$  is called the *conclusion* or *consequent*.

Here is another illustration. The assertion

$$n = 1 \Rightarrow (n - 1)(n - 2) = 0 \quad (\text{for integers } n)$$

is certainly true, for if we substitute  $n = 1$  into  $(n - 1)(n - 2)$  we obtain  $(1 - 1)(1 - 2) = 0 \times (-1) = 0$ . If we write  $P(n)$  for the statement ' $n = 1$ ' and  $Q(n)$  for the statement ' $(n - 1)(n - 2) = 0$ ' we can consider the

truth or falsehood of these two statements for different values of  $n$  as follows.

|               | $P(n)$ | $Q(n)$ |
|---------------|--------|--------|
| $n = 1$       | $T$    | $T$    |
| $n = 2$       | $F$    | $T$    |
| $n \neq 1, 2$ | $F$    | $F$    |

Thus  $n = 1$  illustrates line 1 of Table 2.1.1,  $n = 2$  illustrates line 3, and  $n \neq 1, 2$  illustrates line 4. Again there is no value of  $n$  for which  $P(n)$  is true and  $Q(n)$  is false.

It is important to realize that in mathematics we use the idea of implication in this particularly precise way (as defined by the truth table) extending everyday usage. Consider the following statements.

- (i)  $(\pi < 4) \Rightarrow (1 + 1 = 2)$
- (ii)  $(\pi < 4) \Rightarrow (1 + 1 = 3)$
- (iii)  $(\pi < 3) \Rightarrow (1 + 1 = 2)$
- (iv)  $(\pi < 3) \Rightarrow (1 + 1 = 3)$

From the truth table it is easily seen that the second of these statements is false whereas the other three are true (since  $3 < \pi < 4$ ). However, these statements all seem a little odd from an everyday point of view since the size of the number  $\pi$  really has nothing to do with the effect of adding the number 1 to itself. In everyday speech when we say that  $P$  implies  $Q$  this suggests that the statement  $P$  causes the statement  $Q$  to be true. The idea of causation is difficult to make precise and so it is eliminated from the mathematical use of the word ‘implies’. This difference from everyday usage is clarified in mathematical logic where a statement of the form  $P \Rightarrow Q$  is referred to as a *conditional statement* rather than an implication. In this book we have adopted the language used by most mathematicians.

Another way of motivating our use of implication is to consider its negative, the statement that  $P$  does not imply  $Q$ , usually written  $P \not\Rightarrow Q$ . This will be true precisely when  $P \Rightarrow Q$  is false and from the truth table this is when  $P$  is true and  $Q$  is false. If you think about it these are precisely the circumstances when we would wish to say that  $P$  does not imply  $Q$ . Notice that this means that the statements ' $P \not\Rightarrow Q$ ' and ' $P$  and (not  $Q$ )' are logically equivalent, i.e. have the same meaning (cf. the truth table for ' $P$  and (not  $Q$ )' in the solution to Exercise 1.2).

Using the fact that '(not  $P$ ) or  $Q$ ' is the negation of ' $P$  and (not  $Q$ )' we can say this in a different way: the statements ' $P \Rightarrow Q$ ' and '(not

$P$ ) or  $Q'$  are logically equivalent (again see Exercise 1.2). This logical equivalence does arise in everyday speech. For example, when we say ‘Read the lecture notes or you won’t understand the lecture’ this has the same meaning as ‘If you don’t read the lecture notes then you won’t understand the lecture.’ Here  $P$  is the statement ‘you don’t read the lecture notes’ and  $Q$  is the statement ‘you won’t understand the lecture’.

### *Universal implications*

Strictly speaking the statements ‘ $n > 3 \Rightarrow n > 0$ ’ and ‘ $n = 1 \Rightarrow (n - 1)(n - 2) = 0$ ’ are predicates and give different propositions for each value of  $n$ . However, we often use statements of this form in the way we have been doing to mean that the predicate leads to a true proposition whatever value is assigned to the free variable  $n$  – this is called a *universal* statement. The possible range of values for the free variables may be explained in the text by a phrase like ‘for integers  $n$ ’ or ‘where  $n$  is an integer’, or it may be intended to be clear from the context; it is usually good practice to make the range of values explicit.

A statement of this type is false if it is not a universal statement, in other words there is at least one possible value for the variable for which it fails. Consider the statement ‘ $x > 0 \Rightarrow x \geq 1$ ’ for real numbers  $x$ . We can construct a truth table as follows.

**Table 2.1.2**

|             | $x > 0$ | $x \geq 1$ | $x > 0 \Rightarrow x \geq 1$ |
|-------------|---------|------------|------------------------------|
| $x \leq 0$  | F       | F          | T                            |
| $0 < x < 1$ | T       | F          | F                            |
| $x \geq 1$  | T       | T          | T                            |

Here the entries in the last column are determined by the entries in the previous two columns using the truth table for implication, Table 2.1.1. We see from this that there are values of  $x$  for which  $x > 0$  and  $x \not\geq 1$ , for example  $x = 1/2$ , so that  $x > 0 \Rightarrow x \geq 1$  is not universally true. This means that, for real numbers  $x$ ,  $x > 0$  does not imply that  $x \geq 1$ , which may be written

$$x > 0 \not\Rightarrow x \geq 1.$$

This is an *existence* statement: it claims that there is a value of  $x$  for which the hypothesis is true and the conclusion false.

There will be a fuller discussion of universal statements and existence statements in Chapter 7.

Notice that in these examples we have adopted the common practice of using letters in the middle of the alphabet (such as  $n$ ) to denote integers and letters at the end of the alphabet (such as  $x$ ) to denote real numbers.

If we now reconsider Example 1.2.3 we see that statement (i) there is the universal implication

$$f(a) = 0 \Rightarrow a > 0 \quad (\text{for real numbers } a).$$

The negation of this is the assertion that the implication is false for some real number  $a$ , in other words  $f(a) = 0$  and  $a \not> 0$  for some real number  $a$ . This is precisely the statement (viii) which we identified as the negation.

### *Reading implications*

There are many different ways of reading the statement  $P \Rightarrow Q$ , which may also be written  $Q \Leftarrow P$ , and some of the most common are listed below.

- (i) If  $P$  then  $Q$ .
- (ii)  $P$  implies  $Q$ .
- (iii)  $Q$  if  $P$ .
- (iv)  $P$  only if  $Q$ .
- (v)  $Q$  whenever  $P$ .
- (vi)  $P$  is sufficient for  $Q$ .
- (vii)  $Q$  is necessary for  $P$ .

Take care with the third and fourth of these and also with the last two for it is important to appreciate the difference between  $P \Rightarrow Q$  and  $Q \Rightarrow P$ . It is quite possible for one of these to hold without the other. For example, let us return to the examples already considered. We have seen that, for integers  $n$ ,  $n > 3 \Rightarrow n > 0$ ; but  $n > 0 \not\Rightarrow n > 3$  since, for example, when  $n = 1$ , the hypothesis  $n > 0$  is true whereas the conclusion  $n > 3$  is false. Similarly,  $n = 1 \Rightarrow (n - 1)(n - 2) = 0$  but  $(n - 1)(n - 2) = 0 \not\Rightarrow n = 1$  since, when  $n = 2$ ,  $(n - 1)(n - 2) = 0$  but  $n \neq 1$ .

The statement  $Q \Rightarrow P$  is called the *converse* of the statement  $P \Rightarrow Q$ . As an illustration, in Example 1.2.3, statement (v) is the converse of (i) (and in addition (iv) and (vi) are converse to each other).

When both implications do hold we write  $P \Leftrightarrow Q$ . Thus we define

$$P \Leftrightarrow Q \text{ means } (P \Rightarrow Q) \text{ and } (Q \Rightarrow P).$$

From the truth tables we see that this means that either  $P$  and  $Q$  are both true or they are both false (see Exercise 2.3).

We read  $P \Leftrightarrow Q$  as follows.

- (i)  $P$  is equivalent to  $Q$ .
- (ii)  $P$  is necessary and sufficient for  $Q$ .
- (iii)  $P$  if and only if  $Q$  (sometimes written  $P$  iff  $Q$ ).
- (iv)  $P$  precisely when  $Q$ .

## 2.2 Arithmetic

The basic principles of mathematical reasoning apply throughout mathematics but in this book we are going to explore them mainly in the context of number theory or arithmetic since the basic ideas here are familiar from early schooldays. For the moment it is assumed that the reader is familiar with the *integers*, both positive and negative, i.e. the numbers  $\dots, -2, -1, 0, 1, 2, 3, \dots$ , and their basic arithmetic properties (under addition and multiplication) and order properties. We will also make use of the *rational numbers* (or fractions) and the *real numbers* (or infinite decimals) and these will be discussed in more detail in Chapter 13.

As we do mathematics new vocabulary may be introduced by means of a *definition*. Quite often this gives a precise meaning to an everyday word. Then when we wish to understand the meaning of a statement involving that word it is necessary to invoke the definition. Humpty Dumpty's principle could well stand at the head of any piece of mathematical writing.

'When *I* use a word,' Humpty Dumpty said, in rather a scornful tone, 'it means what *I* choose it to mean – neither more nor less.'

Lewis Carroll, *Through the looking glass and what Alice found there*.

This reminds us that when we meet a new word, or a familiar word in a new setting, we need to discover what the writer means by it, not to decide how *we* might have defined it!

Here is a familiar idea to illustrate this.

*Odd and even integers*

**Definition 2.2.1** Given two integers  $a$  and  $b$ , we say that  $b$  divides  $a$  or  $a$  is a multiple of  $b$  to mean that there is an integer  $q$  such that  $a = bq$ .

Thus for example 3 divides 6 since  $6 = 3 \times 2$ ,  $-14$  divides 28 since  $28 = (-14) \times (-2)$ , and  $b$  divides 0 whatever the value of the integer  $b$  since  $0 = b \times 0$ .

**Definition 2.2.2** Given an integer  $a$  we say that  $a$  is even to mean that 2 divides  $a$ .

**Definition 2.2.3** Given an integer  $a$  we say that  $a$  is odd to mean that it is not even.

Notice how the third definition makes use of the second and the second definition makes use of the first. It is very common to get chains of definitions; you have to work back through the chain to make use of them.

Now consider the following statement.

**Proposition 2.2.4** 101 is an odd integer.

Of course, you know that this is true. But what we are interested in here is how we might prove it from the arithmetic and order properties of the integers and our definition of the word ‘odd’. If we cannot prove it from the definition then our definition isn’t much use.

Incidentally, the word ‘Proposition’ used in this way indicates that I am claiming that this result is true. It is really a synonym for ‘Theorem’ but that word is usually restricted to results of greater significance than this one.

*Proof* From Definition 2.2.3, ‘101 is odd’ means the same as ‘101 is not even’ which from Definition 2.2.2 means that 2 does not divide 101. By Definition 2.2.1, this means that there is no integer  $q$  such that  $2q = 101$ . Now we clearly cannot prove this by exhausting all the possibilities one at a time. However, for an integer  $q$ , either  $q \leq 50$  or  $q \geq 51$ . Hence either  $2q \leq 100$  or  $2q \geq 102$ , so that  $2q \neq 101$ . Hence 101 is not even and so is odd as required.  $\square$

The symbol  $\square$  at the end of this proof is there simply to mark the completion of the proof. I recommend that proofs are concluded by stating what has been proved as occurs here.

Notice that the key step in the proof is provided by the two universal implications

$$q \leq 50 \Rightarrow 2q \leq 100 \quad (\text{for integers } q)$$

and

$$q \geq 51 \Rightarrow 2q \geq 102 \quad (\text{for integers } q).$$

This proof might seem over-elaborate, but if you think about it you will see that it is simply spelling out a simple reason why 101 is an odd number. You might like to try to find a better proof based simply on the above definitions.<sup>†</sup>

### 2.3 Mathematical truth

It is reasonable to ask how we can get started in establishing the truth of mathematical statements if the only method of doing this is to prove them starting from other mathematical statements known to be true.

When the deductive method was first used, for example in the *Elements* of Euclid written about 300 B.C., proofs began from certain ‘axioms’ or ‘postulates’ which were viewed as self-evident truths. For example, Euclid’s Postulate 1 may be reformulated as ‘Given two distinct points, there is a unique straight line passing through them’ which most people would consider to be obvious. This approach sees mathematics as a body of facts about real objects: points, lines, numbers. These facts are determined by using certain accepted self-evident rules of deduction from certain accepted obvious facts, the axioms. There are a number of philosophical difficulties about this point of view. For one thing mathematical objects appear to be ideal abstract representations of objects in the real world: for example you can’t see a point and can only see a blob having some size. For another it has been discovered that different sets of axioms are possible and give rise to different theories which appear to be equally valid mathematically. This was first realized in geometry when in the early nineteenth century ‘non-Euclidean’ geometries were discovered. We can then ask what is the ‘true’ geometry of the universe.

<sup>†</sup> You may feel that the most obvious argument is simply to state that  $101/2 = 50\frac{1}{2}$  which is not an integer. This is of course a valid argument but assumes the properties of the rational numbers. For the present purpose of illustrating the use of definitions it is appropriate to seek an argument based simply on the integers.

This is not a mathematical question but is one for physicists and astronomers. However, the awareness that there is more than one possible geometry was enormously liberating and has given rise to much interesting mathematics as well as physics. Without it the geometry which provides the language used in general relativity theory would not have been discovered.

Most mathematicians do appear to take the view that mathematics deals with real objects. However, in the modern axiomatic approach axioms are seen not necessarily as ‘self-evident truths’ but simply as statements which we are assuming to be true. We do mathematics by exploring what follows from the truth of these axioms using certain accepted rules of deduction. When we come to apply the mathematics we must confirm that these axioms are ‘true’ in an appropriate sense in the area of application or acknowledge that this is an assumption that we are making. The reader will meet good examples of this axiomatic method in the study of algebra.

Since this book is primarily concerned with introducing the reader to mathematical reasoning and exposition it seems best to keep the mathematical context very familiar and so we will mainly restrict ourselves to arithmetic. Most mathematicians probably would consider the basic properties of numbers to be ‘self-evident truths’. It would be cumbersome to develop these properties formally from a list of axioms although this can be done.†. We will simply assume that the reader is familiar with the basic algebraic properties of numbers (dealing with addition, subtraction, multiplication and division) and take these for granted. These can be summarized as follows.

**Properties 2.3.1** *Given two real numbers  $a$  and  $b$  they have a sum  $a + b$  and a product  $ab$  (sometimes denoted by  $a \cdot b$  or  $a \times b$ ) which are also real numbers. The operations of sum (or addition) and product (or multiplication) of real numbers have the following properties.*

- (i) Commutativity.  $a + b = b + a$  and  $ab = ba$ . (This means that the

† Formally the real numbers and the rational numbers have the algebraic structure of a *field* and the integers have the algebraic structure of an *integral domain*. In a precise sense the integers are characterized by the statement that they form an ordered integral domain in which the positive elements satisfy the induction property (see Axiom 5.1.1). For details of such an axiomatic approach see for example G. Birkhoff and S. Mac Lane *A survey of modern algebra*, Macmillan, Fourth edition 1977. The positive integers can be characterized by a particularly simple set of axioms, Peano’s axioms, and these will be described in Chapter 9 (see Axioms 9.4.2).

*order in which we write numbers to be added or multiplied doesn't matter.)*

- (ii) **Associativity.**  $(a + b) + c = a + (b + c)$  and  $(ab)c = a(bc)$ . (*This means that we can write  $a+b+c$  and  $abc$  without brackets without ambiguity.*)
- (iii) **Distributivity.**  $a(b + c) = ab + ac$  and  $(a + b)c = ac + bc$ . (*This tells us how to remove brackets.*)
- (iv) **Zero.**  $a + 0 = a = 0 + a$ .
- (v) **Unity.**  $a \times 1 = a = 1 \times a$ .
- (vi) **Subtraction.** *The equation  $a + x = 0$  has the unique solution  $x = -a$  and so  $a + x = b \Leftrightarrow x = b + (-a) = b - a$ . (This means that we can cancel:  $a + x_1 = a + x_2 \Rightarrow x_1 = x_2$ .)*
- (vii) **Division.** *If  $a \neq 0$ , then equation  $ax = b$  has the unique solution  $x = b/a = ba^{-1}$ . (This means that we can cancel:  $ax_1 = ax_2 \Rightarrow x_1 = x_2$  so long as  $a \neq 0$ .)*

All the usual algebraic properties of numbers can be deduced from the above statements. In particular it follows that  $a \times 0 = 0 = 0 \times a$  and we can also deduce the rule of signs  $(-a)b = -(ab) = a(-b)$  and  $(-a)(-b) = ab$  (see Problems I, Question 6).

If  $a$  and  $b$  are rational numbers (fractions) then so are  $a + b$ ,  $ab$ ,  $b - a$  and (for  $a \neq 0$ )  $b/a$ . The distinction between rational numbers and real numbers will be considered in Chapter 13. If  $a$  and  $b$  are integers then so are  $a + b$ ,  $ab$  and  $b - a$ , but  $b/a$  is an integer only if  $a$  divides  $b$ . Most of this book will be concerned with the integers.

We will also make use of the order properties of the real numbers (and have already made use of these in the proof of Proposition 2.2.4); these may be less familiar and so a formal set of axioms about order (inequalities) will be given and discussed in the next chapter.

### Exercises

**2.1** Which of the following universal statements are true and which are false for integers  $n$ ?

- (i)  $n = 2$  only if  $n^2 - n - 2 = 0$ .
- (ii)  $n = 2$  if  $n^2 - n - 2 = 0$ .
- (iii)  $n = 2$  is sufficient for  $n^2 - n - 2 = 0$ .
- (iv)  $n = 2$  is necessary for  $n^2 - n - 2 = 0$ .
- (v)  $n^2 - n - 2 = 0 \Rightarrow (n = 2 \text{ and } n = -1)$ .
- (vi)  $n^2 - n - 2 = 0 \Rightarrow (n = 2 \text{ or } n = -1)$ .

- (vii)  $n^2 - n - 2 = 0 \Leftrightarrow (n = 2 \text{ or } n = -1)$ .
- (viii)  $n^2 - n - 2 = 0 \Leftrightarrow (n = 2 \text{ and } n = -1)$ .
- (ix)  $(n^2 - n - 2 = 0 \Rightarrow n = 2) \text{ or } (n^2 - n - 2 = 0 \Rightarrow n = -1)$ .
- (x)  $(n^2 - n - 2 = 0 \Leftarrow n = 2) \text{ or } (n^2 - n - 2 = 0 \Leftarrow n = -1)$ .
- (xi)  $(n^2 - n - 2 = 0 \Leftarrow n = 2) \text{ and } (n^2 - n - 2 = 0 \Leftarrow n = -1)$ .

**2.2** By constructing a truth table exhausting the possibilities of  $n$  and using the truth table for ‘implies’, prove that  $n > 0 \Rightarrow n \geq 1$  for integers  $n$ .

**2.3** Complete the following truth table for  $P \Leftrightarrow Q$  using the tables for ‘ $\Rightarrow$ ’ and ‘and’.

| $P$ | $Q$ | $P \Rightarrow Q$ | $Q \Rightarrow P$ | $P \Leftrightarrow Q$ |
|-----|-----|-------------------|-------------------|-----------------------|
| $T$ | $T$ |                   |                   |                       |
| $T$ | $F$ |                   |                   |                       |
| $F$ | $T$ |                   |                   |                       |
| $F$ | $F$ |                   |                   |                       |

**2.4** By using truth tables prove that, for all statements  $P$  and  $Q$ , the following statements are true.

- (i)  $P \Rightarrow (P \text{ or } Q)$ .
- (ii)  $(P \text{ and } Q) \Rightarrow P$ .

**2.5** By using truth tables prove that, for all statements  $P$  and  $Q$ ,

- (i) the statements ‘ $P \Rightarrow Q$ ’ and ‘(not  $Q$ )  $\Rightarrow$  (not  $P$ )’ are equivalent,
- (ii) the statements ‘ $P$  or  $Q$ ’ and ‘(not  $P$ )  $\Rightarrow$   $Q$ ’ are equivalent.

**2.6** Use the method of proof of Proposition 2.2.4 to prove that 7 does not divide 100.

# 3

## Proofs

A proof of a mathematical statement is a logical argument which establishes the truth of the statement. The steps of the logical argument are provided by implications. One of the main aims of this book is to describe a variety of methods of proof so that you can follow these when you meet them and also construct proofs for yourself.

No doubt anyone reading this book will have been seeing and understanding proofs for years. At university you are expected to be able to construct your own proofs and, as importantly, to write them out carefully so that other people can understand them – or even so that you can understand them yourself when you come to look back at your work some months later. One real difficulty is that we do not normally discover proofs in the polished form in which they are presented. It is important to realize that you will usually spend time constructing a proof before you then write out a formal proof. You can think of this as erecting a sort of scaffolding for the purpose of constructing the proof. When the proof has been constructed the scaffolding is removed so that the proof can be admired in all its economical beautiful simplicity! However, one difficulty for the person encountering the proof for the first time is that it can be hard to make sense of. To read and understand the proof we may have to reconstruct the scaffolding for ourselves from the formal proof. This can be difficult – but not usually as difficult as thinking of the proof in the first place unless the proof is very badly written. This is a problem not just for beginning undergraduates but also for professional mathematicians when they read mathematics.

You may ask why then the scaffolding is not retained. The difficulty is that if every detail is given then mathematical arguments become enormously long and cluttered. The aim is to pitch your writing at the level of the expected reader so that there is just enough information

to enable the reconstruction of the scaffolding if necessary but not so much that it would mask the essence of the argument. Too much detail can make a proof based possibly on one simple idea appear enormously complicated. To avoid detail it is quite common to use statements like ‘it is easy to show that [some statement] is true’ or ‘it now readily follows that [some statement] is true’ and then the reader has to confirm that this is indeed the case. It is, however, important not to use such phrases as a lazy way of avoiding thinking about the details.

It is also the case that excessive pedantic precision can sometimes make mathematics hard to read. Writing mathematics is not like writing a computer program; what is written will be read by a human being who has much common experience with the writer and so is able to anticipate to some extent what the writer intends. As Gila Hanna has written,

The student of mathematics has to develop a tolerance for ambiguity. Pedantry can be the enemy of insight.†

Putting this into practice is a matter of fine judgement: ambiguous statements are only acceptable in contexts which resolve the ambiguity almost immediately. While learning to write good mathematics it is probably better to err on the side of pedantry.

In this book arguments will often be presented first with lots of scaffolding (as ‘constructing a proof’) and then with the scaffolding removed (as ‘(formal) proofs’). You should ensure that in each case you do understand why the ‘proof’ does prove the result as claimed. When you read most mathematics books you need to work with pencil in hand reconstructing the detail behind the proofs provided. You cannot normally read a mathematics book like a conventional novel.

### 3.1 Direct proofs

Let us begin by thinking about one of the simplest forms of mathematical result. Very many theorems are of the form  $P \Rightarrow Q$ . How do we set about proving such a statement? Since the statement is necessarily true if  $P$  is false (remember Table 2.1.1) we only need consider the case when  $P$  is true. Then from the truth table we see that  $P \Rightarrow Q$  is true so long as  $Q$  is also true.

So to prove that  $P \Rightarrow Q$  is true, it is sufficient to assume that  $P$  is true and deduce  $Q$ . This is the direct form of proof. Here is an example.

† In David Tall (editor), *Advanced mathematical thinking*, Kluwer, 1991.

**Proposition 3.1.1** For positive real numbers  $a$  and  $b$ ,  $a < b \Rightarrow a^2 < b^2$ .

**Constructing a proof.** We can summarize what is needed in the following way using a given-goal diagram.†

| Given                        | Goal                          |
|------------------------------|-------------------------------|
| $a, b$ positive real numbers | $a < b \Rightarrow a^2 < b^2$ |

The direct method of proof of an implication is to add the hypothesis to the given statements and to set a new goal of proving the conclusion. This leads to the following new given-goal diagram.

| Given                                   | Goal        |
|---|-------------|
| $a, b$ positive real numbers<br>$a < b$ | $a^2 < b^2$ |

We now start to think how we can obtain something like the goal from the given statements. We see that we want  $a^2$  and  $b^2$  in the goal and this suggests multiplying the given inequality through by  $a$  and by  $b$ . Thus

$$a < b \Rightarrow a^2 < ab \tag{3.1}$$

and

$$a < b \Rightarrow ab < b^2 \tag{3.2}$$

using the fact that  $a$  and  $b$  are both positive. Hence, if we suppose that  $a < b$ , it follows that  $a^2 < ab$  and  $ab < b^2$ . But now we recognize that the two inequalities we have obtained both involve  $ab$  and that

$$[(a^2 < ab) \text{ and } (ab < b^2)] \Rightarrow a^2 < b^2 \tag{3.3}$$

so that if  $a < b$  then  $a^2 < b^2$  as required.

The reader might well ask where the three numbered implications have come from; why are these true? The answer is that these statements follow immediately from fundamental properties of the real numbers which can be encapsulated in the inequality (or order) axioms. We have already commented that it would be unwieldy to work from a complete axiom system in this book and readers are unlikely to have difficulties

† The idea of using these ‘given–goal diagrams’ comes from the book by Daniel J. Velleman, *How to prove it, a structured approach*, Cambridge University Press, 1994. It extends usefully the approach taught to the author at school of starting by stating clearly what we are ‘required to prove’ or ‘required to find’. Velleman provides many illustrative examples of different methods of proof some of which are also used in this book.

with the algebraic properties assumed. However, inequalities are less familiar and so it seems useful to list the basic properties to be assumed.

### Axioms 3.1.2

- (i) Trichotomy law. *For each pair of real numbers  $a$  and  $b$ , one and only one of the three possibilities  $a < b$ ,  $a = b$ ,  $a > b$  is true.*
- (ii) Addition law. *For real numbers  $a$ ,  $b$  and  $c$ ,*

$$a < b \Leftrightarrow a + c < b + c.$$

- (iii) Multiplication law. *For real numbers  $a$ ,  $b$  and  $c$ ,*

$$\begin{aligned} a < b &\Leftrightarrow ac < bc && \text{if } c > 0, \\ a < b &\Leftrightarrow ac > bc && \text{if } c < 0. \end{aligned}$$

- (iv) Transitive law. *For real numbers  $a$ ,  $b$  and  $c$ ,*

$$a < b \text{ and } b < c \Rightarrow a < c.$$

Now statements (3.1) and (3.2) follow from the multiplication law and statement (3.3) from the transitive law.

We can now write out a formal proof of Proposition 3.1.1 as follows.

*Proof* Given positive real numbers  $a$  and  $b$  suppose that  $a < b$ . Then  $a^2 < ab$  (multiplying through by  $a > 0$ ) and  $ab < b^2$  (multiplying through by  $b > 0$ ). Hence  $a^2 < b^2$ . It follows that  $a < b \Rightarrow a^2 < b^2$ .  $\square$

Notice that this proof is written as a sequence of sentences with words like ‘then’, ‘hence’ and ‘it follows’ indicating how the sentences are related. Of course some symbols are used but it is a good practice when writing mathematics to read it out aloud to check that when the symbols are converted into words (‘ $a$  is less than  $b$ ’ in the first sentence of the above proof for example) what you have written is a sensible piece of prose. We do sometimes use more symbols so that the above proof might have been written out as follows.

*Proof* For positive integers  $a$  and  $b$ ,  $a < b \Rightarrow (a^2 < ab \text{ and } ab < b^2) \Rightarrow a^2 < b^2$ . Hence  $a < b \Rightarrow a^2 < b^2$ .  $\square$

This presentation highlights the fact that in constructing a proof as a chain of implications we repeatedly use

$$[(P \Rightarrow Q) \text{ and } (Q \Rightarrow R)] \Rightarrow (P \Rightarrow R)$$

which is readily checked by using a truth table.<sup>†</sup>

However, I would encourage the reader to beware of using too many symbols in writing out proofs at this stage as it makes it much harder to be clear that what you are writing is conveying what you intend.

Notice how the ‘and’ statement was handled in this proof. An implication of the form ‘ $P \Rightarrow (Q \text{ and } R)$ ’ is logically equivalent to ‘ $(P \Rightarrow Q) \text{ and } (P \Rightarrow R)$ ’. Thus ‘ $a < b \Rightarrow (a^2 < ab \text{ and } ab < b^2)$ ’ is proved by proving (3.1) and (3.2).

One problem in writing out proofs is to decide how much detail to give and what can be assumed. There is no simple answer to this. Although the above proof did start from the inequality axioms this was not explicitly referred to in the formal proof. You always do have to start somewhere. But it is cumbersome to reduce everything to a set of axioms and there is usually a wide body of results which it is reasonable to assume. For example the result of Proposition 2.2.4 that 101 is an odd number is something which normally would simply be taken for granted. It was explored in Chapter 2 simply to illustrate the role of definitions. One very useful piece of advice is always to be sceptical, for if your proof doesn’t convince you, then it is unlikely to convince anyone else. Notice that the above proof did refer to the facts that  $a > 0$  and  $b > 0$  at the points where they were needed (in applying the multiplication law of inequalities). It is good style to indicate where the hypotheses in the result being proved are used in the proof. Also if some steps in the argument are only valid under certain conditions then you should verify that these conditions are indeed satisfied.

#### *Proof by cases*

When proving universal implications it is often very difficult to consider together all the objects satisfying the hypothesis. Here is a very simple example.

**Example 3.1.3** *If  $a = 1$  or  $a = 2$  then  $a^2 - 3a + 2 = 0$ .*

*Proof* If  $a = 1$  then  $a^2 - 3a + 2 = 1 - 3 + 2 = 0$ . If  $a = 2$  then  $a^2 - 3a + 2 = 4 - 6 + 2 = 0$ . Hence, if  $a = 1$  or  $a = 2$  then  $a^2 - 3a + 2 = 0$ .  $\square$

Notice how the ‘or’ statement was handled in constructing this simple proof. An implication of the form ‘ $(P \text{ or } Q) \Rightarrow R$ ’ is logically equivalent

<sup>†</sup> This step is known in formal logic as a ‘syllogism’:  $P$  is the ‘minor term’,  $Q$  the ‘middle term’ and  $R$  the ‘major term’;  $P \Rightarrow Q$  is the ‘minor premise’ and  $Q \Rightarrow R$  the ‘major premise’.

to the statement ' $(P \Rightarrow R)$  and  $(Q \Rightarrow R)$ ' (this is common usage but can be checked by a truth table argument) and so is proved by proving the two implications in this 'and' statement.

In this example the 'cases' were simply the two numbers satisfying the hypothesis. It can be viewed as 'proof by enumeration' working through the possibilities one by one. This is not always possible. Recall the proof of Proposition 2.2.4 on page 16. The aim was to prove the universal implication 'if  $q$  is an integer then  $2q \neq 101$ '. It is impossible to work through the integers one by one. The proof was achieved by considering two cases: integers  $q$  such that  $q \leq 50$  and integers  $q$  such that  $q \geq 51$ . This argument works because these two cases are *exhaustive*: for every integer  $q$  either  $q \leq 50$  or  $q \geq 51$ .

Here is another example deriving a familiar property of the integers from the inequality axioms.

**Proposition 3.1.4** *For non-zero real numbers  $a$ ,  $a^2 > 0$ .*

**Constructing a proof.** Summarizing what is needed gives the following.

| Given             | Goal      |
|-------------------|-----------|
| $a$ a real number |           |
| $a \neq 0$        | $a^2 > 0$ |

The trichotomy law is the only inequality axiom which refers to the equality of real numbers. If we apply this to the real numbers  $a$  and 0 it can be rewritten

$$a \neq 0 \Rightarrow a > 0 \text{ or } a < 0. \quad (3.4)$$

However, from the multiplication law,

$$a > 0 \Rightarrow a^2 > 0 \quad (3.5)$$

and

$$a < 0 \Rightarrow a^2 > 0. \quad (3.6)$$

Hence

$$a > 0 \text{ or } a < 0 \Rightarrow a^2 > 0. \quad (3.7)$$

Finally, putting together statements (3.4) and (3.7) we obtain

$$a \neq 0 \Rightarrow a^2 > 0$$

as required.

We can now write out the formal proof as follows.

*Proof* Suppose that  $a$  is a non-zero real number. Then (by the trichotomy law) either  $a > 0$  or  $a < 0$ . In either case, from the multiplication law,  $a^2 > 0$ . Hence, if  $a \neq 0$  then  $a^2 > 0$  as required.  $\square$

Again notice how the ‘or’ statement was handled in constructing this proof. To prove statement (3.7) we must prove both statement (3.5) and statement (3.6).†

### 3.2 Constructing proofs backwards

The grand thing is to be able to reason backwards.

Arthur Conan Doyle, *A study in scarlet*.

The process of mathematical investigation is extremely complex. Even when you know what you are trying to prove it is quite unusual in practice to be able construct proofs in the direct way which has been illustrated so far. One analogy which may be helpful is that of creating or discovering a route up a mountain. You may succeed by simply heading off towards the top. But often there will be a number of false starts and it may be helpful to stand back and take an overview of what is possible. You may find that for certain routes your technique is inadequate and you need to refine this on less ambitious projects. One possible approach is to develop your route by mentally working back from the summit, in other words by planning the route backwards. This is frequently a sensible approach to constructing mathematical proofs.

It may be worth remarking that this analogy seems useful when we ponder the nature of mathematical invention: is mathematics discovered or created? A new route up a mountain is both discovered and created; it is to some extent rediscovered and recreated every time someone subsequently uses it – it may even be slightly different each time. But the route appears to have an objective existence independent of those who use it. The same is true of mathematical ideas.

Here is an example requiring a little more thought than the results considered so far.

**Proposition 3.2.1** *For real numbers  $a$  and  $b$ ,  $a < b \Rightarrow 4ab < (a + b)^2$ .*

† It is necessary to be careful about the way the words ‘and’ and ‘or’ are used in normal speech. For example statement (3.7) might be formulated using the word ‘and’ as ‘ $a > 0$  and  $a < 0$  (both) imply that  $a^2 > 0$ ’. This is really a shorthand for ‘ $a > 0$  implies that  $a^2 > 0$  and  $a < 0$  implies that  $a^2 > 0$ ’.

**Constructing a proof.** We can summarize what is needed for a direct proof as follows.

| Given                          | Goal              |
|--------------------------------|-------------------|
| $a, b$ real numbers<br>$a < b$ | $4ab < (a + b)^2$ |

It is difficult to see how to proceed. One difficulty is that the goal is more complicated than the given statement and it is not immediately clear how to reach this more complicated statement. In such a situation it is often best to start with the more complicated statement and to simplify it. In this case simplifying the goal leads to the construction of a proof backwards.

$$\begin{aligned}
 4ab < (a + b)^2 &\Leftarrow 4ab < a^2 + 2ab + b^2 \\
 &\Leftarrow 0 < a^2 - 2ab + b^2 \\
 &\Leftarrow 0 < (a - b)^2 \\
 &\Leftarrow a - b \neq 0 \text{ by Proposition 3.1.4} \\
 &\Leftarrow a \neq b \\
 &\Leftarrow a < b.
 \end{aligned}$$

Hence  $a < b \Rightarrow 4ab < (a + b)^2$ .

It is essential for the proof that these implications go in the direction shown. Of course, apart from the last step, in this case they do in fact go in both directions, and it would be usual to write ' $\Leftrightarrow$ ' instead of ' $\Leftarrow$ ' in those cases. But the fact that the implications ' $\Rightarrow$ ' are true is quite irrelevant for the proof. The above sequence of implications is perfectly satisfactory as a formal proof for it makes it clear that the truth of the goal is *implied* by what is given. If preferred the proof can be presented as a direct proof as follows.

*Proof*  $a < b \Rightarrow a \neq b \Rightarrow a - b \neq 0 \Rightarrow 0 < (a - b)^2 \Rightarrow 0 < a^2 - 2ab + b^2 \Rightarrow 4ab < a^2 + 2ab + b^2 \Rightarrow 4ab < (a + b)^2$ . Hence  $a < b \Rightarrow 4ab < (a + b)^2$ .

□

However, presenting the proof in this way disguises where it has come from. Whilst readers of the proof may accept each of the steps and so be forced to accept the proof they would probably find the earlier backwards presentation more satisfying because it demonstrates how the proof was found.

**Exercises**

**3.1** Prove that for all real numbers  $a$ ,  $b$  and  $c$ ,

- (i)  $(a + b - c)^2 = (a + b)^2 + (a - c)^2 + (b - c)^2 - a^2 - b^2 - c^2$ ,
- (ii)  $bc + ac + ab \leq a^2 + b^2 + c^2$ .

**3.2** Prove that for all integers  $a$ ,  $b$  and  $c$ ,

$$(a \text{ divides } b) \text{ and } (b \text{ divides } c) \Rightarrow a \text{ divides } c.$$

**3.3** Prove that the square of an even integer is even.

**3.4** Prove that 0 divides an integer  $a$  if and only if  $a = 0$ .

**3.5** Prove, from the inequality axioms, that if  $a$ ,  $b$  and  $c$  are real numbers with  $a > 0$ , then  $b \geq c \Rightarrow ab \geq ac$ .

**3.6** Prove that for negative real numbers  $a$  and  $b$ ,  $a < b \Rightarrow a^2 > b^2$ .

**3.7** Proposition 3.2.1 states that  $a < b$  is a sufficient condition for  $4ab < (a + b)^2$ . Is this condition also necessary? If so, prove it. If not, find a necessary and sufficient condition.

**3.8** Prove that for all real numbers  $a$  and  $b$ ,

- (i)  $0 \leq a < b \Rightarrow a^2 < b^2$ ,
- (ii)  $|a| < |b| \Rightarrow a^2 < b^2$ ,
- (iii)  $|a| = |b| \Rightarrow a^2 = b^2$ ,
- (iv)  $|a| \leq |b| \Rightarrow a^2 \leq b^2$ .

# 4

## Proof by contradiction

In the previous chapter we considered how to construct some simple direct proofs. However, the direct method can be inconvenient and does not always work. In this chapter we consider a logically more elaborate but very common and powerful method of proof: proof by contradiction.

### 4.1 Proving negative statements by contradiction

The direct method of proof is often hard to use when we are proving negative statements. Consider the following example.

**Proposition 4.1.1** *There do not exist integers  $m$  and  $n$  such that*

$$14m + 20n = 101.$$

This is a simple example of a *non-existence result* and results of this type are very common in more advanced mathematics.

It is difficult to see how to prove Proposition 4.1.1 directly for we clearly cannot consider all the possibilities for  $m$  and  $n$  one at a time. The method of proof by contradiction is to demonstrate that if what we are trying to prove were false then this would lead to a statement which is known not to be true, a contradiction. A *contradiction* is a statement of the form ' $P$  and (not  $P$ )'. For example, if  $a$  is an integer, the statement ' $a$  is even and  $a$  is odd' is a contradiction since 'odd' is defined to mean 'not even'. Having obtained a contradiction we know that our initial assumption must have been wrong and so the result we are trying to prove must be true.

This may sound confusing and so before we consider in general how this method can be justified let us illustrate it by considering the proof of Proposition 4.1.1.

**Constructing a proof.** The proposition states that whatever integers  $m$  and  $n$  we choose these will not satisfy the equation, i.e.  $14m + 20n \neq 101$ . We can sum this up as follows.

| Given           | Goal                 |
|-----------------|----------------------|
| $m, n$ integers | $14m + 20n \neq 101$ |

A proof by contradiction is given by showing that if the goal were false then this would lead to something which we know not to be true, a contradiction. Then since this shows that the negative of the goal is impossible it means that the goal must be true in accordance with the following precept of Sherlock Holmes'.

When you have eliminated the impossible, whatever remains, *however improbable*, must be the truth.

Arthur Conan Doyle, *The sign of four*.

The method is carried out by adding the negative of the goal to the givens and making the finding of a contradiction the new goal. In this case this gives the following strategy.

| Given                                | Goal          |
|--------------------------------------|---------------|
| $m, n$ integers<br>$14m + 20n = 101$ | contradiction |

Here we can get a contradiction by using the idea of even and odd numbers. Notice that both  $14 = 2 \times 7$  and  $20 = 2 \times 10$  are even numbers and so any number of the form  $14m + 20n$ , where  $m$  and  $n$  are integers, is also even since  $14m + 20n = 2(7m + 10n)$ . Thus if  $14m + 20n = 101$  this means that 101 is an even number. But 101 is an odd number (see Proposition 2.2.4 for a formal proof but we would usually assert this without proof) and so we have two contradictory statements about 101 giving a contradiction as required.

The proof is quite short when written out in final form. Notice that we include the phrase ‘for contradiction’ in the first sentence in order to indicate the method we are using.

*Proof* Suppose for contradiction that  $m$  and  $n$  are integers such that  $14m + 20n = 101$ . Then, since 14 is even and 20 is even,  $101 = 14m + 20n = 2(7m + 10n)$  is even. But this is not true since 101 is odd. Hence such integers  $m$  and  $n$  cannot exist, as required. □

We can justify this method using a truth table. Suppose that we are trying to prove statement  $P$ . The method of contradiction is to prove

the statement  $(\text{not } P) \Rightarrow Q$  where  $Q$  is a false statement. Now look at the following truth table.

| $P$ | $\text{not } P$ | $Q$ | $(\text{not } P) \Rightarrow Q$ |
|-----|-----------------|-----|---------------------------------|
| $T$ | $F$             | $T$ | $T$                             |
| $T$ | $F$             | $F$ | $T$                             |
| $F$ | $T$             | $T$ | $T$                             |
| $F$ | $T$             | $F$ | $F$                             |

If we know that  $(\text{not } P) \Rightarrow Q$  is true and  $Q$  is false, then the only possibility in the above table is the second line. But this shows that  $P$  is true.

### *A template for proofs by contradiction*

It may be useful to present a template for writing out these proofs. Suppose that we are writing out a formal proof of the statement  $P$  using the contradiction method. We would set out the proof as follows.

*Proof* Suppose, for contradiction, that the statement  $P$  is false.  
 Then [present some argument which leads to a contradiction of some sort].  
 Hence our assumption that  $P$  is false must be false.  
 Thus  $P$  is true as required. □

A common failing in writing out proofs by contradiction is for it to be unclear precisely what the contradiction is. Take care to make this clear.

As a second example here is an alternative proof of Proposition 2.2.4 using this method.

**Proposition 2.2.4** 101 is an odd integer.

**Constructing a proof.** This appears to be a positive statement but since the word ‘odd’ means ‘not even’ it too is a non-existence result since it asserts that 101 is not a multiple of 2. We can describe this as follows.

| Given          | Goal          |
|----------------|---------------|
| $q$ an integer | $2q \neq 101$ |

In Chapter 2 the goal was achieved by considering all the possibilities for  $q$ . Although it is not possible to consider these one at a time we could take account of all the possibilities by observing that for any integer  $q$  either  $q \leq 50$  or  $q \geq 51$ . It was observed in Chapter 3 that this is an example of a proof by cases.

As an alternative we can apply the method of proof by contradiction which leads to the following strategy.

| Given          | Goal          |
|----------------|---------------|
| $q$ an integer |               |
| $2q = 101$     | contradiction |

We get a contradiction if we remember how we did division in our early school-days. Then we would have divided 2 into 101 giving 50 with remainder 1. This suggests

$$101 = 2q \Rightarrow 1 = 101 - 100 = 2q - 2 \times 50 = 2(q - 50) \geq 2 \Rightarrow 1 \geq 2$$

and we have a contradiction since it is certainly true that  $1 < 2$ .

*Proof* Suppose for contradiction that 101 is even so that  $101 = 2q$  for some integer  $q$ . Then  $1 = 2(q - 50)$  where  $q - 50$  must be a positive integer. But  $2(q - 50) \geq 2$  so that  $1 \geq 2$  contradicting the fact that  $1 < 2$ . Hence 101 is not even and so must be odd.  $\square$

## 4.2 Proving implications by contradiction

We sometimes find that the direct method of proof of statements of the form  $P \Rightarrow Q$  does not work. Consider the following result.

**Proposition 4.2.1** *If  $a, b, c$  are integers such that  $a > b$ , then*

$$ac \leq bc \Rightarrow c \leq 0.$$

**Constructing a proof.** It is difficult to know how to make use of the multiplicative law of inequalities (see Axioms 3.1.2) in a direct proof since this depends on the sign of  $c$  which is what we are trying to determine.

The easiest thing to do in this case is again a proof by contradiction. Notice from the truth table for ‘implies’ (Table 2.1.1) that if  $P \Rightarrow Q$  is false then we must have  $P$  true and  $Q$  false. So we can prove that  $P \Rightarrow Q$  is true by showing that  $P$  true and  $Q$  false together imply a contradiction.

What we are asked for in Proposition 4.2.1 may be summarized as follows.

| Given              | Goal                              |
|--------------------|-----------------------------------|
| $a, b, c$ integers |                                   |
| $a > b$            | $ac \leq bc \Rightarrow c \leq 0$ |

An attempt at a direct proof of the proposition leads to the following strategy.

| Given              | Goal       |
|--------------------|------------|
| $a, b, c$ integers |            |
| $a > b$            |            |
| $ac \leq bc$       | $c \leq 0$ |

The method of proof by contradiction is to add the negative of the desired conclusion to the ‘givens’ and to seek a contradiction. This gives the following strategy.

| Given              | Goal          |
|--------------------|---------------|
| $a, b, c$ integers |               |
| $a > b$            |               |
| $ac \leq bc$       |               |
| $c > 0$            | contradiction |

It is easy to obtain a contradiction. For by the multiplicative law of inequalities (Axioms 3.1.2)

$$a > b \text{ and } c > 0 \Rightarrow ac > bc.$$

Hence, since our given statements include the statement  $ac \leq bc$ , they imply that  $ac > bc$  and  $ac \leq bc$  which is a contradiction.

The proof is quite short when written out in final form as follows.

*Proof* For integers  $a, b$  and  $c$ , with  $a > b$ , suppose that  $ac \leq bc$  but, for contradiction, that  $c > 0$ . Then the given statement  $a > b$  implies that  $ac > bc$ , contradicting the statement that  $ac \leq bc$ . Hence our assumption that  $c > 0$  must be false, i.e.  $c \leq 0$ . Thus  $ac \leq bc \Rightarrow c \leq 0$ .

□

### 4.3 Proof by contrapositive

In the proof of Proposition 4.2.1 the contradiction we obtained involved the hypothesis in the proposition. This is a particular form of proof by

contradiction. It essentially uses the fact that, given any statements  $P$  and  $Q$ , the statements ' $P \Rightarrow Q$ ' and its *contrapositive* '(not  $Q$ )  $\Rightarrow$  (not  $P$ )' are equivalent, as can be seen by examining their truth tables (see Exercise 2.5). Thus if one is true then so is the other. Now Proposition 4.2.1 is a statement about integers  $a$ ,  $b$  and  $c$  such that  $a > b$ , of the form  $P \Rightarrow Q$  where

$P$  is the statement  $ac \leq bc$ ,  
 $Q$  is the statement  $c \leq 0$ .

To write down the contrapositive notice that

(not  $P$ ) is the statement  $ac > bc$ ,  
(not  $Q$ ) is the statement  $c > 0$ ,

and so the statement (not  $Q$ )  $\Rightarrow$  (not  $P$ ) reads

$$c > 0 \Rightarrow ac > bc,$$

where  $a$ ,  $b$  and  $c$  are integers such that  $a > b$ . But this is just the multiplicative law of inequalities included in Axioms 3.1.2. Hence we could have proved Proposition 4.2.1 by simply observing that its contrapositive is true. We would write out this proof as follows.

*Proof of Proposition 4.2.1* The contrapositive of the statement

$$ac \leq bc \Rightarrow c \leq 0$$

is the statement

$$c > 0 \Rightarrow ac > bc$$

which is just the multiplicative law of inequalities since we are given that  $a > b$ . Thus the proposition is true.  $\square$

#### 4.4 Proving ‘or’ statements

Constructing a proof for composite statements involving ‘or’ usually makes use of a logical construction rather similar to proof by contradiction which it is convenient to consider at this point. Again let us consider an example.

**Proposition 4.4.1** *If  $a$  and  $b$  are real numbers, then*

$$ab = 0 \Leftrightarrow a = 0 \text{ or } b = 0.$$

**Constructing a proof.** Recall from Chapter 2 that the statement ' $P \Leftrightarrow Q$ ' means ' $(P \Rightarrow Q)$  and  $(Q \Rightarrow P)$ '. So there are two parts to the proof: a proof of the ' $\Leftarrow$ ' statement and a proof of the ' $\Rightarrow$ ' statement. We consider these in turn.

' $\Leftarrow$ ': This statement is a restatement of a basic property of the number 0, that  $0 \times b = 0 = a \times 0$  for real numbers  $a$  and  $b$ . This property was mentioned in the discussion at the end of Chapter 2.<sup>†</sup>

' $\Rightarrow$ ': Adopting the direct strategy for proving the converse statement gives the following strategy.

| Given                           | Goal               |
|---------------------------------|--------------------|
| $a, b$ real numbers<br>$ab = 0$ | $a = 0$ or $b = 0$ |

Now recall that ' $P$  or  $Q$ ' is logically equivalent to '(not  $P$ )  $\Rightarrow$   $Q$ ' (see Exercise 2.5) so that the goal may be rewritten ' $a \neq 0 \Rightarrow b = 0$ '. Adopting the direct strategy for proving this gives the following.

| Given   | Goal    |
|---|---------|
| $a, b$ real numbers<br>$ab = 0$<br>$a \neq 0$ | $b = 0$ |

Since  $a \neq 0$ , we can divide through by  $a$  (or multiply through by  $1/a$ ) so that  $ab = 0 \Rightarrow b = 0$  as required.

A formal proof could be set out as follows.

*Proof* It is a basic property of 0 that  $0 \times b = 0 = a \times 0$ . Therefore  $a = 0$  or  $b = 0 \Rightarrow ab = 0$ .

For the converse, suppose that  $a$  and  $b$  are real numbers such that  $ab = 0$ . To see that it follows that  $a = 0$  or  $b = 0$  suppose that  $a \neq 0$ . Then dividing  $ab = 0$  through by  $a$  gives  $b = 0$ , as required. Hence  $ab = 0 \Rightarrow a = 0$  or  $b = 0$ .  $\square$

An alternative method of obtaining this result is by proof by cases. The trichotomy law of inequalities tells us that, for any real number  $a$ ,  $a < 0$  or  $a = 0$  or  $a > 0$ . This gives three possibilities for  $a$  and similarly there are three possibilities for  $b$ , leading to nine possibilities

<sup>†</sup> Problems I, Question 6 asks for a formal proof of this result from Properties 2.3.1.

in all. These can be summarized in a table as follows.

|         | $a < 0$  | $a = 0$  | $a > 0$  |
|---------|----------|----------|----------|
| $b < 0$ | $ab > 0$ | $ab = 0$ | $ab < 0$ |
| $b = 0$ | $ab = 0$ | $ab = 0$ | $ab = 0$ |
| $b > 0$ | $ab < 0$ | $ab = 0$ | $ab > 0$ |

This comes from the multiplicative law of inequalities and the fact that  $ab = 0$  if  $a = 0$  or  $b = 0$ .

Now since all possibilities are included in this table we can read off Proposition 4.4.1 and also the following useful results.

**Proposition 4.4.2** *If  $a$  and  $b$  are real numbers, then  $ab > 0$  if and only if  $a$  and  $b$  have the same sign, i.e. ( $a > 0$  and  $b > 0$ ) or ( $a < 0$  and  $b < 0$ ).*

**Proposition 4.4.3** *If  $a$  and  $b$  are real numbers, then  $ab < 0$  if and only if  $a$  and  $b$  have opposite signs, i.e. ( $a > 0$  and  $b < 0$ ) or ( $a < 0$  and  $b > 0$ ).*

Formally the right to left implications in these results come from the multiplicative law of inequalities and the left to right implications are proved by contradiction using the above table which shows, for example, that if  $a$  and  $b$  do not have the same sign then  $ab$  is not positive.

Proposition 4.4.1 is used when we solve polynomial equations and in the same way we use the other two results to solve polynomial inequalities (see Exercise 6.2).

### Exercises

**4.1** Prove by contradiction that there do not exist integers  $m$  and  $n$  such that  $14m + 21n = 100$ .

**4.2** Prove by contradiction that for any integer  $n$

$$n^2 \text{ is odd} \Rightarrow n \text{ is odd.}$$

**4.3** Prove the result of the previous exercise by writing down its contrapositive.

**4.4** Prove, using the method given above for proving an ‘or’ statement, that if  $a$  is a real number such that  $a^2 \geqslant 7a$  then  $a \leqslant 0$  or  $a \geqslant 7$ .

**4.5** Prove by contradiction from the trichotomy law that, for any real numbers  $a$  and  $b$ ,

$$a \leq b \text{ and } b \leq a \Rightarrow a = b.$$

**4.6** Write down the contrapositive of the statement of Exercise 3.8(iv). Hence prove that, for all real numbers  $a$  and  $b$ ,

- (i)  $|a| < |b|$  if and only if  $a^2 < b^2$ ,
- (ii)  $|a| = |b|$  if and only if  $a^2 = b^2$ .

**4.7** Prove that, for all real numbers  $a$  and  $b$ ,

$$|a + b| \leq |a| + |b|.$$

Give a necessary and sufficient condition for equality.

# 5

## The induction principle

The essence of the natural number concept is ... closure under the successor operation.<sup>†</sup>

Richard Dedekind (1888)

In this chapter we discuss a special proof technique which is particularly useful when proving statements about the positive integers.

### 5.1 Proof by induction

Suppose that we wish to prove that some property holds for all the positive integers  $1, 2, 3, 4, \dots$ . It is often difficult, or even impossible, to prove such statements simply from basic rules of arithmetic. What these rules fail to capture is the fact that the positive integers come in a sequence with any number obtainable by starting from the number 1 and adding 1 to it enough times. The integer  $n + 1$  is called the *successor* of the integer  $n$ . Thus, if we start with the integer 1 and form its successor, and then its successor, and so on, then given any positive integer eventually it will be reached. This idea can be formulated more precisely as follows.

**Axiom 5.1.1 (The induction principle)** *Suppose, that  $P(n)$  is a statement involving a general positive integer  $n$ . Then  $P(n)$  is true for all positive integers  $1, 2, 3, \dots$  if*

- (i)  *$P(1)$  is true, and*
- (ii)  *$P(k) \Rightarrow P(k + 1)$  for all positive integers  $k$ .*

Before discussing the induction principle further, here is an example of how it is used.

<sup>†</sup> The significance of this quotation will be discussed further in Section 9.4.

**Proposition 5.1.2** *For all positive integers  $n$  we have the inequality  $n \leq 2^n$ .*

**Constructing a proof.** When using the induction principle it is important to be clear about what constitutes the predicate  $P(n)$ . In this case  $P(n)$  is the statement

$$n \leq 2^n.$$

Let us begin by comparing the two numbers  $n$  and  $2^n$  in some particular cases.

|       |   |   |   |    |    |    |
|-------|---|---|---|----|----|----|
| $n$   | 1 | 2 | 3 | 4  | 5  | 6  |
| $2^n$ | 2 | 4 | 8 | 16 | 32 | 64 |

From this we can see that the result certainly holds for  $n = 1, 2, 3, 4, 5$  and  $6$  and it seems highly plausible that the result will hold for all positive integers  $n$ . However, testing any number of cases does not constitute a proof that the result will always hold. In this example the above table suggests more than just that the result holds for certain particular values of  $n$ . It demonstrates that the numbers in the second row are growing much more rapidly than the numbers in the first row and if this continues then it would be a strong indication that the result  $P(n)$  holds for all positive integers  $n$ . The proof of the result by induction makes this argument precise.

Referring to Axiom 5.1.1, we see that it asserts that two statements will lead to the proposition.

- (i) The statement  $P(1)$  simply says that  $1 \leq 2$  which is certainly true. This is called the *base case*.
- (ii) The second statement that is needed is called the *inductive step*. We are to prove for each positive integer  $k$  that  $P(k) \Rightarrow P(k+1)$ , in other words that if  $P(k)$  is true then so is  $P(k+1)$ . Here the statement  $P(k)$  is often referred to as the *inductive hypothesis*, for the aim is to prove  $P(k+1)$  under the hypothesis that  $P(k)$  is true.

We can sum up what is needed for a direct proof of the inductive step in this case as follows.

| Given                                  | Goal               |
|--|--------------------|
| $k$ a positive integer<br>$k \leq 2^k$ | $k+1 \leq 2^{k+1}$ |

In order to obtain the goal we try to relate it to the inductive hypothesis. So we start with one side of the inequality required and express it in

terms of one side of the inequality of the hypothesis. One possibility is the following.

$$\begin{aligned} k+1 &\leqslant 2^k + 1 \text{ (by inductive hypothesis)} \\ &\leqslant 2^k + k \text{ since } k \geqslant 1 \\ &\leqslant 2^k + 2^k \text{ (by inductive hypothesis)} \\ &= 2 \times 2^k = 2^{k+1}. \end{aligned}$$

Alternatively we can start on the other side as follows.

$$\begin{aligned} 2^{k+1} = 2 \times 2^k &\geqslant 2k \text{ (by inductive hypothesis)} \\ &= k + k \\ &\geqslant k + 1. \end{aligned}$$

We write out the formal proof (using the second approach) as follows.

*Proof* We use induction on  $n$ .

Base case: For  $n = 1$ ,  $2^n = 2$  and so, since  $1 \leqslant 2$ ,  $n \leqslant 2^n$ .

Inductive step: Suppose now as inductive hypothesis that  $k \leqslant 2^k$  for a positive integer  $k$ . Then  $2^{k+1} = 2 \times 2^k \geqslant 2k$  (by inductive hypothesis)  $= k + k \geqslant k + 1$  and so  $2^{k+1} \geqslant k + 1$  as required.

Conclusion: Hence, by induction,  $n \leqslant 2^n$  for all positive integers  $n$ .  $\square$

The idea of induction is that, since  $P(1)$  is true (explicitly checked) and also  $P(1) \Rightarrow P(2)$  (special case of inductive step), we know that  $P(2)$  is true, and now since  $P(2) \Rightarrow P(3)$  we know that  $P(3)$  is true, and so on:  $P(3) \Rightarrow P(4) \Rightarrow P(5) \Rightarrow \dots$ . Eventually this process will reach  $P(n)$  for any specified positive integer  $n$ . It might be thought that this is ‘obvious’ and in a way it is. But this is because our intuitive ideas about what the integers are include more than the fact that they can be added, multiplied and compared in size. The induction principle is simply a formulation of something which we take for granted about the integers: that we can reach any positive integer by starting from 1 and repeatedly adding 1. Suppose that we think of the integers lined up like dominoes. The inductive step tells us that they are close enough for each domino to knock over the next one, the base case tells us that the first domino falls over, the conclusion is that they all fall over.<sup>†</sup>

<sup>†</sup> The fault in this analogy is that it takes time for each domino to fall and so a domino which is a long way along the line won’t fall over for a long time. Mathematical implication is outside time.

Formally, the induction principle is another axiom about the integers, in addition to the algebraic and order axioms.

It is important to realize that we cannot *prove* a result like Proposition 5.1.2 by considering each case in turn. For any specific value of  $n$  we can check the result by calculation. For example, consider  $n = 8$ : since  $2^8 = 256$  and  $8 \leq 256$ , the statement  $P(8)$  is true. But however many individual cases we checked numerically we would not know that the result always held. Indeed there are mathematical statements (such as statement (iv) in Section 1.1) which are known to be true for a huge number of special cases but which have not yet been proved. Proof by induction is not the only way of proving general statements about the positive integers but it does provide an enormously powerful technique.

In writing out proofs by induction it is important to make it clear that this is the method being used and to be absolutely clear what is the statement  $P(n)$  that is being proved. If you do no more in developing a proof before writing out a formal proof you should certainly write down what the statement  $P(n)$  is.

### *A template for proofs by induction*

It is a good approach for your proofs to follow a standard pattern. First identify the statement  $P(n)$  and be clear what the statements  $P(1)$ ,  $P(k)$ ,  $P(k + 1)$  each say. Then when you write out the formal proof use the following template.

*Proof* We use induction on  $n$ .

Base case: [Prove the statement  $P(1)$ ]

Inductive step: Suppose now as inductive hypothesis that [ $P(k)$  is true] for some positive integer  $k$ . Then [deduce that  $P(k + 1)$  is true]. This proves the inductive step.

Conclusion: Hence, by induction, [ $P(n)$  is true] for all positive integers  $n$ . □

There are many variants of this layout. In particular, the conclusion line is often omitted and including the words ‘base case’, ‘inductive step’ and ‘conclusion’ is not usual. However, as you begin to apply the inductive method, using this template helps to emphasize what such a proof involves. The first two parts (base case and inductive step) are

concerned with verifying the conditions in Axiom 5.1.1 and then the last part (conclusion) invokes the axiom.

Here is another example with the proof laid out according to the above template [with a few comments in brackets].

**Proposition 5.1.3** *For all positive integers  $n$  the number  $n^2 + n$  is even.*

*Proof* We use induction on  $n$ .

[In this case the statement  $P(n)$  to be proved for all positive integers is ' $n^2 + n$  is even' or equivalently '2 divides  $n^2 + n$ ' which means that ' $n^2 + n = 2q$  for some integer  $q$ ' (using Definitions 2.2.1 and 2.2.2).]

Base case: For  $n = 1$ ,  $n^2 + n = 1 + 1 = 2 = 2 \times 1$ , an even number, as required.

Inductive step: Suppose now as inductive hypothesis that  $k^2 + k$  is even for some positive integer  $k$ . Then  $k^2 + k = 2q$  for some integer  $q$ .

[Notice now that the statement  $P(k + 1)$  is ' $(k + 1)^2 + (k + 1)$  is even'.]

Then  $(k+1)^2+(k+1)=k^2+2k+1+k+1=(k^2+k)+2k+2=2q+2k+2$  (by inductive hypothesis)  $= 2(q+k+1) = 2p$ , where  $p$  is the integer  $q+k+1$ , and so  $(k+1)^2+(k+1)$  is even as required.

Conclusion: Hence, by induction,  $n^2 + n$  is even for all positive integers  $n$ .  $\square$

Often the symbol  $n$  is used again when dealing with the inductive step instead of introducing the new symbol  $k$ , i.e. the inductive step reads  $P(n) \Rightarrow P(n + 1)$  for all positive integers  $n$ . The reason for introducing a new letter is that it is easier to write down  $P(k + 1)$  by substituting  $k + 1$  for  $n$  in  $P(n)$ . Substituting  $n + 1$  for  $n$  to obtain  $P(n + 1)$  can lead to confusion and a common error in inductive proofs is getting  $P(k + 1)$  wrong. I would recommend avoiding the double use of the symbol  $n$  at least until you are very familiar with inductive proofs.

The examples given in this chapter are quite simple so that we can concentrate on the method. There will be many more examples in this book and the method is very common. Quite a bit of ingenuity may be required in proving the inductive step and even, on occasion, in formulating the statement  $P(n)$  properly. The reader will sometimes meet inductive proofs presented informally using phrases like 'and so on' with a comment that a formal argument can be written out using induction; in these cases it can be useful for the reader to try to do this.

## 5.2 Changing the base case

The induction principle is used to prove a statement about positive integers by first proving the statement for the integer 1 as the base case and then proving that if it holds for some positive integer then it necessarily holds for its successor. However, the same idea can be used starting from any integer as the base case. Suppose that  $n_0$  is an integer, positive, negative or zero. Induction can be used to prove that a statement  $P(n)$  is true for all integers  $n$  such that  $n \geq n_0$ . The base case is now  $P(n_0)$  and the inductive step is  $P(k) \Rightarrow P(k+1)$  for  $k \geq n_0$ . The basic template given above becomes the following.

*Proof* We use induction on  $n$ .

Base case: [Prove the statement  $P(n_0)$ ]

Inductive step: Suppose now as inductive hypothesis that [ $P(k)$  is true] for some integer  $k$  such that  $k \geq n_0$ . Then [deduce that  $P(k+1)$  is true]. This proves the inductive step.

Conclusion: Hence, by induction, [ $P(n)$  is true] for all integers  $n \geq n_0$ . □

This is illustrated in the following result comparing the sizes of  $n^2$  and  $2^n$ . Calculating these numbers for low values of  $n$  gives the following table.

|       |   |   |   |    |    |    |     |     |
|-------|---|---|---|----|----|----|-----|-----|
| $n$   | 1 | 2 | 3 | 4  | 5  | 6  | 7   | 8   |
| $n^2$ | 1 | 4 | 9 | 16 | 25 | 36 | 49  | 64  |
| $2^n$ | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 |

This suggests the following result.

**Proposition 5.2.1** *For all integers  $n$  such that  $n \geq 4$ , we have the inequality  $n^2 \leq 2^n$ .*

**Constructing a proof.** The inductive step of a proof by induction on  $n$  reads

$$k^2 \leq 2^k \Rightarrow (k+1)^2 \leq 2^{k+1}$$

for integers  $k \geq 4$ . We can achieve the right-hand side of the conclusion inequality by multiplying the hypothesis inequality by 2 giving

$$2k^2 \leq 2^{k+1}.$$

Now, if we can prove that  $(k+1)^2 \leq 2k^2$  then we can complete the proof of the inductive step using

$$(k+1)^2 \leq 2k^2 \leq 2^{k+1} \Rightarrow (k+1)^2 \leq 2^{k+1}.$$

But this inequality is readily proved for  $k \geq 4$  by expanding the brackets and simplifying.

*Proof* We use induction on  $n$ .

Base case: For  $n = 4$ ,  $n^2 = 16$  and  $2^n = 16$  and so  $n^2 \leq 2^n$ .

Inductive step: Suppose now as inductive hypothesis that  $k^2 \leq 2^k$  for some  $k \geq 4$ . Then  $2^{k+1} = 2 \times 2^k \geq 2k^2$  (by inductive hypothesis). So we will have proved that  $2^{k+1} \geq (k+1)^2$  if we can prove that  $2k^2 \geq (k+1)^2$ . But  $2k^2 \geq (k+1)^2 \Leftrightarrow 2k^2 \geq k^2 + 2k + 1 \Leftrightarrow k^2 \geq 2k + 1$  and, since  $k \geq 4$ ,  $k^2 \geq 4k \geq 2k + 2 \geq 2k + 1$  so that  $k^2 \geq 2k + 1$ . Hence  $2k^2 \geq (k+1)^2$  and so we have deduced that  $(k+1)^2 \leq 2^{k+1}$  as required to complete the inductive step.

Conclusion: Hence, by induction,  $n^2 \leq 2^n$  for all  $n \geq 4$ .  $\square$

Notice that induction starting from a base case other than 1 is not really an extension of the induction principle in Axiom 5.1.1. Consider the statement that  $P(n)$  is true for all  $n \geq n_0$ . If we put  $m = n - n_0 + 1$  then  $n \geq n_0$  if and only if  $m \geq 1$ . Thus, since  $n = m + n_0 - 1$ ,  $P(n)$  is true for all  $n \geq n_0$  if and only if  $P(m + n_0 - 1)$  is true for  $m \geq 1$ . The inductive proof of  $P(n)$  with base case  $n = n_0$  is essentially identical to the inductive proof of  $P(m + n_0 - 1)$  with base case  $m = 1$ . For example, the result of Proposition 5.2.1 can be rewritten as  $(m+3)^2 \leq 2^{m+3}$  for  $m \geq 1$ . In this form, the result can be proved by induction with base case  $m = 1$  and if the reader writes out such a proof it will be clear that it is almost identical to the proof given above.

### 5.3 Definition by induction

Consider the following result.

**Proposition 5.3.1** *The sum of the first  $n$  positive integers  $1+2+\dots+n$  is equal to  $\frac{1}{2}n(n+1)$ .*

Whenever you see a line of dots indicating ‘and so on’ this indicates a *definition* by induction (sometimes called a definition by recursion).

In this case, a more precise notation for the sum of the first  $n$  positive integers is  $\sum_{i=1}^n i$ .

**Definition 5.3.2** Given a sequence of numbers  $a(1), a(2), \dots$ , the numbers  $\boxed{\sum_{i=1}^n a(i)}$  for positive integers  $n$  are defined inductively by the following statements:

- (i)  $\sum_{i=1}^1 a(i) = a(1)$ ; and
- (ii)  $\sum_{i=1}^{k+1} a(i) = \sum_{i=1}^k a(i) + a(k+1)$  for  $k \geq 1$ .

Again we have a base case which tells us what the notation means in the case  $n = 1$  and an inductive step which tells us what it means for  $n = k + 1$  in terms of what it means for  $n = k$ . For any specific value of  $n$  we can evaluate the expression by repeated use of the inductive step. Thus, for example,

$$\begin{aligned}\sum_{i=1}^3 a(i) &= \sum_{i=1}^2 a(i) + a(3) \quad (\text{by (ii) for } k = 2) \\ &= \sum_{i=1}^1 a(i) + a(2) + a(3) \quad (\text{by (ii) for } k = 1) \\ &= a(1) + a(2) + a(3) \quad (\text{by (i)}).\end{aligned}$$

Using this notation we can rewrite Proposition 5.3.1 as follows.

**Proposition 5.3.1** For positive integers  $n$ ,

$$\sum_{i=1}^n i = \frac{1}{2}n(n+1).$$

In this case  $\sum_{i=1}^n i$  is defined using Definition 5.3.2 with  $a(i) = i$ .

**Constructing a proof.** We have to make use of the inductive definition in constructing a proof by induction. In this case the statement  $P(n)$  is

the equation in the proposition. So  $P(k)$  is

$$\sum_{i=1}^k i = \frac{1}{2}k(k+1)$$

and  $P(k+1)$  is

$$\sum_{i=1}^{k+1} i = \frac{1}{2}(k+1)((k+1)+1) = \frac{1}{2}(k+1)(k+2).$$

Let us proceed directly to the formal proof.

*Proof* We use induction on  $n$ .

Base case: For  $n = 1$ ,  $\sum_{i=1}^n i = 1$  and  $\frac{1}{2}n(n+1) = \frac{1}{2} \cdot 1 \cdot 2 = 1$  and therefore  $\sum_{i=1}^n i = \frac{1}{2}n(n+1)$ .

Inductive step: Suppose now as inductive hypothesis that

$$\sum_{i=1}^k i = \frac{1}{2}k(k+1)$$

for some positive integer  $k$ . Then

$$\begin{aligned} \sum_{i=1}^{k+1} i &= \sum_{i=1}^k i + (k+1) \quad (\text{by definition}) \\ &= \frac{1}{2}k(k+1) + (k+1) \quad (\text{by inductive hypothesis}) \\ &= \frac{1}{2}(k+1)(k+2) \end{aligned}$$

and so

$$\sum_{i=1}^{k+1} i = \frac{1}{2}(k+1)(k+2)$$

as required.

Conclusion: Hence, by induction,  $\sum_{i=1}^n i = \frac{1}{2}n(n+1)$  for all positive integers  $n$ .  $\square$

Inductive definitions are implicit in the definitions of several very common functions involving the non-negative integers.

Here are some familiar examples.

**Definition 5.3.3** For any real number  $x$ , the powers  $\boxed{x^n}$  for non-negative† integers  $n$  are defined inductively by:

- (i)  $x^0 = 1$ ; and
- (ii)  $x^{k+1} = x \cdot x^k$  for non-negative integers  $k$ .

If you look back to the proof of Proposition 5.1.2 you will see that we made use of this in the form  $2^{k+1} = 2 \times 2^k$ .

In the expression  $x^n$ , the number  $n$  is called the *index* or the *exponent* and the number  $x$  is called the *base*.

**Definition 5.3.4** For non-negative integers  $n$ , the numbers  $\boxed{\text{factorial } n}$ , written  $\boxed{n!}$ , are defined inductively by:

- (i)  $0! = 1$ ; and
- (ii)  $(k+1)! = (k+1) \times k!$  for non-negative integers  $k$ .

Notice that  $n = 0$  is the base case in this definition.

Indeed, even the basic operations of addition and multiplication of integers can be defined inductively starting from the notion of successor (see Definition 9.4.3).

#### 5.4 The strong induction principle

For completeness at this point we consider another variant of the inductive method although if the reader is meeting this method of proof for the first time the remainder of this chapter (apart from the exercises) should be omitted at first reading. The method will not be used subsequently until the proof of Proposition 23.1.2.

Sometimes we discover that  $P(k+1)$  is not implied by  $P(k)$  alone but is implied by the truth of  $P(k)$  together with some or all of  $P(1)$ ,  $P(2)$ , ...,  $P(k-1)$ . In this case we work with the following version of induction.

**Axiom 5.4.1 (The strong induction principle)** Suppose that  $P(n)$

† This definition allows  $x = 0$  and  $n = 0$  and sets  $0^0$  to be 1. This is the convention only in certain contexts such as Exercise 5.4 where we use the notation  $\sum_{i=0}^n x^i$  to represent the geometric progression  $1 + x + x^2 + \dots + x^n$  even when  $x = 0$ . However, some care is required and in certain other contexts it is not possible to attach a sensible meaning to  $0^0$ . A little more is said about this difficulty in the solution to Exercise 5.6 and in a footnote to Example 9.2.4(b).

is a statement involving a general positive integer  $n$ . Then  $P(n)$  is true for all positive integers  $n$  if

- (i)  $P(1)$  is true, and
- (ii)  $[P(n) \text{ holds for all positive integers } n \leq k] \Rightarrow P(k+1)$ , for all positive integers  $k$ .

It is not difficult to see that this is equivalent to Axiom 5.1.1.

In this case the basic template is as follows.

*Proof* We use (strong) induction on  $n$ .

Base case: [Prove the statement  $P(1)$ ]

Inductive step: Suppose now as inductive hypothesis that  $[P(n) \text{ is true for all positive integers } n \leq k]$  for some positive integer  $k$ . Then [deduce that  $P(k+1)$  is true]. This proves the inductive step.

Conclusion: Hence, by induction,  $[P(n) \text{ is true}]$  for all positive integers  $n$ . □

As an illustration of the use of this form of induction we introduce the Fibonacci numbers.

**Definition 5.4.2** For each positive integer  $n$  define the number  $u_n$  inductively as follows.

$$\begin{aligned} u_1 &= 1, \\ u_2 &= 1, \\ u_{k+1} &= u_{k-1} + u_k \quad \text{for } k \geq 2. \end{aligned}$$

The beginning of this sequence of numbers is 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, ..., and the sequence is known as the *Fibonacci sequence*. The important observation which concerns us now is that each number

† Leonardo of Pisa (ca. 1180–1250), who published his best known book *Liber abaci* in 1202, wrote under the name of Fibonacci. He was largely responsible for introducing Hindu–Arabic algebra and numerals to Europe and is often considered the greatest European mathematician of the Middle Ages. His name was attached to these numbers by the nineteenth century number theorist Edouard Lucas because of a problem in the *Liber abaci* concerned with the reproduction of rabbits whose solution involves them (see Problems I, Question 24). The early numbers occur frequently in nature, for example the number of petals in most flowers is a Fibonacci number. Also the sequence has many beautiful mathematical properties and there is even a scholarly journal *The Fibonacci Quarterly* devoted to articles about the sequence and related topics.

is determined not simply by the previous one as in a simple inductive definition but by the previous *two* numbers. This means that we must specify the first two numbers as the base case of the definition.

**Proposition 5.4.3 (The Binet† formula)** *The Fibonacci numbers are given by the following formula:*

$$u_n = (\alpha^n - \beta^n)/\sqrt{5},$$

where  $\alpha = (1 + \sqrt{5})/2$  and  $\beta = (1 - \sqrt{5})/2$ .

It should be observed that  $\alpha$  and  $\beta$  are the roots of the equation  $x^2 - x - 1 = 0$  and so  $\alpha^2 = \alpha + 1$  and  $\beta^2 = \beta + 1$ . The reader may find it remarkable that this formula for these *integers* involves  $\sqrt{5}$  which is most certainly not an integer; in fact, as we will see in Chapter 13, it is not even a rational number! There is in fact a general procedure for finding a general formula for sequences defined in this way (these are called *linear recursive sequences* or *linear difference equations*) which bears a close resemblance to elementary methods for solving linear differential equations. This method leads to a quadratic equation when each term of the sequence depends on the previous two terms.

The number  $\alpha$  was called *the ratio* by the Greeks and in the sixteenth century became known as *the golden ratio*. A rectangle whose sides are in this ratio is called a *golden rectangle* and a construction for such a rectangle is given in Book Six of the *Elements* of Euclid. Such a rectangle was considered to have ideal proportions: it is characterized by the fact that on the removal from the end of the rectangle of a square with sides equal to the shorter sides of the rectangle the remaining rectangle is similar to the original rectangle (and so is also a golden rectangle). As  $n$  increases, the ratio of successive Fibonacci numbers  $u_{n+1}/u_n$  approaches the number  $\alpha$ .

*Proof* The above formula may be proved by strong induction on  $n$ . In this case  $P(n)$  is the statement in the proposition.

Base case: For  $n = 1$ , the formula gives  $(\alpha - \beta)/\sqrt{5} = ((1 + \sqrt{5}) - (1 - \sqrt{5}))/2\sqrt{5} = 1 = u_1$ .

Since the inductive formula does not apply until  $u_3$  we have also to deal with the case  $n = 2$  by calculation. For  $n = 2$ , the formula gives  $(\alpha^2 - \beta^2)/\sqrt{5} = (\alpha - \beta)/\sqrt{5} = 1 = u_2$  using the facts that  $\alpha^2 = \alpha + 1$  and  $\beta^2 = \beta + 1$ .

† J.P.M. Binet was a French mathematician of the early nineteenth century.

Inductive step: Now suppose as inductive hypothesis that the formula holds for all positive integers  $n$  such that  $n \leq k$  for some positive integer  $k \geq 2$ . Then

$$\begin{aligned} u_{k+1} &= u_{k-1} + u_k \quad (\text{by definition of the sequence}) \\ &= ((\alpha^{k-1} - \beta^{k-1}) + (\alpha^k - \beta^k)) / \sqrt{5} \\ &\quad (\text{by inductive hypothesis}) \\ &= ((\alpha^{k-1}(1 + \alpha) - \beta^{k-1}(1 + \beta)) / \sqrt{5} \\ &= (\alpha^{k+1} - \beta^{k+1}) / \sqrt{5} \\ &\quad (\text{again using } \alpha^2 = \alpha + 1 \text{ and } \beta^2 = \beta + 1) \end{aligned}$$

as required to prove the formula for  $n = k + 1$ .

Conclusion: Hence, by induction, the formula holds for all positive integers  $n$ .  $\square$

### Exercises

**5.1** Prove by induction on  $n$  that, for all positive integers  $n$ ,  $n^3 - n$  is divisible by 3.

**5.2** Prove by induction on  $m$  that  $m^3 \leq 2^m$  for  $m \geq 10$ .

**5.3** Prove by induction on  $n$  that, for all positive integers  $n$ ,  $n \geq 1$ .

**5.4** Prove by induction on  $n$  that, for any real number  $x \neq 1$  and for integers  $n \geq 0$ ,

$$\sum_{i=0}^n x^i = \frac{1 - x^{n+1}}{1 - x}.$$

**5.5** Prove that, for any real numbers  $a$  and  $b$  and for integers  $n \geq 0$ ,

$$\sum_{i=0}^n (a + ib) = \frac{1}{2}(n + 1)(2a + nb).$$

**5.6** For non-negative integers  $n$  define the number  $u_n$  inductively as follows.

$$\begin{aligned} u_0 &= 0, \\ u_{k+1} &= 3u_k + 3^k \quad \text{for } k \geq 0. \end{aligned}$$

Prove that  $u_n = n3^{n-1}$  for all non-negative integers  $n$ .

**5.7** Prove by induction on  $n$  that, for any real numbers  $x$  and  $y$  and for non-negative integers  $m$  and  $n$ :

- (i)  $x^n y^n = (xy)^n$ ;
- (ii)  $x^{m+n} = x^m x^n$ ;
- (iii)  $(x^m)^n = x^{mn}$ .

[These statements are known as the laws of exponents or the laws of indices.]

## Problems I: Mathematical statements and proofs

1. By using truth tables prove that, for all statements  $P$  and  $Q$ , the statement ' $P \Rightarrow Q$ ' and its *contrapositive* ' $(\text{not } Q) \Rightarrow (\text{not } P)$ ' are equivalent. In Example 1.2.3 identify which statement is the contrapositive of statement (i)  $(f(a) = 0 \Rightarrow a > 0)$ . Find another pair of statements in that list which are the contrapositives of each other.
2. By using truth tables prove that, for all statements  $P$  and  $Q$ , the three statements (i) ' $P \Rightarrow Q$ ', (ii) ' $(P \text{ or } Q) \Leftrightarrow Q$ ' and (iii) ' $(P \text{ and } Q) \Leftrightarrow P$ ' are equivalent.
3. Prove that the three basic connectives 'or', 'and' and 'not' can all be written in terms of the single connective 'notand' where ' $P$  notand  $Q$ ' is interpreted as 'not  $(P \text{ and } Q)$ '.
4. Prove the following statements concerning positive integers  $a$ ,  $b$  and  $c$ .
  - (i)  $(a \text{ divides } b) \text{ and } (a \text{ divides } c) \Rightarrow a \text{ divides } (b + c)$ .
  - (ii)  $(a \text{ divides } b) \text{ or } (a \text{ divides } c) \Rightarrow a \text{ divides } bc$ .
5. Which of the following conditions are *necessary* for the positive integer  $n$  to be divisible by 6 (proofs not necessary)?
  - (i) 3 divides  $n$ .
  - (ii) 9 divides  $n$ .
  - (iii) 12 divides  $n$ .
  - (iv)  $n = 12$ .
  - (v) 6 divides  $n^2$ .
  - (vi) 2 divides  $n$  and 3 divides  $n$ .

- (vii) 2 divides  $n$  or 3 divides  $n$ .

Which of these conditions are *sufficient*?

- 6.** Use the properties of addition and multiplication of real numbers given in Properties 2.3.1 to deduce that, for all real numbers  $a$  and  $b$ ,

- (i)  $a \times 0 = 0 = 0 \times a$ ,
- (ii)  $(-a)b = -ab = a(-b)$ ,
- (iii)  $(-a)(-b) = ab$ .

- 7.** Prove by contradiction the following statement concerning an integer  $n$ .

$$n^2 \text{ is even} \Rightarrow n \text{ is even.}$$

[You may suppose that an integer  $n$  is odd if and only if  $n = 2q + 1$  for some integer  $q$ . This is proved later as Proposition 11.3.4.]

- 8.** Prove the following statements concerning a real number  $x$ .

- (i)  $x^2 - x - 2 = 0 \Leftrightarrow x = -1 \text{ or } x = 2$ .
- (ii)  $x^2 - x - 2 > 0 \Leftrightarrow x < -1 \text{ or } x > 2$ .

- 9.** Prove by contradiction that there does not exist a largest integer.

[Hint: Observe that for any integer  $n$  there is a greater one, say  $n + 1$ . So begin your proof]

Suppose for contradiction that there is a largest integer. Let this integer be  $n$ . ...]

- 10.** What is wrong with the following proof that 1 is the largest integer?

Let  $n$  be the largest integer. Then, since 1 is an integer we must have  $1 \leq n$ . On the other hand, since  $n^2$  is also an integer we must have  $n^2 \leq n$  from which it follows that  $n \leq 1$ . Thus, since  $1 \leq n$  and  $n \leq 1$  we must have  $n = 1$ . Thus 1 is the largest integer as claimed.

What does this argument prove?

- 11.** Prove by contradiction that there does not exist a smallest positive real number.

- 12.** Prove by induction on  $n$  that, for all positive integers  $n$ , 3 divides  $4^n + 5$ .

**13.** Prove by induction on  $n$  that  $n! > 2^n$  for all integers  $n$  such that  $n \geq 4$ .

**14.** Prove Bernoulli's inequality

$$(1 + x)^n \geq 1 + nx$$

for all non-negative integers  $n$  and real numbers  $x > -1$ .

**15.** For which non-negative integer values of  $n$  is  $n! \geq 3^n$ ?

**16.** Prove by induction on  $n$  that

$$\sum_{i=1}^n \frac{1}{i(i+1)} = \frac{n}{n+1},$$

for all positive integers  $n$ .

**17.** For a positive integer  $n$  the number  $a_n$  is defined inductively by

$$\begin{aligned} a_1 &= 1, \\ a_{k+1} &= \frac{6a_k + 5}{a_k + 2} \quad \text{for } k \text{ a positive integer.} \end{aligned}$$

Prove by induction on  $n$  that, for all positive integers, (i)  $a_n > 0$  and (ii)  $a_n < 5$ .

**18.** Given a sequence of numbers  $a(1), a(2), \dots$ , the number  $\prod_{i=1}^n a(i)$  is defined inductively by

$$(i) \prod_{i=1}^1 a(i) = a(1), \text{ and}$$

$$(ii) \prod_{i=1}^{k+1} a(i) = \left( \prod_{i=1}^k a(i) \right) a(k+1) \text{ for } k \geq 1.$$

Prove that  $\prod_{i=1}^n (1 + x^{2^{i-1}}) = (1 - x^{2^n})/(1 - x)$  for  $x \neq 1$ .

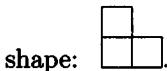
What happens if  $x = 1$ ?

**19.** Prove that

$$\prod_{i=2}^n \left(1 - \frac{1}{i^2}\right) = \frac{n+1}{2n}$$

for integers  $n \geq 2$ .

**20.** Prove that, for a positive integer  $n$ , a  $2^n \times 2^n$  square grid with any one square removed can be covered using L-shaped tiles of the following



**21.** Suppose that  $x$  is a real number such that  $x + 1/x$  is an integer. Prove by induction on  $n$  that  $x^n + 1/x^n$  is an integer for all positive integers  $n$ .

[For the inductive step consider  $(x^k + 1/x^k)(x + 1/x)$ .]

**22.** Prove that

$$\frac{1}{n} \sum_{i=1}^n x_i \geq \left( \prod_{i=1}^n x_i \right)^{1/n}$$

for positive integers  $n$  and positive real numbers  $x_i$ .

[It does not seem to be possible to give a direct proof of this result using induction on  $n$ . However, it can be proved for  $n = 2^m$  for  $m \geq 0$  by induction on  $m$ . The general result now follows by proving the converse of the usual inductive step: if the result holds for  $n = k + 1$ , where  $k$  is a positive integer, then it holds for  $n = k$ .]

**23.** For non-zero real numbers  $x$  we may extend Definition 5.3.3 to a definition of powers  $x^n$  for all integers  $n$  by defining  $x^{-m} = 1/x^m$  for integers  $m > 0$ . With this definition prove the laws of exponents for any non-zero real numbers  $x$  and  $y$  and integers  $m$  and  $n$ :

- (i)  $x^n y^n = (xy)^n$ ;
- (ii)  $x^{m+n} = x^m x^n$ ;
- (iii)  $(x^m)^n = x^{mn}$ .

[Hint: Start from Exercise 5.7.]

**24.** Fibonacci's rabbit problem may be stated as follows:

How many pairs of rabbits will be produced in a year, beginning with a single pair, if in every month each pair bears a new pair which become productive from the second month on?

Assuming that no rabbits die, express the number after  $n$  months as a Fibonacci number and hence answer the problem. Using a calculator and the Binet formula (Proposition 5.4.3) find the number after three years.

**25.** Let  $u_n$  be the  $n$ th Fibonacci number (Definition 5.4.2). Prove, by induction on  $n$  (without using the Binet formula Proposition 5.4.3), that

$$u_{m+n} = u_{m-1}u_n + u_m u_{n+1}$$

for all positive integers  $m$  and  $n$ .

Deduce, again using induction on  $n$ , that  $u_m$  divides  $u_{mn}$ .

**26.** Suppose that  $n$  points on a circle are all joined in pairs. The points are positioned so that no three joining lines are concurrent in the interior of the circle. Let  $a_n$  be the number of regions into which the interior of the circle is divided. Draw diagrams to find  $a_n$  for  $n \leq 6$ .

Prove that  $a_n$  is given by the following formula.

$$\begin{aligned} a_n &= n + \binom{n-1}{2} + \binom{n-1}{3} + \binom{n-1}{4} \\ &= 1 + n(n-1)(n^2 - 5n + 18)/24. \end{aligned}$$

[Here  $\binom{m}{r}$  denotes the binomial coefficient  $\frac{m!}{r!(m-r)!}$  discussed in more detail in Chapter 12.]



# **Part II**

## Sets and functions



# 6

## The language of set theory

The students' task in learning set theory is to steep themselves in unfamiliar and essentially shallow generalities till they become so familiar that they can be used with almost no conscious effort.

Paul R. Halmos, *Naive set theory* (adapted slightly).

The language of set theory is used throughout mathematics. Many general results involve 'an integer  $n$ ' or 'a real number  $a$ ' and, to start with, set theory notation provides a simple way of asserting for example that  $n$  is an integer. However, it turns out that this language is remarkably flexible and powerful and in much mathematics it is indispensable for a proper expression of the ideas involved.

In the second part of this book we introduce the basic vocabulary of the language of sets and functions. The third part of the book will then provide some experience in using this language as we use it to give a precise formulation of the idea of counting, one of the earliest mathematical concepts.

### 6.1 Sets

At this introductory level it is sufficient to define the notion of *set* as any well-defined collection of objects. We can think of a set as a box containing certain objects. In this section some ways of specifying sets are introduced and also some frequently used notation.

We frequently use a single letter to denote a set. This represents a further stage of mathematical abstraction. The reader will already have accepted the abstract notion of a positive integer, for example 'two' is abstracted from 'two apples' and 'two chairs'. Now we move on to consider the set of all positive integers as a single mathematical object. Each time there is further step of abstraction like this it takes time to

become familiar with new sorts of objects. At the end of this chapter we will make a further abstraction step when we consider sets which are collections of sets!

Particular sets which are often used have standard symbols to represent them. In this book the following will be used.

$\mathbb{Z}$  denotes the set of all integers.<sup>†</sup>

$\mathbb{Z}^+$  denotes the set of all positive integers (natural<sup>‡</sup> or counting numbers): 1, 2, 3, etc.

$\mathbb{Z}^{\geq}$  denotes the set of all non-negative integers: 0, 1, 2, 3, etc.

$\mathbb{Q}$  denotes the set of all rational numbers (fractions).

$\mathbb{R}$  denotes the set of all real numbers (i.e. numbers expressible as infinite decimals).

$\mathbb{R}^+$  denotes the set of all positive real numbers.

$\mathbb{R}^{\geq}$  denotes the set of all non-negative real numbers.

$\mathbb{C}$  denotes the set of all complex numbers.

The objects in a set are called the *elements*, *members* or *points* of the set. We write

$$x \in E$$

to denote the fact that the object  $x$  is an element of the set  $E$ . Thus for example ' $a \in \mathbb{R}$ ' is read ' $a$  is an element of the set of real numbers' or more simply just ' $a$  is a real number'. The symbol ' $\in$ ', first used in this way by the Italian mathematician Giuseppe Peano towards the end of the nineteenth century, is a variant of the Greek letter epsilon and care should normally be taken to distinguish it from that letter which is usually written ' $\varepsilon$ ' or ' $\epsilon$ '. However, some books do not make this distinction and use epsilon in place of ' $\in$ '. It is common to use upper case letters to represent sets and lower case to represent elements but this is not always appropriate: in geometry it is usual to use upper case letters to represent points and lower case letters to represent lines (which are sets of points); another exception occurs when a set is itself considered as an element of another set.

The negation of the statement  $x \in A$  is written  $x \notin A$ . Thus  $\sqrt{2} \notin \mathbb{Q}$  is the statement that  $\sqrt{2}$  is not a rational number.

There are basically three ways of specifying a set: we can list the

<sup>†</sup> This symbol comes from 'Zahlen', the German word for 'numbers'.

<sup>‡</sup> Some people include the number 0 in the set of natural numbers but this seems to me unnatural as we usually start counting at 1. Because of this ambiguity, in this book we will normally refer to 'the non-negative integers' or 'the positive integers' depending on whether or not the number 0 is included.

elements, specify a condition for membership, or give a formula or algorithm constructing the elements of the set.

*(a) Listing the elements of a set*

When we list the elements of a set we denote the set by enclosing the elements in curly brackets. Thus

$$A = \{1, 3, \pi, -14\}$$

is the statement that  $A$  is the set whose four elements are 1, 3,  $\pi$  and  $-14$ . In this case we have  $1 \in A$  and  $2 \notin A$ . Notice that the order in which the elements are listed is unimportant and repeating an element makes no difference. Thus, for the above set  $A$  we also have

$$A = \{\pi, 3, -14, 1\} = \{\pi; 3, \pi, 1, -14, -14\}.$$

On the face of it the listing notation is only practical for sets with a small number of elements. However it can be extended to large or even infinite sets. For example we can write

$$\mathbb{Z}^+ = \{1, 2, 3, \dots\}$$

where as usual the dots are read ‘and so on’.

*(b) The conditional definition of a set*

Alternatively, a set may be described by specifying some condition which determines whether or not an object is an element of the set. Let us illustrate this with an example. Consider the set  $B$  defined as follows:

$$B = \{n \in \mathbb{Z} \mid 0 < n < 6\}.$$

Here we write  $n \in \mathbb{Z}$  simply to indicate the sort of objects that we are considering. The statement  $0 < n < 6$  is a predicate. The definition means that  $B$  is the set of integers which when substituted in this predicate give a true proposition, i.e. given an integer  $n$ ,

$$n \in B \Leftrightarrow 0 < n < 6.$$

Of course we could have described this particular set by listing the elements,

$$B = \{1, 2, 3, 4, 5\},$$

but in some cases this is difficult or impossible.

In the conditional definition of a set the vertical line† ‘|’ is read as ‘such that’ and so the above definition would be read as ‘ $B$  is the set of integers  $n$  such that  $0 < n < 6$ ’ or just ‘ $B$  is the set of integers between 0 and 6’. This last reading, which makes no reference to the variable ‘ $n$ ’, demonstrates that the symbol  $n$  in this definition represents a *dummy variable*: its only rôle is to indicate the internal logic of the definition and it can be replaced by any other symbol (not already in use) without any change in meaning. Thus we could equally write

$$B = \{ \alpha \in \mathbb{Z} \mid 0 < \alpha < 6 \}.$$

Notice that, although  $0 < n < 6$  is a predicate and so a mathematical statement in the sense of Chapter 1,  $\{ n \in \mathbb{Z} \mid 0 < n < 6 \}$  is not a statement; it is simply a mathematical object and can be no more true or false than the number 2 can be true or false. However, we can make statements about this object, such as ‘ $2 \in \{ n \in \mathbb{Z} \mid 0 < n < 6 \}$ ’ which is true, or ‘ $m \in \{ n \in \mathbb{Z} \mid 0 < n < 6 \}$ ’ which is equivalent to ‘ $m \in \mathbb{Z}$  and  $0 < m < 6$ ’.

### *(c) The constructive definition of a set*

The other systematic method of describing a set is to give a formula (or more generally an algorithm) constructing the elements of the set. For example

$$\{ n^2 \mid n \in \mathbb{Z} \} = \{ 0, 1, 4, 9, 16, \dots \}$$

is the set of integer squares, which means that the formula  $n^2$  constructs the elements of the set as  $n$  takes all possible integer values, i.e. an element is in the set if and only if it can be written as  $n^2$  for some integer  $n$ . Notice that in this case all non-zero elements in the set arise twice. This makes no difference to the set so that for example  $\{ n^2 \mid n \in \mathbb{Z}^{\geq} \}$  defines the same set.

Similarly,

$$\{ 2q \mid q \in \mathbb{Z} \} = \{ \dots, -4, -2, 0, 2, 4, 6, \dots \}$$

is the set of even integers, and

$$\{ a/b \mid a, b \in \mathbb{Z}, b \neq 0 \}$$

is the set of rational numbers,  $\mathbb{Q}$ . This last example illustrates the

† Alternative notations are to use a colon ‘:’ or a semi-colon ‘;’ here instead of ‘|’ as follows:  $B = \{ n \in \mathbb{Z} : 0 < n < 6 \}$  or  $\{ n \in \mathbb{Z} ; 0 < n < 6 \}$ .

convention that when more than one condition is listed after the symbol ‘|’ this means that all the conditions must be satisfied for an object to be in the set so that the word ‘and’ is understood.

### ***Equality of sets***

Quite commonly problems in mathematics take the form of seeking to pass between different ways of specifying the elements of a set. For example when we learn how to solve quadratic equations of the form  $ax^2 + bx + c = 0$ , where  $a$ ,  $b$  and  $c$  are given real numbers, we are learning how to list the elements of the set  $\{x \in \mathbb{R} \mid ax^2 + bx + c = 0\}$ , or possibly the set  $\{x \in \mathbb{C} \mid ax^2 + bx + c = 0\}$  if we allow complex solutions.

**Definition 6.1.1** Two sets  $A$  and  $B$  are **equal**, written  $A = B$ , if they have precisely the same elements, i.e.  $A = B$  means  $x \in A \Leftrightarrow x \in B$ .

To put it another way: a set is determined by its elements. Notice that this means that to show that two sets  $A$  and  $B$  are equal it is necessary to prove two things (although they can often be done together in simple cases): every element of  $A$  is an element of  $B$  and conversely every element of  $B$  is an element of  $A$ .

**Example 6.1.2**  $\{x \in \mathbb{R} \mid x^2 - x - 2 = 0\} = \{-1, 2\}$ .

**Constructing a proof.** This is another way of stating that, for  $x$  a real number,

$$x^2 - x - 2 = 0 \Leftrightarrow x = -1 \text{ or } x = 2.$$

This is easily proved by the factorization method.

*Proof* For  $x$  a real number,  $x^2 - x - 2 = 0 \Leftrightarrow (x - 2)(x + 1) = 0 \Leftrightarrow x - 2 = 0$  or  $x + 1 = 0$  (by Proposition 4.4.1)  $\Leftrightarrow x = 2$  or  $x = -1$ . Hence  $x^2 - x - 2 = 0$  if and only if  $x = -1$  or  $x = 2$ , as required.  $\square$

At this stage it is convenient to introduce two other ideas from set theory.

**Definition 6.1.3** The **empty set** is the unique set which has no elements at all. It is denoted by the symbol  $\emptyset$ .

Thus the statement that the quadratic equation  $x^2 + 2x + 2 = 0$  has no real solutions may be written  $\{x \in \mathbb{R} \mid x^2 + 2x + 2 = 0\} = \emptyset$ .

Take care to distinguish ' $\emptyset$ ' which is a variant of a Scandinavian letter from the Greek letter phi written ' $\phi$ '.

**Definition 6.1.4** Given sets  $A$  and  $B$  we say that  $A$  is a subset of  $B$ ,

written  $A \subseteq B$ , or  $B \supseteq A$ , when every element of  $A$  is an element of  $B$ , i.e.  $x \in A \Rightarrow x \in B$ . If  $A$  and  $B$  are in addition unequal so that  $B$  contains some element not contained in  $A$ , then we say that  $A$  is a proper subset of  $B$  and write  $A \subset B$ .†

Thus

$$A = B \Leftrightarrow (A \subseteq B \text{ and } B \subseteq A).$$

It is important to distinguish between the symbols  $\in$  and  $\subseteq$  although they are closely related as follows:

$$a \in A \Leftrightarrow \{a\} \subseteq A.$$

If sets are defined by predicates then there is a correspondence between the notions of 'implication' and 'subset': the universal statement that  $P(a) \Rightarrow Q(a)$  for all  $a \in A$  is equivalent to the statement that the set  $\{a \in A \mid P(a)\}$  is a subset of  $\{a \in A \mid Q(a)\}$ .

Notice that if  $A \subseteq B$  and  $B \subseteq C$  then  $A \subseteq C$  since  $x \in A \Rightarrow x \in B$  and  $x \in B \Rightarrow x \in C$  together imply that  $x \in A \Rightarrow x \in C$ . Furthermore  $\emptyset \subseteq A$  for all sets  $A$  whereas  $A \subseteq \emptyset$  only if  $A = \emptyset$ .

## 6.2 Operations on sets

**Definition 6.2.1** Given two sets  $A$  and  $B$  we can form the set of elements which lie both in  $A$  and in  $B$ . This is called the intersection of  $A$  and  $B$  and is denoted by  $A \cap B$ . Thus

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

† Some mathematicians use  $\subset$  where  $\subseteq$  is used in this book. In fact when Peano originally introduced this notation he used it the other way round, writing  $A \supset B$  to indicate that  $A$  is a subset of  $B$ !

Two sets  $A$  and  $B$  are said to be disjoint if  $A \cap B = \emptyset$ , i.e.  $A$  and  $B$  have no elements in common.

**Definition 6.2.2** Given two sets  $A$  and  $B$  we can form the set of elements which lie in  $A$  or lie in  $B$ . This is called the union of  $A$  and  $B$  and is denoted by  $A \cup B$ . Thus

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

**Definition 6.2.3** Given two sets  $A$  and  $B$  we can form the set of elements which lie in  $A$  but not in  $B$ . This is called the difference of  $A$  and  $B$  and is denoted by†  $A - B$ . Thus

$$A - B = \{x \mid x \in A \text{ and } x \notin B\}.$$

Notice that  $A \cap A = A = A \cup A$ ,  $A \cap \emptyset = \emptyset$ ,  $A \cup \emptyset = A$ ,  $A - A = \emptyset$  and  $A - \emptyset = A$ .

**Proposition 6.2.4** Given any two sets  $A$  and  $B$ , the three sets  $A \cap B$ ,  $A - B$  and  $B - A$  are pairwise disjoint (i.e. each pair of these sets is disjoint) and

$$A \cup B = (A \cap B) \cup (A - B) \cup (B - A).$$

Probably the simplest way to prove a statement like this is by means of truth tables as follows.

*Proof* Consider the truth tables on the following page.

The fact that the final two columns of the second table are the same tells us that the above equality of sets holds. The fact that no row has more than one  $T$  in the third, fourth and fifth columns of the first table

† The reader should be aware that this notation is sometimes used in algebra to denote the set  $\{a - b \mid a \in A, b \in B\}$ . In this case the difference of the two sets  $A$  and  $B$  is denoted by  $A \setminus B$ .

tells us that the three corresponding sets are disjoint.

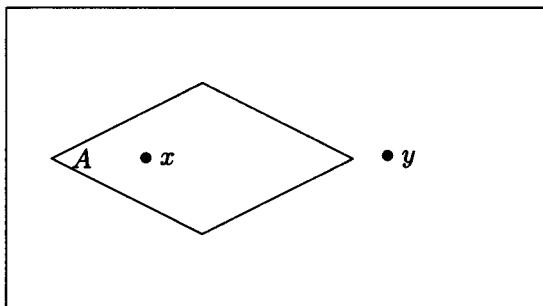
| $x \in A$ | $x \in B$ | $x \in A \cap B$ | $x \in A - B$ | $x \in B - A$ |
|-----------|-----------|------------------|---------------|---------------|
| T         | T         | T                | F             | F             |
| T         | F         | F                | T             | F             |
| F         | T         | F                | F             | T             |
| F         | F         | F                | F             | F             |

| $x \in A$ | $x \in B$ | $x \in (A \cap B) \cup (A - B) \cup (B - A)$ | $x \in A \cup B$ |
|-----------|-----------|--|------------------|
| T         | T         | T  | T                |
| T         | F         | T  | T                |
| F         | T         | T  | T                |
| F         | F         | F  | F                |

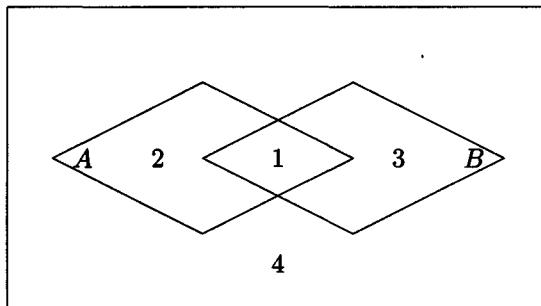
□

In the truth tables in the above proof the columns are headed by statements  $x \in A$ ,  $x \in B$ , etc. which can be true or false, not simply by the names of the sets  $A$ ,  $B$ , etc. which would not be statements.

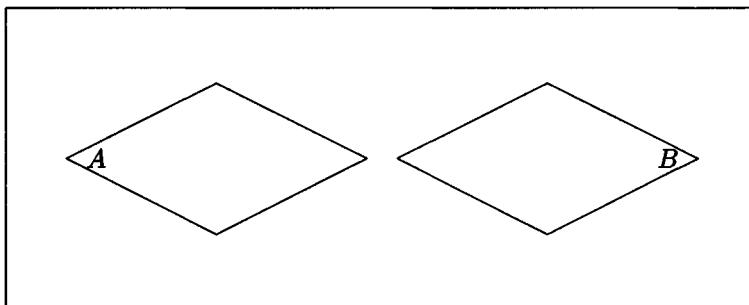
This proof illustrates the close relationship between the logical connectives introduced in Chapter 1 and the operations on sets defined above in terms of those connectives. Proofs of the above type are usually illustrated by a *Venn diagram*. We indicate a set  $A$  by the interior region of some curve drawn on the page so that the elements of the set correspond to points in the region. Thus for example in the following diagram  $x \in A$  but  $y \notin A$ .



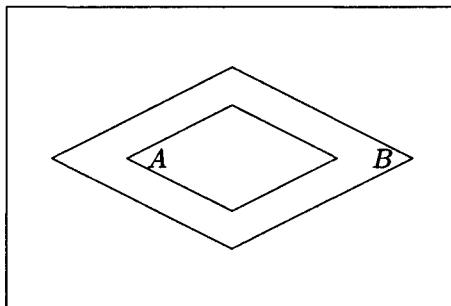
Two general sets  $A$  and  $B$  are described by two overlapping regions. Consider the following diagram.



Here the interior of the left-hand rhombus (regions 1 and 2) denotes  $A$  and the interior of the right-hand rhombus (regions 1 and 3) denotes  $B$ . Then the four rows of the above truth tables correspond to the four regions in the diagram. Thus region 1 corresponds to  $A \cap B$ , region 2 corresponds to  $A - B$ , region 3 corresponds to  $B - A$  and the three regions 1, 2 and 3 together give  $A \cup B$ . The fact that regions 1, 2 and 3 are disjoint corresponds to the fact that the three sets on the right-hand side of Example 6.2.4 are disjoint and the fact that these three regions together make up  $A \cup B$  corresponds to the equality of sets in Proposition 6.2.4. From this point of view the result is essentially obvious. Care is needed in using this sort of diagrammatic proof in more elaborate examples because unless proper care is taken it is sometimes possible to draw a diagram that does not include all possible regions or some feature of the diagram has nothing to do with the set theory (see for example the solution to Exercise 6.6). Notice that this diagram does not imply that  $A \cap B$  is non-empty for it may be that there are no elements in the region 1. However, if we were given that  $A \cap B = \emptyset$  then we could denote this by the following Venn diagram.



We can represent  $A \subseteq B$  as follows.



### 6.3 The power set

**Definition 6.3.1** The power set of a set  $X$ , denoted by  $\mathcal{P}(X)$ , is the set of all subsets of the set  $X$ . Thus  $A \in \mathcal{P}(X)$  is another way of writing  $A \subseteq X$ .

**Example 6.3.2** If  $X = \{a, b, c\}$  then

$$\mathcal{P}(X) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, X\}.$$

Notice that the empty set  $\emptyset$  is an element of the power set  $\mathcal{P}(X)$  for any set  $X$  since  $\emptyset \subseteq X$ . A set like  $\{a\}$  with a single element is called a *singleton*. It is important to distinguish the singleton set  $\{a\}$  from the element  $a$ . In particular the singleton  $\{\emptyset\}$  is to be distinguished from the empty set  $\emptyset$ : a box containing an empty box is not an empty box!

It is often the case that all the sets we are considering are subsets of some fixed set, say the set of real numbers. We then consider this to be the *universal set*.

**Definition 6.3.3** Once we have fixed a universal set  $U$  we can define the complement of any  $A \in \mathcal{P}(U)$ , denoted by  $A^c$ , to be the difference of  $U$  and  $A$ . Thus

$$A^c = U - A = \{x \in U \mid x \notin A\}.$$

For example, if the universal set is  $\mathbb{Z}$ , the set of integers, and  $E$  is the set of even integers, then the complement  $E^c$  is the set of odd integers.

The intersection, union and complement of subsets of some universal set  $U$  correspond to the logical connectives ‘and’, ‘or’ and ‘not’. The relationships between these operations may be summed up in the following theorem.

**Theorem 6.3.4** *Let  $A$ ,  $B$  and  $C$  be subsets of some universal set  $U$  (i.e.  $A, B, C \in \mathcal{P}(U)$ ). Then we have the following identities.*

- (i) associativity:  $A \cup (B \cup C) = (A \cup B) \cup C$ ,  $A \cap (B \cap C) = (A \cap B) \cap C$ .
- (ii) commutativity:  $A \cup B = B \cup A$ ,  $A \cap B = B \cap A$ .
- (iii) distributivity:  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ ,  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ .
- (iv) De Morgan laws:<sup>†</sup>  $(A \cup B)^c = A^c \cap B^c$ ,  $(A \cap B)^c = A^c \cup B^c$ .
- (v) complementation:  $A \cup A^c = U$ ,  $A \cap A^c = \emptyset$ .
- (vi) double complement:  $(A^c)^c = A$ .

These can be proved by truth tables or Venn diagrams. Alternatively they can be proved by using logical argument from the definitions. Let us illustrate this by writing out the proof of one part.

*Proof of  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ .*

Proof of ‘ $\subseteq$ ’: Suppose that  $x \in A \cap (B \cup C)$ . Then  $x \in A$  and  $x \in (B \cup C)$ . Since  $x \in B \cup C$ ,  $x \in B$  or  $x \in C$ . If  $x \in B$  then, since  $x \in A$  as well, we have  $x \in A \cap B$  and so  $x \in (A \cap B) \cup (A \cap C)$  as required. On the other hand, if  $x \notin B$ , then we must have  $x \in C$  and so, since also  $x \in A$ , we have  $x \in A \cap C$  and so  $x \in (A \cap B) \cup (A \cap C)$ .

Proof of ‘ $\supseteq$ ’: Suppose now that  $x \in (A \cap B) \cup (A \cap C)$ . Then  $x \in A \cap B$  or  $x \in A \cap C$ . If  $x \in A \cap B$ , then  $x \in A$  and  $x \in B$  so that  $x \in A$  and  $x \in B \cup C$  which gives  $x \in A \cap (B \cup C)$  as required. On the other hand if  $x \notin A \cap B$  then  $x \in A \cap C$  and again we get  $x \in A \cap (B \cup C)$ .  $\square$

We can now use these results to derive other set identities by algebraic manipulation. Here is an example.

### Proposition 6.3.5

$$(A \cup B) \cap (C \cup D) = (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D).$$

<sup>†</sup> These results was observed independently in the nineteenth century by the British mathematician Augustus De Morgan and the United States mathematician Benjamin Peirce (see C.B. Boyer and U.C. Merzbach, *A history of mathematics*, Wiley, Second edition 1989). They correspond to the equivalence of the statements ‘not ( $P$  and  $Q$ )’ and ‘(not  $P$ ) or (not  $Q$ )’ commented on in the solution to Exercise 1.2.

*Proof* First notice that commutativity means that distributivity on one side implies distributivity on the other side so that Theorem 6.3.4(iii) implies that  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ .

$$\begin{aligned} & (A \cup B) \cap (C \cup D) \\ &= (A \cap (C \cup D)) \cup (B \cap (C \cup D)) \quad \text{by distributivity} \\ &= (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D) \quad \text{by distributivity.} \end{aligned}$$

The associativity of the union operation means that we do not need any additional brackets here.  $\square$

### Exercises

**6.1** The following are standard subsets of the set of real numbers known as the real intervals with endpoints the real numbers  $a$  and  $b$ .

The *open interval*:  $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$ .

The *closed interval*:  $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$ .

The *right half-open interval*:  $[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\}$ .

The *left half-open interval*:  $(a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$ .

- (i) Prove that  $0 \notin (0, 1)$ ,  $0 \in [0, 1]$ ,  $0 \in [0, 1)$  and  $0 \notin (0, 1]$ .
- (ii) Find the elements of the set  $[a, b] - (a, b)$ .
- (iii) Prove that  $(a, b) = \emptyset$  if and only if  $a \geq b$ . [Hint: Prove the contrapositive.]  
Find the corresponding results for the other real intervals with endpoints  $a$  and  $b$ .
- (iv) Prove that, if  $a \leq b$ , then  $[a, b] \subseteq (c, d)$  if and only if  $c < a$  and  $b < d$ .

**6.2** Prove that

- (i)  $\{x \in \mathbb{R} \mid x^2 + x - 2 = 0\} = \{1, -2\}$ ,
- (ii)  $\{x \in \mathbb{R} \mid x^2 + x - 2 < 0\} = (-2, 1)$ ,
- (iii)  $\{x \in \mathbb{R} \mid x^2 + x - 2 > 0\} = \{x \in \mathbb{R} \mid x < -2\} \cup \{x \in \mathbb{R} \mid x > 1\}$ .

**6.3** Find predicates which determine the following subsets of the set of integers  $\mathbb{Z}$ : (i)  $\{3\}$ , (ii)  $\{1, 2, 3\}$ , (iii)  $\{1, 3\}$ .

**6.4** By using a truth table prove that  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ .

Draw a Venn diagram to illustrate the proof.

**6.5** Prove that

- (i)  $A \subseteq B \Leftrightarrow A \cup B = B$ ,
- (ii)  $A \subseteq B \Leftrightarrow A \cap B = A$ .

**6.6** Prove by contradiction that, if  $A \cap B \subseteq C$  and  $x \in B$ , then  $x \notin A - C$ .  
[Work from the definitions of ' $A \cap B$ ', ' $A - C$ ', and ' $\subseteq$ '.]

**6.7** Using the fact that an implication is equivalent to its contrapositive, prove that, for subsets of a universal set  $U$ ,  $A \subseteq B$  if and only if  $B^c \subseteq A^c$ .

# 7

## Quantifiers

In Chapter 1 a predicate was described as an expression containing one or more free variables; it becomes a proposition, and so is true or false, when a specific value is assigned to each free variable. Of course whether this proposition is true or is false usually depends on the values selected.

However, we saw in the last chapter that a proposition can be created from a predicate in another way – by making a statement about the set of values of the free variables which make it true. Many results in mathematics take the form of listing the values of this set, for example when we solve an equation. But often results simply address the question of whether there is any choice of values of the free variables resulting in a true proposition and whether there is any choice resulting in a false proposition. Statements that such values exist are known as existential statements. Statements that they do not can be thought of as universal statements. We met examples of universal statements when discussing implications in Chapter 2.

In this chapter we discuss general universal and existential statements.

### 7.1 Universal statements

Suppose that  $P(a)$  is a predicate with a single free variable  $a$  with possible values in a set  $A$ . Usually  $P(a)$  is true for some elements of the set  $A$  and false for others, and in the last chapter we described how such a predicate could be used to describe a subset of  $A$ , denoted by  $\{a \in A \mid P(a)\}$ , the subset of elements of  $A$  for which the statement  $P(a)$  is true. In certain cases this subset is the whole of  $A$ ; the statement that this occurs is a *universal statement*.

**Definition 7.1.1** The notation  $\boxed{\forall a \in A, P(a)}$  is an alternative way of writing

$$\{ a \in A \mid P(a) \} = A.$$

It is read: ‘for each element  $a$  in the set  $A$  the proposition  $P(a)$  is true’ or ‘ $P(a)$  is true for each  $a$  in the set  $A$ ’.

In this notation, the comma between ‘ $\forall a \in A$ ’ and the predicate ‘ $P(a)$ ’ is included simply to clarify the expression. Sometimes brackets are used. The symbol ‘ $\forall$ ’ is called the *universal quantifier symbol* and is read ‘for each’, ‘for every’, ‘for all’ or ‘for any’. In this book we will adopt the practice of putting quantifiers before the predicates to which they refer. However, we will usually write these statements in words and in this case show more flexibility about the order.

As an example consider Proposition 3.1.4 which asserts that  $a^2 > 0$  for any non-zero real number  $a$ . This is a statement about the set of non-zero real numbers,  $\mathbb{R} - \{0\}$ . The predicate is ‘ $a^2 > 0$ ’. We may write the statement in symbols as

$$\forall a \in \mathbb{R} - \{0\}, a^2 > 0$$

or equivalently

$$\{ a \in \mathbb{R} - \{0\} \mid a^2 > 0 \} = \mathbb{R} - \{0\}.$$

A third way of writing the statement is as the universal implication

$$a \in \mathbb{R} - \{0\} \Rightarrow a^2 > 0$$

and if you look back to Chapter 3 you will see that this is how we proved the statement.

Notice that in all three of these expressions  $a$  is a ‘dummy’ or ‘bound’ variable and can be changed to any other symbol without changing the meaning. For example

$$\forall x \in \mathbb{R} - \{0\}, x^2 > 0$$

is equivalent to the above statements.

## 7.2 Existential statements

We now turn to the negation of a universal statement: the statement that it is false. In Chapter 2 we considered the statement  $x > 0 \not\Rightarrow x \geq 1$  for real numbers  $x$  and explained that this meant that the universal

statement  $x > 0 \Rightarrow x \geq 1$ , which is shorthand for the statement  $\forall x \in \mathbb{R} (x > 0 \Rightarrow x \geq 1)$ , is false. Table 2.1.2 described the behaviour for different values of  $x$  and of course, for any  $x$  such that  $0 < x < 1$ ,  $x > 0$  is true and  $x \geq 1$  is false so that  $x > 0 \Rightarrow x \geq 1$  is false. However, to demonstrate that  $\forall x \in \mathbb{R} (x > 0 \Rightarrow x \geq 1)$  is false we simply have to show that there is a single value of  $x \in \mathbb{R}$  for which  $x > 0 \Rightarrow x \geq 1$  is false, i.e.  $x > 0$  and  $x \not\geq 1$ , in other words the set  $\{x \in \mathbb{R} \mid x > 0 \text{ and } x \not\geq 1\}$  is non-empty. The simplest way to do this is to give a specific element in the set, such as  $1/2$ :  $1/2 > 0$  and  $1/2 \not\geq 1$ . Such a statement, that the subset defined by a predicate is non-empty, is called an *existential statement*.

**Definition 7.2.1** The notation  $\boxed{\exists a \in A, P(a)}$  is an alternative way of writing

$$\{a \in A \mid P(a)\} \neq \emptyset.$$

It is read: ‘for some element  $a$  in the set  $A$  the proposition  $P(a)$  is true’ or ‘ $P(a)$  is true for some  $a$  in the set  $A$ ’.

The symbol ‘ $\exists$ ’ is called the *existential quantifier symbol* and is read ‘for some’, ‘for at least one’ or sometimes ‘there exists ... such that’.

**Remarks 7.2.2** Notice that the word ‘any’ sometimes indicates a universal statement and sometimes an existential statement.

The normal meaning of ‘any’ is ‘every’ as in ‘ $a^2 \geq 0$  for any real number  $a$ ’. This is a universal statement which can be written symbolically as ‘ $\forall a \in \mathbb{R}, a^2 \geq 0$ ’. However, in negative or interrogative statements ‘any’ is used idiomatically to mean ‘some’. For example, ‘There is not any real real number  $a$  such that  $a^2 < 0$ ’ is asserting that the existential statement ‘ $\exists a \in \mathbb{R}, a^2 < 0$ ’ is false. And ‘Is there any real number  $a$  such that  $a^2 = 2$ ?’ is asking whether the existential statement ‘ $\exists a \in \mathbb{R}, a^2 = 2$ ’ is true.

Fowler† gives some non-mathematical examples: ‘Have you any bananas?’ with the possible answers ‘No we haven’t any bananas’ and ‘Yes we have some bananas’.

Real care is required with questions involving ‘any’. ‘Is there any integer  $a$  such that  $a \geq 1$ ?’ seems clear enough and is asking whether ‘ $\exists a \in \mathbb{Z}, a \geq 1$ ’ is true. But ‘Is  $a \geq 1$  for any integer  $a$ ?’ seems less clear

† H.W. Fowler, *A dictionary of modern English usage* (revised by Ernest Gowers), Oxford University Press, Second edition 1968.

to me and might be taken to asking about the same statement as the first question, ' $\exists a \in \mathbb{Z}, a \geq 1$ ' (which is true) but might also be taken to be asking about ' $\forall a \in \mathbb{Z}, a \geq 1$ ' (which is false). Great care is needed in using 'for any' in interrogative statements.

### 7.3 Proving statements involving quantifiers

An enormous number of results in advanced mathematics take the form of asserting the truth or falsehood of some universal or existential statement; this is one of the factors which distinguishes advanced from elementary mathematics and many of the results in this book take this form. This section provides an overview of the main methods of proof but in effect much of the whole book is about proving such results.

#### (a) *Proving statements of the form $\forall a \in A, P(a)$*

We usually prove statements of this form by rewriting them in the form

$$a \in A \Rightarrow P(a).$$

An example of this is the proof of Proposition 3.1.4 which we have already discussed.

#### (b) *Proving statements of the form $\exists a \in A, P(a)$*

We often prove statements of this form by simply exhibiting a particular element  $a \in A$  for which  $P(a)$  is true. This is *proof by example*.

**Example 7.3.1** To prove  $\exists n \in \mathbb{Z}, n^2 = 9$ .

*Solution* Observe that  $3 \in \mathbb{Z}$  and  $3^2 = 9$  and so  $n = 3$  provides an example proving this statement.  $\square$

There are, however, less direct methods of proving existential statements such as the use of the counting arguments which will be considered in Chapter 11.

#### (c) *Proving statements involving both quantifiers*

Very many statements involve both quantifiers. Consider the result of Exercise 3.3.

**Proposition 7.3.2** *For integers  $n$ , if  $n$  is even then  $n^2$  is even.*

This is a universal implication:  $\forall n \in \mathbb{Z} (n \text{ is even} \Rightarrow n^2 \text{ is even})$ . However, the hypothesis that  $n$  is even is an existence statement, which may be written  $\exists q \in \mathbb{Z}, n = 2q$ . We can use this by making use of some specific integer  $q$  such that  $n = 2q$ . Thus we begin the proof of this result, by the direct method, as follows.

Suppose that  $n$  is an even integer. Then  $n = 2q$  for some integer  $q$ .

The conclusion which we are aiming for is the statement that  $n^2$  is even, which may be written, again spelling out the definition,  $\exists q \in \mathbb{Z}, n^2 = 2q$ . Recall that in such statements the symbol ‘ $q$ ’ is a dummy variable and its only rôle is to glue the notation together: in words we could write this as ‘ $n^2$  is twice some integer.’ We could replace  $q$  by any other symbol not already in use, for example  $\exists p \in \mathbb{Z}, n^2 = 2p$ . Indeed in this case we ought to do this since the symbol  $q$  is already in use: when we wrote ‘Then  $n = 2q$  for some integer  $q$ ’ we were using  $q$  to denote some specific integer with the property that  $n = 2q$ . It is only by doing this that we can make use of the existence statement. We can now proceed to find an integer  $p$  such that  $n^2 = 2p$  and complete the proof as follows.

Therefore  $n^2 = (2q)^2 = 4q^2 = 2(2q^2)$  and so, since  $2q^2$  is an integer,  $n^2$  is even.

Hence, if  $n$  is even, then  $n^2$  is even.

This is actually written out without reference to ‘ $p$ ’ although we could have said ‘ $n^2 = 2p$  where  $p = 2q^2$  is an integer and so  $n^2$  is even.’

There is a genuine ambiguity about the statement ‘ $n = 2q$  for some  $q \in \mathbb{Z}$ ’. It may be the existential statement that the integer  $q$  exists and this is what is meant by the statement  $\exists q \in \mathbb{Z}, n = 2q$ . Alternatively, it may be the statement that  $q$  is a specific integer such that  $n = 2q$  as occurs in this proof. The statement with the second meaning is possible because of the statement with the first meaning, and we could make this clear by writing out the proof as follows.

Suppose that  $n$  is even. Then  $\exists q \in \mathbb{Z}, n = 2q$ . So let  $q_1$  be an integer such that  $n = 2q_1$ . Then  $n^2 = (2q_1)^2 = 2(2q_1^2)$ . Hence, since  $2q_1^2$  is an integer,  $\exists p \in \mathbb{Z}, n^2 = 2p$ . Thus  $n^2$  is even.

Hence, if  $n$  is even then  $n^2$  is even.

In practice this distinction is blurred and I would encourage the reader not to worry about it – it is a case where ambiguity is better than

pedantry. Most problems are avoided so long as a different dummy variable is used each time a definition involving a quantifier is used in the proof.

#### 7.4 Disproving statements involving quantifiers

The idea of disproving statements can appear a little strange at first, but to some extent this is a matter of presentation: disproving ' $P$ ' is the same as proving 'not  $P$ '.

##### (a) *Disproving statements of the form $\forall a \in A, P(a)$*

We have already observed that the negation of this statement is the statement

$$\exists a \in A, \text{not } P(a)$$

and so we can disprove it by giving a single example for which it is false. This is called *disproof by counterexample* to  $P(a)$ .

**Example 7.4.1** To disprove the statement  $\forall x \in \mathbb{R}, x^2 > 2$ .

*Solution* A counterexample is provided by  $x = 1$  since  $1 \in \mathbb{R}$  and  $1^2 = 1 \leqslant 2$ .  $\square$

Great care is required in interpreting negatives of universal statements in everyday speech. For example consider the statement 'All the members of the class are not here' which would normally be taken to be the negative of the universal statement 'All the members of the class are here', in other words 'Some member of the class is not here.' This differs from our careful usage. Consider a mathematical statement of the same structure: 'All the numbers in the set are not even.' This must mean the same as 'All the numbers in the set are odd' since 'odd' means the same as 'not even'. But the everyday usage indicated above would give 'Some number in the set is odd.' Take care to use language forms which are not open to misunderstanding: if we wish to indicate an absence from a class then we should say 'Not all the members of the class are here.' But then, away from mathematics, one can have a lot of fun with ambiguity!

(b) *Disproving statements of the form  $\exists a \in A, P(a)$*

The negation of this statement, often written

$$\exists a \in A, P(a),$$

is the statement

$$\forall a \in A, \text{not } P(a)$$

and this gives one way of disproving the statement. We made use of this fact in the proof of Proposition 2.2.4. Here is another very simple and familiar example.

**Proposition 7.4.2** *There does not exist a real number  $x$  such that  $x^2 = -1$ .*

*Proof* We know that, for all  $x \in \mathbb{R}$ , we have the inequality  $x^2 \geq 0$  and so  $x^2 \neq -1$ . Hence there does not exist  $x \in \mathbb{R}$  such that  $x^2 = -1$ .  $\square$

The other way of disproving an existence statement is by contradiction. Here we show that the statement  $P(a)$  where  $a \in A$  necessarily leads to a contradiction. Proposition 4.1.1 which is a non-existence statement was proved using this method.

## 7.5 Proof by induction

We can reformulate the method of proof by induction using the language of set theory. Recall from Chapter 5 that induction is used to prove statements of the form  $\forall n \in \mathbb{Z}^+, P(n)$ . Such a statement can be rewritten as  $\{n \in \mathbb{Z}^+ \mid P(n)\} = \mathbb{Z}^+$ . Thus induction can be thought of as a method for proving that certain subsets of  $\mathbb{Z}^+$ , the set of positive integers, are in fact the whole set. From this point of view we can express the induction principle (Axiom 5.1.1) as follows.

**Axiom 7.5.1 (The induction principle reformulated)** *Suppose that  $A$  is a subset of  $\mathbb{Z}^+$ , the set of positive integers. Then  $A = \mathbb{Z}^+$  if*

- (i)  $1 \in A$ , and
- (ii)  $\forall k \in \mathbb{Z}^+ (k \in A \Rightarrow k + 1 \in A)$ .

This statement reduces to Axiom 5.1.1 if we put  $A = \{n \in \mathbb{Z}^+ \mid P(n)\}$ . It can be deduced from Axiom 5.1.1 if we write  $P(n)$  for the predicate  $n \in A$ .

Induction is quite often formulated in this more formal way.

### 7.6 Predicates involving more than one free variable

We have already met a number of universal and existential statements involving more than one variable. If  $P(a, b)$  is a predicate involving two free variables  $a \in A$  and  $b \in B$  then we can form propositions involving quantifiers as follows.

- (i)  $\forall a \in A, \forall b \in B, P(a, b).$
- (ii)  $\exists a \in A, \exists b \in B, P(a, b).$
- (iii)  $\forall a \in A, \exists b \in B, P(a, b).$
- (iv)  $\exists b \in B, \forall a \in A, P(a, b).$
- (v)  $\forall b \in B, \exists a \in A, P(a, b).$
- (vi)  $\exists a \in A, \forall b \in B, P(a, b).$

The meaning of the first two of these is fairly clear. Examples of these which we have already met are as follows.

**Proposition 3.1.1**  $\forall a, b \in \mathbb{R}^+, a < b \Rightarrow a^2 < b^2.$

**Proposition 3.2.1**  $\forall a, b \in \mathbb{R}, a < b \Rightarrow 4ab < (a + b)^2.$

**Proposition 4.1.1** *It is not true that  $\exists m, n \in \mathbb{Z}, 14m + 20n = 101.$*

Notice in these examples that ' $\forall a, b \in A$ ' is a shorthand for ' $\forall a \in A, \forall b \in A$ ' and similarly for ' $\exists$ '.

Statements involving both quantifiers require some care in particular regarding the order of the quantifiers. Consider for example the predicate ' $m < n$ ' involving positive integers  $m$  and  $n$ .

**Example 7.6.1**  $\forall m \in \mathbb{Z}^+, \exists n \in \mathbb{Z}^+, m < n.$

This is the statement that  $\{m \in \mathbb{Z}^+ \mid \exists n \in \mathbb{Z}^+, m < n\} = \mathbb{Z}^+$  or that  $m \in \mathbb{Z}^+ \Rightarrow (\exists n \in \mathbb{Z}^+, m < n)$ . Notice how the use of the single quantifier ' $\exists n \in \mathbb{Z}^+$ ' leads to a predicate ' $\exists n \in \mathbb{Z}^+, m < n$ ' with a single free variable  $m$ . We can then consider for which values of  $m$  this predicate is true. In this case we are considering the assertion that it holds for all positive integers  $m$ . This means that for each positive integer  $m$ , there exists a greater integer  $n$ . This is clearly the case and we can prove it by example: take  $n = m + 1$ . We can write out a formal proof as follows.

*Proof* This result is true because, given a positive integer  $m$ , if we put  $n = m + 1$  then  $n$  is a positive integer and  $m < n$ .  $\square$

**Example 7.6.2**  $\exists n \in \mathbb{Z}^+, \forall m \in \mathbb{Z}^+, m < n.$

This is the statement that the set  $\{n \in \mathbb{Z}^+ \mid \forall m \in \mathbb{Z}^+, m < n\}$  is non-empty. For a positive integer  $n$  to be in this set we must have  $\forall m \in \mathbb{Z}^+, m < n$ , in other words it must be greater than all positive integers. But it certainly isn't greater than  $n$  itself and so we can disprove the statement  $\forall m \in \mathbb{Z}^+, m < n$  by the counterexample  $m = n$ . Writing this out formally gives the following.

*Proof* This result is false because, for each positive integer  $n$ , if we put  $m = n$  then  $m$  is a positive integer and  $m \not< n$  so that  $m = n$  provides a counterexample to the statement  $\forall m \in \mathbb{Z}^+, m < n$  which is therefore false.  $\square$

Alternatively, the proof might be written more briefly as follows leaving the reader to sort out the quantifiers.

*Proof* This result is false because, for each positive integer  $n$ , if we put  $m = n$  then  $m$  is a positive integer and  $m \not< n$ .  $\square$

**Example 7.6.3**  $\forall n \in \mathbb{Z}^+, \exists m \in \mathbb{Z}^+, m < n.$

Let us consider the set  $\{n \in \mathbb{Z}^+ \mid \exists m \in \mathbb{Z}^+, m < n\}$ . This is the set of positive integers which are strictly greater than some other positive integer. If  $n > 1$  then  $n$  does lie in this set since we can take  $m = n - 1$  but on the other hand 1 is not an element of this set since  $\forall m \in \mathbb{Z}^+, m \geq 1$ . Hence  $\{n \in \mathbb{Z}^+ \mid \exists m \in \mathbb{Z}^+, m < n\} = \mathbb{Z}^+ - \{1\}$ . Since 1 does not lie in this set the universal statement is false. All we need say is the following.

*Proof* This statement is false and a counterexample is  $n = 1$  since  $m \not< 1$  for all positive integers  $m$ .  $\square$

**Example 7.6.4**  $\exists m \in \mathbb{Z}^+, \forall n \in \mathbb{Z}^+, m < n.$

This is the statement that the set  $\{m \in \mathbb{Z}^+ \mid \forall n \in \mathbb{Z}^+, m < n\}$  is non-empty. For a positive integer  $m$  to be in this set we must have  $\forall n \in \mathbb{Z}^+, m < n$ , in other words it must be smaller than all positive integers. But it certainly isn't smaller than itself and so we can disprove the statement  $\forall n \in \mathbb{Z}^+, m < n$  by the counterexample  $n = m$ . Writing this out formally gives the following.

*Proof* This result is false because, for each positive integer  $m$ , if we put  $n = m$  then  $n$  is a positive integer and  $m \not< n$  so that  $n = m$  provides a counterexample to the statement  $\forall n \in \mathbb{Z}^+, m < n$  which is therefore false.  $\square$

The alternative more usual briefer form is as follows.

*Proof* This result is false because, for each positive integer  $m$ , if we put  $n = m$  then  $n$  is a positive integer and  $m \not< n$ .  $\square$

The reader should carefully compare the proofs in Example 7.6.2 and Example 7.6.4. Although in both cases the implication used is that if  $m = n$  then  $m \not< n$ , in the first example we start from a general positive integer  $n$  and then define  $m$  by  $m = n$ , whereas in the second example we start from a general positive integer  $m$  and then define  $n$  by  $n = m$ . The distinction may be clarified when the reader does Exercise 7.2.

### 7.7 The Cartesian product of two sets

At the beginning of this chapter we introduced quantifiers in terms of properties of the subset defined by a predicate involving one free variable. Predicates involving more than one free variable also define subsets – of a set known as the Cartesian product. For simplicity we restrict attention to the case of two free variables.

**Definition 7.7.1** Given sets  $X$  and  $Y$ , the Cartesian product of  $X$  and  $Y$ , denoted by  $X \times Y$ , is the set of all ordered pairs  $(x, y)$  where  $x \in X$  and  $y \in Y$ . Thus

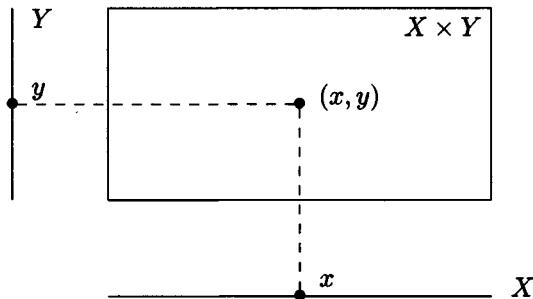
$$X \times Y = \{(x, y) \mid x \in X \text{ and } y \in Y\}.$$

In referring to an ordered pair in this definition we mean that two such pairs,  $(x_1, y_1)$  and  $(x_2, y_2)$ , are equal,  $(x_1, y_1) = (x_2, y_2)$ , if and only if  $x_1 = x_2$  and  $y_1 = y_2$ . We say that the ordered pair  $(x, y)$  has coordinates  $x$  and  $y$ .

When  $Y = X$  we write  $X \times X = X^2$ .

We can picture points in the Cartesian product as follows using lines to represent the sets  $X$  and  $Y$  and a rectangle to represent  $X \times Y$ . It is

customary to use a horizontal line to denote the first set and a vertical line to denote the second set.



**Examples 7.7.2** (a) For  $X = \{a, b, c\}$  and  $Y = \{a, b\}$ ,

$$X \times Y = \{(a, a), (a, b), (b, a), (b, b), (c, a), (c, b)\}$$

and

$$Y \times X = \{(a, a), (a, b), (a, c), (b, a), (b, b), (b, c)\}.$$

Notice that these two sets are different.

(b)  $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$  is the familiar 2-dimensional Euclidean plane.

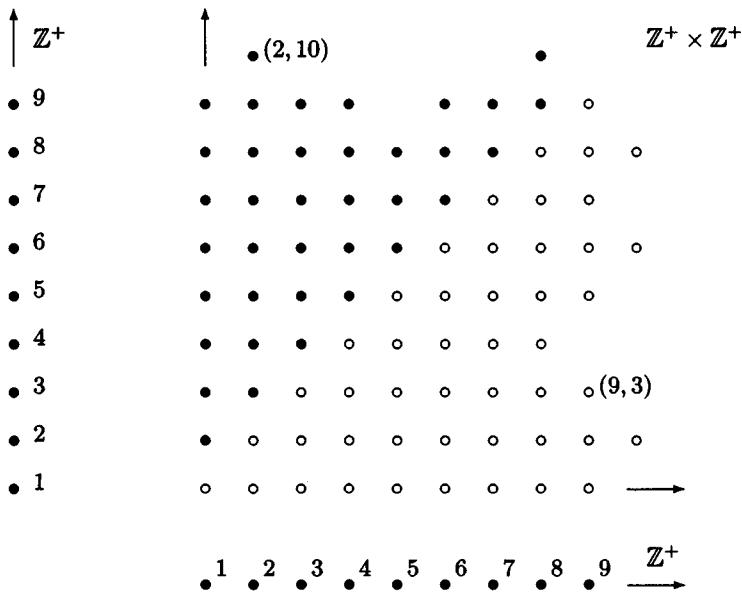
Now given a predicate  $P(a, b)$  involving free variables  $a \in A$  and  $b \in B$ , we can define the subset  $\{(a, b) \in A \times B \mid P(a, b)\}$  of the Cartesian product. This is familiar in  $\mathbb{R}^2$  from the study of plane curves using the methods of Cartesian geometry; we can describe certain subsets of  $\mathbb{R}^2$  by giving an *equation*. For example we say that the circle with centre  $(0, 0)$  and radius 1 has equation  $x^2 + y^2 = 1$  to mean that the set of points in  $\mathbb{R}^2$  which lie on this circle is given by  $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ .

**Example 7.7.3** To illustrate the relationship between universal and existential statements involving a predicate with two free variables and the subset defined by the predicate we look again at the examples considered in the previous section. We can draw part of the picture of the subset

$$\{(m, n) \in \mathbb{Z}^+ \times \mathbb{Z}^+ \mid m < n\} \subseteq \mathbb{Z}^+ \times \mathbb{Z}^+$$

as in the diagram on the next page.

Here the solid dots ( $\bullet$ ) indicate the elements of the Cartesian product for which the predicate  $m < n$  is true and the circles ( $\circ$ ) those for which



it is false. Thus  $(2, 10)$  is in the subset determined by the predicate whereas  $(9, 3)$  is not.

We now consider each example in turn.

**Example 7.6.1**  $\forall m \in \mathbb{Z}^+, \exists n \in \mathbb{Z}^+, m < n$ . This is true because each vertical line of elements in the Cartesian product contains an element in the subset, for example  $\{m\} \times \mathbb{Z}^+$ , the vertical line of points whose first coordinate is  $m$ , contains the point  $(m, m + 1)$ .

**Example 7.6.2**  $\exists n \in \mathbb{Z}^+, \forall m \in \mathbb{Z}^+, m < n$ . This is false because no horizontal line lies entirely in the subset, for example  $\mathbb{Z}^+ \times \{n\}$ , the horizontal line of points whose second coordinate is  $n$ , contains the point  $(n, n)$  which does not lie in the subset.

**Example 7.6.3**  $\forall n \in \mathbb{Z}^+, \exists m \in \mathbb{Z}^+, m < n$ . This is false because there is a horizontal line containing no elements of the subset, namely the line  $\mathbb{Z}^+ \times \{1\}$ .

**Example 7.6.4**  $\exists m \in \mathbb{Z}^+, \forall n \in \mathbb{Z}^+, m < n$ . This is false because each vertical line contains a point not in the subset, for example  $\{m\} \times \mathbb{Z}^+$  contains the point  $(m, m)$ .

To conclude this chapter here are some standard results relating the Cartesian product to union and intersection.

**Proposition 7.7.4** *For all sets  $A$ ,  $B$ ,  $C$  and  $D$  the following hold:*

- (i)  $A \times (B \cup C) = (A \times B) \cup (A \times C)$ ;
- (ii)  $A \times (B \cap C) = (A \times B) \cap (A \times C)$ ;
- (iii)  $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$ ;
- (iv)  $(A \times B) \cup (C \times D) \subseteq (A \cup C) \times (B \cup D)$ .

These statements may be proved directly from the definitions. As an illustration here is a proof of part (ii). (Part (iv) appears as Exercise 7.7 and part (iii) appears in Problems II, Question 13.)

*Proof of part (ii)* A proof that two sets are equal requires us to prove two set inclusions. In this case we can do them together as follows.

$$\begin{aligned} (x, y) \in A \times (B \cap C) &\Leftrightarrow x \in A \text{ and } y \in B \cap C \\ &\Leftrightarrow x \in A \text{ and } y \in B \text{ and } y \in C \\ &\Leftrightarrow (x, y) \in A \times B \text{ and } (x, y) \in A \times C \\ &\Leftrightarrow (x, y) \in (A \times B) \cap (A \times C). \end{aligned}$$

Thus  $(x, y) \in A \times (B \cap C) \Leftrightarrow (x, y) \in (A \times B) \cap (A \times C)$  as required.  $\square$

## Exercises

**7.1** Determine the following sets:

- (i)  $\{ m \in \mathbb{Z}^+ \mid \exists n \in \mathbb{Z}^+, m \leq n \}$ ,
- (ii)  $\{ m \in \mathbb{Z}^+ \mid \forall n \in \mathbb{Z}^+, m \leq n \}$ ,
- (iii)  $\{ n \in \mathbb{Z}^+ \mid \exists m \in \mathbb{Z}^+, m \leq n \}$ ,
- (iv)  $\{ n \in \mathbb{Z}^+ \mid \forall m \in \mathbb{Z}^+, m \leq n \}$ .

**7.2** Prove or disprove the following statements.

- (i)  $\forall m, n \in \mathbb{Z}^+, m \leq n$ .
- (ii)  $\exists m, n \in \mathbb{Z}^+, m \leq n$ .
- (iii)  $\forall m \in \mathbb{Z}^+, \exists n \in \mathbb{Z}^+, m \leq n$ .
- (iv)  $\exists m \in \mathbb{Z}^+, \forall n \in \mathbb{Z}^+, m \leq n$ .
- (v)  $\forall n \in \mathbb{Z}^+, \exists m \in \mathbb{Z}^+, m \leq n$ .
- (vi)  $\exists n \in \mathbb{Z}^+, \forall m \in \mathbb{Z}^+, m \leq n$ .

**7.3** Prove or disprove the following statements.

- (i)  $\forall m, n \in \mathbb{Z}, m \leq n.$
- (ii)  $\exists m, n \in \mathbb{Z}, m \leq n.$
- (iii)  $\forall m \in \mathbb{Z}, \exists n \in \mathbb{Z}, m \leq n.$
- (iv)  $\exists m \in \mathbb{Z}, \forall n \in \mathbb{Z}, m \leq n.$
- (v)  $\forall n \in \mathbb{Z}, \exists m \in \mathbb{Z}, m \leq n.$
- (vi)  $\exists n \in \mathbb{Z}, \forall m \in \mathbb{Z}, m \leq n.$

**7.4** Prove or disprove each of the following statements.

- (i)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, x + y = 0.$
- (ii)  $\exists y \in \mathbb{R}, \forall x \in \mathbb{R}, x + y = 0.$
- (iii)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, xy = 0.$
- (iv)  $\exists y \in \mathbb{R}, \forall x \in \mathbb{R}, xy = 0.$
- (v)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, xy = 1.$
- (vi)  $\exists y \in \mathbb{R}, \forall x \in \mathbb{R}, xy = 1.$
- (vii)  $\forall n \in \mathbb{Z}^+, (n \text{ is even or } n \text{ is odd}).$
- (viii)  $(\forall n \in \mathbb{Z}^+, n \text{ is even}) \text{ or } (\forall n \in \mathbb{Z}^+, n \text{ is odd}).$

**7.5** Prove the following:

$$(\exists q \in \mathbb{Z}, n = 2q + 1) \Rightarrow (\exists p \in \mathbb{Z}, n^2 = 2p + 1).$$

**7.6** Write the following universal statement in terms of quantifiers and prove it.

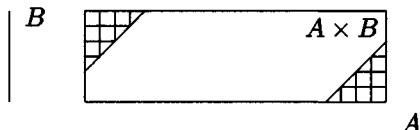
For integers  $a$  and  $b$ , if  $a$  and  $b$  are even then so is  $a + b$ .

**7.7** For sets  $A, B, C$  and  $D$  prove that

$$(A \times B) \cup (C \times D) \subseteq (A \cup C) \times (B \cup D),$$

and give an example to show that these sets are not always equal.

**7.8** Suppose that the set  $\{(a, b) \mid P(a, b)\}$  is given by the triangular regions in the diagram below.



Decide whether each of the following statements is true or false.

- (i)  $\forall a \in A, \exists b \in B, P(a, b).$
- (ii)  $\exists b \in B, \forall a \in A, P(a, b).$
- (iii)  $\forall b \in B, \exists a \in A, P(a, b).$
- (iv)  $\exists a \in A, \forall b \in B, P(a, b).$

**7.9** Find a reformulation of Axiom 5.4.1 (the strong induction principle) as a method of proving that subsets of  $\mathbb{Z}^+$  are the whole set (similar to Axiom 7.5.1).

# 8

## Functions

The notion of *function* is one of the most fundamental in mathematics. It is probably familiar to the reader in the context of calculus, and it is here that the concept was first clarified and the difference between a function and a formula properly understood in the early years of the nineteenth century. Today, the language of functions is used throughout mathematics.

In this chapter the function concept is introduced. We discuss various ways of defining a function and consider the graph of a function.

### 8.1 Functions and formulae

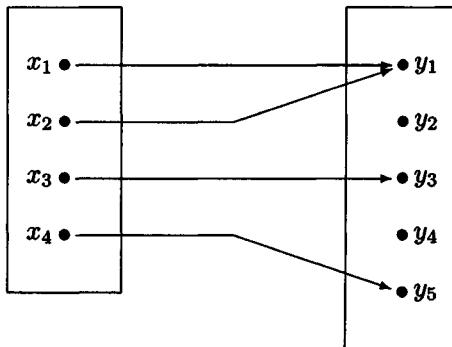
**Definition 8.1.1** Suppose that  $X$  and  $Y$  are sets. A function, map or mapping from the set  $X$  to the set  $Y$  is the assignment of a unique element of  $Y$  to each element of  $X$ . If  $f$  is a function from  $X$  to  $Y$  we write  $f: X \rightarrow Y$  and denote the element of  $Y$  assigned to an element  $x \in X$  by  $f(x)$  writing  $x \mapsto f(x)$ . The element  $f(x) \in Y$  is called the value of  $f$  at  $x \in X$  or the image of  $x$  under  $f$ . The set  $X$  is called the domain of the function  $f$  and the set  $Y$  is called the codomain.

**Example 8.1.2** The most basic way of describing a function is by listing the values. For example, given  $X = \{x_1, x_2, x_3, x_4\}$  and  $Y = \{y_1, y_2, y_3, y_4, y_5\}$ , the following table determines a function  $f: X \rightarrow Y$ .

|        |       |       |       |       |
|--------|-------|-------|-------|-------|
| $x$    | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
| $f(x)$ | $y_1$ | $y_1$ | $y_3$ | $y_5$ |

We determine the value of the function at an element of  $X$  by locating that element in the first row and reading down to the second row. Notice that each element of the domain  $X$  occurs precisely once in the first row so that this is a well-defined procedure. On the other hand elements of the codomain  $Y$  may occur once in the second row, but they may also occur more than once or not at all: an element of the codomain may be the value of the function at several elements of the domain or may not be a value at all.

We can picture this function as follows.



Here we find the value of  $f$  at an element of the domain  $X$  by following the arrow which begins at that element. Each element of the domain lies at the beginning of just one arrow so this is a well-defined procedure. However, elements in the codomain may lie at the end of one arrow, several arrows or no arrow at all.

Yet another way of describing a function is as follows. Think of the elements of the codomain set as boxes. The function describes how to place the objects of the domain set in these boxes. Thus the function under consideration would be pictured as follows.

| $y_1$           | $y_2$ | $y_3$ | $y_4$ | $y_5$ |
|-----------------|-------|-------|-------|-------|
| $x_1 \quad x_2$ |       | $x_3$ |       | $x_4$ |

Again notice that each element of the domain set occurs precisely once

in the picture. However, any given box, corresponding to an element of the codomain, may contain one object, several objects or none at all.

The reader may wonder at this variety of ways of thinking about a function. It is important to realize that mathematicians develop a variety of ways of thinking about the rather abstract concepts they deal with. Any given way may suit one person better than another and each leads to different insights: it can be useful to be aware of several.

**Example 8.1.3** Suppose that  $X = \{a, b, c\}$  and  $Y = \{d, e\}$ . Then there are precisely eight functions from the set  $X$  to the set  $Y$  given as follows.

| $x$ | $f_1(x)$ | $f_2(x)$ | $f_3(x)$ | $f_4(x)$ | $f_5(x)$ | $f_6(x)$ | $f_7(x)$ | $f_8(x)$ |
|-----|----------|----------|----------|----------|----------|----------|----------|----------|
| $a$ | $d$      | $d$      | $d$      | $d$      | $e$      | $e$      | $e$      | $e$      |
| $b$ | $d$      | $d$      | $e$      | $e$      | $d$      | $d$      | $e$      | $e$      |
| $c$ | $d$      | $e$      | $d$      | $e$      | $d$      | $e$      | $d$      | $e$      |

Here the functions  $f_1$  and  $f_8$  are examples of *constant functions*. Given sets  $X$  and  $Y$  and any element  $y_0 \in Y$  there is a *constant function*  $c_{y_0}: X \rightarrow Y$  given by  $c_{y_0}(x) = y_0$  for all  $x \in X$ .

Of course a function can have the same set as domain and codomain as in the next example.

**Example 8.1.4** Suppose that  $Z = \{a, b\}$ . Then there are precisely four functions from  $Z$  to  $Z$  as follows.

| $x$ | $g_1(x)$ | $g_2(x)$ | $g_3(x)$ | $g_4(x)$ |
|-----|----------|----------|----------|----------|
| $a$ | $a$      | $a$      | $b$      | $b$      |
| $b$ | $a$      | $b$      | $a$      | $b$      |

In this case the functions  $g_1$  and  $g_4$  are constant functions. The function  $g_2$  is an example of an *identity function*. Given a set  $X$ , the *identity function* on  $X$ ,  $I_X: X \rightarrow X$ , is given by  $I_X(x) = x$  for all  $x \in X$ .

Describing a function by listing the values is only practical when the domain set is small and impossible if the domain set is infinite. The most common way of describing a function  $f$  is by giving a *formula* for  $f$  which provides a procedure for finding the value of the function at each element of the domain. When defining a function using a formula it is important to be clear about which sets are the domain and the codomain of the function.

**Example 8.1.5** Recall that  $\mathbb{R}^{\geq 0}$  denotes the subset of  $\mathbb{R}$ , the set of real numbers, given by  $\{x \in \mathbb{R} \mid x \geq 0\}$ . Consider the following four functions:

- (i)  $f_1: \mathbb{R} \rightarrow \mathbb{R}$  given by  $f_1(x) = x^2$ ;
- (ii)  $f_2: \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}$  given by  $f_2(x) = x^2$ ;
- (iii)  $f_3: \mathbb{R} \rightarrow \mathbb{R}^{\geq 0}$  given by  $f_3(x) = x^2$ ;
- (iv)  $f_4: \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$  given by  $f_4(x) = x^2$ .

Notice that, although the formula  $f_i(x) = x^2$  is the same for each of these four functions, they are considered to be four distinct functions since the domain and codomain are part of the definition of the function. We will see later that these four functions have different properties even though they are given by the same formula.

We need to take care that the formula makes sense for each element of the domain.

**Example 8.1.6** The formula  $f_1(x) = \frac{x^2 + x - 2}{x - 1}$  does not define a function  $\mathbb{R} \rightarrow \mathbb{R}$  since it gives no value for  $x = 1$ . When working with the set of real numbers it is quite common not to specify the domain and codomain of a function given by a formula. The convention is that, if these are not specified, then we take as the domain the subset of  $\mathbb{R}$  consisting of the numbers for which the formula makes sense and we take  $\mathbb{R}$  as the codomain. Using this convention the above formula for  $f_1$  defines a function  $f_1: \mathbb{R} - \{1\} \rightarrow \mathbb{R}$ .

If we really want a function with domain  $\mathbb{R}$  then in an example like this there are two basic techniques for extending  $f_1$ .

*Rewriting the formula:* We can rewrite the formula in such a way that it makes sense for all real numbers  $x$  using

$$(x^2 + x - 2)/(x - 1) = (x - 1)(x + 2)/(x - 1) = x + 2$$

for  $x \neq 1$ . Then  $f_2(x) = x + 2$  does define a function on  $\mathbb{R}$ , i.e. with domain  $\mathbb{R}$ , extending the function  $f_1$ .

*Explicit definition:* Alternatively we can explicitly specify the value of the function at the element 1 where the formula for  $f_1$  does not work. Thus

$$f_3(x) = \begin{cases} \frac{x^2 + x - 2}{x - 1} & \text{if } x \neq 1, \\ 3 & \text{if } x = 1, \end{cases}$$

defines a function  $f_3: \mathbb{R} \rightarrow \mathbb{R}$ .

Notice that we may specify the values at individual points any way we like. It doesn't have to relate in a sensible way to the values elsewhere. In defining  $f_3(1)$  we selected that value given by the formula  $x + 2$  so that  $f_2$  and  $f_3$  are in fact the same function. However, we could extend the function  $f_1$  to the whole of  $\mathbb{R}$  in a different way. For example

$$f_4(x) = \begin{cases} \frac{x^2 + x - 2}{x - 1} & \text{if } x \neq 1, \\ 42 & \text{if } x = 1, \end{cases}$$

is another way of extending the function  $f_1$  to the whole of  $\mathbb{R}$ .

**Example 8.1.7** In some cases, for example  $x \mapsto 1/x$ , there appears to be no way of rewriting the formula to include in the domain of definition points where it gives no value (in this case  $x = 0$ ). But in these cases we can still use the second method. For example

$$g(x) = \begin{cases} 1/x & \text{if } x \neq 0, \\ 73 & \text{if } x = 0, \end{cases}$$

is a perfectly respectable function  $g: \mathbb{R} \rightarrow \mathbb{R}$ .

We can use different formulae for different parts of the domain as in the following example.

**Example 8.1.8** The *modulus function*  $x \mapsto |x|$  is defined on  $\mathbb{R}$  by

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

Notice here that the value of the function at 0 is given twice, since  $0 \geq 0$  and  $0 \leq 0$ . This isn't a problem since the two formulae ( $x \mapsto x$  and  $x \mapsto -x$ ) give the same value (namely 0) for  $x = 0$ . But when specifying a function in this way we do need to be careful to check that it is *well-defined* (i.e. there is a uniquely specified value) at each point.

**Definition 8.1.9** Two functions  $f: X \rightarrow Y$  and  $g: X \rightarrow Y$  are equal, written  $f = g$ , when they have the same value at each point of the domain  $X$ , i.e.  $f(x) = g(x)$  for all  $x \in X$ . Notice that it is implicit in this definition that two equal functions have the same domain and the same codomain.

**Example 8.1.10** The functions  $f_2$  and  $f_3$  of Example 8.1.6 are equal even though they are defined in different ways. Thus it is not the process leading to the values which determines the function but what the values are. Of course, in defining  $f_3$  the value of  $f_3(1)$  was selected precisely so that they would be equal.

We have seen in Example 8.1.6 how to extend the domain of a function. We can also make the domain smaller.

**Definition 8.1.11** Suppose that  $f: X \rightarrow Y$  is a function and  $A$  is a subset of  $X$ , i.e.  $A \subseteq X$ . Then we can define a function  $g: A \rightarrow Y$  by  $g(a) = f(a)$  for all  $a \in A$ . This function is called the restriction of  $f$  to  $A$  and is denoted by  $f|A$ .

**Examples 8.1.12** (a) In Example 8.1.3  $f_1|\{a, b\} = f_2|\{a, b\}$  and is the constant function taking the value  $d$ .

(b) In Example 8.1.5  $f_2 = f_1|\mathbb{R}^{\geq}$  and  $f_4 = f_3|\mathbb{R}^{\geq}$ .

(c) In Example 8.1.6  $f_2|(\mathbb{R} - \{1\}) = f_4|(\mathbb{R} - \{1\}) = f_1$ .

## 8.2 Composition of functions

Suppose that  $f: X \rightarrow Y$  and  $g: Y \rightarrow Z$ . Then, given an element  $x \in X$ , the function  $f$  assigns to it an element  $y = f(x) \in Y$  and now the function  $g$  assigns to this an element  $g(y) = g(f(x)) \in Z$ . Thus using  $f$  and  $g$  an element of  $Z$  has been assigned to  $x$ . This process defines a function with domain  $X$  and codomain  $Z$  called the composite of  $f$  and  $g$ .

**Definition 8.2.1** Given two functions  $f: X \rightarrow Y$  and  $g: Y \rightarrow Z$  the composite of  $f$  and  $g$ , denoted by  $g \circ f: X \rightarrow Z$  or simply  $gf: X \rightarrow Z$ , is defined by

$$g \circ f(x) = g(f(x)) \quad \text{for all } x \in X.$$

The order of the  $f$  and  $g$  in this definition should be carefully noted. Since we have adopted the convention (usual† in most areas of mathematics) of writing the symbol for the function on the left of the element

† But not used universally. In some branches of algebra it is quite common to write the value  $f(x)$  as  $xf$  or even  $x^f$ .

where it is being evaluated (viz.  $f(x)$ ) it is natural to denote the function obtained by applying  $f$  and then applying  $g$  by the symbol  $g \circ f$  which presents them (from left to right) in the opposite order to their application. We can write the following to indicate this:

$$g \circ f: X \xrightarrow{f} Y \xrightarrow{g} Z.$$

**Examples 8.2.2** (a) The function  $x \mapsto (x + 1)^2$  from  $\mathbb{R}$  to  $\mathbb{R}$  is the composite of the function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(x) = x + 1$  and the function  $g: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $g(x) = x^2$  because

$$g \circ f(x) = g(f(x)) = (x + 1)^2 \quad \text{for all } x \in X.$$

Notice that the order does matter here because

$$f \circ g(x) = f(g(x)) = x^2 + 1;$$

composition of functions is not in general *commutative*. It is not usually a good idea to refer simply to ‘the composite of  $f$  and  $g$ ’ when both orders are possible since there is no well-established convention about which order these words indicate. It is best to use the symbolism  $g \circ f$  or  $f \circ g$  in order to be quite clear.

(b) Suppose that  $A$  is a subset of a set  $X$ . Then we may define a function  $i: A \rightarrow X$  by  $i(a) = a$  for all  $a \in A$ . This function is called the *inclusion function* of  $A$  into  $X$ . If now  $f: X \rightarrow Y$  is a function, then the composite  $f \circ i: A \rightarrow Y$  is equal to the restriction  $f|_A: A \rightarrow Y$  of  $f$  to  $A$ .

(c) Suppose that  $f: \mathbb{R} \rightarrow \mathbb{R}$  is defined by  $f(x) = x^2 + 1$  and  $g: \mathbb{R} - \{0\} \rightarrow \mathbb{R}$  is defined by  $g(x) = 1/x$ . Then strictly speaking the composite  $g \circ f$  is not defined since the codomain of  $f$  is not the same as the domain of  $g$ .

However, every value of  $f$  lies in the domain of  $g$  since  $x^2 + 1 \geq 1 > 0$  for all real numbers  $x$  so that  $g$  can be evaluated at each value of  $f$ . In a case like this we can still attach a meaning to the composite  $g \circ f$  as the function given by  $g \circ f(x) = g(f(x))$ . This gives the function  $\mathbb{R} \rightarrow \mathbb{R}$  determined by  $g \circ f(x) = 1/(x^2 + 1)$ .

**Proposition 8.2.3** Suppose that  $f: X \rightarrow Y$ ,  $g: Y \rightarrow Z$  and  $h: Z \rightarrow W$  are functions. Then

- (i)  $(h \circ g) \circ f = h \circ (g \circ f): X \rightarrow Z$ ,
- (ii)  $f \circ I_X = f = I_Y \circ f: X \rightarrow Y$ .

*Proof* These results are proved simply by evaluating the functions. For example, given  $x \in X$ , both the functions in (i) assign  $h(g(f(x)))$  to this element and so the functions are equal.  $\square$

Notice that the first part of Proposition 8.2.3 means that we can write  $h \circ g \circ f$  without ambiguity. Composition of functions is *associative*.

### 8.3 Sequences

**Definition 8.3.1** A function  $f : \mathbb{Z}^+ \rightarrow A$  is called a sequence in the set  $A$ .

The study of sequences in  $\mathbb{R}$  or  $\mathbb{C}$  is of great importance in advanced calculus and numerical analysis. Sequences may be defined by a formula, such as  $n \mapsto n^2$ ,  $n \mapsto 1/n$ ,  $n \mapsto 2^n$ ,  $n \mapsto (1 + 1/n)^n$  and  $n \mapsto 1/n!$ . But more interesting are those sequences defined inductively. We saw in Chapter 5 that strictly speaking formulae involving the exponent  $n$  or the factorial  $n!$  really have to be defined inductively. Another example of an inductively defined sequence considered there was the Fibonacci sequence.

It is common in advanced calculus to solve numerical problems by generating a sequence of numbers  $x_n$  which provide increasingly good approximations to the solution as  $n$  increases – the sequence converges to a limit which is the required solution.<sup>†</sup>

One important reason for becoming adept at handling quantifiers is to be able to handle the definition of the limit of a sequence. To illustrate the ideas we introduce the idea of a null sequence, a sequence with limit 0.

**Definition 8.3.2** Given a sequence  $f : \mathbb{Z}^+ \rightarrow \mathbb{R}$  of real numbers, we say that the sequence is null, written  $\lim f = 0$  or  $\lim_{n \rightarrow \infty} f(n) = 0$ , when

$$\forall \varepsilon \in \mathbb{R}^+, \exists N \in \mathbb{Z}^+, \forall n \in \mathbb{Z}^+ (n \geq N \Rightarrow |f(n)| < \varepsilon).$$

<sup>†</sup> A good example is the Newton–Raphson method for solving an equation, described in most books on advanced calculus (see for example G.B. Thomas and R.S. Finney, *Calculus and analytic geometry*, Addison-Wesley, Eighth edition 1992).

The use of  $\varepsilon$  for a general positive real number in this definition is traditional. This is not the place to investigate the full implications of this definition which involves *three* quantifiers! Here is a simple example of its use.

**Example 8.3.3** The sequence  $n \mapsto 1/\sqrt{n}$  is null.

**Constructing a proof.** We are required to prove that

$$\forall \varepsilon \in \mathbb{R}^+, \exists N \in \mathbb{Z}^+, \forall n \in \mathbb{Z}^+ (n \geq N \Rightarrow 1/\sqrt{n} < \varepsilon).$$

[Notice that  $|1/\sqrt{n}| = 1/\sqrt{n}$ .] So suppose we are given  $\varepsilon \in \mathbb{R}^+$ . Then to demonstrate the existence of the integer  $N$  we examine when  $1/\sqrt{n} < \varepsilon$ .

$$1/\sqrt{n} < \varepsilon \Leftrightarrow 1/n < \varepsilon^2 \Leftrightarrow n > 1/\varepsilon^2.$$

Hence so long as we choose a positive integer  $N$  such that  $N > 1/\varepsilon^2$  we will have  $n \geq N \Rightarrow n > 1/\varepsilon^2 \Rightarrow 1/\sqrt{n} < \varepsilon$  as required.

So for example, if  $\varepsilon = 1$  then we must choose  $N > 1$ , if  $\varepsilon = 1/2$  then we must choose  $N > 4$ , and if  $\varepsilon = 1/100$  then  $N > 10000$ . The smaller the number  $\varepsilon$  the greater the number  $N$  is required to be. But whatever the positive real number  $\varepsilon$  is the condition  $N > 1/\varepsilon^2$  tells us how large  $N$  has to be.

*Proof* Given a positive real number  $\varepsilon$ ,  $1/\sqrt{n} < \varepsilon$  if and only if  $n > 1/\varepsilon^2$ . Hence if we choose  $N > 1/\varepsilon^2$  then  $n \geq N$  implies that  $n > 1/\varepsilon^2$  so that  $1/\sqrt{n} < \varepsilon$  is required.  $\square$

This type of proof involving multiple quantifiers is quite elaborate and this is in fact an extremely simple example. More complicated examples are not considered in this book but are dealt with in detail in books on infinite sequences and series, advanced calculus and analysis.<sup>†</sup>

#### 8.4 The image of a function

Given a function  $f: X \rightarrow Y$ , it is not necessary that every element of  $Y$  is a value of the function. For example the function  $\mathbb{R} \rightarrow \mathbb{R}$  given by  $x \mapsto x^2$  does not have  $-1$  as a value. Thus we can obtain a subset of  $Y$  by considering those elements which are values.

<sup>†</sup> See, for example, R. Haggerty, *Fundamentals of mathematical analysis*, Addison-Wesley, Second edition 1993.

**Definition 8.4.1** Given a function  $f: X \rightarrow Y$ , the subset of the codomain  $Y$  consisting of those elements which are values of  $f$  is called the image of  $f$  and is denoted by Im  $f$ . Thus

$$\text{Im } f = \{ f(x) \mid x \in X \}.$$

This means that the function  $f$  provides a constructive definition of the set  $\text{Im } f$ , as discussed in Chapter 6.

### 8.5 The graph of a function

**Definition 8.5.1** Suppose that  $f: X \rightarrow Y$  is a function. Then we define the graph of  $f$  to be the subset of the Cartesian product  $X \times Y$  given by

$$G_f = \{ (x, y) \in X \times Y \mid y = f(x) \} = \{ (x, f(x)) \mid x \in X \}.$$

**Remarks 8.5.2** In the case that  $X$  and  $Y$  are subsets of  $\mathbb{R}$  this is the usual idea of the graph of a function. Work in calculus makes it clear that the graph gives a great deal of information about the function and that it is useful to develop skill in drawing graphs.

**Example 8.5.3** The graphs of the functions  $f_1$  and  $f_2$  of Example 8.1.3 are as follows.



Here we indicate the elements of the graph by solid dots ( $\bullet$ ).

The graph of the function  $f: X \rightarrow Y$  is a particular example of a subset of the Cartesian product  $X \times Y$  determined by a predicate: the predicate is  $y = f(x)$ . Not every subset of  $X \times Y$  arises as the graph of a function. Each column  $\{x_0\} \times Y$  must contain a single element, namely

$(x_0, f(x_0))$ . However, the function  $f$  is determined by its graph  $G_f$ , for, given  $x_0 \in X$ , there is a unique element  $(x, y) = (x_0, y_0) \in G_f$  such that  $x = x_0$ , and then  $f(x_0) = y_0$ .†

### Exercises

**8.1** Define functions  $f$  and  $g: \mathbb{R}^2 \rightarrow \mathbb{R}$  by:

$$\begin{aligned}f(x, y) &= \frac{x+y}{2} + \frac{|x-y|}{2}; \\g(x, y) &= \begin{cases} x & \text{if } x \geq y, \\ y & \text{if } x \leq y. \end{cases}\end{aligned}$$

Prove that the function  $g$  is well-defined. Prove that  $f = g$ .

**8.2** Define functions  $f$  and  $g: \mathbb{R} \rightarrow \mathbb{R}$  by  $f(x) = x^3$  and  $g(x) = 1 - x$ . Find the functions (i)  $f \circ f$ , (ii)  $f \circ g$ , (iii)  $g \circ f$ , (iv)  $g \circ g$ .

List the elements of the set  $\{x \in \mathbb{R} \mid fg(x) = gf(x)\}$ .

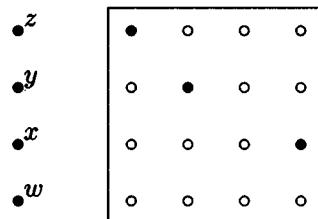
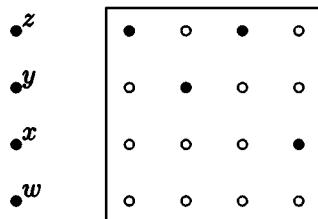
**8.3** Find functions  $f_i: \mathbb{R} \rightarrow \mathbb{R}$  with images as follows:

- (i)  $\text{Im } f_1 = \mathbb{R}$ ;
- (ii)  $\text{Im } f_2 = \mathbb{R}^+$ ;
- (iii)  $\text{Im } f_3 = \mathbb{R} - \mathbb{Z}$ ;
- (iv)  $\text{Im } f_4 = \mathbb{Z}$ .

**8.4** Prove that the sequence  $n \mapsto 1/n$  is null.

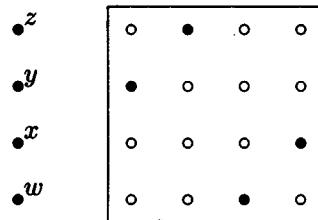
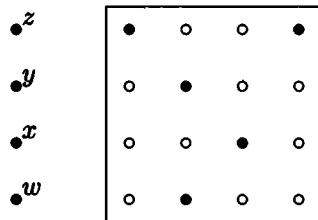
**8.5** Let  $X = \{a, b, c, d\}$  and  $Y = \{w, x, y, z\}$ . Which of the following subsets of  $X \times Y$  is the graph of a function  $f: X \rightarrow Y$ ? For those which are write down a table for the corresponding function as in Example 8.1.3. Explain why the others are not graphs of functions.

† Our definition of a function (Definition 8.1.1) is usually considered rather informal in more advanced mathematics. There a function is defined to be a graph: in other words a function  $X \rightarrow Y$  is defined to be a subset of  $X \times Y$  in which each element of  $X$  occurs as the first coordinate of an element precisely once (see for example Daniel J. Velleman, *How to prove it, a structured approach*, Cambridge University Press, 1994).



(i)       $\bullet^a$      $\bullet^b$      $\bullet^c$      $\bullet^d$

(ii)       $\bullet^a$      $\bullet^b$      $\bullet^c$      $\bullet^d$



(iii)       $\bullet^a$      $\bullet^b$      $\bullet^c$      $\bullet^d$

(iv)       $\bullet^a$      $\bullet^b$      $\bullet^c$      $\bullet^d$

# 9

## Injections, surjections and bijections

In defining a function  $f: X \rightarrow Y$  we insist that a unique element of  $Y$  is assigned to each element of  $X$ . However, we do not require that each element of  $Y$  is assigned to some element of  $X$  nor do we prevent the possibility of the same element of  $Y$  being assigned to several (or even all the) elements of  $X$ . By imposing additional conditions concerning the number of elements of  $X$  to which elements of  $Y$  are assigned we get functions with particular properties. In this chapter we consider functions with particularly good properties and in particular functions which are *bijections* for which we can define an inverse function.

### 9.1 Properties of functions

**Definition 9.1.1** Suppose that  $f: X \rightarrow Y$  is a function.

(i) If no element of  $Y$  is assigned to more than one element of  $X$ , i.e. the function takes a different value for each point of the domain, then we say that the function  $f$  is an **injection** (or that it is injective or one-to-one). In symbols we can write this

$$\forall x_1, x_2 \in X, (x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2))$$

or equivalently using the contrapositive

$$\forall x_1, x_2 \in X, (f(x_1) = f(x_2) \Rightarrow x_1 = x_2).$$

(ii) If each element of  $Y$  is assigned to some element of  $X$ , i.e. each point of the codomain is a value of the function, then we say that the

function  $f$  is a surjection (or that it is surjective or onto). In symbols we can write this

$$\forall y \in Y, \exists x \in X, y = f(x).$$

(iii) If  $f$  is both an injection and a surjection then we say that it is a bijection (or that it is bijective or one-to-one and onto).

We can reformulate these definitions using the notion of a *pre-image*.

**Definition 9.1.2** For a function  $f: X \rightarrow Y$ , given an element  $y \in Y$  a pre-image of  $y$  (under  $f$ ) is an element  $x \in X$  such that  $y = f(x)$ .

So  $f$  is an injection if and only if every element of  $Y$  has at most one pre-image;  $f$  is a surjection if and only if every element of  $Y$  has at least one pre-image; and  $f$  is a bijection if and only if every element of  $Y$  has precisely one pre-image.

These ideas are simply expressed in terms of the various models for a function introduced in the previous chapter. If a function is described by listing the values of the function as in Example 8.1.2 then we find the pre-images of an element of the codomain by locating its occurrences in the list of values and reading up to the corresponding elements of the first row. Thus a function is injective when no element occurs more than once in the list of values, it is surjective when every element of the codomain does occur in the list of values and bijective when every element of the codomain occurs precisely once in the list of values.

In terms of the picture representing the function by arrows from the domain to the codomain (see page 90) we find pre-images by following arrows backwards. Thus, for example, a function is bijective if every point of the codomain is at the end of precisely one arrow.

Finally, if we think of a function as a procedure for placing the elements of the set  $X$  into a set of boxes  $Y$  (see page 90) then a pre-image of a particular box  $y \in Y$  is simply an element which is placed in that box. From this point of view,  $f$  is injective when no two elements are placed in the same box, it is surjective when no box is left empty, and it is bijective when each box contains precisely one element.

**Examples 9.1.3** (a) In Example 8.1.3 all the functions apart from the constant functions  $f_1$  and  $f_8$  are surjections. None of the functions are

injections.

(b) In Example 8.1.4 the non-constant functions  $f_2$  and  $f_3$  are bijections. Notice that, for any set  $X$ , the identity function  $I_X$  is a bijection.

(c) In Example 8.1.5 the function  $f_1$  is neither an injection (since  $(-1)^2 = 1^2 = 1$ ) nor a surjection (since there is no real number  $x$  such that  $x^2 = -1$ ); the function  $f_2$  is an injection (since  $0 \leq x_1 < x_2 \Rightarrow x_1^2 < x_2^2$ ) but not a surjection; the function  $f_3$  is a surjection but not an injection; and the function  $f_4$  is a bijection. This example illustrates the importance of the domain and the codomain. For example, if a result has been proved for injections then it can be applied to  $f_2$  and to  $f_4$  but not to  $f_1$  and  $f_3$ .

**Remarks 9.1.4** Notice that we can easily convert any function into a surjection by changing the codomain. Recall that, given a function  $f: X \rightarrow Y$ , the *image* of  $f$ , denoted by  $\text{Im } f$ , is the set of values of  $f$ . The assignment determining the function  $f$  also determines a function  $f^+: X \rightarrow \text{Im } f$  (i.e.  $f^+(x) = f(x)$ ) which is a surjection.

Since the definitions of injectivity and surjectivity involve universal and existential statements, proofs and disproofs of these properties usually follow the format discussed in Chapter 7.

**Example 9.1.5** To determine whether the function  $f_1: \mathbb{R} \rightarrow \mathbb{R}$  given by  $f_1(x) = x + 1$  is an injection, a surjection or a bijection.

For injectivity, it is usually easiest to use the second formulation:

$$f_1(x_1) = f_1(x_2) \Rightarrow x_1 = x_2.$$

Filling in the definition of the function this gives

$$x_1 + 1 = x_2 + 1 \Rightarrow x_1 = x_2$$

which is certainly true so that  $f_1$  is injective.

For surjectivity, we rewrite the definition as usual as a universal implication:

$$y \in \mathbb{R} \Rightarrow \exists x \in \mathbb{R}, y = f_1(x).$$

If we put the definition of the function into this then it is easy to prove the existential statement by example. For

$$y = f_1(x) \Leftrightarrow y = x + 1 \Leftrightarrow x = y - 1$$

which tells us that, given  $y \in \mathbb{R}$ , then  $y = f_1(x)$  if  $x = y - 1$  so that  $y$  does have a pre-image under  $f_1$ . Hence  $f_1$  is surjective and so a bijection.

In fact, what is written above is rather inefficient; we can often prove bijectivity in one fell swoop. The best way to discover whether a function is injective or surjective can be to investigate the pre-images of a general element in the codomain using the approach used above for surjectivity. In this case if  $y$  is a general element of the codomain  $\mathbb{R}$  then

$$\begin{aligned} x \text{ is a pre-image of } y &\Leftrightarrow y = f_1(x) \\ &\Leftrightarrow y = x + 1 \\ &\Leftrightarrow x = y - 1. \end{aligned}$$

This shows that each element  $y$  of  $\mathbb{R}$  has precisely one pre-image, namely  $y - 1$ , so that the function is a bijection.

We can write out the solution to this problem quite briefly as follows.

*Solution* Since  $y = f_1(x) \Leftrightarrow y = x + 1 \Leftrightarrow x = y - 1$ , for  $x, y \in \mathbb{R}$ , we see that each element  $y$  of  $\mathbb{R}$  has precisely one pre-image under  $f_1$ . Hence  $f_1$  is a bijection.  $\square$

**Example 9.1.6** To determine whether the function  $f_2: \mathbb{R}^+ \rightarrow \mathbb{R}^+$  given by  $f_2(x) = x + 1$  is injective, surjective or bijective.

This function is given by the same formula as the previous example. If we try the pre-image approach we get almost the same argument:

$$\begin{aligned} x \text{ is a pre-image of } y &\Leftrightarrow y = f_2(x) \\ &\Leftrightarrow y = x + 1 \\ &\Leftrightarrow x = y - 1. \end{aligned}$$

The difference becomes clear when we ask whether the formula for the pre-image makes sense. Remember that if  $x$  is to be a pre-image then it must certainly lie in the domain  $\mathbb{R}^+$  (otherwise  $f_2(x)$  isn't even defined). But, given  $y \in \mathbb{R}^+$ ,  $y - 1 \in \mathbb{R}^+$  if and only if  $y - 1 > 0$  (by the definition of  $\mathbb{R}^+$ ), i.e.  $y > 1$ . So we see that  $y$  has a pre-image if and only if  $y > 1$ . This proves that the image of  $f_2$  is the set  $\{y \in \mathbb{R}^+ \mid y > 1\}$  and, since this is not the whole of  $\mathbb{R}^+$ ,  $f_2$  is not surjective. As usual, in the formal proof we give an explicit counterexample, an element of  $\mathbb{R}^+$  which is not a value of the function.

However, the function is injective and the proof is just the same as in the previous example.

*Solution* The function is injective since, for  $x_1, x_2 \in \mathbb{R}^+$ ,  $f_2(x_1) = f_2(x_2) \Rightarrow x_1 + 1 = x_2 + 1 \Rightarrow x_1 = x_2$ . However, if  $x \in \mathbb{R}^+$  then  $x > 0$  so that  $f_2(x) = x + 1 > 1$ . Thus  $1/2$  is not a value of  $f_2$  which is not then surjective.  $\square$

**Example 9.1.7** To determine whether the function  $f_3: \mathbb{R} \rightarrow \mathbb{R}$  given by  $f_3(x) = 4x^2 - 4x + 2$  is injective, surjective or bijective.

This time the pre-image approach gives

$$\begin{aligned} y = f_3(x) &\Leftrightarrow y = 4x^2 - 4x + 2 \\ &\Leftrightarrow y = (2x - 1)^2 + 1 \\ &\Leftrightarrow x = (1 \pm \sqrt{y - 1})/2. \end{aligned}$$

Again we must check whether our formula for the pre-image makes sense. Now if  $y < 1$  then  $y - 1$  has no (real) square root and so in that case the formula has no meaning. Looking back through the steps above we see what the problem is:  $y = (2x - 1)^2 + 1$  tells us that if  $y$  is a value then  $y \geq 1$ . This tells us that  $f_3$  is not surjective.

The other thing we notice in the formula for the pre-image is the symbol  $\pm$ . This means that  $x$  is a pre-image for  $y$  if and only if  $x = (1 + \sqrt{y - 1})/2$  or  $x = (1 - \sqrt{y - 1})/2$ . In other words if  $\sqrt{y - 1}$  exists and is non-zero, then  $y$  has two pre-images. For example, taking  $y = 2$  for which  $\sqrt{y - 1} = 1$ ,† we obtain  $f_3(x) = 2 \Leftrightarrow x = (1 \pm 1)/2$ , i.e.  $x = 1$  or  $x = 0$ . So  $f_3$  is not injective.

*Solution* Because  $f_3(0) = f_3(1) = 2$ , the function is not injective.

For a real number  $x$ ,  $f_3(x) = 4x^2 - 4x + 2 = (2x - 1)^2 + 1 \geq 1$ . Thus  $0$  is not a value of  $f_3$  and so it is not surjective.  $\square$

Notice that in presenting this proof no indication is given of how we found the counterexample to injectivity. Formally this is not required for the proof. In a case like this it is not difficult to spot such a counterexample (for example from the graph of the function or from rewriting the formula as  $(2x - 1)^2 + 1$ ) but it is often a good idea to help the reader by indicating how you found the counterexample.

† We adopt the usual convention that if  $y$  is a non-negative real number then  $\sqrt{y}$  represents the non-negative square root of  $y$ , i.e.  $x = \sqrt{y}$  if and only if  $y = x^2$  and  $x \geq 0$ .

**Example 9.1.8** To determine whether the function  $f_4: \mathbb{R}^+ \rightarrow \{x \in \mathbb{R}^+ \mid x \geq 1\}$  given by  $f_4(x) = 4x^2 - 4x + 2$  is injective, surjective or bijective.

The pre-image approach gives just the same formula as in the previous example. However, if  $y$  is in the codomain then  $y \geq 1$  and so the square root in the formula for a pre-image element does determine a real number. Furthermore  $(1 + \sqrt{y-1})/2 > 1/2$  and so this gives one pre-image for  $y$  showing that  $f_4$  is surjective. For  $y > 1$ , we have a second pre-image precisely when  $(1 - \sqrt{y-1})/2 > 0$  which occurs when  $\sqrt{y-1} < 1$ , i.e.  $y < 2$ . This proves that  $f_4$  is not injective since elements  $y$  in the codomain have two pre-images if  $y < 2$ . However, it is probably more satisfying to give two specific elements of the domain where the function takes the same value; we can get these by taking  $y = 5/4$  which has  $\sqrt{y-1} = 1/2$  and so the pre-images  $(1 \pm 1/2)/2$ , i.e.  $3/4$  and  $1/4$ .

Sketching the graph of this function is a great aid to seeing what is going on. The reader should do this.

*Solution* Notice that  $y = f_4(x) \Leftrightarrow x = (1 \pm \sqrt{y-1})/2$ . The function is surjective since given  $y \geq 1$ ,  $y = f_4(x)$  for  $x = (1 + \sqrt{y-1})/2 \in \mathbb{R}^+$ . The function is not injective since  $f_4(1/4) = f_4(3/4) = 5/4$ .  $\square$

## 9.2 bijections and inverses

**Definition 9.2.1** A function  $f: X \rightarrow Y$  is called **invertible** if† there exists a function  $g: Y \rightarrow X$  such that

$$y = f(x) \Leftrightarrow x = g(y) \quad \text{for all } x \in X \text{ and for all } y \in Y.$$

In this case we call  $g$  an **inverse (function)** of  $f$  and write  $g = f^{-1}$ .

The symmetry of the definition shows that in this case  $g$  is also invertible and  $f$  is an inverse of  $g$ .

This means then that corresponding to the assignment of an element of  $Y$  to each element of  $X$  there is a reverse assignment of an element of  $X$  to each element of  $Y$ .

† The reader should note the way that the word ‘if’ is used in this definition. It really means ‘if and only if’ since we are defining the meaning of the word ‘invertible’. Saying that  $f$  is invertible means precisely the same as saying that there is a function  $g$  with the properties given. This usage in definitions is very common although it has on the whole been avoided so far in this book.

**Example 9.2.2** Suppose that  $f: \mathbb{R} \rightarrow \mathbb{R}$  is defined by  $f(x) = 2x + 1$  and  $g: \mathbb{R} \rightarrow \mathbb{R}$  is defined by  $g(x) = (x - 1)/2$ . Then, for  $x, y \in \mathbb{R}$ ,

$$y = f(x) \Leftrightarrow y = 2x + 1 \Leftrightarrow x = (y - 1)/2 \Leftrightarrow x = g(y).$$

Thus  $f$  and  $g$  are invertible and each is the inverse of the other.

**Theorem 9.2.3** *Let  $f: X \rightarrow Y$ . Then  $f$  is invertible if and only if it is a bijection. Furthermore, if it is invertible, then its inverse function is unique.*

**Constructing a proof.** Intuitively this result is pretty clear. For example if we think of the model of the function given by placing elements from the set  $X$  in the set of boxes  $Y$  then we have observed that if the function is bijective there is precisely one element of  $X$  in each box. In this case the inverse function is given by associating to each box the element which it contains. Furthermore this procedure is only possible if there is only one element in each box which means that  $f$  is bijective.

The formal proof makes these ideas a bit more precise in terms of the formal definitions of bijectivity and inverse and is constructed as so often by simply spelling out these definitions. It may seem quite long, but if you are clear about what you are trying to do there is really only one possible way forward at each stage and so the proof more or less writes itself.

The last part of the theorem asserts that  $f$  has a unique inverse. This means that  $f$  has an inverse (for this is what is meant by stating that it is invertible) but only one inverse. Uniqueness statements are very common. The usual approach to proving them is illustrated in this case: we demonstrate that if  $g_1$  and  $g_2$  are inverses of  $f$  then  $g_1 = g_2$ . Arguments of this type have already been met in proofs of the injectivity of a function  $f$ , for this asserts that each value of the function has a unique pre-image: this is proved by demonstrating that if  $x_1$  and  $x_2$  are pre-images of the same element then  $x_1 = x_2$  (see for example Example 9.1.5).

*Proof* For the first statement two things have to be proved.

(a) ' $f$  invertible  $\Rightarrow f$  bijective': Suppose that  $f$  is invertible. Then, by definition, there is an inverse function  $g: Y \rightarrow X$  so that

$$y = f(x) \Leftrightarrow x = g(y) \quad \text{for all } x \in X \text{ and for all } y \in Y.$$

Now to prove that  $f$  is bijective we must prove that  $f$  is surjective and that it is injective.

For injectivity, suppose that  $x_1, x_2 \in X$  are such that  $f(x_1) = f(x_2)$ ; then we must prove that  $x_1 = x_2$ . Put  $y_0 = f(x_1) = f(x_2)$ . Then  $y_0 = f(x_1) \Rightarrow x_1 = g(y_0)$  (from the definition of the inverse) but on the other hand  $y_0 = f(x_2) \Rightarrow x_2 = g(y_0)$ . Hence  $x_1 = g(y_0) = x_2$  as required.

For surjectivity suppose that  $y_0 \in Y$ . We must demonstrate that there exists an element  $x_0 \in X$  such that  $y_0 = f(x_0)$ . Put  $x_0 = g(y_0)$ . Then it is immediate from the definition of an inverse function that  $y_0 = f(x_0)$  as required.

(b) ' $f$  bijective  $\Rightarrow f$  invertible': Suppose that  $f$  is bijective. To show that it is invertible we must construct an inverse function  $g: Y \rightarrow X$ . Suppose that  $y_0 \in Y$ . Since  $f$  is a bijection,  $y_0$  has precisely one pre-image, say  $x_0 \in X$ , such that  $f(x_0) = y_0$ . So we can define a function by the rule that, for each  $y \in Y$ ,  $g(y)$  is the unique element  $x \in X$  such that  $f(x) = y$ . It follows from this definition that

$$y = f(x) \Leftrightarrow x = g(y) \quad \text{for all } x \in X \text{ and for all } y \in Y$$

and so  $g$  is an inverse of  $f$ .

(c) ' $f$  invertible  $\Rightarrow$  it has a unique inverse': Suppose that  $f$  is invertible and  $g_1: Y \rightarrow X$  and  $g_2: Y \rightarrow X$  are inverse functions for  $f$ . To demonstrate that  $g_1 = g_2$  we must show that  $g_1(y) = g_2(y)$  for all  $y \in Y$ . Let  $y_0 \in Y$ . Then put  $x_1 = g_1(y_0)$  and  $x_2 = g_2(y_0)$  so that we are required to prove that  $x_1 = x_2$ . This is easy because  $x_1 = g_1(y_0) \Rightarrow y_0 = f(x_1)$  and  $x_2 = g_2(y_0) \Rightarrow y_0 = f(x_2)$  so that  $f(x_1) = f(x_2)$ . But now, since  $f$  is invertible it is bijective (by (a)) and so in particular injective. It follows that  $x_1 = x_2$ , i.e.  $g_1(y_0) = g_2(y_0)$  as required.  $\square$

Because of this theorem the words ‘bijective’ and ‘invertible’ are used interchangeably when referring to functions. Many standard functions are not actually bijections but by suitable restrictions of the domain and codomain can be converted into bijections. Here are some examples from calculus.

**Examples 9.2.4** (a) The function  $\sin: \mathbb{R} \rightarrow \mathbb{R}$  is neither injective (since, for example,  $\sin 0 = \sin \pi$ ) nor surjective (since, for example, there is no real number  $x$  such that  $\sin x = 2$ ). To define an inverse function  $\sin^{-1}$  we restrict the domain and the codomain of  $\sin$  so that it is bijective. Surjectivity is easy: we simply change the codomain to the image of  $\sin$ ,

i.e.  $\text{Im}(\sin)$ , which is the set  $[-1, 1] = \{x \in \mathbb{R} \mid -1 \leq x \leq 1\}$  (since we know that each number in this set arises as a value of the sine function on  $\mathbb{R}$ ). Injectivity is not much harder for we know that as  $x$  varies from  $-\pi/2$  to  $\pi/2$  the value of  $\sin x$  increases steadily from  $-1$  to  $1$  (sketch the graph). So we take as our domain  $[-\pi/2, \pi/2] = \{x \in \mathbb{R} \mid -\pi/2 \leq x \leq \pi/2\}$ . To sum up, the function†

$$\sin: [-\pi/2, \pi/2] \rightarrow [-1, 1]$$

is a bijection. By the above theorem it has an inverse

$$\sin^{-1}: [-1, 1] \rightarrow [-\pi/2, \pi/2].$$

In making this definition we had to choose a subset of  $\mathbb{R}$  on which the sine function was injective. There are many possibilities, for example  $[\pi/2, 3\pi/2]$ , and each has its own inverse function. The above choice is the most natural and the function it leads to is called the *principal value* of the inverse.

In the same way we can restrict the cosine and tangent functions to obtain inverse functions

$$\cos^{-1}: [-1, 1] \rightarrow [0, \pi]$$

and

$$\tan^{-1}: \mathbb{R} \rightarrow (-\pi/2, \pi/2).$$

(b) Given  $n \in \mathbb{Z}^+$ , the function  $x \mapsto x^n$  is a bijection  $\mathbb{R} \rightarrow \mathbb{R}$  for  $n$  odd and so in this case we do have an inverse function  $\mathbb{R} \rightarrow \mathbb{R}$ , the  *$n$ th root function*, denoted by‡  $x \mapsto x^{1/n}$  or  $x \mapsto \sqrt[n]{x}$ .

† Strictly speaking this function should have a different name in order to indicate the change of domain and codomain. However, it is usual to use the same name.

‡ The notation  $x^{1/n}$  is selected so that the law of exponents holds:  $(x^{1/n})^n = (x^n)^{1/n} = x^1 = x$ . With this definition it is possible to give some indication of the problem in defining  $0^0$  referred to in Definition 5.3.3. Notice that  $0^{1/n}$  is defined to be 0. Furthermore, we have observed (Exercise 8.4) that the sequence  $1/n$  is null,  $\lim 1/n = 0$ , which suggests that we ought to define  $0^0 = 0$ . However, if the law of indices is to hold then we must have  $x^0 = 1$  for non-zero  $x$  (see the solution to Exercise 5.7) and this suggests that  $0^0 = 1$ , the convention adopted in this book. It is not possible to satisfy these conflicting demands. For  $(x, y) \in \mathbb{R}^{\geq} \times \mathbb{R} - \{(0, 0)\}$  it is possible to define  $x^y$  extending the definition for integer exponents so that the laws of exponents are satisfied and so that the function of two variables  $(x, y) \mapsto x^y$  is ‘continuous’. It is not possible to extend this ‘continuous’ function to the point  $(0, 0)$ . A full explanation of this, and the definition of ‘continuous’, is beyond the scope of this book (see for example K.G. Binmore, *Mathematical analysis, a straightforward approach*, Cambridge University Press, Second edition 1982).

On the other hand it is neither an injection nor a surjection when  $n$  is even. [The case  $n = 2$  is typical and is discussed in Example 9.1.3(c).] In this case we restrict the domain and the codomain to  $\mathbb{R}^{\geq}$  so that we have a bijection. Thus the inverse function  $x \mapsto x^{1/n} = \sqrt[n]{x}$  is a function  $\mathbb{R}^{\geq} \rightarrow \mathbb{R}^{\geq}$ . The convention is that for even  $n$  the expressions  $x^{1/n}$  and  $\sqrt[n]{x}$  represent the non-negative root of a non-negative real number  $x$ .

The following equivalent formulation of Definition 9.2.1 is often useful.

**Proposition 9.2.5** *The functions  $f: X \rightarrow Y$  and  $g: Y \rightarrow X$  are inverses of each other if and only if  $g \circ f = I_X$  and  $f \circ g = I_Y$ .*

*Proof* ‘ $\Rightarrow$ ’: Suppose that  $f$  and  $g$  are inverse to each other. Then

$$y = f(x) \Leftrightarrow x = g(y) \quad \text{for all } x \in X \text{ and all } y \in Y.$$

Now, given  $x_0 \in X$ , put  $y_0 = f(x_0)$  so that  $x_0 = g(y_0)$ . Then  $g \circ f(x_0) = g(y_0) = x_0$ , i.e.  $g \circ f = I_X$ .

Similarly  $f \circ g = I_Y$ .

‘ $\Leftarrow$ ’: Suppose that  $g \circ f = I_X$  and  $f \circ g = I_Y$ . Suppose that  $y_0 = f(x_0)$ . Then  $g(y_0) = g \circ f(x_0) = I_X(x_0) = x_0$ . Thus  $y = f(x) \Rightarrow x = g(y)$  for all  $x \in X, y \in Y$ .

Similarly,  $x = g(y) \Rightarrow y = f(x)$  and so we have proved that  $f$  and  $g$  are inverses of each other.  $\square$

### 9.3 Functions and subsets

In the previous section it has been demonstrated that a function  $f$  has an inverse  $f^{-1}$  if and only if it is a bijection. However, the reader will discover that the notation  $f^{-1}$  which has been used for the inverse function is used even when the function  $f$  is not a bijection. This chapter ends with a brief description of this more general use.

To begin with recall the definition of the power set (Definition 6.3.1): the power set  $\mathcal{P}(X)$  of a set  $X$  is the set of subsets of  $X$  so that  $A \in \mathcal{P}(X)$  means simply  $A \subseteq X$ . We now describe how given a function  $f: X \rightarrow Y$  we can associate to it two functions between the power sets of  $X$  and  $Y$  (one in each direction) each of which captures the function  $f$  in a different way.

**Definition 9.3.1** *Suppose that  $f: X \rightarrow Y$  is a function.*

(i) The function  $\vec{f} : \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$  is defined by

$$\vec{f}(A) = \{ f(x) \mid x \in A \}$$

for  $A \in \mathcal{P}(X)$ .

(ii) The function  $\overleftarrow{f} : \mathcal{P}(Y) \rightarrow \mathcal{P}(X)$  is defined by

$$\overleftarrow{f}(B) = \{ x \in X \mid f(x) \in B \}$$

for  $B \in \mathcal{P}(Y)$ .

**Remarks 9.3.2** The notations  $\vec{f}$  and  $\overleftarrow{f}$  are non-standard. Most writers denote the function  $\vec{f}$  simply by  $f$  and denote the function  $\overleftarrow{f}$  by  $f^{-1}$ . With experience this does not lead to confusion (and is another example of how ambiguity can be better than pedantry). However, in this book the notations  $\vec{f}$  and  $\overleftarrow{f}$  will be used since it seems potentially confusing to have two different functions with the same name when meeting these ideas for the first time.

In the case of  $\vec{f}$  notice that each element  $x_0$  of  $X$  corresponds to an element of  $\mathcal{P}(X)$ , namely the singleton subset  $\{x_0\}$ , and similarly for  $Y$ . So we can think of  $\vec{f}$  as an extension of  $f$  in the sense that

$$\vec{f}(\{x_0\}) = \{ f(x) \mid x \in \{x_0\} \} = \{ f(x) \mid x = x_0 \} = \{ f(x_0) \}.$$

Notice that  $\vec{f}(X)$  is the image of  $f$ ,  $\text{Im}(f)$  (see Definition 8.4.1).

In the case of  $\overleftarrow{f}$ ,

$$\overleftarrow{f}(\{y_0\}) = \{ x \in X \mid f(x) \in \{y_0\} \} = \{ x \in X \mid f(x) = y_0 \}$$

for each  $y_0 \in Y$ . Thus  $\overleftarrow{f}(\{y_0\})$  gives the set of pre-images of  $y_0$ . In the box model of a function (page 90) this gives the set of elements in the box  $y_0$ . Now, if  $f$  is a bijection with inverse  $f^{-1}$ , then  $f(x) = y_0$  if and only if  $x = f^{-1}(y_0)$  so that  $\overleftarrow{f}(\{y_0\}) = \{f^{-1}(y_0)\}$  and we can think of  $\overleftarrow{f}$  as an extension of  $f^{-1}$ .

If  $f$  is not a bijection then  $\overleftarrow{f}(\{y\})$  will not be a singleton subset for some elements of  $Y$ : if it is not a surjection then it will be empty for elements  $y$  not in the image, and if not an injection then it will contain more than one element for some  $y$ .

### 9.4 Peano's axioms for the natural numbers

The notion of the successor of an integer  $n$  was introduced in Section 5.1. Chapter 5 was headed by a quotation from Richard Dedekind suggesting that this idea captured the essence of the natural numbers (or positive integers). The necessary language is now available so that some indication of the ideas behind the Dedekind quotation can be given.

**Definition 9.4.1** *The*  $\boxed{\text{successor function}}$ ,  $\boxed{s: \mathbb{Z}^+ \rightarrow \mathbb{Z}^+}$ , *is defined by*  
 $s(n) = n + 1$  *for*  $n \in \mathbb{Z}^+$ .

Dedekind observed that the existence of the successor function together with the number 1 completely captures the properties of the natural numbers. These ideas are now usually associated with the name of the Italian mathematician Giuseppe Peano.<sup>†</sup>

**Axioms 9.4.2 (Peano)** *The set of positive integers  $\mathbb{Z}^+$  is a set with a function  $s: \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$  and an element  $1 \in \mathbb{Z}^+$  such that*

- (i)  *$s$  is an injection,*
- (ii)  *$1$  is not in the image of  $s$ , and*
- (iii) *for  $A \subseteq \mathbb{Z}^+$ , if  $1 \in A$ , and  $n \in A \Rightarrow s(n) \in A$ , then  $A = \mathbb{Z}^+$ .*

Notice that (iii) is just the induction axiom as reformulated in Axiom 7.5.1.

<sup>†</sup> The first exposition of these ideas was by Richard Dedekind in his book *Was sind und was sollen die Zahlen?* published in 1888. Peano formulated his axioms for the natural numbers in his book *Arithmetices principia, nova methodo exposita* published in 1889. Peano did not see Dedekind's work until his own book was in press and Peano's exposition using the language of set theory is considered to have been much more influential. It was in this book that he introduced notations for set membership (the modern symbol  $\in$ ) and set inclusion (he used an inverted 'C' which he also used for implication – recall that it was observed in Section 6.1 that there is a strong connection between set inclusion and implication) and he seems to have been the first to clarify the distinction between set membership and set inclusion. Peano was a remarkable mathematician. It is curious that his 1889 book was written in Latin rather than the Italian or French he usually used. In his book about Peano (Reidel, 1980), Hubert Kennedy describes it as 'the unique romantic act of his scientific career' and suggests that it reflected Peano's awareness that he was doing something historic – after all, the great classics such as Isaac Newton's *Principia* had been written in Latin. The following year brought Peano's most striking achievement, his construction of a space-filling curve; this is a curve in the plane  $\mathbb{R}^2$  given by continuous functions  $x = f(t)$ ,  $y = g(t)$  such that, as  $t$  varies over the interval  $[0, 1]$ ,  $(x, y)$  passes through every point of the unit square  $[0, 1] \times [0, 1]$ . This was perhaps the first indication of the subtlety of the notion of dimension, confounding intuition, whose exploration has now led to the modern study of fractal sets. (See for example Donald M. Davis, *The nature and power of mathematics*, Princeton University Press, 1993.)

It is not difficult to prove (by induction) that, given any set  $X$  with a function  $s : X \rightarrow X$  and a distinguished element  $1 \in X$  satisfying these axioms, there is a bijection  $X \rightarrow \mathbb{Z}^+$  under which the distinguished elements 1 and the successor functions correspond.† This is what is meant by saying that these statements provide axioms for the positive integers.

We can use  $s$  to define algebraic operations on  $\mathbb{Z}^+$  inductively as follows.

**Definition 9.4.3 (Peano)** *The sum  $m + n$  of positive integers  $m$  and  $n$  may be defined by induction on  $n$  by*

- (i)  $m + 1 = s(m)$ , and
- (ii) for  $k \in \mathbb{Z}^+$ ,  $m + s(k) = s(m + k)$ .

*The product of positive integers  $mn$  or  $m \times n$  is now defined (making use of addition) by induction on  $n$  by*

- (i)  $m \times 1 = m$ , and
- (ii) for  $k \in \mathbb{Z}^+$ ,  $m \times s(k) = m \times k + m$ .

It is an interesting exercise to prove the basic algebraic properties of addition and multiplication starting from these definitions, for example the commutativity of addition  $m + n = n + m$ .

### Exercises

**9.1** Determine whether each of the following functions  $\mathbb{R} \rightarrow \mathbb{R}$  is injective, surjective or bijective.

- (i)  $f_1(x) = 2x + 5$ .
- (ii)  $f_2(x) = x^2 + 2x + 1$ .
- (iii)  $f_3(x) = x^2 - 2x$ .
- (iv)  $f_4(x) = \begin{cases} 1/x & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$

**9.2** In each of the above examples consider the effect of changing the domain and the codomain to  $\mathbb{R}^+$ .

† See for example G. Birkhoff and S. Mac Lane, *A survey of modern algebra*, Macmillan, Fourth edition 1977.

**9.3** Find inverses for the following functions:

- (i)  $f_1: \mathbb{R} \rightarrow \mathbb{R}$  given by  $f_1(x) = 3x + 2$ ;
- (ii)  $f_2: \mathbb{R} \rightarrow \mathbb{R}$  given by  $f_2(x) = x^3 + 1$ .

**9.4** Suppose that  $f: X \rightarrow Y$  and  $g: Y \rightarrow Z$  are injections. Prove that  $g \circ f: X \rightarrow Z$  is an injection.

**9.5** Suppose that  $f: X \rightarrow Y$  and  $g: Y \rightarrow Z$  are bijections of sets. Prove that the composite  $g \circ f: X \rightarrow Z$  is also a bijection and that

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}: Z \rightarrow Y \rightarrow X.$$

**9.6** Let  $f: X \rightarrow Y$  be a function with graph  $G_f \subseteq X \times Y$ . Prove that  $f$  is surjective if and only if  $\forall y \in Y$ ,  $(X \times \{y\} \cap G_f) \neq \emptyset$ .

**9.7** Let  $f: X \rightarrow Y$  be a function and  $B_1, B_2 \in \mathcal{P}(Y)$ . Prove that

- (i)  $B_1 \subseteq B_2 \Rightarrow \overleftarrow{f}(B_1) \subseteq \overleftarrow{f}(B_2)$ ,
- (ii)  $\overleftarrow{f}(B_1 \cap B_2) = \overleftarrow{f}(B_1) \cap \overleftarrow{f}(B_2)$ ,
- (iii)  $\overleftarrow{f}(B_1 \cup B_2) = \overleftarrow{f}(B_1) \cup \overleftarrow{f}(B_2)$ .

Prove that the converse of the first of these statements is not universally true (by constructing a counterexample).

## Problems II: Sets and functions

**1** Prove the following statements:

- (i)  $\{x \in \mathbb{R} \mid x^2 - 3x + 2 = 0\} = \{x \in \mathbb{Z} \mid 0 < x < 3\};$
- (ii)  $\{x \in \mathbb{R} \mid x^2 - 3x + 2 < 0\} = \{x \in \mathbb{R} \mid x < 2\} \cap \{x \in \mathbb{R} \mid x > 1\};$
- (iii)  $\{x \in \mathbb{R} \mid x^2 - 3x + 2 > 0\} = \{x \in \mathbb{R} \mid x > 2\} \cup \{x \in \mathbb{R} \mid x < 1\}.$

**2.** By using a truth table prove that, for sets  $A$ ,  $B$  and  $C$ ,

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

Draw a Venn diagram to illustrate the proof.

**3.** Prove the absorption laws:

- (i)  $A \cap (A \cup B) = A;$
- (ii)  $A \cup (A \cap B) = A.$

**4.** Prove by contradiction or otherwise that  $A \cap B = A \cap C$  and  $A \cup B = A \cup C$  if and only if  $B = C$ .

**5.** Using truth tables, prove that for sets  $A$ ,  $B$  and  $C$ ,

- (i)  $(A \cup C) - B \subseteq (A - B) \cup C,$
- (ii)  $(A \cap C) - B = (A - B) \cap C.$

Draw Venn diagrams to illustrate the proofs.

Prove that there is equality in the first of these results if and only if  $B \cap C = \emptyset$ .

Deduce from the second of these results that

$$(A - B) \cap C = \emptyset \text{ if and only if } A \cap C \subseteq B.$$

**6.** Use the distributivity law to prove that

$$(A \cap B) \cup (B \cap C) \cup (C \cap A) = (A \cup B) \cap (B \cup C) \cap (C \cup A).$$

**7.** For subsets of a universal set  $U$  prove that  $B \subseteq A^c$  if and only if  $A \cap B = \emptyset$ . By taking complements deduce that  $A^c \subseteq B$  if and only if  $A \cup B = U$ . Deduce that  $B = A^c$  if and only if  $A \cap B = \emptyset$  and  $A \cup B = U$ .

**8.** Given sets  $A, B \in \mathcal{P}(X)$ , their *symmetric difference* is defined by

$$A \Delta B = (A - B) \cup (B - A) = (A \cup B) - (A \cap B).$$

Prove that

- (i) the symmetric difference is associative,  $(A \Delta B) \Delta C = A \Delta (B \Delta C)$  for all  $A, B, C \in \mathcal{P}(X)$ ,
- (ii) there exists a unique set  $N \in \mathcal{P}(X)$  such that  $A \Delta N = A$  for all  $A \in \mathcal{P}(X)$   
[Hint: Guess what  $N$  is!],
- (iii) for each  $A \in \mathcal{P}(X)$ , there exists a unique  $A' \in \mathcal{P}(X)$  such that  $A \Delta A' = N$ ,
- (iv) for each  $A, B \in \mathcal{P}(X)$ , there exists a unique set  $C$  such that  $A \Delta C = B$ .

**9.** Using the notation of the previous problem, prove that for sets  $A, B, C, D \in \mathcal{P}(X)$

$$A \Delta B = C \Delta D \Leftrightarrow A \Delta C = B \Delta D.$$

**10.** We define half-infinite† intervals as follows:

$$\begin{aligned} (a, \infty) &= \{x \in \mathbb{R} \mid x > a\}; \\ [a, \infty) &= \{x \in \mathbb{R} \mid x \geq a\}. \end{aligned}$$

Prove that

- (i)  $(a, \infty) \subseteq [b, \infty) \Leftrightarrow a \geq b$ ,
- (ii)  $[a, \infty) \subseteq (b, \infty) \Leftrightarrow a > b$ .

† The symbol ‘ $\infty$ ’ used in the notation in this question does not represent a number ‘infinity’. The expression  $(a, \infty)$  is used by analogy with  $(a, b)$ , where  $b$  is a real number, but the definitions are different (cf. Exercise 6.1). In this book no meaning is attached to the symbol ‘ $\infty$ ’ on its own and it is only used in the notations in this question and in Definition 8.3.2. Further discussion of this use of the symbol can be found in books on analysis, for example R. Haggerty, *Fundamentals of mathematical analysis*, Addison-Wesley, Second edition 1993.

**11.** Give a proof or a counterexample for each of the following statements:

- (i)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, x + y > 0;$
- (ii)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, x - y > 0;$
- (iii)  $\exists x \in \mathbb{R}, \forall y \in \mathbb{R}, x + y > 0;$
- (iv)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, xy > 0;$
- (v)  $\exists x \in \mathbb{R}, \forall y \in \mathbb{R}, xy > 0;$
- (vi)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, xy \geq 0;$
- (vii)  $\exists x \in \mathbb{R}, \forall y \in \mathbb{R}, xy \geq 0;$
- (viii)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, (x + y > 0 \text{ or } x + y = 0);$
- (ix)  $\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, (x + y > 0 \text{ and } x + y = 0);$
- (x)  $(\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, x + y > 0) \text{ and } (\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, x + y = 0).$

**12.** Suppose that  $A \subseteq \mathbb{Z}$ . Write the following statement entirely in symbols using the quantifiers  $\forall$  and  $\exists$ . Write out the negative of this statement in symbols.

There is a greatest number in the set  $A$ .

Give an example of a set  $A$  for which this statement is true. Give another example for which it is false.

**13.** Prove that, for sets  $A, B, C$  and  $D$ ,

- (i)  $A \times (B \cup C) = (A \times B) \cup (A \times C),$
- (ii)  $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D),$

**14.** Define functions  $f$  and  $g: \mathbb{R} \rightarrow \mathbb{R}$  by  $f(x) = x^2$  and  $g(x) = x^2 - 1$ .

Find the functions  $f \circ f$ ,  $f \circ g$ ,  $g \circ f$ ,  $g \circ g$ .

List the elements of the set  $\{x \in \mathbb{R} \mid fg(x) = gf(x)\}$ .

**15.** Given  $A \in \mathcal{P}(X)$  define the *characteristic function*  $\chi_A: X \rightarrow \{0, 1\}$  by

$$\chi_A(x) = \begin{cases} 0 & \text{if } x \notin A, \\ 1 & \text{if } x \in A. \end{cases}$$

Suppose that  $A$  and  $B$  are subsets of  $X$ .

- (i) Prove that the function  $x \mapsto \chi_A(x)\chi_B(x)$  (multiplication of integers) is the characteristic function of the intersection  $A \cap B$ .
- (ii) Find the subset  $C$  whose characteristic function is given by

$$\chi_C(x) = \chi_A(x) + \chi_B(x) - \chi_A(x)\chi_B(x).$$

**16.** Determine which of the following functions  $f_i: \mathbb{R} \rightarrow \mathbb{R}$  are injective, which are surjective and which are bijective. Write down an inverse function of each of the bijections.

- (i)  $f_1(x) = x - 1;$
- (ii)  $f_2(x) = x^3;$
- (iii)  $f_3(x) = x^3 - x;$
- (iv)  $f_4(x) = x^3 - 3x^2 + 3x - 1;$
- (v)  $f_5(x) = e^x;$
- (vi)  $f_6(x) = \begin{cases} x^2 & \text{if } x \geq 0, \\ -x^2 & \text{if } x \leq 0. \end{cases}$

**17.** Functions  $f: \mathbb{R} \rightarrow \mathbb{R}$  and  $g: \mathbb{R} \rightarrow \mathbb{R}$  are defined as follows.

$$\begin{aligned} f(x) &= \begin{cases} x+2 & \text{if } x < -1, \\ -x & \text{if } -1 \leq x \leq 1, \\ x-2 & \text{if } x > 1. \end{cases} \\ g(x) &= \begin{cases} x-2 & \text{if } x < -1, \\ -x & \text{if } -1 \leq x \leq 1, \\ x+2 & \text{if } x > 1. \end{cases} \end{aligned}$$

Find the functions  $f \circ g$  and  $g \circ f$ . Is  $g$  the inverse of the function  $f$ ? Is  $f$  injective or surjective? How about  $g$ ? Sketch and compare the graphs of these functions.

**18.** Suppose that  $f: X \rightarrow Y$  and  $g: Y \rightarrow Z$  are surjections. Prove that the composite  $g \circ f: X \rightarrow Z$  is a surjection.

**19.** Let  $f: X \rightarrow Y$  be a function. Prove that there exists a function  $g: Y \rightarrow X$  such that  $f \circ g = I_Y$  if and only if  $f$  is a surjection. [ $g$  is called a *right inverse* of  $f$ .]

**20.** Let  $f: X \rightarrow Y$  be a function and  $A_1, A_2 \in \mathcal{P}(X)$ .

- (i) Prove that  $A_1 \subseteq A_2 \Rightarrow \vec{f}(A_1) \subseteq \vec{f}(A_2)$ . Prove that the converse is not universally true. Give a simple condition on  $f$  which is equivalent to the converse.
- (ii) Prove that  $\vec{f}(A_1 \cap A_2) \subseteq \vec{f}(A_1) \cap \vec{f}(A_2)$ . Prove that equality is not universally true.
- (iii) Prove that  $\vec{f}(A_1 \cup A_2) = \vec{f}(A_1) \cup \vec{f}(A_2)$ .

21. Let  $f: X \rightarrow Y$  be a function. Prove that

- (i)  $f$  is injective  $\Leftrightarrow \overline{f}$  is injective  $\Leftrightarrow \overleftarrow{f}$  is surjective,
- (ii)  $f$  is surjective  $\Leftrightarrow \overline{f}$  is surjective  $\Leftrightarrow \overleftarrow{f}$  is injective.

22. Starting from Peano's axioms prove that if  $n \in \mathbb{Z}^+$  and  $n \neq 1$  then  $n$  is a successor, i.e.  $s(a) = n$  for some  $a \in \mathbb{Z}^+$ .

[Let  $A = \text{Im}(s) \cup \{1\}$  and prove that  $A = \mathbb{Z}^+$ .]

23. Starting from Definition 9.4.3 prove that addition of positive integers is

- (i) associative, i.e.  $(a + b) + c = a + (b + c)$  for positive integers  $a, b, c$ ,
- (ii) commutative, i.e.  $a + b = b + a$  for positive integers  $a, b$ .

[For (ii), prove first of all that  $a + 1 = 1 + a$  by induction on  $a$ .]



# **Part III**

## Numbers and counting



# 10

## Counting

One, two, buckle my shoe;  
Three, four, knock at the door;  
Five, six, pick up sticks;  
Seven, eight, lay them straight;  
Nine, ten, a big fat hen.  
Eleven, twelve, dig and delve;  
Thirteen, fourteen, maids a-courtin';  
Fifteen, sixteen, maids a-kissing;  
Seventeen, eighteen, maids a-waitin';  
Nineteen, twenty, my plate's empty.

*Traditional.*

Perhaps the oldest mathematical activity is counting. Some of the earliest words we learn as children are the names of the first few natural numbers and counting games are very common. In this chapter the counting process is formulated mathematically as the constructing of a bijection from a standard set of positive integers to the set of the elements being counted. This allows us to consider precisely what we mean by the number of elements in a set and what it means for a set to be finite.

### 10.1 Counting finite sets

Recall that we have suggested that a set can be thought of as a box containing the elements of the set. The process of counting the elements in the set or box can be carried out by removing them in turn one by one while reciting the names of the positive integers in order: ‘one’, ‘two’, ‘three’, and so on. When we have exhausted the elements, we say that the last number recited is the number of elements in the set, or the *cardinality* of the set. Sometimes, instead of removing the elements we simply point to them in turn as when we count the number of people in a

room, but the idea is the same. What this amounts to is the construction of a bijection from the standard set

$$\mathbb{N}_n = \{1, 2, \dots, n\} = \{i \in \mathbb{Z} \mid 1 \leq i \leq n\},$$

the set of the first  $n$  positive integers for some positive integer  $n$ , to the set being counted. As we count a set  $X$ , the counting process effectively defines a bijection  $f: \mathbb{N}_n \rightarrow X$  by setting  $f(1)$  equal to the first element of  $X$  which we count,  $f(2)$  equal to the second element of  $X$  which we count, and so on. The function  $f$  is an injection because we do not count any element twice and it is a surjection because every element is counted. When we have constructed the bijection  $f: \mathbb{N}_n \rightarrow X$  we can list the elements of the set  $X = \{x_1, x_2, \dots, x_n\}$  where we write  $x_i = f(i)$ . Thus the function  $f$  is described by the following table.

|        |       |       |       |         |           |       |
|--------|-------|-------|-------|---------|-----------|-------|
| $i$    | 1     | 2     | 3     | $\dots$ | $n - 1$   | $n$   |
| $f(i)$ | $x_1$ | $x_2$ | $x_3$ | $\dots$ | $x_{n-1}$ | $x_n$ |

Because  $f$  is a bijection each element of  $X$  occurs precisely once in the second row.

We can formulate this as a mathematical definition. It is natural to extend the definition to include the empty set, the set with no elements.<sup>†</sup>

**Definition 10.1.1** *Given a set  $X$ , if there is a bijection  $f: \mathbb{N}_n \rightarrow X$  then we say that the cardinality of  $X$ , or the number of elements in  $X$ , is  $n$  and write  $|X| = n$ .*

*The cardinality of the empty set is defined to be 0,  $|\emptyset| = 0$ .*

**Examples 10.1.2** (a) The cardinality of the set  $A = \{k \in \mathbb{Z} \mid 25 < k \leq 30\}$  is 5. When we list the elements of the set, say  $\{29, 26, 30, 27, 28\}$ , we are essentially constructing a bijection  $f: \mathbb{N}_5 \rightarrow A$ . The bijection corresponding to this listing is given by the following table.

|        |    |    |    |    |    |
|--------|----|----|----|----|----|
| $i$    | 1  | 2  | 3  | 4  | 5  |
| $f(i)$ | 29 | 26 | 30 | 27 | 28 |

This is not the most obvious way of counting this set. The bijection

<sup>†</sup> In fact we could define  $\mathbb{N}_0$  to be the empty set so that the case of cardinality 0 is covered by the bijection definition. However, the reader meeting set theory for the first time may be worried by the bijection between two empty sets (it is the function whose graph is the empty subset of  $\emptyset \times \emptyset = \emptyset!$ ) and so it seems best to keep this case separate.

$g: \mathbb{N}_5 \rightarrow A$  corresponding to the obvious way of counting it is given by the following table.

| $i$    | 1  | 2  | 3  | 4  | 5  |
|--------|----|----|----|----|----|
| $g(i)$ | 26 | 27 | 28 | 29 | 30 |

(b) The cardinality of the set  $\mathbb{N}_n$  is  $n$  for the identity map provides a bijection  $\mathbb{N}_n \rightarrow \mathbb{N}_n$ . Another bijection  $f: \mathbb{N}_n \rightarrow \mathbb{N}_n$  is given by  $f(i) = n - i$ .

There is one rather subtle difficulty about this definition: we need to be sure that if we count a set in two different ways then we get the same answer. For example, in Example 10.1.2(a), we gave two ways of counting the set  $\{ k \in \mathbb{Z} \mid 25 < k \leq 30 \}$  and there are others. However, whatever order we use to list this set of integers we find that the fifth element is the last one.

It is common to count a set more than once if we want to be sure about the number of elements in it. If a set is counted twice, for example when counting the number of people in a room, and the answers differ then we assume either that a mistake was made on one or both occasions (in other words the person counting failed to construct a bijection; it can be difficult to be sure that you have an injection and that you have a surjection if people keep moving about!) or that it wasn't the same set (somebody entered or left the room and so was counted on one occasion but not the other). In other words we assume that if mistakes are avoided then the answer will be the same even though we do not assume that the bijections will be the same. We can formulate this assumption as follows.

**Proposition 10.1.3** *Suppose that  $f: \mathbb{N}_m \rightarrow X$  and  $g: \mathbb{N}_n \rightarrow X$  are bijections with the same codomain. Then  $m = n$ .*

We need to know this in order that the above definition of the number of elements in a set gives a *well-defined* number. Very often in mathematics definitions are made which involve some choices. Demonstrating that the notion defined is well-defined amounts to proving that it is independent of these choices, in other words different choices lead to the same answer. The notion of cardinality provides a good example of such a definition.

At this point the reader may consider that the author has taken leave of his senses. Surely it is absolutely obvious that the number of elements

in a set is a well-defined notion and two different orders of counting cannot lead to a different answer. This may be so, although our experience is only based on relatively small sets and modern physics suggests that care may be needed when we come to deal, say, with a set of electrons. In Chapter 14 we will look briefly at infinite sets and see there how misleading our intuition can be. It is often the case in mathematics that we know what properties we want some notion to have. The real question is how to define the notion mathematically. If we were unable to prove Proposition 10.1.3 then this would call into question whether Definition 10.1.1 is the right one. It has often taken many years to get a definition right, giving an appropriate mathematical formulation of the meaning of a word which successfully captures the intuitive concept. For example the search for the most useful definition of continuity when applied to functions  $\mathbb{R} \rightarrow \mathbb{R}$  took over a hundred years.

**Constructing a proof of Proposition 10.1.3.** In spite of being intuitively ‘obvious’ this is not an easy result; it would be remarkable for a student to find a proof without some assistance. One difficulty is that it involves a set  $X$  about which we know very little. Let us summarize what we are trying to achieve.

| Given  | Goal    |
|--|---------|
| bijection $f: \mathbb{N}_m \rightarrow X$<br>bijection $g: \mathbb{N}_n \rightarrow X$ | $m = n$ |

To proceed, notice that the goal does not involve the set  $X$ . However, since  $f$  and  $g$  are bijections they are invertible with inverses  $f^{-1}: X \rightarrow \mathbb{N}_m$  and  $g^{-1}: X \rightarrow \mathbb{N}_n$  which are also bijections. So the composite

$$g^{-1} \circ f: \mathbb{N}_m \xrightarrow{f} X \xrightarrow{g^{-1}} \mathbb{N}_n$$

is also a bijection (with inverse  $f^{-1} \circ g$ ). Thus we can use the bijections  $f$  and  $g$  to construct a bijection  $\mathbb{N}_m \rightarrow \mathbb{N}_n$ . Now the statement that such a bijection exists does not involve the set  $X$ . This suggests the following strategy for proving the proposition.

| Given   | Goal    |
|---|---------|
| bijection $\mathbb{N}_m \rightarrow \mathbb{N}_n$ | $m = n$ |

If we can prove this then we will be home since the given statement follows from the hypotheses in the proposition. Notice that this implication is just a special case of the result we are trying to prove with  $X = \mathbb{N}_n$  for we then automatically have a bijection  $g: \mathbb{N}_n \rightarrow \mathbb{N}_n$  given

by the identity map. It is quite common to prove a general result by reducing it to some special case.

Now to tackle this implication we look back to the definition of a bijection: a function is a bijection if and only if it is an injection and a surjection. So we might begin by examining what can be deduced simply from the injectivity of a function  $\mathbb{N}_m \rightarrow \mathbb{N}_n$ . Intuitively this implies that  $m \leq n$ , suggesting the following first step.

| Given   | Goal       |
|---|------------|
| $\text{injection } \mathbb{N}_m \rightarrow \mathbb{N}_n$ | $m \leq n$ |

In fact this implication on its own would be sufficient for our needs. For a bijection  $h: \mathbb{N}_m \rightarrow \mathbb{N}_n$  is an injection and has an inverse  $h^{-1}: \mathbb{N}_n \rightarrow \mathbb{N}_m$  which is also a bijection and so an injection. Applying the implication to  $h$  gives  $m \leq n$  and applying it to  $h^{-1}$  gives  $n \leq m$ . But  $m \leq n$  and  $n \leq m$  only if  $m = n$  (see Exercise 4.5).

This process of reducing a result to be proved to successively simpler or more concrete statements is common in constructing proofs. To keep track of what we are doing it can be useful to formulate these other statements as propositions or theorems in their own right; sometimes the word ‘lemma’ is used instead, indicating that a result’s primary interest is as a stepping stone in the proof of some more striking result or results.

In this case it turns out that the implication we have reached is of sufficient interest to be stated formally as a lemma. So we do this now and write out the formal proof that Proposition 10.1.3 follows from it. We will then return to the proof of the lemma in the next chapter.

*Proof of Proposition 10.1.3* This result follows from the following result to be proved in the next chapter.

**Lemma 10.1.4** *If there exists an injection  $\mathbb{N}_m \rightarrow \mathbb{N}_n$  then  $m \leq n$ .*

For, given bijections  $f: \mathbb{N}_m \rightarrow X$  and  $g: \mathbb{N}_n \rightarrow X$ , they are invertible with inverses  $f^{-1}: X \rightarrow \mathbb{N}_m$  and  $g^{-1}: X \rightarrow \mathbb{N}_n$  which are also bijections. So the composite  $g^{-1} \circ f: \mathbb{N}_m \rightarrow \mathbb{N}_n$  is a bijection, and so an injection. It follows from the lemma that  $m \leq n$ .

But now we can reverse the rôles of  $f$  and  $g$ . The composite  $f^{-1} \circ g: \mathbb{N}_n \rightarrow \mathbb{N}_m$  is an injection and so again by the lemma  $n \leq m$ .

Since we have proved that  $m \leq n$  and  $n \leq m$  it follows that  $m = n$  as required.  $\square$

Of course not all sets can be counted. For example if we start counting the positive integers in the obvious way we never come to an end. It is useful to distinguish sets for which the counting process does end from those which do not; to distinguish finite sets from infinite sets.

**Definition 10.1.5** *Given a set  $X$ , if  $|X| = n$  for some non-negative integer  $n$  then we say that the set is finite. If this is not the case then we say that the set is infinite.*

Thus saying that a set  $X$  is finite means that either it is empty or there exists a bijection  $\mathbb{N}_n \rightarrow X$  for some positive integer  $n$ .

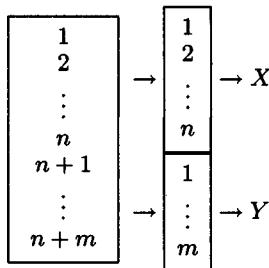
## 10.2 Two basic counting principles

There are two basic counting principles which are frequently used and extremely powerful. The first seems almost too simple to state.

**Theorem 10.2.1 (The addition principle)** *Suppose that  $X$  and  $Y$  are disjoint finite sets. Then  $X \cup Y$  is finite and  $|X \cup Y| = |X| + |Y|$ .*

**Constructing a proof.** The form of words used in the conclusion of this result sometimes causes confusion. Because it is an ‘and’ statement the reader might imagine that it is necessary to prove the first statement and then to prove the second statement. But in fact the first statement is an immediate consequence of the second and, with what we have done so far, the only way to prove that a set is finite is to prove that it has cardinality equal to some non-negative integer. Since  $X$  and  $Y$  are finite sets,  $|X|$  and  $|Y|$  are non-negative integers and thus so is  $|X| + |Y|$ . We obtain the result by simply proving that  $X \cup Y$  has cardinality equal to  $|X| + |Y|$ .

Notice that since the set  $X$  is finite then either  $X$  is empty or there is a bijection  $\mathbb{N}_n \rightarrow X$  for some positive integer  $n$ . It is necessary to consider these two cases separately and similarly for  $Y$ . The cases when  $X$  or  $Y$  has cardinality 0 are trivial. In the other cases the result seems obvious since we simply count the elements in the set  $X$ , say  $n$ , and then count the elements of  $Y$  starting at  $n + 1$ . The following proof simply translates this intuitive idea into the language of bijections, using the bijections to  $X$  and  $Y$  to give a bijection to  $X \cup Y$  as indicated in the diagram on the next page.



*Proof* Let  $|X| = n$  and  $|Y| = m$ . We are required to prove that  $|X \cup Y| = n + m$ . If  $n = 0$  then  $X$  is the empty set and so  $X \cup Y = Y$  so that automatically  $|X \cup Y| = |Y| = m = 0 + m$ . Similarly the result is clear if  $m = 0$ . So let us consider the case when  $n$  and  $m$  are both positive.

Since  $|X| = n$  and  $|Y| = m$  there are bijections  $f: \mathbb{N}_n \rightarrow X$  and  $g: \mathbb{N}_m \rightarrow Y$ . We can now define a bijection  $h: \mathbb{N}_{n+m} \rightarrow X \cup Y$  as follows.

$$h(i) = \begin{cases} f(i) & \text{if } 1 \leq i \leq n, \\ g(i-n) & \text{if } n+1 \leq i \leq n+m. \end{cases}$$

This shows that  $|X \cup Y| = m + n$ . Notice that  $h$  is an injection because  $f$  and  $g$  are injections and  $X \cap Y = \emptyset$ ;  $h$  is a surjection because  $f$  and  $g$  are surjections.  $\square$

At the end of this proof some details have been omitted. The reader might like to consider how to supply these details justifying the last sentence of the proof. This sentence does indicate how the hypotheses in the theorem are brought into play in the proof; it is important when writing out a proof to indicate at which point(s) hypotheses are used.

**Corollary 10.2.2** *For a positive integer  $n$ , suppose that  $X_1, X_2, \dots, X_n$  is a collection of  $n$  pairwise disjoint finite sets (i.e.  $i \neq j \Rightarrow X_i \cap X_j = \emptyset$ ).*

*Then  $X_1 \cup X_2 \cup \dots \cup X_n = \bigcup_{i=1}^n X_i$  is a finite set and*

$$|X_1 \cup X_2 \cup \dots \cup X_n| = |X_1| + |X_2| + \dots + |X_n|.$$

*Proof* This is an easy induction proof using the theorem. The details are left as an exercise.  $\square$

The second basic counting principle is a consequence of the first.

**Theorem 10.2.3 (The multiplication principle)** Suppose that  $X$  and  $Y$  are finite sets such that  $|X| = n$  and  $|Y| = m$ . Then the Cartesian product  $X \times Y$  is a finite set and  $|X \times Y| = mn$ .

*Proof* If either of  $X$  and  $Y$  is empty then so is  $X \times Y$  and the result follows. So suppose that both sets are non-empty.

Let  $f: \mathbb{N}_n \rightarrow X$  be a bijection and write  $x_i = f(i)$ . Then

$$X = \{x_1, x_2, \dots, x_n\} = \{x_1\} \cup \{x_2\} \cup \dots \cup \{x_n\}$$

and so

$$X \times Y = (\{x_1\} \times Y) \cup (\{x_2\} \times Y) \cup \dots \cup (\{x_n\} \times Y).$$

But now, if  $g: \mathbb{N}_m \rightarrow Y$  is a bijection, then  $i \mapsto (x_k, g(i))$  gives a bijection  $\mathbb{N}_m \rightarrow \{x_k\} \times Y$  so that  $|\{x_k\} \times Y| = m$ . It follows from the previous corollary that  $|X \times Y| = mn$ .  $\square$

**Examples 10.2.4** (a) An element of  $\mathbb{N}_9 \times \mathbb{N}_9$  is an ordered pair of positive integers each no greater than 9. The number of such pairs is  $9 \times 9 = 81$  by the multiplication principle.

(b) What then is the number of ordered pairs of *distinct* positive integers no greater than 9? The function  $n \mapsto (n, n)$  provides a bijection  $\mathbb{N}_9 \rightarrow \{(n, n) \mid n \in \mathbb{N}_9\}$  showing that there are nine ordered pairs of equal integers. This set of ordered pairs is called the *diagonal set* in  $\mathbb{N}_9 \times \mathbb{N}_9$  and is denoted by  $\Delta(\mathbb{N}_9)$ . We are interested in the cardinality of the complement of the diagonal set,  $\mathbb{N}_9 \times \mathbb{N}_9 - \Delta(\mathbb{N}_9)$ . However,  $\mathbb{N}_9 \times \mathbb{N}_9$  is the disjoint union  $\Delta(\mathbb{N}_9) \cup (\mathbb{N}_9 \times \mathbb{N}_9 - \Delta(\mathbb{N}_9))$  and so, by the addition principle,

$$|\mathbb{N}_9 \times \mathbb{N}_9| = |\Delta(\mathbb{N}_9)| + |\mathbb{N}_9 \times \mathbb{N}_9 - \Delta(\mathbb{N}_9)|$$

so that the number of ordered pairs of distinct positive integers no greater than 9,  $|\mathbb{N}_9 \times \mathbb{N}_9 - \Delta(\mathbb{N}_9)|$ , is  $81 - 9 = 72$ .

There are of course other ways of obtaining the same result. For example, there are nine possibilities for the first coordinate and for each of these there are then eight possibilities for the second coordinate so that the number of ordered distinct pairs is  $9 \times 8 = 72$ .

From this result we can deduce that the number of *unordered* distinct pairs is  $72/2 = 36$  since each pair may be ordered in two ways.

### 10.3 The inclusion–exclusion principle

The addition principle in the last section leads on to an extremely powerful combinatorial technique known as the inclusion–exclusion principle. This is concerned with a situation where we have a finite number of finite sets of known cardinality (not in general disjoint) and we wish to know the number of elements which lie in their union. Let us consider the simplest example, two sets.

**Proposition 10.3.1** *Suppose that  $X$  and  $Y$  are (not necessarily disjoint) finite sets. Then  $X \cup Y$  is a finite set and*

$$|X \cup Y| = |X| + |Y| - |X \cap Y|.$$

*Proof* Recall from Proposition 6.2.4 that  $X \cup Y = (X - Y) \cup (Y - X) \cup (X \cap Y)$ , a union of disjoint sets, so that by the addition principle

$$|X \cup Y| = |X - Y| + |Y - X| + |X \cap Y|.$$

Similarly,  $X = (X - Y) \cup (X \cap Y)$  is a union of disjoint sets and so

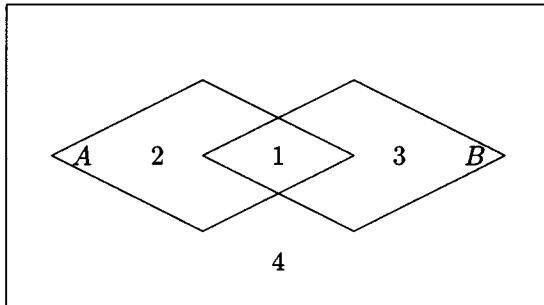
$$|X| = |X - Y| + |X \cap Y|.$$

Similarly  $|Y| = |Y - X| + |X \cap Y|$ . Thus

$$|X \cup Y| = (|X| - |X \cap Y|) + (|Y| - |X \cap Y|) + |X \cap Y| = |X| + |Y| - |X \cap Y|$$

as required.  $\square$

This result is really obvious from the following Venn diagram.



Recall that  $A \cup B$  consists of regions 1, 2 and 3;  $A$  consists of regions 1 and 2;  $B$  consists of 1 and 3; and  $A \cap B$  consists of region 1. If we count  $A$  and then count  $B$  giving  $|A| + |B|$  then we will have counted the elements in  $A \cap B$  twice and so we take account of this by subtracting  $|A \cap B|$ .

The inclusion–exclusion principle is the generalization of this result to an arbitrary finite collection of sets. Consider the case of three sets.

**Proposition 10.3.2** *Suppose that  $X$ ,  $Y$  and  $Z$  are finite sets. Then  $X \cup Y \cup Z$  is a finite set and*

$$|X \cup Y \cup Z| = |X| + |Y| + |Z| - |X \cap Y| - |X \cap Z| - |Y \cap Z| + |X \cap Y \cap Z|.$$

*Proof* We apply Proposition 10.3.1 to  $X \cup Y \cup Z = (X \cup Y) \cup Z$  and then apply it again to the sets which arise.

$$|(X \cup Y) \cup Z| = |X \cup Y| + |Z| - |(X \cup Y) \cap Z|; \quad (10.1)$$

$$|X \cup Y| = |X| + |Y| - |X \cap Y|; \quad (10.2)$$

$$\begin{aligned} |(X \cup Y) \cap Z| &= |(X \cap Z) \cup (Y \cap Z)| \\ &= |X \cap Z| + |Y \cap Z| - |(X \cap Z) \cap (Y \cap Z)| \\ &= |X \cap Z| + |Y \cap Z| - |X \cap Y \cap Z|. \end{aligned} \quad (10.3)$$

Substituting from equations (10.2) and (10.3) into equation (10.1) gives the result.  $\square$

The general inclusion–exclusion principle is given as Problems III, Question 4. This proposition illustrates the inductive step in the proof. A number of applications appear later in the book.

### Exercises

**10.1** Suppose that  $X$  is a non-empty finite set of cardinality  $n$ . Prove that a set  $Y$  has  $|Y| = n$  if and only if there exists a bijection  $X \rightarrow Y$ .

**10.2** Prove Corollary 10.2.2.

**10.3** Let  $X = \{a, b, c\}$  and  $Y = \{d, e\}$ . What is the cardinality of the Cartesian product  $X \times Y$ ? Write down an explicit bijection  $\mathbb{N}_n \rightarrow X \times Y$  where  $n = |X \times Y|$ .

**10.4** Each of a collection of 144 tiles is either triangular or square, either red or blue, and either wooden or plastic. Given that there are 68 wooden tiles, 69 red tiles, 75 triangular tiles, 36 red wooden tiles, 40 triangular wooden tiles, 38 red triangular tiles, and 23 red wooden triangular tiles, how many blue plastic square tiles are there?

# 11

## Properties of finite sets

There are a number of special techniques available for proving results which involve finite sets. It is clearly possible to determine whether or not a finite set contains an element with a certain property simply by trying each element in turn. More interestingly it is possible on occasion to use ‘counting arguments’ to demonstrate the existence of an element with a certain property without considering the properties of individual elements at all. We consider these techniques in this chapter.

### 11.1 The pigeonhole principle

We begin by recalling the key step used in the last chapter to prove that the cardinality of a finite set is a well-defined concept. We left for this chapter the proof of the following result.

**Lemma 10.1.4** *If there exists an injection  $\mathbb{N}_m \rightarrow \mathbb{N}_n$  then  $m \leq n$ .*

**Constructing a proof.** Since this is a statement involving natural numbers a proof by induction is suggested. There are two natural numbers  $m$  and  $n$  in the statement and so there is a choice: induction on  $m$  or induction on  $n$ ? The following proof is by induction on  $n$ . The task of constructing an alternative proof by induction on  $m$  is left to the reader (see Problems III, Question 8).

The predicate  $P(n)$  in this case is the following universal implication.

$$\forall m \in \mathbb{Z}^+ (\exists \text{ an injection } \mathbb{N}_m \rightarrow \mathbb{N}_n \Rightarrow m \leq n).$$

So the steps of the inductive proof are as follows.

**Base case:** It is not difficult to prove by contradiction that if there exists an injection  $\mathbb{N}_m \rightarrow \mathbb{N}_1$  then  $m = 1$ . (Note that  $m \leq 1 \Leftrightarrow m = 1$  for a positive integer  $m$ .)

Inductive step: For each  $k \geq 1$  we must prove an implication which may be summarized as follows.

| Given   | Goal  |
|---|---|
| $\forall m \in \mathbb{Z}^+ (\exists \text{ injection } N_m \rightarrow N_k \\ \Rightarrow m \leq k)$ | $\forall m \in \mathbb{Z}^+ (\exists \text{ injection } N_m \rightarrow N_{k+1} \\ \Rightarrow m \leq k + 1)$ |

The variable  $m$  is a ‘dummy’ variable in both the given statement and the goal. To avoid confusion it is better as usual to replace  $m$  by some other symbol, say  $m_1$ , in the goal statement so that it becomes  $\forall m_1 \in \mathbb{Z}^+ (\exists \text{ injection } N_{m_1} \rightarrow N_{k+1} \Rightarrow m_1 \leq k + 1)$ . The direct method of proving this is to suppose that  $f: N_{m_1} \rightarrow N_{k+1}$  is an injection and then to seek to deduce that  $m_1 \leq k + 1$ . This may be summarized as follows.

| Given   | Goal             |
|---|------------------|
| $\forall m \in \mathbb{Z}^+ (\exists \text{ injection } N_m \rightarrow N_k \\ \Rightarrow m \leq k)$ |                  |
| $m_1 \in \mathbb{Z}^+$<br>injection $f: N_{m_1} \rightarrow N_{k+1}$                                  | $m_1 \leq k + 1$ |

It is not possible to apply the given universal implication directly to the injection  $f$  since its codomain is  $N_{k+1}$ : to apply the given implication we need an injection into  $N_k$  and this may require a change of domain. The details of the proof follow [with additional comments in brackets].

*Proof* We prove Lemma 10.1.4 by induction on  $n$ .

Base case: This statement is trivial for  $n = 1$  since, if  $f: N_m \rightarrow N_1$ , we must have  $f(i) = 1$  for all  $i \in N_m$  and this is not an injection if  $m > 1$  since then, for example,  $f(1) = f(2) = 1$ . Thus if  $f$  is an injection then  $m = 1 = n$ .

[Notice that this is really a proof by contrapositive. What is actually proved is ‘ $m > 1 \Rightarrow f: N_m \rightarrow N_1$  is not an injection’ whose contrapositive is ‘ $f: N_m \rightarrow N_1$  is an injection  $\Rightarrow m \leq 1$ ’.]

Inductive step: Suppose, for some  $k \geq 1$ , that the result is true for  $n = k$ . We must deduce that it holds for  $n = k + 1$  so suppose that  $f: N_{m_1} \rightarrow N_{k+1}$  is an injection.

[The aim now is to construct an injection  $f_1: N_m \rightarrow N_k$  for some  $m$  to which we can apply the inductive hypothesis.]

We consider two cases.

(i) Suppose that  $f(i) < k + 1$  for all  $i \in N_m$ . Then we can restrict the

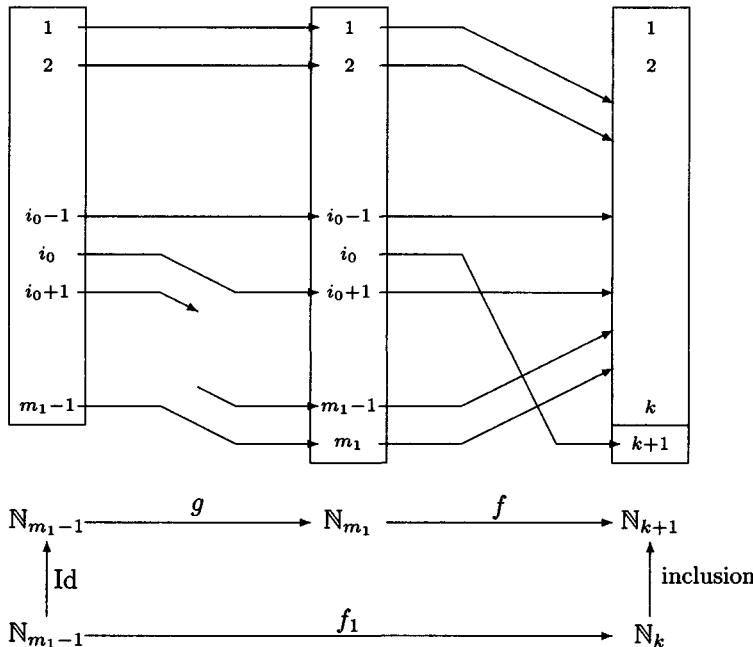
codomain and define a function  $f_1: \mathbb{N}_{m_1} \rightarrow \mathbb{N}_k$  by  $f_1(i) = f(i)$ . The function  $f_1$  is also an injection and so, by inductive hypothesis,  $m_1 \leq k$  and so certainly  $m_1 \leq k + 1$  as required.

[The function  $f_1$  is an injection since  $f_1(i_1) = f_1(i_2) \Rightarrow f(i_1) = f(i_2) \Rightarrow i_1 = i_2$  since  $f$  is an injection.]

(ii) On the other hand, suppose that  $k+1$  is a value of  $f$ , say  $f(i_0) = k+1$  where  $i_0 \in \mathbb{N}_{m_1}$ . In this case we can define an injection  $g: \mathbb{N}_{m_1-1} \rightarrow \mathbb{N}_{m_1}$  by

$$g(i) = \begin{cases} i & \text{for } i < i_0, \\ i+1 & \text{for } i \geq i_0, \end{cases}$$

and then define  $f_1 = f \circ g: \mathbb{N}_{m_1-1} \rightarrow \mathbb{N}_k$ .



[Notice that if  $i \neq i_0$  then  $f(i) \neq f(i_0) = k+1$  since  $f$  is an injection so that  $f(i) \in \mathbb{N}_k$ .]

Then this function  $f_1$  is also an injection (since it is the composite of two injections) and so, by inductive hypothesis,  $m_1 - 1 \leq k$ , i.e.  $m_1 \leq k + 1$ , as required.

Since one or other of these cases must be true for each function  $f$  this

completes the proof of the inductive step.

Conclusion: Hence, by induction on  $n$ , if  $f: \mathbb{N}_m \rightarrow \mathbb{N}_n$  is an injection then  $m \leq n$ .  $\square$

We can immediately extend this lemma to general finite sets.

**Corollary 11.1.1** *Suppose that  $X$  and  $Y$  are non-empty finite sets. If there exists an injection  $f: X \rightarrow Y$  then  $|X| \leq |Y|$ .*

So for example if a group of people are all sitting in a room (on separate chairs) then the number of people is no greater than the number of chairs.

**Constructing a proof.** The hypothesis tells us that there are bijections  $g_1: \mathbb{N}_m \rightarrow X$  and  $g_2: \mathbb{N}_n \rightarrow Y$  for some positive integers  $m$  and  $n$ . Then  $|X| = m$  and  $|Y| = n$  and we can obtain the result by applying the lemma to the composite

$$g_2^{-1} \circ f \circ g_1: \mathbb{N}_m \rightarrow X \rightarrow Y \rightarrow \mathbb{N}_n$$

since this is an injection (because it is a composite of injections).

This is a good example of how we can deduce a result about general finite sets from the special case about the model sets  $\mathbb{N}_n$  of the same cardinality.

*Proof* By hypothesis  $|X| = m$  and  $|Y| = n$  for positive integers  $m$  and  $n$ . Hence there are bijections  $g_1: \mathbb{N}_m \rightarrow X$  and  $g_2: \mathbb{N}_n \rightarrow Y$ . Then  $g_2^{-1} \circ f \circ g_1: \mathbb{N}_m \rightarrow \mathbb{N}_n$  is an injection. Hence, by Lemma 10.1.4,  $m \leq n$ , i.e.  $|X| \leq |Y|$ , as required.  $\square$

This result is usually stated in the following form. It was first formulated in the first half of the nineteenth century by Lejeune Dirichlet who called it the ‘drawer principle’.

**Theorem 11.1.2 (The pigeonhole principle)** *Suppose that  $f: X \rightarrow Y$  is a function between non-empty finite sets such that  $|X| > |Y|$ . Then  $f$  is not an injection, i.e. there exist distinct elements  $x_1$  and  $x_2 \in X$  such that  $f(x_1) = f(x_2)$ .*

*Proof* This is the contrapositive of Corollary 11.1.1 and so follows from that result.  $\square$

**Examples 11.1.3** (a) There are many simple applications of this result. For example, in any group of more than twelve people there must be two people with birthdays in the same month.

(b) For a slightly more elaborate application consider the following. Suppose that  $X$  is a set of at least two people. Then it can be shown that within this set there are two people with the same number of friends in the set. (Here we are assuming that if  $x_1$  is a friend of  $x_2$  then  $x_2$  is a friend of  $x_1$ .)

To see this let  $|X| = n$  and define a function  $f: X \rightarrow \{0, 1, 2, \dots, n-1\}$  by setting  $f(x)$  equal to the number of friends of  $x$  in  $X$ . Now notice that if  $n-1$  is a value of  $f$  then there exists  $x_0 \in X$  such that  $f(x_0) = n-1$ . This means that the person  $x_0$  is a friend of everyone else and so no one is friendless, i.e. 0 is not a value of  $f$ . Hence the numbers 0 and  $n-1$  (distinct since  $n \geq 2$ ) cannot both be values of  $f$ . Hence the image  $\text{Im } f$  has cardinality less than  $n$  and so by the pigeonhole principle  $f$  cannot be an injection, i.e. there exist distinct elements  $x_1$  and  $x_2$  such that  $f(x_1) = f(x_2)$  as claimed.

Some more mathematical applications will come later.

We can mimic the proof of Lemma 10.1.4 to prove the following more general result. The details are left to the reader (see Problems III, Question 9). (A similar result appears as Exercise 11.1.)

**Proposition 11.1.4** *Suppose that  $f: X \rightarrow \mathbb{N}_n$  is an injection. Then  $X$  is a finite set and  $|X| \leq n$ .*

This has the following consequence which is intuitively completely obvious. However, a *proof* of it has to be based on our definitions.

**Corollary 11.1.5** *Suppose that  $X \subseteq Y$  where  $Y$  is a finite set. Then  $X$  is also a finite set and  $|X| \leq |Y|$ .*

*Proof* Suppose that  $|Y| = n$  so that there is a bijection  $f: \mathbb{N}_n \rightarrow Y$ . Let  $i: X \rightarrow Y$  be the inclusion function  $i(x) = x$ . This is clearly an injection and so the composite  $f^{-1} \circ i: X \rightarrow Y \rightarrow \mathbb{N}_n$  is an injection. Hence, by Proposition 11.1.4,  $X$  is finite and  $|X| \leq n$ , as required.  $\square$

In the above we have concentrated on the property of injectivity. It is not difficult to deduce similar results about surjectivity such as the following.

**Theorem 11.1.6** Suppose that  $f: X \rightarrow Y$  is a function between non-empty finite sets such that  $|X| < |Y|$ . Then  $f$  is not a surjection, i.e. there exists an element of  $Y$  which is not a value of the function.

This can be proved by similar methods to the pigeonhole principle. Alternatively it can be deduced from the pigeonhole principle by observing that from a surjection  $X \rightarrow Y$  it is possible to construct an injection  $Y \rightarrow X$  (see for example Problems II, Question 19).

If two finite sets are known to have the same size, then it becomes easier to check whether a function between them is a bijection.

**Theorem 11.1.7** Suppose that  $X$  and  $Y$  are non-empty finite sets of the same cardinality. Then a function  $X \rightarrow Y$  is an injection if and only if it is a surjection. Thus to prove bijectivity it is only necessary to check one of injectivity and surjectivity.

The proof of this result is left to the reader (Problems III, Question 11).

## 11.2 Finite sets of real numbers

One advantage of working with finite sets is that when we are seeking an object with a certain property each of the elements can be tested in turn and this task will terminate. As an example of this consider the problem of seeking the greatest element in a set of real numbers. Of course there are sets of numbers without a greatest, for example the set of integers (see Problems I, Question 10). It will now be shown that any finite non-empty set of real numbers does have a greatest element.

**Definition 11.2.1** Let  $A$  be a set of real numbers, i.e.  $A \subseteq \mathbb{R}$ . Then  $b$  is a minimum element of  $A$  or  $b = \min A$  when

- (i)  $b$  is an element of  $A$ , i.e.  $b \in A$ ,
- (ii)  $b$  is less than or equal to every element of  $A$ , i.e.  $a \in A \Rightarrow b \leq a$ .

Similarly,  $c$  is a maximum element of  $A$  or  $c = \max A$  when

- (i)  $c$  is an element of  $A$ , i.e.  $c \in A$ ,
- (ii)  $c$  is greater than or equal to every element of  $A$ , i.e.  $a \in A \Rightarrow c \geq a$ .

Notice that if a set of real numbers,  $A$ , has a maximum then this maximum is unique so that the number  $\max A$  is well-defined. Writing out a formal proof of this uniqueness is left as an exercise (Exercise 11.2). Similarly the minimum is unique if it exists.

**Examples 11.2.2** (a) For  $A = \{-1, 17, 12, -78, 8\}$ ,  $\min A = -78$  and  $\max A = 17$ .

(b) The set  $\mathbb{Z}^+$  of positive integers has  $\min \mathbb{Z}^+ = 1$  but there is no maximum element.

(c) Any non-empty set of positive integers has a minimum element. This result is known as *the well-ordering principle* and is equivalent to the induction principle (see Exercise 11.6 and Problems III, Question 15).

(d) The set  $\mathbb{R}^+ = \{x \in \mathbb{R} \mid x > 0\}$  of positive real numbers has no minimum element and no maximum element. Notice that, given any element  $a \in \mathbb{R}^+$  there is a smaller one, for example  $a/2 \in \mathbb{R}^+$  (see Problems I, Question 11).

**Proposition 11.2.3** *Let  $A$  be a finite non-empty set of real numbers. Then  $A$  has a minimum element and a maximum element.*

When this result is applied it is always important to check that the set under consideration is non-empty. The empty set is finite but does not have a minimum or maximum element since it has no elements at all!

**Constructing a proof.** If we think why this is intuitively obvious we come up with the following ideas.

Suppose that  $|A| = n \in \mathbb{Z}^+$ . (This makes use of the fact that  $A$  is finite.) Then we can write  $A = \{a_1, a_2, \dots, a_n\}$ . To find the maximum element we try  $a_1$  and begin successively comparing it with the other elements  $a_2, a_3$  and so on. If we hit something greater we start using that in place of  $a_1$  and carry on working through the list repeating the replacement process whenever we reach a greater element. When we have completed the list, the element we have is the maximum element.

To write out this argument more formally we notice that the phrase ‘and so on’ suggests that this is really an induction argument even though the statement of the proposition does not explicitly refer to a positive integer. However, there is an implicit integer, namely the cardinality of the set  $A$ .

*Proof* We write out the proof for ‘maximum element’ leaving the proof for ‘minimum element’ as an exercise.

We prove by induction on  $n$  that, for all positive integers  $n$ , all non-empty finite sets of real numbers with cardinality  $n$  have a maximum element. This is clearly equivalent to the proposition because any non-empty finite set has cardinality  $n$  for some positive integer  $n$ , by definition of ‘finite’.

Base case: For  $n = 1$ , the single element of the set is the maximum (and the minimum) element!

Inductive step: Suppose that for some  $k \in \mathbb{Z}^+$  the result holds for  $n = k$  so that all sets of real numbers of cardinality  $k$  have a maximum element. Let  $A = \{a_1, a_2, \dots, a_k, a_{k+1}\}$  be a set of real numbers of cardinality  $k + 1$ . By inductive hypothesis, the set  $A' = \{a_1, a_2, \dots, a_k\}$  has a maximum element, say  $a'$ . If  $a' \geq a_{k+1}$ , then  $a'$  is the maximum element of  $A$ . If  $a' < a_{k+1}$ , then  $a_{k+1}$  is the maximum element of  $A$ . In either case  $A$  has a maximum element and so we have deduced the result for  $n = k + 1$ , completing the inductive step.

Hence by induction the result holds for all  $n \in \mathbb{Z}^+$  and the proposition is proved.  $\square$

### 11.3 Two applications of finiteness

#### *The greatest common divisor*

In Definition 2.2.1 we introduced the notion of one integer dividing another. Recall that a non-zero integer  $d$  divides  $a$  when there is an integer  $q$  such that  $a = dq$ . In this case we say that  $d$  is a *divisor* or a *factor* of  $a$ . Alternatively we say that  $a$  is a *multiple* of  $d$ .

Every non-zero integer divides 0 (since we can take  $q = 0$ ). However, if  $a$  is non-zero then, if  $a = dq$ ,  $q$  is also non-zero and so, since  $|q| \geq 1$ , we have  $|a| \geq |d|$ . Thus the set of divisors of  $a$ ,

$$D(a) = \{n \in \mathbb{Z} \mid n \text{ divides } a\},$$

is finite if  $a$  is non-zero with  $\min D(a) = -|a|$  and  $\max D(a) = |a|$ .

If now we are given two integers  $a$  and  $b$ , not both zero, then the set of *common divisors* of  $a$  and  $b$  is given by the intersection  $D(a) \cap D(b)$ . Notice that  $1 \in D(a) \cap D(b)$  so that  $D(a) \cap D(b)$  is a non-empty finite set and so must then have a maximum element: the *greatest common divisor*.

**Definition 11.3.1** *Let  $a$  and  $b$  be two integers, at least one of which*

is non-zero. Then the greatest common divisor or highest common factor of  $a$  and  $b$  is the unique positive integer  $d$  such that

- (i)  $d$  is a common divisor, i.e.  $d$  divides  $a$  and  $d$  divides  $b$ ,
- (ii)  $d$  is greater than every other common divisor:

$$c \text{ divides } a \text{ and } c \text{ divides } b \Rightarrow c \leq d.$$

We denote the greatest common divisor of  $a$  and  $b$  by  $(a, b)$ <sup>†</sup> or in case of possible ambiguity  $\gcd(a, b)$ .

**Definition 11.3.2** Two integers  $a$  and  $b$ , not both zero, are called coprime or relatively prime when  $(a, b) = 1$ , in other words their only common factors are 1 and  $-1$ .

**Example 11.3.3** The set of divisors of 30 is

$$D(30) = \{-30, -15, -10, -6, -5, -3, -2, -1, 1, 2, 3, 5, 6, 10, 15, 30\}$$

and the set of divisors of 72 is

$$\begin{aligned} D(72) &= \{-72, -36, -24, -18, -12, -9, -8, -6, -4, \\ &\quad -3, -2, -1, 1, 2, 3, 4, 6, 8, 9, 12, 18, 24, 36, 72\}. \end{aligned}$$

Hence the set of common divisors is

$$D(30, 72) = \{-6, -3, -2, -1, 1, 2, 3, 6\}.$$

Thus the greatest common divisor  $(30, 72) = 6$ .

### *Odd and even integers*

We can use these ideas to prove a result which was used in Problems I, Question 7. It is a special case of the *division theorem* which will be reached in Chapter 15.

<sup>†</sup> The reader should be aware (and beware) of the way in which some notation is used in a variety of ways. Here we are using  $(a, b)$  for the greatest common divisor of two integers, but we have previously used this notation to denote an element of the Cartesian product  $\mathbb{Z} \times \mathbb{Z}$  and to denote the open interval  $\{x \in \mathbb{R} \mid a < x < b\}$ . It is usually clear from the context which of these meanings is intended. The reader just has to get used to this sort of ambiguity.

**Proposition 11.3.4** *An integer  $a$  is odd if and only if  $a = 2q + 1$  for some integer  $q$ .*

Recall that we have defined ‘odd’ to mean not divisible by 2 (see Definition 2.2.3). Since this result is an ‘if and only if’ statement we have two things to prove: that the condition  $a = 2q + 1$  for some  $q$  is both necessary and sufficient for  $a$  to be odd. The proof of sufficiency is a generalization of the proof of Proposition 2.2.4 in Section 4.1 so let us get that out of the way.

*Proof of the sufficiency of  $a = 2q + 1$*  Suppose that  $a = 2q + 1$  where  $q \in \mathbb{Z}$ . To prove that  $a$  is odd, suppose for contradiction that  $a$  is even. Then  $a = 2p$  for some  $p \in \mathbb{Z}$ . But then  $2p = a = 2q + 1$  so that  $1 = 2(p - q)$  which implies  $1 \geq 2$  giving the required contradiction.

Hence  $2q + 1$  is not even and so is odd by definition.  $\square$

**Constructing a proof of the necessity of  $a = 2q + 1$ .** The difficulty of the proof of necessity is that we somehow have to get our hands on the element  $q$ . One common way of constructing an element in results of this type is to define it in terms of some set – in this case we define it as the maximum member of a finite set. If you think of how you would in practice find this number  $q$  you realize that you would keep doubling integers until you got as close to  $a$  as possible. When we formalize this into a mathematical argument it gives the proof. It is easier to work with positive integers first and then extend to negative ones.

*Proof of the necessity of  $a = 2q + 1$  (a)* Suppose that  $a$  is a positive odd integer. Set

$$A = \{ k \in \mathbb{Z} \mid k \geq 0 \text{ and } a \geq 2k \}.$$

This is a non-empty set since  $0 \in A$  and it is a finite set since  $k \in A \Rightarrow 2k \leq a \Rightarrow k \leq a$ . So we can define  $q = \max A$ .

To prove that  $a = 2q + 1$  notice first of all that, since  $q \in A$ ,  $a \geq 2q$  and, since  $q$  is odd,  $a \neq 2q$  and so  $a \geq 2q + 1$ . On the other hand, since  $q = \max A$ ,  $q + 1 \notin A$  so that  $a < 2(q + 1) = 2q + 2$ , i.e.  $a \leq 2q + 1$ . Thus we have proved that  $a \geq 2q + 1$  and  $a \leq 2q + 1$  from which it follows that  $a = 2q + 1$  as required.

(b) To complete the proof suppose that  $a$  is a negative odd integer. Then  $-a$  is a positive odd integer so that from what we have proved  $-a = 2q_1 + 1$ . But then  $a = -2q_1 - 1 = 2(-q_1 - 1) + 1 = 2q + 1$  as required if we put  $q = -q_1 - 1$ .  $\square$

**Exercises**

**11.1** Suppose that  $\mathbb{N}_n \rightarrow X$  is a surjection. Prove, by induction on  $n$ , that  $X$  is a finite set and that  $|X| \leq n$ .

[This is similar to Proposition 11.1.4.]

**11.2** Use Definition 11.2.1 to prove that if a set of real numbers has a maximum element then this element is unique.

[Hint: Suppose that  $c_1$  and  $c_2$  are two maximum elements of a set  $A$  and use the definition to prove that  $c_1 = c_2$ .]

**11.3** Find the greatest common divisor of 88 and 136.

**11.4** Prove that, if  $a$  and  $b$  are non-zero integers with  $\gcd(a, b) = d$ , then the integers  $a/d$  and  $b/d$  are coprime.

**11.5** Suppose that  $X$  and  $Y$  are finite sets such that  $X \subseteq Y$ . Prove that

- (i)  $X = Y \Leftrightarrow |X| = |Y|$ ,
- (ii)  $X \subset Y \Leftrightarrow |X| < |Y|$ .

**11.6** Prove by induction on  $n$  that if  $A$  is a set of positive integers without a least element then  $\mathbb{N}_n \subseteq \mathbb{Z}^+ - A$  for every  $n$  so that  $A$  is the empty set.

Deduce the well-ordering principle: every non-empty set of positive integers has a least element.

# 12

## Counting functions and subsets

Very many counting problems can be formulated in terms of counting the number of functions between two sets (possibly satisfying certain properties) or counting the number of subsets of a given set (possibly satisfying certain properties). In this chapter we give a brief introduction to these ideas. They naturally lead to the *binomial coefficients*, one of the most important families of numbers in all mathematics.

### 12.1 Counting sets of functions

Suppose that  $X$  and  $Y$  are finite sets. It is natural to ask how many different functions there are with domain  $X$  and codomain  $Y$ . For example, if  $X$  is a set of people and  $Y$  is a set of dishes on a menu, then each function  $X \rightarrow Y$  represents a choice of (one) dish for each person, so that the number of functions represents the number of possible orders by the set of people. If  $S$  is a set of students and  $T$  is a set of tutors then a function  $S \rightarrow T$  is an assignment of a tutor for each student.

There is another increased level of abstraction here. The idea of treating a set, for example the set of integers, as a single mathematical object is one stage of abstraction. A further stage is viewing a function from one set to another as a single mathematical object. But now we are forming a new set whose elements are the functions from one set to another and asking questions about this set.

Let us return to the problem we are considering to give an example.

**Example 12.1.1** Let  $X = \{a, b, c\}$  and  $Y = \{d, e\}$ . Then the eight functions  $X \rightarrow Y$  are listed in Example 8.1.3. In order to define a function  $f: X \rightarrow Y$ , we must list the elements  $f(a)$ ,  $f(b)$  and  $f(c)$ . Intuitively, since  $f(a)$  can be chosen in two ways ( $d$  and  $e$ ),  $f(b)$  can be

chosen in two ways and  $f(c)$  can be chosen in two ways,  $f$  can be chosen in  $2 \times 2 \times 2 = 8$  ways and so there are eight functions. This example illustrates the following general result.

**Proposition 12.1.2** *Suppose that  $X$  and  $Y$  are non-empty finite sets with  $|X| = m$  and  $|Y| = n$ . Then the number of functions  $X \rightarrow Y$  is given by  $n^m$ .*

**Constructing a proof.** We can generalize the above argument as follows. By the hypotheses in the proposition we can list the elements of the sets:  $X = \{x_1, x_2, \dots, x_m\}$  and  $Y = \{y_1, y_2, \dots, y_n\}$ . A function  $f: X \rightarrow Y$  is determined by listing the values  $f(x_1), f(x_2)$ , and so on up to  $f(x_m)$ . There are  $n$  possibilities for  $f(x_1)$ ,  $n$  possibilities for  $f(x_2)$ , and so on. Hence the number of functions is  $n \times n \times \dots \times n$  ( $m$  factors)  $= n^m$ , as required.

The phrase ‘and so on’ in this argument indicates that it is really an argument by induction – in this case an induction on  $m$ . This is a good example of a result with a convincing informal proof, as given above, for which the formal proof is rather complicated. The reader should endeavour to see how the simple idea in the informal proof underpins the formal proof. When writing out a formal proof it is convenient to have a name for the set of functions  $X \rightarrow Y$ .

**Definition 12.1.3** *Given sets  $X$  and  $Y$  we denote the set of functions from  $X$  to  $Y$  by  $\boxed{\text{Fun}(X, Y)}$ .*

*Proof of Proposition 12.1.2* The proof is by induction on the positive integer  $m$ .

Base case: For  $m = 1$  suppose that  $X = \{x_1\}$ . Then for any set  $Y$  there is a bijection  $\text{Fun}(X, Y) \rightarrow Y$  given by  $f \mapsto f(x_1)$ . Hence  $|\text{Fun}(X, Y)| = |Y| = n$  as required.

Inductive step: Suppose that, for some  $k \geq 1$ , the result holds for  $m = k$ . To deduce the result for  $m = k+1$  suppose that  $X_1$  and  $Y$  are non-empty finite sets with  $|X_1| = k+1$  and  $|Y| = n$ . Let  $X_1 = \{x_1, x_2, \dots, x_{k+1}\}$  and  $Y = \{y_1, y_2, \dots, y_n\}$ . Then there are  $n$  possibilities for  $f(x_{k+1})$  and

$$\text{Fun}(X_1, Y) = \bigcup_{i=1}^n \{ f: X_1 \rightarrow Y \mid f(x_{k+1}) = y_i \},$$

a disjoint union. We can determine the cardinality of each of the sets in

the union using the bijection

$$\{ f: X_1 \rightarrow Y \mid f(x_{k+1}) = y_i \} \rightarrow \text{Fun}(\{x_1, x_2, \dots, x_k\}, Y)$$

defined by  $f \mapsto f|(X_1 - \{x_{k+1}\})$ . [Notice that if we know that  $f(x_{k+1}) = y_i$  then  $f$  is determined by its values on  $x_1, \dots, x_k$ .] For then, by inductive hypothesis,

$$|\{ f: X_1 \rightarrow Y \mid f(x_{k+1}) = y_i \}| = |\text{Fun}(\{x_1, x_2, \dots, x_k\}, Y)| = n^k.$$

Hence, by the addition principle,

$$|\text{Fun}(X_1, Y)| = \sum_{i=1}^n n^k = n \times n^k = n^{k+1}$$

as required to deduce the result for  $m = k + 1$  and so completing the inductive step.

Conclusion: Hence, by induction on  $m$ , the result holds for all positive integers  $m$ .  $\square$

For certain applications we do not wish to know the total number of functions, but just the number which have some specific property. Quite commonly we want to know the number of injections.

**Proposition 12.1.4** *Suppose that  $X$  and  $Y$  are non-empty sets with  $|X| = m$  and  $|Y| = n$ . Then the number of injections  $X \rightarrow Y$  is given by  $n(n-1)\dots(n-m+1)$ .*

The number  $n(n-1)\dots(n-m+1)$  is usually called a *falling factorial* and is denoted by  $(n)_m$  in books on combinatorics. When  $m = n$  it is just the usual factorial denoted by  $n!$ . When  $m > n$  the falling factorial vanishes since one of the factors is 0: this shows that Proposition 12.1.4 is a generalization of the pigeonhole principle which asserts that in this case there are no injections. If  $m \leq n$ , then  $(n)_m = n!/(n-m)!$ .

The proof of this result is almost identical to that of Proposition 12.1.2 except that as we choose the values  $f(x_1), f(x_2)$  and so on we must avoid elements of the codomain  $Y$  which have already been selected. To write it out it is again useful to have a name for the set we are counting.

**Definition 12.1.5** *Given sets  $X$  and  $Y$  we denote the set of injections from  $X$  to  $Y$  by  $\boxed{\text{Inj}(X, Y)}$ .*

*Proof of Proposition 12.1.4* The proof is by induction on the positive integer  $m$ .

Base case: For  $m = 1$  suppose that  $X = \{x_1\}$ . Then for any set  $Y$  there is a bijection  $\text{Inj}(X, Y) = \text{Fun}(X, Y) \rightarrow Y$  given by  $f \mapsto f(x_1)$ . Hence  $|\text{Inj}(X, Y)| = |Y| = n$  as required.

Inductive step: Suppose that, for some  $k \geq 1$ , the result holds for  $m = k$ . To deduce the result for  $m = k + 1$  suppose that  $X_1$  is a set of  $k + 1$  elements. Let  $X_1 = \{x_1, x_2, \dots, x_{k+1}\}$  and  $Y = \{y_1, y_2, \dots, y_n\}$ . Then

$$\text{Inj}(X_1, Y) = \bigcup_{i=1}^n \{f \in \text{Inj}(X_1, Y) \mid f(x_{k+1}) = y_i\},$$

a disjoint union. We can determine the cardinality of each of the sets in the union using the bijection

$$\{f \in \text{Inj}(X_1, Y) \mid f(x_{k+1}) = y_i\} \rightarrow \text{Inj}(\{x_1, x_2, \dots, x_k\}, Y - \{y_i\})$$

defined by  $f \mapsto f|(X_1 - \{x_{k+1}\})$ . [Notice that if  $f$  is an injection and  $f(x_{k+1}) = y_i$  then none of the values  $f(x_1), \dots, f(x_k)$  can be  $y_i$ .] For then, by inductive hypothesis,

$$\begin{aligned} |\{f \in \text{Inj}(X_1, Y) \mid f(x_{k+1}) = y_i\}| &= |\text{Inj}(\{x_1, x_2, \dots, x_k\}, Y - \{y_i\})| \\ &= (n-1)_k \\ &= (n-1)(n-2)\dots(n-k). \end{aligned}$$

Hence, by the addition principle,

$$\begin{aligned} |\text{Inj}(X_1, Y)| &= \sum_{i=1}^n (n-1)(n-2)\dots(n-k) \\ &= n \times (n-1)\dots(n-k) = (n)_{k+1} \end{aligned}$$

as required to deduce the result for  $m = k + 1$  and so completing the inductive step.

Conclusion: Hence, by induction on  $m$ , the result holds for all positive integers  $m$ .  $\square$

In the case when  $m = n$ , an injection is necessarily a bijection (see Theorem 11.1.7). Bijections with the same domain and codomain are of particular interest.

**Definition 12.1.6** Given a set  $X$ , a bijection  $X \rightarrow X$  is called a

permutation of the set  $X$ .

**Corollary 12.1.7** *Given a finite non-empty set  $X$  of cardinality  $n$ , the number of permutations of  $X$  is given by the factorial  $n!$ .*

*Proof* This is just the special case of Proposition 12.1.4 when  $m = n$  for in this case an injection is automatically a bijection (by Theorem 11.1.7).  $\square$

## 12.2 Counting sets of subsets

Recall from Definition 6.3.1 that the *power set* of  $X$  is the set of subsets of  $X$ , i.e.

$$\mathcal{P}(X) = \{ A \mid A \subseteq X \}.$$

**Proposition 12.2.1** *Suppose that  $X$  is a set of cardinality  $n$ . Then the power set  $\mathcal{P}(X)$  is a finite set of cardinality  $2^n$ :*

$$|\mathcal{P}(X)| = 2^{|X|}.$$

**Constructing a proof.** A subset  $A \subseteq X$  is determined by deciding whether each of the  $n$  elements of  $X$  is in the subset: there are clearly two possibilities for each element  $x$ , namely  $x \in A$  and  $x \notin A$ . Hence the total number of possibilities for the subset  $A$  is given by  $2 \times 2 \times \dots \times 2 = 2^n$ .

This rather informal argument can be made precise by using the idea of the *characteristic function* of a subset which appeared in Problems II, Question 15.

**Definition 12.2.2** *For any set  $X$ , we can associate to each subset  $A$  of  $X$  ( $A \in \mathcal{P}(X)$ ) a characteristic function  $\chi_A: X \rightarrow \{0, 1\}$  defined by*

$$\chi_A(x) = \begin{cases} 0 & \text{if } x \notin A, \\ 1 & \text{if } x \in A. \end{cases}$$

**Lemma 12.2.3** *The function  $\mathcal{P}(X) \rightarrow \text{Fun}(X, \{0, 1\})$  given by  $A \mapsto \chi_A$  is a bijection.*

*Proof* The inverse is given by  $\chi \mapsto \{x \in X \mid \chi(x) = 1\}$  ( $= \overleftarrow{\chi}(\{1\})$ ).  $\square$

*Proof of Proposition 12.2.1* Using the result on counting functions in the last section,

$$|\mathcal{P}(X)| = |\text{Fun}(X, \{0, 1\})| = 2^{|X|},$$

as required.  $\square$

**Definition 12.2.4** Given a set  $X$  and a non-negative integer  $r$ , an  $r$ -subset of  $X$  is a subset  $A \subseteq X$  of cardinality  $r$ . We denote the set of  $r$ -subsets of  $X$  by  $\mathcal{P}_r(X)$ , i.e.

$$\mathcal{P}_r(X) = \{ A \subseteq X \mid |A| = r \}.$$

We define the binomial coefficient or binomial number  $\binom{n}{r}$  (read ' $n$  choose  $r$ ') to be the cardinality of the set  $\mathcal{P}_r(X)$  when  $|X| = n$ .

The reason for this name will be apparent when we come to the binomial theorem later in this chapter. Notice that given two finite sets  $X_1$  and  $X_2$  of the same cardinality, there is a bijection  $f: X_1 \rightarrow X_2$ . Such a bijection will induce a bijection  $\vec{f}: \mathcal{P}_r(X_1) \rightarrow \mathcal{P}_r(X_2)$  (see Definition 9.3.1) and so the binomial coefficients are well-defined.

**Example 12.2.5** Suppose that  $X = \{a, b, c, d\}$ ; then

$$\begin{aligned}\mathcal{P}_0(X) &= \{\emptyset\}, \\ \mathcal{P}_1(X) &= \{\{a\}, \{b\}, \{c\}, \{d\}\}, \\ \mathcal{P}_2(X) &= \{\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}\}, \\ \mathcal{P}_3(X) &= \{\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}\}, \\ \mathcal{P}_4(X) &= \{X\}, \\ \mathcal{P}_r(X) &= \emptyset \quad \text{for } r > 4.\end{aligned}$$

Thus  $\binom{4}{0} = 1$ ,  $\binom{4}{1} = 4$ ,  $\binom{4}{2} = 6$ ,  $\binom{4}{3} = 4$ ,  $\binom{4}{4} = 1$  and  $\binom{4}{r} = 0$  for  $r > 4$ .

This example illustrates the following general results.

**Proposition 12.2.6** For  $n$  and  $r$  non-negative integers

- (i)  $\binom{n}{r} = 0$  if  $r > n$ ,
- (ii)  $\binom{n}{0} = 1$ ,  $\binom{n}{1} = n$ ,  $\binom{n}{n} = 1$ ,

$$(iii) \quad \binom{n}{r} = \binom{n}{n-r} \text{ for } 0 \leq r \leq n.$$

*Proof* Suppose that  $X$  is a finite set with  $|X| = n$ .

(i)  $X$  has no subsets of cardinality greater than  $n$  and so  $\binom{n}{r} = 0$  for  $r > n$ .

(ii)  $\mathcal{P}_0(X) = \{\emptyset\}$  and so  $\binom{n}{0} = 1$ .  $\mathcal{P}_n(X) = \{X\}$  and so  $\binom{n}{n} = 1$ .

If  $X = \{x_1, x_2, \dots, x_n\}$  then  $\mathcal{P}_1(X) = \{\{x_1\}, \{x_2\}, \dots, \{x_n\}\}$  and so  $\binom{n}{1} = n$ .

(iii) There is a bijection  $\mathcal{P}_r(X) \rightarrow \mathcal{P}_{n-r}(X)$  given by mapping each subset to its complement  $A \mapsto A^c$ . Hence  $\binom{n}{r} = \binom{n}{n-r}$ .  $\square$

### Proposition 12.2.7

$$\sum_{i=0}^n \binom{n}{i} = 2^n.$$

*Proof* This is immediate from Definition 12.2.4 using Proposition 12.2.1 and Proposition 12.2.6(i).  $\square$

### Evaluating binomial coefficients

The method used in Example 12.2.5, calculating binomial coefficients by enumerating subsets, is extremely inefficient. There is an ancient inductive method for evaluating these numbers and the key to this is the following result.

### Proposition 12.2.8

For positive integers  $n$  and  $r$  such that  $1 \leq r \leq n$

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}.$$

*Proof* Let  $X$  be a set of cardinality  $n$  and choose  $x_1 \in X$ . Then we may define a bijection  $f: \mathcal{P}_r(X) \rightarrow \mathcal{P}_{r-1}(X - \{x_1\}) \cup \mathcal{P}_r(X - \{x_1\})$  as follows. Let  $A \in \mathcal{P}_r(X)$ . Then

$$f(A) = \begin{cases} A - \{x_1\} \in \mathcal{P}_{r-1}(X - \{x_1\}) & \text{if } x_1 \in A, \\ A \in \mathcal{P}_r(X - \{x_1\}) & \text{if } x_1 \notin A. \end{cases}$$

We can prove that this is a bijection by writing down the inverse

$$g: \mathcal{P}_{r-1}(X - \{x_1\}) \cup \mathcal{P}_r(X - \{x_1\}) \rightarrow \mathcal{P}_r(X)$$

as follows:

$$g(A) = \begin{cases} A \cup \{x_1\} & \text{if } A \in \mathcal{P}_{r-1}(X - \{x_1\}), \\ A & \text{if } A \in \mathcal{P}_r(X - \{x_1\}). \end{cases}$$

The result now follows by the addition principle (Theorem 10.2.1).  $\square$

**Remarks 12.2.9** This relation gives an inductive method for calculating binomial coefficients which appears to have been first discovered in China by the end of the eleventh century but is now known in the west as Pascal's triangle after the French mathematician Blaise Pascal (1623–1662) who connected the study of probability with the binomial coefficients and made use of this method of calculating them.

We may display the binomial coefficients in a triangle beginning as follows.

|  |  |  |   |   |    |    |    |    |
|--|--|--|---|---|----|----|----|----|
|  |  |  | 1 |   |    |    |    |    |
|  |  |  | 1 | 2 | 1  |    |    |    |
|  |  |  | 1 | 3 | 3  | 1  |    |    |
|  |  |  | 1 | 4 | 6  | 4  | 1  |    |
|  |  |  | 1 | 5 | 10 | 10 | 5  | 1  |
|  |  |  | 1 | 6 | 15 | 20 | 15 | 6  |
|  |  |  | 1 | 7 | 21 | 35 | 35 | 21 |

Here the numbers in the  $(n + 1)$ st row are the binomial coefficients  $\binom{n}{r}$  for  $r = 0, 1, 2, \dots, n$ . By Proposition 12.2.6 the border consists entirely of the number 1. By Proposition 12.2.8 each number is the sum of the two adjacent numbers in the row above. This is not a very convenient method of working out a single binomial coefficient but there is an explicit formula as follows.

**Theorem 12.2.10** *For non-negative integers  $n$  and  $r$  such that  $r \leq n$ ,*

$$\binom{n}{r} = \frac{n(n-1)\dots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}.$$

**Constructing a proof.** The most direct proof of this formula is by induction using Proposition 12.2.8. The details of this argument are given below. However, although this is probably the simplest argument to write out clearly it is not the most enlightening argument: it gives no

indication why the result is true. Mathematicians prefer proofs which give some understanding although this is not always achieved.<sup>†</sup>

In this case we can find an alternative and, I would argue, a better argument based on the definition of the binomial number as the cardinality of a set of subsets of a given cardinality and intuitive ideas about how we choose such a subset.

Suppose that  $X$  is a set of  $n$  elements. Each element  $A$  of  $\mathcal{P}_r(X)$  may be specified by listing the  $r$  distinct elements of  $X$  which belong to  $A$ . For  $r > 0$  this amounts to constructing an injection  $f: \mathbb{N}_r \rightarrow X$  so that  $\text{Im}(f) = A$ . Thus the function

$$\phi: \text{Inj}(\mathbb{N}_r, X) \rightarrow \mathcal{P}_r(X)$$

given by  $\phi(f) = \text{Im}(f)$  is a surjection. However, it is not a bijection since different injections may have the same image: we may list the elements of a subset  $A$  in a variety of orders. Each ordering of the elements of  $A$  corresponds to an element of the set  $\text{Inj}(\mathbb{N}_r, A)$  and, by Proposition 12.1.4, this set has cardinality  $r!$ . Hence, since the cardinality of  $\text{Inj}(\mathbb{N}_r, X)$  is  $n!/(n-r)!$  (also by Proposition 12.1.4), it follows that the cardinality of  $\mathcal{P}_r(X)$  is  $(n!/(n-r)!)/r! = n!/r!(n-r)!$  as claimed in the theorem.

*Proof* The proof is by induction on  $n$ .

Base case: If  $n = 0$  then the only possibility for  $r$  is  $r = 0$  and we know that  $\binom{0}{0} = 1$  from Proposition 12.2.6. When  $n = 0$  and  $r = 0$ ,  $n!/r!(n-r)! = 0!/0!0! = 1$  (recall that  $0! = 1$ ) and so the result is true for  $n = 0$ .

Inductive step: Suppose as inductive hypothesis that, for some  $k \geq 0$ , the result is true for  $n = k$ . To deduce it for  $n = k + 1$ , first observe that, for  $r = 0$  and  $r = k + 1$ ,  $\binom{k+1}{r} = 1$  (by Proposition 12.2.6) and  $(k+1)!/r!(k+1-r)! = (k+1)!/0!(k+1)! = 1$  and so the result holds in these cases. Now suppose that  $1 \leq r \leq k$ . Then, by Proposition 12.2.8,

$$\binom{k+1}{r} = \binom{k}{r-1} + \binom{k}{r}$$

<sup>†</sup> The eminent mathematician Pierre Deligne once wrote at the end of a very formal proof of his own ‘I would be grateful if anyone who has understood this proof would explain it to me’, ‘Je serais reconnaissant à toute personne ayant compris cette démonstration de me l’expliquer’ (see *Théorie des topos et cohomologie étale des schémas, Tome 3*, Lecture Notes in Mathematics, 305, Springer-Verlag, 1973, page 584). This is quoted in D. Alibert and M. Thomas, ‘Research into mathematical proof’ in David Tall (editor), *Advanced mathematical thinking*, Kluwer, 1991.

$$\begin{aligned}
 &= \frac{k!}{(r-1)!(k-r+1)!} + \frac{k!}{r!(k-r)!} \\
 &\quad \text{by inductive hypothesis} \\
 &= \frac{k!r}{r!(k-r+1)!} + \frac{k!(k-r+1)}{r!(k-r+1)!} \\
 &= \frac{k!(r+k-r+1)}{r!(k-r+1)!} = \frac{(k+1)!}{r!(k+1-r)!}
 \end{aligned}$$

as required to deduce the result for  $n = k + 1$ . This proves the inductive step.

Hence the result is true for all positive integers  $n$ .  $\square$

### 12.3 The binomial theorem

[Professor Moriarty] is . . . endowed by nature with a phenomenal mathematical faculty. At the age of twenty-one he wrote a treatise upon the binomial theorem, which has had a European vogue. On the strength of it he won the mathematical chair at one of our smaller universities, and had, to all appearances, a most brilliant career before him.

Arthur Conan Doyle, *Memoirs of Sherlock Holmes* (The final problem).

The name ‘binomial coefficients’ for the family of numbers considered in the previous section comes from its rôle in the binomial theorem. This theorem describes how to remove the brackets from an expression of the form  $(a + b)^n$ . A *binomial* is the sum of two terms and so  $a + b$  is a binomial.

**Theorem 12.3.1 (The binomial theorem)** *For all real numbers  $a$  and  $b$  and non-negative integers  $n$ ,*

$$\begin{aligned}
 (a + b)^n &= \sum_{i=0}^n \binom{n}{i} a^{n-i} b^i \\
 &= a^n + \dots + \binom{n}{i} a^{n-i} b^i + \dots + b^n.
 \end{aligned}$$

This theorem is only the simplest form of the binomial theorem. There are versions of the theorem when the exponent is a negative integer or a non-integer but in these cases the finite sum is replaced by an infinite series.<sup>†</sup>

<sup>†</sup> The more general theorem was discovered by the English mathematician Isaac Newton in the seventeenth century. Details can be found in books on advanced calculus or mathematical analysis – see for example the excellent book R. Courant and F. John, *Introduction to calculus and analysis*, Volume I, Wiley, 1965. No doubt Professor Moriarty’s treatise was on this more general theorem.

**Examples 12.3.2** Using the binomial theorem we can read off the following identities from Pascal's triangle:

$$\begin{aligned}(a+b)^2 &= a^2 + 2ab + b^2; \\ (a+b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3; \\ &\vdots \\ (a+b)^7 &= a^7 + 7a^6b + 21a^5b^2 + 35a^4b^3 + 35a^3b^4 + 21a^2b^5 + 7ab^6 + b^7.\end{aligned}$$

*Proof of the binomial theorem* Recall (from Definition 5.3.3) that we have adopted the convention that  $x^0 = 1$  for all non-zero real numbers  $x$  (including  $x = 0$ ).

The proof is by induction on  $n$ . Let  $P(n)$  be the statement we are required to prove.

Base case: For  $n = 0$  this statement reads  $(a+b)^0 = \binom{0}{0}a^0b^0$  which is true since  $(a+b)^0 = a^0 = b^0 = \binom{0}{0} = 1$ .

Inductive step: Suppose now as inductive hypothesis that, for some non-negative integer  $k$ ,  $P(n)$  holds for  $n = k$ . Then, for all  $a$  and  $b$ ,

$$\begin{aligned}(a+b)^{k+1} &= (a+b)^k(a+b) \\ &= \sum_{i=0}^k \binom{k}{i} a^{k-i}b^i(a+b) \quad \text{by inductive hypothesis} \\ &= \sum_{i=0}^k \binom{k}{i} a^{k-i+1}b^i + \sum_{i=0}^k \binom{k}{i} a^{k-i}b^{i+1} \\ &= \sum_{i=0}^k \binom{k}{i} a^{k-i+1}b^i + \sum_{j=1}^{k+1} \binom{k}{j-1} a^{k+1-j}b^j \quad (\text{writing } j \text{ for } i+1) \\ &= \sum_{i=0}^k \binom{k}{i} a^{k+1-i}b^i + \sum_{i=1}^{k+1} \binom{k}{i-1} a^{k+1-i}b^i \quad (\text{writing } i \text{ for } j) \\ &= a^{k+1}b^0 + \sum_{i=1}^k \binom{k}{i} a^{k+1-i}b^i + \sum_{i=1}^k \binom{k}{i-1} a^{k+1-i}b^i + a^0b^{k+1} \\ &\qquad\qquad\qquad \text{since } \binom{k}{0} = \binom{k}{k} = 1 \\ &= a^{k+1} + \sum_{i=1}^k \binom{k+1}{i} a^{k+1-i}b^i + b^{k+1} \quad \text{by Proposition 12.2.8} \\ &= \sum_{i=0}^{k+1} \binom{k+1}{i} a^{k+1-i}b^i\end{aligned}$$

as required to prove  $P(k + 1)$ .

Conclusion: Hence by induction  $P(n)$  holds for all positive integers  $n$ .  $\square$

This provides an alternative proof of Proposition 12.2.7.

### Corollary 12.3.3

$$\sum_{i=0}^n \binom{n}{i} = 2^n.$$

*Proof* Simply put  $a = b = 1$  in the binomial theorem.  $\square$

### Exercises

**12.1** Write down the next three rows of Pascal's triangle (page 151) and hence find the coefficient of  $a^3b^7$  in  $(a + b)^{10}$ .

**12.2** Find the coefficient of  $a^{98}b^2$  in  $(a + b)^{100}$ .

**12.3** Three people each select a different card from a single pack of 52 distinct cards. How many choices are possible (i) if we record who selected which card, and (ii) if we forget who selected which card?

**12.4** Three people each select a main dish from a menu of five items. How many choices are possible (i) if we record who selected which dish (as the waiter should), and (ii) if we ignore who selected which dish (as the chef could)?

**12.5** By equating the coefficients of  $x^n$  in  $(1 + x)^{2n} = (1 + x)^n(1 + x)^n$  prove that

$$\binom{2n}{n} = \sum_{i=0}^n \binom{n}{i}^2.$$

**12.6** Prove that the product of any  $n$  consecutive positive integers is divisible by  $n!$ .

**12.7** Use  $(1+i)^n$  to prove that

$$\sum_{0 \leq 2r \leq n} (-1)^r \binom{n}{2r} = \begin{cases} (-4)^k & \text{if } n = 4k, \\ (-4)^k & \text{if } n = 4k+1, \\ 0 & \text{if } n = 4k+2, \\ \frac{1}{2}(-4)^{k+1} & \text{if } n = 4k+3. \end{cases}$$

Obtain a similar result for  $\sum_{0 \leq 2r+1 \leq n} (-1)^r \binom{n}{2r+1}$ .

[This question uses complex numbers. The binomial theorem, Theorem 12.3.1, holds also for complex numbers  $a$  and  $b$ , with the same proof.]

# 13

## Number systems

Numbers originally arose for the purpose of counting. This makes use of the positive integers  $\mathbb{Z}^+ = \{1, 2, 3, \dots\}$ , often called the natural or counting numbers. In order to write down numbers we need some system of *numerals*, a system of symbols to represent numbers. At first the notation used was simply some sort of tally or repeated mark but in time more elaborate systems of numerals were devised leading eventually to our present Hindu–Arabic decimal system of numerals which makes use of ‘cypherization’ (i.e. different symbols 1, 2, 3, 4, 5, 6, 7, 8, 9 for different numbers) and a ‘place value system’ (e.g. in 121 the first ‘1’ denotes 100, the ‘2’ denotes  $2 \times 10 = 20$ , and the second ‘1’ denotes 1 so that the number represented is the sum  $100 + 20 + 1$ ).

An effective place value system requires a symbol to denote the absence of a positive digit and it is for this purpose that a symbol like ‘0’ first appeared (so that, for example, 102 and 12 represent different numbers). In due course zero was recognized as a number in its own right and then much later negative integers were introduced giving the complete system of integers  $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$  in which any two elements may be added or subtracted.

However, numbers are also used for purposes of measurement, originally of lengths but then of many other quantities. Before 1800 B.C. the Babylonians had developed a system of sexagesimal fractions rather like our modern decimals but using a base of 60 rather than 10. These were used until relatively modern times, particularly in astronomy, and indeed are still used in the measurement of angles by degrees, minutes and seconds and time by hours, minutes and seconds. For example if we state that the size of an angle is  $18^\circ 15' 53''$  we are saying that it is

$$18 + \frac{15}{60} + \frac{53}{(60)^2} = 18 \frac{953}{3600} \text{ degrees. Other ancient cultures had different}$$

techniques for dealing with fractions.

The classical Greek mathematicians took the view that the *numbers* were simply the positive integers.<sup>†</sup> Indeed, they took the view that the essence of all things is explainable in terms of the positive integers and their ratios. The Pythagoreans (founded by Pythagoras about 500 B.C.) assumed at first that any two lengths were *commensurable*, meaning that, given any two line segments, some unit of length could be found so that the length of each of the lines is a whole number times this unit. If then the first line is  $m$  units and the second is  $n$  units for positive integers  $m$  and  $n$ , this means that the ratio of the lengths of the two lines is  $m/n$  for some positive integers  $m$  and  $n$ , i.e. it is a rational number.

A simple consequence of Pythagoras's theorem ('the square on the hypotenuse of a right-angled triangle is the sum of the squares on the other two sides') is that the length of the diagonal of a unit square is a number whose square is 2. Some time around 400 B.C. the Pythagoreans discovered that the diagonal and side of a square are *incommensurable* or, equivalently, that there is no rational number whose square is 2. This undermined their view of the place of the positive integers in the world and was kept a secret; there is a story that the person who eventually revealed this secret of the Pythagoreans was executed by drowning. In due course an elegant way round this dilemma was discovered by Eudoxus whose theory of proportion<sup>‡</sup> appears in Book Five of the *Elements* of Euclid (about 300 B.C.). This involved a separation of the concepts of 'number' (which meant a positive integer) and 'magnitude' (which meant the length of a line). It was not until the nineteenth century that the two concepts were satisfactorily brought back together again through the work of mathematicians such as Richard Dedekind in Germany.

This book is not the place to go into further details about these later developments. However, it is important that the reader realizes that there are real difficulties in developing a satisfactory system of numbers and numerals.

In this chapter we review what is involved in the algebra of *rational numbers* and go on to prove the classic result that there is no rational square root of 2. A numeral system sufficient to measure all lengths is

<sup>†</sup> To be precise, the Greeks did not regard 1 as a number. Euclid (Book Seven of the *Elements*) refers to 1 as a *unit* and then defines a *number* to be a multitude composed of units.

<sup>‡</sup> See, for example, C.B. Boyer and U.C. Merzbach, *A history of mathematics*, Wiley, Second edition 1989.

provided by *infinite decimals* and we conclude the chapter by discussing this system and indicating some of its subtleties.

Further extensions of the number system are possible. The reader may well be familiar with one possibility: the system of *complex numbers* which first arose in sixteenth century Italy in the solution of cubic and quartic equations (but was not explained properly until the nineteenth century). Complex numbers are not described in this book although a few problems make use of them. The reader will meet them in the course of the study of algebra or advanced calculus.

### 13.1 The rational numbers

The rational number system provides a set of numbers which can be added and multiplied and which satisfy all the basic algebraic properties of numbers as listed in Properties 2.3.1. The difficulty with the integers is that we cannot always do division. For integers  $b \neq 0$  and  $a$  we can only find an integer solution of the equation  $bx = a$  if  $b$  divides  $a$ . However, there is always a unique rational number which is a solution. Furthermore the set of rational numbers provide just enough numbers for this purpose: given a rational number  $q$ , there are integers  $b \neq 0$  and  $a$  such that  $bq = a$ .

A satisfactory definition of rational numbers involves ideas which will be discussed later in this book – such a definition will be indicated in Chapter 22 (see Example 22.3.5). For the moment the reader is asked to accept that a system of rational numbers with the properties described above exists. We now consider how these properties lead to the description of the numbers and their properties in terms of fractions. While doing this it is convenient to make a distinction between a rational number and a fraction: each rational number is representable by a fraction but different fractions may represent the same rational number.

**Definition 13.1.1** We will use the word **fraction** to denote an expression of the form  $a/b$  where  $a$  and  $b$  are integers and  $b$  is non-zero;  $a$  is the **numerator** and  $b$  the **denominator** of the fraction.

Given integers  $a$  and (non-zero)  $b$  we say that the fraction  $a/b$  represents the **rational number**  $q$  which has the property that  $bq = a$ .

As a temporary notation we indicate that the rational number  $q$  is

represented by the fraction  $a/b$  by writing

$$q = \langle a/b \rangle.$$

The equation  $1x = a$ , for  $a$  an integer, has an integer solution, namely  $a$ , and so the integers can be thought of as a subset of the rational numbers: an integer  $a$  can be represented as a fraction by  $a = \langle a/1 \rangle$ .

**Proposition 13.1.2** *Two fractions  $a_1/b_1$  and  $a_2/b_2$  represent the same rational number  $q$ ,  $\langle a_1/b_1 \rangle = \langle a_2/b_2 \rangle = q$ , if and only if  $a_1b_2 = a_2b_1$ .*

*Proof* The fraction  $a_1/b_1$  represents the rational number  $q_1$  such that  $b_1q_1 = a_1$ , and the fraction  $a_2/b_2$  represents  $q_2$  such that  $b_2q_2 = a_2$ . Using the basic properties of multiplication (commutativity, associativity and cancelling) it follows that

$$\begin{aligned} b_1q_1 = a_1 &\Leftrightarrow b_1b_2q_1 = a_1b_2, \\ b_2q_2 = a_2 &\Leftrightarrow b_1b_2q_2 = a_2b_1. \end{aligned}$$

Hence

$$q_1 = q_2 \Leftrightarrow b_1b_2q_1 = b_1b_2q_2 \Leftrightarrow a_1b_2 = a_2b_1.$$

□

Thus for example  $\langle 2/3 \rangle = \langle 8/12 \rangle = \langle (-14)/(-21) \rangle$ .

By multiplying numerator and denominator by  $-1$  if necessary, we can represent any rational number by a fraction with positive denominator. Then, by dividing the numerator and denominator by their greatest common divisor, we see that every rational number may be written in the form  $q = \langle a/b \rangle$ , where  $b > 0$  and the greatest common divisor  $(a, b) = 1$  (see Exercise 11.4).

**Definition 13.1.3** *The fraction  $a/b$  is in lowest terms when  $b$  is positive and  $a$  and  $b$  are coprime.*

Thus every rational number may be represented by a fraction in lowest terms.

To see how to describe the sum of two rational numbers in terms of fractions suppose that  $q_1 = \langle a/b \rangle$  and  $q_2 = \langle c/d \rangle$  so that  $bq_1 = a$  and  $dq_2 = c$  (where  $a, b, c$  and  $d$  are integers with  $b$  and  $d$  non-zero). Then  $bdq_1 = ad$  and  $bdq_2 = bc$  so that (using distributivity)  $bd(q_1 + q_2) = ad + bc$  which means that  $q_1 + q_2$  is represented by  $(ad+bc)/bd$ . Similarly, since

$bdq_1q_2 = ac$ ,  $q_1q_2$  is represented by  $ac/bd$ . This discussion motivates the following definition.

**Definition 13.1.4** We define  $\boxed{\text{sum}}$  and  $\boxed{\text{product}}$  of two fractions by

$$\begin{aligned}\frac{a}{b} + \frac{c}{d} &= \frac{ad + bc}{bd}, \\ \frac{a}{b} \times \frac{c}{d} &= \frac{ac}{bd}.\end{aligned}$$

Now it is necessary to check that these definitions lead to *well-defined* definitions of addition and multiplication of rational numbers. This procedure is always necessary when we have different expressions for the same object and then make a definition using these expressions. In this case we need to check that if different fractions are used to represent the rational numbers then adding or multiplying the fractions leads to fractions representing the same rational number.

**Proposition 13.1.5** *Addition and multiplication of rational numbers are well-defined by the above formulae.*

**Constructing a proof.** We deal simply with addition. The proof for multiplication is similar and easier. Let us look at what we have to do.

| Given   | Goal  |
|---|---|
| $\langle a_1/b_1 \rangle = \langle a_2/b_2 \rangle$ |   |
| $\langle c_1/d_1 \rangle = \langle c_2/d_2 \rangle$ | $\langle a_1/b_1 + c_1/d_1 \rangle = \langle a_2/b_2 + c_2/d_2 \rangle$ |

When we spell out what these various statements mean, this leads to the following.

| Given             | Goal                        |
|-------------------|-----------------------------|
| $a_1b_2 = a_2b_1$ | $(a_1d_1 + b_1c_1)b_2d_2$   |
| $c_1d_2 = c_2d_1$ | $= (a_2d_2 + b_2c_2)b_1d_1$ |

But now we can obtain the goal from the given statements by some simple algebraic manipulation as follows:

$$\begin{aligned}(a_1d_1 + b_1c_1)b_2d_2 &= a_1b_2d_1d_2 + b_1b_2c_1d_2 \\ &= a_2b_1d_1d_2 + b_1b_2c_2d_1 \quad \text{by the given statements} \\ &= (a_2d_2 + b_2c_2)b_1d_1.\end{aligned}$$

*Proof* Suppose that  $\langle a_1/b_1 \rangle = \langle a_2/b_2 \rangle$  and  $\langle c_1/d_1 \rangle = \langle c_2/d_2 \rangle$  so that  $a_1b_2 = a_2b_1$  and  $c_1d_2 = c_2d_1$ . Then  $\langle a_1/b_1 + c_1/d_1 \rangle = \langle (a_1d_1 + b_1c_1)/b_1d_1 \rangle$  is the same as  $\langle a_2/b_2 + c_2/d_2 \rangle = \langle (a_2d_2 + b_2c_2)/b_2d_2 \rangle$  since  $(a_1d_1 + b_1c_1)b_2d_2 = (a_2d_2 + b_2c_2)b_1d_1$  (using the manipulation shown above). This shows that addition is well-defined.

The proof that multiplication is well-defined is similar.  $\square$

Notice that if  $q = \langle a/b \rangle$  then  $bq = \langle b/1 \times a/b \rangle = \langle ab/b \rangle = \langle a/1 \rangle = a$  and so this definition of multiplication is consistent with Definition 13.1.1.

From now on we abandon the notation  $\langle a/b \rangle$  and simply write  $a/b$  for the rational number represented by the fraction  $a/b$  as usual. Thus  $a_1/b_1 = a_2/b_2$  if and only if  $a_1b_2 = a_2b_1$ . The set of rational numbers is denoted by the symbol  $\mathbb{Q}$ .

The above discussion has described how addition and multiplication are defined in  $\mathbb{Q}$ . But of course subtraction of any element and division by non-zero elements are always possible in  $\mathbb{Q}$  since

$$\frac{a}{b} - \frac{c}{d} = \frac{a}{b} + \frac{-c}{d}$$

and, for  $c \neq 0$ ,

$$\frac{a}{b}/\frac{c}{d} = \frac{a}{b} \times \frac{d}{c}.$$

### 13.2 The irrationality of $\sqrt{2}$

The previous section has described the rational numbers and their familiar arithmetic. It was suggested above that a good number system should enable us to represent the length of any line. Pythagoras's theorem tells us that the length of the diagonal of a unit square is a number whose square is 2, usually denoted by  $\sqrt{2}$ . The next result tells us that there is no such rational number.

**Theorem 13.2.1** *There does not exist a rational number whose square is 2.*

**Constructing a proof.** Since every non-zero rational number can be written in the form  $a/b$  where  $a$  and  $b$  are non-zero integers we can write the result we are trying to prove as follows:

$$\exists a, b \in \mathbb{Z} - \{0\}, 2 = (a/b)^2.$$

Whenever a result takes the form of stating that something does not happen it is very difficult to envisage a direct proof and this suggests

very strongly that we need a proof by contradiction. Remember that this requires us to suppose that our goal is false and then to seek a contradiction. The negation of the goal is

$$\exists a, b \in \mathbb{Z} - \{0\}, 2 = (a/b)^2.$$

We use this existential statement by selecting elements  $a$  and  $b$  so that the predicate is true. This gives the following strategy.

| Given  | Goal          |
|--|---------------|
| $a, b \in \mathbb{Z} - \{0\}$<br>$2 = (a/b)^2$ | contradiction |

We can rewrite this as follows.

| Given   | Goal          |
|---|---------------|
| $a, b \in \mathbb{Z} - \{0\}$<br>$a^2 = 2b^2$ | contradiction |

Thus we see that the theorem we are trying to prove is really one about the integer solutions to an equation:

The only integer solution to the equation  $m^2 = 2n^2$  is the trivial one  $m = 0$ ,  $n = 0$ .

The one method we have used for proving that an equation has no integer solutions (see for example Proposition 4.1.1) is to use the idea of divisibility. So let us try that here. Given that  $a^2 = 2b^2$  then we know that  $a^2$  must be an even number. What does this tell us about  $a$ ?

We know that  $a$  is either even or odd and it follows by a proof by contradiction that in fact  $a$  must be even (see Problems I, Question 7).

Returning to our argument, we have seen that  $a^2 = 2b^2$  implies that  $a$  is even. Hence we can write  $a = 2a_1$ . Substituting this into  $a^2 = 2b^2$  we get  $4a_1^2 = 2b^2$  so that  $2a_1^2 = b^2$ . Hence  $b^2$  is even and so by the same argument  $b$  is even. So we have concluded that, if  $a^2 = 2b^2$ , then the numbers  $a$  and  $b$  must both be even.

So what? This in itself is no contradiction but it does suggest how we might get one, for if  $a$  and  $b$  are both even this means that the fraction  $a/b$  which we began with is not written in its lowest terms since  $a$  and  $b$  have the common factor 2. But we know that every rational number can be represented by a fraction in lowest terms. So had we thought to start by insisting that  $a/b$  be the lowest terms fraction representing a square root of 2 then we would now have obtained a contradiction. So we go back and place the requirement that  $(a, b) = 1$  on  $a$  and  $b$ .

This gives the following effective strategy for proving the theorem by contradiction.

| Given                         | Goal                                     |
|-------------------------------|--|
| $a, b \in \mathbb{Z} - \{0\}$ |  |
| $(a, b) = 1$                  | contradiction                            |
| $a^2 = 2b^2$                  | [i.e. $(a, b) = 1$ and $(a, b) \neq 1$ ] |

*Proof* Suppose for contradiction that there is a rational number  $q$  such that  $q^2 = 2$ . Write  $q$  as a fraction in lowest terms,  $q = a/b$ , so that  $a$  and  $b$  are integers such that  $(a, b) = 1$ . Now  $q^2 = 2 \Rightarrow a^2/b^2 = 2 \Rightarrow a^2 = 2b^2 \Rightarrow a^2$  is even  $\Rightarrow a$  is even  $\Rightarrow a = 2a_1$  for some  $a_1 \in \mathbb{Z}$ . But then  $a^2 = 2b^2 \Rightarrow 4a_1^2 = 2b^2 \Rightarrow 2a_1^2 = b^2 \Rightarrow b^2$  is even  $\Rightarrow b$  is even. Hence 2 divides  $a$  and 2 divides  $b$  implying that  $(a, b) \neq 1$ . But this contradicts the choice of  $a$  and  $b$  and so provides the required contradiction.

Hence there does not exist a rational number whose square is 2.  $\square$

### 13.3 Real numbers and infinite decimals

The result in the previous section demonstrates that the rational number system is not adequate to represent all lengths in terms of some standard unit length since the square of the length of the diagonal of a unit square is 2 and there is no rational number whose square is 2. The *real numbers* provide a further extension of the number system which is adequate to represent the lengths of all line segments. A real number which is not a rational number is said to be *irrational*. So Theorem 13.2.1 can be reformulated as the statement ‘ $\sqrt{2}$  is irrational’.

The definition of a real number is beyond the scope of this book since it requires a full treatment of the theory of limits of sequences referred to briefly in Chapter 8. This material can be found in any book on infinite sequences and series or mathematical analysis.†

The notation used to describe real numbers is that of *infinite decimals*. Thus every non-negative real number may be represented by an infinite decimal

$$a_0.a_1a_2\dots a_i\dots, \quad \text{where } a_i \in \mathbb{Z}^{\geq 0} \text{ and } 0 \leq a_i \leq 9 \text{ for } i \geq 1$$

and every infinite decimal represents a real number. The aim in this sec-

† See for example R. Haggerty, *Fundamentals of mathematical analysis*, Addison-Wesley, Second edition 1993.

tion is describe how the real number represented by an infinite decimal may be specified.

First of all, notice that a finite decimal  $a_0.a_1a_2\dots a_n$  (i.e. an infinite decimal with  $a_i = 0$  for  $i > n$ ) is equal to a rational number:

$$\begin{aligned} a_0.a_1a_2\dots a_n \\ = a_0 + a_1/10 + a_2/100 + \dots + a_n/(10)^n \\ = (a_0(10)^n + a_1(10)^{n-1} + a_2(10)^{n-2} + \dots + a_n)/(10)^n. \end{aligned}$$

So the set of finite decimals is a subset of the set of rational numbers.

The intuitive idea of an infinite decimal is that as we take more and more decimal places we get finite decimals which provide a better and better approximation to the number. Before we give a precise definition let us consider an example.

**Example 13.3.1** We have already proved that there is no rational number whose square is 2. Let us consider how to find an infinite decimal representing the positive real number whose square is 2. We construct the integers  $a_0, a_1, \dots$  making up the decimal by an inductive process. The key observation which enables us to do this is that, for non-negative real numbers  $a$  and  $b$ ,  $a < b$  if and only if  $a^2 < b^2$  (see Exercise 4.6).

Step 0: To find the integer part we find the largest integer whose square is less than 2. Since  $1^2 = 1 < 2 < 2^2 = 4$  we know that  $1 < \sqrt{2} < 2$  and so we put  $a_0 = 1$ .

Step 1: To find the first entry after the decimal point we square each of the numbers  $1.0, 1.1 = 11/10, 1.2 = 12/10, \dots, 1.9 = 19/10$  in turn and find the largest of them whose square is less than 2. (Notice that we cannot get equality since 2 is not a rational square). Since

$$(14/10)^2 = 196/100 < 2 < (15/10)^2 = 225/100$$

we know that  $1.4 < \sqrt{2} < 1.5$  and so put  $a_1 = 4$ .

Step 2: For the second decimal place, since

$$(141/100)^2 = 19881/10000 < 2 < (142/100)^2 = 20164/10000$$

so that  $1.41 < \sqrt{2} < 1.42$  we put  $a_2 = 1$ .

This process continues indefinitely giving an infinite decimal whose first few decimal places are 1.41421356237309504880 – which means that

$$1.41421356237309504880 < \sqrt{2} < 1.41421356237309504881.$$

However, there is no apparent pattern to the numbers which emerge. By Step  $n$  we have found a finite decimal  $a_0.a_1a_2\dots a_n$  such that

$$(a_0.a_1a_2\dots a_n)^2 < 2 < (a_0.a_1a_2\dots a_n + 1/(10)^n)^2$$

which means that

$$a_0.a_1a_2\dots a_n < \sqrt{2} < a_0.a_1a_2\dots a_n + 1/(10)^n.$$

For example, saying that the first seven decimal places of  $\sqrt{2}$  are given by 1.4142135 means that

$$1.4142135 < \sqrt{2} < 1.4142136$$

and, if we do the arithmetic, we find that it is indeed the case that

$$1.4142135^2 = 1.99999982358225 < 2 < 1.4142136^2 = 2.00000010642496.$$

This example illustrates the formal definition.

**Definition 13.3.2** *Given an infinite decimal  $a_0.a_1a_2\dots a_i\dots$ , we say that it represents a real number  $a$ , written  $a = a_0.a_1a_2\dots a_i\dots$ , when*

$a_0.a_1a_2\dots a_n \leq a \leq a_0.a_1a_2\dots a_n + 1/(10)^n$  for each  $n \in \mathbb{Z}^+$ .

Notice that we use ' $\leq$ ' rather than ' $<$ ' in this definition. This is necessary so that all rational numbers may be represented by an infinite decimal (possibly with  $a_i = 0$  for  $i > n$ , some  $n$ ).

The statement that every infinite decimal represents *some* real number is one version of the *completeness* or *continuity axiom* for the real numbers.

To see that an infinite decimal is equal to a *unique* real number we observe that the inequalities in the definition uniquely specify the number  $a$ . Given two distinct real numbers  $a$  and  $a'$  we have  $a - a' \neq 0$ . Now we choose an integer  $n$  so that  $1/(10)^n < |a - a'|$  by taking  $n > 1/|a - a'|$  (for then  $1/(10)^n < 1/n < |a - a'|$ ). [The fact that for each real number there is a greater integer is sometimes called the *Archimedean axiom* for the real numbers.] This means that infinite decimals representing  $a$  and  $a'$  will certainly differ by the  $n$ th decimal place.

**Example 13.3.3** Let us consider which real number is represented by the infinite decimal 0.999...99... each of whose entries after the decimal point is 9. An infinite decimal like this is called a *recurring* or repeating decimal, and we use the notation  $0.\overline{9}$  to represent it with a

dot over the integer which recurs. Recurring decimals may have a block of digits recurring, for example  $7.320008140081400814\dots$ . In this case we indicate the repeating block by placing a dot over the first and last digits in the block, viz.  $7.\overline{32000814}.$ <sup>†</sup>

The finite decimals  $0.99\dots 9$  are certainly less than 1. Indeed,

$$0.\underbrace{99\dots 9}_n = \underbrace{99\dots 9}_n / (10)^n = \frac{(10)^n - 1}{(10)^n} = 1 - \frac{1}{(10)^n}.$$

Intuition may suggest that this means that the infinite decimal  $0.\dot{9}$  is also less than 1 and this is a common misconception. I have known people become quite belligerent in their defence of this view. This is not an effective way of defending any point of view but emotion is particularly inappropriate in a mathematical argument! As always, if in doubt apply Humpty Dumpty's principle and examine what the words have been declared to mean. So we return to the definition. Let  $a = 0.\dot{9}$ . From the definition,

$$\begin{aligned} a = 0.\dot{9} &\Leftrightarrow 1 - 1/(10)^n \leq a \leq 1 \\ &\Leftrightarrow 0 \leq 1 - a \leq 1/(10)^n \end{aligned} \tag{13.1}$$

for all positive integers  $n$ .

Thus  $1 - a > 0$  or  $1 - a = 0$ . We eliminate the first of these possibilities by contradiction. If  $x$  is a positive number then as above we can find a positive integer  $k$  such that  $1/(10)^k < x$  by taking  $k > 1/x$ . Applying this to  $x = 1 - a$  demonstrates that if  $1 - a > 0$  then  $1 - a$  does not satisfy the condition (13.1) for  $n = k$ , giving a contradiction. The conclusion must be that  $1 - a$  is not positive and so  $1 - a = 0$  or in other words  $a = 1$ . This proves that

$$0.\dot{9} = 1.$$

If this result still worries you (and after all why not, since it is based on Definition 13.3.2 which is rather sophisticated?) then you may feel happier when you remember the most familiar recurring decimal,  $1/3 = 0.\dot{3}$ , which you probably are comfortable with. Multiplying by 3 gives  $1 = 0.\dot{9}$ .

<sup>†</sup> Alternatively, we place a dot over each digit which recurs,  $7.320\dot{0}\dot{8}\dot{1}\dot{4}$  or a line over the repeating block,  $7.\overline{32000814}$ .

*Recurring decimals and rationals*

The arithmetic of infinite decimals is quite complicated.<sup>†</sup> However, multiplication by 10 is easy for

$$\begin{aligned} 10 \times a_0.a_1a_2\dots a_n &= 10(a_0 + a_1/10 + a_2/100 + \dots + a_n/(10)^n) \\ &= 10a_0 + a_1 + a_2/10 + \dots + a_n/(10)^{n-1} \\ &= (10a_0 + a_1).a_2\dots a_n. \end{aligned}$$

So multiplication of a finite decimal by 10 corresponds to moving the decimal point one place to the right. This means that the same is true of infinite decimals since

$$\begin{aligned} a_0.a_1a_2\dots a_n &\leq a \leq a_0.a_1a_2\dots a_n + 1/(10)^n \\ \Leftrightarrow (10a_0 + a_1).a_2\dots a_n &\leq 10a \leq (10a_0 + a_1).a_2\dots a_n + 1/(10)^{n-1}. \end{aligned}$$

This observation enables us to prove the following result.

**Theorem 13.3.4** *A recurring infinite decimal is equal to a rational number.*

The idea of the proof of this result is very simple but writing out a general proof leads to great notational complexity. The proof consists of an algorithm (or procedure) for finding the rational number. If the length of the recurring block is  $k$  we multiply through by  $10^k$  and subtract the number we started with. The result is a finite decimal and so a rational number. This enables us to write the recurring decimal as a rational number. Rather than writing out a general proof the method is illustrated by giving an example.

**Example 13.3.5** To find the rational number  $12.79317317317\dots = 12.79\dot{3}\dot{1}\dot{7}$ .

*Solution* Let  $x = 12.79\dot{3}\dot{1}\dot{7}$ . Since the length of the recurring part is 3 we multiply through by  $10^3$  giving  $1000x = 12793.\dot{1}\dot{7}\dot{3}$  (since the decimal point moves three places to the right)  $= 12793.17\dot{3}\dot{1}\dot{7}$ . From the definition of the value of an infinite decimal we can split off an initial finite decimal:

$$x = 12.79 + 0.00\dot{3}\dot{1}\dot{7},$$

<sup>†</sup> An excellent discussion of the arithmetic of infinite decimals, a topic bypassed in most treatments of infinite decimals, may be found in A. Gardiner, *Infinite processes, background to analysis*, Springer-Verlag, 1982.

$$1000x = 12793.17 + 0.00\dot{3}1\dot{7}.$$

Subtracting we obtain

$$999x = 12793.17 - 12.79 = 12780.38 = 1278038/100$$

and so  $x = 1278038/99900$ , a rational number as required.  $\square$

We will see later on that the converse of this theorem is also true (see Problems IV, Question 5). This means that the rational numbers correspond precisely to the recurring decimals (a finite decimal can be thought of as a recurring infinite decimal:  $a_0.a_1 \dots a_n = a_0.a_1 \dots a_n\dot{0}$ ).

### Exercises

**13.1** Prove that there does not exist a rational number whose square is 3.

[You may assume that  $a^2$  is divisible by 3 if and only if  $a$  is divisible by 3 (proved later as Proposition 15.2.1).]

**13.2** Find the first two decimal places of an infinite decimal representing  $\sqrt{3}$ .

**13.3** Try to mimic the proof of Theorem 13.2.1 to show that there does not exist a rational number whose square is 4. Where does this go wrong?

**13.4** Prove from Definition 13.3.2 that  $1/3 = 0.\dot{3}$ .

**13.5** Find the rational number equal to the recurring infinite decimal 1.7782345.

# 14

## Counting infinite sets

In Chapter 10 a non-empty set  $X$  was defined to have cardinality  $n$ , where  $n$  is a positive integer, when there is a bijection  $N_n \rightarrow X$  from the standard set  $N_n = \{1, 2, \dots, n\}$ . Such sets, together with the empty set, were defined as *finite* and all others as *infinite*.

We saw (Exercise 10.1) that, given a finite set  $X$ , another set  $Y$  is also finite of the same cardinality if and only if there is a bijection  $X \rightarrow Y$ . In this case we say that  $X$  and  $Y$  are *equipotent*. A bijection  $X \rightarrow Y$  guarantees that  $X$  and  $Y$  have the same cardinality.

These ideas may be extended to infinite sets. We can think of any two sets as having the same cardinality if they are equipotent so that there is a bijection between them. We can then describe this cardinality by introducing additional standard sets. For example, when we embark on counting a set  $X$  we assign a distinct element of the set to each positive integer in turn. If we exhaust the elements on reaching the integer  $n$  then we have constructed a bijection  $N_n \rightarrow X$  and so the set is finite with cardinality  $n$ . If we never exhaust the elements we may nevertheless reach each element of the set eventually so that we have a bijection  $\mathbb{Z}^+ \rightarrow X$ . In this case the set is equipotent to the standard set  $\mathbb{Z}^+$  and we say that the set, although infinite, is *denumerable*.

In this chapter we will explore the properties of infinite sets, considering which properties of finite sets can be extended to such sets. It will become clear that not all do so and this should help explain why such care was needed in developing the theory of finite sets. The fact that when we come to infinite sets various ‘obvious’ results are no longer true suggests that these results may not be quite so obvious after all! In particular we will see that care must be taken when comparing the ‘size’ of infinite sets but that it is possible to do this. It is even possible in certain cases to use ‘counting arguments’ with infinite sets.

The material in this chapter is somewhat harder than most of the rest of the book. It is not used subsequently in the book but is included to provide a context for the earlier material on finite sets and because the ideas are so striking and should be part of any general mathematical education.

### 14.1 Countable sets

**Definition 14.1.1** Two sets  $X$  and  $Y$  are **equipotent** if there is a bijection  $X \rightarrow Y$ .

Thus  $X$  is finite of cardinality  $n \in \mathbb{Z}^+$  if and only if  $X$  is equipotent to  $\mathbb{N}_n$ .

**Definition 14.1.2** A set  $X$  is said to be **denumerable** or enumerable if there is a bijection  $\mathbb{Z}^+ \rightarrow X$ , so that it is equipotent to  $\mathbb{Z}^+$ . A set is **countable** if either it is finite or it is denumerable. A set is **uncountable** if it is not countable.

If the set  $X$  is denumerable then it is said to have **cardinality**  $\aleph_0$  and this is written  $|X| = \aleph_0$ .

The notation for the cardinality of a denumerable set uses the Hebrew letter aleph and ' $\aleph_0$ ' is usually read 'aleph null'.

Conventions vary a little and in some books the word 'countable' is used where we are using 'denumerable' or conversely.

If a set  $X$  is denumerable then we can list the elements of the set, but the list is infinite. Given a bijection  $f: \mathbb{Z}^+ \rightarrow X$ ,

$$X = \{x_1, x_2, \dots, x_n, \dots\}$$

where we write  $x_n = f(n)$ .

**Examples 14.1.3** (a)  $\mathbb{Z}^+$  is denumerable by taking the identity function  $\mathbb{Z}^+ \rightarrow \mathbb{Z}^+$ . This gives the listing  $\mathbb{Z}^+ = \{1, 2, \dots, n, \dots\}$ .

(b) The set of non-negative integers  $\mathbb{Z}^{\geq}$  is denumerable since a bijection  $\mathbb{Z}^+ \rightarrow \mathbb{Z}^{\geq}$  is defined by  $n \mapsto n - 1$ . This gives the listing  $\mathbb{Z}^{\geq} = \{0, 1, \dots, n - 1, \dots\}$ .

(c) The set of all integers  $\mathbb{Z}$  is denumerable, for a bijection  $\mathbb{Z}^+ \rightarrow \mathbb{Z}$  may be defined by

$$n \mapsto \begin{cases} -(n-1)/2 & \text{if } n \text{ is odd,} \\ n/2 & \text{if } n \text{ is even.} \end{cases}$$

This gives the listing  $\mathbb{Z} = \{0, 1, -1, 2, -2, \dots, n, -n, \dots\}$ .

(d) The set of positive even integers  $X = \{2n \mid n \in \mathbb{Z}^+\}$  is denumerable, for a bijection  $\mathbb{Z}^+ \rightarrow X$  may be given by  $n \mapsto 2n$ . This gives the listing  $\{2, 4, \dots, 2n, \dots\}$ .

(e) The set of positive perfect squares  $X = \{n^2 \mid n \in \mathbb{Z}^+\}$  is denumerable, for a bijection  $\mathbb{Z}^+ \rightarrow X$  may be given by  $n \mapsto n^2$ . This gives the listing  $X = \{1, 4, \dots, n^2, \dots\}$ .

### *'Paradoxes' of the infinite*

Recall from Exercise 11.5 that a finite set can never have the same cardinality as a proper subset: the cardinality of a proper subset of a finite set is necessarily smaller. However, the above examples demonstrate that a denumerable set can have equipotent proper subsets. Example 14.1.3(d) shows that the set of positive integers and the positive even integers are equipotent and so, in that sense, there are the ‘same number’ of even integers as integers even though it appears that only half the integers are even!

The first reference to this ‘paradox’ appears in the work of Galileo Galilei.<sup>†</sup> He observes that, despite the fact that the further you go along the sequence of positive integers the scarcer the perfect squares become, we must say that there are as many squares as there are positive integers because of the correspondence between the positive integers and the squares given by the bijection  $n \mapsto n^2$  (Example 14.1.3(e)). He suggests that the idea of one infinite quantity being greater than another has no meaning.

These ideas seem not to have been developed further until 1850 when a book by the unconventional Czech priest Bernhard Bolzano was published posthumously.<sup>‡</sup> He showed that there were many examples of bijections between infinite sets and proper subsets. He also suggested that

<sup>†</sup> In his 1638 book *Discorsi e dimostrazioni matematiche intorno à due nuove scienze* (written, remarkably for the period, in Italian rather than Latin and often referred to by its English title *The two new sciences*).

<sup>‡</sup> *Paradoxien des Unendlichen* published in English as *Paradoxes of the infinite* (edited by D.A. Steele) in 1950.

nevertheless certain infinite sets are in a real sense bigger than others, anticipating the work of Cantor which we will come to shortly.

In 1872 the German mathematician Richard Dedekind defined an infinite set to be one which is equipotent to a proper subset of itself.<sup>†</sup> We now observe that this is equivalent to our definition of infinite set.

**Theorem 14.1.4 (Dedekind)** *A set  $X$  is infinite if and only if it is equipotent to a proper subset of itself.*

*Proof* If  $X$  is equipotent to a proper set then, by Exercise 11.5,  $X$  is not finite and so must be infinite.

For the converse, first of all we prove that an infinite set  $X$  must contain a denumerable subset. We define this set  $A = \{a_1, a_2, \dots, a_n, \dots\}$  inductively.

Base case: Choose any element of the set  $X$  as  $a_1$ .

Inductive step: Suppose that, for some  $k \geq 1$ , a subset of  $k$  distinct elements  $A_k = \{a_n \mid 1 \leq n \leq k\}$  has been defined. Choose any element of  $X - A_k$  [non-empty since  $A_k$  is finite and  $X$  is infinite] as  $a_{k+1}$ .

This defines a denumerable subset  $A = \{a_1, a_2, \dots, a_n, \dots\} \subseteq X$ .

We may now define a bijection  $f: X \rightarrow X - \{a_1\}$  onto a proper subset of  $X$ , as required, by

$$f(x) = \begin{cases} a_{n+1} & \text{if } x = a_n \in A, \\ x & \text{if } x \notin A. \end{cases}$$

Hence  $X$  and the proper subset  $X - \{a_1\}$  are equipotent. □

## 14.2 Denumerable sets

For two finite sets, their cardinality determines whether or not they are equipotent, i.e. whether or not there is a bijection between them (Exercise 10.1). This property extends to sets of cardinality  $\aleph_0$ , denumerable sets, as follows.

**Proposition 14.2.1** *Given a denumerable set  $X$ , a set  $Y$  is also denumerable if and only if it is equipotent to  $X$ .*

*Proof* Since  $X$  is denumerable there is a bijection  $f: \mathbb{Z}^+ \rightarrow X$ .

<sup>†</sup> In his book *Stetigkeit und irrationale Zahlen*.

Suppose that  $Y$  is also denumerable. Then there is a bijection  $g: \mathbb{Z}^+ \rightarrow Y$ . Since a bijection is invertible we can now deduce that  $X$  and  $Y$  are equipotent by writing down the bijection  $g \circ f^{-1}: X \rightarrow \mathbb{Z}^+ \rightarrow Y$ .

Conversely, if  $X$  and  $Y$  are equipotent then there is a bijection  $h: X \rightarrow Y$ . We can deduce that  $Y$  is denumerable by writing down the bijection  $h \circ f: \mathbb{Z}^+ \rightarrow X \rightarrow Y$ .  $\square$

When we form unions or products of finite sets we normally obtain bigger sets. One aspect of the unexpected behaviour of infinite sets is that for them this is not the case.

**Proposition 14.2.2** *If  $A$  and  $B$  are denumerable sets then so is their union  $A \cup B$ .*

This is left as an exercise.

**Proposition 14.2.3** *If  $A$  and  $B$  are denumerable sets then so is their Cartesian product,  $A \times B$ .*

**Constructing a proof.** Since  $A$  and  $B$  are denumerable we can list their elements:  $A = \{a_1, a_2, \dots, a_m, \dots\}$  and  $B = \{b_1, b_2, \dots, b_n, \dots\}$ . Using these listings we can display the elements of  $A \times B$  in a double array.

$$\begin{array}{ccccccc} & & & & & \vdots & \\ & (a_1, b_4) & & & & \vdots & \\ & (a_1, b_3) & (a_2, b_3) & & & \dots & \\ & (a_1, b_2) & (a_2, b_2) & (a_3, b_2) & & \dots & \\ (a_1, b_1) & (a_2, b_1) & (a_3, b_1) & (a_4, b_1) & & \dots & \end{array}$$

If we start to count these elements by going along the rows, which would be the normal way of counting a finite array (consider for example the solution to Exercise 10.3) then we will never get to the end of the first row and so no element in the second row will be reached. One way of ensuring† that we do eventually reach every element is to count along

† This should be contrasted with the situation for finite sets. For a finite set it doesn't matter how we start: the addition principle ensures that we will complete the count and Proposition 10.1.3 ensures that the answer will always be the same.

the diagonals in the following pattern.

|    |   |     |     |     |
|----|---|-----|-----|-----|
| :  | : |     |     |     |
| 10 | : |     |     |     |
| 6  | 9 | ... |     |     |
| 3  | 5 | 8   | ... |     |
| 1  | 2 | 4   | 7   | ... |

If we count in this way the element  $(a_m, b_n)$  will occur as the  $n$ th element in diagonal number  $m + n - 1$ . So (since there are  $k$  elements in the  $k$ th diagonal) it will be element number

$$1 + 2 + \dots + (m + n - 2) + n = \frac{1}{2}(m + n - 2)(m + n - 1) + n$$

(using Proposition 5.3.1).

*Proof* Let  $A = \{a_1, a_2, \dots, a_m, \dots\}$  and  $B = \{b_1, b_2, \dots, b_n, \dots\}$ . Then a bijection  $A \times B \rightarrow \mathbb{Z}^+$  may be defined by

$$(a_m, b_n) \mapsto \frac{1}{2}(m + n - 2)(m + n - 1) + n.$$

□

**Corollary 14.2.4** *If  $A$  is a denumerable set then so is  $A^n$  for every positive integer  $n$ .*

*Proof* This is an easy induction proof. The proposition gives the inductive step. □

**Proposition 14.2.5** *A subset of a denumerable set is countable (i.e. either finite or denumerable).*

**Constructing a proof.** As with finite sets we can prove statements about general denumerable sets by first proving them for the standard denumerable set  $\mathbb{Z}^+$ .

The conclusion of this result is an ‘or’ statement and so we use the standard strategy introduced in Chapter 4: we take a non-finite subset,  $A$ , of  $\mathbb{Z}^+$  in other words an infinite subset, and verify that it must be denumerable. We can construct a bijection  $\mathbb{Z}^+ \rightarrow A$  inductively essentially by listing the elements of  $A$  in increasing order starting with the least.

*Proof* (a) We begin by considering subsets of  $\mathbb{Z}^+$ , the set of positive integers. Suppose that  $A$  is an infinite subset of  $\mathbb{Z}^+$ . Then to prove that  $A$  is denumerable we must construct a bijection  $f: \mathbb{Z}^+ \rightarrow A$ . We define  $f(n)$  inductively as follows.

Base case: Let  $f(1)$  be the least element of the set  $A$ . [Recall that by the well-ordering principle (see Example 11.2.2(c)) any non-empty set of positive integers has a least element.]

Inductive step: Suppose that  $f(k)$  has been defined for some  $k \geq 1$ . Let  $f(k+1)$  be the least element of the set  $\{a \in A \mid a > f(k)\}$ . [This set is non-empty since  $A$  is infinite.]

This defines a function  $f: \mathbb{Z}^+ \rightarrow A$ .

This function is an injection since  $f(k) < f(k+1)$  for all  $k \geq 1$ .

To see that  $f$  is a surjection notice that  $f(n) \geq n$  (easily proved by induction). Given an element  $a \in A$  then let  $m$  be the least integer such that  $f(m) \geq a$  [we must have  $m \leq a$  since  $f(a) \geq a$ ]. Then by the definition of  $f$  we have  $f(m) = a$ .

Thus we have constructed a bijection  $f: \mathbb{Z}^+ \rightarrow A$  showing that  $A$  is denumerable.

Hence each subset of  $\mathbb{Z}^+$  is finite or denumerable.

(b) Now consider the general case. Let  $A$  be a subset of a denumerable set  $X$ . Since  $X$  is denumerable there is a bijection  $g: \mathbb{Z}^+ \rightarrow X$ . Then  $\overleftarrow{g}(A)$  is a subset of  $\mathbb{Z}^+$  and so finite or denumerable. But the restriction of  $g$  gives a bijection  $\overleftarrow{g}(A) \rightarrow A$  and so  $A$  is finite or denumerable.  $\square$

**Theorem 14.2.6 (Cantor (1874))** *The set of rationals  $\mathbb{Q}$  is denumerable.*

*Proof* We may define a function

$$f: \mathbb{Q} \rightarrow \mathbb{Z} \times \mathbb{Z}^+$$

by  $f(q) = (m, n)$  where  $m/n$  is the fraction representing  $q$  in its lowest terms (i.e.  $q = m/n$ ,  $n > 0$  and  $\gcd(m, n) = 1$ ). This is clearly an injection. Its image is an infinite subset of the denumerable set  $\mathbb{Z} \times \mathbb{Z}^+$  and so denumerable. Hence, since  $f$  defines a bijection onto its image,  $\mathbb{Q}$  is denumerable.  $\square$

### 14.3 Uncountable sets

In the previous section we have seen that surprisingly many infinite sets are denumerable and indeed the examples there support Galileo's

suggestion that we can never say that one infinite set is larger than another. However, this is not the case. There are uncountable sets and we can give meaning to the statement that these are larger than the countable sets.

In 1874, Georg Cantor published a remarkable paper in which he proved not only that the set of rational numbers is denumerable but also that there is a hierarchy of infinite sets ordered by size (or cardinality).

**Theorem 14.3.1 (Cantor (1874))** *The set of real numbers  $\mathbb{R}$  is uncountable.*

**Constructing a proof.** Notice that the statement that  $\mathbb{R}$  is uncountable is a non-existence statement: there does not exist a bijection  $\mathbb{Z}^+ \rightarrow \mathbb{R}$ . The method is to show how given any function  $\mathbb{Z}^+ \rightarrow \mathbb{R}$  we can construct an element of  $\mathbb{R}$  which is not in the image of the function, not a value, so that the function is not a surjection and so not a bijection.

Each element  $x \in \mathbb{R}$  may be represented as an infinite decimal

$$x = a_0.a_1a_2 \dots a_n \dots$$

where  $a_i \in \mathbb{Z}$  and, for  $i \geq 1$ ,  $0 \leq a_i \leq 9$ . Some numbers (such as  $(1.\dot{0} = 0.\dot{9})$ ) have two decimal expansions, one ending in recurring 0's (and so a finite decimal) and one ending in recurring 9's (see Problems III, Questions 26 and 27). For these we always use the first representation.

The method is now to start with a function  $\mathbb{Z}^+ \rightarrow \mathbb{R}$  and to list the values of the function as infinite decimals. Having done this we have a beautifully simple, but ingenious, method (known as ‘Cantor’s diagonal process’) for writing down an infinite decimal not in the list and not representing the same real number as any member of the list. This infinite decimal represents a real number which is not a value of the function.

*Proof* Suppose that  $f: \mathbb{Z}^+ \rightarrow \mathbb{R}$  is a function. Let

$$f(m) = a_{m0}.a_{m1}a_{m2} \dots a_{mn} \dots$$

as an infinite decimal (not ending in recurring 9’s).

Put  $b = 0.b_1b_2 \dots b_n \dots$  where

$$b_n = \begin{cases} 1 & \text{if } a_{nn} = 0, \\ 0 & \text{if } a_{nn} \neq 0. \end{cases}$$

Notice that this infinite decimal representing  $b$  certainly does not end

in recurring 9's. Furthermore, for each  $n$ , the  $n$ th decimal place of  $b$ , namely  $b_n$ , certainly differs from the  $n$ th decimal place of  $f(n)$ , namely  $a_{nn}$ . So  $b \neq f(n)$ . Since we have constructed an element of  $\mathbb{R}$  which is not a value of  $f$  this shows that  $f$  is not a surjection. Hence  $\mathbb{R}$  is not denumerable.  $\square$

This result shows that the set of real numbers is in a real sense larger than the set of integers (or the set of rationals). We can make this precise as follows.

**Definition 14.3.2** *Two sets  $X$  and  $Y$  have the same cardinality, written*

$|X| = |Y|$ , if they are equipotent, i.e. there is a bijection  $X \rightarrow Y$ .

If there is an injection  $X \rightarrow Y$  then we write  $|X| \leq |Y|$ . We write

$|X| < |Y|$  to mean that  $|X| \leq |Y|$  and  $|X| \neq |Y|$  and say that  $X$  has smaller cardinality than  $Y$ .

Thus Theorem 14.3.1 may be stated as  $|\mathbb{R}| > |\mathbb{Z}^+|$  or  $|\mathbb{R}| > \aleph_0$ . One natural question now is whether there is a set intermediate in size between  $\mathbb{Z}^+$  and  $\mathbb{R}$ . In 1878, Cantor conjectured that such sets did not exist, a statement which became known as the continuum hypothesis. Here the word 'continuum' is simply a name for the set of real numbers  $\mathbb{R}$ .

**The continuum hypothesis.** Every uncountable set of real numbers has the same cardinality as the set of all real numbers:

$$X \subseteq \mathbb{R} \text{ and } |X| > \aleph_0 \Rightarrow |X| = |\mathbb{R}|.$$

In 1963 the United States mathematician Paul Cohen proved the remarkable result that it is not possible either to prove or to disprove this hypothesis starting from the generally accepted assumptions (axioms) of set theory. This means that there are two alternative versions of set theory, one in which the continuum hypothesis is true and one in which it is false.

The other natural question arising from Theorem 14.3.1 is whether there is a set larger than  $\mathbb{R}$ . Cantor was able to show that there is such a set and indeed that, given any set, its power set is necessarily larger.

**Theorem 14.3.3** For any set  $X$ ,  $|X| < |\mathcal{P}(X)|$ .

**Constructing a proof.** Recall that  $\mathcal{P}(X)$ , the power set of  $X$ , is the set of all subsets of  $X$  (Definition 6.3.1).

It is easy to define an injection  $X \rightarrow \mathcal{P}(X)$  by sending each element to the corresponding singleton subset. The interesting part is in proving that there is no bijection  $X \rightarrow \mathcal{P}(X)$ . The method is similar to that of Theorem 14.3.1: starting from any function  $f: X \rightarrow \mathcal{P}(X)$  we construct an element of  $\mathcal{P}(X)$ , in other words a subset of  $X$ , which is not a value of  $f$ . It is easy to describe such an element but the reason why it is not a value is quite subtle.

*Proof* An injection  $X \rightarrow \mathcal{P}(X)$  may be defined by  $x \mapsto \{x\}$  and so  $|X| \leq |\mathcal{P}(X)|$ .

We prove that there is no bijection  $X \rightarrow \mathcal{P}(X)$  by showing that any function  $X \rightarrow \mathcal{P}(X)$  is not a surjection. Suppose that  $f: X \rightarrow \mathcal{P}(X)$  is a function. We prove that this function is not a surjection and so not a bijection by exhibiting an element of  $\mathcal{P}(X)$  which is not a value. Define

$$A = \{x \in X \mid x \notin f(x)\}.$$

[Notice that, for each  $x \in X$ ,  $f(x) \in \mathcal{P}(X)$  is a subset of  $X$  and so either  $x \in f(x)$  or  $x \notin f(x)$ . Thus ' $x \notin f(x)$ ' is a predicate with a free variable taking values in  $X$ . The set of elements of  $X$  for which this predicate is true is a well-defined subset of  $X$ : this is the set  $A$ .]

To prove that  $A$  is not in the image of the function  $f$  we suppose for contradiction that it is, say  $A = f(a)$  for some  $a \in X$ . Then  $A$  is a subset of  $X$  and  $a$  is an element of  $X$ , and so either  $a \in A$  or  $a \notin A$ . We consider the possibilities in turn.

If  $a \in A$  then, referring back to the predicate defining  $A$ ,  $a \notin f(a)$ , which can be written  $a \notin A$  contradicting our assumption that  $a \in A$ .

On the other hand, if  $a \notin A$ , then it does not satisfy the predicate defining  $A$  and so  $a \in f(a)$ , which can be written  $a \in A$  contradicting our assumption that  $a \notin A$ .

So in either case we have a contradiction and so our initial assumption that  $A$  is in the image of  $f$  must be false and so  $f$  is not a surjection.

Hence, since there is no bijection  $X \rightarrow \mathcal{P}(X)$  but there is an injection,  $|X| < |\mathcal{P}(X)|$ .  $\square$

This is probably the most subtle proof in this book. However, once you have understood it, it appears to have the same glorious inevitability as the great Euclidean proofs by contradiction.

The most important result on finite sets was the pigeonhole principle (Theorem 11.1.2). This result does still hold for infinite sets.

**Theorem 14.3.4 (Cantor–Schröder–Bernstein theorem)** *Suppose that  $X$  and  $Y$  are non-empty sets such that  $|X| > |Y|$ . Then any function  $f: X \rightarrow Y$  is not an injection, i.e. there exist distinct elements  $x_1$  and  $x_2 \in X$  such that  $f(x_1) = f(x_2)$ .*

The proof is omitted.†

This material on counting infinite sets may appear very abstract and obscure and indeed it was many years before it gained general acceptance. To give some idea of its power we will conclude by describing one result which Cantor obtained using his ideas.

**Definition 14.3.5** *A real number is called algebraic if it satisfies a polynomial equation*

$$x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n = 0$$

*with the coefficients  $a_i$  all rational. It is called transcendental if it is not algebraic.*

Thus all rational numbers are algebraic (since  $m/n$  satisfies  $nx - m = 0$ ) and  $\sqrt{2}$  is algebraic (since it satisfies  $x^2 = 2$ ). In 1844 the French mathematician Joseph Liouville gave an explicit construction of certain transcendental numbers‡ but it remained a problem to determine whether specific numbers are transcendental. In 1873, Charles Hermite (French) proved that  $e$ , the base of the natural logarithms, is transcendental, and these ideas were developed by Ferdinand Lindemann (German) to prove in 1882 that  $\pi$  is transcendental. This result finally proved that the ancient search for a ruler and compass construction to give a square of the same area as a given circle has no solution since it can be shown that any length which can be constructed from a given length is an algebraic multiple of the given length§. These arguments about specific numbers are quite delicate and it is still not known whether, for example,  $\pi + e$  is transcendental.

† A proof may be found in D.J. Velleman, *How to prove it, a structured approach*, Cambridge University Press, 1994.

‡ A description of this construction may be found in R. Courant and H. Robbins, *What is mathematics?* Oxford University Press, 1941.

§ See R. Courant and H. Robbins, *op. cit.*

Cantor's remarkable result was that *most* real numbers are transcendental! He proved, using similar methods to the above proof of Theorem 14.2.6, that the set of algebraic numbers is denumerable (see Problems III, Question 28). It follows from Theorem 14.3.1 and Exercise 14.2 that the set of transcendental numbers is uncountable and it is in that sense that most real numbers are transcendental.

### Exercises

**14.1** Prove that if  $A$  is finite and  $B$  is denumerable then  $A \cup B$  is denumerable.

**14.2** Prove that if  $A$  and  $B$  are denumerable then  $A \cup B$  is denumerable. Hence prove by contradiction that, if  $X$  is uncountable and a subset  $A$  is denumerable, then  $X - A$  is uncountable.

**14.3** Prove that if  $\{A_n \mid n \in \mathbb{Z}^+\}$  is a denumerable set of pairwise disjoint denumerable sets then the union

$$\bigcup_{n \in \mathbb{Z}^+} A_n = \{x \mid x \in A_n \text{ for some } n \in \mathbb{Z}^+\}$$

is also denumerable.

**14.4** Use the Cantor–Schröder–Bernstein theorem to prove that if  $|X| \leq |Y|$  and  $|Y| \leq |X|$  then  $|X| = |Y|$ .

## Problems III: Numbers and counting

1. Of the 182 students who are taking three first year core Mathematics modules (Reasoning, Algebra and Calculus), 129 like Reasoning, 129 like Algebra, 129 like Calculus, 85 like Reasoning and Algebra, 89 like Reasoning and Calculus, 86 like Algebra and Calculus, and 54 like all three modules. How many of the students like none of the core modules?
2. Of the 170 students who took all the first year core modules last year, 124 liked Reasoning, 124 liked Algebra, 124 liked Calculus, 10 liked only Reasoning, nobody liked only Algebra, 4 liked only Calculus and 2 liked none of the modules. How many students liked all three core modules?
3. At an international conference of 100 people, 75 speak English, 60 speak Spanish and 45 speak Swahili (and everyone present speaks at least one of these languages).
  - (i) What is the maximum possible number of these people who can speak only one language? In this case how many people speak only English, how many speak only Spanish, how many speak only Swahili, and how many speak all three?
  - (ii) What is the maximum number of people who speak only English? In this case what can be said about the number who speak only Spanish and the number who speak only Swahili?
  - (iii) Prove that the greater the number of people who speak all three languages, the greater the number of people who speak only one language.
4. Prove (by induction on  $n$ ) the general inclusion–exclusion principle which may be stated as follows.

Let  $A_1, A_2, \dots, A_n$  be finite sets. For  $I = \{i_1, i_2, \dots, i_r\} \subseteq \mathbb{N}_n$ , write

$$A_I = \bigcap_{i \in I} A_i = A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_r}.$$

Then

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{\emptyset \neq I \subseteq \mathbb{N}_n} (-1)^{|I|-1} |A_I|,$$

summing over all non-empty subsets of  $\mathbb{N}_n$ .

5. Prove that a finite non-empty set of real numbers has a minimum element.

[This is the part of the proof of Proposition 11.2.3 left as an exercise.]

6. Suppose that  $A$  and  $B$  are non-empty finite sets of real numbers such that  $A \subseteq B$ . Prove that

$$\min B \leq \min A \leq \max A \leq \max B.$$

7. What is wrong with the following argument purporting to prove that for every non-empty finite set  $A$  of real numbers  $\min A = \max A$ ?

*Proof* We prove by induction on  $n$  that if  $|A| = n$  then  $\min A = \max A$ .

Base case: If  $|A| = 1$  then  $A$  is a singleton set  $A = \{a_1\}$  and  $\min A = a_1 = \max A$ .

Inductive step: Suppose that for some positive integer  $k$  the result is true for  $n = k$ . Let  $A$  be a set of real numbers with  $|A| = k + 1$ . Let  $a_1 = \min A$  and  $a_2 = \max A$ . Then

$$\begin{aligned} a_1 &= \min(A - \{a_2\}) \\ &= \max(A - \{a_2\}) \quad (\text{by inductive hypothesis}) \\ &\geq \max(A - \{a_1, a_2\}) \quad (\text{by Question 6}) \\ &\geq \min(A - \{a_1, a_2\}) \quad (\text{by Question 6}) \\ &\geq \min(A - \{a_1\}) \quad (\text{by Question 6}), \\ &= \max(A - \{a_1\}) \quad (\text{by inductive hypothesis}) \\ &= a_2. \end{aligned}$$

Hence  $a_1 \geq a_2$  and since clearly  $a_1 \leq a_2$  this proves that  $a_1 = a_2$  as required to deduce the result for  $n = k + 1$  and so proving the inductive step.

Conclusion: Hence by induction on  $n$ ,  $\min A = \max A$  for a set of cardinality  $n$  for all positive integers  $n$ , and so for all finite sets.  $\square$

8. Prove Lemma 10.1.4 by induction on  $m$ .

**9.** Prove Proposition 11.1.4: if  $X \rightarrow \mathbb{N}_n$  is an injection, then  $X$  is finite and  $|X| \leq n$ .

**10.** Prove Theorem 11.1.6: given non-empty finite sets  $X$  and  $Y$  with  $|X| < |Y|$  there does not exist a surjection  $X \rightarrow Y$ .

**11.** Use the pigeonhole principle and proof by contradiction to prove Theorem 11.1.7: given non-empty finite sets  $X$  and  $Y$  with  $|X| = |Y|$ , a function  $X \rightarrow Y$  is an injection if and only if it is a surjection.

**12.** Suppose that there is an injection  $f: \mathbb{Z}^+ \rightarrow X$ . Prove by contradiction that  $X$  is an infinite set.

[Use Corollary 11.1.1 noting that, for any  $n \geq 1$ ,  $f$  restricts to give an injection  $\mathbb{N}_{n+1} \rightarrow X$ .]

**13.** Find all the divisors of 126 and 180 and hence find the greatest common divisor (126, 180).

**14.** For  $n \in \mathbb{Z}^+$ , suppose that  $A \subseteq \mathbb{N}_{2n}$  and  $|A| = n + 1$ . Prove that  $A$  contains a pair of distinct integers  $a, b$  such that  $a$  divides  $b$ .

[Let  $f(a)$  be the greatest odd integer which divides  $a$  and apply the pigeonhole principle to  $f$ .]

**15.** Prove the induction principle from the well-ordering principle (see Example 11.2.2(c)).

[Prove the induction principle in the form of Axiom 7.5.1 by contradiction.]

**16.** Use the inclusion–exclusion principle (Question 4) to prove that the number of surjections  $\mathbb{N}_m \rightarrow \mathbb{N}_n$  is given by

$$n^m - \binom{n}{1}(n-1)^m + \binom{n}{2}(n-2)^m - \dots + (-1)^{n-1} \binom{n}{n-1} 1^m.$$

[For each  $i \in \mathbb{N}_n$ , let  $A_i = \{ f: \mathbb{N}_m \rightarrow \mathbb{N}_n \mid i \text{ is not a value of } f \}$ .]

Deduce that

$$n^n - \binom{n}{1}(n-1)^n + \binom{n}{2}(n-2)^n - \dots + (-1)^{n-1} \binom{n}{n-1} 1^n = n!$$

- 17.** A *derangement* of  $\mathbb{N}_n$  is a bijection  $f: \mathbb{N}_n \rightarrow \mathbb{N}_n$  with no fixed points, i.e.  $f(i) \neq i$  for all  $i \in \mathbb{N}_n$ . Use the inclusion-exclusion principle to prove that the number of derangements of  $\mathbb{N}_n$  is

$$n! \left( \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots + (-1)^n \frac{1}{n!} \right).$$

[This means that the proportion of all permutations of  $\mathbb{N}_n$  which are derangements is  $1/2! - 1/3! + 1/4! - \dots + (-1)^n 1/n! = 1 - 1/1! + 1/2! - 1/3! + 1/4! - \dots + (-1)^n 1/n!$ . The reader may recognize this as the first part of the series for  $e^{-1}$ . It follows that for large values of  $n$  the probability that a permutation is a derangement is approximately  $1/e$ .]

- 18.** Suppose that  $X$  and  $Y$  are disjoint sets. Prove that the function

$$\bigcup_{i=0}^k \mathcal{P}_i(X) \times \mathcal{P}_{k-i}(Y) \rightarrow \mathcal{P}_k(X \cup Y)$$

defined by  $(A, B) \mapsto A \cup B$  is a bijection. Deduce that

$$\binom{m+n}{k} = \sum_{i=0}^n \binom{m}{i} \binom{n}{k-i}.$$

Give an alternative proof of this result by induction on  $n$  using Proposition 12.2.8.

- 19.** Prove Leibniz's rule for higher order derivatives of products,

$$\frac{d^n(uv)}{dx^n} = \sum_{i=0}^n \binom{n}{i} \frac{d^i u}{dx^i} \frac{d^{n-i}v}{dx^{n-i}} \quad \text{for } n \in \mathbb{Z}^+,$$

by induction on  $n$ .

- 20.** Use the pigeonhole principle to prove that, given ten distinct positive integers less than 107, there exist two disjoint subsets with the same sum.

- 21.** Define the notion of inequality between two fractions by

$$\frac{a_1}{b_1} < \frac{a_2}{b_2} \Leftrightarrow \begin{cases} a_1 b_2 < a_2 b_1 & \text{if } b_1 b_2 > 0, \\ a_1 b_2 > a_2 b_1 & \text{if } b_1 b_2 < 0. \end{cases}$$

Prove that this definition leads to a well-defined notion of inequality between two rational numbers satisfying Axioms 3.1.2.

**22.** For a positive real number  $x$  and a rational number  $q = m/n$  (where  $m$  and  $n$  are integers), define the power  $x^q$  to be  $(x^{1/n})^m$ . In this expression the integral power is defined in Problems I, Question 23 and  $x^{1/n}$  is defined in Example 9.2.4(b).

Prove that this definition of rational powers is well-defined. With this definition prove the laws of exponents for any positive real numbers  $x$  and  $y$  and rational numbers  $p$  and  $q$ :

- (i)  $x^q y^q = (xy)^q$ ;
- (ii)  $x^{p+q} = x^p x^q$ ;
- (iii)  $(x^p)^q = x^{pq}$ .

**23.** Prove that there does not exist a rational number whose square is 10.

**24.** Find the first four decimal places of the infinite decimal representing  $\sqrt{10}$  (without using a calculator!).

**25.** Find the rational number equal to the recurring infinite decimal 2.10012097̇.

**26.** Prove that each finite decimal may be written as an infinite decimal in two distinct ways:

$$\begin{aligned} a_0.a_1a_2\dots a_{n-1}a_n &= a_0.a_1a_2\dots a_{n-1}a_n\dot{0} \\ &= a_0.a_1a_2\dots a_{n-1}(a_n - 1)\dot{9}, \end{aligned}$$

where  $a_n > 0$  if  $n > 0$ .

**27.** Prove that each real number is represented by a unique infinite decimal unless it is representable by a finite decimal, in which case it is representable by precisely two infinite decimals as described in the previous problem.

[Hint: Suppose that two infinite decimals represent the same number:  $a_0.a_1a_2\dots a_n\dots = b_0.b_1b_2\dots b_n\dots$ . Attempt to prove by induction that  $a_n = b_n$  for all  $n \geq 0$ .]

- 28.** Prove that the number of polynomials of degree  $n$  with rational coefficients is denumerable. Deduce that the set of algebraic numbers (see Definition 14.3.5) is denumerable.

[Recall that a polynomial equation  $x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n = 0$  has at most  $n$  real roots.]



# **Part IV**

## Arithmetic



# 15

## The division theorem

The remainder of this book illustrates how the methods of proof and the language of sets and functions introduced so far are applied in arithmetic, the theory of numbers. We begin in this chapter by discussing the elementary ideas of division and remainders which we all met at an early stage of our schooling. The formal treatment of this generalizes the discussion of odd and even integers in Section 11.3.

### 15.1 The division theorem

The main result of this chapter is the following.

**Theorem 15.1.1 (The division theorem)** *Let  $a$  and  $b$  be integers with  $b > 0$ . Then there are unique integers  $q$  and  $r$  such that*

$$a = bq + r \quad \text{and} \quad 0 \leq r < b.$$

This result should be very familiar from early school days. There, in the case of positive  $a$ , we meet the idea of ‘sharing’  $a$  objects equally amongst  $b$  people: they each receive  $q$  objects with  $r$  ‘left over’. The number  $q$  is called the *quotient* and the number  $r$  is called the *remainder*.

In practice when we are given two numbers  $a$  and  $b$  we can find  $q$  either by division using various tricks we learnt at school or more commonly these days by means of a calculator. The calculator will give us a finite decimal which approximates the rational number  $a/b$ . The quotient  $q$  is then the greatest integer less than or equal to  $a/b$  which is given by the integer part of  $a/b$  when  $a$  is positive. We can then find  $r$  as  $a - bq$ .

**Example 15.1.2** To find the quotient and remainder when  $a = 1781293$  and  $b = 1481$ .

For these numbers a calculator gives  $a/b \approx 1202.7637$  (or possibly 1202.7636 depending on the rounding system). Thus  $q = 1202$  and so  $r = 1781293 - 1481 \times 1202 = 1131$ .

**Example 15.1.3** To find the quotient and the remainder when  $a = -7856123$  and  $b = 9812$ .

For these numbers a calculator gives  $a/b \approx -800.66480$ . Thus  $q = -801$  (notice that we *increase* the absolute value of a negative number when we round *down* to the next integer) and so  $r = a - bq = -7856123 + 9812 \times 801 = 3289$ .

These examples demonstrate only that the division theorem is true in these specific numerical examples. We cannot base a general proof on this method since the construction of a decimal corresponding to a rational number depends in general on repeated use of the division theorem, the result we are trying to prove (see Exercise 15.6)! The proof of the theorem requires some care, if only because the result seems so obvious. It is not necessary to have a complete understanding of the proof in order to apply the theorem. It is common in learning mathematics to gain experience in how a theorem is used before fully understanding the proof; seeing how the theorem is used can help to clarify the full meaning of the theorem and so help in exploring the proof.

In the case of  $b = 2$  we have already proved Theorem 15.1.1 as Proposition 11.3.4 for we proved there that a number  $a$  is odd if and only if  $a = 2q + 1$  for some integer  $q$ . Since by definition  $a$  is even if and only if  $a = 2q$  for some integer  $q$  and ‘odd’ is defined to mean ‘not even’, this result means that each integer  $a$  can be written uniquely in the form  $a = 2q + r$  where  $r = 0$  or  $r = 1$ . The number is even if and only if  $r = 0$  and the number is odd if and only if  $r = 1$ .

The method of proof of the division theorem is a direct generalization of the proof of Proposition 11.3.4.

**Constructing a proof.** Notice that there are really two separate things to prove: that the numbers  $q$  and  $r$  exist (this is an example of an *existence theorem*) and that they are unique (a *uniqueness theorem*). It is common for students to omit the uniqueness part when asked to state the division theorem but as will be seen this is a very important part of the result for many applications.

But let us start with existence: it is necessary to show that given  $a$  and  $b$  as in the theorem there exist integers  $q$  and  $r$  such that  $a = bq + r$  and  $0 \leq r < b$ .

First of all notice that  $q$  and  $r$  determine each other and in particular  $r = a - bq$ . So we can remove all reference to  $r$  by restating the condition  $0 \leq r < b$  as  $0 \leq a - bq < b$  or equivalently  $bq \leq a < b(q + 1)$ . Thus we can describe what we need as follows.

| Given                 | Goal   |
|-----------------------|--|
| $a, b \in \mathbb{Z}$ |  |
| $b > 0$               | $\exists q \in \mathbb{Z}, bq \leq a < b(q + 1)$ |

As in the proof of Proposition 11.3.4 it is convenient to deal first with the case of  $a$  non-negative. In this case as in the earlier proof we can get hold of the element  $q$  by defining it to be the maximum element of a certain finite set.

Observe that, if  $a$  is non-negative, then there is no difficulty in finding an integer  $k$  such that  $bk \leq a$  for we can simply take  $k = 0$ . In fact if  $a < b$  then we can simply take  $q = 0$  giving  $r = a$ . Otherwise, we try the non-negative integers,  $0, 1, 2, \dots$ , in turn until we reach the first integer  $q$  such that  $a < b(q + 1)$ . (This is more or less what we do in practice when performing division although the long division methods we learnt at school give some tricks so that we do not have to try every number.) The point of the existence part of the theorem is that this process will eventually succeed.

To see this we introduce the set

$$A = \{ k \in \mathbb{Z} \mid k \geq 0 \text{ and } bk \leq a \}.$$

This is a non-empty set since  $0 \in A$  and it is a finite set since  $k \in A \Rightarrow bk \leq a \Rightarrow k \leq a/b$ . Hence  $A$  will have a maximum element so that we can define  $q = \max A$ .

To ensure that this  $q$  is the number we want we must prove that  $bq \leq a < b(q + 1)$ . Let us sum up where we have got to.

| Given        | Goal                          |
|--------------|-------------------------------|
| $q = \max A$ | $bq \leq a$<br>$a < b(q + 1)$ |

The first goal is immediate since  $q = \max A$  implies that  $q \in A$ . The second goal is a little more elusive. However, when we rewrite it in the form  $b(q + 1) \not\leq a$  we see that it is actually the statement that  $q + 1 \notin A$ . So we can rewrite our strategy for this second goal as follows.

| Given        | Goal             |
|--------------|------------------|
| $q = \max A$ | $q + 1 \notin A$ |

This is really obvious from the definition of what is meant by the maximum element of a set and this is normally all that would be said to complete the proof. For a formal proof notice that the negative reformulation of the goal strongly suggests trying a proof by contradiction giving the following strategy.

| Given                         | Goal          |
|-------------------------------|---------------|
| $q = \max A$<br>$q + 1 \in A$ | contradiction |

Now observe that if  $q + 1 \in A$  then  $q + 1 \leq \max A = q$ , providing the required contradiction since we know that  $q + 1 > q$ .

To complete the proof it is necessary to consider the case when  $a$  is a negative integer.<sup>†</sup> As in the case of  $b = 2$  in Proposition 11.3.4 we can prove the result for negative  $a$  by using the result for  $-a$  which is positive.

We now write this proof out formally.

*Proof of existence* Suppose that  $a \geq 0$ . Consider the set

$$A = \{ k \in \mathbb{Z} \mid k \geq 0 \text{ and } bk \leq a \}.$$

This set is non-empty since  $0 \in A$  and is finite since  $k \in A \Rightarrow 0 \leq k \leq a$ . So it contains a maximum element,  $q = \max A$ .

Now put  $r = a - bq$  so that  $a = bq + r$ . To prove that  $q$  and  $r$  are as claimed in the theorem it remains to prove that  $0 \leq r < b$ . It is clear that  $r \geq 0$  since, by the definition of  $q$ ,  $bq \leq a$ . To prove that  $r < b$  suppose for contradiction that  $r \geq b$ . Then  $a - bq \geq b$  so that  $b(q + 1) \leq a$ , i.e.  $q + 1 \in A$  contradicting the maximality of  $q$ . Hence  $r < b$  as required to complete the proof of existence.

To complete the proof suppose that  $a$  is a negative integer. Then  $-a$  is positive so that from what we have proved  $-a = bq_1 + r_1$  for integers  $q_1$  and  $r_1$  with  $0 \leq r_1 < b$ . Thus  $a = b(-q_1) - r_1$ . We now consider two cases.

(i) If  $r_1 = 0$ , then  $a = b(-q_1)$  and so we have an expression of the

<sup>†</sup> The argument given for non-negative  $a$  can be amended to include the case of negative  $a$  by replacing  $A$  by  $A' = \{ k \in \mathbb{Z} \mid bk \leq a \}$ , in other words allowing negative integers in the set  $A'$ . We can then define  $q = \max A'$  once we are clear that  $A'$  has a maximum element. In this case the set  $A'$  is not a finite set for it contains all large negative integers but it can still be shown to have a maximum using the well-ordering principle. It seems simpler in this book to deal only with maximum values of finite sets and to prove this result for negative  $a$  by making use of the result for  $-a$  since in practice this would be the way in which  $q$  and  $r$  would be found, as illustrated in Example 15.1.3.

required form with  $q = -q_1$  and  $r = 0$ .

(ii) If  $r_1 > 0$ , then  $a = b(-q_1) - r_1 = b(-q_1 - 1) + (b - r_1)$  and so we have an expression of the required form with  $q = -q_1 - 1$  and  $r = b - r_1$ . Notice that, since  $0 < r_1 < b$ , we also have  $0 < b - r_1 < b$ .  $\square$

**Constructing a proof of uniqueness.** The other part of the theorem is uniqueness. This says that if we have integers  $q_1, q_2, r_1, r_2$  such that

$$a = bq_1 + r_1 = bq_2 + r_2, \quad 0 \leq r_1 < b \quad \text{and} \quad 0 \leq r_2 < b$$

then  $q_1 = q_2$  and  $r_1 = r_2$ . As before,  $q_i$  determines  $r_i$  and so  $q_1 = q_2 \Rightarrow r_1 = r_2$  which means that it is sufficient to prove that  $q_1 = q_2$ . We can reorder the pairs if necessary so that  $q_1 \geq q_2$  which means that  $r_1 = a - bq_1 \leq r_2 = a - bq_2$ . This leaves us with the following strategy.

| Given  | Goal        |
|--|-------------|
| $q_1, q_2 \in \mathbb{Z}$<br>$0 \leq a - bq_1 \leq a - bq_2 < b$ | $q_1 = q_2$ |

But this is easy since  $a - bq_1$  and  $a - bq_2$  differ by a multiple of  $b$  and there isn't room for two distinct non-negative integers differing by a multiple of  $b$  within the set of non-negative integers less than  $b$ . Let us write the proof out properly straight away.

*Proof of uniqueness* Suppose that  $a = bq_1 + r_1 = bq_2 + r_2$  with  $0 \leq r_1, r_2 < b$ . We must prove that  $q_1 = q_2, r_1 = r_2$ . To do this suppose, without loss of generality, that  $q_1 \geq q_2$ . Then  $0 \leq r_1 = a - bq_1 \leq r_2 = a - bq_2 < b$ . This means that  $0 \leq (a - bq_2) - (a - bq_1) < b$  so that  $0 \leq b(q_1 - q_2) < b$  which implies that  $0 \leq q_1 - q_2 < 1$ . But since  $q_1 - q_2$  is an integer it follows that  $q_1 - q_2 = 0$ , i.e.  $q_1 = q_2$  from which it follows also that  $r_1 (= a - bq_1) = r_2 (= a - bq_2)$ .  $\square$

## 15.2 Some applications

The division theorem will be fundamental for the number theory in the remainder of this book. In this chapter we content ourselves with a couple of applications concerning squares.

First of all we prove a result which was asserted without proof in Exercise 13.1.

**Proposition 15.2.1** *Let  $a$  be an integer. Then  $a^2$  is divisible by 3 if and only if  $a$  is divisible by 3.*

**Constructing a proof.** The proof of ‘3 divides  $a \Rightarrow 3$  divides  $a^2$ ’ is easy.

What is required for the converse can be summarized as follows.

| Given                                 | Goal          |
|---------------------------------------|---------------|
| $a \in \mathbb{Z}$<br>3 divides $a^2$ | 3 divides $a$ |

If we try a direct argument we get something like: since 3 divides  $a^2$ ,  $a^2 = 3q$  for some integer  $q$  so that  $a = 3(q/a)$ . To complete the proof we need to know that  $a$  divides  $q$  which seems to be no easier than what we are trying to prove. The problem is that it is difficult to get back to  $a$  from  $a^2$  and this suggests trying a proof by contradiction. This gives the following strategy.

| Given  | Goal          |
|--|---------------|
| $a \in \mathbb{Z}$<br>3 divides $a^2$<br>3 does not divide $a$ | contradiction |

Now the uniqueness part of the division theorem enables us to reformulate the negative statement that 3 does not divide  $a$  as the positive statement that the remainder after dividing  $a$  by 3 is 1 or 2, i.e.  $a = 3q+1$  or  $a = 3q+2$  for some  $q \in \mathbb{Z}$ . From this it easily follows that  $a^2$  is not divisible by 3 giving the required contradiction.

It may be useful to formulate the way in which we are using the uniqueness part of the division theorem as an explicit corollary.

**Corollary 15.2.2** *Let  $a$  and  $b$  be integers with  $b > 0$  and suppose that  $a = bq + r$  for integers  $q$  and  $r$  with  $0 \leq r < b$ . Then  $b$  divides  $a$  if and only if  $r = 0$  (or equivalently,  $b$  does not divide  $a$  if and only if  $r > 0$ ).*

**Constructing a proof.** The point here is that if  $b$  divides  $a$  then, for some integer  $q_1$ ,  $a = bq_1 = bq_1 + r_1$  with  $r_1 = 0$ . So if  $a = bq + r$  with  $0 < r < q$  then  $(q, r)$  and  $(q_1, 0)$  would be two distinct solutions to the division problem, contradicting the uniqueness part of the division theorem.

*Proof* This follows immediately from the uniqueness part of the division theorem, for  $b$  divides  $a$  if and only if  $a = bq$  for some integer  $q$  if and only if  $r = 0$ . □

*Proof of Proposition 15.2.1* First of all notice that, if  $a$  is divisible by 3, then  $a = 3q$  and so  $a^2 = 9q^2 = 3(3q^2)$  is divisible by 3.

For the converse suppose that  $a^2$  is divisible by 3 and, for contradiction, that  $a$  is not divisible by 3. Then, by the division theorem,  $a = 3q + 1$  or  $a = 3q + 2$ . However,

$$a = 3q + 1 \Rightarrow a^2 = (3q + 1)^2 = 9q^2 + 6q + 1 = 3(3q^2 + 2q) + 1,$$

and similarly

$$a = 3q + 2 \Rightarrow a^2 = 3(3q^2 + 4q + 1) + 1.$$

Hence in either case the remainder is non-zero and so, by the division theorem,  $a^2$  is not divisible by 3, giving the required contradiction. Hence, if  $a^2$  is divisible by 3 then so is  $a$  as required.  $\square$

If we examine the above proof we see that it contains the demonstration that, for all integers  $a$ , the integer  $a^2$  has remainder 0 or 1 after division by 3, i.e. it never has remainder 2. We can use this fact to recognize that an integer is not a square.

**Proposition 15.2.3** *If  $n \in \mathbb{Z}^+$  is a perfect square (i.e. the square of an integer) then  $n = 3q$  or  $n = 3q + 1$  for some  $q \in \mathbb{Z}$ .*

*Proof* If  $n$  is a perfect square then  $n = a^2$  for some  $a \in \mathbb{Z}$ . By the division theorem  $a = 3q_0$  or  $a = 3q_0 + 1$  or  $a = 3q_0 + 2$  for some  $q_0 \in \mathbb{Z}$ . Hence  $a^2 = 3(3q_0^2)$  or  $a^2 = 3(3q_0^2 + 2q_0) + 1$  or  $a^2 = 3(3q_0^2 + 4q_0 + 1) + 1$  so that  $n = 3q$  or  $n = 3q + 1$  as required.  $\square$

**Example 15.2.4** The integer 1111111111 is not the square of an integer. For  $1111111111 = 3 \times 3703703703 + 2$  so that 1111111111 has remainder 2 after division by 3. Hence by the proposition it is not a perfect square.

### Exercises

**15.1** For  $a$  and  $b$  as follows write down  $q$  and  $r$  such that  $a = bq + r$  and  $0 \leq r < b$ .

- (i)  $a = 100, b = 3;$
- (ii)  $a = 3, b = 100;$
- (iii)  $a = 100, b = 7;$
- (iv)  $a = -100, b = 7;$

- (v)  $a = -105, b = 7;$
- (vi)  $a = -3, b = 105;$
- (vii)  $a = 7684, b = 4148;$
- (viii)  $a = -7684, b = 4148;$
- (ix)  $a = 1234567, b = 1357;$
- (x)  $a = 0, b = 17.$

**15.2** Prove that an integer  $a$  is divisible by 5 if and only if  $a^2$  is divisible by 5.

**15.3** Prove that, for integers  $a$ ,  $a^2$  is divisible by 3 if and only if it is divisible by 9.

**15.4** Prove that 98765432 is not the square of an integer.

**15.5** Prove that if  $n$  is a perfect square then  $n = 4q$  or  $n = 4q + 1$  for some  $q \in \mathbb{Z}$ . Deduce that 1234567 is not a perfect square.

**15.6** By repeated use of the division theorem find the infinite decimal representing the rational number  $5/7$ .

# 16

## The Euclidean algorithm

In this chapter we return to the idea of the greatest common divisor introduced in Chapter 11, introducing an extremely efficient way of calculating this number known as the Euclidean algorithm. As well as being an efficient method of calculation this method has quite far-reaching consequences some of which will be considered in subsequent chapters. This approach can be applied in several other situations in algebra and number theory which the reader will meet if these areas are explored further.

### 16.1 Finding the greatest common divisor

First of all recall the following definition.

**Definition 11.3.1** *Let  $a$  and  $b$  be two integers, at least one of which is non-zero. Then the greatest common divisor of  $a$  and  $b$  is the unique positive integer  $d$  such that*

- (i)  *$d$  is a common divisor, i.e.  $d$  divides  $a$  and  $d$  divides  $b$ ,*
- (ii)  *$d$  is greater than every other common divisor:*

$$c \text{ divides } a \text{ and } c \text{ divides } b \Rightarrow c \leq d.$$

We denote the greatest common divisor of  $a$  and  $b$  by  $\gcd(a, b)$  or simply by  $(a, b)$ .

In Chapter 11 we showed how to find the greatest common divisor by finding all the common divisors. This method is extremely tedious if the numbers are at all large. A more efficient process, involving repeated

application of the division theorem, is given in Book Seven of the *Elements* of Euclid and was discovered independently in China. It is now known as *the Euclidean algorithm*.

To begin with notice that  $d$  divides  $a$  if and only if  $d$  divides  $-a$  and so it is sufficient to deal with the greatest common factor of two non-negative integers. Furthermore, if  $a$  is positive then  $\gcd(a, 0) = a$  and so we may as well restrict attention to strictly positive integers. Thus we consider how to find the greatest common divisor of two positive integers  $a$  and  $b$ .

The method is based on two key observations. The first of these is very simple.

**Lemma 16.1.1** *If a positive integer  $b$  divides  $a$  then  $b$  must be the greatest common divisor of  $a$  and  $b$ , i.e.  $\gcd(a, b) = b$ .*

*Proof* If  $b$  divides  $a$ , then  $b$  is a common divisor since it clearly also divides  $b$ . However, nothing larger than  $b$  can divide  $b$  and so it must be the greatest common divisor.  $\square$

On the other hand, if  $b$  does not divide  $a$ , then the division theorem enables us to write

$$a = bq + r \quad \text{where } 0 < r < b.$$

The second key observation is the following result.

**Lemma 16.1.2** *For non-zero integers  $a$  and  $b$  suppose that*

$$a = bq + r \quad \text{where } q, r \in \mathbb{Z}.$$

*Then  $\gcd(a, b) = \gcd(b, r)$ .*

Before considering the proof of this result let us consider an example of how these two observations may be used to find greatest common divisors.

**Example 16.1.3** To find the greatest common divisor of 72 and 30 (Example 11.3.3 again).

We begin by dividing 30 into 72 which gives

$$72 = 30 \times 2 + 12$$

and so, by Lemma 16.1.2,  $\boxed{\gcd(72, 30) = \gcd(30, 12)}$ .

Now we can repeat the process dividing 12 into 30:

$$30 = 12 \times 2 + 6$$

giving  $(30, 12) = (12, 6)$ .

Finally, since 6 divides 12, we have  $(12, 6) = 6$  by Lemma 16.1.1.

Putting all this together we get

$$(72, 30) = (30, 12) = (12, 6) = 6$$

and so the greatest common divisor of 72 and 30 is 6.

This calculation is an example of the Euclidean algorithm in action. The Euclidean algorithm is stated below as Theorem 16.2.1. But before coming to this we prove Lemma 16.1.2.

### Constructing a proof of Lemma 16.1.2.

| Given        | Goal                      |
|--------------|---------------------------|
| $a = bq + r$ | $\gcd(a, b) = \gcd(b, r)$ |

The difficulty here is that the definition of greatest common divisor is quite complicated and does not give  $(a, b)$  as a simple formula in terms of  $a$  and  $b$ . In fact this is one of those cases when it easier to prove something slightly stronger than needed. Recall that

$$D(a, b) = \{ c \in \mathbb{Z} \mid c \text{ divides } a \text{ and } c \text{ divides } b \},$$

the set of common divisors of  $a$  and  $b$ , and similarly

$$D(b, r) = \{ c \in \mathbb{Z} \mid c \text{ divides } b \text{ and } c \text{ divides } r \},$$

the set of common divisors of  $b$  and  $r$ . If we can prove that  $D(a, b) = D(b, r)$  then we are home because  $\gcd(a, b)$  is the greatest element of the first set and  $\gcd(b, r)$  is the greatest element of the second set.

But this is easy. Recall that proving two sets equal amounts to proving that each is a subset of the other, in this case  $D(a, b) \subseteq D(b, r)$  and  $D(b, r) \subseteq D(a, b)$ . To show that  $D(a, b) \subseteq D(b, r)$  we need to prove that  $c \in D(a, b) \Rightarrow c \in D(b, r)$ . We can summarize what this involves as follows.

| Given                  | Goal                   |
|------------------------|------------------------|
| $a = bq + r$           |                        |
| $c \in \mathbb{Z}$     |                        |
| $c \text{ divides } a$ | $c \text{ divides } b$ |
| $c \text{ divides } b$ | $c \text{ divides } r$ |

One of the goals,  $c$  divides  $b$ , is already given. For the other we need to remember the definition of ‘divides’.

| Given        | Goal                               |
|--------------|------------------------------------|
| $a = bq + r$ |                                    |
| $a = q_1 c$  |                                    |
| $b = q_2 c$  | $\exists k \in \mathbb{Z}, r = kc$ |

This is immediate since  $r = a - bq = q_1 c - q_2 cq = (q_1 - q_2 q)c$  and so we can take  $k = q_1 - q_2 q$ .

The proof that  $D(b, r) \subseteq D(a, b)$  is almost identical (see Exercise 16.4).

When we come to write out a formal proof it is quite short.

*Proof* If  $c$  divides  $a$  and  $c$  divides  $b$  then  $a = cq_1$  and  $b = cq_2$  for some integers  $q_1$  and  $q_2$ . In this case  $r = a - bq = q_1 c - cq_2 q = c(q_1 - q_2 q)$  so that  $c$  divides  $r$ . Thus  $c$  is a common divisor of  $a$  and  $b$  implies that  $c$  is a common divisor of  $b$  and  $r$ . Conversely, if  $c$  is a common divisor of  $b$  and  $r$  then  $c$  is a common divisor of  $a$  and  $b$ . Hence the common divisors of  $a$  and  $b$  are the same as the common divisors of  $b$  and  $r$  and so their greatest common divisors are the same.  $\square$

We have now justified the calculation in Example 16.1.3. So we go on to describe the method in general as an algorithm.

## 16.2 The Euclidean algorithm

**Theorem 16.2.1 (The Euclidean algorithm)** Suppose that  $a$  and  $b$  are positive integers. The following procedure defines a finite sequence of positive integers  $a_0, a_1, \dots, a_n$  such that  $a_n = (a, b)$ , the greatest common divisor of  $a$  and  $b$ .

Put  $a_0 = a, a_1 = b$ .

Step  $k$ : For  $k \geq 1$ , suppose that  $a_0, a_1, \dots, a_k$  have been defined. Using the division theorem write

$$a_{k-1} = a_k q_k + r_k \quad \text{where } q_k, r_k \in \mathbb{Z} \text{ and } 0 \leq r_k < a_k.$$

If  $r_k = 0$ , stop.

Otherwise, put  $a_{k+1} = r_k$  and continue with step  $k + 1$ .

**Remarks 16.2.2** An *algorithm*† is a sequence of steps necessarily leading to a desired conclusion. Notice that since  $b = a_1 > a_2 > \dots > a_n > 0$  the Euclidean algorithm necessarily terminates after at most  $b$  steps. If there were examples where the number of steps was as many as this then it would be a poor algorithm. However, it is known‡ that the number of steps required to find the greatest common divisor of  $a$  and  $b$  using the Euclidean algorithm does not exceed five times the number of decimal digits in the smaller of  $a$  and  $b$ .

Numerical algorithms are very well suited to being programmed on a computer. We can formulate the Euclidean algorithm as a BASIC program as follows.

```

10      INPUT "FIRST NUMBER =", A
20      INPUT "SECOND NUMBER =", B
30      Q=INT(A/B)
40      R=A-B*Q
50      IF R > 0 THEN GOTO 60 ELSE GOTO 90
60      A=B
70      B=R
80      GOTO 30
90      PRINT "GCD =" B

```

Notice that BASIC has within it a command INT which gives the quotient in the division theorem. If we did not have this command then it would be easy to write a subprogram to achieve that (see Problems IV, Question 9). Even if you are unfamiliar with programming you should be able to work through this program. When you run it the first two lines ask you to input the two numbers whose greatest common divisor is sought. Then lines 30 to 50 describe a single step of the algorithm. Line 50 examines whether the last remainder is zero in order to decide whether the algorithm is complete: if it is then the program goes to line 90 and prints out the answer; if it is not then lines 70 and 80 redefine the numbers A and B before applying the division algorithm again in order to

† This word derives from the name of the ninth century arabic scholar Mohammed ibn-Musa al-Khowarizmi whose two books on arithmetic and algebra were very influential in the spread of the Hindu–Arabic numeral system to Europe. The word ‘algebra’ also comes from this same source, from the title of his book *Al-jabr wa'l muqâbalah* through which arabic algebra passed to Europe (see C.B. Boyer and U.C. Merzbach, *A history of mathematics*, Wiley, Second edition 1989).

‡ This result was proved by Gabriel Lamé in 1845 (see Problems IV, Question 8).

obtain the next remainder. Rather than using all those subscripts that appear in the statement of the theorem we change the meaning of A, B, Q and R with each step of the algorithm. For example, in line 60 'A=B' means 'define A to be the number B'. As a result the previous meaning of A is forgotten – this is a special use of the symbol '=' in computer programs†.

In terms of this program the theorem states simply that if the program is run then on inputting any two positive numbers the program will terminate by printing the greatest common divisor.

Notice that Example 11.3.3 illustrates an alternative algorithm for the greatest common divisor: simply try all the positive integers less than or equal to the smaller of  $a$  and  $b$  in turn. However, this is essentially useless for all but the smallest numbers: imagine using it for two six figure numbers. The Euclidean algorithm can deal with six figure numbers quite easily by hand (see Exercise 16.2).

**Example 16.2.3** To find the greatest common divisor of 232 and 136.

The Euclidean algorithm gives the following.

$$\begin{aligned} \text{Step 1: } 232 &= 136 \times 1 + 96 \\ \text{Step 2: } 136 &= 96 \times 1 + 40 \\ \text{Step 3: } 96 &= 40 \times 2 + 16 \\ \text{Step 4: } 40 &= 16 \times 2 + 8 \\ \text{Step 5: } 16 &= 8 \times 2 + 0 \end{aligned}$$

Hence  $(232, 136) = 8$ .

In this case  $a_0 = 232$ ,  $a_1 = 136$ ,  $a_2 = 96$ ,  $a_3 = 40$ ,  $a_4 = 16$ ,  $a_5 = 8$ .

In practice when we are applying the division theorem in these calculations we can use a calculator to find an approximate value for  $a/b$ . Then the quotient  $q$  is the greatest integer less than or equal to  $a/b$  and  $r = a - bq$  as illustrated in Examples 15.1.2 and 15.1.3.

Observe that it is sufficient to write down simply the numbers  $a_k$  in a column. At each stage we simply subtract enough multiples of the last number written down from the previous one in order to get a smaller number. Thus for the above example we could write the following (where the information in brackets is there simply to indicate what has been

† It is even correct to write 'A=A+1' in a computer program, meaning that the value of A should be increased by 1. It is not correct to write 'put  $a = a + 1$ ' when writing mathematics – although such barbarisms are sometimes used.

done).

$$\begin{array}{r}
 232 \\
 136 \\
 96 \quad (= 232 - 136 \times 1) \\
 40 \quad (= 136 - 96 \times 1) \\
 16 \quad (= 96 - 40 \times 2) \\
 8 \quad (= 40 - 16 \times 2) \\
 0 \quad (= 16 - 8 \times 2)
 \end{array}$$

Or more briefly we could write

$$\begin{array}{r}
 232 \\
 136 \quad (-1) \\
 96 \quad (-1) \\
 40 \quad (-2) \\
 16 \quad (-2) \\
 8 \quad (-2) \\
 0
 \end{array}$$

where the numbers in brackets indicate what multiple of each number is subtracted from the one above in order to obtain the next number – namely the quotient when you divide that number into the one above.

If we do write out the calculation in this way, as we will in this book, it is important to realize that it may be meaningless to a reader who has not had the conventions explained. In almost all mathematical writing there is a lot of convention, if only about the meanings of symbols, and Humpty Dumpty's principle applies here. In a case like this it is useful to at least indicate what we are doing, say by prefacing the calculation by stating 'Applying the Euclidean algorithm gives the following sequence of remainders.'

**Constructing a proof of Theorem 16.2.1.** The above example illustrates the general proof. Notice how repeated use of Lemma 16.1.2 implies that 8 is the greatest common divisor:

$$(232, 136) = (136, 96) = (96, 40) = (40, 16) = (16, 8) = 8.$$

Turning to the general case, first of all notice that repeating the steps in the theorem does lead to a *finite* sequence of positive integers  $a_0, a_1, \dots, a_n$  where  $n \leq b$  since  $a_{k+1} < a_k$ .

Now, for  $1 \leq k < n$ ,  $a_{k-1} = a_k q_k + a_{k+1}$ . Hence, just as in the example, by Lemma 16.1.2,  $(a_{k-1}, a_k) = (a_k, a_{k+1})$ . Furthermore, since

$a_{n-1} = a_n q_n$ , it follows (Lemma 16.1.1) that  $(a_{n-1}, a_n) = a_n$ . Hence  $(a, b) = (a_0, a_1) = (a_1, a_2) = \dots = (a_{n-1}, a_n) = a_n$ .

The dots in this proof indicating ‘and so on’ should alert us to the fact that strictly speaking this proof is making use of the induction principle. The style of presentation used in the above paragraph is normally considered acceptable but we could write this out as a formal proof by induction as follows.

*Proof of Theorem 16.2.1.* Let  $P(n)$  be the statement that, given any sequence of positive integers numbers  $a_0, a_1, \dots, a_n$  generated by  $n - 1$  steps of the Euclidean algorithm (but not necessarily terminating after  $n$  steps),  $(a_0, a_1) = (a_{n-1}, a_n)$ .

We prove that  $P(n)$  is true for all positive integers  $n$  by induction on  $n$ . Base case:  $P(1)$  simply says  $(a_0, a_1) = (a_0, a_1)$  which is certainly true! Inductive step: Suppose now that for some  $k \geq 1$  the result is true for  $n = k$ . Then, given a sequence of numbers  $a_0, a_1, \dots, a_{k+1}$  generated by  $k$  steps of the Euclidean algorithm,  $a_{k-1} = a_k q_k + a_{k+1}$  and so by Lemma 16.1.2  $(a_{k-1}, a_k) = (a_k, a_{k+1})$ . But, by inductive hypothesis,  $(a_{k-1}, a_k) = (a_0, a_1)$  and so  $(a_k, a_{k+1}) = (a_0, a_1)$  as required to prove the result for  $n = k + 1$ .

Hence, by induction the statement  $P(n)$  is true for all positive integers  $n$ .

Now to complete the proof, using the notation of the theorem it follows that  $(a, b) = (a_0, a_1) = (a_{n-1}, a_n)$ . But since  $r_n = 0$  it follows that  $a_n$  divides  $a_{n-1}$  and so  $(a_{n-1}, a_n) = a_n$ , i.e.  $(a, b) = a_n$ .  $\square$

### Exercises

**16.1** Use the Euclidean algorithm to find the greatest common divisor of 7684 and 4148.

**16.2** Use the Euclidean algorithm to find the greatest common divisor of 110233442 and 1102246.

**16.3** What difference does the order of the two numbers  $a$  and  $b$  make when we calculate their greatest common divisor using the Euclidean algorithm?

**16.4** Suppose that  $a, b, q$  and  $r$  are integers such that  $a = bq + r$ . Prove that  $D(b, r) \subseteq D(a, b)$ .

# 17

## Consequences of the Euclidean algorithm

In the previous chapter it was explained how the Euclidean algorithm provides an extremely efficient method for calculating the greatest common divisor of two integers. However, it does much more than this. In particular it enables us to see how to write the greatest common divisor in terms of the numbers we started with as an *integral linear combination*. We will see that this result is of both numerical and theoretical interest.

### 17.1 Integral linear combinations

**Theorem 17.1.1** *Let  $a, b$  be integers at least one of which is non-zero with greatest common divisor  $\gcd(a, b)$ . Then there exist integers  $m$  and  $n$  such that*

$$\gcd(a, b) = am + bn.$$

**Definition 17.1.2** *Given integers  $a$  and  $b$ , we say that an integer  $c$  is an integral linear combination of  $a$  and  $b$  if there exist integers  $m$  and  $n$  such that*

$$c = am + bn.$$

Thus Theorem 17.1.1 asserts that the greatest common divisor of two integers is an integral linear combination of the two integers.

Before giving the general proof we illustrate the result by considering an example.

**Example 17.1.3** Consider Example 16.2.3 in which it was shown that  $\gcd(232, 136) = 8$ . The full details of the Euclidean algorithm were as follows.

$$\begin{aligned} \text{Step 1: } 232 &= 136 \times 1 + 96 \\ \text{Step 2: } 136 &= 96 \times 1 + 40 \\ \text{Step 3: } 96 &= 40 \times 2 + 16 \\ \text{Step 4: } 40 &= 16 \times 2 + 8 \\ \text{Step 5: } 16 &= 8 \times 2 + 0 \end{aligned}$$

If we examine these equations we see that we can use them to provide an expressions for each term in the sequence of remainders generated by the algorithm as a linear combination of the previous two terms.

$$\begin{aligned} \text{Step 1: } 96 &= 232 \times 1 - 136 \times 1 \\ \text{Step 2: } 40 &= 136 \times 1 - 96 \times 1 \\ \text{Step 3: } 16 &= 96 \times 1 - 40 \times 2 \\ \text{Step 4: } 8 &= 40 \times 1 - 16 \times 2 \end{aligned}$$

Now if we substitute each of these equations in the next we get an expression for 8 as a linear combination of 232 and 136.

There are two ways of organizing this calculation.

**Method 1:** The usual method is to work up through the equations as follows.

$$\begin{aligned} 8 &= 40 \times 1 - 16 \times 2 \text{ (by Step 4)} \\ &= 40 \times 1 - (96 \times 1 - 40 \times 2) \times 2 \text{ (by Step 3)} \\ &= 96 \times (-2) + 40 \times 5 \text{ (collecting up terms)} \\ &= 96 \times (-2) + (136 \times 1 - 96 \times 1) \times 5 \text{ (by Step 2)} \\ &= 136 \times 5 + 96 \times (-7) \\ &= 136 \times 5 + (232 \times 1 - 136 \times 1) \times (-7) \text{ (by Step 1)} \\ &= 232 \times (-7) + 136 \times 12 \end{aligned}$$

Personally I find this method rather cumbersome. Whatever method is used care with signs is required but I find that slips are less likely using a second method.

**Method 2:** This method is to write each of the numbers generated by the Euclidean algorithm in turn as an integral linear combination of the

## Chapter 17: Consequences of the Euclidean algorithm 209

required form working down through the above equations. The numbers we start with, in this case 232 and 136, are easy to deal with since

$$232 = 232 \times 1 + 136 \times 0$$

and

$$136 = 232 \times 0 + 136 \times 1.$$

Step 1 of the Euclidean algorithm gives the following.

$$\begin{aligned} 96 &= 232 - 136 \\ &= 232 \times 1 + 136 \times (-1). \end{aligned}$$

Step 2 gives the following.

$$\begin{aligned} 40 &= 136 - 96 \\ &= (232 \times 0 + 136 \times 1) - (232 \times 1 + 136 \times (-1)) \\ &= 232 \times (0 - 1) + 136(1 - (-1)) \\ &= 232 \times (-1) + 136 \times 2. \end{aligned}$$

And continuing in this way we obtain the following sequence of equations.

$$\begin{array}{llll} 232 &= 232 \times 1 &+ 136 \times 0 \\ 136 &= 232 \times 0 &+ 136 \times 1 &(-1) \\ \text{Step 1: } 96 &= 232 \times 1 &+ 136 \times (-1) &(-1) \\ \text{Step 2: } 40 &= 232 \times (-1) + 136 \times 2 &&(-2) \\ \text{Step 3: } 16 &= 232 \times 3 &+ 136 \times (-5) &(-2) \\ \text{Step 4: } 8 &= 232 \times (-7) + 136 \times 12 &&(-2) \\ \text{Step 5: } 0 &&& \end{array}$$

Hence, an expression as required is  $8 = 232 \times (-7) + 136 \times 12$ . We stop the calculation at Step 4 because we have obtained what we want.

Remember that the numbers in brackets tell what multiple of each line must be subtracted from the previous line in order to get the next. If we set out the calculation in this way then at each step we do just the same calculation with the multipliers of 232 and 136. So for example, for Step 2, we simply get  $0 - 1 = -1$  as the multiplier of 232 and  $1 - (-1) = 2$  as the multiplier of 136; and, for Step 3,  $1 - (-1) \times 2 = 3$  and  $-1 - 2 \times 2 = -5$ .

Just as the first column of figures was all that it was necessary to write down when finding the greatest common divisor, this list of equations is all that is needed when finding the integral linear combination coefficients.

Notice that you can readily check whether you have successfully manoeuvred yourself through the calculation by checking that the values you have obtained for  $m$  and  $n$  actually work. In this case  $232 \times (-7) + 136 \times 12 = -1624 + 1632 = 8$ . This illustrates the fact that the best way of proving that you have a solution to an equation is by substituting in the values you have found and confirming that they do satisfy the equation.

In this book I will use the second method. Whatever method you use it is important when doing numerical examples always to check that your solutions work since numerical slips are so easy.

**Constructing a proof of Theorem 17.1.1** The second method of organizing the above calculation can be used to prove that each of the numbers in the sequence  $a_0, a_1, \dots, a_n$  generated by the Euclidean algorithm (and in particular the greatest common divisor) is an integral linear combination of  $a$  and  $b$ .

As in the proof of the Euclidean algorithm itself the proof is by induction on the length of a sequence of numbers generated by the Euclidean algorithm.

*Proof of Theorem 17.1.1* We first of all prove the result in the case when  $a$  and  $b$  are non-negative.

We prove by induction on  $n$  that a sequence of positive integers  $a_0, a_1, \dots, a_n$  generated by the first  $n - 1$  steps of the Euclidean algorithm are integral linear combinations of  $a$  and  $b$ .

Base case: The numbers  $a_0$  and  $a_1$  have this form since  $a_0 = a = a \times 1 + b \times 0$  and  $a_1 = b = a \times 0 + b \times 1$ .

Inductive step: Now suppose as inductive hypothesis that, for some  $k \geq 1$ , the result holds for  $n = k$  so that, for  $0 \leq i \leq k$ ,  $a_i = am_i + bn_i$  where  $m_i, n_i \in \mathbb{Z}$ . To prove the result for  $n = k + 1$  observe that

$$\begin{aligned} a_{k+1} &= a_{k-1} - a_k q_k \quad \text{using the notation of Theorem 16.2.1} \\ &= (am_{k-1} + bn_{k-1}) - (am_k + bn_k)q_k \\ &\quad \text{by inductive hypothesis} \\ &= a(m_{k-1} - m_k q_k) + b(n_{k-1} - n_k q_k) \end{aligned}$$

expressing  $a_{k+1}$  as an integral linear combination of  $a$  and  $b$  and completing the inductive step.

Hence by induction the result is true for all positive integers  $n$ .

In particular the greatest common divisor, the final term in the com-

plete sequence of numbers generated by the Euclidean algorithm, is an integral linear combination of  $a$  and  $b$ .

To obtain the general result observe for example that if  $a$  is negative and  $b$  non-negative then by the above argument, since  $-a$  is positive, there are integers  $m_1$  and  $n_1$  such that

$$(-a)m_1 + bn_1 = \gcd(-a, b) = \gcd(a, b).$$

This means that the required integers are given by  $m = -m_1$ ,  $n = n_1$ . Similar methods work for  $b$  negative or  $a$  and  $b$  both negative.  $\square$

## 17.2 An alternative definition of the greatest common divisor

The definition of greatest common factor which was introduced in Chapter 11 is not entirely standard. It reads (essentially) as follows.

**Definition 11.3.1** *Let  $a$  and  $b$  be two integers, at least one of which is non-zero. Then the greatest common divisor of  $a$  and  $b$  is the unique positive integer  $d$  such that*

- (i)  *$d$  is a common divisor, i.e.  $d$  divides  $a$  and  $d$  divides  $b$ ,*
- (ii)  *$d$  is greater than every other common divisor:*

$$c \text{ divides } a \text{ and } c \text{ divides } b \Rightarrow c \leq d.$$

We denote the greatest common divisor of  $a$  and  $b$  by  $\gcd(a, b)$  or sometimes  $\gcd(a, b)$ .

It is usual to have the following stronger condition in place of (ii) in the above:

- (ii)'  *$d$  is a multiple of any other common divisor:*

$$c \text{ divides } a \text{ and } c \text{ divides } b \Rightarrow c \text{ divides } d.$$

This definition is not used in this book because it is not immediately obvious from it that two integers will have a greatest common divisor. It also seems simplest to have a definition which reflects as closely as possible what the words ‘greatest common divisor’ might reasonably be expected to mean!

In fact these two definitions are equivalent since the apparently more stringent condition (ii)' is automatically satisfied by a number  $d$  which satisfies our definition of a greatest common divisor. This is a simple consequence of Theorem 17.1.1 as follows.

**Corollary 17.2.1** Let  $d = \gcd(a, b)$  be the greatest common divisor of the integers  $a$  and  $b$ . Then  $c$  is a common divisor of  $a$  and  $b$  if and only if  $c$  divides  $d$ , i.e.  $D(a, b) = D(d)$ .

**Example 17.2.2** Recall from Example 11.3.3 that the set of common divisors of 30 and 72 is

$$D(30, 72) = \{-6, -3, -2, -1, 1, 2, 3, 6\}.$$

This is precisely  $D(6)$ , the set of divisors of the greatest common divisor  $(30, 72) = 6$ .

*Proof of Corollary 17.2.1* Notice that the statement ‘if and only if’ indicates that two separate results must be proved.

Proof of ‘if’: Suppose that  $c$  divides  $d$ . By definition  $d$  divides  $a$  and  $d$  divides  $b$ . But  $c$  divides  $d$  which divides  $a$  immediately implies from the definition of ‘divides’ that  $c$  divides  $a$  (see Exercise 3.2) and in the same way  $c$  divides  $b$  so that  $c$  is a common divisor as required.

Proof of ‘only if’: By Theorem 17.1.1,  $d = am + bn$  for some  $m, n \in \mathbb{Z}$ . Now if  $c$  is a common divisor of  $a$  and  $b$  then  $c$  divides  $a$  and  $c$  divides  $b$  so that  $a = cq_1$  and  $b = cq_2$  for some integers  $q_1$  and  $q_2$ . In this case  $d = (cq_1)m + (cq_2)n = c(q_1m + q_2n)$  so that  $c$  divides  $d$  as required.  $\square$

### 17.3 Coprime pairs

Recall the following definition.

**Definition 11.3.2** Two integers  $a$  and  $b$ , not both zero, are called coprime when  $(a, b) = 1$ , in other words their only common factors are 1 and  $-1$ .

**Proposition 17.3.1** Two non-zero integers  $a$  and  $b$  are coprime if and only if there exist integers  $m$  and  $n$  such that

$$am + bn = 1.$$

**Constructing a proof.** The necessity of the existence of  $m$  and  $n$  is a special case of Theorem 17.1.1.

In the other direction we are starting from an existence statement. The simplest way to use such a statement is to take particular elements which satisfy the predicate. So in this case we use integers  $m_0$  and  $n_0$

## Chapter 17: Consequences of the Euclidean algorithm 213

such that  $am_0 + bn_0 = 1$  – we are supposing that such integers exist. It was explained in Chapter 7 that it is usual to use the same symbols as in the existence statement for specific elements for which the predicate is true, in this case  $m$  and  $n$  rather than the new symbols  $m_0$  and  $n_0$ , but here it seems clearer to use different symbols for specific elements.

This leads to the following summary of what is required.

| Given                                    | Goal         |
|--|--------------|
| integers $m_0, n_0$<br>$am_0 + bn_0 = 1$ | $(a, b) = 1$ |

To proceed we need to unpack the definition of coprime. To show that  $(a, b) = 1$  we must show that the only common divisors of  $a$  and  $b$  are  $\pm 1$ , i.e.

$$c \text{ divides } a \text{ and } c \text{ divides } b \Rightarrow c = \pm 1 \text{ for integers } c.$$

If we now try the direct method of proof this gives the following strategy.

| Given  | Goal        |
|--|-------------|
| $m_0, n_0 \in \mathbb{Z}$<br>$am_0 + bn_0 = 1$<br>$c \in \mathbb{Z}$<br>$c \text{ divides } a$<br>$c \text{ divides } b$ | $c = \pm 1$ |

Having got this far we can move on to the formal proof which is constructed easily by spelling out the definition of ‘divides’.

*Proof* If  $a$  and  $b$  are coprime then the existence of integers  $m$  and  $n$  as required is a special case of Theorem 17.1.1.

For the converse suppose that  $m_0$  and  $n_0$  are integers such that  $am_0 + bn_0 = 1$ . Suppose that  $c$  is a common divisor of  $a$  and  $b$ . Then, since  $c$  divides  $a$ ,  $a = cq_1$  for some  $q_1 \in \mathbb{Z}$  and, since  $c$  divides  $b$ ,  $b = cq_2$  for some  $q_2 \in \mathbb{Z}$ . Hence

$$1 = am_0 + bn_0 = cq_1m_0 + cq_2n_0 = c(q_1m_0 + q_2n_0).$$

Thus  $c$  divides 1 which implies that  $c = \pm 1$ . Thus the only common divisors of  $a$  and  $b$  are  $\pm 1$  and so the greatest common divisor is 1, as required to prove that  $a$  and  $b$  are coprime.  $\square$

**Theorem 17.3.2** Suppose that  $a, b$  and  $c$  are positive integers with  $a$

and  $b$  coprime. Then

$$a \text{ divides } bc \Rightarrow a \text{ divides } c.$$

**Constructing a proof.** This innocuous appearing result is the key to many basic results in number theory. It is surprisingly difficult to prove. Although in this book I am attempting to demonstrate that many proofs arise naturally from the definitions, this is not the case with all proofs: some proofs are just clever and although we can suggest that the steps arise naturally this is only with the benefit of hindsight. Although this particular proof is not difficult to understand it is rather ingenious: an example of a proof which we might be surprised that anyone was able to think of in the first place!

What we are looking for can be summarized as follows.

| Given                                   | Goal                   |
|---|------------------------|
| $(a, b) = 1$<br>$a \text{ divides } bc$ | $a \text{ divides } c$ |

It is not clear how to proceed. So, as always in such a situation, we go back to the definitions to simplify the statements involved. The ‘divides’ statements are simple enough but, as we have noted before, the definition of the greatest common divisor is rather complicated and difficult to use. It turns out (and this is the first clever step) that a good way to proceed which gives us a logically much simpler statement is to use the previous result which characterized coprime integers. Since  $a$  and  $b$  are coprime,  $am_0 + bn_0 = 1$  for some  $m_0, n_0 \in \mathbb{Z}$ . This gives us the following strategy.

| Given                          | Goal                                 |
|--------------------------------|--------------------------------------|
| $am_0 + bn_0 = 1$<br>$bc = aq$ | $c = ak$ for some $k \in \mathbb{Z}$ |

Even now it is not absolutely obvious what to do. The second clever step is to multiply the first given statement by  $c$  which, if you think about it, is the only way of getting  $c$  out of the given statements using addition or multiplication (division is no use because it may not lead to integers). This gives the following steps:

$$\begin{aligned} c &= (am_0 + bn_0)c \quad (\text{using the first given statement}) \\ &= acm_0 + bcn_0 \\ &= acm_0 + aqn_0 \quad (\text{using the second given statement}) \\ &= a(cm_0 + qn_0). \end{aligned}$$

Thus  $k = cm_0 + qn_0$  gives us the integer we need.

This remarkable proof is very short when written out formally.

*Proof* Suppose that  $a$  divides  $bc$  so that  $bc = aq$  for some  $q \in \mathbb{Z}$ . Since  $(a, b) = 1$ , by Proposition 17.3.1 there are integers  $m_0$  and  $n_0$  such that  $1 = am_0 + bn_0$ . But then  $c = (am_0 + bn_0)c = acm_0 + bcn_0 = acm_0 + aqn_0 = a(cm_0 + qn_0)$  so that  $a$  divides  $c$  as required.  $\square$

### Exercises

**17.1** From Exercise 16.1,  $\gcd(7684, 4148) = 68$ . Find integers  $m$  and  $n$  such that  $68 = 7648m + 4148n$ .

**17.2** Find integers  $m$  and  $n$  such that  $68 = 7648m - 4148n$ .

**17.3** From Exercise 16.2,  $\gcd(11033442, 1102246) = 578$ . Find integers  $m$  and  $n$  such that  $578 = 11033442m + 1102246n$ .

**17.4** Let  $\gcd(a, b, c)$  denote the greatest common divisor of the three non-zero integers  $a$ ,  $b$  and  $c$ . Use Corollary 17.2.1 to prove that

$$\gcd(a, b, c) = \gcd(\gcd(a, b), c).$$

Hence find the greatest common divisor of 11033442, 1102246 and 6035.

**17.5** Find integers  $m$ ,  $n$  and  $p$  such that

$$11033442m + 1102246n + 6035p = \gcd(11033442, 1102246, 6035).$$

**17.6** Suppose that  $a_1, a_2, b_1, b_2$  are positive integers such that  $a_1$  and  $b_1$  are coprime and  $a_2$  and  $b_2$  are coprime. Prove that

$$a_1b_2 = a_2b_1 \Rightarrow a_1 = a_2 \text{ and } b_1 = b_2.$$

# 18

## Linear diophantine equations

In this chapter we present a striking application of the Euclidean algorithm. The following very ancient problem will be solved.

**Problem 18.0.1** *Given integers  $a$ ,  $b$  and  $c$ , find all integers  $m$  and  $n$  such that*

$$am + bn = c.$$

Notice that Proposition 4.1.1 shows that for certain choices of  $a$ ,  $b$  and  $c$  there are no solutions at all since it was proved there that there are no solutions for  $a = 14$ ,  $b = 20$ ,  $c = 101$ . The first part of the solution to the problem is to give a necessary and sufficient condition for a solution to exist. It turns out that the idea used in proving Proposition 4.1.1 gives such a condition although the proof that this condition is sufficient makes use of the material on the Euclidean algorithm. The second part of the solution is to describe all the solutions when the condition is satisfied.

### 18.1 Diophantine equations

Problem 18.0.1 is an example of a *diophantine equation*. This term refers to an equation in one or more unknowns which is to be solved in the integers. The study of such problems is one of the oldest branches of mathematics and examples are to be found in extremely ancient documents. They are named after the Greek mathematician Diophantus†

† Exercise (from the fourth century): Diophantus passed one sixth of his life in childhood, one twelfth in youth, and one seventh as a bachelor. Five years after his marriage was born a son who died four years before his father, living half as long as his father. How long did Diophantus live for?

who lived in the third century A.D.; his book the *Arithmetica* is usually considered to be the high point of classical Greek number theory. Diophantus was usually content to find only one solution and did not allow negative solutions. A complete solution to Problem 18.0.1 was apparently first obtained by the Indian mathematician Brahmagupta in the seventh century.<sup>†</sup>

One of the most famous of all problems in mathematics takes the form of a diophantine equation. It was when reading Diophantus's *Arithmetica* (which had been recently rediscovered and translated into Latin by Bachet) in 1637 that Pierre de Fermat (1601–1665) made the following note (in Latin) in the margin of his copy.

On the other hand, it is impossible for a cube to be written as a sum of two cubes or a fourth power to be written as a sum of two fourth powers or, in general, for any power which is a power greater than the second to be written as a sum of two like powers. I have a truly marvellous demonstration of this proposition which this margin is too narrow to contain.<sup>‡</sup>

To understand this assertion recall that, by Pythagoras's theorem, the lengths of the sides of a right-angled triangle satisfy the equation  $x^2 + y^2 = z^2$ . It is easy to find integers which satisfy this equation, for example  $x = 3$ ,  $y = 4$ ,  $z = 5$ , or  $x = 5$ ,  $y = 12$ ,  $z = 13$ , and a complete description of all the solutions, known as Pythagorean triples, has been known since the time of Euclid. Fermat's note was written next to Book II, Problem 5 the *Arithmetica* which asks which squares can be written as the sum of two squares and so can be solved using knowledge of Pythagorean triples.

Fermat considered the equation  $x^n + y^n = z^n$  for integers  $n \geq 3$  and his note asserts that in this case there are no positive integers<sup>§</sup> which satisfy the equation. Fermat himself did provide elsewhere a proof of

<sup>†</sup> See C.B. Boyer and U.C. Merzbach, *A history of mathematics*, Wiley, Second edition 1989.

<sup>‡</sup> Various versions of this note appear in the literature. This one comes from H.M. Edwards, *Fermat's last theorem, a genetic introduction to algebraic number theory*, Springer-Verlag, 1977. The actual book containing the margin note has been lost. When Fermat's son, Samuel de Fermat, went through his father's papers he found the annotated copy of Diophantus's book and in 1670 published a new edition of the Bachet translation with Fermat's marginal notes as an appendix of 48 Observations. The 'last theorem' was Observation 2.

Another excellent and less advanced source of information about the problem is D.M. Burton, *Elementary number theory*, Allyn and Bacon, 1976.

<sup>§</sup> Actually, Diophantus worked entirely with rational numbers and so, strictly speaking, Fermat was asserting that there are no positive rational solutions. However, it is sufficient to consider only integer solutions because if one had a rational solution then an appropriate multiple of it would be an integer solution (see Problems IV, Question 19).

this result in the case that  $n = 4$ . A proof for  $n = 3$  was first given by the Swiss mathematician Leonhard Euler in 1770 (by which time at the age of 63 he had already been blind for four years); this was incomplete but Adrien-Marie Legendre (French) plugged the gap. A proof for  $n = 5$  was found around 1825 by Lejeune Dirichlet (also of French descent) and by Legendre. Since then, much modern algebra has been developed in attempting to prove Fermat's last theorem and many mathematicians have worked on it. Recently computer calculations have been used and by these means the theorem was shown to be true for values of  $n$  less than 25000. But there is no hope of proving the general theorem by checking integers  $n$  one by one.

However, in 1993 the British mathematician Andrew Wiles of Princeton University announced a proof of Fermat's last theorem. There was intense interest in this work and at the end of the year a gap was found in the proof. By October 1994 this gap had been bridged and the proof is now published.<sup>†</sup> It appears that this most famous mathematical problem has at last been solved.

Notice that Theorem 13.2.1, which asserted that there is no rational number whose square is 2, is a statement about a diophantine equation: it is equivalent to the assertion that the only solution to the diophantine equation  $m^2 = 2n^2$  is the trivial one  $m = 0, n = 0$ .

## 18.2 A condition for the existence of solutions

We now begin the discussion of Problem 18.0.1 stated at the beginning of the chapter.

For simplicity we suppose that  $a, b, c$  are positive. It is easy to extend the methods to the general case (see Exercise 18.2).

**Theorem 18.2.1** *For positive integers  $a, b$  and  $c$ , there exist integers  $m$  and  $n$  such that*

$$am + bn = c$$

<sup>†</sup> The proof appears in Andrew Wiles, Modular elliptic curves and Fermat's last theorem, *Annals of Mathematics*, second series, volume 141, Princeton University Press, 1995 (pages 443–551). This is a highly technical paper beyond the reach even of most professional mathematicians, but books have now appeared attempting to make the background to the proof and some of its ideas more widely available: A.D. Aczel, *Fermat's last theorem, unlocking the secret of an ancient mathematical problem*, Viking, 1997, and S. Singh, *Fermat's last theorem, the story of a riddle that confounded the world's greatest minds for 358 years*, Fourth Estate, 1997.

*if and only if the greatest common divisor  $\gcd(a, b)$  divides  $c$ .*

**Constructing a proof of the necessity of the divisibility condition.** It is always a good idea to think about a specific numerical example. In Proposition 4.1.1 we showed that there are no integers  $m$  and  $n$  such that  $14m + 20n = 101$  by observing that, since 2 divides 14 and 2 divides 20, for any integers  $m$  and  $n$ , 2 divides  $14m + 20n$ . In just the same way, in the general case, if  $d$  divides  $a$  and  $d$  divides  $b$  then  $d$  divides  $am + bn$  and in particular  $\gcd(a, b)$  divides  $am + bn$ . The details follow.

*Proof of the necessity of the divisibility condition* Suppose that  $m_0$  and  $n_0$  are integers such that  $am_0 + bn_0 = c$ . Let  $d = \gcd(a, b)$ . Since  $d$  divides  $a$  and  $d$  divides  $b$ ,  $a = dq_1$  and  $b = dq_2$  for some natural numbers  $q_1$  and  $q_2$ . Therefore  $c = (dq_1)m_0 + (dq_2)n_0 = d(q_1m_0 + q_2n_0)$  so that  $d$  divides  $c$  as required.  $\square$

**Constructing a proof of the sufficiency of the divisibility condition.** This is the interesting part. It is often helpful to think about one particular simple case. For this problem the easiest case of  $\gcd(a, b)$  dividing  $c$  is when  $c = \gcd(a, b)$ . In this case we need to show that there are integers  $m$  and  $n$  such that

$$am + bn = (a, b).$$

But Theorem 17.1.1 is precisely the statement that such integers exist.

Now in the general case of  $\gcd(a, b)$  dividing  $c$  we have  $c = (a, b)q$  for some  $q \in \mathbb{Z}$ . But if  $am_0 + bn_0 = (a, b)$  then  $a(m_0q) + b(n_0q) = (a, b)q = c$  and so we have the required integer solution of  $am + bn = c$ .

*Proof of the sufficiency of the divisibility condition* If  $(a, b)$  divides  $c$  then  $c = (a, b)q$  for some  $q \in \mathbb{Z}$ . By Theorem 17.1.1 we can choose integers  $m_0$  and  $n_0$  such that

$$am_0 + bn_0 = (a, b).$$

Hence

$$a(m_0q) + b(n_0q) = (a, b)q = c.$$

Thus we see that a solution is provided by  $m = m_0q$ ,  $n = n_0q$ .  $\square$

**Example 18.2.2** Prove that the diophantine equation

$$140m + 63n = 35$$

has a solution and write down one such solution.

*Solution* We apply the Euclidean algorithm to the numbers  $a = 140$ ,  $b = 63$ , obtaining the following sequence of remainders.

$$\begin{array}{rcl} 140 & & \\ 63 & (-2) & \\ 14 & (-4) & \\ 7 & (-2) & \\ 0 & & \end{array}$$

This shows that  $(140, 63) = 7$ . Since 7 divides 35 the equation has a solution.

To find a solution we write the sequence of remainders as integral linear combinations.

$$\begin{array}{rcl} 140 & = & 140 \times 1 & + 63 \times 0 \\ 63 & = & 140 \times 0 & + 63 \times 1 & (-2) \\ 14 & = & 140 \times 1 & + 63 \times (-2) & (-4) \\ 7 & = & 140 \times (-4) & + 63 \times 9 & (-2) \\ 0 & & & & \end{array}$$

This gives  $140 \times (-4) + 63 \times 9 = 7$  and so multiplying through by  $35/7 = 5$  we obtain  $140 \times (-20) + 63 \times 45 = 35$ . Thus a solution is given by  $m = -20$ ,  $n = 45$ .  $\square$

### 18.3 Finding all the solutions – the homogeneous case

So far we have seen how to decide whether the diophantine equation

$$am + bn = c \tag{18.1}$$

has a solution and, if it has, how the Euclidean algorithm enables us to find one solution. However, there is no reason to believe that there is only one solution and we now consider how we can find all the solutions to the equation. We will see that if the equation has one solution then it has infinitely many.

We start with the special case of  $c = 0$  for which the equation certainly does have a solution, namely  $m = 0$ ,  $n = 0$ .

The diophantine equation

$$am + bn = 0 \tag{18.2}$$

is called a *homogeneous* equation. Equation (18.2) is said to be the homogeneous equation *associated with* the equation (18.1) (which is called *inhomogeneous* if  $c \neq 0$ ).

The method of finding all the solutions to a homogeneous linear diophantine equation is best illustrated by means of an example. We take the homogeneous equation associated with the equation of Example 18.2.2.

**Example 18.3.1** Find all the solutions to the diophantine equation

$$140m + 63n = 0. \quad (18.3)$$

*Solution* Recall from Example 18.2.2 that  $\gcd(140, 63) = 7$ . So we can begin by dividing through by 7. Then, for  $m, n \in \mathbb{Z}$ ,

$$\begin{aligned} 140m + 63n = 0 &\Leftrightarrow 20m + 9n = 0 \\ &\Leftrightarrow 20m = -9n \\ &\Rightarrow 9 \text{ divides } 20m \end{aligned}$$

since it is clear that  $-9n = 9(-n)$  is divisible by 9. From this we can deduce that 9 divides  $m$  using Proposition 17.3.2 since 9 and 20 are coprime. [Notice that, by Exercise 11.4, it is always the case that dividing  $am + bn = 0$  by  $\gcd(a, b)$  leads to an equation with coprime coefficients.]

It now follows that  $m = 9q$  for some  $q \in \mathbb{Z}$ . Substituting this back into equation (18.3) we get  $20 \times 9q = -9n$  which gives  $n = -20q$ . So we see that the solutions to the diophantine equation are given by

$$m = 9q, \quad n = -20q \quad \text{for } q \in \mathbb{Z}.$$

There is one solution for each integer  $q$ . Notice that a solution is a pair of integers  $(m, n)$ , in other words an element of the set  $\mathbb{Z} \times \mathbb{Z} = \mathbb{Z}^2$ .

The following table gives some examples.

| $q$ | $(m, n)$  |
|-----|-----------|
| -2  | (-18, 40) |
| -1  | (-9, 20)  |
| 0   | (0, 0)    |
| 1   | (9, -20)  |
| 2   | (18, -40) |
| 3   | (27, -60) |

□

The key part of this argument can be abstracted into the following result.

**Proposition 18.3.2** *Suppose that  $a$  and  $b$  are coprime non-zero integers. Then, for  $(m, n) \in \mathbb{Z}^2$ ,*

$$am + bn = 0 \Leftrightarrow (m, n) = (bq, -aq) \text{ for some } q \in \mathbb{Z}.$$

*Proof* We use the same argument as in the example.

The proof of ' $\Leftarrow$ ' is simply by substitution:

$$(m, n) = (bq, -aq) \Rightarrow am + bn = a(bq) + b(-aq) = 0.$$

To prove ' $\Rightarrow$ ' observe that  $am + bn = 0$  implies that  $am = -bn$  so that  $b$  divides  $am$ . By invoking Proposition 17.3.2 it follows that  $b$  divides  $m$  since  $a$  and  $b$  are coprime. Hence  $m = bq$  for some integer  $q$  and substituting this back into the equation gives  $n = -aq$ , as required.  $\square$

**Example 18.3.3** Using this result we could write out the solution to Example 18.3.1 as follows.

*Solution* For  $m, n \in \mathbb{Z}$ ,

$$\begin{aligned} 140m + 63n = 0 &\Leftrightarrow 20m + 9n = 0 \\ &\Leftrightarrow (m, n) = (9q, -20q) \text{ for } q \in \mathbb{Z}, \end{aligned}$$

since 9 and 20 are coprime.  $\square$

However, when using Proposition 18.3.2, care must be taken to get the numbers  $a$  and  $b$  the correct way round and the signs correct. It is usually best to check that the solutions you are claiming actually are solutions – simply by substituting into the equation.

#### 18.4 Finding all the solutions – the general case

Now we return to the general case of equation (18.1). Once we know that such an equation has a single solution then we can write down all the solutions by solving the associated homogeneous equation (18.2). The following result sums up why this is the case.

**Proposition 18.4.1** *Suppose that  $(m, n) = (m_0, n_0) \in \mathbb{Z}^2$  is a solution to the diophantine equation  $am + bn = c$ , i.e.  $am_0 + bn_0 = c$ . Then, for*

$(m, n) \in \mathbb{Z}^2$ ,

$$am + bn = c \Leftrightarrow a(m - m_0) + b(n - n_0) = 0.$$

*Proof*

$$\begin{aligned} am + bn = c &\Leftrightarrow am + bn = am_0 + bn_0 \\ &\Leftrightarrow a(m - m_0) + b(n - n_0) = 0. \end{aligned}$$

□

Let us illustrate how this is used by considering an example.

**Example 18.4.2** Find all the solutions to the diophantine equation

$$140m + 63n = 35.$$

*Solution* In Example 18.2.2 we found the particular solution  $m = -20$ ,  $n = 45$ . Thus

$$\begin{aligned} 140m + 63n = 35 &\Leftrightarrow 140(m + 20) + 63(n - 45) = 0 \\ &\quad \text{by Proposition 18.4.1} \\ &\Leftrightarrow (m + 20, n - 45) = (9q, -20q) \text{ for some } q \in \mathbb{Z} \\ &\quad \text{by Example 18.3.1} \\ &\Leftrightarrow (m, n) = (-20 + 9q, 45 - 20q) \text{ for some } q \in \mathbb{Z}. \end{aligned}$$

The following table gives some solutions.

| $q$ | $m$ | $n$ |
|-----|-----|-----|
| -2  | -38 | 85  |
| -1  | -29 | 65  |
| 0   | -20 | 45  |
| 1   | -11 | 25  |
| 2   | -2  | 5   |
| 3   | 7   | -15 |

Notice that had we started with a different particular solution, say  $m = -2$ ,  $n = 5$ , then we would have obtained that the solutions are  $(m, n) = (-2 + 9q, 5 - 20q)$  for  $q \in \mathbb{Z}$ . But putting in some values for  $q$  illustrates that we get just the same solution set.

| $q$ | $m$ | $n$ |
|-----|-----|-----|
| -4  | -38 | 85  |
| -3  | -29 | 65  |
| -2  | -20 | 45  |
| -1  | -11 | 25  |
| 0   | -2  | 5   |
| 1   | 7   | -15 |

This problem is considered again as Example 20.2.4 where an alternative method is explained.  $\square$

### Exercises

**18.1** Solve the diophantine equation

$$7684m + 4148n = 272.$$

**18.2** Solve the diophantine equation

$$7684m - 4148n = 272.$$

**18.3** Solve the diophantine equation

$$516m + 564n = 6432.$$

Prove that there are unique positive integers satisfying this equation and find these integers.

## Problems IV: Arithmetic

1. Prove that if an integer  $n$  is the sum of two squares ( $n = a^2 + b^2$  for  $a, b \in \mathbb{Z}$ ) then  $n = 4q$  or  $n = 4q + 1$  or  $n = 4q + 2$  for some  $q \in \mathbb{Z}$ . Deduce that 1234567 cannot be written as the sum of two squares.
2. Let  $a$  be an integer. Prove that  $a^2$  is divisible by 5 if and only if  $a$  is divisible by 5.
3. Use the result of Question 2 to prove that there does not exist a rational number whose square is 5.
4. Prove that there is no rational number whose square is 98.
5. Prove that every infinite decimal representing a rational number is recurring (the converse of Theorem 13.3.4) and furthermore that if the fraction in lowest terms representing a fraction is  $a/b$  then the number of recurring digits in its decimal representation is less than  $b$ .

[See Exercise 15.6.]

6. Use the Euclidean algorithm to find the greatest common divisors of (i) 165 and 252, (ii) 4284 and 3480.
7. Let  $u_n$  be the  $n$ th Fibonacci number (for the definition see Definition 5.4.2). Prove that the Euclidean algorithm takes precisely  $n$  steps to prove that  $\gcd(u_{n+1}, u_n) = 1$ .

- 8.** Suppose that  $a$  and  $b$  are two positive integers with  $a \geq b$ . Let  $a_0, a_1, \dots, a_n$  be the sequence of integers generated by the Euclidean algorithm so that  $a_n = \gcd(a, b)$ . Prove by induction on  $k$  that  $a_{n-k} \geq u_{k+2}$  where  $u_m$  is the  $m$ th Fibonacci number. Deduce that  $b \geq u_{n+1}$ .

Hence prove Lamé's theorem: if  $b$  has  $r$  decimal digits then  $n \leq 5r$ .

[Prove that  $\log_{10} \alpha > 1/5$  where  $\alpha = (1 + \sqrt{5})/2$ .]

- 9.** Describe an algorithm to output the numbers  $q$  and  $r$  in the division theorem (Theorem 15.1.1) given positive integers  $a$  and  $b$ .

[The discussion on page 193 may help.]

- 10.** In each case of Question 6 write the greatest common divisor as an integral linear combination of the two numbers.

- 11.** Given positive integers  $a$  and  $b$ , suppose that applying the Euclidean algorithm leads to a sequence of positive integers  $a_0, a_1, \dots, a_N$ . Suppose that the procedure of the proof of Theorem 17.1.1 leads to expressions of the form

$$a_k = am_k + bn_k$$

for integers  $m_k$  and  $n_k$ . Prove that  $m_k n_{k-1} - m_{k-1} n_k = (-1)^k$  for  $1 \leq k \leq N$ . Deduce that the numbers  $m_k$  and  $n_k$  are coprime.

- 12.** With the same data as in the previous question suppose that the procedure of the proof of Theorem 17.1.1 is continued for one more step to give

$$0 = am_{N+1} + bn_{N+1}.$$

Prove that the integers  $m_{N+1}$  and  $n_{N+1}$  determine the solution set of the diophantine equation  $am + bn = 0$ :

$$am + bn = 0 \Leftrightarrow (m, n) = (m_{N+1}q, n_{N+1}q) \text{ for some } q \in \mathbb{Z}.$$

- 13.** The *least common multiple* of two non-zero integers  $a$  and  $b$  is the unique positive integer  $m$  such that

- (i)  $m$  is a common multiple, i.e.  $a$  divides  $m$  and  $b$  divides  $m$ ,
- (ii)  $m$  is less than any other common multiple:

$$a \text{ divides } n \text{ and } b \text{ divides } n \Rightarrow m \leq n.$$

We denote the least common multiple of  $a$  and  $b$  by  $[a, b]$  or  $\text{lcm}[a, b]$ .

Give a proof by contradiction that if a positive integer  $n$  is a common multiple of  $a$  and  $b$  then  $[a, b]$  divides  $n$ .

[Use the division theorem. If  $[a, b]$  does not divide  $n$  then  $n = [a, b]q + r$  where  $0 < r < [a, b]$ . Now prove that  $r$  is a common multiple of  $a$  and  $b$ .]

This means that  $ab/[a, b]$  is an integer. Prove that this integer is a common divisor of  $a$  and  $b$ . Deduce that  $ab/[a, b] \leq (a, b)$ , the greatest common divisor of  $a$  and  $b$ .

Prove that  $ab/(a, b)$  is a common multiple of  $a$  and  $b$ . Deduce that  $(a, b)[a, b] = ab$  if  $a$  and  $b$  are positive.

Hence find  $\text{lcm}[612, 696]$ .

**14.** Let  $a$  and  $b$  be positive integers. By the well-ordering principle the non-empty set of positive integers

$$\{ am + bn \mid m, n \in \mathbb{Z} \text{ and } am + bn > 0 \}$$

has a minimum element  $c$ . Prove by contradiction that  $c$  is a common divisor of  $a$  and  $b$  and hence give an alternative proof of Theorem 17.1.1.

**15.** Solve the linear diophantine equations

- (i)  $165m + 252n = 15$ ,
- (ii)  $165m + 252n = 20$ ,
- (iii)  $4284m + 3480n = 60$ ,
- (iv)  $4284m - 3480n = 36$ .

**16.** Solve the linear diophantine equation

$$336m + 238n = 5558.$$

Prove that there is a unique pair of *positive* integers  $m$  and  $n$  satisfying this equation and find this solution.

**17.** Find all positive integers which satisfy the diophantine equation

$$12563m + 6052n = 9922169.$$

**18.** Solve the linear diophantine equation

$$6m + 10n + 15p = 1.$$

- 19.** Prove that if there are no non-zero integer solutions to the equation  $x^n + y^n = z^n$  then there are no non-zero rational solutions.

[Prove the contrapositive: show how a rational solution leads to an integer solution.]

# **Part V**

## Modular arithmetic



# 19

## Congruence of integers

Throughout this chapter  $m$  will denote a fixed positive integer.

One of the most useful distinctions amongst the integers is that between even and odd numbers. We have already seen a number of mathematical arguments which make use of this. The distinction between these two types of number can be expressed in terms of the remainder when the number is divided by 2: the remainder is 0 in the case of an even number and 1 for an odd number. The idea of congruence is a generalization of this: we classify integers according to the remainder after division by some fixed positive integer, not necessarily 2.

To be more precise, suppose that  $m$  is a positive integer. Two integers  $a$  and  $b$  are *congruent modulo  $m$*  if they have the same remainder after division by  $m$ . This divides the integers into  $m$  different subsets each consisting of the set of integers having a certain remainder. These subsets are called *congruence classes* of integers and they will be investigated more fully in Chapter 21.

This idea is familiar to everyone in the measurement of time. For example when we use the days of the week we are counting days modulo 7. Most calendars list days in columns with the Sundays in the first column, the Mondays in the second and so on up to the Saturdays in the seventh column. This means that two days are in the same column if and only if they are on the same day of the week so that when we count the days successively from some fixed starting point then the numbers of the two days are congruent modulo 7. Thus the columns form the congruence classes. Many of our activities are determined by which column a day lies in, by the day of the week: think of timetables (of trains or lectures) and shop opening hours.

As an illustration, 1 January 1998 is a Thursday. If we want to find the day of the week of 6 September 1998 then we calculate what day of the year it is:  $31 \times 5 + 30 \times 2 + 28 \times 1 + 6 = 155 + 60 + 28 + 6 = 249$ . Applying the division theorem,  $249 = 7 \times 35 + 4$ . Hence 249 is congruent modulo 7 to 4 so that 6 September 1998 is the same day of the week as the fourth day of the year, 4 January 1998, namely Sunday. We will see shortly how to simplify this calculation.

### 19.1 Basic definitions

It is convenient to phrase the formal definition of congruence in a slightly different way from that described above. This formulation of the definition uses only the idea of divisibility and does not rely on the division theorem.

**Definition 19.1.1** Two integers  $a$  and  $b$  are congruent modulo  $m$  when  $a - b$  is divisible by  $m$ , i.e. there exists an integer  $q$  such that  $a - b = mq$ . We write  $a \equiv b \pmod{m}$ . If the modulus  $m$  is clear we may sometimes omit it and just write  $a \equiv b$ .

This definition (and the notation  $\equiv$ ) was introduced by the German mathematician Carl Friedrich Gauss in his book *Disquisitiones arithmeticae* published in 1801 when he was only 24. This book is considered to have laid the foundations of modern number theory. It is available in an English translation† and is beautifully written and readily understood by a modern reader.

To illustrate the definition observe that  $14 \equiv 2 \pmod{3}$  since  $14 - 2 = 12 = 3 \times 4$ , and  $14 \equiv -10 \pmod{3}$  since  $14 - (-10) = 24 = 3 \times 8$ . On the other hand,  $14 \not\equiv 1 \pmod{3}$  since  $14 - 1 = 13$  is not a multiple of 3. Given any two integers they are congruent modulo 1 since every integer is divisible by 1. Because of this it is usual to take  $m > 1$ .

Notice that  $a \equiv 0 \pmod{m}$  is another way of saying that  $m$  divides  $a$ , or equivalently  $a$  is a multiple of  $m$ .

The relation of congruence has many properties which are similar to those of equality.

† C.F. Gauss, *Disquisitiones arithmeticae* (translated by A.A. Clarke), Springer-Verlag, 1986.

**Proposition 19.1.2**

- (i) Reflexive property. *For all integers  $a$ ,  $a \equiv a \pmod{m}$ .*
- (ii) Symmetric property. *If  $a$  and  $b$  are integers such that  $a \equiv b \pmod{m}$ , then  $b \equiv a \pmod{m}$ .*
- (iii) Transitive property. *If  $a$ ,  $b$  and  $c$  are integers such that  $a \equiv b \pmod{m}$  and  $b \equiv c \pmod{m}$ , then  $a \equiv c \pmod{m}$ .*

We usually use these properties without comment.

*Proof* (i) This is immediate since  $a - a = 0 = m \times 0$ .

(ii) If  $a \equiv b \pmod{m}$ , then  $m$  divides  $(a - b)$  and so  $a - b = mq$  for some  $q \in \mathbb{Z}$ . But then  $b - a = -mq = m \times (-q)$  so that  $m$  divides  $(b - a)$ , i.e.  $b \equiv a \pmod{m}$ .

(iii) If  $a \equiv b \pmod{m}$  then  $a - b = mq_1$  for some  $q_1 \in \mathbb{Z}$ . Similarly  $b \equiv c \pmod{m} \Rightarrow b - c = mq_2$  for some  $q_2 \in \mathbb{Z}$ . But then  $a - c = (a - b) + (b - c) = mq_1 + mq_2 = m(q_1 + q_2)$  and so, since  $q_1 + q_2 \in \mathbb{Z}$ ,  $m$  divides  $a - c$ , i.e.  $a \equiv c \pmod{m}$ .  $\square$

The real power of the notion of congruence arises from the fact that it respects the basic operations of arithmetic, addition, subtraction and multiplication, as follows.

**Proposition 19.1.3 (Modular arithmetic)** *Suppose that  $a_1$ ,  $a_2$ ,  $b_1$  and  $b_2$  are integers such that  $a_1 \equiv a_2 \pmod{m}$  and  $b_1 \equiv b_2 \pmod{m}$ . Then*

- (i)  $a_1 + b_1 \equiv a_2 + b_2 \pmod{m}$ ,
- (ii)  $a_1 - b_1 \equiv a_2 - b_2 \pmod{m}$ ,
- (iii)  $a_1 b_1 \equiv a_2 b_2 \pmod{m}$ .

*Proof* Suppose that  $a_1 - a_2 = mq_1$  and  $b_1 - b_2 = mq_2$ . Then we have the following.

(i)  $(a_1 + b_1) - (a_2 + b_2) = (a_1 - a_2) + (b_1 - b_2) = mq_1 + mq_2 = m(q_1 + q_2)$ . Therefore  $m$  divides  $[(a_1 + b_1) - (a_2 + b_2)]$ , i.e.  $a_1 + b_1 \equiv a_2 + b_2 \pmod{m}$  as required.

(ii) Left as an exercise (Exercise 19.2).

(iii) Since  $a_1 = a_2 + mq_1$  and  $b_1 = b_2 + mq_2$ ,  $a_1 b_1 = (a_2 + mq_1)(b_2 + mq_2) = a_2 b_2 + a_2 mq_2 + mq_1 b_2 + m^2 q_1 q_2 = a_2 b_2 + m(a_2 q_2 + q_1 b_2 + mq_1 q_2)$ . Hence  $a_1 b_1 - a_2 b_2 = m(a_2 q_2 + q_1 b_2 + mq_1 q_2)$  so that  $a_1 b_1 - a_2 b_2$  is divisible by  $m$ , i.e.  $a_1 b_1 \equiv a_2 b_2 \pmod{m}$ .  $\square$

This result means that if we consider two integers to be interchangeable if they are congruent then as far as addition, subtraction and multiplication are concerned no problems will arise. Division is more complicated and we will begin to look at this at the end of this chapter.

Modular arithmetic leads to chains of congruences just as ordinary arithmetic leads to chains of equalities.

**Examples 19.1.4** (a) Working modulo 5,  $97 \equiv 2$  since  $97 - 2 = 95 = 5 \times 19$  and  $144 \equiv 4$  since  $144 - 4 = 140 = 5 \times 28$ . Hence  $97^3 + 144^2 \equiv 2^3 + 4^2 \equiv 8 + 16$ . But now  $8 \equiv 3$  and  $16 \equiv 1$  and so  $8 + 16 \equiv 3 + 1 \equiv 4$ . The conclusion is that  $97^3 + 144^2 \equiv 4 \pmod{5}$ .

(b) For all  $n \geq 1$ ,  $4^n + 5 \equiv 1^n + 2 \equiv 1 + 2 \equiv 0 \pmod{3}$  so that 3 divides  $4^n + 5$ . This provides a much simpler proof of the result of Problems I, Question 12.

(c) The calculation in the introduction to this chapter leading to the day of the week of 6 September 1998 can be much simplified using modular arithmetic. The day of the year of 6 September 1998 is  $31 \times 5 + 30 \times 2 + 28 \times 1 + 6 \equiv 3 \times 5 + 2 \times 2 + 0 \times 1 + 6 = 15 + 4 + 6 = 25 \equiv 4 \pmod{7}$ . Hence the day of the week is the same as that of 4 January 1998, namely Sunday.

## 19.2 The remainder map

We had better now confirm that the idea used to motivate the concept of congruence in the introduction to this chapter does agree with the formal definition. There we stated that two integers are congruent modulo  $m$  if they lead to the same remainder after division by  $m$ .

**Definition 19.2.1** *The set of integers*

$$R_m = \{0, 1, 2, \dots, m-1\} = \{i \in \mathbb{Z} \mid 0 \leq i < m\}$$

*is called the set of* (least, non-negative) remainders modulo  $m$  *or, alternatively, the set of residues modulo  $m$ .*

**Proposition 19.2.2** *Given an integer  $a$ , there is a unique  $r \in R_m$  such that  $a \equiv r \pmod{m}$ .*

*Proof* This is an immediate consequence of the division theorem (Theorem 15.1.1) since  $a \equiv r \pmod{m}$  if and only if  $a = mq + r$  for some integer  $q$ .  $\square$

**Definition 19.2.3** We define the remainder map  $r_m: \mathbb{Z} \rightarrow R_m$  by

$$r_m(a) = r \Leftrightarrow a \equiv r \pmod{m} \text{ and } r \in R_m.$$

This is well-defined by Proposition 19.2.2.

Notice that since we are here giving a conditional definition of the values of a function it is necessary to confirm that the condition determines a unique element of the codomain for each element of the domain. This is what is meant by the assertion that the function is ‘well-defined’.

In terms of this notation we can state the equivalence of the definition of congruence in Definition 19.1.1 and the idea of congruence discussed in the introduction as follows.

**Proposition 19.2.4** Two integers  $a$  and  $b$  are congruent modulo  $m$  if and only if  $r_m(a) = r_m(b)$ .

*Proof* Suppose that  $r_m(a) = r'$  and  $r_m(b) = r''$ .

$\Leftarrow$ : By definition of  $r_m$ ,  $a \equiv r' \pmod{m}$  and  $b \equiv r'' \pmod{m}$ . Hence if  $r' = r''$  then  $a \equiv r' \equiv r'' \equiv b \pmod{m}$  so that  $a \equiv b \pmod{m}$ .

$\Rightarrow$ : Similarly, if  $a \equiv b \pmod{m}$  then  $r' \equiv a \equiv b \equiv r'' \pmod{m}$  so that  $r' \equiv r'' \pmod{m}$ . But clearly  $r' \equiv r' \pmod{m}$ . Hence, by the uniqueness part of Proposition 19.2.2,  $r' = r''$ .  $\square$

The last step in this proof is an example of how uniqueness results are often used: given that there is only one object with some property, if you know that two objects have this property then they must be equal.

We now illustrate how the language of congruence can simplify the presentation of some earlier proofs.

**Example 19.2.5** Prove that, for an integer  $a$ ,  $a$  is even if and only if  $a^2$  is even. [Exercise 3.3 and Problems I, Question 7]

*Solution* An integer  $n$  is even if and only if  $n \equiv 0 \pmod{2}$ . The set of least non-negative remainders modulo 2 is just  $R_2 = \{0, 1\}$  and so

$n \equiv 0 \pmod{2}$  if and only if  $n \not\equiv 1 \pmod{2}$ . From modular arithmetic,

$$\begin{aligned} a \equiv 0 \pmod{2} &\Rightarrow a^2 \equiv 0 \pmod{2}, \\ a \equiv 1 \pmod{2} &\Rightarrow a^2 \equiv 1 \pmod{2}. \end{aligned}$$

The contrapositive of the second statement is

$$a^2 \equiv 0 \pmod{2} \Rightarrow a \equiv 0 \pmod{2}.$$

Hence  $a \equiv 0 \pmod{2} \Leftrightarrow a^2 \equiv 0 \pmod{2}$ , i.e.  $a$  is even if and only if  $a^2$  is even.  $\square$

**Example 19.2.6** Prove that there do not exist integers  $a$  and  $b$  such that  $a^2 + b^2 = 1234567$ . [Problems IV, Question 1]

*Solution* The set of remainders modulo 4,  $R_4$ , is  $\{0, 1, 2, 3\}$  and so working modulo 4 we have four possibilities to consider. From modular arithmetic,

$$\begin{aligned} a \equiv 0 \pmod{4} &\Rightarrow a^2 \equiv 0 \pmod{4}, \\ a \equiv 1 \pmod{4} &\Rightarrow a^2 \equiv 1 \pmod{4}, \\ a \equiv 2 \pmod{4} &\Rightarrow a^2 \equiv 4 \equiv 0 \pmod{4}, \\ a \equiv 3 \pmod{4} &\Rightarrow a^2 \equiv 9 \equiv 1 \pmod{4}. \end{aligned}$$

Thus for any  $a \in \mathbb{Z}$ ,  $a^2 \equiv 0$  or  $1 \pmod{4}$ . Therefore given two integers  $a$  and  $b$ ,  $a^2 + b^2 \equiv 0$  or  $1$  or  $2 \pmod{4}$ , i.e.  $a^2 + b^2 \not\equiv 3 \pmod{4}$ .

Suppose now for contradiction that  $a$  and  $b$  are integers such that  $a^2 + b^2 = 1234567$ . Then, since  $1234567 \equiv 3 \pmod{4}$ , we have  $a^2 + b^2 \equiv 3 \pmod{4}$  giving the required contradiction. Hence such integers cannot exist.  $\square$

In neither case is this essentially a new proof. It is just that the congruence notation simplifies the presentation of the old proof. This demonstrates one way in which introducing a new concept like congruence is useful: it helps to simplify and unify earlier arguments. However, even more importantly, when we become familiar with the new concept it can then lead to new ideas; the final chapter of the book will include some proofs which seem to arise from a good understanding of the concept of congruence.

### 19.3 Division in congruences

In stating Proposition 19.1.3 we observed that congruences work well with respect to addition, subtraction and multiplication. However, more care is needed with division.

As an illustration observe that  $30 \equiv 2 \pmod{4}$ . If now we divide both sides by 2 we obtain 15 and 1, which are not congruent modulo 4;  $15 \not\equiv 1 \pmod{4}$  since  $15 - 1 = 14$  is not divisible by 4. Let us look for what has gone wrong. We have the congruence  $30 \equiv 2 \pmod{4}$  because  $30 - 2 = 4 \times 7$ . If we divide this equation by 2 we obtain  $15 - 1 = 4 \times (7/2)$  and we do not get a congruence modulo 4 between 15 and 1 because  $7/2$  is not an integer. We can, however, write  $15 - 1 = (4/2) \times 7 = 2 \times 7$  which shows that we do have a congruence modulo 2:  $15 \equiv 1 \pmod{2}$ .

This example shows how we do division when the divisor also divides the modulus. The general result is as follows.

**Proposition 19.3.1** *Suppose that  $a$  is an integer which divides  $m$  (i.e.  $m/a$  is an integer). Then*

$$ab_1 \equiv ab_2 \pmod{m} \Leftrightarrow b_1 \equiv b_2 \pmod{m/a}.$$

*Proof* From the definition of congruence,  $ab_1 \equiv ab_2 \pmod{m} \Leftrightarrow a(b_1 - b_2) = mq$  for some  $q \in \mathbb{Z} \Leftrightarrow b_1 - b_2 = (m/a)q$  for some  $q \in \mathbb{Z} \Leftrightarrow b_1 \equiv b_2 \pmod{m/a}$ .

[Notice that we make use of the fact that  $m/a$  is an integer when we write ‘mod  $(m/a)$ ’.]  $\square$

At the other extreme, division by numbers which are coprime to the modulus is also easy.

**Proposition 19.3.2** *Suppose that  $a$  is an integer such that  $a$  and  $m$  are coprime (i.e. the greatest common divisor  $(a, m) = 1$ ). Then*

$$ab_1 \equiv ab_2 \pmod{m} \Leftrightarrow b_1 \equiv b_2 \pmod{m}.$$

*Proof* ‘ $\Rightarrow$ ’:  $ab_1 \equiv ab_2 \pmod{m} \Rightarrow m$  divides  $a(b_1 - b_2) \Rightarrow m$  divides  $(b_1 - b_2)$  [by Theorem 17.3.2 using the fact that  $(a, m) = 1$ ]  $\Rightarrow b_1 \equiv b_2 \pmod{m}$ .

‘ $\Leftarrow$ ’: This is a special case of Proposition 19.1.3(iii).  $\square$

Division in ordinary arithmetic first arises when we try to solve the linear equation  $ax = b$ . Division in modular arithmetic arises when we

try to solve the linear congruence  $ax \equiv b \pmod{m}$ , in other words seek all integers which satisfy this congruence.

We can go quite a long way by simply using the above results in turn. We use the first until we run out of divisors of  $m$  and then move on to the second.

**Example 19.3.3** Solve the congruence  $4x \equiv 12 \pmod{14}$ , i.e. find all integers  $x$  such that  $4x \equiv 12 \pmod{14}$ .

*Solution* We simply say

$$\begin{aligned} 4x \equiv 12 \pmod{14} &\Leftrightarrow 2x \equiv 6 \pmod{7} \text{ (by Proposition 19.3.1)} \\ &\Leftrightarrow x \equiv 3 \pmod{7} \text{ (by Proposition 19.3.2).} \end{aligned}$$

Thus  $4x \equiv 12 \pmod{14} \Leftrightarrow x \equiv 3 \pmod{7}$ . □

Notice that as in solving equations it is important to have ' $\Leftrightarrow$ ' in the conclusion. In this case the predicate  $x \equiv 3 \pmod{7}$  provides a necessary and sufficient condition for an integer  $x$  to satisfy the congruence.

We normally leave the solution in this form which provides a conditional description of the solution set,  $S = \{x \in \mathbb{Z} \mid x \equiv 3 \pmod{7}\}$ . As an alternative we could give a constructive description of the set by spelling out the notion of congruence which gives  $4x \equiv 12 \pmod{14} \Leftrightarrow x = 3 + 7q$  for some  $q \in \mathbb{Z}$ ; this shows that the solution set  $S = \{3 + 7q \mid q \in \mathbb{Z}\}$ .

However, the question was couched in terms of the modulus 14 and so it might be appropriate to give the solution in terms of this modulus. If we consider the elements of the modulo 14 remainder set,  $R_{14}$ , we see that for  $r \in R_{14}$ ,  $r \equiv 3 \pmod{7}$  if and only if  $r = 3$  or  $r = 10$ . So we conclude that  $x \equiv 3 \pmod{7}$  if and only if  $x \equiv 3 \pmod{14}$  or  $x \equiv 10 \pmod{14}$ . Thus  $4x \equiv 12 \pmod{14} \Leftrightarrow x \equiv 3 \pmod{14}$  or  $x \equiv 10 \pmod{14}$ .

**Example 19.3.4** Solve the congruence  $6x \equiv 15 \pmod{21}$ .

*Solution* This time we obtain

$$\begin{aligned} 6x \equiv 15 \pmod{21} &\Leftrightarrow 2x \equiv 5 \pmod{7} \text{ (by Proposition 19.3.1)} \\ &\Leftrightarrow 2x \equiv 12 \pmod{7} \\ &\quad (\text{using } 5 \equiv 12 \pmod{7} \text{ to get an even number}) \\ &\Leftrightarrow x \equiv 6 \pmod{7} \text{ (by Proposition 19.3.2).} \end{aligned}$$

Hence  $6x \equiv 15 \pmod{21} \Leftrightarrow x \equiv 6 \pmod{7}$ . □

Since for  $r \in R_{21}$ ,  $r \equiv 6 \pmod{7} \Leftrightarrow r = 6, 13 \text{ or } 20$ , we can state this solution as  $6x \equiv 15 \pmod{21} \Leftrightarrow x \equiv 6, 13 \text{ or } 20 \pmod{21}$ .

The *ad hoc* methods used in these two examples are adequate for problems of this type involving only small numbers but if the numbers are large then something more systematic is needed. In the next chapter we will consider a general method for solving linear congruences.

### Exercises

**19.1** What day of the week is 6 September 2045?

**19.2** Prove that  $a_1 \equiv a_2 \pmod{m}$  and  $b_1 \equiv b_2 \pmod{m} \Rightarrow a_1 - b_1 \equiv a_2 - b_2 \pmod{m}$ .

**19.3** Suppose that a positive integer is written in decimal notation as  $n = a_k a_{k-1} \dots a_2 a_1 a_0$  where  $0 \leq a_i \leq 9$ . Prove that  $n$  is divisible by 3 if and only if the sum of its digits  $a_k + a_{k-1} + \dots + a_1 + a_0$  is divisible by 3.

**19.4** Solve the following congruences:

- (i)  $6x \equiv 24 \pmod{45}$ ;
- (ii)  $10x \equiv 15 \pmod{45}$ .

State the results in terms of the modulus 45.

Describe the solution sets without using the notion of congruence.

**19.5** Prove that

$$ab_1 \equiv ab_2 \pmod{m} \Leftrightarrow b_1 \equiv b_2 \pmod{(m/(a, m))}.$$

[Here  $(a, m)$  denotes the greatest common divisor of  $a$  and  $m$  which necessarily divides  $m$  so that  $m/(a, m)$  is an integer.]

# 20

## Linear congruences

Throughout this chapter  $m$  will denote a fixed positive integer.

In this chapter we discuss the following problem.

**Problem 20.0.1** Given a positive integer  $m$  and integers  $a$  and  $b$ , solve the congruence

$$ax \equiv b \pmod{m},$$

i.e. find all integers  $x$  which satisfy this congruence.

Such a problem is called a *linear congruence* and examples have already been considered in the previous chapter. In this chapter we will discuss the general solution. We will make use of a close relationship between this problem and the linear diophantine equation problem discussed in Chapter 18.

### 20.1 A criterion for the existence of solutions

We considered some examples of linear congruences in the previous chapter. The results we used there were as follows.

**Proposition 19.3.1** If  $a$  divides  $m$ , then

$$ab_1 \equiv ab_2 \pmod{m} \Leftrightarrow b_1 \equiv b_2 \pmod{m/a}.$$

**Proposition 19.3.2** If  $(a, m) = 1$ , then

$$ab_1 \equiv ab_2 \pmod{m} \Leftrightarrow b_1 \equiv b_2 \pmod{m}.$$

These results enable us to divide through the linear congruence  $ax \equiv b \pmod{m}$  by common factors of  $a$  and  $b$ . Examples of this were given in Example 19.3.3 and Example 19.3.4. The first result tells us what to do if this common factor is also a factor of the modulus  $m$ . A natural question which arises is what happens if  $a$  and  $m$  have a common factor which does not divide  $b$ . In fact this tells us that the congruence has no solutions at all as we will now explain beginning with an example.

**Example 20.1.1** Prove that the linear congruence  $6x \equiv 14 \pmod{21}$  has no solutions.

**Constructing a proof.** We can summarize what is required as follows.

| Given              | Goal                         |
|--------------------|------------------------------|
| $x \in \mathbb{Z}$ | $6x \not\equiv 14 \pmod{21}$ |

Since the goal here is negative a proof by contradiction is suggested.

| Given  | Goal          |
|--|---------------|
| $x \in \mathbb{Z}$<br>$6x \equiv 14 \pmod{21}$ | contradiction |

This is not a difficult argument to construct once we make use of the definition of congruence. If  $6x \equiv 14 \pmod{21}$  then  $6x - 14 = 21q$  for some integer  $q$ . Hence  $14 = 6x - 21q$ . But now our work on linear diophantine equations enables us to see immediately how to obtain a contradiction using divisibility by 3 for we can rewrite this equation as  $14 = 3(2x - 7q)$  implying that 3 divides 14 which is certainly not the case and so gives the necessary contradiction.

The reader should write out this argument as a formal proof by contradiction using the template of Chapter 4.

The idea used in this example may be stated generally as follows.

**Proposition 20.1.2** *Let  $a$  and  $b$  be integers. If there exists a common divisor of  $a$  and  $m$  which does not divide  $b$ , then the linear congruence  $ax \equiv b \pmod{m}$  has no solutions.*

The proof of this result is by contradiction following the pattern of the previous example.

*Proof* Suppose that  $c$  is a common divisor of  $a$  and  $m$  which does not

divide  $b$ . Suppose for contradiction that there is an integer  $x_0$  which is a solution to the congruence, i.e.  $ax_0 \equiv b \pmod{m}$ . Then  $ax_0 - b = mq$  for some  $q \in \mathbb{Z}$ . Hence  $b = ax_0 - mq$ . But  $c$  divides  $a$  so that  $a = ca_1$  for some  $a_1 \in \mathbb{Z}$  and, since  $c$  divides  $m$ ,  $m = cm_1$  for some  $m_1 \in \mathbb{Z}$ . Hence  $b = ax_0 - mq = ca_1x_0 - cm_1q = c(a_1x_0 - m_1q)$  so that  $c$  divides  $b$  contradicting the given statement that  $c$  does not divide  $b$ . Hence our assumption must be false and so there is no integer solution to  $ax \equiv b \pmod{m}$ .  $\square$

**Example 20.1.3** Prove that the linear congruence

$$12468x \equiv 34567 \pmod{48732}$$

has no solutions.

*Solution* There are no solutions because 3 is a common divisor of 12468 and 48732 but does not divide 34567. [Use Exercise 19.3.]  $\square$

We would like a criterion for the existence of solutions, in other words a necessary and sufficient condition. Proposition 20.1.2 provides us with a necessary condition: if there is a solution then every common divisor of  $a$  and  $m$  must divide  $b$ . This is in fact a necessary and sufficient condition but in order to state an efficient criterion it is useful to restate the proposition in a more specific way: the reference to ‘every common divisor’ suggests that there is a lot to check. Recall that by Corollary 17.2.1 all the common divisors of  $a$  and  $m$  divide the greatest common divisor  $\gcd(a, m)$ . So the necessary condition for solutions given in Proposition 20.1.2 can be entirely captured by considering divisibility by the greatest common divisor.

**Corollary 20.1.4** *A necessary condition for the existence of solutions of the linear congruence  $ax \equiv b \pmod{m}$  is that the greatest common divisor  $\gcd(a, m)$  divides  $b$ .*

*Proof* This is just a special case of the proposition for, by definition of the greatest common divisor,  $(a, m)$  divides  $a$  and  $(a, m)$  divides  $m$ .  $\square$

The condition in Corollary 20.1.4 is also sufficient for the existence of solutions. The key to proving this is the special case when  $a$  and  $m$  are coprime, in which case, since  $\gcd(a, m) = 1$ , the condition is automatically satisfied.

**Theorem 20.1.5** Let  $m$  be a positive integer. Suppose that  $a$  and  $b$  are integers such that  $a$  and  $m$  are coprime. Then the linear congruence

$$ax \equiv b \pmod{m}$$

has a solution. This solution is unique up to congruence modulo  $m$ .

We will defer the proof of this result to the next section. To see the significance of the uniqueness part of this theorem observe that if  $x_0$  is a solution of the linear congruence then so is any integer congruent to  $x_0$  for, by modular arithmetic, if  $x \equiv x_0 \pmod{m}$  then  $ax \equiv ax_0 \pmod{m}$ . The meaning of the statement that the solution is unique modulo  $m$  is that these are the only solutions: given any solution,  $x_0$ , an integer  $x$  is a solution if and only if it is congruent to  $x_0$ . This is illustrated in the following examples.

**Examples 20.1.6** (a) In Example 19.3.4 it was observed that  $2x \equiv 5 \pmod{7} \Leftrightarrow x \equiv 6 \pmod{7}$ .

(b) In Solution 19.4(ii) it was observed that  $2x \equiv 3 \pmod{9} \Leftrightarrow x \equiv 6 \pmod{9}$ .

(c) The linear congruence  $3x \equiv 8 \pmod{11}$  has a unique solution modulo 11 by the theorem. To find the solution we may simply try in turn the elements  $R_{11}$ . The following table lists the values of  $r_{11}(3x)$ .

|              |   |   |   |   |   |   |   |    |   |   |    |
|--------------|---|---|---|---|---|---|---|----|---|---|----|
| $r_{11}(x)$  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7  | 8 | 9 | 10 |
| $r_{11}(3x)$ | 0 | 3 | 6 | 9 | 1 | 4 | 7 | 10 | 2 | 5 | 8  |

From this we see that  $3x \equiv 8 \pmod{11} \Leftrightarrow x \equiv 10 \pmod{11}$ .

In this case it might have been quicker to observe that  $8 \equiv -3 \pmod{11}$  so that  $3x \equiv 8 \pmod{11} \Leftrightarrow 3x \equiv -3 \pmod{11} \Leftrightarrow x \equiv -1 \pmod{11}$  (since 3 and 11 are coprime)  $\Leftrightarrow x \equiv 10 \pmod{11}$ .

(d) Consider the linear congruence  $290x \equiv 5 \pmod{357}$ .

Applying the Euclidean algorithm we discover that 290 and 357 are coprime and so according to Theorem 20.1.5 this congruence has a unique solution modulo 357. It is a bit daunting to go through all the possibilities and it is also harder to see any tricks with these larger numbers. For this sort of example we need something which is systematic but is more efficient. Such a method will be considered in the next section and will provide a method for proving Theorem 20.1.5.

(e) In Example 19.3.3 it was proved that  $4x \equiv 12 \pmod{14} \Leftrightarrow x \equiv 3$  or  $x \equiv 10 \pmod{14}$ . In this case  $a = 4$  and  $m = 14$  are not coprime and

the solution is not unique modulo 14: there are two solutions modulo 14.

If we put together Theorem 20.1.5, Proposition 19.3.1 and Corollary 20.1.4 we obtain the following result which provides the desired criterion for the existence of solutions to a linear congruence and also a measure of how many solutions there are.

**Theorem 20.1.7** *The linear congruence  $ax \equiv b \pmod{m}$  has a solution if and only if the greatest common divisor  $(a, m)$  divides  $b$ . In this case the number of different solutions modulo  $m$  is  $\gcd(a, m)$ .*

**Constructing a proof.** The proof that the divisibility criterion is necessary and sufficient follows easily from previously stated results.

In order to understand the result about the number of solutions the reader may find it useful to look back to Example 19.3.4 which illustrates it. There we found that the congruence  $6x \equiv 15 \pmod{21}$  has a unique solution modulo 7, namely  $x \equiv 6 \pmod{7}$ . To obtain a result formulated in terms of the modulus 21 we observed that the remainders modulo 21 which are congruent to 6 modulo 7 are  $6 + 7q$  for  $q = 0, 1$  and  $2$ , i.e. 6, 13 and 20. The same idea works in general.

*Proof* The necessity of the divisibility condition is Corollary 20.1.4.

For sufficiency suppose that  $\gcd(a, m)$  divides  $b$ . Then since  $(a, m)$  divides  $a$  and  $m$

$$ax \equiv b \pmod{m} \Leftrightarrow \frac{a}{(a, m)}x \equiv \frac{b}{(a, m)} \pmod{\frac{m}{(a, m)}},$$

by Proposition 19.3.1. Now the numbers  $a/(a, m)$  and  $m/(a, m)$  are coprime (Exercise 11.4) and so by Theorem 20.1.5 this congruence has a (unique) solution modulo  $m/(a, m)$  and so certainly a solution.

Now we determine the number of incongruent solutions. Put  $m_1 = m/(a, m)$ , an integer since  $(a, m)$  divides  $m$ . Suppose that the unique solution to the linear congruence modulo  $m_1$  is represented by  $r_1 \in R_{m_1}$  so that an integer  $x$  is a solution if and only if  $x \equiv r_1 \pmod{m_1}$ . Suppose that an integer  $x$  is congruent modulo  $m$  to  $r \in R_m$ . Then  $x$  is a solution if and only if  $r \equiv r_1 \pmod{m_1}$ . This means that  $r = r_1 + m_1q$ . Now  $r \in R_m \Leftrightarrow 0 \leq r < m \Leftrightarrow 0 \leq r_1 + m_1q < m_1(a, m) \Leftrightarrow 0 \leq q < (a, m)$ . So  $x$  is a solution if and only if  $x \equiv r_1 + m_1q \pmod{m}$  for  $0 \leq q < (a, m)$ , giving the required  $(a, m)$  solutions modulo  $m$ .  $\square$

## 20.2 Linear congruences and diophantine equations

We now present a systematic method of finding the solutions of a linear congruence which will enable us to prove Theorem 20.1.5. The key idea is to relate linear congruences to linear diophantine equations.

Let us go back to the definition of congruence modulo  $m$ . Recall that  $ax \equiv b \pmod{m}$  if and only if  $b - ax$  is a multiple of  $m$  which means that there is some integer, say  $y$ , such that  $b - ax = my$ , i.e.  $ax + my = b$ . This means that solving the linear congruence  $ax \equiv b \pmod{m}$  is equivalent to solving the diophantine equation  $ax + my = b$ .

We can make this more precise in the following way. Notice that a solution to the diophantine equation  $ax + my = b$  consists of an ordered pair of integers  $(x_0, y_0)$  such that  $x = x_0$ ,  $y = y_0$  satisfies the equation. So we can think of a solution as being an element of the Cartesian product  $\mathbb{Z} \times \mathbb{Z}$  or  $\mathbb{Z}^2$ . The *solution set* of the diophantine equation is then the subset of  $\mathbb{Z}^2$  given by  $\{(x, y) \in \mathbb{Z}^2 \mid ax + my = b\}$ . Of course, if there are no solutions, then this is the empty set.

Similarly, the solution set of the linear congruence  $ax \equiv b \pmod{m}$  is the subset  $\{x \in \mathbb{Z} \mid ax \equiv b \pmod{m}\}$  of the integers.

**Proposition 20.2.1** *For integers  $a$ ,  $b$ , and positive integers  $m$ , there is a bijection*

$$f: \{(x, y) \in \mathbb{Z}^2 \mid ax + my = b\} \rightarrow \{x \in \mathbb{Z} \mid ax \equiv b \pmod{m}\}$$

given by  $f(x, y) = x$ .

*Proof* Suppose that  $ax_0 \equiv b \pmod{m}$ . Then, for  $(x, y) \in \mathbb{Z}^2$  such that  $ax + my = b$ ,  $f(x, y) = x_0$  if and only if  $x = x_0$  which means that  $y = (b - ax_0)/m$ . [Notice that  $(b - ax_0)/m \in \mathbb{Z}$  since  $ax_0 \equiv b \pmod{m}$ .] This means that  $x_0$  has the unique pre-image  $(x_0, (b - ax_0)/m)$  and so  $f$  is a bijection with inverse map given by  $g(x) = (x, (b - ax)/m)$ .  $\square$

In Chapter 18 a systematic way of finding solutions to linear diophantine equations using the Euclidean algorithm was described. Using the above correspondence this can be used to solve linear congruences.

**Example 20.2.2** Solve the linear congruence  $290x \equiv 5 \pmod{357}$ .

*Solution* We solve the diophantine equation  $290x + 357y = 5$ . Applying

the Euclidean algorithm we get the following sequence of remainders.

$$\begin{array}{r}
 357 \\
 290 \quad (-1) \\
 67 \quad (-4) \\
 22 \quad (-3) \\
 1 \quad (-22) \\
 0
 \end{array}$$

Thus 290 and 357 are coprime so that the diophantine equation has a solution. To find a solution we use the Euclidean algorithm to write 1 as an integral linear combination of 290 and 357.

$$\begin{array}{rcl}
 357 & = & 357 \times 1 + 290 \times 0 \\
 290 & = & 357 \times 0 + 290 \times 1 \quad (-1) \\
 67 & = & 357 \times 1 + 290 \times (-1) \quad (-4) \\
 22 & = & 357 \times (-4) + 290 \times 5 \quad (-3) \\
 1 & = & 357 \times 13 + 290 \times (-16) \quad (-22) \\
 0
 \end{array}$$

Thus a solution to the diophantine equation is given by  $x = (-16) \times 5 = -80$ ,  $y = 13 \times 5 = 65$ . Hence  $x = -80$  is a solution to the linear congruence.

Finally, to see that this is the unique solution modulo 357 observe that

$$\begin{aligned}
 290x \equiv 5 \pmod{357} &\Leftrightarrow 290x \equiv 290 \times (-80) \pmod{357} \\
 &\Leftrightarrow x \equiv -80 \pmod{357},
 \end{aligned}$$

since 290 and 357 are coprime (using Proposition 19.3.2).

Now since  $-80 \equiv 277 \pmod{357}$  we conclude that  $290x \equiv 5 \pmod{357}$  if and only if  $x \equiv 277 \pmod{357}$ . (It is customary to express the solution in terms of the least non-negative remainders.)  $\square$

If we are only interested in solving the congruence we can work with congruences throughout in the Euclidean algorithm. For example we could set out the calculation in the above proof as follows working modulo 357.

$$\begin{array}{rcl}
 357 & \equiv & 290 \times 0 \\
 290 & \equiv & 290 \times 1 \quad (-1) \\
 67 & \equiv & 290 \times (-1) \quad (-4) \\
 22 & \equiv & 290 \times 5 \quad (-3) \\
 1 & \equiv & 290 \times (-16) \quad (-22) \\
 0
 \end{array}$$

Thus  $290x \equiv 5 \pmod{357}$  if  $x \equiv (-16) \times 5 = -80 \equiv 277 \pmod{357}$ . The proof of uniqueness now goes as before.

**Example 20.2.3** Solve the linear congruence  $255x \equiv 15 \pmod{621}$ .

This is an example where  $a$  and  $m$  are not coprime. However, the method of working is very similar to the previous example.

*Solution* Applying the Euclidean algorithm to 255 and 621 gives the following sequence of remainders.

$$\begin{array}{r} 621 \\ 255 \quad (-2) \\ 111 \quad (-2) \\ 33 \quad (-3) \\ 12 \quad (-2) \\ 9 \quad (-1) \\ 3 \quad (-3) \\ 0 \end{array}$$

So  $\gcd(255, 621) = 3$  and, since this divides 15, solutions to the linear congruence will exist according to Theorem 20.1.7. To find them we write each element in the Euclidean algorithm in turn as a multiple of 255 modulo 621.

$$\begin{array}{rcl} 621 & \equiv & 255 \times 0 \\ 255 & \equiv & 255 \times 1 \quad (-2) \\ 111 & \equiv & 255 \times (-2) \quad (-2) \\ 33 & \equiv & 255 \times 5 \quad (-3) \\ 12 & \equiv & 255 \times (-17) \quad (-2) \\ 9 & \equiv & 255 \times 39 \quad (-1) \\ 3 & \equiv & 255 \times (-56) \quad (-3) \\ 0 & & \end{array}$$

Thus  $255x \equiv 15 \pmod{621}$  if  $x \equiv -56 \times 5 = -280 \equiv 341 \pmod{621}$ .

To find all the solutions observe that

$$\begin{aligned} 255x \equiv 15 \pmod{621} &\Leftrightarrow 255x \equiv 255 \times 341 \pmod{621} \\ &\Leftrightarrow 85x \equiv 85 \times 341 \pmod{207} \\ &\Leftrightarrow x \equiv 341 \equiv 134 \pmod{207}, \end{aligned}$$

since 85 and 207 are coprime (as we have divided through by 3, the

greatest common divisor of 255 and 621). If desired we can write the solutions modulo 621 as

$$255x \equiv 15 \pmod{621} \Leftrightarrow x \equiv 134 \text{ or } 341 \text{ or } 548 \pmod{621}.$$

□

This process can be reversed and we can use techniques for solving linear congruences to solve linear diophantine equations.

**Example 20.2.4** Solve the diophantine equation  $140x + 63y = 35$ . [This was considered in Example 18.4.2.]

*Solution* From Proposition 20.2.1 there is a bijection

$$\{(x, y) \in \mathbb{Z}^2 \mid 140x + 63y = 35\} \rightarrow \{x \in \mathbb{Z} \mid 140x \equiv 35 \pmod{63}\}$$

given by  $(x, y) \mapsto x$ . However,

$$\begin{aligned} 140x \equiv 35 \pmod{63} &\Leftrightarrow 14x \equiv 35 \pmod{63} \\ &\quad (\text{since } 140 \equiv 14 \pmod{63}) \\ &\Leftrightarrow 2x \equiv 5 \pmod{9} \\ &\Leftrightarrow 2x \equiv 14 \pmod{9} \\ &\Leftrightarrow x \equiv 7 \pmod{9} \\ &\Leftrightarrow x = 7 + 9q \text{ (for some } q \in \mathbb{Z}). \end{aligned}$$

Substituting back into the equation we see that  $x = 7 + 9q \Leftrightarrow y = -15 - 20q$  and so the solutions to the diophantine equation are given by  $(x, y) = (7 + 9q, -15 - 20q)$  for  $q \in \mathbb{Z}$ , the same solution set as was obtained in Example 18.4.2, but expressed differently. □

If the numbers are small this is usually the best method of solving linear congruences and linear diophantine equations. When they are larger the Euclidean algorithm method is usually best.

Using the ideas of these examples it is now easy to derive Theorem 20.1.5 from previous results on linear diophantine equations.

*Proof of Theorem 20.1.5* Suppose that  $a$  and  $m$  are coprime integers. Then by Theorem 18.2.1 the diophantine equation  $ax + my = b$  has a solution. It follows that the congruence  $ax \equiv b \pmod{m}$  has a solution.

For uniqueness modulo  $m$ , suppose that a solution is given by  $x = x_0$ . Then  $ax \equiv b \pmod{m} \Leftrightarrow ax \equiv ax_0 \pmod{m} \Leftrightarrow x \equiv x_0 \pmod{m}$  since  $a$  and  $m$  are coprime (Proposition 19.3.2). □

The next chapter will include an alternative and very different proof of this result.

### Exercises

**20.1** Solve the following linear congruences:

- (i)  $3x \equiv 5 \pmod{11}$ ;
- (ii)  $10x \equiv 16 \pmod{35}$ ;
- (iii)  $10x \equiv 5 \pmod{35}$ ;
- (iv)  $10x \equiv 13 \pmod{21}$ .

**20.2** Solve the following linear congruences:

- (i)  $341x \equiv 15 \pmod{912}$ ;
- (ii)  $341x \equiv 15 \pmod{913}$ ;
- (iii)  $345x \equiv 15 \pmod{912}$ .

**20.3** Interpret Theorem 20.1.7 in the case when  $a = 0$ .

# 21

## Congruence classes and the arithmetic of remainders

Throughout this chapter  $m$  will denote a fixed positive integer.

The reader will be familiar with the idea that the set of integers is the disjoint union of two subsets: the set of even integers and the set of odd integers. In this case two integers lie in the same subset if and only if they are congruent modulo 2. In the same way, given any positive integer  $m$  we can divide the set of integers into  $m$  disjoint subsets, called the *congruence classes modulo  $m$* : two integers lie in the same class if and only if they are congruent modulo  $m$ .

For some purposes, rather than using congruence of integers it is more convenient to use equality of congruence classes of integers. This is a typical process of mathematical abstraction: the set of all even integers is a more abstract notion than an even integer. In principle everything that is said about congruence classes can be reformulated in terms of congruent integers. However, as so often in mathematics the process of abstraction is enormously powerful because of the way in which it liberates our thinking. In this case one benefit of working with congruence classes is that they are finite in number so that we can make use of counting arguments like the pigeonhole principle introduced in Chapter 11. The power of this approach will be illustrated by giving an alternative proof of Theorem 20.1.5 which demonstrates that from an appropriate point of view the result is essentially ‘obvious’. A further application with far-reaching results known as Fermat’s little theorem will be described in Chapter 24.

A necessary ingredient of this proof is the idea of adding and multiplying congruence classes which can be thought of as describing an arithmetic of remainders. The idea of working in this way using con-

gruence classes arose in the work of Richard Dedekind (whose work on infinite sets was referred to in Chapter 14) in 1857.

### 21.1 Congruence classes

**Definition 21.1.1** Given an integer  $a$ , the congruence class of  $a$  modulo  $m$  is the set of integers which are congruent to  $a$  modulo  $m$ . We denote this congruence class by  $[a]_m$ . Thus

$$[a]_m = \{x \in \mathbb{Z} \mid x \equiv a \pmod{m}\}.$$

**Examples 21.1.2** (a) In the case when  $m = 2$ ,

$$\begin{aligned}[0]_2 &= \{\dots, -4, -2, 0, 2, 4, \dots\} = \{2q \mid q \in \mathbb{Z}\} = [2]_2 = [-16]_2, \\ [1]_2 &= \{\dots, -3, -1, 1, 3, 5, \dots\} = \{2q + 1 \mid q \in \mathbb{Z}\} = [3]_2 = [-53]_2.\end{aligned}$$

Thus there are just two congruence classes modulo 2: the set of all even integers and the set of all odd integers; the congruence class of an even number is the set of all even numbers, and the congruence class of an odd number is the set of all odd numbers.

(b) In the case when  $m = 6$ , there are six congruence classes:

$$\begin{aligned}[0]_6 &= \{\dots, -12, -6, 0, 6, 12, \dots\} = \{6q \mid q \in \mathbb{Z}\} = [36]_6 = [-594]_6, \\ [1]_6 &= \{6q + 1 \mid q \in \mathbb{Z}\} = [49]_6, \\ [2]_6 &= \{6q + 2 \mid q \in \mathbb{Z}\}, \\ [3]_6 &= \{6q + 3 \mid q \in \mathbb{Z}\}, \\ [4]_6 &= \{6q + 4 \mid q \in \mathbb{Z}\}, \\ [5]_6 &= \{6q + 5 \mid q \in \mathbb{Z}\}.\end{aligned}$$

If we list the integers in order in rows of six –

|    |    |    |     |    |    |    |
|----|----|----|-----|----|----|----|
|    |    |    | ... | -9 | -8 | -7 |
| -6 | -5 | -4 | -3  | -2 | -1 |    |
| 0  | 1  | 2  | 3   | 4  | 5  |    |
| 6  | 7  | 8  | ... |    |    |    |

– then we see that the six columns form the six congruence classes modulo 6.

These examples demonstrate that the  $m$  congruence classes of integers modulo  $m$  provide a *partition* of the set of integers. This is formulated precisely in the next result. The significance of this result is that, given two integers  $a$  and  $b$ , the congruence classes  $[a]_m$  and  $[b]_m$  are either equal or disjoint. Hence the set of integers is the disjoint union of the  $m$  congruence classes modulo  $m$ .

**Proposition 21.1.3** *For integers  $a, b$ ,*

- (i)  $a \equiv b \pmod{m} \Leftrightarrow [a]_m = [b]_m$ ,
- (ii)  $a \not\equiv b \pmod{m} \Leftrightarrow [a]_m \cap [b]_m = \emptyset$ .

*Thus two congruence classes modulo  $m$  are either equal or disjoint.*

**Constructing a proof.** Notice that, since the congruence classes  $[a]_m$  and  $[b]_m$  are subsets of the set of integers, the statement that  $[a]_m = [b]_m$  is the assertion that two subsets are equal:

$$x \in [a]_m \Leftrightarrow x \in [b]_m$$

or equivalently

$$x \equiv a \pmod{m} \Leftrightarrow x \equiv b \pmod{m}.$$

The statement that  $[a]_m \cap [b]_m = \emptyset$  is a non-existence statement and  $a \not\equiv b \pmod{m}$  is a negative statement and therefore it is natural to attempt proofs by contradiction for both implications in part (ii) of the proposition.

*Proof* (i) ‘ $\Rightarrow$ ’: Suppose that  $a \equiv b \pmod{m}$ . Then  $x \in [a]_m \Rightarrow x \equiv a \pmod{m} \Rightarrow x \equiv b \pmod{m}$  (by Proposition 19.1.2(iii))  $\Rightarrow x \in [b]_m$  so that  $[a]_m \subseteq [b]_m$ . Similarly,  $[b]_m \subseteq [a]_m$  and so  $[a]_m = [b]_m$ .

‘ $\Leftarrow$ ’: Suppose that  $[a]_m = [b]_m$ . Then, since  $a \in [a]_m$ , we have  $a \in [b]_m$  so that  $a \equiv b \pmod{m}$ .

(ii) ‘ $\Rightarrow$ ’ Suppose that  $a \not\equiv b \pmod{m}$ . Suppose also, for contradiction, that  $[a]_m \cap [b]_m \neq \emptyset$ . Then we can choose  $x_0 \in [a]_m \cap [b]_m$ . But then  $a \equiv x_0 \equiv b \pmod{m}$  so that  $a \equiv b \pmod{m}$  contradicting our hypothesis. Hence our assumption must be false and  $[a]_m \cap [b]_m = \emptyset$ .

‘ $\Leftarrow$ ’: Suppose that  $[a]_m \cap [b]_m = \emptyset$  and, for contradiction, that  $a \equiv b \pmod{m}$ . Then  $a \in [a]_m \cap [b]_m$  and so  $[a]_m \cap [b]_m \neq \emptyset$  giving the required contradiction. Hence  $a \not\equiv b \pmod{m}$ .  $\square$

The reader may find an alternative picture of the congruence classes modulo  $m$  useful. Recall from page 90 the box model of a function. The

box model of the remainder map  $r_m: \mathbb{Z} \rightarrow R_m$  (Definition 19.2.3) may be pictured as follows.

**Table 21.1.4**

| 0        | 1        | 2        | 3        | ... | $m - 1$  |
|----------|----------|----------|----------|-----|----------|
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |     | $\vdots$ |
| $-m$     |          |          |          |     | $-1$     |
| 0        | 1        | 2        | 3        | ... | $m - 1$  |
| $m$      | $m + 1$  | $m + 2$  | $m + 3$  | ... | $2m - 1$ |
| $2m$     | $2m + 1$ |          |          |     | $3m - 1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |     | $\vdots$ |

Here the boxes are labelled by the  $m$  remainders  $0, 1, 2, \dots, m-1$ . The contents of the box  $r \in R_m$  form the set  $\overline{r}_m(\{r\}) = \{x \in \mathbb{Z} \mid r_m(x) = r\}$  which is just the congruence class  $[r]_m$ . Thus this picture shows the partition of the set of integers  $\mathbb{Z}$  into the  $m$  congruence classes modulo  $m$ , generalizing the picture for  $m = 6$  at the end of Example 21.1.2(b) or the layout of a calendar in the case  $m = 7$ . For a general integer  $a$ , the congruence class  $[a]_m$  is equal to  $[r]_m$  if and only if  $r_m(a) = r$ .

**Definition 21.1.5** We write  $\boxed{\mathbb{Z}_m}$  for the set of congruence classes modulo  $m$ .

Notice that an *element* of the set  $\mathbb{Z}_m$  is a *subset* of the set of integers.

**Proposition 21.1.6** The set of congruence classes modulo  $m$ ,  $\mathbb{Z}_m$ , is a finite set of cardinality  $m$ . The  $m$  elements are given by  $[r]_m$  for integers  $r$  such that  $0 \leq r < m$ .

**Constructing a proof.** We can simply count the boxes in the above picture. If we work from left to right then we get the function given in the following proof.

**Proof** We may define a bijection  $f: \mathbb{N}_m \rightarrow \mathbb{Z}_m$  by  $f(i) = [i-1]_m$ . This is a bijection because it has an inverse given by  $g([a]_m) = r_m(a) + 1$ .

This means that

$$\begin{aligned}\mathbb{Z}_m &= \{f(1), f(2), \dots, f(m)\} \\ &= \{[0]_m, [1]_m, \dots, [m-1]_m\}.\end{aligned}$$

□

The reader should consider what is involved in the assertions that this function  $g$  is well-defined and is an inverse to  $f$  – see Exercise 21.6.

## 21.2 The arithmetic of congruence classes

The reader is probably familiar with the following properties of the addition and multiplication of even and odd integers.

| $+$  | even | odd  | $\times$ | even | odd  |
|------|------|------|----------|------|------|
| even | even | odd  | even     | even | even |
| odd  | odd  | even | odd      | even | odd  |

These tables describe the way odd and even integers behave under addition and multiplication. We can see them as a way of adding and multiplying congruence classes modulo 2, a description of arithmetic in the set  $\mathbb{Z}_2$  as follows.

| $+$              | [0] <sub>2</sub> | [1] <sub>2</sub> | $\times$         | [0] <sub>2</sub> | [1] <sub>2</sub> |
|------------------|------------------|------------------|------------------|------------------|------------------|
| [0] <sub>2</sub> | [0] <sub>2</sub> | [1] <sub>2</sub> | [0] <sub>2</sub> | [0] <sub>2</sub> | [0] <sub>2</sub> |
| [1] <sub>2</sub> | [1] <sub>2</sub> | [0] <sub>2</sub> | [1] <sub>2</sub> | [0] <sub>2</sub> | [1] <sub>2</sub> |

The first table describes how the parity (congruence class modulo 2) of integers  $a$  and  $b$  determines the parity of their sum  $a + b$  and the second gives the same information for the product. We can think of this as defining addition and multiplication in the set  $\mathbb{Z}_2$ .

We are now going to consider arithmetic in the set of congruence classes  $\mathbb{Z}_m$ . Remember that the elements of this set are subsets of the integers, the  $m$  subsets forming the columns in Table 21.1.4. To do arithmetic we need to know how to add and multiply two columns together. The method used to add two columns is to choose any element from each column, add the elements and look which column the result lies in. It turns out that whatever choices are made their sum will lie in the same column and so we can define the resulting column as the sum of the two columns to be added. The definition of subtraction and multiplication is similar. Here is the formal definition.

**Definition 21.2.1** Addition, subtraction and multiplication of elements of  $\mathbb{Z}_m$  are defined by

$$[a]_m + [b]_m = [a + b]_m,$$

$$[a]_m - [b]_m = [a - b]_m,$$

$$[a]_m \times [b]_m = [ab]_m.$$

**Proposition 21.2.2** *Addition, subtraction and multiplication in  $\mathbb{Z}_m$  are well-defined.*

Let us illustrate this in the case of multiplication in  $\mathbb{Z}_8$ . The eight congruence classes modulo 8 are given by the columns in the following picture.

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| :  | :  | :  | :  | :  | :  | :  | :  |
| -8 | -7 |    |    |    |    |    | -1 |
| 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
| 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 |    |    |    |    |    | 23 |
| :  | :  | :  | :  | :  | :  | :  | :  |

We can describe the third column as  $[2]_8$  or as  $[10]_8$ . We can describe the eighth column as  $[7]_8$  or as  $[31]_8$ . Multiplying the first representatives gives  $2 \times 7 = 14$  which lies in the seventh column. Multiplying the second representatives gives  $10 \times 31 = 310 = 8 \times 38 + 6$  which also lies in the seventh column. The reader might like to experiment a little more: whatever elements are chosen from the third column and the eighth column their product will lie in the seventh column. This is what is meant in this case by asserting that the definition of multiplication is well-defined.

This procedure leads to the following multiplication table in which as usual congruence classes are represented by the least non-negative remainders modulo 8.

| $\times$ | $[0]_8$ | $[1]_8$ | $[2]_8$ | $[3]_8$ | $[4]_8$ | $[5]_8$ | $[6]_8$ | $[7]_8$ |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|
| $[0]_8$  | $[0]_8$ | $[0]_8$ | $[0]_8$ | $[0]_8$ | $[0]_8$ | $[0]_8$ | $[0]_8$ | $[0]_8$ |
| $[1]_8$  | $[0]_8$ | $[1]_8$ | $[2]_8$ | $[3]_8$ | $[4]_8$ | $[5]_8$ | $[6]_8$ | $[7]_8$ |
| $[2]_8$  | $[0]_8$ | $[2]_8$ | $[4]_8$ | $[6]_8$ | $[0]_8$ | $[2]_8$ | $[4]_8$ | $[6]_8$ |
| $[3]_8$  | $[0]_8$ | $[3]_8$ | $[6]_8$ | $[1]_8$ | $[4]_8$ | $[7]_8$ | $[2]_8$ | $[5]_8$ |
| $[4]_8$  | $[0]_8$ | $[4]_8$ | $[0]_8$ | $[4]_8$ | $[0]_8$ | $[4]_8$ | $[0]_8$ | $[4]_8$ |
| $[5]_8$  | $[0]_8$ | $[5]_8$ | $[2]_8$ | $[7]_8$ | $[4]_8$ | $[1]_8$ | $[6]_8$ | $[3]_8$ |
| $[6]_8$  | $[0]_8$ | $[6]_8$ | $[4]_8$ | $[2]_8$ | $[0]_8$ | $[6]_8$ | $[4]_8$ | $[2]_8$ |
| $[7]_8$  | $[0]_8$ | $[7]_8$ | $[6]_8$ | $[5]_8$ | $[4]_8$ | $[3]_8$ | $[2]_8$ | $[1]_8$ |

**Constructing a proof of Proposition 21.2.2.** Whenever a definition involves choices it is necessary to check that the outcome is independent of the choice. We indicate the proof for multiplication and leave the proofs for addition and subtraction as exercises (see Exercise 21.1). To prove that multiplication is well-defined it is necessary to demonstrate that

$$[a_1]_m = [a_2]_m \text{ and } [b_1]_m = [b_2]_m \Rightarrow [a_1 b_1]_m = [a_2 b_2]_m.$$

Using Proposition 21.1.3 this can be rewritten

$$a_1 \equiv a_2 \pmod{m} \text{ and } b_1 \equiv b_2 \pmod{m} \Rightarrow a_1 b_1 \equiv a_2 b_2 \pmod{m}.$$

This is one of the parts of Proposition 19.1.3 and so the result holds.

Notice that addition and multiplication in  $\mathbb{Z}_m$  inherit the good properties of addition and multiplication in the integers described in Properties 2.3.1: they are commutative, associative and multiplication is distributive over addition. Notice also that the element  $[0]_m$  plays the rôle of 0 in this arithmetic since it is an additive identity:

$$[a]_m + [0]_m = [a]_m \text{ for all } [a]_m \in \mathbb{Z}_m.$$

In the same way the element  $[1]_m$  plays the rôle of 1 since it is a multiplicative identity:

$$[a]_m \times [1]_m = [a]_m \text{ for all } [a]_m \in \mathbb{Z}_m.$$

### 21.3 The arithmetic of remainders

It is possible to view the arithmetic of congruence classes in a slightly different way which the reader may find more acceptable – as the arithmetic of remainders. Consider the following definitions.

**Definition 21.3.1** Addition, subtraction and multiplication of elements of  $R_m$  are defined by

$$\begin{aligned} a +_m b &= r_m(a + b), \\ a -_m b &= r_m(a - b), \\ a \times_m b &= r_m(ab) \end{aligned}$$

for  $a, b \in R_m$ .

Thus to perform addition of remainders, for example, we simply perform ordinary addition and then take the remainder modulo  $m$ .

We can then write out addition and multiplication tables for  $R_m$ . For example in the case of  $m = 10$  we get the following multiplication table.

| $\times_{10}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---------------|---|---|---|---|---|---|---|---|---|---|
| 0             | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1             | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 2             | 0 | 2 | 4 | 6 | 8 | 0 | 2 | 4 | 6 | 8 |
| 3             | 0 | 3 | 6 | 9 | 2 | 5 | 8 | 1 | 4 | 7 |
| 4             | 0 | 4 | 8 | 2 | 6 | 0 | 4 | 8 | 2 | 6 |
| 5             | 0 | 5 | 0 | 5 | 0 | 5 | 0 | 5 | 0 | 5 |
| 6             | 0 | 6 | 2 | 8 | 4 | 0 | 6 | 2 | 8 | 4 |
| 7             | 0 | 7 | 4 | 1 | 8 | 5 | 2 | 9 | 6 | 3 |
| 8             | 0 | 8 | 6 | 4 | 2 | 0 | 8 | 6 | 4 | 2 |
| 9             | 0 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |

Arithmetic modulo 10 is very familiar since the remainder of a positive integer modulo 10 is simply the last (right-hand) digit when it is written in the usual decimal notation.

It is easy to see the relationship between this remainder arithmetic and the arithmetic of congruences. If we represent the congruence classes modulo  $m$  as  $[a]_m$  for  $a \in R_m$  as described in Proposition 21.1.6 then there is a complete correspondence.

**Proposition 21.3.2** *For  $a$  and  $b \in R_m$ ,*

$$\begin{aligned} [a]_m + [b]_m &= [a +_m b]_m, \\ [a]_m - [b]_m &= [a -_m b]_m, \\ [a]_m \times [b]_m &= [a \times_m b]_m. \end{aligned}$$

This shows that the tables for addition, subtraction and multiplication in  $\mathbb{Z}_m$  can be obtained from those for  $R_m$  simply by replacing  $a \in R_m$  by  $[a]_m \in \mathbb{Z}_m$ . This provides an example of two *isomorphic* algebraic structures. The reader will meet the general concept of isomorphism (meaning ‘same form’) when going on to study abstract algebra.

*Proof* The addition result follows immediately from the fact that  $a + b \equiv a +_m b \text{ mod } m$  and similarly for subtraction and multiplication. (See Exercise 21.2.)  $\square$

### 21.4 Linear diophantine equations

There are two big differences between arithmetic in  $\mathbb{Z}_m$  or  $R_m$  and arithmetic in  $\mathbb{Z}$ . Firstly, the set  $R_m$  is finite and so we can always find all the solutions to equations simply by trying out the possibilities. Secondly, division is rather different. In the integers, the only numbers by which we can always divide are 1 and  $-1$ . But in  $R_m$  there are other numbers which we can always divide by. These differences may be illustrated by revisiting the linear diophantine equation problem.

Recall the main result guaranteeing the existence of solutions.

**Theorem 20.1.5** *Suppose that  $a$  and  $b$  are integers such that  $a$  and  $m$  are coprime. Then the linear congruence*

$$ax \equiv b \pmod{m} \quad (21.1)$$

*has a solution. This solution is unique up to congruence modulo  $m$ .*

This result can be reformulated in the language of congruence classes as follows since  $ax \equiv b \pmod{m}$  if and only if  $[ax]_m = [b]_m$ .

**Theorem 21.4.1** *Suppose that  $a$  and  $b$  are integers such that  $a$  and  $m$  are coprime. Then the equation*

$$[a]_m \times [x]_m = [b]_m \quad (21.2)$$

*has a solution in  $\mathbb{Z}_m$ . This solution is unique.*

This reformulation enables us to illustrate the power of the congruence class approach for, in this formulation, since the set of congruence classes is finite it is possible to give a proof using a counting argument.

**Constructing a proof.** Let  $\text{Row}(a)$  be the set of elements in the row corresponding to  $[a]_m$  in the multiplication table for  $\mathbb{Z}_m$ . Notice the following.

- Equation (21.2) has a solution if and only if the congruence class  $[b]_m$  occurs in  $\text{Row}(a)$ .
- If the elements in  $\text{Row}(a)$  are all distinct (so that there are  $m$  of them) then  $\text{Row}(a) = \mathbb{Z}_m$  (and so certainly  $[b]_m \in \text{Row}(a)$ ). [This is a consequence of the pigeonhole principle – see Exercise 11.5.]
- The elements in  $\text{Row}(a)$  are all distinct if  $(a, m) = 1$  (by Proposition 19.3.2).

These three statements taken together prove that equation (21.2) has a

solution if  $a$  and  $m$  are coprime. The distinctness of the elements in the row implies that the solution is unique.

In writing out the formal proof we go back to the pigeonhole principle (Theorem 11.1.2). To use this we need a function between finite sets.

*Proof* Define a function  $f: \mathbb{Z}_m \rightarrow \mathbb{Z}_m$  by  $f([x]_m) = [a]_m \times [x]_m$ .

Now, by Proposition 19.3.2, since  $(a, m) = 1$ , the function  $f$  is an injection, for

$$\begin{aligned} f([x_1]_m) = f([x_2]_m) &\Rightarrow [ax_1]_m = [ax_2]_m \\ &\Rightarrow ax_1 \equiv ax_2 \pmod{m} \\ &\Rightarrow x_1 \equiv x_2 \pmod{m} \\ &\Rightarrow [x_1]_m = [x_2]_m. \end{aligned}$$

We can now see that it is a surjection using the pigeonhole principle (as in Theorem 11.1.7). For suppose for contradiction that it is not a surjection. Then there is some element  $[y_0]_m \in \mathbb{Z}_m$  which is not a value of  $f$  and so  $f$  is an injection  $\mathbb{Z}_m \rightarrow \mathbb{Z}_m - \{[y_0]_m\}$  from a set of cardinality  $m$  to a set of cardinality  $m-1$  which contradicts the pigeonhole principle. Hence  $f$  is a surjection and so a bijection.

Thus there is a unique element  $[x_0]_m \in \mathbb{Z}_m$  such that  $f([x_0]_m) = [b]_m$ , i.e. such that  $[a]_m \times [x_0]_m = [b]_m$ .  $\square$

Notice that this proof is non-constructive, in other words it does not give a means of finding the solutions to the congruence but simply guarantees that these exist. This is a common feature of proofs involving counting arguments. Of course since there are only  $m$  congruence classes modulo  $m$  we can try them in turn and this will inevitably lead to the solution. However, if  $m$  is large this may not be very efficient; then the Euclidean algorithm method described in the previous chapter can be used.

The reader may feel that the proof based on the Euclidean algorithm is preferable, particularly since it does give an algorithm for finding the solution. However, it can be argued that the proof using congruence classes makes it clearer why the result is true. In the end both proofs have something to offer in the exploration of the idea of congruence. This illustrates the way in which a proof is not simply a logical argument demonstrating that a claim is true: a good proof illuminates the relationships between the mathematical ideas involved and different proofs can provide different illumination.

**Example 21.4.2** If we look at the row corresponding to 7 in the multiplication table for  $R_{10}$  then we see that every remainder modulo 10 occurs precisely once in this row. This means that we can always divide by 7: the equation  $7 \times_{10} x = b$  for  $b \in R_{10}$ , or equivalently the equation  $[7]_{10} \times [x]_{10} = [b]_{10}$ , always has a unique solution. If we examine for which rows this is true we find that it is the rows corresponding to 1, 3, 7, 9 and these are precisely the elements of  $R_{10}$  which are coprime to 10.

Furthermore these are the only rows containing the element 1. We now show that this is always the case. An element is called invertible when the corresponding row in the multiplication table contains 1.

**Definition 21.4.3** The element  $[a]_m \in \mathbb{Z}_m$  is called invertible (or  $a \in \mathbb{Z}$  is invertible modulo  $m$ ) if there is an integer  $a' \in \mathbb{Z}$  such that  $[a]_m \times [a']_m = [1]_m$  (or equivalently  $aa' \equiv 1 \pmod{m}$ ).  $[a']_m \in \mathbb{Z}_m$  is called the inverse of  $[a]_m$  and we write  $[a']_m = ([a]_m)^{-1}$ . Alternatively, we say that  $a'$  is an inverse of  $a$  modulo  $m$ .

For example, in  $\mathbb{Z}_{10}$  the invertible elements are  $[1]_{10}$ ,  $[3]_{10}$ ,  $[7]_{10}$  and  $[9]_{10}$  and the inverses are given by  $([1]_{10})^{-1} = [1]_{10}$ ,  $([3]_{10})^{-1} = [7]_{10}$ ,  $([7]_{10})^{-1} = [3]_{10}$  and  $([9]_{10})^{-1} = [9]_{10}$ .

**Proposition 21.4.4** The integer  $a$  is invertible modulo  $m$  if and only if every linear congruence  $ax \equiv b \pmod{m}$  has a unique solution modulo  $m$ .

Notice that this means that  $\text{Row}(a) = \mathbb{Z}_m$  in the multiplication table for  $\mathbb{Z}_m$ : each congruence class occurs precisely once in the row. This result is an immediate consequence of the next result.

**Proposition 21.4.5** Suppose that  $a'$  is an inverse of  $a$  modulo  $m$ . Then

$$ax \equiv b \pmod{m} \Leftrightarrow x \equiv a'b \pmod{m}.$$

*Proof* Since  $a'$  is an inverse modulo  $m$  of  $a$ ,  $aa' \equiv 1 \pmod{m}$ . Thus

$$ax \equiv b \pmod{m} \Rightarrow a'ax \equiv a'b \pmod{m} \Rightarrow x \equiv a'b \pmod{m}$$

and

$$x \equiv a'b \pmod{m} \Rightarrow ax \equiv aa'b \pmod{m} \Rightarrow ax \equiv b \pmod{m}$$

as required.  $\square$

Notice that this means that once we have found an inverse modulo  $m$  to  $a$  then we can compute the solution to any linear congruence  $ax \equiv b \pmod{m}$  by multiplication. This is illustrated in Exercise 21.5.

### Exercises

**21.1** Use Proposition 19.1.3 to confirm that addition and subtraction in  $\mathbb{Z}_m$  are well-defined by Definition 21.3.2.

**21.2** Prove that  $a +_m b \equiv a + b \pmod{m}$ ,  $a -_m b \equiv a - b \pmod{m}$  and  $a \times_m b \equiv ab \pmod{m}$ . [This was used in the proof of Proposition 21.3.2.]

**21.3** Write down the multiplication table for  $R_{12}$ . From the table identify the invertible elements modulo 12 and their inverses.

**21.4** Prove that  $a$  is invertible modulo  $m$  if and only if  $a$  and  $m$  are coprime.

**21.5** Prove that 290 is invertible modulo 357 and find its inverse. Hence solve the linear congruences:

- (i)  $290x \equiv 2 \pmod{357}$ ;
- (ii)  $290x \equiv 5 \pmod{357}$ ;
- (iii)  $290x \equiv 355 \pmod{357}$ .

**21.6** Prove that the function  $g: \mathbb{Z}_m \rightarrow N_m$  used in the proof of Proposition 21.1.6 is well-defined and an inverse to the function  $f$ .

# 22

## Partitions and equivalence relations

In the previous chapter we saw that congruence of integers can be viewed in two ways: as a relation between certain integers and as a partition of the set of integers. This is an example of a very general concept, known as an equivalence relation, which the reader will encounter frequently in more advanced pure mathematics.

An equivalence relation on a set corresponds to a partition of the set. In this chapter this correspondence is explained and a number of examples discussed.

### 22.1 Partitions

We saw in Chapter 21 that the relation of congruence modulo  $m$  partitions the set of integers  $\mathbb{Z}$  into  $m$  disjoint *congruence classes* (see Proposition 21.1.3). Let us make the notion of a partition of a set precise.

**Definition 22.1.1** *Let  $X$  be a set. Recall that the power set of  $X$ , denoted  $\mathcal{P}(X)$ , is the set of all subsets of  $X$ . A partition of the set  $X$  is a subset  $\Pi$  of  $\mathcal{P}(X)$ , i.e. a set of subsets of  $X$ , such that*

- (i) *the subsets in  $\Pi$  are non-empty, i.e.  $A \in \Pi \Rightarrow A \neq \emptyset$ ,*
- (ii) *the subsets in  $\Pi$  are disjoint, i.e.*  
 $\forall A_1, A_2 \in \Pi, (A_1 \neq A_2 \Rightarrow A_1 \cap A_2 = \emptyset),$
- (iii) *the subsets in  $\Pi$  cover  $X$ :  $\forall x \in X, \exists A \in \Pi, x \in A$ .*

**Examples 22.1.2** (a) By Proposition 21.1.3, the set of congruence classes modulo  $m$ ,

$$\mathbb{Z}_m = \{ [a]_m \mid a \in \mathbb{Z} \} = \{ [a]_m \mid a \in R_m \},$$

is a partition of  $\mathbb{Z}$  into  $m$  subsets.

(b) A set of people may be partitioned according to the year of their birth.

(c) A partition of the set  $N_{10}$  is given by

$$\Pi = \{\{1, 3, 7, 8\}, \{2, 9, 10\}, \{4\}, \{5, 6\}\}.$$

(d) A partition of the set  $\mathbb{Z}$  is given by

$$\Pi = \{\mathbb{Z}^+, \{0\}, \mathbb{Z}^- = \{n \in \mathbb{Z} \mid n < 0\}\}.$$

(e) The set of subsets

$$\{\{1, 2, 3, 4, 5\}, \{5, 6, 7, 8, 9\}\}$$

is not a partition of  $N_9$  since the element 5 lies in both subsets.

(f) The set of subsets

$$\{\{1, 2, 3, 4\}, \{6, 7, 8, 9\}\}$$

is not a partition of  $N_9$  since the element 5 lies in neither subset.

We have already seen that congruence classes modulo  $m$  are obtained by partitioning the integers according to the remainder after division by  $m$ . It is common to describe a partition by means of some property of the elements such as the remainder after division by  $m$  which can be used for Example 22.1.2(a), or the date of birth used in Example 22.1.2(b). This amounts to partitioning according to the value of some function at each point.

**Proposition 22.1.3** *Let  $f: X \rightarrow Y$  be a surjection. Then  $X$  may be partitioned according to the value of the function  $f$  giving the partition*

$$\Pi = \{\overleftarrow{f}(\{y\}) \mid y \in Y\}.$$

Recall from Definition 9.3.1 that  $\overleftarrow{f}(\{y_0\}) = \{x \in X \mid f(x) = y_0\}$ .

*Proof* The proof of this result is immediate from the definitions: the sets of the partition are non-empty since  $f$  is a surjection, they are disjoint because the function  $f$  is well-defined, and they cover  $X$  because  $f$  has a value in  $Y$  at each point of  $X$ .  $\square$

**Examples 22.1.4** (a) The partition of the integers by congruence classes modulo  $m$  corresponds to the remainder map  $r_m: \mathbb{Z} \rightarrow R_m$  discussed in

## Chapter 19.

(b) The partition of the integers of Example 22.1.2(d) corresponds to the function

$$f: \mathbb{Z} \rightarrow \{-1, 0, 1\}$$

defined by

$$f(x) = \begin{cases} x/|x| & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

## 22.2 Equivalence relations

The idea of a partition takes as its starting point the set to be partitioned or divided up. We can view the same idea from the point of view of the elements. If a set of people is physically partitioned into subsets with possibly each subset put in a separate room, each person may not have a good idea of the overall picture but they are able to see who is in the same subset as themselves by looking round the room. A partition of a set  $X$  determines a *relation* on the set: two elements are related if they are in the same subset of the partition.

A *relation* on a set  $X$  is determined by a property which may or may not be satisfied by each ordered pair of elements of the set. If an ordered pair of elements  $a$  and  $b$  satisfies the property then we say that the elements are related and write  $a \sim b$  to indicate this. We write  $a \not\sim b$  to denote the fact that the ordered pair  $(a, b)$  does not satisfy the property and so the elements are not related.

**Proposition 22.2.1** *Suppose that  $\Pi$  is a partition of the set  $X$ . Then we may define a relation on the set  $X$  as follows. For elements  $a$  and  $b$  of  $X$ ,  $a$  is related to  $b$ , written  $a \sim_{\Pi} b$ , if and only if  $a$  and  $b$  belong to the same subset of the partition. Then this relation has the following properties.*

- (i) Reflexive property. *For all  $a \in X$ ,  $a \sim_{\Pi} a$ .*
- (ii) Symmetric property. *If  $a$  and  $b$  are elements of  $X$  such that  $a \sim_{\Pi} b$ , then  $b \sim_{\Pi} a$ .*
- (iii) Transitive property. *If  $a$ ,  $b$  and  $c$  are elements of  $X$  such that  $a \sim_{\Pi} b$  and  $b \sim_{\Pi} c$ , then  $a \sim_{\Pi} c$ .*

This result should be compared with Proposition 19.1.2. Starting from the partition of the integers into congruence classes modulo  $m$  we obtain the relation of congruence modulo  $m$ .

*Proof* The relation is clearly reflexive and symmetric. For transitivity, suppose that  $a \sim_{\Pi} b$  and  $b \sim_{\Pi} c$ . Then  $a$  and  $b$  belong to some  $A_1 \in \Pi$  and  $b$  and  $c$  belong to some  $A_2 \in \Pi$ . Now  $b \in A_1 \cap A_2$  and so these two subsets are not disjoint. Thus  $A_1 = A_2$  and so  $a \sim_{\Pi} c$ .  $\square$

A relation with these three properties is called an *equivalence relation*. In order to clarify the significance of the properties we consider a variety of relations.

**Examples 22.2.2** (a) The relation ' $\leq$ ' on  $\mathbb{Z}$  the set of integers.

- (b) The relation ' $<$ ' on  $\mathbb{Z}$ .
- (c) The relation ' $=$ ' on  $\mathbb{Z}$ .
- (d) We can define a relation on  $\mathbb{Z}$  by  $a \sim b \Leftrightarrow ab > 0$ .
- (e) We can define a relation on  $\mathbb{Z}$  by  $a \sim b \Leftrightarrow ab = 0$ .
- (f) We can define a relation on the set  $\mathbb{Z} \times (\mathbb{Z} - \{0\})$  by  $(a_1, b_1) \sim (a_2, b_2) \Leftrightarrow a_1 b_2 = a_2 b_1$ .
- (g) For each positive integer  $m$ , congruence modulo  $m$  defines a relation on the set  $\mathbb{Z}$ .

We now consider the particular features of relations arising from a partition of a set described in Proposition 22.2.1.

**Definition 22.2.3** Suppose that  $\sim$  is a relation on a set  $X$ .

- (i) The relation is said to be reflexive when, for all  $x \in X$ ,  $x \sim x$ .
- (ii) The relation is said to be symmetric when, for all elements  $x, y$  of  $X$ ,  $x \sim y \Rightarrow y \sim x$ .
- (iii) The relation is said to be transitive when, for all elements  $x, y, z$  of  $X$ ,  $x \sim y$  and  $y \sim z \Rightarrow x \sim z$ .

An equivalence relation on the set  $X$  is a relation which is reflexive, symmetric and transitive.

Let us look at the above examples.

**Examples 22.2.4** (a) The relation ' $\leq$ ' on  $\mathbb{Z}$  is reflexive (since  $\forall a \in \mathbb{Z}, a \leq a$ ), is not symmetric (since  $1 \leq 2$  but  $2 \not\leq 1$ ) and is transitive (since  $a \leq b$  and  $b \leq c \Rightarrow a \leq c$ ).

- (b) The relation ' $<$ ' on  $\mathbb{Z}$  is not reflexive (since  $1 \not< 1$ ), is not symmetric (since  $1 < 2$  but  $2 \not< 1$ ), and is transitive ( $a < b$  and  $b < c \Rightarrow a < c$ ).
- (c) The relation '=' on  $\mathbb{Z}$  is reflexive ( $\forall a \in \mathbb{Z}, a = a$ ), is symmetric ( $\forall a, b \in \mathbb{Z}, a = b \Rightarrow b = a$ ) and is transitive ( $\forall a, b, c \in \mathbb{Z}, a = b$  and  $b = c \Rightarrow a = c$ ). This is an equivalence relation.
- (d) The relation  $a \sim b \Leftrightarrow ab > 0$  on  $\mathbb{Z}$  is not reflexive ( $0 \not\sim 0$  since  $0^2 = 0 \not> 0$ ), is symmetric ( $ab > 0 \Rightarrow ba > 0$ ) and is transitive (since  $ab > 0$  and  $bc > 0 \Rightarrow ab^2c > 0 \Rightarrow ac > 0$ ).
- (e) The relation  $a \sim b \Leftrightarrow ab = 0$  on  $\mathbb{Z}$  is not reflexive ( $1 \not\sim 1$  since  $1^2 = 1 \neq 0$ ), is symmetric (since  $ab = 0 \Rightarrow ba = 0$ ) and is not transitive (since  $1 \sim 0$  and  $0 \sim 1$  but  $1 \not\sim 1$ ).
- (f) The relation  $(a_1, b_1) \sim (a_2, b_2) \Leftrightarrow a_1b_2 = a_2b_1$  on  $\mathbb{Z} \times (\mathbb{Z} - \{0\})$  is reflexive, symmetric and transitive (since  $(a_1, b_1) \sim (a_2, b_2) \Leftrightarrow a_1/b_1 = a_2/b_2 \in \mathbb{Q}$ ) and so an equivalence relation.
- (g) Proposition 19.1.2 asserts that congruence modulo  $m$  is an equivalence relation on the set of integers.
- (h) More generally, Proposition 22.2.1 asserts that the relation  $\sim_\Pi$  arising from a partition  $\Pi$  of the set  $X$  is an equivalence relation. We will demonstrate in the next section that every equivalence relation arises from a partition.

Notice that these properties of a relation are universal statements and so to prove that one fails it is only necessary to give a single counter-example.

### 22.3 Equivalence relations and partitions

We saw in Chapter 21 that the partition of the set of integers into congruence classes modulo  $m$  arises from the relation of congruence modulo  $m$  (see Proposition 21.1.3). This is a quite general phenomenon. Any equivalence relation on a set  $X$  determines a partition of the set  $X$  to which it corresponds. The construction of the partition is an immediate generalization of Proposition 21.1.3.

**Definition 22.3.1** Suppose that  $\sim$  is an equivalence relation on a non-empty set  $X$ . For each  $a \in X$  we define the equivalence class of  $a$  to be the set of elements of  $X$  which are equivalent to  $a$ . We denote this

class by  $\boxed{[a]}$ . Thus

$$[a] = \{x \in X \mid x \sim a\}.$$

Observe that each equivalence class is a subset of  $X$ . We denote the set of all equivalence classes by  $X/\sim$ , i.e.

$$X/\sim = \{[a] \mid a \in X\} \subseteq \mathcal{P}(X).$$

This is a straightforward generalization of the idea of congruence class; when the equivalence relation is congruence modulo  $m$  on the set of integers then the equivalence classes are the congruence classes modulo  $m$ . We can now generalize Proposition 21.1.3 and its proof.

**Theorem 22.3.2** Suppose that  $\sim$  is an equivalence relation on the set  $X$ . Then, for  $a, b \in X$ ,

- (i)  $a \sim b \Leftrightarrow [a] = [b]$ ,
- (ii)  $a \not\sim b \Leftrightarrow [a] \cap [b] = \emptyset$ .

*Proof* (i) ' $\Rightarrow$ ': Suppose that  $a \sim b$ . Then  $x \in [a] \Rightarrow x \sim a \Rightarrow x \sim b$  (by transitivity)  $\Rightarrow x \in [b]$  so that  $[a] \subseteq [b]$ . Similarly, since  $a \sim b \Rightarrow b \sim a$  (symmetry),  $[b] \subseteq [a]$  and so  $[a] = [b]$ .

' $\Leftarrow$ ': Suppose that  $[a] = [b]$ . Then, since  $a \in [a]$  (by reflexivity),  $a \in [b]$  so that  $a \sim b$ .

(ii) ' $\Rightarrow$ ' Suppose that  $a \not\sim b$ . Suppose for contradiction that  $[a] \cap [b] \neq \emptyset$ . Then we can choose  $x_0 \in [a] \cap [b]$ . But then  $x_0 \sim a$  and  $x_0 \sim b \Rightarrow a \sim x_0$  and  $x_0 \sim b$  (by symmetry)  $\Rightarrow a \sim b$  (by transitivity) contradicting our hypothesis. Hence our assumption must be false and  $[a] \cap [b] = \emptyset$ .

' $\Leftarrow$ ': Suppose that  $[a] \cap [b] = \emptyset$  and, for contradiction, that  $a \sim b$ . Then  $a \in [a]$  (by reflexivity) and  $a \in [b]$  (by assumption) so that  $a \in [a] \cap [b]$  and so  $[a] \cap [b] \neq \emptyset$  giving the required contradiction. Hence  $a \not\sim b$ .  $\square$

**Corollary 22.3.3** Suppose that  $\sim$  is an equivalence relation on a non-empty set  $X$ . Then the set of equivalence classes  $\Pi = X/\sim$  is a partition of  $X$ . The relation arising from this partition is  $\sim$ .

*Proof* The theorem shows that distinct equivalence classes are disjoint. They cover  $X$  because, for each  $x \in X$ ,  $x \in [x]$  (reflexivity). They are non-empty because by definition an equivalence class is  $[a]$  for some  $a \in X$  and again reflexivity ensures that  $a \in [a]$ .

Part (i) of the theorem is the statement that  $\sim$  is equal to  $\sim_{\Pi}$  the relation arising from the partition.  $\square$

This is the converse to Proposition 22.2.1 and demonstrates that the notions of an equivalence relation on the set  $X$  and a partition of the set  $X$  correspond and are simply different ways of thinking about the same concept.

We saw earlier that a surjection with domain  $X$  gives rise to a partition of  $X$ . As a final result in this chapter we reformulate this in terms of equivalence relations.

**Proposition 22.3.4** *Let  $f: X \rightarrow Y$  be a surjection. Define an equivalence relation on  $X$  by  $x_1 \sim x_2 \Leftrightarrow f(x_1) = f(x_2)$ . Then the map  $f$  induces a bijection  $X/\sim \rightarrow Y$  by  $[x] \mapsto f(x)$ .*

*Proof* Define  $F: X/\sim \rightarrow Y$  by  $F([x]) = f(x)$  for each  $x \in X$ . We must check that this function is well-defined. Suppose that  $x_1, x_2 \in X$ . Then

$$\begin{aligned} [x_1] = [x_2] &\Rightarrow x_1 \sim x_2 \\ &\Rightarrow f(x_1) = f(x_2) \quad \text{by definition of } \sim. \end{aligned}$$

The function  $F$  is a surjection since  $f$  is a surjection. The proof that it is injective is as follows. Suppose that  $[x_1], [x_2] \in X/\sim$ . Then

$$\begin{aligned} F([x_1]) = F([x_2]) &\Rightarrow f(x_1) = f(x_2) \quad \text{by definition of } F \\ &\Rightarrow x_1 \sim x_2 \quad \text{by definition of } \sim \\ &\Rightarrow [x_1] = [x_2]. \end{aligned}$$

Thus  $F$  is a bijection.  $\square$

**Example 22.3.5** Define a function  $f: \mathbb{Z} \times (\mathbb{Z} - \{0\}) \rightarrow \mathbb{Q}$  by  $f(a, b) = a/b$ . Then  $f(a_1, b_1) = f(a_2, b_2) \Leftrightarrow a_1/b_1 = a_2/b_2 \Leftrightarrow a_1b_2 = a_2b_1$  and so the equivalence relation  $\sim$  on  $\mathbb{Z} \times \mathbb{Z} - \{0\}$  induced by the function  $f$  as in Theorem 22.3.4 is the equivalence relation of Example 22.2.2(f). It follows that there is a bijection  $F: (\mathbb{Z} \times (\mathbb{Z} - \{0\}))/\sim \rightarrow \mathbb{Q}$ .

This formalizes the relationship between fractions and rationals discussed in Chapter 13. There we observed that two fractions  $a_1/b_1$  and  $a_2/b_2$  represent the same rational number if and only if  $a_1b_2 = a_2b_1$ , i.e.  $(a_1, b_1) \sim (a_2, b_2)$ . We can think of the set of fractions as  $\mathbb{Z} \times (\mathbb{Z} - \{0\})$  with  $(a, b)$  written as  $a/b$ .

Now with hindsight we can see how to recast the material in Section 13.1 in the language of equivalence relations. We start with the

equivalence relation on the set  $\mathbb{Z} \times (\mathbb{Z} - \{0\})$ :

$$(a_1, b_1) \sim (a_2, b_2) \Leftrightarrow a_1 b_2 = a_2 b_1.$$

Then we *define* the set of rational numbers to be the set of equivalence classes of  $\mathbb{Z} \times (\mathbb{Z} - \{0\})$  under this equivalence relation. Now it can be seen that the notation  $q = \langle a/b \rangle$  indicating that the fraction  $a/b$  represents the rational  $q$  simply anticipated the notation  $q = [(a, b)]$  indicating that the pair  $(a, b)$  represents the equivalence class  $q$ .

Addition and multiplication of rational numbers (that is equivalence classes) may now be defined by

$$\begin{aligned} [(a_1, b_1)] + [(a_2, b_2)] &= [(a_1 b_2 + a_2 b_1, b_1 b_2)], \\ [(a_1, b_1)] \times [(a_2, b_2)] &= [(a_1 a_2, b_1 b_2)], \end{aligned}$$

and as in Proposition 13.1.5 these operations are well-defined.

The set of integers may be embedded in the set of rationals by letting  $a \in \mathbb{Z}$  correspond to  $[(a, 1)]$ . It is now possible to check that the above definitions of addition and multiplication of rationals extend the definitions for integers and satisfy all the basic algebraic properties of numbers discussed in Chapter 2. It follows that  $[(a, b)]$  is the unique rational solution to the equation  $bx = a$ . The notion of inequality between integers can also be extended to the rationals (see Problems III, Question 21).

Having done this we can agree to use the fraction  $a/b$  to represent the equivalence class  $[(a, b)]$  and so obtain the usual notation for the rational numbers.

### Exercises

**22.1** For each of the following relations on the set  $X$  determine whether it is reflexive, whether it is symmetric and whether it is transitive. For those which are equivalence relations, describe the equivalence classes.

- (i) For  $X = \mathbb{Z}$ , put  $a \sim b \Leftrightarrow a + b$  is even.
- (ii) For  $X = \mathbb{Z}$ , put  $a \sim b \Leftrightarrow a + b$  is odd.
- (iii) For  $X = \mathbb{Z}$ , put  $a \sim b \Leftrightarrow a + b$  is divisible by 3.
- (iv) For  $X = \mathbb{Z}$ , put  $a \sim b \Leftrightarrow a^2 = b^2$ .
- (v) For  $X = \{1, 2\}$ , put  $a \sim b \Leftrightarrow a = 1$  and  $b = 1$ .
- (vi) For  $X = \{1, 2\}$ , put  $a \sim b \Leftrightarrow a = 1$  or  $b = 1$ .
- (vii) For  $X = \mathbb{R} \times \mathbb{R}$ ,  $(a_1, a_2) \sim (b_1, b_2) \Leftrightarrow a_1 = b_1$ .
- (viii) For  $X = \mathbb{R} \times \mathbb{R}$ ,  $(a_1, a_2) = (b_1, b_2) \Leftrightarrow a_1^2 + a_2^2 = b_1^2 + b_2^2$ .

**22.2** Describe the equivalence relation on  $\mathbb{Z}$  determined by the partition  $\Pi = \{A, B, C\}$  where  $A = \mathbb{Z}^+$ ,  $B = \mathbb{Z}^-$  and  $C = \{0\}$ .

**22.3** Define a function  $f: \mathbb{Z}^+ \times \mathbb{Z}^+ \rightarrow \mathbb{Z}$  by  $f(x_1, x_2) = x_1 - x_2$ . Prove that this is a surjection. Identify the equivalence classes of the equivalence relation defined by  $(x_1, x_2) \sim (y_1, y_2) \Leftrightarrow f(x_1, x_2) = f(y_1, y_2)$ .

## Problems V: Modular arithmetic

1. Prove, for positive integers  $n$ , that 7 divides  $6^n + 1$  if and only if  $n$  is odd.
2. Prove that, for all integers  $a$  and  $b$ ,  $a^2 + b^2 \equiv 0, 1, 2, 4$  or  $5$  modulo 8. Deduce that there do not exist integers  $a$  and  $b$  such that  $a^2 + b^2 = 12345790$ .
3. Suppose that a positive integer is written in decimal notation as  $n = a_k a_{k-1} \dots a_2 a_1 a_0$  where  $0 \leq a_i \leq 9$ . Prove that  $n$  is divisible by 9 if and only if the sum of its digits  $a_k + a_{k-1} + \dots + a_1 + a_0$  is divisible by 9.
4. Suppose that a positive integer is written in decimal notation as  $n = a_k a_{k-1} \dots a_2 a_1 a_0$  where  $0 \leq a_i \leq 9$ . Prove that  $n$  is divisible by 11 if and only if the alternating sum of its digits  $a_0 - a_1 + \dots + (-1)^k a_k$  is divisible by 11.
5. Find a test similar to those in the previous questions for divisibility by 99.
6. Prove that the diophantine equation  $3x^2 + 4y^2 = 5z^2$  has no non-trivial (i.e.  $(x, y, z) \neq (0, 0, 0)$ ) solutions.

[Give a proof by contradiction. Obtain a contradiction by proving that if there is a non-trivial solution then there is a solution  $(x_1, y_1, z_1)$  with  $x_1 \not\equiv 0 \pmod{5}$  or  $y_1 \not\equiv 0 \pmod{5}$ .]

Deduce that the equation  $3x^2 + 4y^2 = 5$  has no rational solutions.

- 7.** What is the last digit of  $2^{1000}$ ?
- 8.** Prove that the Fibonacci number  $u_n$  (see Definition 5.4.2) is divisible by 3 if and only if  $n$  is divisible by 4.
- 9.** Solve the following linear congruences:
- (i)  $3x \equiv 15 \pmod{18}$ ;
  - (ii)  $3x \equiv 16 \pmod{18}$ ;
  - (iii)  $4x \equiv 16 \pmod{18}$ ;
  - (iv)  $4x \equiv 14 \pmod{18}$ .
- 10.** Solve the following linear congruences:
- (i)  $23x \equiv 16 \pmod{107}$ ;
  - (ii)  $234x \equiv 20 \pmod{366}$ ;
  - (iii)  $234x \equiv 6 \pmod{366}$ ;
  - (iv)  $234x \equiv 36 \pmod{366}$ .
- 11.** Using the result of Question 10(iv), solve the diophantine equation
- $$234x + 366y = 36.$$
- 12.** Write down the multiplication table for  $R_{15}$ . From the table identify the invertible elements modulo 15 and their inverses.
- 13.** Prove that each row in the multiplication table for  $R_m$  either includes the element 1 or includes the element 0 more than once (but not both).
- 14.** Using a counting argument, prove from the multiplication table for  $R_m$  that, for each integer  $a$ , there exists an integer  $b_1$  such that  $ax \equiv b_1 \pmod{m}$  has no solution if and only if there exists an integer  $b_2$  such that  $ax \equiv b_2 \pmod{m}$  has more than one solution modulo  $m$ .
- 15.** Prove that 245 is invertible modulo 666 and find its inverse. Hence solve the linear congruences:
- (i)  $245x \equiv 3 \pmod{666}$ ;
  - (ii)  $245x \equiv 656 \pmod{666}$ .
- 16.** Identify the invertible elements of  $\mathbb{Z}_{18}$  and write down their inverses.

17. For each of the following relations on the set  $X$  determine whether it is reflexive, whether it is symmetric and whether it is transitive. For those which are equivalence relations, describe the equivalence classes.

- (i) For  $X = \mathbb{Z}$ , put  $a \sim b \Leftrightarrow ab \neq 0$ .
- (ii) For  $X = \mathbb{Z}$ , put  $a \sim b \Leftrightarrow ab \geq 0$ .
- (iii) For  $X = \mathbb{Z}^+$ , put  $a \sim b \Leftrightarrow ab > 0$ .
- (iv) For  $X = \mathbb{Z} - \{0\}$ , put  $a \sim b \Leftrightarrow ab > 0$ .
- (v) For  $X = \mathbb{Z}^+$ , put  $a \sim b \Leftrightarrow ab < 0$ .
- (vi) For  $X = \mathbb{Z} - \{0\}$ , put  $a \sim b \Leftrightarrow ab < 0$ .

18. Which of the following formulae define a (well-defined) function  $f: \mathbb{Q} \rightarrow \mathbb{Q}$ ?

- (i)  $f(a/b) = a^2/b^2$ .
- (ii)  $f(a/b) = a^2/b^3$ .
- (iii)  $f(a/b) = b/a$ .
- (iv)  $f(a/b) = a + b$ .
- (v)  $f(a/b) = (a - b)/2b$ .

19. Which of the following formulae define a (well-defined) function  $f: \mathbb{Z}_6 \rightarrow \mathbb{Z}_4$ ?

- (i)  $f([a]_6) = [a + 1]_4$ .
- (ii)  $f([a]_6) = [2a]_4$ .
- (iii)  $f([a]_6) = [a^2]_4$ .



# **Part VI**

## Prime numbers



# 23

## The sequence of prime numbers

To conclude this book we now spend some time exploring the theory of prime numbers, one of the most fascinating and oldest areas of number theory and one in which some apparently very simple questions still remain unanswered. It is a branch of mathematics which until recently seemed remote from any applications but this has now changed. In this chapter we consider the sequence of prime numbers and factorization into primes.

### 23.1 Definition and basic properties

**Definition 23.1.1** A positive integer  $n$  is said to be **prime** when  $n > 1$  and the only positive divisors of  $n$  are 1 and  $n$ . If an integer  $n > 1$  is not prime then it is said to be **composite**.

Thus  $n > 1$  is composite if and only if it can be written as a product  $n = ab$  of integers  $a$  and  $b$  such that  $1 < a < n$  and  $1 < b < n$ . If a prime number  $n$  is the product of two positive integers,  $n = ab$ , then either  $a = 1$  and  $b = n$  or  $a = n$  and  $b = 1$ .

To some extent it is a matter of convention whether we call the number 1 prime but it is convenient to exclude it for reasons which will become clear in the next chapter. Thus the positive integers are partitioned into three subsets: the set of primes, the set of composites, and the set of units (which just consists of 1).

The first few prime numbers are as follows:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, ... .

**Proposition 23.1.2** *Every integer greater than 1 can be written as a product of prime numbers.*

Here we think of a prime number as a product of a single prime number, namely itself! We can extend this result to include 1 by adopting the convention that the product of no numbers at all is equal to 1 (remember that  $x^0 = 1$ ).

The proof of Proposition 23.1.2 is *the* classic example of a proof by strong induction (see Section 5.4).

*Proof* The proof is by induction. We prove by strong induction on  $n$  the statement that, for  $n \geq 2$ , the number  $n$  may be written as a product of primes.

Base case: 2 is a prime number and so a product of a single prime.

Inductive step: Suppose now as inductive hypothesis that, for some  $k \geq 2$ , if  $2 \leq n \leq k$  then  $n$  may be written as a product of primes. Then if the number  $k + 1$  is prime it is a product of a (single) prime. If the number  $k + 1$  is not prime then it is composite in which case we can write  $k + 1 = ab$  where  $2 \leq a, b \leq k$ . But then by inductive hypothesis  $a$  and  $b$  may be written as products of primes so that by combining these products we write  $k + 1$  as a product of primes. Hence in both cases  $k + 1$  is a product of primes as required to prove the inductive step.

Conclusion: Hence, by induction,  $n$  is a product of primes for all  $n \geq 2$ .

□

The key property of prime numbers is the following result which appears in Book Seven of the *Elements* of Euclid.

**Theorem 23.1.3** *Suppose that  $p$  is a prime number and  $p$  divides  $ab$  for positive integers  $a$  and  $b$ . Then  $p$  divides  $a$  or  $p$  divides  $b$ .*

**Constructing a proof.** This is an example of a result where the conclusion is an ‘or’ statement. What is required can be summarized as follows.

| Given                                  | Goal                               |
|--|------------------------------------|
| $p$ a prime number<br>$p$ divides $ab$ | $p$ divides $a$ or $p$ divides $b$ |

Recall from Section 4.4 that we prove an ‘or’ statement by proving that if the first statement is false then the second must be true. This gives

the following strategy.

| Given                   | Goal            |
|-------------------------|-----------------|
| $p$ a prime number      |                 |
| $p$ divides $ab$        |                 |
| $p$ does not divide $a$ | $p$ divides $b$ |

This is easy to achieve once we recognize that, when  $p$  is prime, its only divisors are  $\pm 1$  and  $\pm p$  and so these are the only candidates for common divisors of  $p$  and another integer  $a$ . So if we are told that  $p$  (and so  $-p$ ) does not divide  $a$  the only common divisors are  $\pm 1$  and so the numbers are coprime. Theorem 17.3.2 now gives the result.

*Proof* Suppose that  $p$  divides  $ab$ . Now if  $p$  does not divide  $a$  then, since  $p$  is a prime number,  $p$  and  $a$  are coprime. It follows from Theorem 17.3.2 that  $p$  divides  $b$ . Thus either  $p$  divides  $a$  or  $p$  divides  $b$  as required.  $\square$

This property of prime numbers is usually taken as the definition of a prime number in more advanced work. Notice that if an integer  $n$  greater than 1 is not prime then it does not satisfy this property. For there are then integers  $a$  and  $b$  such that  $1 < a, b < n$  and  $n = ab$ . But now  $n$  divides  $ab$  but  $n$  does not divide  $a$  and  $n$  does not divide  $b$ . This shows that the property of prime numbers described in Theorem 23.1.3 characterizes prime numbers amongst the integers greater than 1.

## 23.2 The sieve of Eratosthenes

The problem of determining whether or not a positive integer is prime and the factorization of composite numbers into primes has always intrigued mathematicians. Recently methods of public-key cryptography† have been discovered which make use of the product of two large primes. It is relatively easy to find large prime numbers but if then they are multiplied together recovering the prime factors (which as we will see in the next section are unique) is an enormously time consuming task even with the help of a computer and is impractical using known techniques if the prime numbers are large enough. The codes are public in the sense that the knowledge of how to encode a message is made public. The coding technique uses remainders modulo a product of two large primes. To

† Accessible accounts of some of the basic ideas may be found in D.M. Davis, *The nature and power of mathematics*, Princeton University Press, 1993, and J.F. Humphreys and M.Y. Prest, *Numbers, groups and codes*, Cambridge University Press, 1989.

decode a message it is necessary to make use of the prime factors which are not made public. So anyone can encode a message but only someone knowing the prime numbers can decode it.

In this book we restrict attention to a basic and ancient method for listing primes named after Eratosthenes of Cyrene (ca. 276-ca. 194 B.C.) who was director of the library in Alexandria.

The most obvious approach to deciding whether a positive integer  $n$  is prime is to try all smaller positive integers one by one. If none (except 1) divides  $n$  then it must be prime. As soon as one does divide  $n$  then we know that  $n$  must be composite. This task is shortened once we realize that we do not have to try every integer less than  $n$ .

**Proposition 23.2.1** If  $n \geq 2$  is a composite number, then it has a prime factor not exceeding  $\sqrt{n}$ .

**Proof** Since  $n$  is composite we can write  $n = ab$  where  $a$  and  $b$  are integers such that  $1 < a, b < n$ . If  $a \leq \sqrt{n}$ , then we choose a prime factor of  $a$  to obtain a prime factor of  $n$  which is less than or equal to  $\sqrt{n}$ . On the other hand, if  $a > \sqrt{n}$ , then  $b = n/a < \sqrt{n}$  and so we can take a prime factor of  $b$ .  $\square$

So to decide whether  $n$  is prime or composite it is sufficient to try all prime numbers not exceeding  $\sqrt{n}$ . We can organize this into a systematic method for finding all the prime numbers not exceeding some given number, an algorithm called the ‘Sieve of Eratosthenes’. The simplest way to explain the algorithm is to give an example.

**Example 23.2.2** To find all prime numbers not exceeding 100.

First of all write out the sequence of integers from 2 to 100 as follows.

The first number in the list, the number 2, is necessarily prime but all multiples of 2 are necessarily composite so we cross them out.

The next of the remaining numbers is 3, which must also be prime. Again, all multiples of 3 are necessarily composite and so we cross out those which remain.

At each stage the next remaining number is necessarily prime since all multiples of lower primes have been crossed out. We then cross out all multiples of this number. So here we proceed to 5 and then 7.

Now, since there are no more primes not greater than  $\sqrt{100} = 10$ , the positive integers which remain must all be prime by Proposition 23.2.1 and so we have found all the primes less than 100, namely those integers not crossed out below.

| 2             | 3             | 4             | 5             | 6             | 7             | 8             | 9             | 10            |
|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 11            | <del>12</del> | 13            | <del>14</del> | <del>15</del> | <del>16</del> | 17            | <del>18</del> | 19            |
| <del>21</del> | <del>22</del> | 23            | <del>24</del> | <del>25</del> | <del>26</del> | 27            | <del>28</del> | 29            |
| 31            | <del>32</del> | <del>33</del> | <del>34</del> | <del>35</del> | <del>36</del> | 37            | <del>38</del> | <del>39</del> |
| 41            | <del>42</del> | 43            | <del>44</del> | <del>45</del> | <del>46</del> | 47            | <del>48</del> | <del>49</del> |
| <del>51</del> | <del>52</del> | 53            | <del>54</del> | <del>55</del> | <del>56</del> | <del>57</del> | <del>58</del> | 59            |
| 61            | <del>62</del> | <del>63</del> | <del>64</del> | <del>65</del> | <del>66</del> | 67            | <del>68</del> | <del>69</del> |
| 71            | <del>72</del> | 73            | <del>74</del> | <del>75</del> | <del>76</del> | <del>77</del> | <del>78</del> | 79            |
| <del>81</del> | <del>82</del> | 83            | <del>84</del> | <del>85</del> | <del>86</del> | <del>87</del> | <del>88</del> | 89            |
| <del>91</del> | <del>92</del> | <del>93</del> | <del>94</del> | <del>95</del> | <del>96</del> | 97            | <del>98</del> | <del>99</del> |
|               |               |               |               |               |               |               |               | 100           |

Thus the primes less than 100 are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89 and 97 (25 in all).

### 23.3 The fundamental theorem of arithmetic

We have shown (Proposition 23.1.2) that every positive integer greater than 1 is prime or a product of primes. In this section it is demonstrated that this factorization is essentially unique, a result known as the fundamental theorem of arithmetic. It is a curiosity of history that, although all the results about prime numbers proved so far are very ancient, the fundamental theorem is a relatively modern theorem and its first clear statement and proof only appeared in the nineteenth century in Gauss's *Disquisitiones arithmeticae*.

**Theorem 23.3.1 (Fundamental theorem of arithmetic)** *Every positive integer greater than 1 can be written uniquely as a product of prime*

*numbers with the prime factors in the product written in non-decreasing order.*

The point of the final part about the order of the factors is simply that in general the expression as a product of primes is clearly not unique because we can reorder the factors. For example we have  $12 = 2 \times 2 \times 3 = 2 \times 3 \times 2 = 3 \times 2 \times 2$  giving three ways of writing 12 as a product of primes. However, these expressions are obtained simply by reordering the factors and we can uniquely specify an order by requiring the factors to be in non-decreasing order, in this case giving the first expression.

We can write the theorem more symbolically by saying that each integer  $n > 1$  can be written uniquely in the form

$$n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$$

where  $r \geq 1$ ,  $k_i \in \mathbb{Z}^+$  and each  $p_i$  is a prime number with  $p_1 < p_2 < \dots < p_r$ . We call this the *standard prime factorization* of  $n$ . If we allow  $r = 0$  we can include the number 1 in this by adopting the convention that it is the product of no prime numbers.

We have already proved the existence of the prime factorization (Proposition 23.1.2) and so our task now is to prove uniqueness. To do this we generalize Theorem 23.1.3.

**Proposition 23.3.2** *Suppose that  $p$  is a prime number and  $p$  divides a product of positive integers  $a_1 a_2 \dots a_m$ . Then  $p$  divides  $a_i$  for some  $i$  such that  $1 \leq i \leq m$ .*

*Proof* The proof is by induction on  $m$ .

Base case: For  $m = 1$  this result says  $p$  divides  $a_1 \Rightarrow p$  divides  $a_1$  which is certainly true.

Inductive step: Suppose that, for some  $k \geq 1$ , the result is true for  $m = k$ . Then we can deduce it for  $m = k + 1$  as follows.

$$\begin{aligned} p \text{ divides } a_1 a_2 \dots a_k a_{k+1} \\ \Rightarrow p \text{ divides } (a_1 a_2 \dots a_k) a_{k+1} \\ \Rightarrow p \text{ divides } a_1 a_2 \dots a_k \text{ or } p \text{ divides } a_{k+1} \quad (\text{by Theorem 23.1.3}) \\ \Rightarrow p \text{ divides } a_i \text{ for some } i \text{ such that } 1 \leq i \leq k \\ \quad (\text{by inductive hypothesis}) \text{ or } p \text{ divides } a_{k+1} \\ \Rightarrow p \text{ divides } a_i \text{ for some } i \text{ such that } 1 \leq i \leq k + 1. \end{aligned}$$

This proves the inductive step.

Conclusion: Hence the result is true for all positive integers  $m$ .  $\square$

*Proof of Theorem 23.3.1* We have already proved that factorization into primes is possible (Proposition 23.1.2). We prove uniqueness by contradiction. So suppose, for contradiction, that there is some integer  $n > 1$  which has two different factorizations into primes:

$$n = p_1 p_2 \dots p_r = q_1 q_2 \dots q_s,$$

where the  $p_i$ s and the  $q_j$ s are all primes and  $p_1 \leq p_2 \leq \dots \leq p_r$ ,  $q_1 \leq q_2 \leq \dots \leq q_s$ . We can now cancel any common primes from the two factorizations so that we get an equality as above with  $p_i \neq q_j$  for all  $i$  and  $j$ . [Notice that not all the primes will cancel since we were supposing that the factorizations were different.] However, this contradicts Proposition 23.3.2 since clearly  $p_1$  divides  $p_1 p_2 \dots p_r$  and so  $p_1$  divides  $q_1 q_2 \dots q_s$  whereas  $p_1$  does not divide any  $q_j$  since  $q_j$  is prime and different from  $p_1$ . This contradiction shows that our original hypothesis is false and so we have proved the theorem.  $\square$

### 23.4 Applications of the fundamental theorem of arithmetic

A direct application of the fundamental theorem of arithmetic is to find all the divisors of a number.

**Proposition 23.4.1** Suppose that  $a = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$  is the standard prime factorization of a positive integer  $a$ . Then  $b$  is a divisor of  $a$  if and only if it can be written in the form  $b = p_1^{l_1} p_2^{l_2} \dots p_r^{l_r}$  where  $0 \leq l_i \leq k_i$  for  $1 \leq i \leq r$ .

**Example 23.4.2** The standard prime factorization of 72 is  $72 = 2^3 \times 3^2$ . Hence the divisors of 72 are  $2^{l_1} \times 3^{l_2}$  where  $0 \leq l_1 \leq 3$  and  $0 \leq l_2 \leq 2$ . This gives  $2^0 \times 3^0 = 1$ ,  $2^1 \times 3^0 = 2$ ,  $2^2 \times 3^0 = 4$ ,  $2^3 \times 3^0 = 8$ ,  $2^0 \times 3^1 = 3$ ,  $2^1 \times 3^1 = 6$ ,  $2^2 \times 3^1 = 12$ ,  $2^3 \times 3^1 = 24$ ,  $2^0 \times 3^2 = 9$ ,  $2^1 \times 3^2 = 18$ ,  $2^2 \times 3^2 = 36$  and  $2^3 \times 3^2 = 72$ .

*Proof of Proposition 23.4.1* Recall that  $b$  is a divisor of  $a$  if and only if there is an integer  $q$  such that  $a = bq$ . Now the standard prime factorizations of  $b$  and  $q$  must combine to give the standard prime factorization of  $a$ , by the uniqueness part of the fundamental theorem. It is clear then that if  $a = bq$  the standard prime factorization of  $b$  must be obtained from

the standard prime factorization of  $a$  by simply omitting some factors. Thus  $b$  has the given form.

Conversely if  $b$  has the given form then  $b$  divides  $a$  since we can take  $q = p_1^{k_1-l_1} p_2^{k_2-l_2} \dots p_r^{k_r-l_r}$ .  $\square$

**Corollary 23.4.3** *Given positive integers  $a$  and  $b$  suppose that  $p_1 < p_2 < \dots < p_r$  are the primes occurring in the prime factorization of  $a$  or in the prime factorization of  $b$  so that  $a = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$  and  $b = p_1^{l_1} p_2^{l_2} \dots p_r^{l_r}$  where  $k_i$  and  $l_i$  are non-negative integers. Then the greatest common divisor of  $a$  and  $b$  is given by  $p_1^{m_1} p_2^{m_2} \dots p_r^{m_r}$  where  $m_i$  is the minimum of the two numbers  $k_i$  and  $l_i$ .*

*Proof* This is immediate from the proposition.  $\square$

**Example 23.4.4** Consider the greatest common divisor  $(72, 30)$  (Example 11.3.3). The standard prime factorizations are  $72 = 2^3 \times 3^2 = 2^3 \times 3^2 \times 5^0$ ,  $30 = 2 \times 3 \times 5 = 2^1 \times 3^1 \times 5^1$ . Hence the greatest common divisor is  $2^1 \times 3^1 \times 5^0 = 6$ .

It is important to realize that this is not a practical way of finding the greatest common divisor of two numbers of substantial size since factorization into primes is in general very time-consuming.

We can also use the fundamental theorem to give an alternative proof of the irrationality of  $\sqrt{2}$ .

**Theorem 13.2.1** *There does not exist a rational number whose square is 2.*

*Proof* This proof is also by contradiction. Suppose for contradiction that there is a rational number  $q \in \mathbb{Q}$  such that  $q^2 = 2$ . Since  $q^2 = (-q)^2$  we may suppose that  $q$  is positive. Write  $q = a/b$  where  $a, b \in \mathbb{Z}^+$ . Then  $a^2 = 2b^2$ . Let  $2 < p_2 < \dots < p_r$  be the primes occurring as factors of  $a$  or of  $b$ . Then  $a = 2^{k_1} p_2^{k_2} \dots p_r^{k_r}$  and  $b = 2^{l_1} p_2^{l_2} \dots p_r^{l_r}$  where  $k_i$  and  $l_i$  are non-negative integers. This means that  $a^2 = 2^{2k_1} p_2^{2k_2} \dots p_r^{2k_r}$  and  $2b^2 = 2^{2l_1+1} p_2^{2l_2} \dots p_r^{2l_r}$ . But by the fundamental theorem these expressions must be the same and so in particular  $2k_1 = 2l_1 + 1$  which is impossible for integers  $k_1$  and  $l_1$ . This provides the necessary contradiction demonstrating that our hypothesis that there exists a rational square root of 2 is false and so proving the theorem.  $\square$

### 23.5 The distribution of prime numbers

The sequence of primes is very irregular and it is extremely difficult to see any pattern in it. We can, however, prove that there are infinitely many primes. The following result is also from the *Elements* of Euclid (Book Nine) and is one of the great classic proofs by contradiction.

**Theorem 23.5.1** *There are infinitely many prime numbers.*

*Proof* The proof is by contradiction. Suppose for contradiction that the set  $P$  of primes is finite, say of cardinality  $n$ . Then we can list the elements of  $P$ :

$$P = \{ p_1, p_2, \dots, p_n \}.$$

Now consider the positive integer  $m = p_1 p_2 \dots p_n + 1$ . This number is not divisible by  $p_i$  for  $1 \leq i \leq n$  since the remainder is 1. But, by Proposition 23.1.2,  $m$  may be written as a product of primes and so is divisible by some prime  $p$  which must then not be an element of  $P$ . However,  $P$  was the set of all primes and so this is a contradiction showing that our assumption that the set of primes is finite must be false. Hence the set of primes must be infinite.  $\square$

If we carry on with the Sieve of Eratosthenes we find that the number of primes gradually thins out. It is natural to ask whether it is possible to give a formula for the  $n$ th prime number or less ambitiously for an estimate for the number of primes at each stage. Mathematicians carried out enormous calculations by hand in the investigation of the distribution of prime numbers. Attempts were made to find simple expressions which took prime values; for example Leonhard Euler proved that  $n^2 - n + 41$  is prime for  $1 \leq n \leq 40$ . Gauss investigated the distribution of the primes up to 3000000 and found 216745 of them! There are in fact 216817 but a few slips were inevitable in such an enormous hand calculation.

**Definition 23.5.2** For  $n \in \mathbb{Z}^+$  we write  $\boxed{\pi(n)}$  for the number of primes not exceeding  $n$ .

Thus  $\pi(2) = 1$ ,  $\pi(10) = 4$ ,  $\pi(100) = 25$  and calculation shows that  $\pi(1000) = 168$ ,  $\pi(10^6) = 78498$ . One of the most remarkable theorems of the nineteenth century gives an estimate for  $\pi(n)$  for large values of  $n$ .

**Theorem 23.5.3 (Prime number theorem)** *The ratio of  $\pi(n)$  and  $n/\log_e n$  approaches 1 as  $n$  grows without bound.*

Here  $\log_e n$  is the natural logarithm of  $n$  which is the area under the graph of the function  $x \mapsto 1/x$  between  $x = 1$  and  $x = n$ .

The theorem says that  $\pi(n)$  and  $n/\log_e n$  are *asymptotic*. The precise statement requires the idea of a null sequence introduced in Definition 8.3.2: it is the statement that the sequence  $n \mapsto (\pi(n) \log_e n)/n - 1$  is null. This means that  $n/\log_e n$  gives a good estimate for  $\pi(n)$  if  $n$  is large enough. To get some idea of how good an estimate consider the following table.<sup>†</sup>

| $n$       | $\pi(n)$    | $n/\log_e n$  | ratio |
|-----------|-------------|---------------|-------|
| $10^3$    | 168         | 144.8         | 1.160 |
| $10^6$    | 78498       | 72382.4       | 1.085 |
| $10^9$    | 50847534    | 48254942.4    | 1.054 |
| $10^{12}$ | 37607912018 | 36191206825.3 | 1.039 |

The first person to suggest publicly that a result along these lines was true was the French mathematician Adrien-Marie Legendre who suggested in 1798 that  $\pi(n)$  is approximately equal to  $n/(\log_e n - 1.08366)$  on the basis of extensive calculations. Gauss had already in 1793 reached the conclusion that  $\pi(n)$  is asymptotic to  $n/\log_e n$  according to unpublished work found amongst his papers. However, he found an even better approximation. As a result of his calculations he estimated that the number of primes in the interval from  $a$  to  $b$  is approximately  $\int_a^b \frac{dx}{\log_e x}$ . The agreement is striking: for example between 2600000 and 2700000 Gauss found 6762 primes whereas  $\int_{2600000}^{2700000} \frac{dx}{\log_e x} = 6761.332$ . He discusses this estimate in a letter he wrote in 1849.<sup>‡</sup>

The prime number theorem was finally proved in 1896 independently by the French mathematician Jacques Hadamard and by the Belgian mathematician C.J. de la Vallée-Poussin making use of ideas from ad-

<sup>†</sup> Extracted from Kenneth Rosen, *Elementary number theory and its applications*, Addison-Wesley, 1988. The values of  $\pi(n)$  for large  $n$  are not found by the methods described in this book which involve testing all the numbers up to  $n$  but by evaluating some approximate formula for  $\pi(n)$  to an accuracy of a half – which determines  $\pi(n)$  since it is obviously an integer.

<sup>‡</sup> An account of Gauss's work on the distribution of primes including a translation of this letter is to be found in an article by L.J. Goldstein, A history of the prime number theorem, *American Mathematical Monthly*, volume 80, Mathematical Association of America, 1973 (pages 599–615).

vanced calculus.†. De la Vallée-Poussin showed in addition that  $\pi(n)$  is more accurately represented by the ‘logarithmic integral’  $\text{li}(n) = \int_0^n \frac{dx}{\log_e x}$  (which can be given a well-defined value in spite of the fact that  $\log_e 1 = 0$ ) than by any expression of the form  $n/(\log_e n - A)$  where  $A$  is a constant, confirming conjectures suggested by Gauss’s calculations.

Before leaving this topic it is worth remarking on a remarkable theorem proved the British mathematician J.E. Littlewood in 1914. He showed that, although for all values of  $n$  for which it has been calculated  $\text{li}(n) - \pi(n)$  is positive, this difference takes positive and negative values infinitely often as  $n$  runs through the positive integers. This is an example of a ‘pure existence’ result and the proof gave no indication at all how to find a value of  $n$  for which the difference is negative. In 1933, Littlewood’s student S. Skewes proved that there was a value of  $n$  for which the difference is negative before you reach  $\exp \exp \exp \exp 79$  (using the exponential function  $\exp(x) = e^x$  inverse to the natural logarithm  $\log_e$ ). This reduces the search for  $n$  for which the difference is negative to a finite search – but illustrates how knowing that a task is finite may be no help at all in carrying it out. The number arising here is normally considered to be larger than any number which had previously arisen in a pure mathematical argument. Subsequent work by Skewes in 1955 marginally reduced the bound – to a number with about  $10^{10^{100}}$  decimal places!

### Exercises

**23.1** Use the Sieve of Eratosthenes to find all the prime numbers less than 200.

**23.2** Write 4148 and 7684 as products of prime numbers. Hence find  $\gcd(4148, 7684)$ .

**23.3** Prove that there exists an integer  $n$  such that  $n^2 - n + 41$  (Euler’s formula) is not prime.

**23.4** Prove that for each positive integer  $n$  there is a sequence of  $n$  consecutive integers all of which are composite.

[Hint: Consider  $(n+1)! + i$ .]

† A beautiful argument motivating the prime number theorem and using only the basic ideas of the calculus may be found in R. Courant and H. Robbins, *What is mathematics?* Oxford University Press, 1941 (pages 482–486).

**23.5** Euler's totient function  $\phi: \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$  is defined by putting  $\phi(n)$  equal to the number of positive integers not exceeding  $n$  which are co-prime to  $n$ . Prove that if  $p$  is prime then  $\phi(p^k) = p^k - p^{k-1}$ .

**23.6** Use the method of proof of this chapter to prove that there does not exist a rational number  $q$  such that  $q^3 = 2$ .

**23.7** Consider the set of positive integers of the form  $4q + 1$ . Notice that this set of numbers is closed under multiplication. Call an element  $n$  of this set a pseudo-prime when  $n > 1$  and the only divisors of  $n$  in this set are 1 and  $n$ .

Prove that every  $n > 1$  in the set can be written as a product of pseudo-primes but that this factorization is not always unique. So the analogue of the fundamental theorem of arithmetic is false in this set of numbers.

# 24

## Congruence modulo a prime

In this final chapter we consider a number of results concerning expressions which are automatically divisible by some given prime number. The results given here have their origins in the seventeenth century but it is convenient to formulate and prove them using the idea of congruence.

Although the results may appear only to be of theoretical interest these ideas do in fact provide an important ingredient in the techniques of public-key cryptography using prime numbers referred to in the previous chapter. Furthermore the ideas introduced in this chapter led to the development of group theory, the study of symmetry, which is now an enormously rich subject with many applications both within mathematics and in many branches of science.

### 24.1 Fermat's little theorem

We first consider a striking result known as Fermat's 'little' theorem to distinguish it from his 'great' or 'last' theorem which was discussed in Chapter 18. As so often with Fermat's results it is not known how he proved the result although he mentioned that he had found a proof in a letter dated 18 October 1640. Euler was the first to publish a proof, in 1736, and a variant of this proof appears as Problems VI, Question 17. Apparently this proof also appears in an unpublished manuscript by Gottfried Leibniz some time before 1683.

In Chapter 21 we introduced operations of addition, subtraction and multiplication on the set  $\mathbb{Z}_m$  of congruence classes of the integers modulo  $m$  and observed that an element  $[a]_m \in \mathbb{Z}_m$  has a (unique) multiplicative inverse  $[a']_m$  (i.e.  $[a]_m \times [a']_m = [1]_m$  or equivalently  $aa' \equiv 1 \pmod{m}$ ) if

and only if  $a$  is coprime to  $m$ . For example, when  $m = 10$  we saw that  $[1]_{10}$ ,  $[3]_{10}$ ,  $[7]_{10}$  and  $[9]_{10}$  are invertible in  $\mathbb{Z}_{10}$ .

In the case when the modulus is a prime number, say  $p$ , then every non-zero class in  $\mathbb{Z}_p$  is invertible because all numbers  $a$  such that  $0 < a < p$  are coprime to  $p$ . This makes arithmetic modulo  $p$ , or in  $\mathbb{Z}_p$ , particularly simple. This fact can be used to prove some striking results in number theory and Fermat's little theorem is an example.

**Theorem 24.1.1 (Fermat's little theorem)** *If  $p$  is a prime number and  $a$  is a positive number which is not a multiple of  $p$ , then  $a^{p-1} \equiv 1 \pmod{p}$ .*

Equivalently we can assert that  $a^{p-1} - 1$  is divisible by  $p$ . The proof which follows was first given by the Scottish mathematician James Ivory in 1806.<sup>†</sup>

*Proof* We make use of the function  $f: \mathbb{Z}_p \rightarrow \mathbb{Z}_p$  used in the proof of Theorem 21.4.1. Since  $a$  is not a multiple of  $p$ ,  $a$  and  $p$  are coprime. We define  $f: \mathbb{Z}_p \rightarrow \mathbb{Z}_p$  by  $f([x]_p) = [a]_p \times [x]_p = [ax]_p$ . As in the proof of Theorem 21.4.1, since  $(a, p) = 1$ , this function is an injection so that, by the pigeonhole principle, it is a bijection.

Now  $f[0]_p = [0]_p$  and so the  $(p - 1)$  classes  $[a]_p, [2a]_p, \dots, [(p - 1)a]_p$  are simply the classes  $[1]_p, [2]_p, \dots, [p - 1]_p$  in a different order. Hence

$$[a]_p \times [2a]_p \times \cdots \times [(p - 1)a]_p = [1]_p \times [2]_p \times \cdots \times [p - 1]_p.$$

This means that

$$a \times 2a \times \cdots \times (p - 1)a \equiv 1 \times 2 \times \cdots \times (p - 1) \pmod{p}$$

or

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

But now, since  $p$  is prime,  $(p - 1)!$  and  $p$  are coprime and so, by Proposition 19.3.2, we can divide through by  $(p - 1)!$  to obtain  $a^{p-1} \equiv 1 \pmod{p}$  as required.  $\square$

**Corollary 24.1.2** *If  $p$  is prime then, for all integers  $a$ ,  $a^p \equiv a \pmod{p}$ .*

*Proof* There are two cases. If  $a$  is a multiple of  $p$  then so is  $a^p$  and so

<sup>†</sup> According to H. Davenport, *The higher arithmetic, an introduction to the theory of numbers*, Cambridge University Press, Sixth edition 1992.

$a^p \equiv 0 \equiv a \pmod{p}$ . If  $a$  is not a multiple of  $p$  then it is immediate from the theorem that  $a^p \equiv a \pmod{p}$ .  $\square$

This theorem can be used to determine the congruence class of a high power modulo a prime.

**Example 24.1.3** To determine the congruence class of  $3^{203} \pmod{11}$ .

*Solution* By Fermat's little theorem  $3^{10} \equiv 1 \pmod{11}$ . Hence  $3^{203} = (3^{10})^{20} \times 3^3 \equiv 1 \times 27 \equiv 5 \pmod{11}$ .  $\square$

It should be observed that Fermat's little theorem does not necessarily give the minimal positive power of  $a$  which is congruent to 1 modulo  $p$  and this may be less than  $p - 1$ . This is clear when  $a = 1$ . For a more interesting example consider  $p = 7$ . Here  $2^3 \equiv 1 \pmod{7}$ .

The smallest positive integer  $n$  such that  $a^n \equiv 1 \pmod{p}$  is called the *order* of  $a$  modulo  $p$ . It is not difficult to deduce from Fermat's theorem that the order of any element divides  $p - 1$  (see Problems VI, Question 14).

It is also true, but harder to prove, that for each prime  $p$  there is some number  $a$  with order  $p - 1$  modulo  $p$ . Such a number is called a *primitive root* modulo  $p$ . The proof is another counting argument.<sup>†</sup>

## 24.2 Wilson's theorem

Another attractive result concerning divisibility by a prime  $p$  is the following.

**Theorem 24.2.1 (Wilson's theorem)** *If  $p$  is prime, then  $(p - 1)! \equiv -1 \pmod{p}$ .*

Equivalently,  $(p - 1)! + 1$  is divisible by  $p$ . The theorem is named after the English mathematician John Wilson who conjectured it on the basis of numerical calculations. He was a student of Edward Waring who included it in his book *Meditationes algebraicae* in 1770 suggesting that it would be very hard to prove. The following year the French mathematician Joseph Lagrange published a proof.

**Example 24.2.2** Consider the case of  $p = 11$ . Then  $10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2$ . We could determine the remainder after dividing

<sup>†</sup> See for example D.M. Burton, *Elementary number theory*, Allyn and Bacon, 1976.

by 11 by a direct calculation. However, notice the following congruences modulo 11:

$$10 \equiv -1;$$

$$9 \times 5 = 45 \equiv 1;$$

$$8 \times 7 = 56 \equiv 1;$$

$$6 \times 2 = 12 \equiv 1;$$

$$4 \times 3 = 12 \equiv 1.$$

Using these we see that  $10! = 10 \times (9 \times 5) \times (8 \times 7) \times (6 \times 2) \times (4 \times 3) \equiv (-1) \times 1 \times 1 \times 1 \times 1 = -1$  as claimed.

In this calculation we paired off each number between 2 and 9 with its inverse modulo 11. This method works for any prime modulus and provides a general proof.

*Proof of Theorem 24.2.1* For  $p = 2$ ,  $(p - 1)! = 1! = 1 \equiv -1 \pmod{2}$  and so the result is true.

Suppose now that  $p$  is prime and  $p \geq 3$ . Then each integer  $a$  such that  $1 \leq a \leq p - 1$  is coprime to  $p$  (since  $p$  is prime) and so by Theorem 20.1.5 there is a unique integer  $a' \in R_p$  such that  $aa' \equiv 1 \pmod{p}$ . Using the language of Definition 21.4.3,  $a'$  is an inverse of  $a$  modulo  $p$ .

If  $a = 1$  or  $a = p - 1$  then  $a' = a$ . For other values of  $a$ ,  $a' \neq a$ . To see this observe that

$$\begin{aligned} x^2 \equiv 1 \pmod{p} &\Leftrightarrow x^2 - 1 \equiv 0 \pmod{p} \\ &\Leftrightarrow (x - 1)(x + 1) \equiv 0 \pmod{p} \\ &\Leftrightarrow p \text{ divides } (x - 1)(x + 1) \\ &\Leftrightarrow p \text{ divides } x - 1 \text{ or } p \text{ divides } x + 1 \\ &\quad \text{since } p \text{ is prime (using Theorem 23.1.3)} \\ &\Leftrightarrow x \equiv 1 \pmod{p} \text{ or } x \equiv -1 \pmod{p}. \end{aligned}$$

Thus the  $p - 3$  integers  $2, 3, \dots, p - 2$  may be divided into pairs  $a$  and  $a'$  such that  $aa' \equiv 1 \pmod{p}$ . This means that their product  $2 \times 3 \times \dots \times (p - 2) \equiv 1 \pmod{p}$  and so  $(p - 1)! \equiv p - 1 \equiv -1 \pmod{p}$  as required.  $\square$

In fact Wilson's theorem provides a criterion for a number to be prime: an integer  $n > 1$  is prime if and only if  $(n - 1)! \equiv -1 \pmod{n}$  (see Exercise 24.4) but this is not a useful criterion since the calculation of  $(n - 1)!$  is very time-consuming if  $n$  is large.

### 24.3 Looking for primes

Fermat's little theorem gives some indication of how it might be possible to prove that a number is prime without trying all possible factors or that it is composite without actually factorizing it. Notice that the theorem can be restated in the following way.

**Corollary 24.3.1** *Given a positive integer  $n$ , if  $a^n \not\equiv a \pmod{n}$  for some integer  $a$ , then  $n$  is not prime.*

*Proof* This is simply the contrapositive of Corollary 24.1.2. □

**Example 24.3.2** For example, taking  $n = 63$ ,  $2^{63} = 2^{60} \times 2^3 = (2^6)^{10} \times 2^3 = (64)^{10} \times 8 \equiv 1^{10} \times 8 = 8 \pmod{63}$  and so  $2^{63} \not\equiv 2 \pmod{63}$ . Hence 63 is not prime.

It is of course easier to prove that 63 is composite by observing that  $63 = 7 \times 9$  and this particular result is not an efficient way of proving that a number is composite. It does, however, point the way to seeking more efficient tests which show that numbers are composite.

It is natural to ask whether we can do something similar to prove that a number is prime. Is there some sort of converse to Corollary 24.3.1?

Apparently, at one time the Chinese believed that, if  $2^n \equiv 2 \pmod{n}$ , then  $n$  is prime, but in fact this is false as can be seen by the following.

**Example 24.3.3**  $2^{341} \equiv 2 \pmod{341}$  but  $341 = 11 \times 31$  is not prime.

To see this observe that, by Fermat's theorem,  $2^{10} \equiv 1 \pmod{11}$ . Also  $2^5 = 32 \equiv 1 \pmod{31}$  so that  $2^{10} \equiv 1 \pmod{31}$ . This means that  $2^{10} - 1$  is divisible by 11 and by 31 and so these primes are amongst its prime factorization. Thus  $2^{10} - 1$  is a multiple of  $11 \times 31 = 341$ , i.e.  $2^{10} \equiv 1 \pmod{341}$ . It follows that  $2^{340} \equiv 1 \pmod{341}$  so that  $2^{341} \equiv 2 \pmod{341}$ .

This conjecture was perhaps a bit optimistic. More realistic might be to hope that, if  $a^{n-1} \equiv 1 \pmod{n}$  for all  $a \in \mathbb{Z}$  which are coprime to  $n$ , then  $n$  is prime. In fact this fails too.

**Example 24.3.4** We have  $a^{560} \equiv 1 \pmod{561}$  for all integers  $a$  coprime to 561 but  $561 = 3 \times 11 \times 17$  is not prime.

First of all notice that if  $a$  is coprime to 561, then  $(a, 3) = 1$ ,  $(a, 11) = 1$  and  $(a, 17) = 1$ . Then, by Fermat's theorem we have  $a^2 \equiv 1 \pmod{3}$  so that  $a^{560} \equiv 1 \pmod{3}$ ,  $a^{10} \equiv 1 \pmod{11}$  so that  $a^{560} \equiv 1 \pmod{11}$  and

$a^{16} \equiv 1 \pmod{17}$  so that  $a^{560} \equiv 1 \pmod{17}$ . This means that  $a^{560} - 1$  is divisible by 3, by 11 and by 17 and so, as in Example 24.3.3, it is divisible by  $561 = 3 \times 11 \times 17$ , i.e.  $a^{560} \equiv 1 \pmod{561}$ .

Nevertheless, these ideas can be developed to give tests for primality and this is an important area of active mathematical research.<sup>†</sup>

### Exercises

- 24.1** Find the remainder after  $2^{1000000}$  is divided by 17.
- 24.2** Suppose that  $p$  is an odd prime. Prove that  $a^{2p-1} \equiv a \pmod{2p}$ .
- 24.3** Use Fermat's theorem to find the inverse of 7 modulo 17 in  $R_{17}$ , the set of remainders modulo 17.
- 24.4** Prove that, for an integer  $n > 2$ , if  $(n-1)! \equiv -1 \pmod{n}$  then  $n$  is prime.
- 24.5** Prove that, if  $p$  and  $q$  are distinct prime numbers, then

$$p^{q-1} + q^{p-1} \equiv 1 \pmod{pq}.$$

<sup>†</sup> See K.H. Rosen, *Elementary number theory and its applications*, Addison-Wesley, Second edition 1988.

## Problems VI: Prime numbers

1. Find the prime factorizations of 3480 and 4284 and hence find their greatest common divisor. Compare the amount of work with the solution of Problems IV, Question 6(ii).
2. Prove that there are infinitely many prime numbers which are congruent to 3 modulo 4.
3. Let  $E \subseteq \mathbb{Z}$  be the set of even integers. Call an even integer *E-prime* if it is not the product of two other even integers. Prove that 60 can be written as the product of *E-primes* in two quite distinct ways.
4. Re-prove Corollary 17.2.1 using the fundamental theorem of arithmetic.
5. Re-prove the result of Problems IV, Question 13 that  $(a, b)[a, b] = ab$  for positive integers  $a$  and  $b$  using the fundamental theorem of arithmetic.
6. Find the number of positive divisors of the number  $p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$  (where  $p_1, p_2, \dots, p_r$  are distinct primes).
7. Prove that if  $a$  and  $b$  are coprime, and  $a$  and  $c$  are coprime, then  $a$  and  $bc$  are coprime.
8. Prove that, if  $a$  and  $b$  are coprime, then  $ab$  is a perfect square if and only if  $a$  and  $b$  are both perfect squares.

9. Prove that, if  $m_1$  and  $m_2$  are coprime positive integers, then

$$a \equiv b \pmod{m_1 m_2} \Leftrightarrow a \equiv b \pmod{m_1} \text{ and } a \equiv b \pmod{m_2}.$$

10. Use the inclusion-exclusion principle (Problems III, Question 4) to deduce from Exercise 23.5 that if  $n$  has the standard prime factorization  $n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$  then

$$\phi(n) = n \left(1 - \frac{1}{p_1}\right) \left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right)$$

where  $\phi$  is Euler's totient function.

11. Use the inclusion-exclusion principle to find the number of positive integers no greater than 100 which are perfect powers ( $n = m^r$  for positive integers  $m$  and  $r$  with  $r \geq 2$ ). Check the result of your calculation by listing these numbers. Do the same calculation for positive integers no greater than 1000.

12. Prove that if the polynomial equation  $a_n x^n + \dots + a_1 x + a_0 = 0$  (with integer coefficients) has a rational solution  $x = r/s$  (in its lowest terms) then  $r$  divides  $a_0$  and  $s$  divides  $a_n$ .

Deduce that every rational solution of  $x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0$  is an integer.

Hence show that the equation  $x^3 - 8x + a = 0$  has no rational solutions if  $a \equiv \pm 1 \pmod{5}$ .

13. Recall that, given a prime  $p$ , the order modulo  $p$  of an integer  $a \not\equiv 0 \pmod{p}$  is the least positive integer  $n$  such that  $a^n \equiv 1 \pmod{p}$ . Find the orders modulo  $p$  of all the non-zero elements of  $R_p$  for  $p = 5$  and  $p = 11$ .

14. Use Fermat's little theorem to prove, by contradiction, that for a prime  $p$  the order modulo  $p$  of each integer  $a \not\equiv 0 \pmod{p}$  divides  $p - 1$ .

[Try a proof by contradiction using the division theorem.]

15. Suppose that  $p$  is prime and does not divide the integer  $a$ . Prove that in the infinite decimal representation of the rational number  $a/p$  the number of recurring digits is given by the order of 10 modulo  $p$  (and so by the previous question divides  $p - 1$ ).

**16.** Find the remainders when

- (i)  $29!$  is divided by  $31$ ,
- (ii)  $18!$  is divided by  $23$ ,
- (iii)  $18!$  is divided by  $437$ .

**17.** [Euler's proof of Fermat's little theorem.] Suppose that  $p$  is a prime number. Prove that, for  $0 < i < p$ , the binomial coefficient  $\binom{p}{i}$  is a multiple of  $p$ .

[By Theorem 12.2.10  $p! = i!(p-i)!\binom{p}{i}$ . Now use the fundamental theorem of arithmetic.]

Deduce that  $(a+b)^p \equiv a^p + b^p \pmod{p}$ . Hence prove, by induction on  $a$ , that  $a^p \equiv a$  for all  $a \in \mathbb{Z}$ . Deduce Fermat's little theorem.

**18.** [Gauss's proof of Fermat's little theorem.] Let  $p$  be a prime and  $a$  a non-zero element of  $R_p$ .

(i) Prove that there exists a positive integer  $m$  such that  $a^m \equiv 1 \pmod{p}$  (and so there is a least such, the order of  $a$  modulo  $p$ ).

[Notice that the set  $\{[a^k]_p \mid k \in \mathbb{Z}^+\}$  is a subset of  $\mathbb{Z}_p$  and so finite.]

(ii) Prove that an equivalence relation may be defined on  $R_p - \{0\}$  by  $x_1 \sim x_2 \Leftrightarrow x_1 \equiv x_2 a^k$  for some  $k \in \mathbb{Z}^+$ .

(iii) Prove that the number of elements in each equivalence class is  $n$ , the order of  $a$ .

(iv) Prove that  $n$  divides  $p-1$  (proving the result of Question 14 without using Fermat's theorem).

(v) Deduce Fermat's theorem.



## Solutions to exercises

The reader should not imagine that the solutions which follow are the only possible correct solutions of the exercises. Of course for some, such as completing the truth tables in Exercises 1, there is a unique ‘right answer’ and anything else is wrong. But most exercises ask for a proof and for these there may be more than one possible proof and in any case the same proof may be set out in a variety of ways. If the reader’s solution is different from the author’s then the reader should pause and check whether both solutions are correct.

### Exercises 1 (page 8)

| 1.1 | $P$ | $Q$ | $P$ and $Q$ |
|-----|-----|-----|-------------|
|     | $T$ | $T$ | $T$         |
|     | $T$ | $F$ | $F$         |
|     | $F$ | $T$ | $F$         |
|     | $F$ | $F$ | $F$         |

| 1.2 (i) | $P$ | $Q$ | $P$ and $Q$ | not ( $P$ and $Q$ ) |
|---------|-----|-----|-------------|---------------------|
|         | $T$ | $T$ | $T$         | $F$                 |
|         | $T$ | $F$ | $F$         | $T$                 |
|         | $F$ | $T$ | $F$         | $T$                 |
|         | $F$ | $F$ | $F$         | $T$                 |

| (ii) | $P$ | $Q$ | not $P$ | not $Q$ | (not $P$ ) or (not $Q$ ) |
|------|-----|-----|---------|---------|--------------------------|
|      | $T$ | $T$ | $F$     | $F$     | $F$                      |
|      | $T$ | $F$ | $F$     | $T$     | $T$                      |
|      | $F$ | $T$ | $T$     | $F$     | $T$                      |
|      | $F$ | $F$ | $T$     | $T$     | $T$                      |

| (iii) and (iv) |     | $P$ | $Q$ | not $P$ | not $Q$ | $P$ and (not $Q$ ) | (not $P$ ) or $Q$ |
|----------------|-----|-----|-----|---------|---------|--------------------|-------------------|
| $T$            | $T$ | $F$ | $F$ |         |         | $F$                | $T$               |
| $T$            | $F$ | $F$ | $T$ |         |         | $T$                | $F$               |
| $F$            | $T$ | $T$ | $F$ |         |         | $F$                | $T$               |
| $F$            | $F$ | $T$ | $T$ |         |         | $F$                | $T$               |

Notice that the final columns of the first two tables are the same. This tells us that the statements ‘not ( $P$  and  $Q$ )’ and ‘(not  $P$ ) or (not  $Q$ )’ are logically equivalent, or have the same meaning.

| 1.3 | $a$ | $b$ | $a < b$ | $a = b$ | $a \leq b$ |
|-----|-----|-----|---------|---------|------------|
|     | 1   | 1   | $F$     | $T$     | $T$        |
|     | 2   | 1   | $F$     | $F$     | $F$        |
|     | 1   | 2   | $T$     | $F$     | $T$        |
|     | -2  | 1   | $T$     | $F$     | $T$        |

1.4 The easiest way of thinking about this question is to imagine filling in a table as follows.

|         |      |       |
|---------|------|-------|
| good    | 1    | 2     |
| neither | 3    | 4     |
| bad     | 5    | 6     |
|         | boys | girls |

The idea is that each child will be placed in this table according to sex and mathematical ability (we assume that the assessment of ability does not present any problems). Thus in region 1 is the list of boys who are good at mathematics, in region 2 is the list of girls who are good at mathematics, region 2 together with region 4 provides the list of girls who are not bad at mathematics, and so on. We can now interpret each of the given statements as statements about this table.

- (i) Regions 4 and 6 are empty.
- (ii) Regions 2 and 4 are empty.
- (iii) Region 2 is empty.
- (iv) Region 6 is not empty.
- (v) Regions 4 and 6 are not both empty.
- (vi) Region 1 is empty.
- (vii) Regions 4 and 6 are empty.

From this we see that statement (v) is the negation of statement (i) whereas statement (vii) is logically equivalent to statement (i).

**1.5** Recall that  $|a|$  is defined by  $|a| = a$  if  $a \geq 0$  and  $|a| = -a$  if  $a \leq 0$ . So we consider these two cases separately.

If  $a \geq 0$ , then  $|a|^2 = a^2$  by definition.

If  $a \leq 0$ , then  $|a|^2 = (-a)^2 = a^2$ .

Since for every real number  $a$ ,  $a \geq 0$  or  $a \leq 0$  we have proved that  $|a|^2 = a^2$ .

Notice that if  $a = 0$  then  $a \geq 0$  and  $a \leq 0$ . This is not a problem since in this case both definitions ( $a$  and  $-a$ ) give 0. We say that the modulus of a real number is *well-defined* (see Example 8.1.8).

### Exercises 2 (page 19)

**2.1** (i) True. (ii) False. (iii) True. (iv) False. (v) False. (vi) True. (vii) True. (viii) True. (ix) False. (x) True. (xi) True.

Notice that statement (viii) is true because the statement ' $n = 2$  and  $n = -1$ ' is false for all integers  $n$ .

[In fact there is a certain ambiguity about the last three parts of this question. Consider statement (ix). As explained in the text, the statement  $n^2 - n - 2 = 0 \Rightarrow n = 2$  is strictly speaking a predicate giving a different proposition for each value of  $n$ . It is usual to interpret such a predicate as a universal statement meaning that the implication is true whatever value is assigned to the free variable  $n$ . This particular universal implication is false since, when  $n = -1$ ,  $n^2 - n - 2 = 0$  but  $n \neq 2$ . In the same way the universal implication  $n^2 - n - 2 = 0 \Rightarrow n = -1$  is false. This leads to the conclusion that statement (ix) is false and this would be the normal interpretation.]

However the reader should be aware that statement (ix) can be viewed differently. The whole statement can be viewed as a predicate with free variable  $n$  and we can consider whether this predicate is true whatever integer value is assigned to  $n$ . This is indeed the case for, although the first implication is false when  $n = -1$  and the second one is false when  $n = 2$ , one or other of them is true whatever value is assigned to  $n$ .

The precise meaning of more complicated universal statements such as this may be made clear by using the quantifier ' $\forall$ ' to be discussed in Chapter 7: the symbols ' $\forall n$ ' are read 'for all  $n$ '. Statement (ix) is false if it is interpreted as

$$(\forall n, n^2 - n - 2 = 0 \Rightarrow n = 2) \text{ or } (\forall n, n^2 - n - 2 = 0 \Rightarrow n = -1).$$

It is true if it is interpreted as

$$\forall n, (n^2 - n - 2 = 0 \Rightarrow n = 2) \text{ or } (n^2 - n - 2 = 0 \Rightarrow n = -1).$$

Statements (x) and (xi) are true whichever way they are interpreted.]

**2.2** Notice that if  $n$  is an integer then either  $n \leq 0$  or  $n \geq 1$ .

|            | $n > 0$ | $n \geq 1$ | $n > 0 \Rightarrow n \geq 1$ |
|------------|---------|------------|------------------------------|
| $n \leq 0$ | F       | F          | T                            |
| $n \geq 1$ | T       | T          | T                            |

Here the last column follows from the previous two using the truth table for ‘implies’. Since both entries in the last column are ‘T’ we have proved the universal implication  $n > 0 \Rightarrow n \geq 1$  for integers  $n$ .

From the discussion in the chapter, the corresponding universal implication for a real number is false (see Table 2.1.2). This illustrates the importance of being clear about the range of values of the variables in a predicate.

| 2.3 | $P$ | $Q$ | $P \Rightarrow Q$ | $Q \Rightarrow P$ | $P \Leftrightarrow Q$ |
|-----|-----|-----|-------------------|-------------------|-----------------------|
|     | T   | T   | T                 | T                 | T                     |
|     | T   | F   | F                 | T                 | F                     |
|     | F   | T   | T                 | F                 | F                     |
|     | F   | F   | T                 | T                 | T                     |

| 2.4 | $P$ | $Q$ | $P \text{ or } Q$ | $P \text{ and } Q$ | $P \Rightarrow (P \text{ or } Q)$ | $(P \text{ and } Q) \Rightarrow P$ |
|-----|-----|-----|-------------------|--------------------|-----------------------------------|------------------------------------|
|     | T   | T   | T                 | T                  | T                                 | T                                  |
|     | T   | F   | T                 | F                  | T                                 | T                                  |
|     | F   | T   | T                 | F                  | T                                 | T                                  |
|     | F   | F   | F                 | F                  | T                                 | T                                  |

Since all entries in the final two columns are ‘T’ these statements are true regardless of the truth values of the statements  $P$  and  $Q$ .

| 2.5 (i) | $P$ | $Q$ | $P \Rightarrow Q$ | not $Q$ | not $P$ | $(\text{not } Q) \Rightarrow (\text{not } P)$ |
|---------|-----|-----|-------------------|---------|---------|---|
|         | T   | T   | T                 | F       | F       | T   |
|         | T   | F   | F                 | T       | F       | F   |
|         | F   | T   | T                 | F       | T       | T   |
|         | F   | F   | T                 | T       | T       | T   |

Since the entries in the third and sixth columns are identical the corresponding statements are logically equivalent: an implication is equivalent to its contrapositive.

| (ii) | $P$ | $Q$ | $P \text{ or } Q$ | $\text{not } P$ | $(\text{not } P) \Rightarrow Q$ |
|------|-----|-----|-------------------|-----------------|---------------------------------|
|      | $T$ | $T$ | $T$               | $F$             | $T$                             |
|      | $T$ | $F$ | $T$               | $F$             | $T$                             |
|      | $F$ | $T$ | $T$               | $T$             | $T$                             |
|      | $F$ | $F$ | $F$               | $T$             | $F$                             |

Since the entries in the third and-fifth columns are identical the corresponding statements are logically equivalent.

**2.6** We must prove that there is no integer  $k$  such that  $7k = 100$ . For any integer  $k$ ,  $k \leq 14$  or  $k \geq 15$ . Hence  $7k \leq 98$  or  $7k \geq 105$  so that  $7k \neq 100$ .

### Exercises 3 (page 29)

**3.1** (i) We simplify each side by multiplying out.

$$\begin{aligned} (a+b-c)^2 &= a^2 + b^2 + c^2 + 2ab - 2ac - 2bc, \\ (a+b)^2 + (a-c)^2 + (b-c)^2 - a^2 - b^2 - c^2 &= a^2 + b^2 + c^2 + 2ab - 2ac - 2bc. \end{aligned}$$

Therefore  $(a+b-c)^2 = (a+b)^2 + (a-c)^2 + (b-c)^2 - a^2 - b^2 - c^2$ .

(ii) As with Proposition 3.2.1 we start with the goal and work backwards taking care to get the implications in the right direction.

$$\begin{aligned} bc + ac + ab \leq a^2 + b^2 + c^2 &\Leftarrow 0 \leq a^2 + b^2 + c^2 - bc - ac - ab \\ &\Leftarrow 0 \leq 2a^2 + 2b^2 + 2c^2 - 2bc - 2ac - 2ab \\ &\Leftarrow 0 \leq (b-c)^2 + (a-c)^2 + (a-b)^2 \\ &\Leftarrow 0 \leq (b-c)^2 \text{ and } 0 \leq (a-c)^2 \\ &\quad \text{and } 0 \leq (a-b)^2 \end{aligned}$$

and this final statement is certainly true for all real numbers  $a$ ,  $b$  and  $c$ . Hence  $bc + ac + ab \leq a^2 + b^2 + c^2$ .

**3.2** This is a case of spelling out the definitions. Suppose that  $a$  divides  $b$  and  $b$  divides  $c$ . Then  $b = aq$  for some integer  $q$  and  $c = bp$  for some integer  $p$ . [Notice that we cannot use the same letter  $q$  in both cases since there is no reason to expect that these two numbers will be equal.] Hence  $c = bp = (aq)p = a(qp)$  so that  $a$  divides  $c$  since  $qp$ , the product of two integers, is an integer. Thus, if  $a$  divides  $b$  and  $b$  divides  $c$  then  $a$  divides  $c$ .

**3.3** This can be proved directly from the definitions as follows.

Suppose that  $n$  is an even integer. Then  $n = 2q$  for some integer  $q$ . Therefore  $n^2 = (2q)^2 = 4q^2 = 2(2q^2)$  and so, since  $2q^2$  is an integer,  $n^2$  is even. Hence, if  $n$  is even then  $n^2$  is even.†

Alternatively, this may be deduced from the previous result since, for any even integer  $a$ , 2 divides  $a$  and  $a$  divides  $a^2$ .

**3.4** Again from the definitions, if 0 divides  $a$ , then there is an integer  $q$  such that  $a = 0 \times q$ . Since  $0 \times q = 0$  this means that  $a = 0$ .

Conversely,  $0 = 0 \times 0$  and so 0 divides 0.

**3.5** Remember that  $b \geq c$  is shorthand for  $b > c$  or  $b = c$  so we must prove that  $b > c \Rightarrow ab \geq ac$  and  $b = c \Rightarrow ab \geq ac$ . The first follows from the inequality axioms:  $b > c \Rightarrow ab > ac$  (since  $a > 0$ )  $\Rightarrow ab \geq ac$ . The second is even easier:  $b = c \Rightarrow ab = ac \Rightarrow ab \geq ac$ .

[Notice here the methods for using and proving ‘or’ statements. To use an ‘or’ statement it is necessary to show that each component statement separately leads to the desired conclusion. To prove an ‘or’ statement it is sufficient to prove one of the component statements. Proofs of ‘or’ statements are considered further in the next chapter.]

**3.6** Given negative real numbers  $a$  and  $b$  suppose that  $a < b$ . Then  $a^2 > ab$  (multiplying through by  $a < 0$ ) and  $ab > b^2$  (multiplying through by  $b < 0$ ). Hence  $a^2 > b^2$ . It follows that  $a < b \Rightarrow a^2 > b^2$ .

[This proof mimics the proof of Proposition 3.1.1. We could have deduced the result from Proposition 3.1.1 by observing that if  $a$  and  $b$  are negative then  $-a$  and  $-b$  are positive and that  $a < b \Rightarrow -b < -a$ . The reader should write out the details of this proof.]

**3.7** The condition is not necessary. To prove that  $4ab < (a+b)^2 \nRightarrow a < b$  it is sufficient to give a single example of numbers  $a$  and  $b$  for which the implication fails, say  $a = 2$ ,  $b = 1$  for which  $4ab = 8 < (a+b)^2 = 9$  but  $a \not< b$ . On examining the proof of Proposition 3.2.1 we find that each implication is reversible apart from the last step. This means that  $4ab < (a+b)^2 \Leftrightarrow a \neq b$  and so a necessary and sufficient condition is  $a \neq b$ .

† This proof is discussed in more detail in Chapter 7 (Proposition 7.3.2).

**3.8** (i) Suppose that  $a$  and  $b$  are real numbers such that  $0 \leq a < b$ ; then either  $a = 0$  or  $a > 0$ . If  $a > 0$  then we have already proved (Proposition 3.1.1) that  $a^2 < b^2$ . If  $a = 0$  then  $a^2 = 0$  and since  $b > 0$  implies that  $b^2 > 0$  we again have  $a^2 < b^2$ . So in either case  $a^2 < b^2$ . Hence  $0 \leq a < b \Rightarrow a^2 < b^2$ .

(ii) By definition the modulus of any real number is non-negative. Hence, if  $|a| < |b|$  then  $0 \leq |a| < |b|$  and so by the first part  $|a|^2 < |b|^2$ . But (by Exercise 1.5)  $|a|^2 = a^2$  and  $|b|^2 = b^2$  so that  $a^2 < b^2$  as required.

(iii)  $|a| = |b| \Rightarrow a^2 = |a|^2 = |b|^2 = b^2$  (using Exercise 1.5 again).

(iv) If  $|a| \leq |b|$  then  $|a| < |b|$  or  $|a| = |b|$ . In the first case  $a^2 < b^2$  by (ii) and in the second case  $a^2 = b^2$  by (iii) and so in either case  $a^2 \leq b^2$ . Thus  $|a| \leq |b| \Rightarrow a^2 \leq b^2$  as required.

### Exercises 4 (page 37)

**4.1** Suppose for contradiction that  $m$  and  $n$  are integers such that  $14m + 21n = 100$ . Then 7 divides 14 and 7 divides 21 so that, since  $100 = 7(2m + 3n)$ , 7 divides 100. But 7 does not divide 100 [by Exercise 2.6 if a proof is insisted upon] and so we have a contradiction. Hence such integers  $m$  and  $n$  do not exist.

**4.2** Given that  $n$  is an integer such that  $n^2$  is odd, suppose for contradiction that  $n$  is even. Then, by Exercise 3.3,  $n^2$  is even, contradicting the statement that  $n^2$  is odd. Hence our assumption that  $n$  is even must be false, i.e.  $n$  is odd. Thus  $n^2$  is odd  $\Rightarrow n$  is odd, as required.

**4.3** The contrapositive of the required result reads

$$n \text{ is even} \Rightarrow n^2 \text{ is even.}$$

This is true by Exercise 3.3 and so the required result is true.

**4.4** Suppose that  $a$  is an integer such that  $a^2 \geq 7a$ . To see that it follows that  $a \leq 0$  or  $a \geq 7$  suppose that  $a \not\leq 0$ . Then  $a > 0$  so that  $a^2 \geq 7a \Rightarrow a \geq 7$  by dividing through by the positive number  $a$ . Hence  $a^2 \geq 7a \Rightarrow a \leq 0$  or  $a \geq 7$ .

**4.5** Suppose that  $a$  and  $b$  are real numbers such that  $a \leq b$  and  $b \leq a$ . Suppose in addition, for contradiction, that  $a \neq b$ . Then, by the trichotomy law (Axioms 3.1.2(i))  $a < b$  or  $a > b$ . But  $a < b$  contradicts

our hypothesis that  $b \leq a$  and  $a > b$  contradicts our hypothesis that  $a \leq b$  and so in either case we have a contradiction as required. Thus our assumption that  $a \neq b$  must be false and so  $a = b$  as required.

**4.6** The contrapositive of Exercise 3.8(iv) reads

$$a^2 > b^2 \Rightarrow |a| > |b|.$$

(i) Interchanging  $a$  and  $b$  this gives  $a^2 < b^2 \Rightarrow |a| < |b|$  and so combining this with Exercise 3.8(ii) we have proved  $|a| < |b| \Leftrightarrow a^2 < b^2$  as required.

(ii) We have already proved (Exercise 3.8(iii)) that  $|a| = |b| \Rightarrow a^2 = b^2$ . For the converse, suppose that  $a^2 = b^2$  and, for contradiction, that  $|a| \neq |b|$ . Then, by trichotomy, either  $|a| < |b|$  in which case  $a^2 < b^2$  giving a contradiction, or  $|a| > |b|$  in which case  $a^2 > b^2$  giving a contradiction. Since we necessarily obtain a contradiction our assumption that  $|a| \neq |b|$  must be false and so  $|a| = |b|$  as required to prove that  $a^2 = b^2 \Rightarrow |a| = |b|$ .

**4.7** Using the previous question,  $|a + b| \leq |a| + |b| \Leftrightarrow (a + b)^2 \leq (|a| + |b|)^2 \Leftrightarrow a^2 + 2ab + b^2 \leq a^2 + 2|ab| + b^2 \Leftrightarrow ab \leq |ab|$  which is always true. From a similar argument with ' $\leq$ ' replaced by '=' throughout, we see that equality holds if and only if  $ab = |ab|$  and so if and only if  $ab \geq 0$ .

[This result is known as the ‘triangle inequality’ because when generalized to complex numbers it corresponds to the geometric fact that the sum of the lengths of any two sides of a triangle is not less than the length of the third side.]

### Exercises 5 (page 51)

**5.1** In this case  $P(n)$  is the statement ‘3 divides  $(n^3 - n)$ ’ or ‘ $n^3 - n = 3q$  for some integer  $q$ ’.

For the base case, the statement  $P(1)$  simply says that 3 divides 0 which is certainly true since  $0 = 3 \times 0$ .

We can sum up what is needed for the inductive step as follows.

| Given   | Goal                            |
|---|---------------------------------|
| $k$ a positive integer<br>3 divides $k^3 - k$ | 3 divides $(k + 1)^3 - (k + 1)$ |

If we spell out what divisibility means this becomes the following.

| Given   | Goal   |
|---|--|
| $k$ a positive integer<br>$k^3 - k = 3q$ for some integer $q$ | $(k+1)^3 - (k+1) = 3p$<br>for some integer $p$ |

[Notice here that we should not use the letter  $q$  in both the given and the goal statements about divisibility by 3, since there is no reason to expect that they will be the same integer (in fact they can't be). So the letter  $p$  is used in the goal statement.]

This suggests that we let  $q$  be an integer such that  $k^3 - k = 3q$  and try to find the  $p$  we want in terms of it. So we manipulate  $(k+1)^3 - (k+1)$  in the hope that we can relate it to  $k^3 - k$ .

$$\begin{aligned} (k+1)^3 - (k+1) &= k^3 + 3k^2 + 3k + 1 - k - 1 \\ &= (k^3 - k) + 3(k^2 + k) \\ &= 3(q + k^2 + k). \end{aligned}$$

This shows that the number we require is given by  $p = q + k^2 + k$ .

We now write this out formally using the template of section 3.1.

*Proof* We use induction on  $n$ .

Base case: For  $n = 1$ ,  $n^3 - n = 1 - 1 = 0 = 3 \times 0$  and so 3 divides  $n^3 - n$ .

Inductive step: Suppose now as inductive hypothesis that 3 divides  $k^3 - k$  for some positive integer  $k$ . Then  $k^3 - k = 3q$  for some integer  $q$ . Hence  $(k+1)^3 - (k+1) = k^3 + 3k^2 + 3k + 1 - k - 1 = (k^3 - k) + 3(k^2 + k) = 3(q + k^2 + k)$  and so 3 divides  $(k+1)^3 - (k+1)$ , as required.

Conclusion: Hence, by induction,  $n$  divides  $n^3 - n$  for all positive integers  $n$ .  $\square$

**5.2** So far statements proved by induction have used the symbol ' $n$ ' for the free variable in the predicate. This is the most common symbol for an integer free variable but other symbols can be used – this question uses the symbol ' $m$ '.

The inductive step requires us to prove that, for  $k \geq 10$ ,

$$k^3 \leq 2^k \Rightarrow (k+1)^3 \leq 2^{k+1}.$$

To prove this notice first of all that  $k^3 \leq 2^k \Rightarrow 2k^3 \leq 2^{k+1}$  giving us one side of the desired inequality. The desired conclusion will follow if we can prove that  $(k+1)^3 \leq 2k^3$ .

Now starting from the other side we can expand:  $(k+1)^3 = k^3 + 3k^2 + 3k + 1$ . Hence

$$\begin{aligned}(k+1)^3 \leq 2^{k+1} &\Leftarrow k^3 + 3k^2 + 3k + 1 \leq 2k^3 \\ &\Leftrightarrow 3k^2 + 3k + 1 \leq k^3.\end{aligned}$$

But if  $k \geq 10$ ,  $3k^2 + 3k + 1 \leq 3k^2 + 3k^2 + k^2 = 7k^2 \leq k^3$ .

*Proof* We use induction on  $m$ .

Base case: For  $m = 10$ ,  $m^3 = 1000$  and  $2^m = 1024$  and so  $m^3 \leq 2^m$ .

Inductive step: Suppose as inductive hypothesis that  $k^3 \leq 2^k$  for some  $k \geq 10$ . Then  $(k+1)^3 = k^3 + 3k^2 + 3k + 1 \leq k^3 + 3k^2 + 3k^2 + k^2$  (since  $k \geq 1$ ) =  $k^3 + 7k^2 \leq k^3 + k^3$  (since  $k \geq 7$ ) =  $2k^3 \leq 2 \times 2^k$  (by inductive hypothesis) =  $2^{k+1}$  and so the result holds for  $m = k+1$  as required.

Hence, by induction,  $m^3 \leq 2^m$  for all  $m \geq 10$ .  $\square$

[We only needed  $k \geq 7$  in proving the inductive step. Does this mean that the result holds for all  $m \geq 7$ ? Why not?]

### 5.3 We use induction on $n$ .

Base case: For  $n = 1$ ,  $n \geq 1$ .

Inductive step: Suppose as inductive hypothesis that  $k \geq 1$  for some positive integer  $k$ . Then since  $k+1 \geq k$  it follows that  $k+1 \geq 1$  proving the inductive step.

Hence by induction on  $n$ ,  $n \geq 1$  for all positive integers  $n$ .

[The reader may find this a curious exercise. However, although we have made use of this result in examples several times as if it were a 'self-evident truth' it was not included in the basic properties given earlier. It can be deduced from these basic properties, and this is how to do it.]

### 5.4 In this case the statement $P(n)$ to be proved for all integers $n \geq 0$ is $\sum_{i=0}^n x^i = (1 - x^{n+1})/(1 - x)$ .

Thus  $P(k)$  reads  $\sum_{i=0}^k x^i = (1 - x^{k+1})/(1 - x)$ , and  $P(k+1)$  reads  $\sum_{i=0}^{k+1} x^i = (1 - x^{(k+1)+1})/(1 - x)$  or, simplifying,  $\sum_{i=0}^{k+1} x^i = (1 - x^{k+2})/(1 - x)$ .

*Proof* We use induction on  $n$ .

Base case: For  $n = 0$ ,  $\sum_{i=0}^n x^i = x^0 = 1$  and  $(1 - x^{n+1})/(1 - x) = (1 - x)/(1 - x) = 1$  and so the result holds.

Inductive step: Suppose now as inductive hypothesis that  $\sum_{i=0}^k x^i = (1 - x^{k+1})/(1 - x)$  for some  $k \geq 0$ . Then  $\sum_{i=0}^{k+1} x^i = \sum_{i=0}^k x^i + x^{k+1} =$

$(1-x^{k+1})/(1-x) + x^{k+1}$  (by inductive hypothesis)  $= (1-x^{k+1}+x^{k+1}-x^{k+2})/(1-x) = (1-x^{k+2})/(1-x)$  as required.

Conclusion: Hence  $\sum_{i=0}^n x^i = (1-x^{n+1})/(1-x)$  for all  $n \geq 0$ .  $\square$

[This sum is known as a *geometric progression* with ratio  $x$ .]

**5.5** We prove this result by induction on  $n$ .

Base case: For  $n = 0$ ,  $\sum_{i=0}^n (a+ib) = a$  and  $\frac{1}{2}(n+1)(2a+nb) = \frac{1}{2} \times 2a = a$  and so the result holds.

Inductive step: Suppose as inductive hypothesis that  $\sum_{i=0}^k (a+ib) = \frac{1}{2}(k+1)(2a+kb)$  for some  $k \geq 0$ . Then  $\sum_{i=0}^{k+1} (a+ib) = \sum_{i=0}^k (a+ib) + (a+(k+1)b) = \frac{1}{2}(k+1)(2a+kb) + (a+kb+b) = \frac{1}{2}(2ka+k^2b+2a+kb+2a+2kb+2b) = \frac{1}{2}(2ka+k^2b+4a+3kb+2b) = \frac{1}{2}(k+2)(2a+kb+b) = \frac{1}{2}(k+2)(2a+(k+1)b)$  as required to prove the result for  $n = k + 1$ .

Conclusion: Hence the result holds for all  $n \geq 0$ .

Alternatively, this result can be deduced from Proposition 5.3.1 as follows.

$$\begin{aligned}\sum_{i=0}^n (a+ib) &= \sum_{i=1}^n a + b \sum_{i=1}^n i \\ &= (n+1)a + \frac{1}{2}n(n+1)b \\ &= \frac{1}{2}(n+1)(2a+bn).\end{aligned}$$

[This sum is known as an *arithmetic progression* with difference  $b$ .]

**5.6** We can prove this by induction on  $n$  starting at  $n = 0$ . In this case  $P(n)$  is simply the statement ' $u_n = n3^{n-1}$ '.

Base case: For  $n = 0$  the formula gives  $0 \times 3^{-1} = 0$  agreeing with the definition of  $u_0$ .

Inductive step: Now suppose that  $u_k = k3^{k-1}$  for some non-negative integer  $k$ . Then  $u_{k+1} = 3u_k + 3^k$  (by definition of the sequence)  $= 3k3^{k-1} + 3^k$  (by inductive hypothesis)  $= k3^k + 3^k = (k+1)3^k$  as required to prove the formula for  $n = k + 1$ .

Conclusion: Hence, by induction,  $u_n = n3^{n-1}$  for all non-negative integers  $n$ .

**5.7 (i)** The proof is by induction on  $n$  using the inductive definition of  $x^n$  in Definition 5.3.3.

Base case: For  $n = 0$ ,  $x^n y^n = 1 = (xy)^n$  as required.

Inductive step: Suppose as inductive hypothesis that  $x^k y^k = (xy)^k$  for some non-negative integer  $k$ . Then, by the inductive steps in the definitions of  $x^n$ ,  $y^n$  and  $(xy)^n$ ,  $x^{k+1} y^{k+1} = (x \cdot x^k)(y \cdot y^k) = (xy)(xy)^k = (xy)^{k+1}$  as required to prove the result for  $n = k + 1$ .

Conclusion: Hence, by induction,  $x^n y^n = (xy)^n$  for any real numbers  $x$  and  $y$  and non-negative integers  $n$ .

(ii) Base case: For  $n = 0$ ,  $x^{m+n} = x^m = x^m x^0 = x^m x^n$  as required.

Inductive step: Suppose as inductive hypothesis that  $x^{m+k} = x^m x^k$  for some non-negative integer  $k$ . Then  $x^{m+k+1} = x x^{m+k}$  (by the inductive step in the definition) =  $x x^m x^k$  (by inductive hypothesis) =  $x^m x x^k = x^m x^{k+1}$  (by the inductive step in the definition) as required to prove the result for  $n = k + 1$ .

Conclusion: Hence, by induction on  $n$ ,  $x^{m+n} = x^m x^n$  for any real number  $x$  and non-negative integers  $m$  and  $n$ .

[Notice that in order for this identity to hold for  $m = 1$  and  $n = 0$  it is necessary to define  $x^0 = 1$  for  $x \neq 0$ . The identity does not require any particular value for  $x^0$  when  $x = 0$ .]

(iii) Base case: For  $n = 0$ ,  $(x^m)^n = 1 = x^0 = x^{mn}$  as required.

Inductive step: Suppose as inductive hypothesis that  $(x^m)^k = x^{mk}$  for some non-negative integer  $k$ . Then  $(x^m)^{k+1} = x^m (x^m)^k$  (by the inductive step in the definition of  $(x^m)^n$ ) =  $x^m x^{mk}$  (by inductive hypothesis) =  $x^{m+mk}$  (by part (ii) of the question) =  $x^{m(k+1)}$  as required to prove the result for  $n = k + 1$ .

Conclusion: Hence, by induction on  $n$ ,  $(x^m)^n = x^{mn}$  for any real number  $x$  and non-negative integers  $m$  and  $n$ .

### Exercises 6 (page 72)

**6.1** (i)  $0 \not< 0$  and so  $0 \notin (0, 1)$ .

$0 \leq 0 \leq 1$  and so  $0 \in [0, 1]$ .

$0 \leq 0 < 1$  and so  $0 \in [0, 1)$ .

$0 \not< 0$  and so  $0 \notin (0, 1]$ .

(ii)  $[a, b] - (a, b) = \{a, b\}$ .

(iii) Since  $(a, b) = \emptyset$  is a non-existence statement it is hard to work with directly. So we prove this result by proving the equivalent contrapositive statement:

$$(a, b) \neq \emptyset \Leftrightarrow a < b.$$

$\Rightarrow$ : Suppose that  $(a, b) \neq \emptyset$ . Then we may choose an element  $x \in (a, b)$ . This means that  $a < x$  and  $x < b$  which implies that  $a < b$ .

$\Leftarrow$ : Suppose that  $a < b$ . Then  $2a < a+b < 2b$  so that  $a < (a+b)/2 < b$  which means that  $(a+b)/2 \in (a, b)$  so that  $(a, b) \neq \emptyset$ .

The equivalent statements for the other intervals are:

$$[a, b] = \emptyset \Leftrightarrow a > b;$$

$$[a, b) = \emptyset \Leftrightarrow a \geq b;$$

$$(a, b] = \emptyset \Leftrightarrow a \geq b.$$

(iv) Suppose that  $[a, b] \subseteq (c, d)$ . Then since  $a \in [a, b]$  we must have  $a \in (c, d)$  and so in particular  $c < a$ . Similarly  $b \in [a, b] \Rightarrow b \in (c, d) \Rightarrow b < d$ .

Conversely, if  $c < a$  and  $b < d$  then  $x \in [a, b] \Rightarrow a \leq x \leq b \Rightarrow c < x < d \Rightarrow x \in (c, d)$  and so  $[a, b] \subseteq (c, d)$ .

**6.2** (i) For a real number  $a$ ,  $a \in \{x \in \mathbb{R} \mid x^2 + x - 2 = 0\} \Leftrightarrow a^2 + a - 2 = 0 \Leftrightarrow (a-1)(a+2) = 0 \Leftrightarrow a-1 = 0$  or  $a+2 = 0 \Leftrightarrow a = 1$  or  $a = -2 \Leftrightarrow a \in \{1, -2\}$ .

Hence  $\{x \in \mathbb{R} \mid x^2 + x - 2 = 0\} = \{1, -2\}$ .

(ii) For  $a \in \mathbb{R}$ ,  $a \in \{x \in \mathbb{R} \mid x^2 + x - 2 < 0\} \Leftrightarrow a^2 + a - 2 < 0 \Leftrightarrow (a-1)(a+2) < 0 \Leftrightarrow (a-1 > 0 \text{ and } a+2 < 0) \text{ or } (a-1 < 0 \text{ and } a+2 > 0) \Leftrightarrow (a > 1 \text{ and } a < -2) \text{ or } (a < 1 \text{ and } a > -2) \Leftrightarrow a < 1 \text{ and } a > -2$  [Notice that there are no real numbers which are both less than  $-2$  and greater than  $1$ ]  $\Leftrightarrow a \in (-2, 1)$ .

Hence  $\{x \in \mathbb{R} \mid x^2 + x - 2 < 0\} = (-2, 1)$ .

An alternative proof would be to construct a table as follows –

|              | $x - 1$ | $x + 2$ | $x^2 + x - 2 = (x-1)(x+2)$ |
|--------------|---------|---------|----------------------------|
| $x < -2$     | –       | –       | +                          |
| $x = -2$     | –       | 0       | 0                          |
| $-2 < x < 1$ | –       | +       | –                          |
| $x = 1$      | 0       | +       | 0                          |
| $x > 1$      | +       | +       | +                          |

– from which it is immediate that  $x^2 + x - 2 < 0 \Leftrightarrow -2 < x < 1$ .

(iii) This is immediate from the above table or by trichotomy from parts (i) and (ii). Alternatively, a proof may be constructed by writing down a chain of equivalences as in parts (i) and (ii).

- 6.3** (i)  $\{3\} = \{x \in \mathbb{Z} \mid x = 3\}$ .  
(ii)  $\{1, 2, 3\} = \{x \in \mathbb{Z} \mid 0 < x < 4\}$ .  
(iii)  $\{1, 3\} = \{x \in \mathbb{Z} \mid 0 < x < 4 \text{ and } x \text{ is odd}\}$ .

| 6.4 | $x \in A$ | $x \in B$ | $x \in C$ | $x \in B \cup C$ | $x \in A \cap (B \cup C)$ |
|-----|-----------|-----------|-----------|------------------|---------------------------|
|     | T         | T         | T         | T                | T                         |
|     | T         | T         | F         | T                | T                         |
|     | T         | F         | T         | T                | T                         |
|     | T         | F         | F         | F                | F                         |
|     | F         | T         | T         | T                | F                         |
|     | F         | T         | F         | T                | F                         |
|     | F         | F         | T         | T                | F                         |
|     | F         | F         | F         | F                | F                         |

|  | $x \in A$ | $x \in B$ | $x \in C$ | $x \in A \cap B$ | $x \in A \cap C$ | $x \in (A \cap B) \cup (A \cap C)$ |
|--|-----------|-----------|-----------|------------------|------------------|------------------------------------|
|  | T         | T         | T         | T                | T                | T                                  |
|  | T         | T         | F         | T                | F                | T                                  |
|  | T         | F         | T         | F                | T                | T                                  |
|  | T         | F         | F         | F                | F                | F                                  |
|  | F         | T         | T         | F                | F                | F                                  |
|  | F         | T         | F         | F                | F                | F                                  |
|  | F         | F         | T         | F                | F                | F                                  |
|  | F         | F         | F         | F                | F                | F                                  |

Now since the final columns are the same we see that the two sets contain the same elements, i.e. are equal.

The Venn diagram for this result is shown on the next page; region  $i$  corresponds to line  $i$  in the above tables. We easily see that both sets correspond to regions 1, 2 and 3.

**6.5** It is best to prove the implications ' $\Rightarrow$ ' and ' $\Leftarrow$ ' separately in each case.

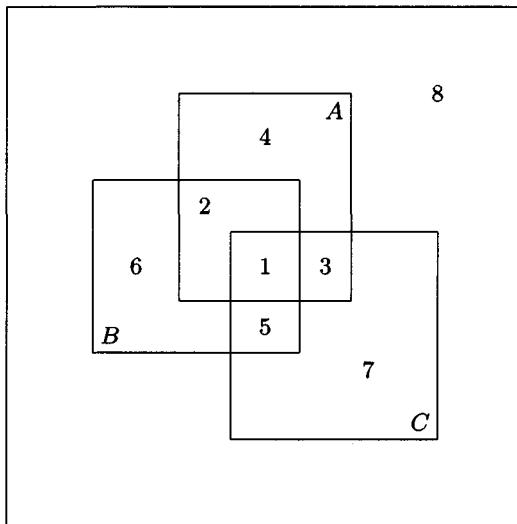
(i) ' $\RightarrowB \subseteq A \cup B$  since  $x \in B \Rightarrow x \in A$  or  $x \in B$ . Conversely, given that  $A \subseteq B$ ,  $x \in A \Rightarrow x \in B$  so that  $x \in A$  or  $x \in B \Rightarrow x \in B$ , i.e.  $A \cup B \subseteq B$ .

' $\Leftarrow$ ' This follows immediately from  $A \subseteq A \cup B$ .

(ii) ' $\RightarrowA \cap B \subseteq A$  since  $x \in A$  and  $x \in B \Rightarrow x \in A$ . Conversely, given that  $A \subseteq B$ ,  $x \in A \Rightarrow x \in B$  so that  $x \in A \Rightarrow x \in A$  and  $x \in B$ , i.e.  $A \subseteq A \cap B$ .

' $\Leftarrow$ ' This follows immediately from  $A \cap B \subseteq B$ .

Alternatively, this result follows from Problems I, Question 2.



**6.6** Let us start by summarizing what is needed.

| Given                               | Goal             |
|-------------------------------------|------------------|
| $A \cap B \subseteq C$<br>$x \in B$ | $x \notin A - C$ |

Since the goal is a negative statement this suggests a proof by contradiction as follows.

| Given  | Goal          |
|--|---------------|
| $A \cap B \subseteq C$<br>$x \in B$<br>$x \in A - C$ | contradiction |

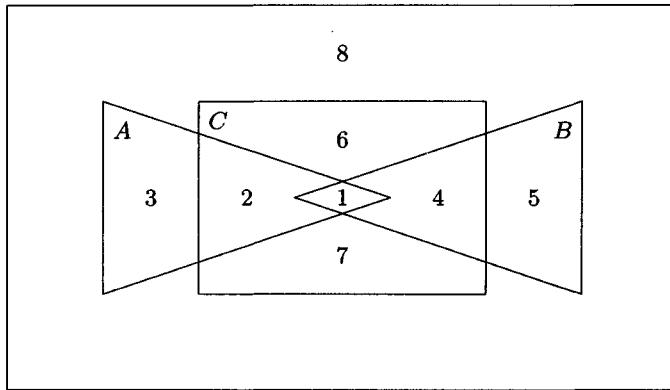
Now if we remember the definition of the difference of two sets this becomes the following.

| Given  | Goal          |
|--|---------------|
| $A \cap B \subseteq C$<br>$x \in B$<br>$x \in A$<br>$x \notin C$ | contradiction |

Now the proof is apparent for  $x \in B$  and  $x \in A \Rightarrow x \in A \cap B \Rightarrow x \in C$  (using  $A \cap B \subseteq C$ ) which contradicts  $x \notin C$ . This can be written out formally as follows.

*Proof* Suppose that  $A \cap B \subseteq C$  and  $x \in B$ . Suppose for contradiction that  $x \in A - C$ . Then  $x \in A$  and so, since also  $x \in B$ , we have  $x \in A \cap B$  which implies that  $x \in C$ . But  $x \in A - C$  implies that  $x \notin C$  and so we have a contradiction. Hence our assumption must be false and  $x \notin A - C$  as required.  $\square$

A Venn diagram illustrating this result is given below.



In the diagram the set  $A - C$  is given by region 3 which clearly contains no points of  $B$  (regions 1, 4 and 5).

Notice that although this seems the natural way of drawing this diagram two regions 6 and 7 together make up  $C - (A \cup B)$  and there is no logical distinction between them. This is a case where the geometry of the diagram does not correspond to anything in the set theory. It indicates why proofs using Venn diagrams are not really entirely satisfactory and the best proof of this result is that given above.

**6.7** Recall that  $A \subseteq B$  is defined to mean

$$x \in A \Rightarrow x \in B. \quad (1)$$

Hence  $B^c \subseteq A^c$  means

$$x \in B^c \Rightarrow x \in A^c$$

or equivalently

$$x \notin B \Rightarrow x \notin A. \quad (2)$$

The two statements (1) and (2) are contrapositives of each other and so equivalent. Hence  $A \subseteq B \Leftrightarrow B^c \subseteq A^c$ .

**Exercises 7 (page 86)**

**7.1** (i)  $\{m \in \mathbb{Z}^+ \mid \exists n \in \mathbb{Z}^+, m \leq n\} = \mathbb{Z}^+$ : given  $m \in \mathbb{Z}^+$  take  $n = m$  to prove that  $\exists n \in \mathbb{Z}^+, m \leq n$ .

(ii)  $\{m \in \mathbb{Z}^+ \mid \forall n \in \mathbb{Z}^+, m \leq n\} = \{1\}$ : it is true that  $\forall n \in \mathbb{Z}^+, 1 \leq n$  but for  $m \neq 1$  a counterexample to  $\forall n \in \mathbb{Z}^+, m \leq n$  is provided by  $n = 1$  since  $m \not\leq 1$ ;

(iii)  $\{n \in \mathbb{Z}^+ \mid \exists m \in \mathbb{Z}^+, m \leq n\} = \mathbb{Z}^+$ : given  $n \in \mathbb{Z}^+$  take  $m = n$  to prove that  $\exists m \in \mathbb{Z}^+, m \leq n$ .

(iv)  $\{n \in \mathbb{Z}^+ \mid \forall m \in \mathbb{Z}^+, m \leq n\} = \emptyset$ : given  $n \in \mathbb{Z}^+$  a counterexample to  $\forall m \in \mathbb{Z}^+, m \leq n$  is provided by  $m = n + 1$  since  $n + 1 \not\leq n$ .

**7.2** (i) False: a counterexample is  $m = 2, n = 1$ .

(ii) True: an example is  $m = n = 1$ .

(iii) True: see 7.1(i).

(iv) True: see 7.1(ii).

(v) True: see 7.1(iii).

(vi) False: see 7.1(iv).

**7.3** (i) False: a counterexample is  $m = 2, n = 1$ .

(ii) True: an example is  $m = n = 1$ .

(iii) True: given  $m \in \mathbb{Z}$  put  $n = m$ .

(iv) False: for each  $m \in \mathbb{Z}$  a counterexample is given by  $n = m - 1$ .

(v) True: given  $n \in \mathbb{Z}$  put  $m = n$ .

(vi) False: for each  $n \in \mathbb{Z}$  a counterexample is given by  $m = n + 1$ .

**7.4** (i) True: for each  $x \in \mathbb{R}$ , take  $y = -x$ .

(ii) False: for each  $y \in \mathbb{R}$ , the predicate is false for  $x = 1 - y$ .

(iii) True: take  $y = 0$ .

(iv) True: take  $y = 0$ .

(v) False: take  $x = 0$ .

(vi) False: take  $x = 0$ .

(vii) True: since ‘odd’ means ‘not even’.

(viii) False: ' $\forall n, n$  is even' is false, e.g.  $n = 1$ , and ' $\forall n, n$  is odd' is false, e.g.  $n = 0$ .

**7.5** Suppose that  $n \in \mathbb{Z}^+$  is such that  $\exists q \in \mathbb{Z}^+, n = 2q + 1$ . Then we can take  $q \in \mathbb{Z}^+$  such that  $n = 2q + 1$ . This means that  $n^2 = (2q + 1)^2 = 4q^2 + 4q + 1 = 2(2q^2 + 2q) + 1$  and so  $n^2 = 2p + 1$  where  $p = 2q^2 + 2q$  so that  $\exists p \in \mathbb{Z}^+, n^2 = 2p + 1$  as required.

[This proof illustrates how we *use* a statement involving  $\exists$  simply by using an element with the stated property and that we *prove* a statement involving  $\exists$  by constructing an element with the stated property.]

**7.6** We can write the statement as follows.

$$\begin{aligned} \forall a, b \in \mathbb{Z}, [(\exists q_1 \in \mathbb{Z}, a = 2q_1) \text{ and } (\exists q_2 \in \mathbb{Z}, b = 2q_2)] \\ \Rightarrow \exists p \in \mathbb{Z}, a + b = 2p. \end{aligned}$$

Suppose that  $a$  and  $b$  are even. Then, by definition,  $a = 2q_1$  and  $b = 2q_2$  for integers  $q_1$  and  $q_2$ . Therefore  $a + b = 2q_1 + 2q_2 = 2(q_1 + q_2)$  and so  $a + b$  is even since  $q_1 + q_2$  is an integer. Hence, if  $a$  and  $b$  are even then  $a + b$  is even.

**7.7** Suppose that  $(x, y) \in (A \times B) \cup (C \times D)$ . Then  $(x, y) \in A \times B$  or  $(x, y) \in C \times D$ . We now examine these two cases in turn.

If  $(x, y) \in A \times B$ , then  $x \in A$  and  $y \in B$ . Thus  $x \in A \cup C$  and  $y \in B \cup D$ . Hence  $(x, y) \in (A \cup C) \times (B \cup D)$  as required. On the other hand, if  $(x, y) \in C \times D$  an almost identical argument demonstrates that  $(x, y) \in (A \cup C) \times (B \cup D)$ .

Thus  $(x, y) \in (A \times B) \cup (C \times D) \Rightarrow (x, y) \in (A \cup C) \times (B \cup D)$  as required.

To give a counterexample to the equality of these sets we try a case of small sets just to see what happens – and it works.

Suppose that all four sets consist of just a single element:  $A = \{a\}$ ,  $B = \{b\}$ ,  $C = \{c\}$  and  $D = \{d\}$ . Then  $A \times B = \{(a, b)\}$  and  $C \times D = \{(c, d)\}$  so that the left-hand side  $(A \times B) \cup (C \times D) = \{(a, b), (c, d)\}$ . However,  $A \cup C = \{a, c\}$  and  $B \cup D = \{b, d\}$  so that  $(A \cup C) \times (B \cup D) = \{(a, b), (a, d), (c, b), (c, d)\}$ . Thus  $(A \cup C) \times (B \cup D)$  is not a subset of  $(A \times B) \cup (C \times D)$  since it contains the element  $(a, d)$ .

**7.8** (i) False. (ii) False. (iii) True. (iv) False.

**7.9** Suppose that  $A$  is a subset of  $\mathbb{Z}^+$ , the set of positive integers. Then  $A = \mathbb{Z}^+$  if

- (i)  $1 \in A$ , and
- (ii)  $\forall k \in \mathbb{Z}^+ ((1 \leq n \leq k \Rightarrow n \in A) \Rightarrow k + 1 \in A)$ .

### Exercises 8 (page 99)

**8.1** Given  $(x, y) \in \mathbb{R}^2$ , by the trichotomy law precisely one of  $x < y$ ,  $x = y$  and  $x > y$  is true. There is no ambiguity in the definition of  $g(x, y)$  in the first and the last cases. If  $x = y$  then  $x \leq y$  and  $x \geq y$  and so two definitions are given,  $y$  in the first case and  $x$  in the second; these values are the same since we are considering the case  $x = y$ . Hence  $g$  is well-defined.

If  $x \geq y$ , then  $x - y \geq 0$  so that  $|x - y| = x - y$ . Thus, in this case,  $f(x, y) = (x + y)/2 + (x - y)/2 = x = g(x, y)$ . On the other hand, if  $x \leq y$ , then  $x - y \leq 0$  so that  $|x - y| = -(x - y) = y - x$ . This implies that, in this case,  $f(x, y) = (x + y)/2 + (y - x)/2 = y = g(x, y)$ . Since, for each  $(x, y) \in \mathbb{R}^2$ ,  $x \geq y$  or  $x \leq y$  this proves that  $f = g$ .

[Strictly speaking the values of the function  $f$  should be written  $f((x, y))$  (and similarly for  $g$ ) since the function is evaluated at the point  $(x, y) \in \mathbb{R}^2$ . It is usual to write the values simply as  $f(x, y)$  and it is common to describe  $f$  as a function of two variables  $x$  and  $y$ .]

**8.2** The four composite functions are given by

- (i)  $f \circ f(x) = (x^3)^3 = x^9$ ,
- (ii)  $f \circ g(x) = (1 - x)^3$ ,
- (iii)  $g \circ f(x) = 1 - x^3$ ,
- (iv)  $g \circ g(x) = 1 - (1 - x) = x$  so that  $g \circ g = I_{\mathbb{R}}$ , the identity function.

For the set,  $fg(x) = gf(x) \Leftrightarrow (1 - x)^3 = 1 - x^3 \Leftrightarrow 1 - 3x + 3x^2 - x^3 = 1 - x^3 \Leftrightarrow x(x - 1) = 0 \Leftrightarrow x = 0$  or  $x = 1$ . Hence  $\{x \in \mathbb{R} \mid fg(x) = gf(x)\} = \{0, 1\}$ .

**8.3** There are lots of possibilities here. The following functions are as required:

- (i)  $f_1 = I_{\mathbb{R}}$ , the identity function;

$$(ii) f_2(x) = \begin{cases} x^2 & \text{if } x \neq 0, \\ 1 & \text{if } x = 0; \end{cases}$$

another example is given by  $f_2(x) = e^x$  (the exponential function);

$$(iii) f_3(x) = \begin{cases} x & \text{if } x \notin \mathbb{Z}, \\ x + 1/2 & \text{if } x \in \mathbb{Z}; \end{cases}$$

$$(iv) f_4(x) = n \in \mathbb{Z} \text{ if } n \leq x < n + 1.$$

**8.4** Suppose that we are given  $\varepsilon \in \mathbb{R}^+$ . Then

$$1/n < \varepsilon \Leftrightarrow n > 1/\varepsilon$$

and so if we choose an integer  $N > 1/\varepsilon$  we have  $n \geq N \Rightarrow n > 1/\varepsilon \Rightarrow |1/n| = 1/n < \varepsilon$  as required to show that  $1/n$  is null.

**8.5** (i) This is the graph of the function whose values are given by the following table.

| $x$ | $f(x)$ |
|-----|--------|
| $a$ | $z$    |
| $b$ | $y$    |
| $c$ | $z$    |
| $d$ | $x$    |

(ii) This is not the graph of a function since no element is specified for the value at  $c$ .

(iii) This is not the graph of a function since two elements are specified for the value at  $b$ .

(iv) This is the graph of the function whose values are given by the following table.

| $x$ | $f(x)$ |
|-----|--------|
| $a$ | $y$    |
| $b$ | $z$    |
| $c$ | $w$    |
| $d$ | $x$    |

### Exercises 9 (page 113)

**9.1** (i)  $y = f_1(x) \Leftrightarrow y = 2x + 5 \Leftrightarrow x = (y - 5)/2$ . So  $f_1$  is a bijection and the inverse is given by  $f_1^{-1}(x) = (x - 5)/2$ .

(ii)  $y = f_2(x) \Leftrightarrow y = x^2 + 2x + 1 = (x + 1)^2$ . So the function is not

injective (for example  $f_2(-2) = f_2(0) = 1$ ) and it is not surjective (since  $(x+1)^2 \geq 0$  for all real  $x$ ).

(iii)  $y = f_3(x) \Leftrightarrow y = x^2 - 2x = (x-1)^2 - 1$ . So again this is not injective (for example  $f_3(0) = f_3(2) = 0$ ) and is not surjective (since  $(x-1)^2 - 1 \geq -1$  for all real  $x$ ).

(iv)  $y = 1/x \Leftrightarrow x = 1/y$  and  $y = 0 \Leftrightarrow x = 0$  and so  $f_4$  is a bijection which is its own inverse.

Sketching the graphs is quite a help in doing these questions.

**9.2** (i) Notice that the inverse function is not defined for  $x \leq 5$  since then  $(x-5)/2 \notin \mathbb{R}^+$ . So  $f_1: \mathbb{R}^+ \rightarrow \mathbb{R}^+$  is an injection but not a surjection since, for example, 5 is not a value.

(ii)  $y = f_2(x) = (x+1)^2 \Leftrightarrow x = -1 \pm \sqrt{y}$ . Here  $x \in \mathbb{R}^+ \Leftrightarrow 1 < \pm\sqrt{y}$  so we need  $y > 1$  and must take the positive square root. Thus the function  $\mathbb{R}^+ \rightarrow \mathbb{R}^+$  is not a surjection (for example 1 is not a value) but is an injection.

(iii) This example does not restrict to a function  $\mathbb{R}^+ \rightarrow \mathbb{R}^+$  since, for example,  $f_3(1) = -1$ .

(iv) This is still a bijection which is its own inverse since  $1/x > 0 \Leftrightarrow x > 0$ .

**9.3** (i) Suppose that  $x, y \in \mathbb{R}$ . Then

$$y = 3x + 2 \Leftrightarrow 3x = y - 2 \Leftrightarrow x = (y - 2)/3.$$

Hence  $f_1^{-1}(x) = (x - 2)/3$ .

(ii) Suppose that  $x, y \in \mathbb{R}$ . Then

$$y = x^3 + 1 \Leftrightarrow x^3 = y - 1 \Leftrightarrow x = \sqrt[3]{y - 1}.$$

Hence  $f_2^{-1}(x) = \sqrt[3]{x - 1}$ .

**9.4** Suppose that  $gf(x_1) = gf(x_2)$  for elements  $x_1, x_2 \in X$ . Since  $g$  is an injection,  $f(x_1) = f(x_2)$ . Since  $f$  is an injection  $x_1 = x_2$  as required to prove that  $g \circ f$  is an injection.

**9.5** We can verify the conditions of Proposition 9.2.5 to prove that  $g \circ f$  and  $f^{-1} \circ g^{-1}$  are inverses of each other as follows.

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

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

Thus  $g \circ f$  is invertible (or equivalently a bijection) with inverse  $f^{-1} \circ g^{-1}$ .

[This result shows that composites of bijections are very easy to deal with. Notice the reversal of order when writing down the inverse.]

**9.6** Suppose that  $f$  is a surjection. Then, given  $y_0 \in Y$  there exists  $x_0 \in X$  such that  $f(x_0) = y_0$ . But then  $(x_0, y_0) \in X \times \{y_0\} \cap G_f$  and so  $X \times \{y_0\} \cap G_f \neq \emptyset$ .

Conversely, suppose that  $\forall y \in Y, (X \times \{y\}) \cap G_f \neq \emptyset$ . Then to show that  $f$  is a surjection we must show that, given  $y_0 \in Y$ , there exists an element  $x_0 \in X$  such that  $f(x_0) = y_0$ .  $(X \times \{y_0\}) \cap G_f \neq \emptyset$  so we can choose  $(x_1, y_1) \in X \times \{y_0\} \cap G_f$ . But then  $(x_1, y_1) \in X \times \{y_0\}$  so that  $x_1 \in X$  and  $y_1 = y_0$ . Also  $(x_1, y_1) \in G_f$  and so  $y_1 = f(x_1)$ . It follows that  $y_0 = f(x_1)$  and so we can take  $x_0 = x_1$ . Hence  $f$  is a surjection.

**9.7** (i) Suppose that  $B_1 \subseteq B_2$ . Then  $x \in \overleftarrow{f}(B_1) \Rightarrow f(x) \in B_1 \Rightarrow f(x) \in B_2 \Rightarrow x \in \overleftarrow{f}(B_2)$ . Hence  $\overleftarrow{f}(B_1) \subseteq \overleftarrow{f}(B_2)$ .

(ii) The result follows from  $x \in \overleftarrow{f}(B_1 \cap B_2) \Leftrightarrow f(x) \in B_1 \cap B_2 \Leftrightarrow f(x) \in B_1$  and  $f(x) \in B_2 \Leftrightarrow x \in \overleftarrow{f}(B_1)$  and  $x \in \overleftarrow{f}(B_2) \Leftrightarrow x \in \overleftarrow{f}(B_1) \cap \overleftarrow{f}(B_2)$ .

(iii) The result follows from  $x \in \overleftarrow{f}(B_1 \cup B_2) \Leftrightarrow f(x) \in B_1 \cup B_2 \Leftrightarrow f(x) \in B_1$  or  $f(x) \in B_2 \Leftrightarrow x \in \overleftarrow{f}(B_1)$  or  $x \in \overleftarrow{f}(B_2) \Leftrightarrow x \in \overleftarrow{f}(B_1) \cup \overleftarrow{f}(B_2)$ .

When we consider the converse of (i) we realize that we have no control over elements of  $B_1$  which are not values of the function. To give a very simple example consider the function  $f: \{a\} \rightarrow \{a, b\}$  defined by  $f(a) = a$ . Take  $B_1 = \{b\}$  and  $B_2 = \{a\}$ . Then  $\overleftarrow{f}(B_1) = \emptyset \subseteq \overleftarrow{f}(B_2) = \{a\}$  whereas clearly  $B_1 \not\subseteq B_2$ . Thus  $\overleftarrow{f}(B_1) \subseteq \overleftarrow{f}(B_2) \not\Rightarrow B_1 \subseteq B_2$ .

[The implication does hold if  $f$  is a surjection and the reader should consider how to write out the proof of this.]

### Exercises 10 (page 132)

**10.1** Since  $X$  is finite of cardinality  $n$  there exists a bijection  $f: \mathbb{N}_n \rightarrow X$ .

Suppose that  $Y$  is also finite of cardinality  $n$ . Then there is a bijection

$g: \mathbb{N}_n \rightarrow Y$ . Since a bijection is invertible we can now write down a bijection as required,  $g \circ f^{-1}: X \rightarrow \mathbb{N}_n \rightarrow Y$ .

Conversely, given a bijection  $h: X \rightarrow Y$  we can deduce that  $|Y|$  is finite of cardinality  $n$  by writing down the bijection  $h \circ f: \mathbb{N}_n \rightarrow X \rightarrow Y$ .

**10.2** The statement  $P(n)$  to be proved by induction is: ‘Given any collection  $X_1, X_2, \dots, X_n$  of pairwise disjoint finite sets, their union is finite and  $|X_1 \cup X_2 \cup \dots \cup X_n| = |X_1| + |X_2| + \dots + |X_n|$ .’

Base case: For  $n = 1$  this just says  $|X_1| = |X_1|$  which is clearly true.

Inductive step: Suppose that the result holds when  $n$  is some positive integer  $k$ . To deduce that the result holds for  $n = k+1$ , suppose that  $X_1, X_2, \dots, X_{k+1}$  is some collection of  $k+1$  finite sets. Then, by inductive hypothesis the union of the first  $k$  of these sets is finite and

$$|X_1 \cup X_2 \cup \dots \cup X_k| = |X_1| + |X_2| + \dots + |X_k|.$$

Hence, by the addition theorem,  $(X_1 \cup \dots \cup X_k) \cup X_{k+1}$  is a finite set and

$$\begin{aligned} |X_1 \cup X_2 \cup \dots \cup X_k \cup X_{k+1}| &= |X_1 \cup X_2 \cup \dots \cup X_k| + |X_{k+1}| \\ &= |X_1| + |X_2| + \dots + |X_k| + |X_{k+1}| \end{aligned}$$

so that the result holds for  $n = k + 1$ .

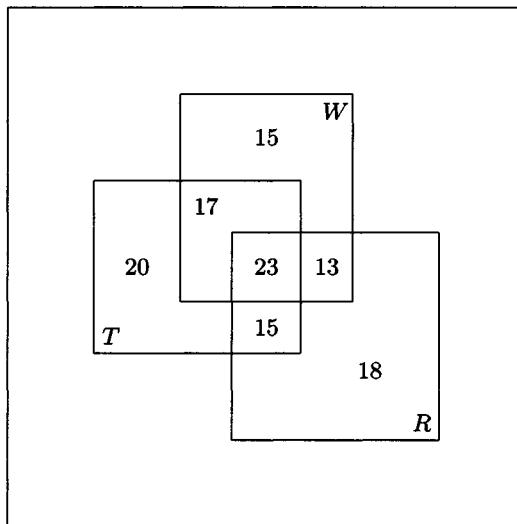
Hence by induction the result holds for all positive integers  $n$ .

**10.3** The cardinality of  $X \times Y$  is  $3 \times 2 = 6$ . An explicit bijection  $f: \mathbb{N}_6 \rightarrow X \times Y$  is given by the following table.

| $i$    | 1        | 2        | 3        | 4        | 5        | 6        |
|--------|----------|----------|----------|----------|----------|----------|
| $f(i)$ | $(a, d)$ | $(a, e)$ | $(b, d)$ | $(b, e)$ | $(c, d)$ | $(c, e)$ |

**10.4** This sort of question is probably best answered using a Venn diagram. Let  $T$  denote the set of triangular tiles,  $R$  denote the set of red tiles and  $W$  denote the set of wooden tiles. Then we have the Venn diagram on the next page.

The total number in the whole square is 144 and we wish to know the number in the outer region representing  $(W \cup R \cup T)^c$ . There are eight disjoint regions and we can calculate the numbers in the other seven regions from the given data as follows. We are given that  $|W \cap R \cap T| = 23$ . Since we are given that  $|W \cap R| = 36$  we see, from the addition theorem, that  $|(W \cap R) - (W \cap R \cap T)| = 13$ . Continuing in this way we can fill in the numbers as shown. Now we can calculate  $|W \cup R \cup T| = 121$ . So



again using the addition theorem we get  $|W \cup R \cup T|^c = 144 - 121 = 23$ . Hence there are 23 blue plastic square tiles.

Alternatively we can use the inclusion-exclusion principle (Proposition 10.3.2) without drawing a diagram. This gives

$$\begin{aligned}
 & |W \cup R \cup T| \\
 &= |W| + |R| + |T| - |W \cap R| - |W \cap T| - |R \cap T| + |W \cap R \cap T| \\
 &= 68 + 69 + 75 - 36 - 40 - 38 + 23 \\
 &= 121
 \end{aligned}$$

which gives the same result as before.

### Exercises 11 (page 143)

**11.1** The proof is by induction on  $n$ .

Base case: For  $n = 1$ , if  $f: \mathbb{N}_1 \rightarrow X$  is a surjection then we must have  $X = \{f(1)\}$  and so  $|X| = 1$ .

Inductive step: Suppose that for some  $k \geq 1$  the result holds for  $n = k$ . To deduce the result for  $n = k + 1$ , suppose that  $f: \mathbb{N}_{k+1} \rightarrow X$  is a surjection. There are two cases.

- (i) Suppose that  $f|_{\mathbb{N}_k}: \mathbb{N}_k \rightarrow X$  is a surjection. Then, by inductive hypothesis,  $X$  is finite and  $|X| \leq k$  so that certainly  $|X| \leq k + 1$ , as required for the inductive step.

(ii) Otherwise  $f(k+1) \notin \vec{f}(\mathbb{N}_k)$  and so we may define a function  $f_1: \mathbb{N}_k \rightarrow X - \{f(k+1)\}$  by  $f_1(i) = f(i)$ . This too is a surjection and so, by inductive hypothesis,  $|X - \{f(k+1)\}| \leq k$ . Hence, by the addition theorem,  $|X| \leq k+1$ , as required for the inductive step.

This completes the proof of the inductive step.

Hence, by induction, the result is true for all positive integers  $n$ .

**11.2** Suppose that  $c_1$  and  $c_2$  are maximum elements of  $A$ , a set of real numbers. Then, since  $c_1$  is a maximum and  $c_2 \in A$ ,  $c_1 \geq c_2$ , and since  $c_2$  is a maximum and  $c_1 \in A$ ,  $c_2 \geq c_1$ . Hence  $c_1 \leq c_2$  and  $c_1 \geq c_1$  so that  $c_1 = c_2$  as required.

**11.3** Checking the possibilities gives:

$$\begin{aligned} D(88) &= \{\pm 1, \pm 2, \pm 4, \pm 8, \pm 11, \pm 22, \pm 44, \pm 88\}; \\ D(136) &= \{\pm 1, \pm 2, \pm 4, \pm 8, \pm 17, \pm 34, \pm 68, \pm 136\}. \end{aligned}$$

Hence  $D(88) \cap D(136) = \{\pm 1, \pm 2, \pm 4, \pm 8\}$  so that the greatest common divisor  $(88, 136) = 8$ .

**11.4** Suppose that  $c$  is a common divisor of  $a/d$  and  $b/d$ . Then, since  $c$  divides  $a/d$ ,  $a/d = cq$  for some  $q \in \mathbb{Z}$  so that  $a = cdq$  and thus  $cd$  divides  $a$ . Similarly, since  $c$  divides  $b/d$  then  $cd$  divides  $b$  [fill in the details]. Hence  $cd$  is a common divisor of  $a$  and  $b$  and so  $cd \leq d$  by the definition of  $d$ . But if  $cd \leq d$  then  $c \leq 1$  (since  $d$  is positive). Thus every common factor of  $a/d$  and  $b/d$  is less than or equal to 1 and so  $(a/d, b/d) = 1$  as required.

**11.5** By Corollary 11.1.5,  $|X| \leq |Y|$ . Thus the statements we are asked to prove are contrapositives. We prove the statements  $\Rightarrow$  and the implications in the other directions follow since they are the contrapositives.

(i) Clearly, if  $X = Y$  then  $|X| = |Y|$ .

(ii) Suppose now that  $X \subset Y$  (in other words  $X$  is a *proper* subset not containing every element of  $Y$ ). Let  $y_0 \in Y$  be an element such that  $y_0 \notin X$ . Then  $X - \{y_0\}$  is a finite set (Corollary 11.1.5 again) and by the addition theorem  $|X - \{y_0\}| = |X| - 1$ . Hence by Corollary 11.1.5  $|Y| \leq |X| - 1$  and so  $|Y| < |X|$ .

**11.6** Let  $A$  be a set of positive integers without a least element. We prove that  $\mathbb{N}_n \subseteq \mathbb{Z}^+ - A$  for all  $n$  by induction.

Base case: If  $1 \in A$  it would be the least element, which is impossible. Hence  $1 \in \mathbb{Z}^+ - A$ .

Inductive step: Suppose as inductive hypothesis that  $\mathbb{N}_k \subseteq \mathbb{Z}^+ - A$  for some positive integer  $k$ . Then if  $k + 1 \in A$  it would be the least element by inductive hypothesis, which is impossible. Hence  $k + 1 \in \mathbb{Z}^+ - A$  and so, using the inductive hypothesis again,  $\mathbb{N}_{k+1} \subseteq \mathbb{Z}^+ - A$ , completing the inductive step.

Hence  $\mathbb{N}_n \subseteq \mathbb{Z}^+ - A$  for all  $n$  so that in particular  $n \notin A$  for all  $n$ , which means that  $A$  is the empty set.

This proves that if a set of positive integers does not have a least element then it is empty. The contrapositive of this statement is the well-ordering principle, which is therefore also proved.

### Exercises 12 (page 155)

**12.1** Continuing Pascal's triangle from the last line on page 151 we obtain the following.

|   |    |    |     |     |     |     |     |    |
|---|----|----|-----|-----|-----|-----|-----|----|
|   | 1  | 7  | 21  | 35  | 35  | 21  | 7   | 1  |
| 1 | 8  | 28 | 56  | 70  | 56  | 28  | 8   | 1  |
| 1 | 9  | 36 | 84  | 126 | 126 | 84  | 36  | 9  |
| 1 | 10 | 45 | 120 | 210 | 252 | 210 | 120 | 45 |

From this we can read off that  $\binom{10}{3} = 120$  and by the binomial theorem this is the required coefficient.

**12.2** By the binomial theorem the required coefficient is  $\binom{100}{2}$  and the formula evaluates this as  $100!/2!98! = (100 \times 99)/(2 \times 1) = 4950$ .

**12.3** (i) Let  $X$  be the set of three people and  $Y$  the pack of 52 cards. Then each group selection corresponds to an injection  $X \rightarrow Y$ . By Proposition 12.1.4 the number of possibilities is  $(52)_3 = 52 \times 51 \times 50 = 132600$ .

(ii) Each set of three cards is an element of  $\mathcal{P}_3(Y)$ . By Definition 12.2.4 the number of possibilities is  $\binom{52}{3} = 52 \times 51 \times 50/3! = 22100$ .

**12.4** (i) Let  $X$  be the set of three people and  $Y$  the set of five dishes. Then each group selection corresponds to a function  $X \rightarrow Y$  (since two or more may select the same dish). By Proposition 12.1.2 the number of possibilities is  $5^3 = 125$ .

(ii) Calculating the number of possible orders for the chef is complicated by the fact that two or more people may select the same dish so that we are not simply counting subsets of  $Y$ . The possibilities are as follows. If the three dishes are distinct then the possibilities are the elements of  $\mathcal{P}_3(Y)$  and so the number is  $\binom{5}{3} = 5!/3!2! = 10$ . If only two distinct dishes are selected then one must be selected once and the other twice so that the possibilities correspond to injections  $\mathbb{N}_2 \rightarrow Y$  and the number is  $(5)_2 = 5 \times 4 = 20$ . If all three people select the same dish then the number of possibilities is 5. So the total number of possibilities is  $10 + 20 + 5 = 35$ .

[This approach to the problem is rather cumbersome and would be difficult to apply if the numbers were a little larger – think about applying it to find the number of possible orders by five people selecting from a menu of eight items. A less direct approach gives a simpler calculation. Let  $Y = \{y_1, y_2, y_3, y_4, y_5\}$ . Then each order corresponds to a triple  $(y_{i_1}, y_{i_2}, y_{i_3})$  where  $i_1 \leq i_2 \leq i_3$ . A bijection from this set of triples to  $\mathcal{P}_3(\mathbb{N}_7)$  is given by  $(y_{i_1}, y_{i_2}, y_{i_3}) \mapsto \{i_1, i_2 + 1, i_3 + 2\}$ . Hence the number of orders is  $\binom{7}{3} = 7!/3!4! = 35$ . This argument can be used to show that the number of possible selections of  $k$  objects (repetitions permitted) from  $n$  objects is given by  $\binom{n+k-1}{k}$ . So, for example, if five people select from a menu of eight items the number is  $\binom{12}{5} = 12!/5!7! = 792$ .]

**12.5** The coefficient of  $x^n$  in  $(1+x)^{2n}$  is  $\binom{2n}{n}$ .

In the product  $(1+x)^n(1+x)^n$  the terms contributing to  $x^{2n}$  have the form  $\binom{n}{i}x^i \times \binom{n}{n-i}x^{n-i}$ . Hence

$$\binom{2n}{n} = \sum_{i=0}^n \binom{n}{i} \binom{n}{n-i} = \sum_{i=1}^n \binom{n}{i}^2$$

as required.

**12.6** Suppose that the  $n$  consecutive integers are  $k+1, k+2, \dots, k+n$ . The binomial coefficient  $\binom{k+n}{n}$  is given by  $(k+n)_n/n!$ . Since this is an integer by definition, it follows that  $n!$  divides

$$(k+n)_n = (k+n)(k+n-1)\dots(k+1),$$

the product of the given set of integers.

**12.7** By the binomial theorem,

$$\begin{aligned}(1+i)^n &= \sum_{j=0}^n \binom{n}{j} i^j \\ &= \sum_{0 \leq 2r \leq n} \binom{n}{2r} (-1)^r + \sum_{0 \leq 2r+1 \leq n} \binom{n}{2r+1} (-1)^r i,\end{aligned}$$

separating out the real and imaginary parts and using  $i^2 = -1$ .

However,  $1+i = \sqrt{2}e^{i\pi/4}$  and so  $(1+i)^n = 2^{n/2} \exp(in\pi/4)$ . The given result follows on equating the real parts of these two expressions for  $(1+i)^n$  using the fact that

$$2^{n/2} \exp(in\pi/4) = \begin{cases} 2^{2k} \exp(ik\pi), & \text{if } n = 4k, \\ & = 4^k (-1)^k = (-4)^k, \\ 2^{(4k+1)/2} \exp(i(4k+1)\pi/4), & \text{if } n = 4k+1, \\ & = 2^{(4k+1)/2} (-1)^k 2^{-1/2} (1+i) \\ & = (-4)^k (1+i), \\ 2^{2k+1} \exp(i(2k+1)\pi/2), & \text{if } n = 4k+2, \\ & = 2^{2k+1} (-1)^k i, \\ 2^{(4k+3)/2} \exp(i(4k+3)\pi/4), & \text{if } n = 4k+3, \\ & = 2^{(4k+3)/2} (-1)^k 2^{-1/2} (-1+i) \\ & = 2^{-1} (-4)^{k+1} (1-i). \end{cases}$$

Similarly, by equating the imaginary parts we obtain

$$\sum_{0 \leq 2r+1 \leq n} (-1)^r \binom{n}{2r+1} = \begin{cases} 0 & \text{if } n = 4k, \\ (-4)^k & \text{if } n = 4k+1, \\ 2(-4)^k & \text{if } n = 4k+2, \\ 2(-4)^k & \text{if } n = 4k+3. \end{cases}$$

### Exercises 13 (page 169)

**13.1** Suppose for contradiction that there is rational number  $q$  such that  $q^2 = 3$ . Write  $q$  as a fraction in lowest terms:  $q = a/b$  so that  $a$  and  $b$  are integers such that  $(a, b) = 1$ . Now  $q^2 = 3 \Rightarrow a^2/b^2 = 3 \Rightarrow a^2 = 3b^2 \Rightarrow a^2$  is divisible by 3. It follows that  $a$  must be divisible by 3.

Thus we can write  $a = 3a_1$  for some  $a_1 \in \mathbb{Z}$ . But then  $a^2 = 3b^2 \Rightarrow 9a_1^2 = 3b^2 \Rightarrow 3a_1^2 = b^2 \Rightarrow b^2$  is divisible by 3  $\Rightarrow b$  is divisible by 3, as above. Hence 3 is a common factor of  $a$  and  $b$  so that  $(a, b) \neq 1$ , contradicting the choice of  $a$  and  $b$  and giving the required contradiction.

Hence there does not exist a rational number whose square is 3.

**13.2** Calculation gives

$$\begin{aligned} 1^2 = 1 &\leqslant 3 \leqslant 2^2 = 4, \\ (1.7)^2 = 2.89 &\leqslant 3 \leqslant (1.8)^2 = 3.24, \\ (1.73)^2 = 2.9929 &\leqslant 3 \leqslant (1.74)^2 = 3.0276. \end{aligned}$$

Hence  $\sqrt{3} = 1.73\dots$

**13.3** The same argument for  $\sqrt{4}$  leads to  $q^2 = 4 \Rightarrow a^2/b^2 = 4 \Rightarrow a^2 = 4b^2 \Rightarrow a^2$  is divisible by 4. However, it does not now follow that  $a$  is divisible by 4, e.g. we can have  $a = 2$ . All we can prove is that 4 divides  $a^2 \Rightarrow 2$  divides  $a$  but this does not lead to a contradiction since  $a = 2a_1 \Rightarrow 4a_1^2 = 4b^2 \Rightarrow a_1^2 = b^2$ .

Of course there *does* exist a rational number whose square is 4, namely  $2=2/1$ .

**13.4** From the definition,  $1/3 = 0.\overline{3}$

$$\begin{aligned} &\Leftrightarrow \underbrace{0.33\dots33}_n \leqslant 1/3 \leqslant \underbrace{0.33\dots34}_n \text{ for each } n \in \mathbb{Z}^+ \\ &\Leftrightarrow \underbrace{33\dots33}_n \leqslant (10)^n/3 \leqslant \underbrace{33\dots34}_n \text{ for each } n \in \mathbb{Z}^+ \\ &\Leftrightarrow \underbrace{99\dots99}_n \leqslant (10)^n \leqslant \underbrace{100\dots02}_n \\ &\Leftrightarrow (10)^n - 1 \leqslant (10)^n \leqslant (10)^n + 2 \end{aligned}$$

which is certainly true.

**13.5** Let  $x = 1.778\dot{2}34\dot{5}$ . Then  $10000x = 17782.345\dot{2}34\dot{5}$  from which it follows that  $9999x = 17782.345 - 1.778 = 17780.567 = 17780567/1000$ . It follows that  $x = 17780567/9999000$ .

### Exercises 14 (page 181)

**14.1** The result is trivial if  $A$  is empty since then  $A \cup B = B$ .

Suppose otherwise that  $|X| = n \in \mathbb{Z}^+$ . Then there is a bijection  $f: \mathbb{N}_n \rightarrow A$ . Since  $B$  is denumerable there is a bijection  $g: \mathbb{Z}^+ \rightarrow B$ . We can now count  $A \cup B$  by first counting the finite set  $A$  and then counting  $B$  (we cannot do it the other way round or we would never get to  $A$ ). This gives a function  $h: \mathbb{Z}^+ \rightarrow A \cup B$  defined by

$$h(i) = \begin{cases} f(i) & \text{if } 1 \leqslant i \leqslant n, \\ g(i-n) & \text{if } i \geqslant n+1. \end{cases}$$

This function is a surjection since  $f$  and  $g$  are surjections. For it to be an injection we need  $A \cap B = \emptyset$  for otherwise any common elements will be counted twice. But so long as  $A \cap B = \emptyset$  then it is an injection since  $f$  and  $g$  are. Thus  $h$  is a bijection as required to prove  $A \cup B$  denumerable.

If  $A$  and  $B$  are not disjoint, replace  $A$  in the above argument by the finite set  $A - B$ .

[This argument should be compared with Theorem 10.2.1 and Proposition 10.3.1.]

**14.2** In the same way as in the previous result we may assume that  $A$  and  $B$  are disjoint by replacing  $A$  by  $A - B$ . By Proposition 14.2.5,  $A - B$  is either finite (in which case we can quote the previous result) or denumerable.

To count two disjoint denumerable sets we cannot count one and then the other since we will never get to the elements in the second. We overcome this by alternating between the sets as we count. Let  $f: \mathbb{Z}^+ \rightarrow X$  and  $g: \mathbb{Z}^+ \rightarrow Y$  be bijections. Then we can define a bijection  $h: \mathbb{Z}^+ \rightarrow X \cup Y$  by

$$h(i) = \begin{cases} f(k) & \text{if } i = 2k - 1, \\ g(k) & \text{if } i = 2k. \end{cases}$$

Suppose that  $X$  is uncountable and  $A \subseteq X$  is denumerable. Suppose for contradiction that  $X - A$  is countable, i.e. finite or denumerable. Then, by what has just been proved,  $X = A \cup (X - A)$  is denumerable, contradicting the uncountability of  $X$ . Hence  $X - A$  is uncountable as required.

**14.3** Using induction we can deduce from the previous result that any finite union of denumerable sets is denumerable. However, this method will not deal with a denumerable set of sets. To see how to do this, use the fact that  $A_n$  is denumerable to list the elements:

$$A_n = \{a_{1n}, a_{2n}, \dots, a_{mn}, \dots\}.$$

We can now display all the elements in the union of the sets in a double array as shown at the top of the next page. This looks very similar to the array in the proof of Proposition 14.2.3 and so we can use the same

diagonal method of counting as indicated.

|          |          |          |          |         |  |
|----------|----------|----------|----------|---------|--|
| $\vdots$ | $\vdots$ |          |          |         |  |
| $a_{14}$ | $\vdots$ |          |          |         |  |
| $a_{13}$ | $a_{23}$ | $\dots$  |          |         |  |
| $a_{12}$ | $a_{22}$ | $a_{32}$ | $\dots$  |         |  |
| $a_{11}$ | $a_{21}$ | $a_{31}$ | $a_{41}$ | $\dots$ |  |
| $\vdots$ | $\vdots$ |          |          |         |  |
| 10       | $\vdots$ |          |          |         |  |
| 6        | 9        | $\dots$  |          |         |  |
| 3        | 5        | 8        | $\dots$  |         |  |
| 1        | 2        | 4        | 7        | $\dots$ |  |

This gives the bijection  $\bigcup A_n \rightarrow \mathbb{Z}^+$ :

$$a_{mn} \mapsto \frac{1}{2}(m+n-2)(m+n-1) + n.$$

[The condition that the sets are disjoint is not necessary for this result. The reader might like to consider how to deduce the general case from the special case of disjoint sets.]

**14.4** Suppose that  $X$  and  $Y$  are two sets such that  $|X| \leq |Y|$  and  $|Y| \leq |X|$ . We prove that  $|X| = |Y|$  by contraction. Suppose, for contradiction, that  $|X| \neq |Y|$ . Then, since  $|X| \leq |Y|$ , we have  $|X| < |Y|$  and, since  $|Y| \leq |X|$ , we have  $|Y| < |X|$ . But, by definition, if  $|X| < |Y|$  there exists an injection  $X \rightarrow Y$  and, by the Cantor–Schröder–Bernstein theorem, if  $|Y| < |X|$ , there does not exist an injection  $X \rightarrow Y$ . Thus we have obtained a contradiction so that our assumption that  $|X| \neq |Y|$  is false. Hence  $|X| = |Y|$  as required.

[The statement proved in this exercise is the more usual statement of the Cantor–Schröder–Bernstein theorem.]

### Exercises 15 (page 197)

- 15.1** (i)  $100 = 3 \times 33 + 1 \Rightarrow q = 33, r = 1$ .
- (ii)  $3 = 100 \times 0 + 3$ .
- (iii)  $100 = 7 \times 14 + 2$ .

- (iv)  $-100 = 7 \times (-15) + 5.$
- (v)  $-105 = 7 \times (-15) + 0.$
- (vi)  $-3 = 105 \times (-1) + 102.$
- (vii)  $7684 = 4148 \times 1 + 3536.$
- (viii)  $-7684 = 4148 \times (-2) + 612.$
- (ix)  $1234567 = 1357 \times 909 + 1054.$
- (x)  $0 = 17 \times 0 + 0.$

**15.2** First of all notice that, if  $a$  is divisible by 5, then  $a = 5q$  and so  $a^2 = 25q^2 = 5(5q^2)$  is divisible by 5.

For the converse suppose that  $a^2$  is divisible by 5 and, for contradiction, that  $a$  is not divisible by 5. Then, by the division theorem,  $a = 5q + 1$ ,  $a = 5q + 2$ ,  $a = 5q + 3$  or  $a = 5q + 4$ . However,

$$\begin{aligned} a = 5q + 1 &\Rightarrow a^2 = (5q + 1)^2 = 25q^2 + 10q + 1 = 5(5q^2 + 2q) + 1, \\ a = 5q + 2 &\Rightarrow a^2 = 5(5q^2 + 4q) + 4, \\ a = 5q + 3 &\Rightarrow a^2 = 5(5q + 6q + 1) + 4, \\ a = 5q + 4 &\Rightarrow a^2 = 5(5q + 8q + 3) = 1. \end{aligned}$$

Hence in each case the remainder is non-zero and so, by the division theorem,  $a^2$  is not divisible by 5, giving the required contradiction. Hence, if  $a^2$  is divisible by 5 then so is  $a$  as required.

**15.3** First of all notice that if  $a^2$  is divisible by 9 then it is certainly divisible by 3. For the converse, if  $a^2$  is divisible by 3 then, by Proposition 15.2.1,  $a$  is divisible by 3, i.e.  $a = 3q$  and so  $a^2 = 9q^2$  is divisible by 9.

**15.4**  $98765432 = 3 \times 32921810 + 2$  has remainder 2 after division by 3 and so by Proposition 15.2.3 it is not a perfect square.

**15.5** Suppose that  $n$  is a perfect square, say  $n = a^2$  for some integer  $a$ . Then  $a$  is either even or odd, i.e.  $a = 2q_0$  or  $a = 2q_0 + 1$  for some integer  $q$ . Hence  $a^2 = 4q_0^2$  or  $a^2 = 4q_0^2 + 4q_0 + 1 = 4(q_0^2 + q_0) + 1$  so that  $n = 4q$  (where  $q = q_0^2$ ) or  $n = 4q + 1$  (where  $q = q_0^2 + q_0$ ) as required.

However,  $1234567 = 4 \times 308641 + 3$  has remainder 3 after division by 4 and so is not a perfect square.

**15.6** To find the first decimal place  $a_1$  observe that  $50 = 7 \times 7 + 1$  so that  $7/10 < 5/7 < 8/10$  and  $a_1 = 7$ . Continuing in this way we obtain the following.

$$\begin{aligned} 50 &= 7 \times 7 + 1 \Rightarrow a_1 = 7 \\ 10 &= 7 \times 1 + 3 \Rightarrow a_2 = 1 \\ 30 &= 7 \times 4 + 2 \Rightarrow a_3 = 4 \\ 20 &= 7 \times 2 + 6 \Rightarrow a_4 = 2 \\ 60 &= 7 \times 8 + 4 \Rightarrow a_5 = 8 \\ 40 &= 7 \times 5 + 5 \Rightarrow a_6 = 5 \\ 50 &= 7 \times 7 + 1 \end{aligned}$$

At this point the calculation goes back to the beginning and so the decimal places begin to recur giving  $5/7 = 0.\dot{7}1428\dot{5}$ .

### Exercises 16 (page 206)

**16.1** Applying the Euclidean algorithm gives the following sequence of remainders.

$$\begin{array}{r} 7684 \\ 4148 \quad (-1) \\ 3536 \quad (-1) \\ 612 \quad (-5) \\ 476 \quad (-1) \\ 136 \quad (-3) \\ 68 \quad (-2) \\ 0 \end{array}$$

Thus  $(7684, 4148) = 68$ .

**16.2** Applying the Euclidean algorithm gives the following sequence of remainders.

$$\begin{array}{r} 11033442 \\ 1102246 \quad (-10) \\ 10982 \quad (-100) \\ 4046 \quad (-2) \\ 2890 \quad (-1) \\ 1156 \quad (-2) \\ 578 \quad (-2) \\ 0 \end{array}$$

Thus  $(11033442, 1102246) = 578$ .

**16.3** In the examples we have done we have always put  $a > b$ , writing down the larger number first. If we change the order of the numbers so that  $a_0 = b < a_1 = a$  then the first step of the Euclidean algorithm will read

$$a_0 = a_1 \times 0 + a_0$$

and so (using the notation of Theorem 16.2.1)  $a_2 = r_1 = a_0$ . Since we now have  $a_1 = a$ ,  $a_2 = b$ , the remaining steps of the algorithm are identical to the steps which arise from writing down the larger number first.

So the order of the numbers does not really matter. If the smaller number is written down first then the first step simply switches the order so that there is this one extra step.

**16.4** Suppose that  $c \in D(b, r)$ . Then, since  $c$  divides  $b$ ,  $b = cq_1$  for some integer  $q_1$ , and, since  $c$  divides  $r$ ,  $r = cq_2$ . It follows that  $a = bq + r = c(q_1q + q_2)$  so that  $c$  divides  $a$  and  $c \in D(a, b)$ . Thus  $D(b, r) \subseteq D(a, b)$ .

### Exercises 17 (see page 215)

**17.1** Extending the table of Solution 16.1 to keep track of the linear combinations we obtain the following.

$$\begin{array}{rcl} 7684 & = & 7684 \times 1 + 4148 \times 0 \\ 4148 & = & 7684 \times 0 + 4148 \times 1 & (-1) \\ 3536 & = & 7684 \times 1 + 4148 \times (-1) & (-1) \\ 612 & = & 7684 \times (-1) + 4148 \times 2 & (-5) \\ 476 & = & 7684 \times 6 + 4148 \times (-11) & (-1) \\ 136 & = & 7684 \times (-7) + 4148 \times 13 & (-3) \\ 68 & = & 7684 \times 27 + 4148 \times (-50) & (-2) \\ 0 & & & \end{array}$$

From this we see that  $68 = 7684 \times 27 + 4148 \times (-50)$  and checking this does give  $207468 - 207400 = 68$ .

Hence one possible solution is  $m = 27$ ,  $n = -50$ .

**17.2** From the previous solution one possible solution is  $m = 27$ ,  $n = 50$ .

**17.3** Extending the table of Solution 16.2 we obtain the following.

$$\begin{array}{rcl}
 11033442 & = & 11033442 \times 1 & + 1102246 \times 0 \\
 1102246 & = & 11033442 \times 0 & + 1102246 \times 1 & (-10) \\
 10982 & = & 11033442 \times 1 & + 1102246 \times (-10) & (-100) \\
 4046 & = & 11033442 \times (-100) & + 1102246 \times 1001 & (-2) \\
 2890 & = & 11033442 \times 201 & + 1102246 \times (-2012) & (-1) \\
 1156 & = & 11033442 \times (-301) & + 1102246 \times 3013 & (-2) \\
 578 & = & 11033442 \times 803 & + 1102246 \times (-8038) & (-2) \\
 0 & & & &
 \end{array}$$

Hence  $11033442 \times 803 + 1102246 \times (-8038) = 578$  and so one possible solution is  $m = 803$ ,  $n = -8038$ .

[It is not so easy to check this solution on a calculator since rather large numbers arise!]

**17.4** As in the proof of Lemma 16.1.2 we prove this result by proving that the set of common divisors of  $a$ ,  $b$  and  $c$  is equal to the set of common divisors of  $\gcd(a, b)$  and  $c$ . We can prove this as follows:

$$\begin{aligned}
 d &\text{ is a common divisor of } a, b \text{ and } c \\
 \Leftrightarrow d &\text{ divides } a, b \text{ and } c \\
 \Leftrightarrow d &\text{ divides } \gcd(a, b) \text{ and } c \text{ (using Corollary 17.2.1).}
 \end{aligned}$$

Now, since the two sets of common divisors are equal they must have the same greatest element as required.

From this and Exercise 16.2 we obtain  $\gcd(11033442, 1102246, 6035) = \gcd(578, 6035)$ . The Euclidean algorithm applied to these two integers gives the following sequence of remainders.

$$\begin{array}{rcl}
 6035 & & \\
 578 & (-10) & \\
 255 & (-2) & \\
 68 & (-3) & \\
 51 & (-1) & \\
 17 & (-3) & \\
 0 & &
 \end{array}$$

Hence  $\gcd(11033442, 1102246, 6035) = \gcd(578, 6035) = 17$ .

**17.5** Extending the table in the previous solution to keep track of the linear combinations we obtain the following.

$$\begin{array}{rcl}
 6035 & = & 578 \times 0 + 6035 \times 1 \\
 578 & = & 578 \times 1 + 6035 \times 0 & (-10) \\
 255 & = & 578 \times (-10) + 6035 \times 1 & (-2) \\
 68 & = & 578 \times 21 + 6035 \times (-2) & (-3) \\
 51 & = & 578 \times (-73) + 6035 \times 7 & (-1) \\
 17 & = & 578 \times 94 + 6035 \times (-9) & (-3) \\
 0 & & 
 \end{array}$$

Hence

$$\begin{aligned}
 17 &= 578 \times 94 + 6035 \times (-9) \\
 &= (11033442 \times 803 + 1102246 \times (-8038)) \times 94 + 6035 \times (-9) \\
 &\quad (\text{using the result of Exercise 17.3}) \\
 &= 11033442 \times 75482 + 1102246 \times (-755572) + 6035 \times (-9).
 \end{aligned}$$

Hence one solution is  $m = 75482$ ,  $n = -755572$ ,  $p = -9$ .

**17.6** If  $a_1 b_2 = a_2 b_1$  then  $a_2$  divides  $a_1 b_2$ . But then, if  $a_2$  and  $b_2$  are coprime,  $a_2$  divides  $a_1$  and so in particular  $a_2 \leq a_1$ . But by symmetry the same argument can be used to prove that  $a_1 \leq a_2$  so that  $a_1 = a_2$ . It follows from  $a_1 b_2 = a_2 b_1$  that  $b_1 = b_2$ .

Notice that this result proves that the fraction in lowest terms representing a rational number (see Definition 13.1.3) is unique since  $a_1 b_2 = a_2 b_1 \Leftrightarrow a_1/b_1 = a_2/b_2$ .

### Exercises 18 (page 224)

**18.1** From Exercise 16.1,  $(7684, 4148) = 68$ . Now since  $272 = 68 \times 4$  we see that 68 divides 272 so that the diophantine equations does have solutions.

From Exercise 17.1,  $7684 \times 27 + 4148 \times (-50) = 68$ . Multiplying through by 4 we see that one solution to the equation is given by  $m = 27 \times 4 = 108$ ,  $n = (-50) \times 4 = -200$ .

Thus, for  $m, n \in \mathbb{Z}$ ,

$$\begin{aligned}
 7684m + 4148n &= 272 \\
 \Leftrightarrow 7684m + 4148n &= 7684 \times 108 + 4148 \times (-200) \\
 \Leftrightarrow 7684(m - 108) + 4148(n + 200) &= 0
 \end{aligned}$$

$$\begin{aligned}
 &\Leftrightarrow 113(m - 108) + 61(n + 200) = 0 \\
 &\quad \text{dividing through by } \gcd(7684, 4148) = 68 \\
 &\Leftrightarrow (m - 108, n + 200) = (61q, -113q) \text{ for some } q \in \mathbb{Z} \\
 &\quad \text{since 61 and 113 are coprime} \\
 &\Leftrightarrow (m, n) = (108 + 61q, -200 - 113q) \text{ for some } q \in \mathbb{Z}.
 \end{aligned}$$

Thus the solutions to the diophantine equation are given by

$$(m, n) = (108 + 61q, -200 - 113q) \text{ for } q \in \mathbb{Z}.$$

**18.2** Notice that  $7684m - 4148n = 272 \Leftrightarrow 7684m + 4148(-n) = 272$ . So the solutions are given by

$$(m, n) = (108 + 61q, 200 + 113q) \text{ for } q \in \mathbb{Z}.$$

**18.3** Applying the Euclidean algorithm to the integers 516 and 564 gives the following sequence of remainders expressed as integer linear combinations of 516 and 564.

$$\begin{array}{rcl}
 564 & = & 516 \times 0 + 564 \times 1 \\
 516 & = & 516 \times 1 + 564 \times 0 \quad (-1) \\
 48 & = & 516 \times (-1) + 564 \times 1 \quad (-10) \\
 36 & = & 516 \times 11 + 564 \times (-10) \quad (-1) \\
 12 & = & 516 \times (-12) + 564 \times 11 \quad (-3) \\
 0 & &
 \end{array}$$

Since 12 divides 6432 there are solutions. Since  $6432/12 = 536$  a solution is given by  $m = -12 \times 536 = -6432$  and  $n = 11 \times 536 = 5896$ . Thus, to find the general solution,

$$\begin{aligned}
 &516m + 564n = 6432 \\
 &\Leftrightarrow 516m + 564n = 516 \times (-6432) + 564 \times 5896 \\
 &\Leftrightarrow 516(m + 6432) + 564(n - 5896) = 0 \\
 &\Leftrightarrow 43(m + 6432) + 47(n - 5896) = 0 \\
 &\quad \text{dividing through by } \gcd(516, 564) = 12 \\
 &\Leftrightarrow (m + 6432, n - 5896) = (47q, -43q) \text{ for some } q \in \mathbb{Z} \\
 &\quad \text{since 43 and 47 are coprime} \\
 &\Leftrightarrow (m, n) = (-6432 + 47q, 5896 - 43q) \text{ for some } q \in \mathbb{Z}.
 \end{aligned}$$

To find positive solutions notice that  $m > 0 \Leftrightarrow -6432 + 47q > 0 \Leftrightarrow 47q > 6432 \Leftrightarrow q > 6432/47 = 136 + 40/47 \Leftrightarrow q \geq 137$ . Similarly,

$n > 0 \Leftrightarrow 5896 - 43q > 0 \Leftrightarrow 43q < 5896 \Leftrightarrow q < 5896/43 = 137 + 5/43 \Leftrightarrow q \leq 137$ . Thus the only solution with  $m$  and  $n$  both positive arises from  $q = 137$ . This gives the unique positive solution  $(m, n) = (7, 5)$ .

[If we are only seeking positive solutions to this particular problem then it is quicker to try the possibilities since necessarily  $n < 6432/564$ , i.e.  $n \leq 11$ . We simply try each positive integer less than 12 in turn to see whether  $6432 - 564n$  is divisible by 516.]

### Exercises 19 (page 239)

**19.1** From 1998 to 2045 there are 12 leap years. Hence the day number of 6 September 2045 is  $366 \times 12 + 365 \times 35 + 31 \times 5 + 30 \times 2 + 28 \times 1 + 6$ . Working modulo 7 in order to obtain the remainder modulo 7 gives  $2 \times 5 + 1 \times 0 + 3 \times 5 + 2 \times 2 + 0 \times 1 + 6 = 10 + 15 + 4 + 6 \equiv 3 + 1 + 4 + 6 = 14 \equiv 0$ . Hence the day of the week is the same as 7 January 1998, namely Wednesday.

[In the Gregorian calendar which is at present the internationally accepted calendar, year  $n$  is a leap year when  $n$  is divisible by 4 but not by 100, or  $n$  is divisible by 400. Thus 2000 will be a leap year but 1900 was not.]

**19.2**  $a_1 \equiv a_2$  and  $b_1 \equiv b_2 \pmod{m} \Rightarrow a_1 - a_2 = mq_1$  and  $b_1 - b_2 = mq_2$  (for some integers  $q_1$  and  $q_2$ )  $\Rightarrow (a_1 - b_1) - (a_2 - b_2) = (a_1 - a_2) - (b_1 - b_2) = mq_1 - mq_2 = m(q_1 - q_2) \Rightarrow a_1 - b_1 \equiv a_2 - b_2 \pmod{m}$ .

**19.3** If  $n = a_k a_{k-1} \dots a_2 a_1 a_0$  in decimal notation, then

$$\begin{aligned} n &= a_k \times 10^k + a_{k-1} \times 10^{k-1} + \dots + a_2 \times 10^2 + a_1 \times 10 + a_0 \\ &\equiv a_k + a_{k-1} + \dots + a_2 + a_1 + a_0 \pmod{3} \end{aligned}$$

since  $10 \equiv 1 \pmod{3}$  so that  $10^i \equiv 1$  for positive integers  $i$ . Since an integer  $m$  is divisible by 3 if and only if  $m \equiv 0 \pmod{3}$  we see that  $n$  is divisible by 3 if and only if  $a_k + a_{k-1} + \dots + a_2 + a_1 + a_0$  is divisible by 3.

† For a discussion of different calendars and a formula for the day of the week of any date see K.H. Rosen, *Elementary number theory and its applications*, Addison-Wesley, Second edition 1988.

**19.4** (i)  $6x \equiv 24 \pmod{45} \Leftrightarrow 2x \equiv 8 \pmod{15} \Leftrightarrow x \equiv 4 \pmod{15}$  since  $(2, 15) = 1$ .

Hence  $6x \equiv 24 \pmod{45} \Leftrightarrow x \equiv 4 \pmod{15}$ , i.e.  $6x \equiv 24 \pmod{45} \Leftrightarrow x \equiv 4, 19$  or  $34 \pmod{45}$ .

We can describe the solution set as  $\{4 + 15q \mid q \in \mathbb{Z}\}$ .

(ii)  $10x \equiv 15 \pmod{45} \Leftrightarrow 2x \equiv 3 \pmod{9} \Leftrightarrow 2x \equiv 12 \pmod{9} \Leftrightarrow x \equiv 6 \pmod{9}$  since  $(2, 9) = 1$ .

Hence  $10x \equiv 15 \pmod{45} \Leftrightarrow x \equiv 6 \pmod{9}$ , i.e.  $10x \equiv 15 \pmod{45} \Leftrightarrow x \equiv 6, 15, 24, 33$  or  $42 \pmod{45}$ .

We can describe the solution set as  $\{6 + 9q \mid q \in \mathbb{Z}\}$ .

**19.5**  $ab_1 \equiv ab_2 \pmod{m} \Leftrightarrow \frac{a}{(a, m)}b_1 \equiv \frac{a}{(a, m)}b_2 \pmod{\frac{m}{(a, m)}}$  by Proposition 19.3.1  $\Leftrightarrow b_1 \equiv b_2 \pmod{\frac{m}{(a, m)}}$  by Proposition 19.3.2, since  $a/(a, m)$  and  $m/(a, m)$  are coprime (by Exercise 11.4).

### Exercises 20 (page 249)

**20.1** (i) Since  $(3, 11) = 1$  there is a unique solution modulo 11. Either try the integers 0 to 10 in turn or say  $3x \equiv 5 \equiv -6 \pmod{11} \Leftrightarrow x \equiv -2 \equiv 9 \pmod{11}$  to see that the solutions are given by  $x \equiv 9 \pmod{11}$ .

(ii) Since  $(10, 35) = 5$  which does not divide 16 there are no solutions.

(iii) Since  $(10, 35) = 5$  divides 5 there are solutions. To find them:  $10x \equiv 5 \pmod{35} \Leftrightarrow 2x \equiv 1 \equiv 8 \pmod{7} \Leftrightarrow x \equiv 4 \pmod{7}$ .

(iv) Since  $(10, 21) = 1$  there is a unique solution modulo 21. Trying the integers from 0 to 20 in turn we find that the solution is  $x \equiv 16 \pmod{21}$ .

Alternatively, we can say something like  $10x \equiv 13 \pmod{21} \Leftrightarrow 10x \equiv 13 + 147 = 160 \pmod{21} \Leftrightarrow x \equiv 16 \pmod{21}$ .

**20.2** (i) Applying the Euclidean algorithm to 341 and 912 and writing each remainder as a multiple of 341 modulo 912 gives the equations at the top of the next page.

Thus 341 and 912 are coprime so that the congruence has a unique solution modulo 912. Since  $341 \times (-115) \equiv 1 \pmod{912}$ , the unique solution is given by  $x \equiv -115 \times 15 = -1725 \equiv 99 \pmod{912}$ .

Hence  $341x \equiv 15 \pmod{912} \Leftrightarrow x \equiv 99 \pmod{912}$ .

$$\begin{array}{rcl}
 912 & \equiv & 341 \times 0 \\
 341 & \equiv & 341 \times 1 & (-2) \\
 230 & \equiv & 341 \times (-2) & (-1) \\
 111 & \equiv & 341 \times 3 & (-2) \\
 8 & \equiv & 341 \times (-8) & (-13) \\
 7 & \equiv & 341 \times 107 & (-1) \\
 1 & \equiv & 341 \times (-115) & (-7) \\
 0
 \end{array}$$

(ii) Applying the Euclidean algorithm to 341 and 913 gives the following sequence of remainders.

$$\begin{array}{rcl}
 913 \\
 341 & & (-2) \\
 231 & & (-1) \\
 110 & & (-2) \\
 11 & & (-10) \\
 0
 \end{array}$$

Hence  $\gcd(341, 913) = 11$  and so the congruence  $341x \equiv 15 \pmod{913}$  has no solutions since 15 is not divisible by 11.

(iii) Applying the Euclidean algorithm to 345 and 912 and writing each remainder as a multiple of 345 modulo 912 gives the following.

$$\begin{array}{rcl}
 912 & \equiv & 345 \times 0 \\
 345 & \equiv & 345 \times 1 & (-2) \\
 222 & \equiv & 345 \times (-2) & (-1) \\
 123 & \equiv & 345 \times 3 & (-1) \\
 99 & \equiv & 345 \times (-5) & (-1) \\
 24 & \equiv & 345 \times 8 & (-4) \\
 3 & \equiv & 345 \times (-37) & (-8) \\
 0
 \end{array}$$

Hence  $\gcd(345, 912) = 3$  which divides 15 so that solutions to the linear congruence exist. Since  $15/3 = 5$  one solution is given by  $x \equiv -37 \times 5 = -185 \pmod{912}$ .

To find all the solutions observe that

$$\begin{aligned}
 345x \equiv 15 \pmod{912} &\Leftrightarrow 345x \equiv 345 \times (-185) \pmod{912} \\
 &\Leftrightarrow 115x \equiv 115 \times (-185) \pmod{304} \\
 &\Leftrightarrow x \equiv -185 \equiv 119 \pmod{304},
 \end{aligned}$$

since 115 and 304 are coprime (because we have divided through by the greatest common divisor).

Thus  $345x \equiv 15 \pmod{912} \Leftrightarrow x \equiv 119 \pmod{304}$  or, if it is desired to write the solution modulo 912,  $345x \equiv 15 \pmod{912} \Leftrightarrow x \equiv 119, 423$  or  $727 \pmod{912}$ .

**20.3** When  $a \equiv 0 \pmod{m}$  the linear congruence becomes simply  $0 \equiv b \pmod{m}$  and so has nothing to do with  $x$ . It will be satisfied (for all values of  $x$ ) when  $b$  is a multiple of  $m$  and will not be satisfied (for any value of  $x$ ) when  $b$  is not a multiple of  $m$ .

This is precisely what Theorem 20.1.7 reduces to in this case. For if  $a = 0$  then  $\gcd(a, m) = m$  so that there is a solution if and only if  $m$  divides  $b$ . In this case there are  $m$  different solutions modulo  $m$  which means that everything is a solution.

[At first glance Theorem 20.1.7 makes no sense at all when  $a = 0$  but after careful consideration we see that it does make sense but tells us something rather obvious. This phenomenon is quite common. If we have obtained some general result then it can be worth while thinking what it means in particularly simple cases. If it doesn't work in such cases then this can indicate that something has been missed in the general situation.]

### Exercises 21 (page 261)

**21.1** Suppose that  $[a_1]_m = [a_2]_m$  and  $[b_1]_m = [b_2]_m$ . Then  $a_1 \equiv a_2$  and  $b_1 \equiv b_2 \pmod{m}$  so that  $a_1 + b_1 \equiv a_2 + b_2 \pmod{m}$  and  $a_1 - b_1 \equiv a_2 - b_2 \pmod{m}$  (by Proposition 19.1.3). Hence  $[a_1 + b_1]_m = [a_2 + b_2]_m$  and  $[a_1 - b_1]_m = [a_2 - b_2]_m$  so that addition and subtraction in  $\mathbb{Z}_m$  are well-defined.

**21.2** These results are immediate since by definition (Definition 19.2.3)  $r_m(c) \equiv c \pmod{m}$  for all integers  $c$  and so in particular for  $a + b$ ,  $a - b$  and  $ab$ .

**21.3** The multiplication table for  $R_{12}$  is given below.

From this, looking which rows contain 1, we see that 1, 5, 7 and 11 are the invertible elements and that each is its own inverse.

[Notice that these are precisely the positive integers less than 12 which are coprime to 12 illustrating the result of the next exercise.]

| $\times$ | 0 | 1  | 2  | 3 | 4 | 5  | 6 | 7  | 8 | 9 | 10 | 11 |
|----------|---|----|----|---|---|----|---|----|---|---|----|----|
| 0        | 0 | 0  | 0  | 0 | 0 | 0  | 0 | 0  | 0 | 0 | 0  | 0  |
| 1        | 0 | 1  | 2  | 3 | 4 | 5  | 6 | 7  | 8 | 9 | 10 | 11 |
| 2        | 0 | 2  | 4  | 6 | 8 | 10 | 0 | 2  | 4 | 6 | 8  | 10 |
| 3        | 0 | 3  | 6  | 9 | 0 | 3  | 6 | 9  | 0 | 3 | 6  | 9  |
| 4        | 0 | 4  | 8  | 0 | 4 | 8  | 0 | 4  | 8 | 0 | 4  | 8  |
| 5        | 0 | 5  | 10 | 3 | 8 | 1  | 6 | 11 | 4 | 9 | 2  | 7  |
| 6        | 0 | 6  | 0  | 6 | 0 | 6  | 0 | 6  | 0 | 6 | 0  | 6  |
| 7        | 0 | 7  | 2  | 9 | 4 | 11 | 6 | 1  | 8 | 3 | 10 | 5  |
| 8        | 0 | 8  | 4  | 0 | 8 | 4  | 0 | 8  | 4 | 0 | 8  | 4  |
| 9        | 0 | 9  | 6  | 3 | 0 | 9  | 6 | 3  | 0 | 9 | 6  | 3  |
| 10       | 0 | 10 | 8  | 6 | 4 | 2  | 0 | 10 | 8 | 6 | 4  | 2  |
| 11       | 0 | 11 | 10 | 9 | 8 | 7  | 6 | 5  | 4 | 3 | 2  | 1  |

**21.4** The integer  $a$  is invertible modulo  $m$  precisely when the linear congruence  $ax \equiv 1 \pmod{m}$  has a solution. By Theorem 20.1.7, this congruence has a solution if and only if  $a$  and  $m$  are coprime and so this is a criterion for  $a$  to be invertible modulo  $m$ .

**21.5** From Example 20.2.2 we see that 290 and 357 are coprime and that  $290 \times (-16) \equiv 1 \pmod{357}$ . Hence  $([290]_{357})^{-1} = [-16]_{357} = [341]_{357}$ .

- (i)  $290x \equiv 2 \pmod{357} \Leftrightarrow x \equiv -16 \times 2 = -32 \equiv 325 \pmod{357} \Leftrightarrow x \equiv 325 \pmod{357}$ .
- (ii)  $290x \equiv 5 \pmod{357} \Leftrightarrow x \equiv -16 \times 5 = -80 \equiv 277 \pmod{357} \Leftrightarrow x \equiv 277 \pmod{357}$ .
- (iii)  $290x \equiv 355 \pmod{357} \Leftrightarrow x \equiv -16 \times 355 \equiv -16 \times (-2) = 32 \pmod{357} \Leftrightarrow x \equiv 32 \pmod{357}$ .

**21.6** To prove that the function  $g$  is well-defined suppose that  $[a_1]_m = [a_2]_m$ . Then  $a_1 \equiv a_2 \pmod{m}$  so that  $r_m(a_1) = r_m(a_2) \pmod{m}$  (by Proposition 19.2.4). It follows that  $r_m(a_1) + 1 = r_m(a_2) + 1$  as required.

To see that  $g$  is the inverse of  $f$  observe that, for  $i \in \mathbb{N}_m$ ,  $[a]_m = f(i) \Leftrightarrow [a]_m = [i-1]_m \Leftrightarrow a \equiv i-1 \pmod{m} \Leftrightarrow r_m(a) = i-1$  (since  $i-1 \in R_m$ )  $\Leftrightarrow r_m(a) + 1 = i \Leftrightarrow g([a]_m) = i$ .

**Exercises 22 (page 269)**

- 22.1** (i) This is the same as congruence modulo 2 and so is an equivalence relation. There are two equivalence classes: the odd integers and the even integers.
- (ii) This is not reflexive ( $0 \not\sim 0$ ), is symmetric (since  $a + b = b + a$ ) and not transitive (since  $0 \sim 1$  and  $1 \sim 0$  but  $0 \not\sim 0$ ).
- (iii) This is not reflexive ( $1 \not\sim 1$ ), is symmetric (since  $a + b = b + a$ ) and not transitive (since  $1 \sim 2$  and  $2 \sim 1$  but  $1 \not\sim 1$ ).
- (iv) This is an equivalence relation. The equivalence classes are  $\{0\}$ ,  $\{1, -1\}$ ,  $\{2, -2\}$ ,  $\{3, -3\}$ , ....
- (v) This is not reflexive ( $2 \not\sim 2$ ), is symmetric  $a \sim b \Leftrightarrow a = 1$  and  $b = 1 \Leftrightarrow b \sim a$  and is transitive ( $a \sim b$  and  $b \sim c \Rightarrow a = b = c = 1 \Rightarrow a \sim c$ ).
- (vi) This is not reflexive ( $2 \not\sim 2$ ), is symmetric and is not transitive ( $2 \sim 1$  and  $1 \sim 2$  but  $2 \not\sim 2$ ).
- (vii) This is an equivalence relation. The equivalence classes are lines parallel to the  $y$ -axis.
- (viii) This is an equivalence relation. The equivalence classes are circles with centre the origin.

**22.2** The equivalence relation may be described by  $a \sim b \Leftrightarrow ab > 0$  or  $a = b = 0$ .

**22.3** To see that  $f$  is a surjection suppose that  $a \in \mathbb{Z}$ . If  $a > 0$  then  $f(a, 0) = a$ ; if  $a = 0$  then  $f(1, 1) = 0$ ; and if  $a < 0$  then  $f(0, -a) = a$ . Thus  $f$  is a surjection. The equivalence relation is given by  $(x_1, x_2) \sim (y_1, y_2) \Leftrightarrow x_1 - x_2 = y_1 - y_2$ . The equivalence classes have the form  $\{(x_1, x_2) \mid x_1 - x_2 = k\}$  for some  $k \in \mathbb{Z}$ .

[This example gives one method of defining the set of integers  $\mathbb{Z}$  starting from the set of positive integers  $\mathbb{Z}^+$ : we may define  $\mathbb{Z}$  to be the set of equivalence classes of the above equivalence relation on  $\mathbb{Z}^+$ .]

**Exercises 23 (page 287)**

**23.1** Applying the sieve of Erastosthenes gives the array on the next page.

Thus the primes less than 200 are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109,

113, 127, 131, 137, 139, 149, 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199 (46 in all).

|     | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  |
| 21  | 22  | 23  | 24  | 25  | 26  | 27  | 28  | 29  | 30  |
| 31  | 32  | 33  | 34  | 35  | 36  | 37  | 38  | 39  | 40  |
| 41  | 42  | 43  | 44  | 45  | 46  | 47  | 48  | 49  | 50  |
| 51  | 52  | 53  | 54  | 55  | 56  | 57  | 58  | 59  | 60  |
| 61  | 62  | 63  | 64  | 65  | 66  | 67  | 68  | 69  | 70  |
| 71  | 72  | 73  | 74  | 75  | 76  | 77  | 78  | 79  | 80  |
| 81  | 82  | 83  | 84  | 85  | 86  | 87  | 88  | 89  | 90  |
| 91  | 92  | 93  | 94  | 95  | 96  | 97  | 98  | 99  | 100 |
| 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 |
| 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 |
| 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 |
| 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 |
| 141 | 142 | 143 | 144 | 145 | 146 | 147 | 148 | 149 | 150 |
| 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 |
| 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 |
| 171 | 172 | 173 | 174 | 175 | 176 | 177 | 178 | 179 | 180 |
| 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 |
| 191 | 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 |

**23.2**  $4148 = 2^2 \times 17 \times 61$ .

$$7684 = 2^2 \times 17 \times 113.$$

$$\text{Hence } \gcd(4148, 7684) = 2^2 \times 17 = 68.$$

[This is the same result as obtained in Exercise 16.1 by the Euclidean algorithm. Even with numbers of this size the method used in Chapter 16 is more efficient.]

**23.3** For  $n = 41$ ,  $n^2 - n + 41 = 41^2 - 41 + 41 = 41^2$ .

**23.4** For  $2 \leq i \leq n+1$  the integer  $(n+1)! + i$  is divisible by  $i$  and so is composite. This provides a sequence of  $n$  consecutive composite numbers.

**23.5** By the fundamental theorem of arithmetic an integer is not coprime to  $p^k$  if and only if it is a multiple of  $p$ . But now  $1 \leq pq \leq p^k$  if and only if  $1 \leq q \leq p^{k-1}$  and so there are  $p^{k-1}$  positive integers not

exceeding  $p^k$  which are not coprime to  $p^k$ . Hence the number which are coprime to  $p^k$  is  $p^k - p^{k-1}$ .

**23.6** Suppose for contradiction that there is a rational number  $q \in \mathbb{Q}$  such that  $q^3 = 2$ . Write  $q = a/b$  where  $a, b \in \mathbb{Z}^+$ . Then  $a^3 = 2b^3$ . Let  $2 < p_2 < \dots < p_r$  be the primes occurring as factors of  $a$  or of  $b$ . Then  $a = 2^{k_1}p_2^{k_2}\dots p_r^{k_r}$  and  $b = 2^{l_1}p_2^{l_2}\dots p_r^{l_r}$  where  $k_i$  and  $l_i$  are non-negative integers. This means that  $a^3 = 2^{3k_1}p_2^{3k_2}\dots p_r^{3k_r}$  and  $2b^3 = 2^{3l_1+1}p_2^{3l_2}\dots p_r^{3l_r}$ . But by the fundamental theorem these expressions must be the same and so in particular  $3k_1 = 3l_1 + 1$  which is impossible for integers  $k_1$  and  $l_1$ . This provides the necessary contradiction demonstrating that our hypothesis that there exists a rational cube root of 2 is false and so proving that no such rational number exists.

**23.7** The proof that every integer  $4q + 1$  for  $q \in \mathbb{Z}^+$  is a product of pseudo-primes is by strong induction on  $q$ .

Base case: 5 ( $q = 1$ ) is a prime and so a pseudo-prime.

Inductive step: Suppose as inductive hypothesis that each number  $4q + 1$  is a product of pseudo-primes for  $1 \leq q \leq k$ . If  $4(k + 1) + 1$  is a pseudo-prime then it is a product of a (single) pseudo-prime. Otherwise we can write  $4(k + 1) + 1 = ab$  where  $a > 1$  and  $b > 1$  and  $a \equiv b \equiv 1 \pmod{4}$ . But then, by inductive hypothesis  $a$  and  $b$  may be written as products of pseudo-primes so that, by combining these products, we write  $4(k + 1) + 1$  as a product of pseudo-primes. Hence in both cases  $4(k + 1) + 1$  is a product of pseudo-primes, proving the inductive step.

Hence by induction,  $4q + 1$  is a product of pseudo-primes for all  $q \geq 1$ .

Notice that a product of two primes congruent to 3 modulo 4 is necessarily a pseudo-prime. So for example  $3 \times 3 = 9$ ,  $3 \times 7 = 21$  and  $7 \times 7 = 49$  are pseudo-primes. Thus  $9 \times 49 = 21 \times 21 = 441$  has two distinct expressions as a product of pseudo-primes (and this is the smallest such number).

### Exercises 24 (page 294)

**24.1** By Fermat's theorem  $2^{16} \equiv 1 \pmod{17}$ . Hence  $2^{1000000} \equiv 1 \pmod{17}$  since 16 divides 1000000 and so the remainder after  $2^{1000000}$  is divided by 17 is 1.

**24.2** By Fermat's theorem,  $a^{p-1} \equiv 1 \pmod{p}$  and  $a^p \equiv a \pmod{p}$  so that  $a^{2p-1} \equiv a \pmod{p}$ . This means that  $p$  is a prime factor of  $a^{2p-1} - a$ .

However,  $a^2 \equiv a \pmod{2}$  and so, by induction on  $n$ ,  $a^n \equiv a \pmod{2}$  for all positive integers  $n$ . In particular 2 is a prime factor of  $a^{2p-1} - a$ . Since  $p$  is odd, it follows that  $2p$  is a divisor of  $a^{2p-1} - a$ , i.e.  $a^{2p-1} \equiv a \pmod{2p}$ .

**24.3** By Fermat's theorem,  $7^{16} \equiv 1 \pmod{17}$  and so an inverse of 7 modulo 17 is given by  $7^{15}$ . Now  $7^2 = 49 \equiv -2$ , so  $7^4 \equiv 4$ ,  $7^8 \equiv 16 \equiv -1$  so that  $7^{15} \equiv -1 \times 4 \times (-2) \times 7 = 56 \equiv 5$ . Hence 5 is the inverse of 7 modulo 17.

**24.4** The proof is by contradiction. Suppose that  $n > 2$ ,  $(n-1)! \equiv -1 \pmod{n}$  but that, for contradiction,  $n$  is not prime. Then there is an integer  $a$  such that  $1 < a < n$  which divides  $n$ .

- (i) Since  $1 < a < n$ , it follows that  $1 < a \leq n-1$  and so  $(n-1)! \equiv 0 \pmod{a}$ .
- (ii) Since  $a$  divides  $n$  and  $(n-1)! \equiv -1 \pmod{n}$  it follows that  $(n-1)! \equiv -1 \pmod{a}$  and so, since  $a > 1$ ,  $(n-1)! \not\equiv 0 \pmod{a}$ .

This gives the necessary contradiction.

Hence our assumption that  $n$  is not prime must be false and so  $n$  is prime as required.

**24.5** First of all notice that  $p^{q-1} + q^{p-1} \equiv q^{p-1} \equiv 1 \pmod{p}$  by Fermat's theorem since  $p$  and  $q$  are coprime being distinct primes. Similarly  $p^{q-1} + q^{p-1} \equiv 1 \pmod{q}$ . Thus  $p$  and  $q$  are amongst the primes in the prime factorization of  $p^{q-1} + q^{p-1} - 1$  which must therefore be divisible by  $pq$ . Hence  $p^{q-1} + q^{p-1} \equiv 1 \pmod{pq}$ .

## Bibliography: suggestions for further reading

- G. Birkhoff and S. Mac Lane, *A survey of modern algebra*, Macmillan, Fourth edition 1977.
- Carl B. Boyer and Uta C. Merzbach, *A history of mathematics*, Wiley, Second edition 1989.
- Victor Bryant, *Yet another introduction to analysis*, Cambridge University Press, 1990.
- David M. Burton, *Elementary number theory*, Allyn and Bacon, 1976.
- Harold Davenport, *The higher arithmetic, an introduction to the theory of numbers*, Cambridge University Press, Sixth edition 1992.
- Donald M. Davis, *The nature and power of mathematics*, Princeton University Press, 1993.
- Keith Devlin, *Sets, functions and logic, an introduction to advanced mathematics*, Chapman and Hall, Second edition 1992.
- A. Gardiner, *Infinite processes, background to analysis*, Springer-Verlag, 1982.
- Carl Friedrich Gauss, *Disquisitiones arithmeticae* (translated by Arthur A. Clarke), Springer-Verlag, 1986.
- Rod Haggerty, *Fundamentals of mathematical analysis*, Addison-Wesley, Second edition 1993.
- J.F. Humphreys and M.Y. Prest, *Numbers, groups and codes*, Cambridge University Press, 1989.
- Kenneth H. Rosen, *Elementary number theory and its applications*, Addison-Wesley, Second edition 1988.
- Alan Tucker, *Applied combinatorics*, Wiley, 1995.
- Daniel J. Velleman, *How to prove it, a structured approach*, Cambridge University Press, 1994.

# List of symbols

|   |  |
|---|--|
| $P(n)$ , $P(m, n)$ predicates 4                           | $\infty$ infinity 116                                    |
| $\pi$ pi 5  | $\chi_A$ characteristic function 117                     |
| $a \leq b$ , $a = \pm b$ 6                                | $\mathbb{N}_n = \{1, 2, \dots, n\}$ 124                  |
| $ a $ modulus of $a$ 7, 93                                | $ X $ cardinality of $X$ 124, 178                        |
| $\Rightarrow$ implies 10                                  | $D(a)$ divisors 140                                      |
| $\not\Rightarrow$ does not imply 12                       | $(a, b)$ , $\gcd(a, b)$ 141                              |
| $\square$ end of proof 17                                 | $\text{Fun}(X, Y)$ functions 145                         |
| $\sum a(i)$ sum 46  | $(n)_m$ falling factorial 146                            |
| $x^n$ power 48, 56, 109, 186                              | $\text{Inj}(X, Y)$ injections 146                        |
| $n!$ factorial 48   | $\chi_A$ characteristic function 148                     |
| $\prod a(i)$ product 55                                   | $\mathcal{P}_r(X)$ $r$ -subset 149                       |
| $\mathbb{Z}, \mathbb{Z}^+, \mathbb{Z}^{\geq}$ integers 62 | $\binom{n}{r}$ binomial coefficient 149                  |
| $\mathbb{Q}$ rationals 62, 162                            | $\aleph_0$ aleph null 171                                |
| $\mathbb{R}, \mathbb{R}^+, \mathbb{R}^{\geq}$ reals 62    | $e$ base of natural logarithms 180                       |
| $\mathbb{C}$ complex numbers 62                           | $[a, b]$ , $\text{lcm}[a, b]$ 227                        |
| $\in, \notin$ set elements 62                             | $a \equiv b \pmod{m}$ congruence 232                     |
| $\emptyset$ empty set 65                                  | $R_m$ remainders 234                                     |
| $\subseteq, \subset$ subset 66                            | $r_m: \mathbb{Z} \rightarrow R_m$ remainder function 235 |
| $A \cap B$ intersection 67                                | $[a]_m$ congruence class 251                             |
| $A \cup B$ union 67                                       | $\mathbb{Z}_m$ congruence classes 253                    |
| $A - B$ difference 67                                     | $a \sim b$ , $a \not\sim b$ relation 264                 |
| $\mathcal{P}(X)$ power set 70                             | $a \sim_{\Pi} b$ partition relation 264                  |
| $A^c$ complement 70                                       | $[a]$ equivalence class 267                              |
| $(a, b), [a, b], [a, b), (a, b]$ real intervals 72        | $X/\sim$ set of equivalence classes 267                  |
| $\forall$ for all 75                                      | $\pi(n)$ primes 285                                      |
| $\exists$ for some 76                                     | $\log_e$ natural logarithm 286                           |
| $(x, y) \in X \times Y$ Cartesian product 83              | $\text{li}(n)$ logarithmic integral 287                  |
| $c_y$ constant function 91                                | $\phi(n)$ Euler's totient function 288                   |
| $I_X$ identity function 91                                |  |
| $f A$ restriction of $f$ 94                               |  |
| $g \circ f$ composite 94                                  |  |
| $\lim f$ limit of sequence 96                             |  |
| $\text{Im } f$ image of $f$ 98, 103                       |  |
| $\sqrt{y}$ square root 105                                |  |
| $f^{-1}$ inverse function 106, 111                        |  |
| $x^{1/n}$ , $\sqrt[n]{x}$ $n$ th root 109                 |  |
| $\overleftarrow{f}, \overleftarrow{f}$ 111                |  |
| $A \Delta B$ symmetric difference 116                     |  |
| $[a, \infty), (a, \infty)$ real intervals 116             |  |

# Index

- absolute value 7
- absorption laws 115
- abstraction 61, 144, 250
- addition 18, 113, 161, 255, 256
  - addition law for inequalities 24
  - addition principle 128
- aleph null 171
- algebra 203
- algebraic number 180, 187
- algebraic number properties 18, 159, 256
  - algebraic structure 18
- algorithm 203, 280
- al-Khowarizmi, Mohammed ibn-Musa 203
- all 75
- ambiguity 22, 141
- analysis 97
- and 7, 25
- antecedent 11
- any 75, 76
- Archimedes 5
  - Archimedean axiom 166
- arithmetic 15, 191
  - arithmetic modulo a prime 290
  - arithmetic of congruence classes 254
  - arithmetic of remainders 256
- arithmetic progression 309
- arrow picture of a function 90
- associativity 19, 71, 96, 119
- asymptotic 286
- axiom 17
  - axiomatic method 18
- Babylonians 157
- base 48
- base case 40, 44
- BASIC 203
- Bernoulli's inequality 55
- bijection 102, 107, 147
- bijective 108
- Binet, J.P.M. 80
- Binet formula 50
- binomial coefficients 57, 149
  - binomial theorem 153
- Bolzano, Bernhard 172
- bound variable 75
- box picture of a function 90
- calendar 231, 336
- Cantor, Georg 177
  - Cantor–Schröder–Bernstein theorem 180, 329
  - Cantor's diagonal process 177
- cardinality 123, 178
- Cartesian product 83, 98, 245
- cause 12
- characteristic function 117, 148
- closed interval 72
- codes 279
- codomain 89
- Cohen, Paul 178
- commensurable 158
- common divisor 140
- commutativity 18, 71, 95, 119
- complement of a set 70
- completeness axiom 166
- complex number 62, 156, 159, 306
- composite number 277
- composition of functions 94
- conclusion 11
- conditional statement 12
  - conditional definition of set 63
- congruence 231, 232
  - congruence class 231, 250, 251, 262
- conjunction 7
- consequent 11
- constant function 91
- constructive definition of set 64
- continuity axiom 166
- continuum hypothesis 178
- contradiction 30
- contrapositive 34

convergence 96  
 converse 14  
 coordinate 83  
 coprime integers 141, 212, 242  
 cosine function 5, 109  
 countable set 171  
 counterexample 79  
 counting 123, 144, 157, 170, 325  
     counting argument 133, 170, 250  
 cryptography 279  
 cypherization 157  
 days of the week 231  
 decimal 164  
     decimal arithmetic 168  
 de la Vallée-Poussin, C.J. 286  
 De Morgan, Augustus 71  
     De Morgan laws 71  
 Dedekind, Richard 39, 112, 158, 173,  
     251  
 definition 15  
     inductive 45  
     recursive 45  
 Deligne, Pierre 152  
 denominator 159  
 denumerable set 171, 174  
 derangement 185  
 diagonal set 130  
 difference of sets 67  
 Diophantus 216  
     diophantine equation 216, 245  
 Dirichlet, Lejeune 136, 218  
 disjoint sets 67  
 disjunction 6  
 disproof by counterexample 79  
 disproof of statements involving  
     quantifiers 79  
 distributivity 19, 71  
 divides 16  
     3 divisibility by 239  
     9 divisibility by 271  
     11 divisibility by 271  
     division 19, 159, 162, 237  
     division theorem 141, 191, 226  
     divisor 140, 283  
 domain 89  
 dummy variable 64, 75  
 $e$  180  
 each 75  
 element 62  
 empty set 65  
 enumerable set 171  
 equal functions 93  
 equal sets 65  
 equipotent sets 171  
 equivalence class 266  
 equivalence relation 265  
 equivalent 15, 300  
 Erastosthenes of Cyrene 280  
     sieve of Erastosthenes 280  
 Euclid 17, 50, 158, 200, 217, 278, 285  
     Euclidean algorithm 200, 202, 225,  
     245  
     Euclidean plane 84  
 Eudoxus 158  
 Euler, Leonhard 4, 218, 285, 289, 297  
     Euler's totient function 288, 296  
 even integer 16, 142, 231, 235, 250, 251,  
     254  
 every 75  
 exhaustive cases 26  
 existence 13, 192, 287  
     existential quantifier 76  
     existential statement 75  
 exponent 48, 56, 109, 186  
     laws of exponents 52, 56, 186  
     exponential function 287  
 factor 140  
 factorial 48  
 falling factorial 146  
 false 6  
 Fermat, Pierre de 217, 289  
     Fermat's last theorem 217  
     Fermat's little theorem 290, 297  
 Fibonacci 49  
     Fibonacci number 49, 57, 225, 272  
     Fibonacci's rabbit problem 57  
 field 18  
 finite decimal 165, 186  
 finite set 128, 133, 136  
 formula 91  
 fraction 159, 268  
 free variable 4, 74  
 function 89, 99, 144  
     of two variables 317  
 fundamental theorem of arithmetic 282  
 Galileo Galilei 172  
 Gauss, Carl Friedrich 232, 281, 285, 297  
 geometric progression 309  
 geometry 17  
 given-goal diagram 23  
 Goldbach, Christian 4  
     Goldbach conjecture 4  
 golden ratio 50  
 graph 98  
 greatest common divisor (gcd) 140, 199,  
     207, 211, 225, 227, 295  
 group theory 289  
 Hadamard, Jacques 286  
 half-infinite interval 116  
 half-open interval 72  
 Hermite, Charles 180  
 highest common factor 140  
 Hindu–Arabic numerals 157  
 homogeneous equation 221

**Humpty Dumpty's principle** 15  
**hypothesis** 11  
**identity function** 91  
**if** 14, 106  
    if and only if 15  
    iff 15  
**image of an element** 89  
**image of a function** 97, 103, 111  
**implication** 10  
    implies 14  
**inclusion function** 95  
**inclusion-exclusion principle** 131, 182,  
    296  
**incommensurable** 158  
**index** 48, 56, 109, 186  
    laws of indices 52, 56, 186  
**induction principle** 39, 139, 184  
    strong induction 48, 88  
**inductive definition** 45  
    inductive hypothesis 40  
    inductive step 40  
**inequality axioms** 23  
**infinity** 116  
    infinite decimal 164, 166, 186, 198,  
        225, 296  
    infinite set 128, 170  
    infinite sequence 97  
**inhomogeneous equation** 221  
**injection** 101, 146  
**integer** 15, 62  
**integral domain** 18  
**intersection of sets** 66  
**interval of real numbers** 72, 116  
    closed 72  
    half-infinite 116  
    half-open 72  
    open 72  
**inverse function** 106, 320  
    right inverse 118  
**inverse modulo  $m$**  260  
**invertible function** 106  
**invertible modulo  $m$**  260  
**irrational number** 164  
**isomorphism** 257  
**Ivory, James** 290  
**Lagrange, Joseph** 291  
**Lamé, Gabriel** 203  
    Lamé's theorem 226  
**least common multiple (lcm)** 226, 295  
**Legendre, Adrien-Marie** 218, 286  
**Leibniz, Gottfried** 289  
    Leibniz's rule 185  
**lemma** 127  
**Leonardo of Pisa** 49  
**limit** 96  
**Lindemann, Ferdinand** 180  
**linear combination** 207  
**linear congruence** 238, 240  
**linear diophantine equation** 216, 258  
**Liouville, Joseph** 180  
**listing elements** 63, 89  
**Littlewood, J.E.** 287  
**logarithm** 286  
    logarithmic integral 287  
**logical connectives** 5  
**lowest terms** 160  
**magnitude** 158  
**map** 89  
**mathematical invention** 27  
**mathematical statement** 3  
**mathematical truth** 17  
**maximum** 138  
**measurement** 157  
**minimum** 138  
**modular arithmetic** 233  
    modulus 231  
**modulus function** 7, 93  
**multiple** 140, 116  
**multiplication** 18, 113, 161, 255, 256  
    multiplication law for inequalities 24  
    multiplication principle 130  
**natural logarithm** 286  
**natural number** 62, 123  
**necessary** 14  
    necessary and sufficient 15  
**negation** 7  
**negative integer** 157  
**Newton, Isaac** 153  
**non-constructive proof** 259  
**non-existence result** 30  
**not** 7  
**null sequence** 96  
**number** 123, 157  
    number theory 15  
**numerator** 159  
**odd integer** 16, 32, 142, 231, 250, 251,  
    254  
**one-to-one function** 101  
**only if** 14  
**onto function** 102  
**open interval** 72  
**or** 5, 25, 27, 35, 304  
**order axioms** 23  
**order modulo a prime** 291, 296  
**ordered pair** 83  
**paradox** 172  
**partition** 252, 262, 266  
**Pascal, Blaise** 151  
    Pascal's triangle 151  
**Peano, Giuseppe** 18, 62, 112  
    Peano's axioms 112  
**Pearce, Benjamin** 71  
**perfect square** 197, 198  
**permutation** 147

pi 5  
 pigeonhole principle 133, 136, 146, 180,  
     250, 259  
 place value system 157  
 positive integers 62, 112  
 power set 70, 148, 178, 262  
 powers 48, 56, 109, 186  
     powers modulo a prime 291  
 predicate 4, 74, 81  
 pre-image 102, 111  
 prime number 4, 277  
     primality test 293  
     prime factor 280  
     prime factorization 282  
     prime number theorem 286  
 primitive root 291  
 principal value of inverse circular  
     functions 109  
 product 55, 113  
 proof 10, 21, 77  
     backwards 27  
     by cases 25  
     by contradiction 30, 163, 285  
     by contrapositive 34  
     by counterexample 79  
     by enumeration 26  
     by example 77  
     by induction 39, 80  
     construction 22  
     direct 21  
     non-constructive 259  
     of statements involving quantifiers 77  
     template 32, 42, 44, 49  
 proper subset 66  
 proposition 3, 16, 74  
 public-key cryptography 279, 289  
 Pythagoras 158  
     Pythagoras's theorem 158, 161  
     Pythagorean triples 217  
     Pythagoreans 158  
 quantifiers 74  
 quotient 191  
 rational number 62, 159, 176, 268  
 real number 62, 164, 166  
 recurring decimal 166, 225, 296  
 reflexivity 233, 264  
 relation 264  
 relatively prime 141  
 remainder 191, 231, 234  
     remainder map 235  
 residue 234  
 restriction 94  
 right inverse 118  
 root 109  
 sequence 96  
 set 61  
 sexagesimal fraction 157  
 sine function 108  
 singleton set 70  
 Skewes, S. 287  
 solution set 65, 245  
 some 76  
 square root 105  
     square root of 2 irrational 162, 165,  
         284  
 squaring the circle 180  
 statement 3  
 strong induction 48, 88, 278  
 subset 66, 148  
 subtraction 19, 162, 255, 256  
 successor 39, 112  
     successor function 112  
 sufficient 14  
 sum 46, 113  
 sum of squares 225, 236  
 surjection 102, 184  
 syllogism 25  
 symbols 24  
 symmetric difference 116  
 symmetry 233, 264  
 tangent function 109  
 then 14  
 transcendental number 180  
 transitivity 24, 233, 264  
 triangle inequality 306  
 trichotomy law 24  
 true 6  
     truth table 6, 67  
     truth value 6  
 uncountable set 171, 176  
 union of sets 67  
 uniqueness 192, 235, 243  
 unit 277  
 unity 19  
 universal quantifier 75  
 universal set 70  
 universal statement 13, 74, 301  
     universal implication 13  
 value 4, 89  
 Venn diagram 68  
 Waring, Edward 291  
 well-defined 93, 125, 161, 186, 235, 255,  
     268, 273, 301  
 well-ordering principle 139, 143, 184  
 Wiles, Andrew 218  
 Wilson, John 291  
 writing mathematics 22  
 zero 19, 157

CORRECTIONS TO  
**An Introduction to Mathematical Reasoning**

BY PETER J. ECCLES  
NOVEMBER 16, 2009

*Thanks to Jerry Folland and John Palmieri for pointing out these corrections.*

- p. 56, problem 21:  $x + 1/x$  is to be interpreted as  $x + (1/x)$ , not  $(x + 1)/x$ . Same goes for  $x^n + 1/x^n$ .
- p. 78, line 6 from bottom:  $a \rightarrow n$ .
- p. 97, last line of proof of Example 8.3.3:  $\text{is} \rightarrow \text{as}$ .
- p. 125, 5th line of text:  $n - i$  should be  $n + 1 - i$ .
- p. 184, problem 14:  $N$  should be  $\mathbb{N}$ .
- p. 185, problem 18, line 4: the summation should be from 0 to  $k$ , not 0 to  $n$ .
- p. 185, problem 20, second line: change “disjoint” to “distinct.”
- p. 267, second displayed equation:  $\mathbb{Z}$  should be  $X$ .
- p. 269, last line:  $(a_1, a_2) = (b_1, b_2)$  should be  $(a_1, a_2) \sim (b_1, b_2)$ .
- p. 270, problem 22.2:  $Z^+$  should be  $\mathbb{Z}^+$ .
- p. 323, last 3 lines: In the last two sentences  $X$  and  $Y$  need to be switched.
- p. 327, solution to problem 14.1, line 2:  $|X|$  should be  $|A|$ .
- p. 341, solution to problem 22.3: Change  $f(a, 0)$  to  $f(a + 1, 1)$ , and  $f(0, -a)$  to  $f(1, 1 - a)$ .

MORE CORRECTIONS TO  
**An Introduction to Mathematical Reasoning**  
BY PETER J. ECCLES  
APRIL 19, 2013

- p. 57, problem 25: the right side of the formula is actually not defined if  $m = 1$ , because  $u_{m-1}$  appears. So either prove this for all positive integers  $n$  and all integers  $m \geq 2$ , or define  $u_0$  to be zero.
- p. 95, Proposition 8.2.3(i): the codomain of the composition should be  $W$ , not  $Z$ .
- p. 117, problem 14: the set in the last line of the problem should be  $\{x \in \mathbb{R} \mid f \circ g(x) = g \circ f(x)\}$ .
- p. 225, problem 7: it should say “takes precisely  $n - 1$  steps” (instead of  $n$  steps), and it should be for  $n \geq 2$ .



# CALCULUS

Third Edition

*Michael Spivak*

# **CALCULUS**

# CALCULUS

Third Edition

*Michael Spivak*

**CALCULUS**  
Third Edition

*Copyright © 1967, 1980, 1994 by Michael Spivak  
All rights reserved  
Library of Congress Catalog Card Number 80-82517*

**Publish or Perish, Inc.**  
1572 West Gray, #377  
Houston, Texas 77019 

*Manufactured in the United States of America  
ISBN 0-914098-89-6*

*Dedicated to the Memory of Y. P.*

# PREFACE

*I hold every man a debtor  
to his profession,  
from the which as men of course  
doe seeke to receive countenance and profit,  
so ought they of duty to endeavour  
themselves by way of amends,  
to be a help and  
ornament thereunto.*

FRANCIS BACON

## PREFACE TO THE FIRST EDITION

Every aspect of this book was influenced by the desire to present calculus not merely as a prelude to but as the first real encounter with mathematics. Since the foundations of analysis provided the arena in which modern modes of mathematical thinking developed, calculus ought to be the place in which to expect, rather than avoid, the strengthening of insight with logic. In addition to developing the students' intuition about the beautiful concepts of analysis, it is surely equally important to persuade them that precision and rigor are neither deterrents to intuition, nor ends in themselves, but the natural medium in which to formulate and think about mathematical questions.

This goal implies a view of mathematics which, in a sense, the entire book attempts to defend. No matter how well particular topics may be developed, the goals of this book will be realized only if it succeeds as a whole. For this reason, it would be of little value merely to list the topics covered, or to mention pedagogical practices and other innovations. Even the cursory glance customarily bestowed on new calculus texts will probably tell more than any such extended advertisement, and teachers with strong feelings about particular aspects of calculus will know just where to look to see if this book fulfills their requirements.

A few features do require explicit comment, however. Of the twenty-nine chapters in the book, two (starred) chapters are optional, and the three chapters comprising Part V have been included only for the benefit of those students who might want to examine on their own a construction of the real numbers. Moreover, the appendices to Chapters 3 and 11 also contain optional material.

The order of the remaining chapters is intentionally quite inflexible, since the purpose of the book is to present calculus as the evolution of one idea, not as a collection of "topics." Since the most exciting concepts of calculus do not appear until Part III, it should be pointed out that Parts I and II will probably require less time than their length suggests—although the entire book covers a one-year course, the chapters are not meant to be covered at any uniform rate. A rather natural dividing point does occur between Parts II and III, so it is possible to reach differentiation and integration even more quickly by treating Part II very briefly, perhaps returning later for a more detailed treatment. This arrangement corresponds to the traditional organization of most calculus courses, but I feel that it will only diminish the value of the book for students who have seen a small amount of calculus previously, and for bright students with a reasonable background.

The problems have been designed with this particular audience in mind. They range from straightforward, but not overly simple, exercises which develop basic techniques and test understanding of concepts, to problems of considerable difficulty and, I hope, of comparable interest. There are about 625 problems in all. Those which emphasize manipulations usually contain many example, numbered

with small Roman numerals, while small letters are used to label interrelated parts in other problems. Some indication of relative difficulty is provided by a system of starring and double starring, but there are so many criteria for judging difficulty, and so many hints have been provided, especially for harder problems, that this guide is not completely reliable. Many problems are so difficult, especially if the hints are not consulted, that the best of students will probably have to attempt only those which especially interest them; from the less difficult problems it should be easy to select a portion which will keep a good class busy, but not frustrated. The answer section contains solutions to about half the examples from an assortment of problems that should provide a good test of technical competence. A separate answer book contains the solutions of the other parts of these problems, and of all the other problems as well. Finally, there is a Suggested Reading list, to which the problems often refer, and a glossary of symbols.

I am grateful for the opportunity to mention the many people to whom I owe my thanks. Jane Bjorkgren performed prodigious feats of typing that compensated for my fitful production of the manuscript. Richard Serkey helped collect the material which provides historical sidelights in the problems, and Richard Weiss supplied the answers appearing in the back of the book. I am especially grateful to my friends Michael Freeman, Jay Goldman, Anthony Phillips, and Robert Wells for the care with which they read, and the relentlessness with which they criticized, a preliminary version of the book. Needles to say, they are not responsible for the deficiencies which remain, especially since I sometimes rejected suggestions which would have made the book appear suitable for a larger group of students. I must express my admiration for the editors and staff of W. A. Benjamin, Inc., who were always eager to increase the appeal of the book, while recognizing the audience for which it was intended.

The inadequacies which preliminary editions always involve were gallantly endured by a rugged group of freshmen in the honors mathematics course at Brandeis University during the academic year 1965–1966. About half of this course was devoted to algebra and topology, while the other half covered calculus, with the preliminary edition as the text. It is almost obligatory in such circumstances to report that the preliminary version was a gratifying success. This is always safe—after all, the class is unlikely to rise up in a body and protest publicly—but the students themselves, it seems to me, deserve the right to assign credit for the thoroughness with which they absorbed an impressive amount of mathematics. I am content to hope that some other students will be able to use the book to such good purpose, and with such enthusiasm.

*Waltham, Massachusetts  
February 1967*

**MICHAEL SPIVAK**

## PREFACE TO THE SECOND EDITION

I have often been told that the title of this book should really be something like “An Introduction to Analysis,” because the book is usually used in courses where the students have already learned the mechanical aspects of calculus—such courses are standard in Europe, and they are becoming more common in the United States. After thirteen years it seems too late to change the title, but other changes, in addition to the correction of numerous misprints and mistakes, seemed called for. There are now separate Appendices for many topics that were previously slighted: polar coordinates, uniform continuity, parameterized curves, Riemann sums, and the use of integrals for evaluating lengths, volumes and surface areas. A few topics, like manipulations with power series, have been discussed more thoroughly in the text, and there are also more problems on these topics, while other topics, like Newton’s method and the trapezoid rule and Simpson’s rule, have been developed in the problems. There are in all about 160 new problems, many of which are intermediate in difficulty between the few routine problems at the beginning of each chapter and the more difficult ones that occur later.

Most of the new problems are the work of Ted Shifrin. Frederick Gordon pointed out several serious mistakes in the original problems, and supplied some non-trivial corrections, as well as the neat proof of Theorem 12-2, which took two Lemmas and two pages in the first edition. Joseph Lipman also told me of this proof, together with the similar trick for the proof of the last theorem in the Appendix to Chapter 11, which went unproved in the first edition. Roy O. Davies told me the trick for Problem 11-66, which previously was proved only in Problem 20-8 [21-8 in the third edition], and Marina Ratner suggested several interesting problems, especially ones on uniform continuity and infinite series. To all these people go my thanks, and the hope that in the process of fashioning the new edition their contributions weren’t too badly botched.

MICHAEL SPIVAK

## PREFACE TO THE THIRD EDITION

The most significant change in this third edition is the inclusion of a new (starred) Chapter 17 on planetary motion, in which calculus is employed for a substantial physics problem.

In preparation for this, the old Appendix to Chapter 4 has been replaced by three Appendices: the first two cover vectors and conic sections, while polar coordinates are now deferred until the third Appendix, which also discusses the polar coordinate equations of the conic sections. Moreover, the Appendix to Chapter 12 has been extended to treat vector operations on vector-valued curves.

Another large change is merely a rearrangement of old material: “The Cosmopolitan Integral,” previously a second Appendix to Chapter 13, is now an Appendix to the chapter on “Integration in Elementary Terms” (previously Chapter 18, now Chapter 19); moreover, those problems from that chapter which used the material from that Appendix now appear as problems in the newly placed Appendix.

A few other changes and renumbering of Problems result from corrections, and elimination of incorrect problems.

I was both startled and somewhat dismayed when I realized that after allowing 13 years to elapse between the first and second editions of the book, I have allowed another 14 years to elapse before this third edition. During this time I seem to have accumulated a not-so-short list of corrections, but no longer have the original communications, and therefore cannot properly thank the various individuals involved (who by now have probably lost interest anyway). I have had time to make only a few changes to the Suggested Reading, which after all these years probably requires a complete revision; this will have to wait until the next edition, which I hope to make in a more timely fashion.

MICHAEL SPIVAK

# CONTENTS

PREFACE vi

**PART I Prologue**

- 1 Basic Properties of Numbers 3
- 2 Numbers of Various Sorts 21

**PART II Foundations**

- 3 Functions 39
  - Appendix. Ordered Pairs* 54
- 4 Graphs 56
  - Appendix 1. Vectors* 75
  - Appendix 2. The Conic Sections* 80
  - Appendix 3. Polar Coordinates* 84
- 5 Limits 90
- 6 Continuous Functions 113
- 7 Three Hard Theorems 120
- 8 Least Upper Bounds 131
  - Appendix. Uniform Continuity* 142

**PART III Derivatives and Integrals**

- 9 Derivatives 147
- 10 Differentiation 166
- 11 Significance of the Derivative 185
  - Appendix. Convexity and Concavity* 216
- 12 Inverse Functions 227
  - Appendix. Parametric Representation of Curves* 241
- 13 Integrals 250
  - Appendix. Riemann Sums* 279
- 14 The Fundamental Theorem of Calculus 282

- 15 The Trigonometric Functions 300
- \*16  $\pi$  is Irrational 321
- \*17 Planetary Motion 327
- 18 The Logarithm and Exponential Functions 336
- 19 Integration in Elementary Terms 359
- Appendix. The Cosmopolitan Integral* 397

**PART IV**

**Infinite Sequences and Infinite Series**

- 20 Approximation by Polynomial Functions 405
- \*21  $e$  is Transcendental 435
- 22 Infinite Sequences 445
- 23 Infinite Series 464
- 24 Uniform Convergence and Power Series 491
- 25 Complex Numbers 517
- 26 Complex Functions 532
- 27 Complex Power Series 546

**PART V**

**Epilogue**

- 28 Fields 571
- 29 Construction of the Real Numbers 578
- 30 Uniqueness of the Real Numbers 591
  - Suggested Reading* 599
  - Answers (to selected problems)* 609
  - Glossary of Symbols* 655
  - Index* 659

# **CALCULUS**

PART I  
PROLOGUE

*To be conscious that  
you are ignorant is a great step  
to knowledge.*

**BENJAMIN DISRAELI**

The title of this chapter expresses in a few words the mathematical knowledge required to read this book. In fact, this short chapter is simply an explanation of what is meant by the “basic properties of numbers,” all of which—addition and multiplication, subtraction and division, solutions of equations and inequalities, factoring and other algebraic manipulations—are already familiar to us. Nevertheless, this chapter is not a review. Despite the familiarity of the subject, the survey we are about to undertake will probably seem quite novel; it does not aim to present an extended review of old material, but to condense this knowledge into a few simple and obvious properties of numbers. Some may even seem too obvious to mention, but a surprising number of diverse and important facts turn out to be consequences of the ones we shall emphasize.

Of the twelve properties which we shall study in this chapter, the first nine are concerned with the fundamental operations of addition and multiplication. For the moment we consider only addition: this operation is performed on a pair of numbers—the sum  $a + b$  exists for any two given numbers  $a$  and  $b$  (which may possibly be the same number, of course). It might seem reasonable to regard addition as an operation which can be performed on several numbers at once, and consider the sum  $a_1 + \dots + a_n$  of  $n$  numbers  $a_1, \dots, a_n$  as a basic concept. It is more convenient, however, to consider addition of pairs of numbers only, and to define other sums in terms of sums of this type. For the sum of three numbers  $a$ ,  $b$ , and  $c$ , this may be done in two different ways. One can first add  $b$  and  $c$ , obtaining  $b + c$ , and then add  $a$  to this number, obtaining  $a + (b + c)$ ; or one can first add  $a$  and  $b$ , and then add the sum  $a + b$  to  $c$ , obtaining  $(a + b) + c$ . Of course, the two compound sums obtained are equal, and this fact is the very first property we shall list:

(P1) If  $a$ ,  $b$ , and  $c$  are any numbers, then

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

The statement of this property clearly renders a separate concept of the sum of three numbers superfluous; we simply agree that  $a + b + c$  denotes the number  $a + (b + c) = (a + b) + c$ . Addition of four numbers requires similar, though slightly more involved, considerations. The symbol  $a + b + c + d$  is defined to mean

- (1)  $((a + b) + c) + d$ ,
- or (2)  $(a + (b + c)) + d$ ,
- or (3)  $a + ((b + c) + d)$ ,
- or (4)  $a + (b + (c + d))$ ,
- or (5)  $(a + b) + (c + d)$ .

This definition is unambiguous since these numbers are all equal. Fortunately, *this* fact need not be listed separately, since it follows from the property P1 already listed. For example, we know from P1 that

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

and it follows immediately that (1) and (2) are equal. The equality of (2) and (3) is a direct consequence of P1, although this may not be apparent at first sight (one must let  $b + c$  play the role of  $b$  in P1, and  $d$  the role of  $c$ ). The equalities (3) = (4) = (5) are also simple to prove.

It is probably obvious that an appeal to P1 will also suffice to prove the equality of the 14 possible ways of summing five numbers, but it may not be so clear how we can reasonably arrange a proof that this is so without actually listing these 14 sums. Such a procedure is feasible, but would soon cease to be if we considered collections of six, seven, or more numbers; it would be totally inadequate to prove the equality of all possible sums of an arbitrary finite collection of numbers  $a_1, \dots, a_n$ . This fact may be taken for granted, but for those who would like to worry about the proof (and it is worth worrying about once) a reasonable approach is outlined in Problem 24. Henceforth, we shall usually make a tacit appeal to the results of this problem and write sums  $a_1 + \dots + a_n$  with a blithe disregard for the arrangement of parentheses.

The number 0 has one property so important that we list it next:

(P2) If  $a$  is any number, then

$$a + 0 = 0 + a = a.$$

An important role is also played by 0 in the third property of our list:

(P3) For every number  $a$ , there is a number  $-a$  such that

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

Property P2 ought to represent a distinguishing characteristic of the number 0, and it is comforting to note that we are already in a position to prove this. Indeed, if a number  $x$  satisfies

$$a + x = a$$

for any one number  $a$ , then  $x = 0$  (and consequently this equation also holds for all numbers  $a$ ). The proof of this assertion involves nothing more than subtracting  $a$  from both sides of the equation, in other words, adding  $-a$  to both sides; as the following detailed proof shows, all three properties P1–P3 must be used to justify this operation.

|       |                                  |
|-------|----------------------------------|
| If    | $a + x = a,$                     |
| then  | $(-a) + (a + x) = (-a) + a = 0;$ |
| hence | $((-a) + a) + x = 0;$            |
| hence | $0 + x = 0;$                     |
| hence | $x = 0.$                         |

As we have just hinted, it is convenient to regard subtraction as an operation derived from addition: we consider  $a - b$  to be an abbreviation for  $a + (-b)$ . It is then possible to find the solution of certain simple equations by a series of steps (each justified by P1, P2, or P3) similar to the ones just presented for the equation  $a + x = a$ . For example:

$$\begin{aligned} \text{If } & x + 3 = 5, \\ \text{then } & (x + 3) + (-3) = 5 + (-3); \\ \text{hence } & x + (3 + (-3)) = 5 - 3 = 2; \\ \text{hence } & x + 0 = 2; \\ \text{hence } & x = 2. \end{aligned}$$

Naturally, such elaborate solutions are of interest only until you become convinced that they can always be supplied. In practice, it is usually just a waste of time to solve an equation by indicating so explicitly the reliance on properties P1, P2, and P3 (or any of the further properties we shall list).

Only one other property of addition remains to be listed. When considering the sums of three numbers  $a$ ,  $b$ , and  $c$ , only two sums were mentioned:  $(a + b) + c$  and  $a + (b + c)$ . Actually, several other arrangements are obtained if the order of  $a$ ,  $b$ , and  $c$  is changed. That these sums are all equal depends on

(P4) If  $a$  and  $b$  are any numbers, then

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

The statement of P4 is meant to emphasize that although the operation of addition of pairs of numbers might conceivably depend on the order of the two numbers, in fact it does not. It is helpful to remember that not all operations are so well behaved. For example, subtraction does not have this property: usually  $a - b \neq b - a$ . In passing we might ask just when  $a - b$  does equal  $b - a$ , and it is amusing to discover how powerless we are if we rely only on properties P1–P4 to justify our manipulations. Algebra of the most elementary variety shows that  $a - b = b - a$  only when  $a = b$ . Nevertheless, it is impossible to derive this fact from properties P1–P4; it is instructive to examine the elementary algebra carefully and determine which step(s) cannot be justified by P1–P4. We will indeed be able to justify all steps in detail when a few more properties are listed. Oddly enough, however, the crucial property involves multiplication.

The basic properties of multiplication are fortunately so similar to those for addition that little comment will be needed; both the meaning and the consequences should be clear. (As in elementary algebra, the product of  $a$  and  $b$  will be denoted by  $a \cdot b$ , or simply  $ab$ .)

(P5) If  $a$ ,  $b$ , and  $c$  are any numbers, then

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c.$$

(P6) If  $a$  is any number, then

$$a \cdot 1 = 1 \cdot a = a.$$

## 6 Prologue

Moreover,  $1 \neq 0$ .

(The assertion that  $1 \neq 0$  may seem a strange fact to list, but we have to list it, because there is no way it could possibly be proved on the basis of the other properties listed—these properties would all hold if there were only one number, namely, 0.)

(P7) For every number  $a \neq 0$ , there is a number  $a^{-1}$  such that

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

(P8) If  $a$  and  $b$  are any numbers, then

$$a \cdot b = b \cdot a.$$

One detail which deserves emphasis is the appearance of the condition  $a \neq 0$  in P7. This condition is quite necessary; since  $0 \cdot b = 0$  for all numbers  $b$ , there is no number  $0^{-1}$  satisfying  $0 \cdot 0^{-1} = 1$ . This restriction has an important consequence for division. Just as subtraction was defined in terms of addition, so division is defined in terms of multiplication: The symbol  $a/b$  means  $a \cdot b^{-1}$ . Since  $0^{-1}$  is meaningless,  $a/0$  is also meaningless—division by 0 is *always* undefined.

Property P7 has two important consequences. If  $a \cdot b = a \cdot c$ , it does not necessarily follow that  $b = c$ ; for if  $a = 0$ , then both  $a \cdot b$  and  $a \cdot c$  are 0, no matter what  $b$  and  $c$  are. However, if  $a \neq 0$ , then  $b = c$ ; this can be deduced from P7 as follows:

$$\begin{aligned} &\text{If } a \cdot b = a \cdot c \text{ and } a \neq 0, \\ &\text{then } a^{-1} \cdot (a \cdot b) = a^{-1} \cdot (a \cdot c); \\ &\text{hence } (a^{-1} \cdot a) \cdot b = (a^{-1} \cdot a) \cdot c; \\ &\text{hence } 1 \cdot b = 1 \cdot c; \\ &\text{hence } b = c. \end{aligned}$$

It is also a consequence of P7 that if  $a \cdot b = 0$ , then either  $a = 0$  or  $b = 0$ . In fact,

$$\begin{aligned} &\text{if } a \cdot b = 0 \text{ and } a \neq 0, \\ &\text{then } a^{-1} \cdot (a \cdot b) = 0; \\ &\text{hence } (a^{-1} \cdot a) \cdot b = 0; \\ &\text{hence } 1 \cdot b = 0; \\ &\text{hence } b = 0. \end{aligned}$$

(It may happen that  $a = 0$  and  $b = 0$  are both true; this possibility is not excluded when we say “either  $a = 0$  or  $b = 0$ ”; in mathematics “or” is always used in the sense of “one or the other, or both.”)

This latter consequence of P7 is constantly used in the solution of equations. Suppose, for example, that a number  $x$  is known to satisfy

$$(x - 1)(x - 2) = 0.$$

Then it follows that either  $x - 1 = 0$  or  $x - 2 = 0$ ; hence  $x = 1$  or  $x = 2$ .

On the basis of the eight properties listed so far it is still possible to prove very little. Listing the next property, which combines the operations of addition and multiplication, will alter this situation drastically.

(P9) If  $a$ ,  $b$ , and  $c$  are any numbers, then

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

(Notice that the equation  $(b + c) \cdot a = b \cdot a + c \cdot a$  is also true, by P8.)

As an example of the usefulness of P9 we will now determine just when  $a - b = b - a$ :

|               |   |
|---------------|---|
| If            | $a - b = b - a$ ,                           |
| then          | $(a - b) + b = (b - a) + b = b + (b - a)$ ; |
| hence         | $a = b + b - a$ ;                           |
| hence         | $a + a = (b + b - a) + a = b + b$ .         |
| Consequently  | $a \cdot (1 + 1) = b \cdot (1 + 1)$ ,       |
| and therefore | $a = b$ .                                   |

A second use of P9 is the justification of the assertion  $a \cdot 0 = 0$  which we have already made, and even used in a proof on page 6 (can you find where?). This fact was not listed as one of the basic properties, even though no proof was offered when it was first mentioned. With P1–P8 alone a proof was not possible, since the number 0 appears only in P2 and P3, which concern addition, while the assertion in question involves multiplication. With P9 the proof is simple, though perhaps not obvious: We have

$$\begin{aligned} a \cdot 0 + a \cdot 0 &= a \cdot (0 + 0) \\ &= a \cdot 0; \end{aligned}$$

as we have already noted, this immediately implies (by adding  $-(a \cdot 0)$  to both sides) that  $a \cdot 0 = 0$ .

A series of further consequences of P9 may help explain the somewhat mysterious rule that the product of two negative numbers is positive. To begin with, we will establish the more easily acceptable assertion that  $(-a) \cdot b = -(a \cdot b)$ . To prove this, note that

$$\begin{aligned} (-a) \cdot b + a \cdot b &= [(-a) + a] \cdot b \\ &= 0 \cdot b \\ &= 0. \end{aligned}$$

It follows immediately (by adding  $-(a \cdot b)$  to both sides) that  $(-a) \cdot b = -(a \cdot b)$ . Now note that

$$\begin{aligned} (-a) \cdot (-b) + [-(a \cdot b)] &= (-a) \cdot (-b) + (-a) \cdot b \\ &= (-a) \cdot [(-b) + b] \\ &= (-a) \cdot 0 \\ &= 0. \end{aligned}$$

Consequently, adding  $(a \cdot b)$  to both sides, we obtain

$$(-a) \cdot (-b) = a \cdot b.$$

The fact that the product of two negative numbers is positive is thus a consequence of P1–P9. In other words, *if we want P1 to P9 to be true, the rule for the product of two negative numbers is forced upon us.*

The various consequences of P9 examined so far, although interesting and important, do not really indicate the significance of P9; after all, we could have listed each of these properties separately. Actually, P9 is the justification for almost all algebraic manipulations. For example, although we have shown how to solve the equation

$$(x - 1)(x - 2) = 0,$$

we can hardly expect to be presented with an equation in this form. We are more likely to be confronted with the equation

$$x^2 - 3x + 2 = 0.$$

The “factorization”  $x^2 - 3x + 2 = (x - 1)(x - 2)$  is really a triple use of P9:

$$\begin{aligned} (x - 1) \cdot (x - 2) &= x \cdot (x - 2) + (-1) \cdot (x - 2) \\ &= x \cdot x + x \cdot (-2) + (-1) \cdot x + (-1) \cdot (-2) \\ &= x^2 + x[(-2) + (-1)] + 2 \\ &= x^2 - 3x + 2. \end{aligned}$$

A final illustration of the importance of P9 is the fact that this property is actually used every time one multiplies arabic numerals. For example, the calculation

$$\begin{array}{r} 13 \\ \times 24 \\ \hline 52 \\ 26 \\ \hline 312 \end{array}$$

is a concise arrangement for the following equations:

$$\begin{aligned} 13 \cdot 24 &= 13 \cdot (2 \cdot 10 + 4) \\ &= 13 \cdot 2 \cdot 10 + 13 \cdot 4 \\ &= 26 \cdot 10 + 52. \end{aligned}$$

(Note that moving 26 to the left in the above calculation is the same as writing  $26 \cdot 10$ .) The multiplication  $13 \cdot 4 = 52$  uses P9 also:

$$\begin{aligned} 13 \cdot 4 &= (1 \cdot 10 + 3) \cdot 4 \\ &= 1 \cdot 10 \cdot 4 + 3 \cdot 4 \\ &= 4 \cdot 10 + 12 \\ &= 4 \cdot 10 + 1 \cdot 10 + 2 \\ &= (4 + 1) \cdot 10 + 2 \\ &= 5 \cdot 10 + 2 \\ &= 52. \end{aligned}$$

The properties P1–P9 have descriptive names which are not essential to remember, but which are often convenient for reference. We will take this opportunity to list properties P1–P9 together and indicate the names by which they are commonly designated.

|      |  |   |
|------|--|---|
| (P1) | (Associative law for addition)           | $a + (b + c) = (a + b) + c.$                                  |
| (P2) | (Existence of an additive identity)      | $a + 0 = 0 + a = a.$  |
| (P3) | (Existence of additive inverses)         | $a + (-a) = (-a) + a = 0.$                                    |
| (P4) | (Commutative law for addition)           | $a + b = b + a.$  |
| (P5) | (Associative law for multiplication)     | $a \cdot (b \cdot c) = (a \cdot b) \cdot c.$                  |
| (P6) | (Existence of a multiplicative identity) | $a \cdot 1 = 1 \cdot a = a; \quad 1 \neq 0.$                  |
| (P7) | (Existence of multiplicative inverses)   | $a \cdot a^{-1} = a^{-1} \cdot a = 1, \text{ for } a \neq 0.$ |
| (P8) | (Commutative law for multiplication)     | $a \cdot b = b \cdot a.$                                      |
| (P9) | (Distributive law)                       | $a \cdot (b + c) = a \cdot b + a \cdot c.$                    |

The three basic properties of numbers which remain to be listed are concerned with inequalities. Although inequalities occur rarely in elementary mathematics, they play a prominent role in calculus. The two notions of inequality,  $a < b$  ( $a$  is less than  $b$ ) and  $a > b$  ( $a$  is greater than  $b$ ), are intimately related:  $a < b$  means the same as  $b > a$  (thus  $1 < 3$  and  $3 > 1$  are merely two ways of writing the same assertion). The numbers  $a$  satisfying  $a > 0$  are called **positive**, while those numbers  $a$  satisfying  $a < 0$  are called **negative**. While positivity can thus be defined in terms of  $<$ , it is possible to reverse the procedure:  $a < b$  can be defined to mean that  $b - a$  is positive. In fact, it is convenient to consider the collection of all positive numbers, denoted by  $P$ , as the basic concept, and state all properties in terms of  $P$ :

- (P10) (Trichotomy law) For every number  $a$ , one and only one of the following holds:

- (i)  $a = 0,$
- (ii)  $a$  is in the collection  $P,$
- (iii)  $-a$  is in the collection  $P.$

- (P11) (Closure under addition) If  $a$  and  $b$  are in  $P$ , then  $a + b$  is in  $P.$

- (P12) (Closure under multiplication) If  $a$  and  $b$  are in  $P$ , then  $a \cdot b$  is in  $P.$

## 10 Prologue

These three properties should be complemented with the following definitions:

$$\begin{aligned} a > b &\quad \text{if } a - b \text{ is in } P; \\ a < b &\quad \text{if } b > a; \\ a \geq b &\quad \text{if } a > b \text{ or } a = b; \\ a \leq b &\quad \text{if } a < b \text{ or } a = b.* \end{aligned}$$

Note, in particular, that  $a > 0$  if and only if  $a$  is in  $P$ .

All the familiar facts about inequalities, however elementary they may seem, are consequences of P10–P12. For example, if  $a$  and  $b$  are any two numbers, then precisely one of the following holds:

- (i)  $a - b = 0$ ,
- (ii)  $a - b$  is in the collection  $P$ ,
- (iii)  $-(a - b) = b - a$  is in the collection  $P$ .

Using the definitions just made, it follows that precisely one of the following holds:

- (i)  $a = b$ ,
- (ii)  $a > b$ ,
- (iii)  $b > a$ .

A slightly more interesting fact results from the following manipulations. If  $a < b$ , so that  $b - a$  is in  $P$ , then surely  $(b + c) - (a + c)$  is in  $P$ ; thus, if  $a < b$ , then  $a + c < b + c$ . Similarly, suppose  $a < b$  and  $b < c$ . Then

$$\begin{aligned} &b - a \text{ is in } P, \\ \text{and } &c - b \text{ is in } P, \\ \text{so } &c - a = (c - b) + (b - a) \text{ is in } P. \end{aligned}$$

This shows that if  $a < b$  and  $b < c$ , then  $a < c$ . (The two inequalities  $a < b$  and  $b < c$  are usually written in the abbreviated form  $a < b < c$ , which has the third inequality  $a < c$  almost built in.)

The following assertion is somewhat less obvious: If  $a < 0$  and  $b < 0$ , then  $ab > 0$ . The only difficulty presented by the proof is the unraveling of definitions. The symbol  $a < 0$  means, by definition,  $0 > a$ , which means  $0 - a = -a$  is in  $P$ . Similarly  $-b$  is in  $P$ , and consequently, by P12,  $(-a)(-b) = ab$  is in  $P$ . Thus  $ab > 0$ .

The fact that  $ab > 0$  if  $a > 0$ ,  $b > 0$  and also if  $a < 0$ ,  $b < 0$  has one special consequence:  $a^2 > 0$  if  $a \neq 0$ . Thus squares of nonzero numbers are always positive, and in particular we have proved a result which might have seemed sufficiently elementary to be included in our list of properties:  $1 > 0$  (since  $1 = 1^2$ ).

\*There is one slightly perplexing feature of the symbols  $\geq$  and  $\leq$ . The statements

$$\begin{aligned} 1 + 1 &\leq 3 \\ 1 + 1 &\leq 2 \end{aligned}$$

are both true, even though we know that  $\leq$  could be replaced by  $<$  in the first, and by  $=$  in the second. This sort of thing is bound to occur when  $\leq$  is used with specific numbers; the usefulness of the symbol is revealed by a statement like Theorem 1—here equality holds for some values of  $a$  and  $b$ , while inequality holds for other values.

The fact that  $-a > 0$  if  $a < 0$  is the basis of a concept which will play an extremely important role in this book. For any number  $a$ , we define the **absolute value**  $|a|$  of  $a$  as follows:

$$|a| = \begin{cases} a, & a \geq 0 \\ -a, & a \leq 0. \end{cases}$$

Note that  $|a|$  is always positive, except when  $a = 0$ . For example, we have  $|-3| = 3$ ,  $|7| = 7$ ,  $|1 + \sqrt{2} - \sqrt{3}| = 1 + \sqrt{2} - \sqrt{3}$ , and  $|1 + \sqrt{2} - \sqrt{10}| = \sqrt{10} - \sqrt{2} - 1$ . In general, the most straightforward approach to any problem involving absolute values requires treating several cases separately, since absolute values are defined by cases to begin with. This approach may be used to prove the following very important fact about absolute values.

**THEOREM 1** For all numbers  $a$  and  $b$ , we have

$$|a + b| \leq |a| + |b|.$$

**PROOF** We will consider 4 cases:

- (1)  $a \geq 0, b \geq 0$ ;
- (2)  $a \geq 0, b \leq 0$ ;
- (3)  $a \leq 0, b \geq 0$ ;
- (4)  $a \leq 0, b \leq 0$ .

In case (1) we also have  $a + b \geq 0$ , and the theorem is obvious; in fact,

$$|a + b| = a + b = |a| + |b|,$$

so that in this case equality holds.

In case (4) we have  $a + b \leq 0$ , and again equality holds:

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

In case (2), when  $a \geq 0$  and  $b \leq 0$ , we must prove that

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

This case may therefore be divided into two subcases. If  $a + b \geq 0$ , then we must prove that

$$\begin{aligned} a + b &\leq a - b, \\ \text{i.e.,} \quad b &\leq -b, \end{aligned}$$

which is certainly true since  $b \leq 0$  and hence  $-b \geq 0$ . On the other hand, if  $a + b \leq 0$ , we must prove that

$$\begin{aligned} -a - b &\leq a - b, \\ \text{i.e.,} \quad -a &\leq a, \end{aligned}$$

which is certainly true since  $a \geq 0$  and hence  $-a \leq 0$ .

Finally, note that case (3) may be disposed of with no additional work, by applying case (2) with  $a$  and  $b$  interchanged. ■

Although this method of treating absolute values (separate consideration of various cases) is sometimes the only approach available, there are often simpler methods which may be used. In fact, it is possible to give a much shorter proof of Theorem 1; this proof is motivated by the observation that

$$|a| = \sqrt{a^2}.$$

(Here, and throughout the book,  $\sqrt{x}$  denotes the *positive* square root of  $x$ ; this symbol is defined only when  $x \geq 0$ .) We may now observe that

$$\begin{aligned} (|a + b|)^2 &= (a + b)^2 = a^2 + 2ab + b^2 \\ &\leq a^2 + 2|a| \cdot |b| + b^2 \\ &= |a|^2 + 2|a| \cdot |b| + |b|^2 \\ &= (|a| + |b|)^2. \end{aligned}$$

From this we can conclude that  $|a + b| \leq |a| + |b|$  because  $x^2 < y^2$  implies  $x < y$ , provided that  $x$  and  $y$  are both nonnegative; a proof of *this* fact is left to the reader (Problem 5).

One final observation may be made about the theorem we have just proved: a close examination of either proof offered shows that

$$|a + b| = |a| + |b|$$

if  $a$  and  $b$  have the same sign (i.e., are both positive or both negative), or if one of the two is 0, while

$$|a + b| < |a| + |b|$$

if  $a$  and  $b$  are of opposite signs.

We will conclude this chapter with a subtle point, neglected until now, whose inclusion is required in a conscientious survey of the properties of numbers. After stating property P9, we proved that  $a - b = b - a$  implies  $a = b$ . The proof began by establishing that

$$a \cdot (1 + 1) = b \cdot (1 + 1),$$

from which we concluded that  $a = b$ . This result is obtained from the equation  $a \cdot (1 + 1) = b \cdot (1 + 1)$  by dividing both sides by  $1 + 1$ . Division by 0 should be avoided scrupulously, and it must therefore be admitted that the validity of the argument depends on knowing that  $1 + 1 \neq 0$ . Problem 25 is designed to convince you that this fact cannot possibly be proved from properties P1–P9 alone! Once P10, P11, and P12 are available, however, the proof is very simple: We have already seen that  $1 > 0$ ; it follows that  $1 + 1 > 0$ , and in particular  $1 + 1 \neq 0$ .

This last demonstration has perhaps only strengthened your feeling that it is absurd to bother proving such obvious facts, but an honest assessment of our present situation will help justify serious consideration of such details. In this chapter we have assumed that numbers are familiar objects, and that P1–P12 are merely explicit statements of obvious, well-known properties of numbers. It would be difficult, however, to justify this assumption. Although one learns how to “work with” numbers in school, just what numbers *are*, remains rather vague. A great deal of this book is devoted to elucidating the concept of numbers, and by the end

of the book we will have become quite well acquainted with them. But it will be necessary to work with numbers throughout the book. It is therefore reasonable to admit frankly that we do not yet thoroughly understand numbers; we may still say that, in whatever way numbers are finally defined, they should certainly have properties P1–P12.

Most of this chapter has been an attempt to present convincing evidence that P1–P12 are indeed basic properties which we should assume in order to deduce other familiar properties of numbers. Some of the problems (which indicate the derivation of other facts about numbers from P1–P12) are offered as further evidence. It is still a crucial question whether P1–P12 actually account for *all* properties of numbers. As a matter of fact, we shall soon see that they do *not*. In the next chapter the deficiencies of properties P1–P12 will become quite clear, but the proper means for correcting these deficiencies is not so easily discovered. The crucial additional basic property of numbers which we are seeking is profound and subtle, quite unlike P1–P12. The discovery of this crucial property will require all the work of Part II of this book. In the remainder of Part I we will begin to see why some additional property is required; in order to investigate this we will have to consider a little more carefully what we mean by “numbers.”

## PROBLEMS

1. Prove the following:

- (i) If  $ax = a$  for some number  $a \neq 0$ , then  $x = 1$ .
- (ii)  $x^2 - y^2 = (x - y)(x + y)$ .
- (iii) If  $x^2 = y^2$ , then  $x = y$  or  $x = -y$ .
- (iv)  $x^3 - y^3 = (x - y)(x^2 + xy + y^2)$ .
- (v)  $x^n - y^n = (x - y)(x^{n-1} + x^{n-2}y + \dots + xy^{n-2} + y^{n-1})$ .
- (vi)  $x^3 + y^3 = (x + y)(x^2 - xy + y^2)$ . (There is a particularly easy way to do this, using (iv), and it will show you how to find a factorization for  $x^n + y^n$  whenever  $n$  is odd.)

2. What is wrong with the following “proof”? Let  $x = y$ . Then

$$\begin{aligned} x^2 &= xy, \\ x^2 - y^2 &= xy - y^2, \\ (x + y)(x - y) &= y(x - y), \\ x + y &= y, \\ 2y &= y, \\ 2 &= 1. \end{aligned}$$

3. Prove the following:

- (i)  $\frac{a}{b} = \frac{ac}{bc}$ , if  $b, c \neq 0$ .
- (ii)  $\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$ , if  $b, d \neq 0$ .

(iii)  $(ab)^{-1} = a^{-1}b^{-1}$ , if  $a, b \neq 0$ . (To do this you must remember the defining property of  $(ab)^{-1}$ .)

(iv)  $\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{db}$ , if  $b, d \neq 0$ .

(v)  $\frac{a}{b} \Big/ \frac{c}{d} = \frac{ad}{bc}$ , if  $b, c, d \neq 0$ .

(vi) If  $b, d \neq 0$ , then  $\frac{a}{b} = \frac{c}{d}$  if and only if  $ad = bc$ . Also determine when  $\frac{a}{b} = \frac{b}{a}$ .

4. Find all numbers  $x$  for which

(i)  $4 - x < 3 - 2x$ .

(ii)  $5 - x^2 < 8$ .

(iii)  $5 - x^2 < -2$ .

(iv)  $(x - 1)(x - 3) > 0$ . (When is a product of two numbers positive?)

(v)  $x^2 - 2x + 2 > 0$ .

(vi)  $x^2 + x + 1 > 2$ .

(vii)  $x^2 - x + 10 > 16$ .

(viii)  $x^2 + x + 1 > 0$ .

(ix)  $(x - \pi)(x + 5)(x - 3) > 0$ .

(x)  $(x - \sqrt[3]{2})(x - \sqrt{2}) > 0$ .

(xi)  $2^x < 8$ .

(xii)  $x + 3^x < 4$ .

(xiii)  $\frac{1}{x} + \frac{1}{1-x} > 0$ .

(xiv)  $\frac{x-1}{x+1} > 0$ .

5. Prove the following:

(i) If  $a < b$  and  $c < d$ , then  $a + c < b + d$ .

(ii) If  $a < b$ , then  $-b < -a$ .

(iii) If  $a < b$  and  $c > d$ , then  $a - c < b - d$ .

(iv) If  $a < b$  and  $c > 0$ , then  $ac < bc$ .

(v) If  $a < b$  and  $c < 0$ , then  $ac > bc$ .

(vi) If  $a > 1$ , then  $a^2 > a$ .

(vii) If  $0 < a < 1$ , then  $a^2 < a$ .

(viii) If  $0 \leq a < b$  and  $0 \leq c < d$ , then  $ac < bd$ .

(ix) If  $0 \leq a < b$ , then  $a^2 < b^2$ . (Use (viii).)

(x) If  $a, b \geq 0$  and  $a^2 < b^2$ , then  $a < b$ . (Use (ix), backwards.)

6. (a) Prove that if  $0 \leq x < y$ , then  $x^n < y^n$ ,  $n = 1, 2, 3, \dots$ .

(b) Prove that if  $x < y$  and  $n$  is odd, then  $x^n < y^n$ .

(c) Prove that if  $x^n = y^n$  and  $n$  is odd, then  $x = y$ .

(d) Prove that if  $x^n = y^n$  and  $n$  is even, then  $x = y$  or  $x = -y$ .

7. Prove that if  $0 < a < b$ , then

$$a < \sqrt{ab} < \frac{a+b}{2} < b.$$

Notice that the inequality  $\sqrt{ab} \leq (a+b)/2$  holds for all  $a, b \geq 0$ . A generalization of this fact occurs in Problem 2-22.

- \*8. Although the basic properties of inequalities were stated in terms of the collection  $P$  of all positive numbers, and  $<$  was defined in terms of  $P$ , this procedure can be reversed. Suppose that P10–P12 are replaced by

- (P'10) For any numbers  $a$  and  $b$  one, and only one, of the following holds:

- (i)  $a = b$ ,
- (ii)  $a < b$ ,
- (iii)  $b < a$ .

- (P'11) For any numbers  $a, b$ , and  $c$ , if  $a < b$  and  $b < c$ , then  
 $a < c$ .

- (P'12) For any numbers  $a, b$ , and  $c$ , if  $a < b$ , then  
 $a + c < b + c$ .

- (P'13) For any numbers  $a, b$ , and  $c$ , if  $a < b$  and  $0 < c$ , then  
 $ac < bc$ .

Show that P10–P12 can then be deduced as theorems.

9. Express each of the following with at least one less pair of absolute value signs.

- (i)  $|\sqrt{2} + \sqrt{3} - \sqrt{5} + \sqrt{7}|$ .
- (ii)  $|(|a+b| - |a| - |b|)|$ .
- (iii)  $|(|a+b| + |c| - |a+b+c|)|$ .
- (iv)  $|x^2 - 2xy + y^2|$ .
- (v)  $|(\sqrt{2} + \sqrt{3}) - (\sqrt{5} - \sqrt{7})|$ .

10. Express each of the following without absolute value signs, treating various cases separately when necessary.

- (i)  $|a+b| - |b|$ .
- (ii)  $|(|x| - 1)|$ .
- (iii)  $|x| - |x^2|$ .
- (iv)  $a - |(a - |a|)|$ .

11. Find all numbers  $x$  for which

- (i)  $|x - 3| = 8$ .
- (ii)  $|x - 3| < 8$ .
- (iii)  $|x + 4| < 2$ .
- (iv)  $|x - 1| + |x - 2| > 1$ .
- (v)  $|x - 1| + |x + 1| < 2$ .

- (vi)  $|x - 1| + |x + 1| < 1$ .  
 (vii)  $|x - 1| \cdot |x + 1| = 0$ .  
 (viii)  $|x - 1| \cdot |x + 2| = 3$ .

**12.** Prove the following:

- (i)  $|xy| = |x| \cdot |y|$ .
- (ii)  $\left| \frac{1}{x} \right| = \frac{1}{|x|}$ , if  $x \neq 0$ . (The best way to do this is to remember what  $|x|^{-1}$  is.)
- (iii)  $\frac{|x|}{|y|} = \left| \frac{x}{y} \right|$ , if  $y \neq 0$ .
- (iv)  $|x - y| \leq |x| + |y|$ . (Give a very short proof.)
- (v)  $|x| - |y| \leq |x - y|$ . (A very short proof is possible, if you write things in the right way.)
- (vi)  $|(|x| - |y|)| \leq |x - y|$ . (Why does this follow immediately from (v)?)
- (vii)  $|x + y + z| \leq |x| + |y| + |z|$ . Indicate when equality holds, and prove your statement.

**13.** The maximum of two numbers  $x$  and  $y$  is denoted by  $\max(x, y)$ . Thus  $\max(-1, 3) = \max(3, 3) = 3$  and  $\max(-1, -4) = \max(-4, -1) = -1$ . The minimum of  $x$  and  $y$  is denoted by  $\min(x, y)$ . Prove that

$$\begin{aligned}\max(x, y) &= \frac{x + y + |y - x|}{2}, \\ \min(x, y) &= \frac{x + y - |y - x|}{2}.\end{aligned}$$

Derive a formula for  $\max(x, y, z)$  and  $\min(x, y, z)$ , using, for example

$$\max(x, y, z) = \max(x, \max(y, z)).$$

- 14.** (a) Prove that  $|a| = |-a|$ . (The trick is not to become confused by too many cases. First prove the statement for  $a \geq 0$ . Why is it then obvious for  $a \leq 0$ ?)  
 (b) Prove that  $-b \leq a \leq b$  if and only if  $|a| \leq b$ . In particular, it follows that  $-|a| \leq a \leq |a|$ .  
 (c) Use this fact to give a new proof that  $|a + b| \leq |a| + |b|$ .

**\*15.** Prove that if  $x$  and  $y$  are not both 0, then

$$\begin{aligned}x^2 + xy + y^2 &> 0, \\ x^4 + x^3y + x^2y^2 + xy^3 + y^4 &> 0.\end{aligned}$$

Hint: Use Problem 1.

**\*16.** (a) Show that

$$\begin{aligned}(x + y)^2 &= x^2 + y^2 \quad \text{only when } x = 0 \text{ or } y = 0, \\ (x + y)^3 &= x^3 + y^3 \quad \text{only when } x = 0 \text{ or } y = 0 \text{ or } x = -y.\end{aligned}$$

- (b) Using the fact that

$$x^2 + 2xy + y^2 = (x + y)^2 \geq 0,$$

show that  $4x^2 + 6xy + 4y^2 > 0$  unless  $x$  and  $y$  are both 0.

- (c) Use part (b) to find out when  $(x + y)^4 = x^4 + y^4$ .
- (d) Find out when  $(x + y)^5 = x^5 + y^5$ . Hint: From the assumption  $(x + y)^5 = x^5 + y^5$  you should be able to derive the equation  $x^3 + 2x^2y + 2xy^2 + y^3 = 0$ , if  $xy \neq 0$ . This implies that  $(x + y)^3 = x^2y + xy^2 = xy(x + y)$ .

You should now be able to make a good guess as to when  $(x + y)^n = x^n + y^n$ ; the proof is contained in Problem 11-57.

17. (a) Find the smallest possible value of  $2x^2 - 3x + 4$ . Hint: "Complete the square," i.e., write  $2x^2 - 3x + 4 = 2(x - 3/4)^2 + ?$   
 (b) Find the smallest possible value of  $x^2 - 3x + 2y^2 + 4y + 2$ .  
 (c) Find the smallest possible value of  $x^2 + 4xy + 5y^2 - 4x - 6y + 7$ .

18. (a) Suppose that  $b^2 - 4c \geq 0$ . Show that the numbers

$$\frac{-b + \sqrt{b^2 - 4c}}{2}, \quad \frac{-b - \sqrt{b^2 - 4c}}{2}$$

both satisfy the equation  $x^2 + bx + c = 0$ .

- (b) Suppose that  $b^2 - 4c < 0$ . Show that there are no numbers  $x$  satisfying  $x^2 + bx + c = 0$ ; in fact,  $x^2 + bx + c > 0$  for all  $x$ . Hint: Complete the square.
- (c) Use this fact to give another proof that if  $x$  and  $y$  are not both 0, then  $x^2 + xy + y^2 > 0$ .
- (d) For which numbers  $\alpha$  is it true that  $x^2 + \alpha xy + y^2 > 0$  whenever  $x$  and  $y$  are not both 0?
- (e) Find the smallest possible value of  $x^2 + bx + c$  and of  $ax^2 + bx + c$ , for  $a > 0$ .

19. The fact that  $a^2 \geq 0$  for all numbers  $a$ , elementary as it may seem, is nevertheless the fundamental idea upon which most important inequalities are ultimately based. The great-granddaddy of all inequalities is the *Schwarz inequality*:

$$x_1y_1 + x_2y_2 \leq \sqrt{x_1^2 + x_2^2} \sqrt{y_1^2 + y_2^2}.$$

(A more general form occurs in Problem 2-21.) The three proofs of the Schwarz inequality outlined below have only one thing in common—their reliance on the fact that  $a^2 \geq 0$  for all  $a$ .

- (a) Prove that if  $x_1 = \lambda y_1$  and  $x_2 = \lambda y_2$  for some number  $\lambda$ , then equality holds in the Schwarz inequality. Prove the same thing if  $y_1 = y_2 = 0$ . Now suppose that  $y_1$  and  $y_2$  are not both 0, and that there is no number

$\lambda$  such that  $x_1 = \lambda y_1$  and  $x_2 = \lambda y_2$ . Then

$$\begin{aligned} 0 &< (\lambda y_1 - x_1)^2 + (\lambda y_2 - x_2)^2 \\ &= \lambda^2(y_1^2 + y_2^2) - 2\lambda(x_1 y_1 + x_2 y_2) + (x_1^2 + x_2^2). \end{aligned}$$

Using Problem 18, complete the proof of the Schwarz inequality.

- (b) Prove the Schwarz inequality by using  $2xy \leq x^2 + y^2$  (how is this derived?) with

$$x = \frac{x_i}{\sqrt{x_1^2 + x_2^2}}, \quad y = \frac{y_i}{\sqrt{y_1^2 + y_2^2}},$$

first for  $i = 1$  and then for  $i = 2$ .

- (c) Prove the Schwarz inequality by first proving that

$$(x_1^2 + x_2^2)(y_1^2 + y_2^2) = (x_1 y_1 + x_2 y_2)^2 + (x_1 y_2 - x_2 y_1)^2.$$

- (d) Deduce, from each of these three proofs, that equality holds only when  $y_1 = y_2 = 0$  or when there is a number  $\lambda$  such that  $x_1 = \lambda y_1$  and  $x_2 = \lambda y_2$ .

In our later work, three facts about inequalities will be crucial. Although proofs will be supplied at the appropriate point in the text, a personal assault on these problems is infinitely more enlightening than a perusal of a completely worked-out proof. The statements of these propositions involve some weird numbers, but their basic message is very simple: if  $x$  is close enough to  $x_0$ , and  $y$  is close enough to  $y_0$ , then  $x+y$  will be close to  $x_0+y_0$ , and  $xy$  will be close to  $x_0y_0$ , and  $1/y$  will be close to  $1/y_0$ . The symbol “ $\varepsilon$ ” which appears in these propositions is the fifth letter of the Greek alphabet (“epsilon”), and could just as well be replaced by a less intimidating Roman letter; however, tradition has made the use of  $\varepsilon$  almost sacrosanct in the contexts to which these theorems apply.

- 20.** Prove that if

$$|x - x_0| < \frac{\varepsilon}{2} \quad \text{and} \quad |y - y_0| < \frac{\varepsilon}{2},$$

then

$$\begin{aligned} |(x + y) - (x_0 + y_0)| &< \varepsilon, \\ |(x - y) - (x_0 - y_0)| &< \varepsilon. \end{aligned}$$

- \*21.** Prove that if

$$|x - x_0| < \min\left(\frac{\varepsilon}{2(|y_0| + 1)}, 1\right) \quad \text{and} \quad |y - y_0| < \frac{\varepsilon}{2(|x_0| + 1)},$$

then  $|xy - x_0y_0| < \varepsilon$ .

(The notation “min” was defined in Problem 13, but the formula provided by that problem is irrelevant at the moment; the first inequality in the hypothesis just means that

$$|x - x_0| < \frac{\varepsilon}{2(|y_0| + 1)} \quad \text{and} \quad |x - x_0| < 1;$$

at one point in the argument you will need the first inequality, and at another point you will need the second. One more word of advice: since the hypotheses only provide information about  $x - x_0$  and  $y - y_0$ , it is almost a foregone conclusion that the proof will depend upon writing  $xy - x_0y_0$  in a way that involves  $x - x_0$  and  $y - y_0$ .)

- \*22. Prove that if  $y_0 \neq 0$  and

$$|y - y_0| < \min\left(\frac{|y_0|}{2}, \frac{\varepsilon|y_0|^2}{2}\right),$$

then  $y \neq 0$  and

$$\left|\frac{1}{y} - \frac{1}{y_0}\right| < \varepsilon.$$

- \*23. Replace the question marks in the following statement by expressions involving  $\varepsilon$ ,  $x_0$ , and  $y_0$  so that the conclusion will be true:

If  $y_0 \neq 0$  and

$$|y - y_0| < ? \quad \text{and} \quad |x - x_0| < ?$$

then  $y \neq 0$  and

$$\left|\frac{x}{y} - \frac{x_0}{y_0}\right| < \varepsilon.$$

This problem is trivial in the sense that its solution follows from Problems 21 and 22 with almost no work at all (notice that  $x/y = x \cdot 1/y$ ). The crucial point is not to become confused; decide which of the two problems should be used first, and don't panic if your answer looks unlikely.

- \*24. This problem shows that the actual placement of parentheses in a sum is irrelevant. The proofs involve “mathematical induction”; if you are not familiar with such proofs, but still want to tackle this problem, it can be saved until after Chapter 2, where proofs by induction are explained.

Let us agree, for definiteness, that  $a_1 + \cdots + a_n$  will denote

$$a_1 + (a_2 + (a_3 + \cdots + (a_{n-2} + (a_{n-1} + a_n)))) \cdots ).$$

Thus  $a_1 + a_2 + a_3$  denotes  $a_1 + (a_2 + a_3)$ , and  $a_1 + a_2 + a_3 + a_4$  denotes  $a_1 + (a_2 + (a_3 + a_4))$ , etc.

- (a) Prove that

$$(a_1 + \cdots + a_k) + a_{k+1} = a_1 + \cdots + a_{k+1}.$$

Hint: Use induction on  $k$ .

- (b) Prove that if  $n \geq k$ , then

$$(a_1 + \cdots + a_k) + (a_{k+1} + \cdots + a_n) = a_1 + \cdots + a_n.$$

Hint: Use part (a) to give a proof by induction on  $k$ .

(c) Let  $s(a_1, \dots, a_k)$  be some sum formed from  $a_1, \dots, a_k$ . Show that

$$s(a_1, \dots, a_k) = a_1 + \dots + a_k.$$

Hint: There must be two sums  $s'(a_1, \dots, a_l)$  and  $s''(a_{l+1}, \dots, a_k)$  such that

$$s(a_1, \dots, a_k) = s'(a_1, \dots, a_l) + s''(a_{l+1}, \dots, a_k).$$

25. Suppose that we interpret “number” to mean either 0 or 1, and  $+$  and  $\cdot$  to be the operations defined by the following two tables.

| $+$ | 0 | 1 | $\cdot$ | 0 | 1 |
|-----|---|---|---------|---|---|
| 0   | 0 | 1 | 0       | 0 | 0 |
| 1   | 1 | 0 | 1       | 0 | 1 |

Check that properties P1–P9 all hold, even though  $1 + 1 = 0$ .

# CHAPTER 2 NUMBERS OF VARIOUS SORTS

In Chapter 1 we used the word “number” very loosely, despite our concern with the basic properties of numbers. It will now be necessary to distinguish carefully various kinds of numbers.

The simplest numbers are the “counting numbers”

$$1, 2, 3, \dots$$

The fundamental significance of this collection of numbers is emphasized by its symbol **N** (for **natural numbers**). A brief glance at P1–P12 will show that our basic properties of “numbers” do not apply to **N**—for example, P2 and P3 do not make sense for **N**. From this point of view the system **N** has many deficiencies. Nevertheless, **N** is sufficiently important to deserve several comments before we consider larger collections of numbers.

The most basic property of **N** is the principle of “mathematical induction.” Suppose  $P(x)$  means that the property  $P$  holds for the number  $x$ . Then the principle of mathematical induction states that  $P(x)$  is true for all natural numbers  $x$  provided that

- (1)  $P(1)$  is true.
- (2) Whenever  $P(k)$  is true,  $P(k + 1)$  is true.

Note that condition (2) merely asserts the truth of  $P(k + 1)$  under the assumption that  $P(k)$  is true; this suffices to ensure the truth of  $P(x)$  for all  $x$ , if condition (1) also holds. In fact, if  $P(1)$  is true, then it follows that  $P(2)$  is true (by using (2) in the special case  $k = 1$ ). Now, since  $P(2)$  is true it follows that  $P(3)$  is true (using (2) in the special case  $k = 2$ ). It is clear that each number will eventually be reached by a series of steps of this sort, so that  $P(k)$  is true for all numbers  $k$ .

A favorite illustration of the reasoning behind mathematical induction envisions an infinite line of people,

person number 1, person number 2, person number 3,  $\dots$ .

If each person has been instructed to tell any secret he hears to the person behind him (the one with the next largest number) and a secret is told to person number 1, then clearly every person will eventually learn the secret. If  $P(x)$  is the assertion that person number  $x$  will learn the secret, then the instructions given (to tell all secrets learned to the next person) assures that condition (2) is true, and telling the secret to person number 1 makes (1) true. The following example is a less facetious use of mathematical induction. There is a useful and striking formula which expresses the sum of the first  $n$  numbers in a simple way:

$$1 + \dots + n = \frac{n(n+1)}{2}.$$

To prove this formula, note first that it is clearly true for  $n = 1$ . Now *assume* that for some natural number  $k$  we have

$$1 + \dots + k = \frac{k(k+1)}{2}.$$

Then

$$\begin{aligned} 1 + \dots + k + (k+1) &= \frac{k(k+1)}{2} + k + 1 \\ &= \frac{k(k+1) + 2k + 2}{2} \\ &= \frac{k^2 + 3k + 2}{2} \\ &= \frac{(k+1)(k+2)}{2}, \end{aligned}$$

so the formula is also true for  $k + 1$ . By the principle of induction this proves the formula for all natural numbers  $n$ . This particular example illustrates a phenomenon that frequently occurs, especially in connection with formulas like the one just proved. Although the proof by induction is often quite straightforward, the method by which the formula was discovered remains a mystery. Problems 5 and 6 indicate how some formulas of this type may be derived.

The principle of mathematical induction may be formulated in an equivalent way without speaking of “properties” of a number, a term which is sufficiently vague to be eschewed in a mathematical discussion. A more precise formulation states that if  $A$  is any collection (or “set”—a synonymous mathematical term) of natural numbers and

- (1)  $1$  is in  $A$ ,
- (2)  $k + 1$  is in  $A$  whenever  $k$  is in  $A$ ,

then  $A$  is the set of all natural numbers. It should be clear that this formulation adequately replaces the less formal one given previously—we just consider the set  $A$  of natural numbers  $x$  which satisfy  $P(x)$ . For example, suppose  $A$  is the set of natural numbers  $n$  for which it is true that

$$1 + \dots + n = \frac{n(n+1)}{2}.$$

Our previous proof of this formula showed that  $A$  contains  $1$ , and that  $k + 1$  is in  $A$ , if  $k$  is. It follows that  $A$  is the set of all natural numbers, i.e., that the formula holds for all natural numbers  $n$ .

There is yet another rigorous formulation of the principle of mathematical induction, which looks quite different. If  $A$  is any collection of natural numbers, it

is tempting to say that  $A$  must have a smallest member. Actually, this statement can fail to be true in a rather subtle way. A particularly important set of natural numbers is the collection  $A$  that contains no natural numbers at all, the “empty collection” or “null set,”\* denoted by  $\emptyset$ . The null set  $\emptyset$  is a collection of natural numbers that has no smallest member—in fact, it has no members at all. This is the only possible exception, however; if  $A$  is a nonnull set of natural numbers, then  $A$  has a least member. This “intuitively obvious” statement, known as the “well-ordering principle,” can be proved from the principle of induction as follows. Suppose that the set  $A$  has no least member. Let  $B$  be the set of natural numbers  $n$  such that  $1, \dots, n$  are all *not* in  $A$ . Clearly  $1$  is in  $B$  (because if  $1$  were in  $A$ , then  $A$  would have  $1$  as smallest member). Moreover, if  $1, \dots, k$  are not in  $A$ , surely  $k + 1$  is not in  $A$  (otherwise  $k + 1$  would be the smallest member of  $A$ ), so  $1, \dots, k + 1$  are all not in  $A$ . This shows that if  $k$  is in  $B$ , then  $k + 1$  is in  $B$ . It follows that every number  $n$  is in  $B$ , i.e., the numbers  $1, \dots, n$  are *not* in  $A$  for any natural number  $n$ . Thus  $A = \emptyset$ , which completes the proof.

It is also possible to prove the principle of induction from the well-ordering principle (Problem 10). Either principle may be considered as a basic assumption about the natural numbers.

There is still another form of induction which should be mentioned. It sometimes happens that in order to prove  $P(k + 1)$  we must assume not only  $P(k)$ , but also  $P(l)$  for all natural numbers  $l \leq k$ . In this case we rely on the “principle of complete induction”: If  $A$  is a set of natural numbers and

- (1)  $1$  is in  $A$ ,
- (2)  $k + 1$  is in  $A$  if  $1, \dots, k$  are in  $A$ ,

then  $A$  is the set of all natural numbers.

Although the principle of complete induction may appear much stronger than the ordinary principle of induction, it is actually a consequence of that principle. The proof of this fact is left to the reader, with a hint (Problem 11). Applications will be found in Problems 7, 17, 20 and 22.

Closely related to proofs by induction are “recursive definitions.” For example, the number  $n!$  (read “ $n$  factorial”) is defined as the product of all the natural numbers less than or equal to  $n$ :

$$n! = 1 \cdot 2 \cdot \dots \cdot (n - 1) \cdot n.$$

This can be expressed more precisely as follows:

- (1)  $1! = 1$
- (2)  $n! = n \cdot (n - 1)!$ .

This form of the definition exhibits the relationship between  $n!$  and  $(n - 1)!$  in an

\*Although it may not strike you as a collection, in the ordinary sense of the word, the null set arises quite naturally in many contexts. We frequently consider the set  $A$ , consisting of all  $x$  satisfying some property  $P$ ; often we have no guarantee that  $P$  is satisfied by *any* number, so that  $A$  might be  $\emptyset$ —in fact often one proves that  $P$  is always false by showing that  $A = \emptyset$ .

explicit way that is ideally suited for proofs by induction. Problem 23 reviews a definition already familiar to you, which may be expressed more succinctly as a recursive definition; as this problem shows, the recursive definition is really necessary for a rigorous proof of some of the basic properties of the definition.

One definition which may not be familiar involves some convenient notation which we will constantly be using. Instead of writing

$$a_1 + \cdots + a_n,$$

we will usually employ the Greek letter  $\Sigma$  (capital sigma, for “sum”) and write

$$\sum_{i=1}^n a_i.$$

In other words,  $\sum_{i=1}^n a_i$  denotes the sum of the numbers obtained by letting  $i = 1, 2, \dots, n$ . Thus

$$\sum_{i=1}^n i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Notice that the letter  $i$  really has nothing to do with the number denoted by  $\sum_{i=1}^n i$ , and can be replaced by any convenient symbol (except  $n$ , of course!):

$$\begin{aligned}\sum_{j=1}^n j &= \frac{n(n+1)}{2}, \\ \sum_{j=1}^i j &= \frac{i(i+1)}{2}, \\ \sum_{n=1}^j n &= \frac{j(j+1)}{2}.\end{aligned}$$

To define  $\sum_{i=1}^n a_i$  precisely really requires a recursive definition:

$$\begin{aligned}(1) \quad \sum_{i=1}^1 a_i &= a_1, \\ (2) \quad \sum_{i=1}^n a_i &= \sum_{i=1}^{n-1} a_i + a_n.\end{aligned}$$

But only purveyors of mathematical austerity would insist too strongly on such precision. In practice, all sorts of modifications of this symbolism are used, and no one ever considers it necessary to add any words of explanation. The symbol

$$\sum_{\substack{i=1 \\ i \neq 4}}^n a_i,$$

for example, is an obvious way of writing

$$a_1 + a_2 + a_3 + a_5 + a_6 + \cdots + a_n,$$

or more precisely,

$$\sum_{i=1}^3 a_i + \sum_{i=5}^n a_i.$$

The deficiencies of the natural numbers which we discovered at the beginning of this chapter may be partially remedied by extending this system to the set of **integers**

$$\dots, -2, -1, 0, 1, 2, \dots$$

This set is denoted by **Z** (from German “Zahl,” number). Of properties P1–P12, only P7 fails for **Z**.

A still larger system of numbers is obtained by taking quotients  $m/n$  of integers (with  $n \neq 0$ ). These numbers are called **rational numbers**, and the set of all rational numbers is denoted by **Q** (for “quotients”). In this system of numbers all of P1–P12 are true. It is tempting to conclude that the “properties of numbers,” which we studied in some detail in Chapter 1, refer to just one set of numbers, namely, **Q**. There is, however, a still larger collection of numbers to which properties P1–P12 apply—the set of all **real numbers**, denoted by **R**. The real numbers include not only the rational numbers, but other numbers as well (the **irrational numbers**) which can be represented by infinite decimals;  $\pi$  and  $\sqrt{2}$  are both examples of irrational numbers. The proof that  $\pi$  is irrational is not easy—we shall devote all of Chapter 16 of Part III to a proof of this fact. The irrationality of  $\sqrt{2}$ , on the other hand, is quite simple, and was known to the Greeks. (Since the Pythagorean theorem shows that an isosceles right triangle, with sides of length 1, has a hypotenuse of length  $\sqrt{2}$ , it is not surprising that the Greeks should have investigated this question.) The proof depends on a few observations about the natural numbers. Every natural number  $n$  can be written either in the form  $2k$  for some integer  $k$ , or else in the form  $2k + 1$  for some integer  $k$  (this “obvious” fact has a simple proof by induction (Problem 8)). Those natural numbers of the form  $2k$  are called **even**; those of the form  $2k + 1$  are called **odd**. Note that even numbers have even squares, and odd numbers have odd squares:

$$(2k)^2 = 4k^2 = 2 \cdot (2k^2),$$

$$(2k + 1)^2 = 4k^2 + 4k + 1 = 2 \cdot (2k^2 + 2k) + 1.$$

In particular it follows that the converse must also hold: if  $n^2$  is even, then  $n$  is even; if  $n^2$  is odd, then  $n$  is odd. The proof that  $\sqrt{2}$  is irrational is now quite simple. Suppose that  $\sqrt{2}$  were rational; that is, suppose there were natural numbers  $p$

and  $q$  such that

$$\left(\frac{p}{q}\right)^2 = 2.$$

We can assume that  $p$  and  $q$  have no common divisor (since all common divisors could be divided out to begin with). Now we have

$$p^2 = 2q^2.$$

This shows that  $p^2$  is even, and consequently  $p$  must be even; that is,  $p = 2k$  for some natural number  $k$ . Then

$$p^2 = 4k^2 = 2q^2,$$

so

$$2k^2 = q^2.$$

This shows that  $q^2$  is even, and consequently that  $q$  is even. Thus both  $p$  and  $q$  are even, contradicting the fact that  $p$  and  $q$  have no common divisor. This contradiction completes the proof.

It is important to understand precisely what this proof shows. We have demonstrated that there is no rational number  $x$  such that  $x^2 = 2$ . This assertion is often expressed more briefly by saying that  $\sqrt{2}$  is irrational. Note, however, that the use of the symbol  $\sqrt{2}$  implies the existence of *some* number (necessarily irrational) whose square is 2. We have not proved that such a number exists and we can assert confidently that, at present, a proof is *impossible* for us. Any proof at this stage would have to be based on P1–P12 (the only properties of **R** we have mentioned); since P1–P12 are also true for **Q** the exact same argument would show that there is a rational number whose square is 2, and this we know is false. (Note that the reverse argument will not work—our proof that there is no rational number whose square is 2 cannot be used to show that there is no real number whose square is 2, because our proof used not only P1–P12 but also a special property of **Q**, the fact that every number in **Q** can be written  $p/q$  for integers  $p$  and  $q$ .)

This particular deficiency in our list of properties of the real numbers could, of course, be corrected by adding a new property which asserts the existence of square roots of positive numbers. Resorting to such a measure is, however, neither aesthetically pleasing nor mathematically satisfactory; we would still not know that every number has an  $n$ th root if  $n$  is odd, and that every positive number has an  $n$ th root if  $n$  is even. Even if we assumed this, we could not prove the existence of a number  $x$  satisfying  $x^5 + x + 1 = 0$  (even though there does happen to be one), since we do not know how to write the solution of the equation in terms of  $n$ th roots (in fact, it is known that the solution cannot be written in this form). And, of course, we certainly do not wish to assume that all equations have solutions, since this is false (no real number  $x$  satisfies  $x^2 + 1 = 0$ , for example). In fact, this direction of investigation is not a fruitful one. The most useful hints about the property distinguishing **R** from **Q**, the most compelling evidence for the necessity of elucidating this property, do not come from the study of numbers alone. In order to study the properties of the real numbers in a more profound way, we

must study more than the real numbers. At this point we must begin with the foundations of calculus, in particular the fundamental concept on which calculus is based—functions.

### PROBLEMS

1. Prove the following formulas by induction.

$$(i) \quad 1^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

$$(ii) \quad 1^3 + \cdots + n^3 = (1 + \cdots + n)^2.$$

2. Find a formula for

$$(i) \quad \sum_{i=1}^n (2i-1) = 1 + 3 + 5 + \cdots + (2n-1).$$

$$(ii) \quad \sum_{i=1}^n (2i-1)^2 = 1^2 + 3^2 + 5^2 + \cdots + (2n-1)^2.$$

Hint: What do these expressions have to do with  $1 + 2 + 3 + \cdots + 2n$  and  $1^2 + 2^2 + 3^2 + \cdots + (2n)^2$ ?

3. If  $0 \leq k \leq n$ , the “binomial coefficient”  $\binom{n}{k}$  is defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}, \quad \text{if } k \neq 0, n$$

$$\binom{n}{0} = \binom{n}{n} = 1. \quad (\text{This becomes a special case of the first formula if we define } 0! = 1.)$$

- (a) Prove that

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.$$

(The proof does not require an induction argument.)

This relation gives rise to the following configuration, known as “Pascal’s triangle”—a number not on one of the sides is the sum of the two numbers above it; the binomial coefficient  $\binom{n}{k}$  is the  $(k+1)$ st number in the  $(n+1)$ st row.

|  |  |  |   |   |    |    |    |
|--|--|--|---|---|----|----|----|
|  |  |  | 1 |   |    |    |    |
|  |  |  | 1 | 1 | 1  |    |    |
|  |  |  | 1 | 1 | 2  | 1  |    |
|  |  |  | 1 | 1 | 3  | 3  | 1  |
|  |  |  | 1 | 4 | 6  | 4  | 1  |
|  |  |  | 1 | 5 | 10 | 10 | 5  |
|  |  |  | 1 | 6 | 15 | 20 | 15 |
|  |  |  | 1 | 6 | 15 | 20 | 15 |
|  |  |  | 1 | 6 | 15 | 20 | 15 |

- (b) Notice that all the numbers in Pascal's triangle are natural numbers. Use part (a) to prove by induction that  $\binom{n}{k}$  is always a natural number. (Your entire proof by induction will, in a sense, be summed up in a glance by Pascal's triangle.)
- (c) Give another proof that  $\binom{n}{k}$  is a natural number by showing that  $\binom{n}{k}$  is the number of sets of exactly  $k$  integers each chosen from  $1, \dots, n$ .
- (d) Prove the "binomial theorem": If  $a$  and  $b$  are any numbers and  $n$  is a natural number, then

$$\begin{aligned}(a+b)^n &= a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + b^n \\ &= \sum_{j=0}^n \binom{n}{j}a^{n-j}b^j.\end{aligned}$$

- (e) Prove that

$$(i) \quad \sum_{j=0}^n \binom{n}{j} = \binom{n}{0} + \cdots + \binom{n}{n} = 2^n.$$

$$(ii) \quad \sum_{j=0}^n (-1)^j \binom{n}{j} = \binom{n}{0} - \binom{n}{1} + \cdots \pm \binom{n}{n} = 0.$$

$$(iii) \quad \sum_{l \text{ odd}} \binom{n}{l} = \binom{n}{1} + \binom{n}{3} + \cdots = 2^{n-1}.$$

$$(iv) \quad \sum_{l \text{ even}} \binom{n}{l} = \binom{n}{0} + \binom{n}{2} + \cdots = 2^{n-1}.$$

4. (a) Prove that

$$\sum_{k=0}^l \binom{n}{k} \binom{m}{l-k} = \binom{n+m}{l}.$$

Hint: Apply the binomial theorem to  $(1+x)^n(1+x)^m$ .

- (b) Prove that

$$\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}.$$

5. (a) Prove by induction on  $n$  that

$$1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r}$$

if  $r \neq 1$  (if  $r = 1$ , evaluating the sum certainly presents no problem).

- (b) Derive this result by setting  $S = 1 + r + \cdots + r^n$ , multiplying this equation by  $r$ , and solving the two equations for  $S$ .

6. The formula for  $1^2 + \cdots + n^2$  may be derived as follows. We begin with the formula

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

Writing this formula for  $k = 1, \dots, n$  and adding, we obtain

$$2^3 - 1^3 = 3 \cdot 1^2 + 3 \cdot 1 + 1$$

$$3^3 - 2^3 = 3 \cdot 2^2 + 3 \cdot 2 + 1$$

.

.

$$(n+1)^3 - n^3 = 3 \cdot n^2 + 3 \cdot n + 1$$

$$\overline{(n+1)^3 - 1} = 3[1^2 + \cdots + n^2] + 3[1 + \cdots + n] + n.$$

Thus we can find  $\sum_{k=1}^n k^2$  if we already know  $\sum_{k=1}^n k$  (which could have been found in a similar way). Use this method to find

(i)  $1^3 + \cdots + n^3$ .

(ii)  $1^4 + \cdots + n^4$ .

(iii)  $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \cdots + \frac{1}{n(n+1)}$ .

(iv)  $\frac{3}{1^2 \cdot 2^2} + \frac{5}{2^2 \cdot 3^2} + \cdots + \frac{2n+1}{n^2(n+1)^2}$ .

- \*7. Use the method of Problem 6 to show that  $\sum_{i=1}^n k^p$  can always be written in the form

$$\frac{n^{p+1}}{p+1} + An^p + Bn^{p-1} + Cn^{p-2} + \cdots$$

(The first 10 such expressions are

$$\begin{aligned}
\sum_{k=1}^n k &= \frac{1}{2}n^2 + \frac{1}{2}n \\
\sum_{k=1}^n k^2 &= \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n \\
\sum_{k=1}^n k^3 &= \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2 \\
\sum_{k=1}^n k^4 &= \frac{1}{5}n^5 + \frac{1}{2}n^4 + \frac{1}{3}n^3 - \frac{1}{30}n \\
\sum_{k=1}^n k^5 &= \frac{1}{6}n^6 + \frac{1}{2}n^5 + \frac{5}{12}n^4 - \frac{1}{12}n^2 \\
\sum_{k=1}^n k^6 &= \frac{1}{7}n^7 + \frac{1}{2}n^6 + \frac{1}{2}n^5 - \frac{1}{6}n^3 + \frac{1}{42}n \\
\sum_{k=1}^n k^7 &= \frac{1}{8}n^8 + \frac{1}{2}n^7 + \frac{7}{12}n^6 - \frac{7}{24}n^4 + \frac{1}{12}n^2 \\
\sum_{k=1}^n k^8 &= \frac{1}{9}n^9 + \frac{1}{2}n^8 + \frac{2}{3}n^7 - \frac{7}{15}n^5 + \frac{2}{9}n^3 - \frac{1}{30}n \\
\sum_{k=1}^n k^9 &= \frac{1}{10}n^{10} + \frac{1}{2}n^9 + \frac{3}{4}n^8 - \frac{7}{10}n^6 + \frac{1}{2}n^4 - \frac{3}{20}n^2 \\
\sum_{k=1}^n k^{10} &= \frac{1}{11}n^{11} + \frac{1}{2}n^{10} + \frac{5}{6}n^9 - 1n^7 + 1n^5 - \frac{1}{2}n^3 + \frac{5}{66}n.
\end{aligned}$$

Notice that the coefficients in the second column are always  $\frac{1}{2}$ , and that after the third column the powers of  $n$  with nonzero coefficients decrease by 2 until  $n^2$  or  $n$  is reached. The coefficients in all but the first two columns seem to be rather haphazard, but there actually is some sort of pattern; finding it may be regarded as a super-perspicacity test. See Problem 27-17 for the complete story.)

8. Prove that every natural number is either even or odd.
9. Prove that if a set  $A$  of natural numbers contains  $n_0$  and contains  $k + 1$  whenever it contains  $k$ , then  $A$  contains all natural numbers  $\geq n_0$ .
10. Prove the principle of mathematical induction from the well-ordering principle.
11. Prove the principle of complete induction from the ordinary principle of induction. Hint: If  $A$  contains 1 and  $A$  contains  $n + 1$  whenever it contains  $1, \dots, n$ , consider the set  $B$  of all  $k$  such that  $1, \dots, k$  are all in  $A$ .
12. (a) If  $a$  is rational and  $b$  is irrational, is  $a + b$  necessarily irrational? What if  $a$  and  $b$  are both irrational?  
(b) If  $a$  is rational and  $b$  is irrational, is  $ab$  necessarily irrational? (Careful!)  
(c) Is there a number  $a$  such that  $a^2$  is irrational, but  $a^4$  is rational?

- (d) Are there two irrational numbers whose sum and product are both rational?
13. (a) Prove that  $\sqrt{3}$ ,  $\sqrt{5}$ , and  $\sqrt{6}$  are irrational. Hint: To treat  $\sqrt{3}$ , for example, use the fact that every integer is of the form  $3n$  or  $3n+1$  or  $3n+2$ . Why doesn't this proof work for  $\sqrt{4}$ ?  
 (b) Prove that  $\sqrt[3]{2}$  and  $\sqrt[3]{3}$  are irrational.
14. Prove that  
 (a)  $\sqrt{2} + \sqrt{6}$  is irrational.  
 (b)  $\sqrt{2} + \sqrt{3}$  is irrational.
15. (a) Prove that if  $x = p + \sqrt{q}$  where  $p$  and  $q$  are rational, and  $m$  is a natural number, then  $x^m = a + b\sqrt{q}$  for some rational  $a$  and  $b$ .  
 (b) Prove also that  $(p - \sqrt{q})^m = a - b\sqrt{q}$ .
16. (a) Prove that if  $m$  and  $n$  are natural numbers and  $m^2/n^2 < 2$ , then  $(m+2n)^2/(m+n)^2 > 2$ ; show, moreover, that
- $$\frac{(m+2n)^2}{(m+n)^2} - 2 < 2 - \frac{m^2}{n^2}.$$
- (b) Prove the same results with all inequality signs reversed.  
 (c) Prove that if  $m/n < \sqrt{2}$ , then there is another rational number  $m'/n'$  with  $m/n < m'/n' < \sqrt{2}$ .
- \*17. It seems likely that  $\sqrt{n}$  is irrational whenever the natural number  $n$  is not the square of another natural number. Although the method of Problem 13 may actually be used to treat any particular case, it is not clear in advance that it will always work, and a proof for the general case requires some extra information. A natural number  $p$  is called a **prime number** if it is impossible to write  $p = ab$  for natural numbers  $a$  and  $b$  unless one of these is  $p$ , and the other 1; for convenience we also agree that 1 is *not* a prime number. The first few prime numbers are 2, 3, 5, 7, 11, 13, 17, 19. If  $n > 1$  is not a prime, then  $n = ab$ , with  $a$  and  $b$  both  $< n$ ; if either  $a$  or  $b$  is not a prime it can be factored similarly; continuing in this way proves that we can write  $n$  as a product of primes. For example,  $28 = 4 \cdot 7 = 2 \cdot 27$ .  
 (a) Turn this argument into a rigorous proof by complete induction. (To be sure, any reasonable mathematician would accept the informal argument, but this is partly because it would be obvious to her how to state it rigorously.)

A fundamental theorem about integers, which we will not prove here, states that this factorization is unique, except for the order of the factors. Thus, for example, 28 can never be written as a product of primes one of which is 3, nor can it be written in a way that involves 2 only once (now you should appreciate why 1 is not allowed as a prime).

- (b) Using this fact, prove that  $\sqrt{n}$  is irrational unless  $n = m^2$  for some natural number  $m$ .

- (c) Prove more generally that  $\sqrt[k]{n}$  is irrational unless  $n = m^k$ .

- (d) No discussion of prime numbers should fail to allude to Euclid's beautiful proof that there are infinitely many of them. Prove that there cannot be only finitely many prime numbers  $p_1, p_2, p_3, \dots, p_n$  by considering  $p_1 \cdot p_2 \cdot \dots \cdot p_n + 1$ .

- \*18.** (a) Prove that if  $x$  satisfies

$$x^n + a_{n-1}x^{n-1} + \dots + a_0 = 0,$$

for some integers  $a_{n-1}, \dots, a_0$ , then  $x$  is irrational unless  $x$  is an integer.  
(Why is this a generalization of Problem 17?)

- (b) Prove that  $\sqrt{6} - \sqrt{2} - \sqrt{3}$  is irrational.

- (c) Prove that  $\sqrt{2} + \sqrt[3]{2}$  is irrational. Hint: Start by working out the first 6 powers of this number.

- 19.** Prove Bernoulli's inequality: If  $h > -1$ , then

$$(1+h)^n \geq 1+nh.$$

Why is this trivial if  $h > 0$ ?

- 20.** The Fibonacci sequence  $a_1, a_2, a_3, \dots$  is defined as follows:

$$\begin{aligned} a_1 &= 1, \\ a_2 &= 1, \\ a_n &= a_{n-1} + a_{n-2} \quad \text{for } n \geq 3. \end{aligned}$$

This sequence, which begins 1, 1, 2, 3, 5, 8, ..., was discovered by Fibonacci (circa 1175–1250), in connection with a problem about rabbits. Fibonacci assumed that an initial pair of rabbits gave birth to one new pair of rabbits per month, and that after two months each new pair behaved similarly. The number  $a_n$  of pairs born in the  $n$ th month is  $a_{n-1} + a_{n-2}$ , because a pair of rabbits is born for each pair born the previous month, and moreover each pair born two months ago now gives birth to another pair. The number of interesting results about this sequence is truly amazing—there is even a Fibonacci Association which publishes a journal, *The Fibonacci Quarterly*. Prove that

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

One way of deriving this astonishing formula is presented in Problem 24-15.

- 21.** The Schwarz inequality (Problem 1-19) actually has a more general form:

$$\sum_{i=1}^n x_i y_i \leq \sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}.$$

Give three proofs of this, analogous to the three proofs in Problem 1-19.

22. The result in Problem 1-7 has an important generalization: If  $a_1, \dots, a_n \geq 0$ , then the “arithmetic mean”

$$A_n = \frac{a_1 + \dots + a_n}{n}$$

and “geometric mean”

$$G_n = \sqrt[n]{a_1 \dots a_n}$$

satisfy

$$G_n \leq A_n.$$

- (a) Suppose that  $a_1 < A_n$ . Then some  $a_i$  satisfies  $a_i > A_n$ ; for convenience, say  $a_2 > A_n$ . Let  $\bar{a}_1 = A_n$  and let  $\bar{a}_2 = a_1 + a_2 - \bar{a}_1$ . Show that

$$\bar{a}_1 \bar{a}_2 \geq a_1 a_2.$$

Why does repeating this process enough times eventually prove that  $G_n \leq A_n$ ? (This is another place where it is a good exercise to provide a formal proof by induction, as well as an informal reason.) When does equality hold in the formula  $G_n \leq A_n$ ?

The reasoning in this proof is related to another interesting proof.

- (b) Using the fact that  $G_n \leq A_n$  when  $n = 2$ , prove, by induction on  $k$ , that  $G_n \leq A_n$  for  $n = 2^k$ .
- (c) For a general  $n$ , let  $2^m > n$ . Apply part (b) to the  $2^m$  numbers

$$a_1, \dots, a_n, \underbrace{A_n, \dots, A_n}_{2^m-n \text{ times}}$$

to prove that  $G_n \leq a_n$ .

23. The following is a recursive definition of  $a^n$ :

$$\begin{aligned} a^1 &= a, \\ a^{n+1} &= a^n \cdot a. \end{aligned}$$

Prove, by induction, that

$$\begin{aligned} a^{n+m} &= a^n \cdot a^m, \\ (a^n)^m &= a^{nm}. \end{aligned}$$

(Don’t try to be fancy: use either induction on  $n$  or induction on  $m$ , not both at once.)

24. Suppose we know properties P1 and P4 for the natural numbers, but that multiplication has never been mentioned. Then the following can be used as a recursive definition of multiplication:

$$\begin{aligned} 1 \cdot b &= b, \\ (a+1) \cdot b &= a \cdot b + b. \end{aligned}$$

Prove the following (in the order suggested!):

$$\begin{aligned} a \cdot (b + c) &= a \cdot b + a \cdot c \text{ (use induction on } a\text{),} \\ a \cdot 1 &= a, \\ a \cdot b &= b \cdot a \text{ (you just finished proving the case } b = 1\text{).} \end{aligned}$$

25. In this chapter we began with the natural numbers and gradually built up to the real numbers. A completely rigorous discussion of this process requires a little book in itself (see Part V). No one has ever figured out how to get to the real numbers without going through this process, but if we do accept the real numbers as given, then the natural numbers can be *defined* as the real numbers of the form  $1, 1+1, 1+1+1$ , etc. The whole point of this problem is to show that there is a rigorous mathematical way of saying “etc.”

- (a) A set  $A$  of real numbers is called **inductive** if

- (1)  $1$  is in  $A$ ,
- (2)  $k + 1$  is in  $A$  whenever  $k$  is in  $A$ .

Prove that

- (i)  $\mathbf{R}$  is inductive.
- (ii) The set of positive real numbers is inductive.
- (iii) The set of positive real numbers unequal to  $\frac{1}{2}$  is inductive.
- (iv) The set of positive real numbers unequal to  $5$  is not inductive.
- (v) If  $A$  and  $B$  are inductive, then the set  $C$  of real numbers which are in both  $A$  and  $B$  is also inductive.

- (b) A real number  $n$  will be called a **natural number** if  $n$  is in *every* inductive set.

- (i) Prove that  $1$  is a natural number.
- (ii) Prove that  $k + 1$  is a natural number if  $k$  is a natural number.

26. There is a puzzle consisting of three spindles, with  $n$  concentric rings of decreasing diameter stacked on the first (Figure 1). A ring at the top of a stack may be moved from one spindle to another spindle, provided that it is not placed on top of a smaller ring. For example, if the smallest ring is moved to spindle 2 and the next-smallest ring is moved to spindle 3, then the smallest ring may be moved to spindle 3 also, on top of the next-smallest. Prove that the entire stack of  $n$  rings can be moved onto spindle 3 in  $2^n - 1$  moves, and that this cannot be done in fewer than  $2^n - 1$  moves.

- \*27. University B. once boasted 17 tenured professors of mathematics. Tradition prescribed that at their weekly luncheon meeting, faithfully attended by all 17, any members who had discovered an error in their published work should make an announcement of this fact, and promptly resign. Such an announcement had never actually been made, because no professor was aware of any errors in her or his work. This is not to say that no errors existed, however. In fact, over the years, in the work of every member of the department at least one error had been found, by some other member of the

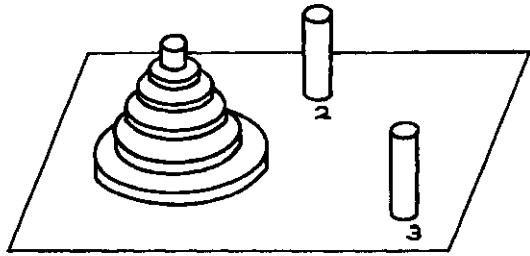


FIGURE 1

department. This error had been mentioned to all other members of the department, but the actual author of the error had been kept ignorant of the fact, to forestall any resignations.

One fateful year, the department was augmented by a visitor from another university, one Prof. X, who had come with hopes of being offered a permanent position at the end of the academic year. Naturally, he was apprised, by various members of the department, of the published errors which had been discovered. When the hoped-for appointment failed to materialize, Prof. X obtained his revenge at the last luncheon of the year. "I have enjoyed my visit here very much," he said, "but I feel that there is one thing that I have to tell you. At least one of you has published an incorrect result, which has been discovered by others in the department." What happened the next year?

- \*\*28.** After figuring out, or looking up, the answer to Problem 27, consider the following: Each member of the department already knew what Prof. X asserted, so how could his saying it change anything?



PART **2**  
FOUNDATIONS

*The statement is so frequently made  
that the differential calculus deals with  
continuous magnitude, and yet  
an explanation of this continuity is  
nowhere given;  
even the most rigorous expositions  
of the differential calculus do not base  
their proofs upon continuity but,  
with more or less consciousness of the fact,  
they either appeal to geometric notions  
or those suggested by geometry,  
or depend upon theorems which are never  
established in a purely arithmetic manner.  
Among these, for example,  
belongs the above-mentioned theorem,  
and a more careful investigation  
convinced me that this theorem, or  
any one equivalent to it, can be regarded  
in some way as a sufficient basis  
for infinitesimal analysis.  
It then only remained to discover its true  
origin in the elements of arithmetic  
and thus at the same time  
to secure a real definition of  
the essence of continuity.  
I succeeded Nov. 24, 1858, and  
a few days afterward I communicated  
the results  
of my meditations to my dear friend  
Durège with whom I had a long  
and lively discussion.*

RICHARD DEDEKIND

# CHAPTER 3 FUNCTIONS

Undoubtedly the most important concept in all of mathematics is that of a function—in almost every branch of modern mathematics functions turn out to be the central objects of investigation. It will therefore probably not surprise you to learn that the concept of a function is one of great generality. Perhaps it will be a relief to learn that, for the present, we will be able to restrict our attention to functions of a very special kind; even this small class of functions will exhibit sufficient variety to engage our attention for quite some time. We will not even begin with a proper definition. For the moment a provisional definition will enable us to discuss functions at length, and will illustrate the intuitive notion of functions, as understood by mathematicians. Later, we will consider and discuss the advantages of the modern mathematical definition. Let us therefore begin with the following:

## PROVISIONAL DEFINITION

A function is a rule which assigns, to each of certain real numbers, some other real number.

The following examples of functions are meant to illustrate and amplify this definition, which, admittedly, requires some such clarification.

*Example 1* The rule which assigns to each number the square of that number.

*Example 2* The rule which assigns to each number  $y$  the number

$$\frac{y^3 + 3y + 5}{y^2 + 1}.$$

*Example 3* The rule which assigns to each number  $c \neq 1, -1$  the number

$$\frac{c^3 + 3c + 5}{c^2 - 1}.$$

*Example 4* The rule which assigns to each number  $x$  satisfying  $-17 \leq x \leq \pi/3$  the number  $x^2$ .

*Example 5* The rule which assigns to each number  $a$  the number 0 if  $a$  is irrational, and the number 1 if  $a$  is rational.

*Example 6* The rule which assigns

to 2 the number 5,  
to 17 the number  $\frac{36}{\pi}$ ,

to  $\frac{\pi^2}{17}$  the number 28,

to  $\frac{36}{\pi}$  the number 28,

and to any  $y \neq 2, 17, \pi^2/17$ , or  $36/\pi$ , the number 16 if  $y$  is of the form  $a + b\sqrt{2}$  for  $a, b$  in  $\mathbb{Q}$ .

*Example 7* The rule which assigns to each number  $t$  the number  $t^3 + x$ . (This rule depends, of course, on what the number  $x$  is, so we are really describing infinitely many different functions, one for each number  $x$ .)

*Example 8* The rule which assigns to each number  $z$  the number of 7's in the decimal expansion of  $z$ , if this number is finite, and  $-\pi$  if there are infinitely many 7's in the decimal expansion of  $z$ .

One thing should be abundantly clear from these examples—a function is *any* rule that assigns numbers to certain other numbers, not just a rule which can be expressed by an algebraic formula, or even by one uniform condition which applies to every number; nor is it necessarily a rule which you, or anybody else, can actually apply in practice (no one knows, for example, what rule 8 associates to  $\pi$ ). Moreover, the rule may neglect some numbers and it may not even be clear to which numbers the function applies (try to determine, for example, whether the function in Example 6 applies to  $\pi$ ). The set of numbers to which a function *does* apply is called the *domain* of the function.

Before saying anything else about functions we badly need some notation. Since throughout this book we shall frequently be talking about functions (indeed we shall hardly ever talk about anything else) we need a convenient way of naming functions, and of referring to functions in general. The standard practice is to denote a function by a letter. For obvious reasons the letter “ $f$ ” is a favorite, thereby making “ $g$ ” and “ $h$ ” other obvious candidates, but any letter (or any reasonable symbol, for that matter) will do, not excluding “ $x$ ” and “ $y$ ”, although these letters are usually reserved for indicating numbers. If  $f$  is a function, then the number which  $f$  associates to a number  $x$  is denoted by  $f(x)$ —this symbol is read “ $f$  of  $x$ ” and is often called the **value of  $f$  at  $x$** . Naturally, if we denote a function by  $x$ , some other letter must be chosen to denote the number (a perfectly legitimate, though perverse, choice would be “ $f$ ,” leading to the symbol  $x(f)$ ). Note that the symbol  $f(x)$  makes sense only for  $x$  in the domain of  $f$ ; for other  $x$  the symbol  $f(x)$  is not defined.

If the functions defined in Examples 1–8 are denoted by  $f, g, h, r, s, \theta, \alpha_x$ , and  $y$ , then we can rewrite their definitions as follows:

$$(1) \quad f(x) = x^2 \quad \text{for all } x.$$

$$(2) \quad g(y) = \frac{y^3 + 3y + 5}{y^2 + 1} \quad \text{for all } y.$$

$$(3) \quad h(c) = \frac{c^3 + 3c + 5}{c^2 - 1} \quad \text{for all } c \neq 1, -1.$$

$$(4) \quad r(x) = x^2 \quad \text{for all } x \text{ such that } -17 \leq x \leq \pi/3.$$

$$(5) \quad s(x) = \begin{cases} 0, & x \text{ irrational} \\ 1, & x \text{ rational.} \end{cases}$$

$$(6) \quad \theta(x) = \begin{cases} 5, & x = 2 \\ \frac{36}{\pi}, & x = 17 \\ 28, & x = \frac{\pi^2}{17} \\ 28, & x = \frac{36}{\pi} \\ 16, & x \neq 2, 17, \frac{\pi^2}{17}, \text{ or } \frac{36}{\pi}, \text{ and } x = a + b\sqrt{2} \text{ for } a, b \in \mathbb{Q}. \end{cases}$$

$$(7) \quad \alpha_x(t) = t^3 + x \quad \text{for all numbers } t.$$

$$(8) \quad y(x) = \begin{cases} n, & \text{exactly } n \text{ 7's appear in the decimal expansion of } x \\ -\pi, & \text{infinitely many 7's appear in the decimal expansion of } x. \end{cases}$$

These definitions illustrate the common procedure adopted for defining a function  $f$ —indicating what  $f(x)$  is for every number  $x$  in the domain of  $f$ . (Notice that this is exactly the same as indicating  $f(a)$  for every number  $a$ , or  $f(b)$  for every number  $b$ , etc.) In practice, certain abbreviations are tolerated. Definition (1) could be written simple

$$(1) \quad f(x) = x^2$$

the qualifying phrase “for all  $x$ ” being understood. Of course, for definition (4) the only possible abbreviation is

$$(4) \quad r(x) = x^2, \quad -17 \leq x \leq \pi/3.$$

It is usually understood that a definition such as

$$k(x) = \frac{1}{x} + \frac{1}{x-1}, \quad x \neq 0, 1$$

can be shortened to

$$k(x) = \frac{1}{x} + \frac{1}{x-1};$$

in other words, *unless the domain is explicitly restricted further, it is understood to consist of all numbers for which the definition makes any sense at all.*

You should have little difficulty checking the following assertions about the functions defined above:

$$f(x+1) = f(x) + 2x + 1;$$

$$g(x) = h(x) \text{ if } x^3 + 3x + 5 = 0;$$

$$r(x+1) = r(x) + 2x + 1 \text{ if } -17 \leq x \leq \frac{\pi}{3} - 1;$$

$$s(x+y) = s(x) \text{ if } y \text{ is rational;}$$

$$\theta\left(\frac{\pi^2}{17}\right) = \theta\left(\frac{36}{\pi}\right);$$

$$\alpha_x(x) = x \cdot [f(x) + 1];$$

$$y\left(\frac{1}{3}\right) = 0, \quad y\left(\frac{7}{9}\right) = -\pi.$$

If the expression  $f(s(a))$  looks unreasonable to you, then you are forgetting that  $s(a)$  is a number like any other number, so that  $f(s(a))$  makes sense. As a matter of fact,  $f(s(a)) = s(a)$  for all  $a$ . Why? Even more complicated expressions than  $f(s(a))$  are, after a first exposure, no more difficult to unravel. The expression

$$f(r(s(\theta(\alpha_3(y(\frac{1}{3})))))),$$

formidable as it appears, may be evaluated quite easily with a little patience:

$$\begin{aligned} & f(r(s(\theta(\alpha_3(y(\frac{1}{3})))))) \\ &= f(r(s(\theta(\alpha_3(0))))) \\ &= f(r(s(\theta(3)))) \\ &= f(r(s(16))) \\ &= f(r(1)) \\ &= f(1) \\ &= 1. \end{aligned}$$

The first few problems at the end of this chapter give further practice manipulating this symbolism.

The function defined in (1) is a rather special example of an extremely important class of functions, the polynomial functions. A function  $f$  is a **polynomial function** if there are real numbers  $a_0, \dots, a_n$  such that

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0, \quad \text{for all } x$$

(when  $f(x)$  is written in this form it is usually tacitly assumed that  $a_n \neq 0$ ). The highest power of  $x$  with a nonzero coefficient is called the **degree** of  $f$ ; for example, the polynomial function  $f$  defined by  $f(x) = 5x^6 + 137x^4 - \pi$  has degree 6.

The functions defined in (2) and (3) belong to a somewhat larger class of functions, the **rational functions**; these are the functions of the form  $p/q$  where  $p$  and  $q$  are polynomial functions (and  $q$  is not the function which is always 0). The rational functions are themselves quite special examples of an even larger class of functions, very thoroughly studied in calculus, which are simpler than many of the functions first mentioned in this chapter. The following are examples of this kind of function:

$$(9) \quad f(x) = \frac{x + x^2 + x \sin^2 x}{x \sin x + x \sin^2 x}$$

$$(10) \quad f(x) = \sin(x^2).$$

$$(11) \quad f(x) = \sin(\sin(x^2)).$$

$$(12) \quad f(x) = \sin^2(\sin(\sin^2(x \sin^2 x^2))) \cdot \sin\left(\frac{x + \sin(x \sin x)}{x + \sin x}\right).$$

By what criterion, you may feel impelled to ask, can such functions, especially a monstrosity like (12), be considered simple? The answer is that they can be built up from a few simple functions using a few simple means of combining functions. In order to construct the functions (9)–(12) we need to start with the “identity function”  $I$ , for which  $I(x) = x$ , and the “sine function”  $\sin$ , whose value  $\sin(x)$  at  $x$  is often written simple  $\sin x$ . The following are some of the important ways in which functions may be combined to produce new functions.

If  $f$  and  $g$  are any two functions, we can define a new function  $f + g$ , called the **sum** of  $f$  and  $g$ , by the equation

$$(f + g)(x) = f(x) + g(x).$$

Note that according to the conventions we have adopted, the domain of  $f + g$  consists of all  $x$  for which “ $f(x) + g(x)$ ” makes sense, i.e., the set of all  $x$  in both domain  $f$  and domain  $g$ . If  $A$  and  $B$  are any two sets, then  $A \cap B$  (read “ $A$  intersect  $B$ ” or “the intersection of  $A$  and  $B$ ”) denotes the set of  $x$  in both  $A$  and  $B$ ; this notation allows us to write  $\text{domain}(f + g) = \text{domain } f \cap \text{domain } g$ .

In a similar vein, we define the **product**  $f \cdot g$  and the **quotient**  $\frac{f}{g}$  (or  $f/g$ ) of  $f$  and  $g$  by

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

and

$$\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}.$$

Moreover, if  $g$  is a function and  $c$  is a number, we define a new function  $c \cdot g$  by

$$(c \cdot g)(x) = c \cdot g(x).$$

This becomes a special case of the notation  $f \cdot g$  if we agree that the symbol  $c$  should also represent the function  $f$  defined by  $f(x) = c$ ; such a function, which has the same value for all numbers  $x$ , is called a **constant function**.

The domain of  $f \cdot g$  is  $\text{domain } f \cap \text{domain } g$ , and the domain of  $c \cdot g$  is simply the domain of  $g$ . On the other hand, the domain of  $f/g$  is rather complicated—it may be written  $\text{domain } f \cap \text{domain } g \cap \{x : g(x) \neq 0\}$ , the symbol  $\{x : g(x) \neq 0\}$  denoting the set of numbers  $x$  such that  $g(x) \neq 0$ . In general,  $\{x : \dots\}$  denotes the set of all  $x$  such that “ $\dots$ ” is true. Thus  $\{x : x^3 + 3 < 11\}$  denotes the set of all numbers  $x$  such that  $x^3 < 8$ , and consequently  $\{x : x^3 + 3 < 11\} = \{x : x < 2\}$ . Either of these symbols could just as well have been written using  $y$  everywhere instead of  $x$ . Variations of this notation are common, but hardly require any discussion. Any one can guess that  $\{x > 0 : x^3 < 8\}$  denotes the set of positive numbers whose cube is less than 8; it could be expressed more formally as  $\{x : x > 0 \text{ and } x^3 < 8\}$ . Incidentally, this set is equal to the set  $\{x : 0 < x < 2\}$ . One

variation is slightly less transparent, but very standard. The set  $\{1, 3, 2, 4\}$ , for example, contains just the four numbers 1, 2, 3, and 4; it can also be denoted by  $\{x : x = 1 \text{ or } x = 3 \text{ or } x = 2 \text{ or } x = 4\}$ .

Certain facts about the sum, product, and quotient of functions are obvious consequences of facts about sums, products, and quotients of numbers. For example, it is very easy to prove that

$$(f + g) + h = f + (g + h).$$

The proof is characteristic of almost every proof which demonstrates that two functions are equal—the two functions must be shown to have the same domain, and the same value at any number in the domain. For example, to prove that  $(f + g) + h = f + (g + h)$ , note that unraveling the definition of the two sides gives

$$\begin{aligned} [(f + g) + h](x) &= (f + g)(x) + h(x) \\ &= [f(x) + g(x)] + h(x) \end{aligned}$$

and

$$\begin{aligned} [f + (g + h)](x) &= f(x) + (g + h)(x) \\ &= f(x) + [g(x) + h(x)], \end{aligned}$$

and the equality of  $[f(x) + g(x)] + h(x)$  and  $f(x) + [g(x) + h(x)]$  is a fact about numbers. In this proof the equality of the two domains was not explicitly mentioned because this is obvious, as soon as we begin to write down these equations; the domain of  $(f + g) + h$  and of  $f + (g + h)$  is clearly domain  $f \cap$  domain  $g \cap$  domain  $h$ . We naturally write  $f + g + h$  for  $(f + g) + h = f + (g + h)$ , precisely as we did for numbers.

It is just as easy to prove that  $(f \cdot g) \cdot h = f \cdot (g \cdot h)$ , and this function is denoted by  $f \cdot g \cdot h$ . The equations  $f + g = g + f$  and  $f \cdot g = g \cdot f$  should also present no difficulty.

Using the operations  $+, \cdot, /$  we can now express the function  $f$  defined in (9) by

$$f = \frac{I + I \cdot I + I \cdot \sin \cdot \sin}{I \cdot \sin + I \cdot \sin \cdot \sin}.$$

It should be clear, however, that we cannot express function (10) this way. We require yet another way of combining functions. This combination, the composition of two functions, is by far the most important.

If  $f$  and  $g$  are any two functions, we define a new function  $f \circ g$ , the **composition** of  $f$  and  $g$ , by

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

the domain of  $f \circ g$  is  $\{x : x \text{ is in domain } g \text{ and } g(x) \text{ is in domain } f\}$ . The symbol “ $f \circ g$ ” is often read “ $f$  circle  $g$ .” Compared to the phrase “the composition of  $f$  and  $g$ ” this has the advantage of brevity, of course, but there is another advantage of far greater import: there is much less chance of confusing  $f \circ g$  with  $g \circ f$ , and

these must *not* be confused, since they are not usually equal; in fact, almost any  $f$  and  $g$  chosen at random will illustrate this point (try  $f = I \cdot I$  and  $g = \sin$ , for example). Lest you become too apprehensive about the operation of composition, let us hasten to point out that composition *is* associative:

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

(and the proof is a triviality); this function is denoted by  $f \circ g \circ h$ . We can now write the functions (10), (11), (12) as

$$(10) \quad f = \sin \circ (I \cdot I),$$

$$(11) \quad f = \sin \circ \sin \circ (I \cdot I),$$

$$(12) \quad f = (\sin \cdot \sin) \circ \sin \circ (\sin \cdot \sin) \circ (I \cdot [(\sin \cdot \sin) \circ (I \cdot I)]) .$$

$$\sin \circ \left( \frac{I + \sin \circ (I \cdot \sin)}{I + \sin} \right).$$

One fact has probably already become clear. Although this method of writing functions reveals their “structure” very clearly, it is hardly short or convenient. The shortest name for the function  $f$  such that  $f(x) = \sin(x^2)$  for all  $x$  unfortunately seems to be “the function  $f$  such that  $f(x) = \sin(x^2)$  for all  $x$ .” The need for abbreviating this clumsy description has been clear for two hundred years, but no reasonable abbreviation has received universal acclaim. At present the strongest contender for this honor is something like

$$x \rightarrow \sin(x^2)$$

(read “ $x$  goes to  $\sin(x^2)$ ” or just “ $x$  arrow  $\sin(x^2)$ ”), but it is hardly popular among writers of calculus textbooks. In this book we will tolerate a certain amount of ellipsis, and speak of “the function  $f(x) = \sin(x^2)$ .” Even more popular is the quite drastic abbreviation: “the function  $\sin(x^2)$ .” For the sake of precision we will never use this description, which, strictly speaking, confuses a number and a function, but it is so convenient that you will probably end up adopting it for personal use. As with any convention, utility is the motivating factor, and this criterion is reasonable so long as the slight logical deficiencies cause no confusion. On occasion, confusion *will* arise unless a more precise description is used. For example, “the function  $x + t^3$ ” is an ambiguous phrase; it could mean either

$$x \rightarrow x + t^3, \text{ i.e., the function } f \text{ such that } f(x) = x + t^3 \text{ for all } x$$

or

$$t \rightarrow x + t^3, \text{ i.e., the function } f \text{ such that } f(t) = x + t^3 \text{ for all } t.$$

As we shall see, however, for many important concepts associated with functions, calculus has a notation which contains the “ $x \rightarrow$ ” built in.

By now we have made a sufficiently extensive investigation of functions to warrant reconsidering our definition. We have defined a function as a “rule,” but it is hardly clear what this means. If we ask “What happens if you break this rule?” it is not easy to say whether this question is merely facetious or actually profound.

A more substantial objection to the use of the word “rule” is that

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

and

$$f(x) = x^2 + 3x + 3 - 3(x + 1)$$

are certainly *different* rules, if by a rule we mean the actual instructions given for determining  $f(x)$ ; nevertheless, we want

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

and

$$f(x) = x^2 + 3x + 3 - c(x + 1)$$

to define the same function. For this reason, a function is sometimes defined as an “association” between numbers; unfortunately the word “association” escapes the objections raised against “rule” only because it is even more vague.

There is, of course, a satisfactory way of defining functions, or we should never have gone to the trouble of criticizing our original definition. But a satisfactory definition can never be constructed by finding synonyms for English words which are troublesome. The definition which mathematicians have finally accepted for “function” is a beautiful example of the means by which intuitive ideas have been incorporated into rigorous mathematics. The correct question to ask about a function is not “What is a rule?” or “What is an association?” but “What does one have to know about a function in order to know all about it?” The answer to the last question is easy—for each number  $x$  one needs to know the number  $f(x)$ ; we can imagine a table which would display all the information one could desire about the function  $f(x) = x^2$ :

| $x$         | $f(x)$  |
|-------------|---------|
| 1           | 1       |
| -1          | 1       |
| 2           | 4       |
| -2          | 4       |
| $\sqrt{2}$  | 2       |
| $-\sqrt{2}$ | 2       |
| $\pi$       | $\pi^2$ |
| $-\pi$      | $\pi^2$ |

It is not even necessary to arrange the numbers in a table (which would actually be impossible if we wanted to list all of them). Instead of a two column array we can consider various pairs of numbers

$$(1, 1), (-1, 1), (2, 4), (-2, 4), (\pi, \pi^2), (\sqrt{2}, 2), \dots$$

simply collected together into a set.\* To find  $f(1)$  we simply take the second number of the pair whose first member is 1; to find  $f(\pi)$  we take the second number of the pair whose first member is  $\pi$ . We seem to be saying that a function might as well be defined as a collection of pairs of numbers. For example, if we were given the following collection (which contains just 5 pairs):

$$f = \{(1, 7), (3, 7), (5, 3), (4, 8), (8, 4)\},$$

then  $f(1) = 7$ ,  $f(3) = 7$ ,  $f(5) = 3$ ,  $f(4) = 8$ ,  $f(8) = 4$  and 1, 3, 4, 5, 8 are the only numbers in the domain of  $f$ . If we consider the collection

$$f = \{(1, 7), (3, 7), (2, 5), (1, 8), (8, 4)\},$$

then  $f(3) = 7$ ,  $f(2) = 5$ ,  $f(8) = 4$ ; but it is impossible to decide whether  $f(1) = 7$  or  $f(1) = 8$ . In other words, a function cannot be defined to be any old collection of pairs of numbers; we must rule out the possibility which arose in this case. We are therefore led to the following definition.

#### DEFINITION

A **function** is a collection of pairs of numbers with the following property: if  $(a, b)$  and  $(a, c)$  are both in the collection, then  $b = c$ ; in other words, the collection must not contain two different pairs with the same first element.

This is our first full-fledged definition, and illustrates the format we shall always use to define significant new concepts. These definitions are so important (at least as important as theorems) that it is essential to know when one is actually at hand, and to distinguish them from comments, motivating remarks, and casual explanations. They will be preceded by the word **DEFINITION**, contain the term being defined in boldface letters, and constitute a paragraph unto themselves.

There is one more definition (actually defining two things at once) which can now be made rigorously:

#### DEFINITION

If  $f$  is a function, the **domain** of  $f$  is the set of all  $a$  for which there is some  $b$  such that  $(a, b)$  is in  $f$ . If  $a$  is in the domain of  $f$ , it follows from the definition of a function that there is, in fact, a *unique* number  $b$  such that  $(a, b)$  is in  $f$ . This unique  $b$  is denoted by  $f(a)$ .

With this definition we have reached our goal: the important thing about a function  $f$  is that a number  $f(x)$  is determined for each number  $x$  in its domain. You may feel that we have also reached the point where an intuitive definition has been replaced by an abstraction with which the mind can hardly grapple. Two consolations may be offered. First, although a function has been defined as a

\*The pairs occurring here are often called “ordered pairs,” to emphasize that, for example,  $(2, 4)$  is not the same pair as  $(4, 2)$ . It is only fair to warn that we are going to define functions in terms of ordered pairs, another undefined term. Ordered pairs *can* be defined, however, and an appendix to this chapter has been provided for skeptics.

collection of pairs, there is nothing to stop you from *thinking* of a function as a rule. Second, neither the intuitive nor the formal definition indicates the best way of thinking about functions. The best way is to draw pictures; but this requires a chapter all by itself.

### PROBLEMS

1. Let  $f(x) = 1/(1+x)$ . What is

- (i)  $f(f(x))$  (for which  $x$  does this make sense?).
- (ii)  $f\left(\frac{1}{x}\right)$ .
- (iii)  $f(cx)$ .
- (iv)  $f(x+y)$ .
- (v)  $f(x)+f(y)$ .
- (vi) For which numbers  $c$  is there a number  $x$  such that  $f(cx) = f(x)$ .  
Hint: There are a lot more than you might think at first glance.
- (vii) For which numbers  $c$  is it true that  $f(cx) = f(x)$  for two different numbers  $x$ ?

2. Let  $g(x) = x^2$ , and let

$$h(x) = \begin{cases} 0, & x \text{ rational} \\ 1, & x \text{ irrational.} \end{cases}$$

- (i) For which  $y$  is  $h(y) \leq y$ ?
- (ii) For which  $y$  is  $h(y) \leq g(y)$ ?
- (iii) What is  $g(h(z)) - h(z)$ ?
- (iv) For which  $w$  is  $g(w) \leq w$ ?
- (v) For which  $\varepsilon$  is  $g(g(\varepsilon)) = g(\varepsilon)$ ?

3. Find the domain of the functions defined by the following formulas.

- (i)  $f(x) = \sqrt{1-x^2}$ .
- (ii)  $f(x) = \sqrt{1-\sqrt{1-x^2}}$ .
- (iii)  $f(x) = \frac{1}{x-1} + \frac{1}{x-2}$ .
- (iv)  $f(x) = \sqrt{1-x^2} + \sqrt{x^2-1}$ .
- (v)  $f(x) = \sqrt{1-x} + \sqrt{x-2}$ .

4. Let  $S(x) = x^2$ , let  $P(x) = 2^x$ , and let  $s(x) = \sin x$ . Find each of the following. In each case your answer should be a *number*.

- (i)  $(S \circ P)(y)$ .
- (ii)  $(S \circ s)(y)$ .
- (iii)  $(S \circ P \circ s)(t) + (s \circ P)(t)$ .
- (iv)  $s(t^3)$ .

5. Express each of the following functions in terms of  $S$ ,  $P$ ,  $s$ , using only  $+$ ,  $\cdot$ , and  $\circ$  (for example, the answer to (i) is  $P \circ s$ ). In each case your

answer should be a *function*.

- (i)  $f(x) = 2^{\sin x}$ .
- (ii)  $f(x) = \sin 2^x$ .
- (iii)  $f(x) = \sin x^2$ .
- (iv)  $f(x) = \sin^2 x$  (remember that  $\sin^2 x$  is an abbreviation for  $(\sin x)^2$ ).
- (v)  $f(t) = 2^{t^c}$ . (Note:  $a^{bc}$  always means  $a^{(bc)}$ ; this convention is adopted because  $(a^b)^c$  can be written more simply as  $a^{bc}$ .)
- (vi)  $f(u) = \sin(2^u + 2^{u^2})$ .
- (vii)  $f(y) = \sin(\sin(\sin(2^{2^{\sin y}})))$ .
- (viii)  $f(a) = 2^{\sin^2 a} + \sin(a^2) + 2^{\sin(a^2 + \sin a)}$ .

Polynomial functions, because they are simple, yet flexible, occupy a favored role in most investigations of functions. The following two problems illustrate their flexibility, and guide you through a derivation of their most important elementary properties.

6. (a) If  $x_1, \dots, x_n$  are distinct numbers, find a polynomial function  $f_i$  of degree  $n - 1$  which is 1 at  $x_i$  and 0 at  $x_j$  for  $j \neq i$ . Hint: the product of all  $(x - x_i)$  for  $j \neq i$ , is 0 at  $x_j$  if  $j \neq i$ . (This product is usually denoted by

$$\prod_{\substack{j=1 \\ j \neq i}}^n (x - x_j),$$

the symbol  $\Pi$  (capital pi) playing the same role for products that  $\Sigma$  plays for sums.)

- (b) Now find a polynomial function  $f$  of degree  $n - 1$  such that  $f(x_i) = a_i$ , where  $a_1, \dots, a_n$  are given numbers. (You should use the functions  $f_i$  from part (a). The formula you will obtain is called the “Lagrange interpolation formula.”)

7. (a) Prove that for any polynomial function  $f$ , and any number  $a$ , there is a polynomial function  $g$ , and a number  $b$ , such that  $f(x) = (x - a)g(x) + b$  for all  $x$ . (The idea is simply to divide  $(x - a)$  into  $f(x)$  by long division, until a constant remainder is left. For example, the calculation

$$\begin{array}{r}
 & x^2 & +x & -2 \\
 x - 1 \overline{)x^3} & & -3x + 1 \\
 & x^3 & -x^2 & \\
 \hline
 & x^2 & -3x & \\
 & x^2 & -x & \\
 \hline
 & & -2x + 1 & \\
 & & -2x + 2 & \\
 \hline
 & & & -1
 \end{array}$$

shows that  $x^3 - 3x + 1 = (x - 1)(x^2 + x - 2) - 1$ . A formal proof is possible by induction on the degree of  $f$ .)

- (b) Prove that if  $f(a) = 0$ , then  $f(x) = (x - a)g(x)$  for some polynomial function  $g$ . (The converse is obvious.)
- (c) Prove that if  $f$  is a polynomial function of degree  $n$ , then  $f$  has at most  $n$  roots, i.e., there are at most  $n$  numbers  $a$  with  $f(a) = 0$ .
- (d) Show that for each  $n$  there is a polynomial function of degree  $n$  with  $n$  roots. If  $n$  is even find a polynomial function of degree  $n$  with no roots, and if  $n$  is odd find one with only one root.

8. For which numbers  $a$ ,  $b$ ,  $c$ , and  $d$  will the function

$$f(x) = \frac{ax + b}{cx + d}$$

satisfy  $f(f(x)) = x$  for all  $x$ ?

9. (a) If  $A$  is any set of real numbers, define a function  $C_A$  as follows:

$$C_A(x) = \begin{cases} 1, & x \text{ in } A \\ 0, & x \text{ not in } A. \end{cases}$$

Find expressions for  $C_{A \cap B}$  and  $C_{A \cup B}$  and  $C_{\mathbf{R} - A}$ , in terms of  $C_A$  and  $C_B$ . (The symbol  $A \cap B$  was defined in this chapter, but the other two may be new to you. They can be defined as follows:

$$\begin{aligned} A \cup B &= \{x : x \text{ is in } A \text{ or } x \text{ is in } B\}, \\ \mathbf{R} - A &= \{x : x \text{ is in } \mathbf{R} \text{ but } x \text{ is not in } A\}. \end{aligned}$$

- (b) Suppose  $f$  is a function such that  $f(x) = 0$  or  $1$  for each  $x$ . Prove that there is a set  $A$  such that  $f = C_A$ .
- (c) Show that  $f = f^2$  if and only if  $f = C_A$  for some set  $A$ .

10. (a) For which functions  $f$  is there a function  $g$  such that  $f = g^2$ ? Hint: You can certainly answer this question if “function” is replaced by “number.”
- (b) For which functions  $f$  is there a function  $g$  such that  $f = 1/g$ ?
  - \*(c) For which functions  $b$  and  $c$  can we find a function  $x$  such that

$$(x(t))^2 + b(t)x(t) + c(t) = 0$$

for all numbers  $t$ ?

- \*(d) What conditions must the functions  $a$  and  $b$  satisfy if there is to be a function  $x$  such that

$$a(t)x(t) + b(t) = 0$$

for all numbers  $t$ ? How many such functions  $x$  will there be?

11. (a) Suppose that  $H$  is a function and  $y$  is a number such that  $H(H(y)) = y$ . What is

$$\underbrace{H(H(H(\cdots(H(y)\cdots)))}_{80 \text{ times}} ?$$

- (b) Same question if 80 is replaced by 81.
  - (c) Same question if  $H(H(y)) = H(y)$ .
  - \*(d) Find a function  $H$  such that  $H(H(x)) = H(x)$  for all numbers  $x$ , and such that  $H(1) = 36$ ,  $H(2) = \pi/3$ ,  $H(13) = 47$ ,  $H(36) = 36$ ,  $H(\pi/3) = \pi/3$ ,  $H(47) = 47$ . (Don't try to "solve" for  $H(x)$ ; there are many functions  $H$  with  $H(H(x)) = H(x)$ . The extra conditions on  $H$  are supposed to suggest a way of finding a suitable  $H$ .)
  - \*(e) Find a function  $H$  such that  $H(H(x)) = H(x)$  for all  $x$ , and such that  $H(1) = 7$ ,  $H(17) = 18$ .
12. A function  $f$  is **even** if  $f(x) = f(-x)$  and **odd** if  $f(x) = -f(-x)$ . For example,  $f$  is even if  $f(x) = x^2$  or  $f(x) = |x|$  or  $f(x) = \cos x$ , while  $f$  is odd if  $f(x) = x$  or  $f(x) = \sin x$ .
- (a) Determine whether  $f + g$  is even, odd, or not necessarily either, in the four cases obtained by choosing  $f$  even or odd, and  $g$  even or odd. (Your answers can most conveniently be displayed in a  $2 \times 2$  table.)
  - (b) Do the same for  $f \cdot g$ .
  - (c) Do the same for  $f \circ g$ .
  - (d) Prove that every even function  $f$  can be written  $f(x) = g(|x|)$ , for infinitely many functions  $g$ .

- \*13. (a) Prove that any function  $f$  with domain  $\mathbf{R}$  can be written  $f = E + O$ , where  $E$  is even and  $O$  is odd.
- (b) Prove that this way of writing  $f$  is unique. (If you try to do part (b) first, by "solving" for  $E$  and  $O$  you will probably find the solution to part (a).)
14. If  $f$  is any function, define a new function  $|f|$  by  $|f|(x) = |f(x)|$ . If  $f$  and  $g$  are functions, define two new functions,  $\max(f, g)$  and  $\min(f, g)$ , by

$$\begin{aligned}\max(f, g)(x) &= \max(f(x), g(x)), \\ \min(f, g)(x) &= \min(f(x), g(x)).\end{aligned}$$

Find an expression for  $\max(f, g)$  and  $\min(f, g)$  in terms of  $| |$ .

15. (a) Show that  $f = \max(f, 0) + \min(f, 0)$ . This particular way of writing  $f$  is fairly useful; the functions  $\max(f, 0)$  and  $\min(f, 0)$  are called the **positive** and **negative parts** of  $f$ .
- (b) A function  $f$  is called **nonnegative** if  $f(x) \geq 0$  for all  $x$ . Prove that any function  $f$  can be written  $f = g - h$ , where  $g$  and  $h$  are nonnegative, in infinitely many ways. (The "standard way" is  $g = \max(f, 0)$  and  $h = -\min(f, 0)$ .) Hint: Any *number* can certainly be written as the difference of two nonnegative *numbers* in infinitely many ways.

- \*16. Suppose  $f$  satisfies  $f(x + y) = f(x) + f(y)$  for all  $x$  and  $y$ .
- (a) Prove that  $f(x_1 + \dots + x_n) = f(x_1) + \dots + f(x_n)$ .
  - (b) Prove that there is some number  $c$  such that  $f(x) = cx$  for all *rational* numbers  $x$  (at this point we're not trying to say anything about  $f(x)$  for irrational  $x$ ). Hint: First figure out what  $c$  must be. Now prove that

$f(x) = cx$ , first when  $x$  is a natural number, then when  $x$  is an integer, then when  $x$  is the reciprocal of an integer and, finally, for all rational  $x$ .

- \*17. If  $f(x) = 0$  for all  $x$ , then  $f$  satisfies  $f(x+y) = f(x) + f(y)$  for all  $x$  and  $y$ , and also  $f(x \cdot y) = f(x) \cdot f(y)$  for all  $x$  and  $y$ . Now suppose that  $f$  satisfies these two properties, but that  $f(x)$  is not always 0. Prove that  $f(x) = x$  for all  $x$ , as follows:
- Prove that  $f(1) = 1$ .
  - Prove that  $f(x) = x$  if  $x$  is rational.
  - Prove that  $f(x) > 0$  if  $x > 0$ . (This part is tricky, but if you have been paying attention to the philosophical remarks accompanying the problems in the last two chapters, you will know what to do.)
  - Prove that  $f(x) > f(y)$  if  $x > y$ .
  - Prove that  $f(x) = x$  for all  $x$ . Hint: Use the fact that between any two numbers there is a rational number.
- \*18. Precisely what conditions must  $f$ ,  $g$ ,  $h$ , and  $k$  satisfy in order that  $f(x)g(y) = h(x)k(y)$  for all  $x$  and  $y$ ?
- \*19. (a) Prove that there do *not* exist functions  $f$  and  $g$  with either of the following properties:
  - $f(x) + g(y) = xy$  for all  $x$  and  $y$ .
  - $f(x) \cdot g(y) = x + y$  for all  $x$  and  $y$ .
 Hint: Try to get some information about  $f$  or  $g$  by choosing particular values of  $x$  and  $y$ .
 (b) Find functions  $f$  and  $g$  such that  $f(x+y) = g(xy)$  for all  $x$  and  $y$ .
- \*20. (a) Find a function  $f$ , other than a constant function, such that  $|f(y) - f(x)| \leq |y - x|$ .
 (b) Suppose that  $f(y) - f(x) \leq (y - x)^2$  for all  $x$  and  $y$ . (Why does this imply that  $|f(y) - f(x)| \leq (y - x)^2$ ?) Prove that  $f$  is a constant function. Hint: Divide the interval from  $x$  to  $y$  into  $n$  equal pieces.
21. Prove or give a counterexample for each of the following assertions:
  - $f \circ (g + h) = f \circ g + f \circ h$ .
  - $(g + h) \circ f = g \circ f + h \circ f$ .
  - $\frac{1}{f \circ g} = \frac{1}{f} \circ g$ .
  - $\frac{1}{f \circ g} = f \circ \left(\frac{1}{g}\right)$ .
 22. (a) Suppose  $g = h \circ f$ . Prove that if  $f(x) = f(y)$ , then  $g(x) = g(y)$ .
 (b) Conversely, suppose that  $f$  and  $g$  are two functions such that  $g(x) = g(y)$  whenever  $f(x) = f(y)$ . Prove that  $g = h \circ f$  for some function  $h$ . Hint: Just try to define  $h(z)$  when  $z$  is of the form  $z = f(x)$  (these are the only  $z$  that matter) and use the hypotheses to show that your definition will not run into trouble.

- 23.** Suppose that  $f \circ g = I$ , where  $I(x) = x$ . Prove that
- if  $x \neq y$ , then  $g(x) \neq g(y)$ ;
  - every number  $b$  can be written  $b = f(a)$  for some number  $a$ .
- \*24.** (a) Suppose  $g$  is a function with the property that  $g(x) \neq g(y)$  if  $x \neq y$ . Prove that there is a function  $f$  such that  $f \circ g = I$ .
- (b) Suppose that  $f$  is a function such that every number  $b$  can be written  $b = f(a)$  for some number  $a$ . Prove that there is a function  $g$  such that  $f \circ g = I$ .
- \*25.** Find a function  $f$  such that  $g \circ f = I$  for some  $g$ , but such that there is no function  $h$  with  $f \circ h = I$ .
- \*26.** Suppose  $f \circ g = I$  and  $h \circ f = I$ . Prove that  $g = h$ . Hint: Use the fact that composition is associative.
- 27.** (a) Suppose  $f(x) = x + 1$ . Are there any functions  $g$  such that  $f \circ g = g \circ f$ ?
- (b) Suppose  $f$  is a constant function. For which functions  $g$  does  $f \circ g = g \circ f$ ?
- (c) Suppose that  $f \circ g = g \circ f$  for all functions  $g$ . Show that  $f$  is the identity function,  $f(x) = x$ .
- 28.** (a) Let  $F$  be the set of all functions whose domain is  $\mathbf{R}$ . Prove that, using  $+$  and  $\cdot$  as defined in this chapter, all of properties P1–P9 except P7 hold for  $F$ , provided 0 and 1 are interpreted as constant functions.
- (b) Show that P7 does not hold.
- \*(c)** Show that P10–P12 cannot hold. In other words, show that there is no collection  $P$  of functions in  $F$ , such that P10–P12 hold for  $P$ . (It is sufficient, and will simplify things, to consider only functions which are 0 except at two points  $x_0$  and  $x_1$ .)
- (d) Suppose we define  $f < g$  to mean that  $f(x) < g(x)$  for all  $x$ . Which of P'10–P'13 (in Problem 1-8) now hold?
- (e) If  $f < g$ , is  $h \circ f < h \circ g$ ? Is  $f \circ h < g \circ h$ ?

## APPENDIX. ORDERED PAIRS

Not only in the definition of a function, but in other parts of the book as well, it is necessary to use the notion of an ordered pair of objects. A definition has not yet been given, and we have never even stated explicitly what properties an ordered pair is supposed to have. The one property which we will require states formally that the ordered pair  $(a, b)$  should be determined by  $a$  and  $b$ , and the order in which they are given:

$$\text{if } (a, b) = (c, d), \text{ then } a = c \text{ and } b = d.$$

Ordered pairs may be treated most conveniently by simply introducing  $(a, b)$  as an undefined term and adopting the basic property as an axiom—since this property is the only significant fact about ordered pairs, there is not much point worrying about what an ordered pair “really” is. Those who find this treatment satisfactory need read no further.

The rest of this short appendix is for the benefit of those readers who will feel uncomfortable unless ordered pairs are somehow defined so that this basic property becomes a theorem. There is no point in restricting our attention to ordered pairs of numbers; it is just as reasonable, and just as important, to have available the notion of an ordered pair of any two mathematical objects. This means that our definition ought to involve only concepts common to all branches of mathematics. The one common concept which pervades all areas of mathematics is that of a set, and ordered pairs (like everything else in mathematics) can be defined in this context; an ordered pair will turn out to be a set of a rather special sort.

The set  $\{a, b\}$ , containing the two elements  $a$  and  $b$ , is an obvious first choice, but will not do as a definition for  $(a, b)$ , because there is no way of determining from  $\{a, b\}$  which of  $a$  or  $b$  is meant to be the first element. A more promising candidate is the rather startling set:

$$\{\{a\}, \{a, b\}\}.$$

This set has two members, both of which are *themselves* sets; one member is the set  $\{a\}$ , containing the single member  $a$ , the other is the set  $\{a, b\}$ . Shocking as it may seem, we are going to define  $(a, b)$  to be this set. The justification for this choice is given by the theorem immediately following the definition—the definition works, and there really isn’t anything else worth saying.

DEFINITION

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

THEOREM 1

If  $(a, b) = (c, d)$ , then  $a = c$  and  $b = d$ .

PROOF

The hypothesis means that

$$\{\{a\}, \{a, b\}\} = \{\{c\}, \{c, d\}\}.$$

Now  $\{\{a\}, \{a, b\}\}$  contains just two members,  $\{a\}$  and  $\{a, b\}$ ; and  $a$  is the only common element of these two members of  $\{\{a\}, \{a, b\}\}$ . Similarly,  $c$  is the unique

common member of both members of  $\{\{c\}, \{c, d\}\}$ . Therefore  $a = c$ . We therefore have

$$\{\{a\}, \{a, b\}\} = \{\{a\}, \{a, d\}\},$$

and only the proof that  $b = d$  remains. It is convenient to distinguish 2 cases.

*Case 1.*  $b = a$ . In this case,  $\{a, b\} = \{a\}$ , so the set  $\{\{a\}, \{a, b\}\}$  really has only one member, namely,  $\{a\}$ . The same must be true of  $\{\{a\}, \{a, d\}\}$ , so  $\{a, d\} = \{a\}$ , which implies that  $d = a = b$ .

*Case 2.*  $b \neq a$ . In this case,  $b$  is in one member of  $\{\{a\}, \{a, b\}\}$  but not in the other. It must therefore be true that  $b$  is in one member of  $\{\{a\}, \{a, d\}\}$  but not in the other. This can happen only if  $b$  is in  $\{a, d\}$ , but  $b$  is not in  $\{a\}$ ; thus  $b = a$  or  $b = d$ , but  $b \neq a$ ; so  $b = d$ . ■

# CHAPTER 4

## GRAPHS

Mention the real numbers to a mathematician and the image of a straight line will probably form in her mind, quite involuntarily. And most likely she will neither banish nor too eagerly embrace this mental picture of the real numbers. “Geometric intuition” will allow her to interpret statements about numbers in terms of this picture, and may even suggest methods of proving them. Although the properties of the real numbers which were studied in Part I are not greatly illuminated by a geometric picture, such an interpretation will be a great aid in Part II.

You are probably already familiar with the conventional method of considering the straight line as a picture of the real numbers, i.e., of associating to each real number a point on a line. To do this (Figure 1) we pick, arbitrarily, a point which we label 0, and a point to the right, which we label 1. The point twice as far to the right is labeled 2, the point the same distance from 0 to 1, but to the left of 0, is labeled  $-1$ , etc. With this arrangement, if  $a < b$ , then the point corresponding to  $a$  lies to the left of the point corresponding to  $b$ . We can also draw rational numbers, such as  $\frac{1}{2}$ , in the obvious way. It is usually taken for granted that the irrational numbers also somehow fit into this scheme, so that every real number can be drawn as a point on the line. We will not make too much fuss about justifying this assumption, since this method of “drawing” numbers is intended solely as a method of picturing certain abstract ideas, and our proofs will never rely on these pictures (although we will frequently use a picture to suggest or help explain a proof). Because this geometric picture plays such a prominent, albeit inessential role, geometric terminology is frequently employed when speaking of numbers—thus a number is sometimes called a *point*, and  $\mathbf{R}$  is often called the *real line*.

The number  $|a - b|$  has a simple interpretation in terms of this geometric picture: it is the distance between  $a$  and  $b$ , the length of the line segment which has  $a$  as one end point and  $b$  as the other. This means, to choose an example whose frequent occurrence justifies special consideration, that the set of numbers  $x$  which satisfy  $|x - a| < \varepsilon$  may be pictured as the collection of points whose distance from  $a$  is less than  $\varepsilon$ . This set of points is the “interval” from  $a - \varepsilon$  to  $a + \varepsilon$ , which may also be described as the points corresponding to numbers  $x$  with  $a - \varepsilon < x < a + \varepsilon$  (Figure 2).

Sets of numbers which correspond to intervals arise so frequently that it is desirable to have special names for them. The set  $\{x : a < x < b\}$  is denoted by  $(a, b)$  and called the **open interval** from  $a$  to  $b$ . This notation naturally creates some ambiguity, since  $(a, b)$  is also used to denote a pair of numbers, but in context it is always clear (or can easily be made clear) whether one is talking about a pair or an interval. Note that if  $a \geq b$ , then  $(a, b) = \emptyset$ , the set with no elements; in prac-

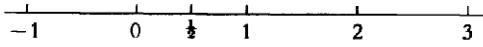


FIGURE 1



FIGURE 2

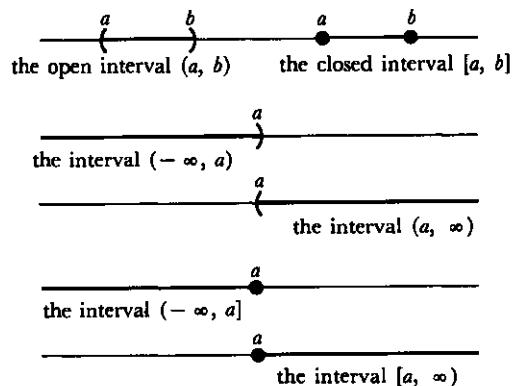


FIGURE 3

tice, however, it is almost always assumed (explicitly if one has been careful, and implicitly otherwise), that whenever an interval  $(a, b)$  is mentioned, the number  $a$  is less than  $b$ .

The set  $\{x : a \leq x \leq b\}$  is denoted by  $[a, b]$  and is called the **closed interval** from  $a$  to  $b$ . This symbol is usually reserved for the case  $a < b$ , but it is sometimes used for  $a = b$ , also. The usual pictures for the intervals  $(a, b)$  and  $[a, b]$  are shown in Figure 3; since no reasonably accurate picture could ever indicate the difference between the two intervals, various conventions have been adopted. Figure 3 also shows certain "infinite" intervals. The set  $\{x : x > a\}$  is denoted by  $(a, \infty)$ , while the set  $\{x : x \geq a\}$  is denoted by  $[a, \infty)$ ; the sets  $(-\infty, a)$  and  $(-\infty, a]$  are defined similarly. At this point a standard warning must be issued: the symbols  $\infty$  and  $-\infty$ , though usually read "infinity" and "minus infinity," are *purely* suggestive; there is no number " $\infty$ " which satisfies  $\infty \geq a$  for all numbers  $a$ . While the symbols  $\infty$  and  $-\infty$  will appear in many contexts, it is always necessary to define these uses in ways that refer only to numbers. The set  $\mathbf{R}$  of all real numbers is also considered to be an "interval," and is sometimes denoted by  $(-\infty, \infty)$ .

Of even greater interest to us than the method of drawing numbers is a method of drawing pairs of numbers. This procedure, probably also familiar to you, requires a "coordinate system," two straight lines intersecting at right angles. To distinguish these straight lines, we call one the *horizontal axis*, and one the *vertical axis*. (More prosaic terminology, such as the "first" and "second" axes, is probably preferable from a logical point of view, but most people hold their books, or at least their blackboards, in the same way, so that "horizontal" and "vertical" are more descriptive.) Each of the two axes could be labeled with real numbers, but we can also label points on the horizontal axis with pairs  $(a, 0)$  and points on the vertical axis with pairs  $(0, b)$ , so that the intersection of the two axes, the "origin" of the coordinate system, is labeled  $(0, 0)$ . Any pair  $(a, b)$  can now be drawn as in Figure 4, lying at the vertex of the rectangle whose other three vertices are labeled  $(0, 0)$ ,  $(a, 0)$ , and  $(0, b)$ . The numbers  $a$  and  $b$  are called the *first* and *second coordinates*, respectively, of the point determined in this way.

Our real concern, let us recall, is a method of drawing functions. Since a function is just a collection of pairs of numbers, we can draw a function by drawing each of the pairs in the function. The drawing obtained in this way is called the **graph** of the function. In other words, the graph of  $f$  contains all the points corresponding to pairs  $(x, f(x))$ . Since most functions contain infinitely many pairs, drawing the graph promises to be a laborious undertaking, but, in fact, many functions have graphs which are quite easy to draw.

Not surprisingly, the simplest functions of all, the constant functions  $f(x) = c$ , have the simplest graphs. It is easy to see that the graph of the function  $f(x) = c$  is a straight line parallel to the horizontal axis, at distance  $c$  from it (Figure 5).

The functions  $f(x) = cx$  also have particularly simple graphs—straight lines through  $(0, 0)$ , as in Figure 6. A proof of this fact is indicated in Figure 7:

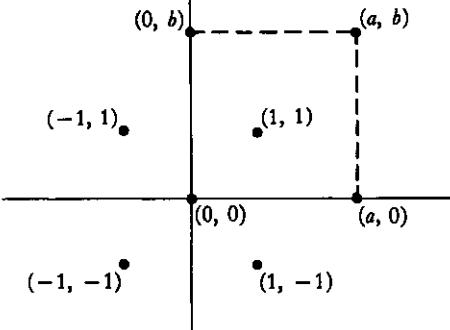


FIGURE 4

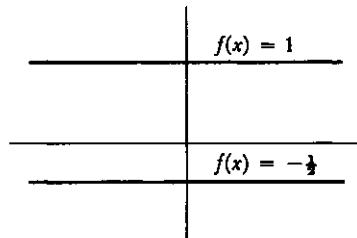


FIGURE 5

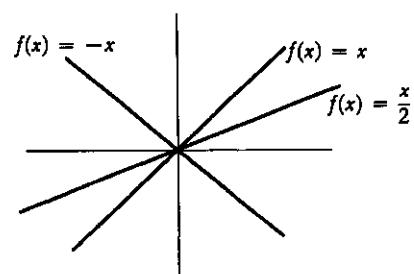


FIGURE 6

Let  $x$  be some number not equal to 0, and let  $L$  be the straight line which passes through the origin  $O$ , corresponding to  $(0, 0)$ , and through the point  $A$ , corresponding to  $(x, cx)$ . A point  $A'$ , with first coordinate  $y$ , will lie on  $L$  when the triangle  $A'B'O$  is similar to the triangle  $ABO$ , thus when

$$\frac{A'B'}{OB'} = \frac{AB}{OB} = c;$$

this is precisely the condition that  $A'$  corresponds to the pair  $(y, cy)$ , i.e., that  $A'$  lies on the graph of  $f$ . The argument has implicitly assumed that  $c > 0$ , but the other cases are treated easily enough. The number  $c$ , which measures the ratio of the sides of the triangles appearing in the proof, is called the *slope* of the straight line, and a line parallel to this line is also said to have slope  $c$ .

This demonstration has neither been labeled nor treated as a formal proof. Indeed, a rigorous demonstration would necessitate a digression which we are not at all prepared to follow. The rigorous proof of *any* statement connecting geometric and algebraic concepts would first require a real proof (or a precisely stated assumption) that the points on a straight line correspond in an exact way to the real numbers. Aside from this, it would be necessary to develop plane geometry as precisely as we intend to develop the properties of real numbers. Now the detailed development of plane geometry is a beautiful subject, but it is by no means a prerequisite for the study of calculus. We shall use geometric pictures only as an aid to intuition; for our purposes (and for most of mathematics) it is perfectly satisfactory to *define* the plane to be the set of all pairs of real numbers, and to *define* straight lines as certain collections of pairs, including, among others, the collections  $\{(x, cx) : x \text{ a real number}\}$ . To provide this artificially constructed geometry with all the structure of geometry studied in high school, one more definition is required. If  $(a, b)$  and  $(c, d)$  are two points in the plane, i.e., pairs of real numbers, we *define* the **distance** between  $(a, b)$  and  $(c, d)$  to be

$$\sqrt{(a - c)^2 + (b - d)^2}.$$

If the motivation for this definition is not clear, Figure 8 should serve as adequate explanation—with this definition the Pythagorean theorem has been built into our geometry.\*

Reverting once more to our informal geometric picture, it is not hard to see (Figure 9) that the graph of the function  $f(x) = cx + d$  is a straight line with slope  $c$ , passing through the point  $(0, d)$ . For this reason, the functions  $f(x) = cx + d$  are called **linear functions**. Simple as they are, linear functions occur frequently, and you should feel comfortable working with them. The following is a typical problem whose solution should not cause any trouble. Given two distinct points  $(a, b)$  and  $(c, d)$ , find the linear function  $f$  whose graph goes through  $(a, b)$  and  $(c, d)$ . This amounts to saying that  $f(a) = b$  and  $f(c) = d$ . If

\* The fastidious reader might object to this definition on the grounds that nonnegative numbers are not yet known to have square roots. This objection is really unanswerable at the moment—the definition will just have to be accepted with reservations, until this little point is settled.

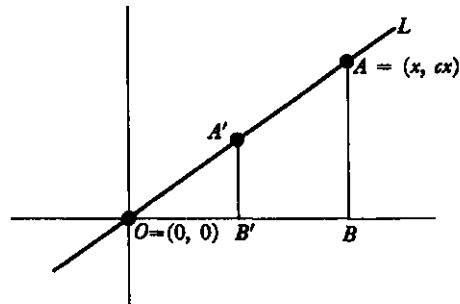


FIGURE 7

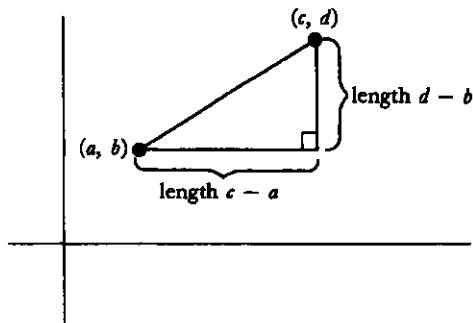


FIGURE 8

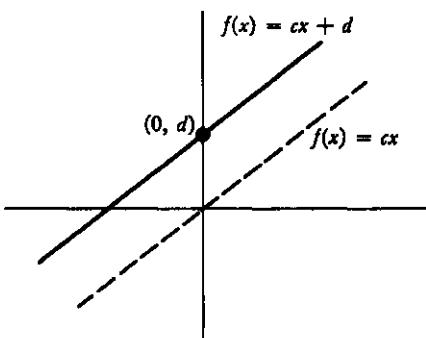


FIGURE 9

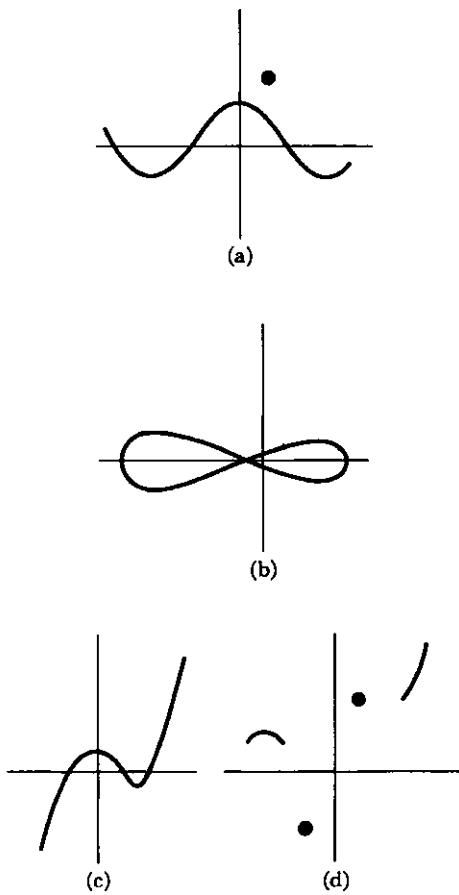


FIGURE 10

$f$  is to be of the form  $f(x) = \alpha x + \beta$ , then we must have

$$\begin{aligned}\alpha a + \beta &= b, \\ \alpha c + \beta &= d;\end{aligned}$$

therefore  $\alpha = (d - b)/(c - a)$  and  $\beta = b - [(d - b)/(c - a)]a$ , so

$$f(x) = \frac{d - b}{c - a}x + b - \frac{d - b}{c - a}a = \frac{d - b}{c - a}(x - a) + b,$$

a formula most easily remembered by using the “point-slope form” (see Problem 6).

Of course, this solution is possible only if  $a \neq c$ ; the graphs of linear functions account only for the straight lines which are not parallel to the vertical axis. The vertical straight lines are not the graph of *any* function at all; in fact, the graph of a function can never contain even two distinct points on the same vertical line. This conclusion is immediate from the definition of a function—two points on the same vertical line correspond to pairs of the form  $(a, b)$  and  $(a, c)$  and, by definition, a function cannot contain  $(a, b)$  and  $(a, c)$  if  $b \neq c$ . Conversely, if a set of points in the plane has the property that no two points lie on the same vertical line, then it is surely the graph of a function. Thus, the first two sets in Figure 10 are not graphs of functions and the last two are; notice that the fourth is the graph of a function whose domain is not all of  $\mathbf{R}$ , since some vertical lines have no points on them at all.

After the linear functions the simplest is perhaps the function  $f(x) = x^2$ . If we draw some of the pairs in  $f$ , i.e., some of the pairs of the form  $(x, x^2)$ , we obtain a picture like Figure 11.

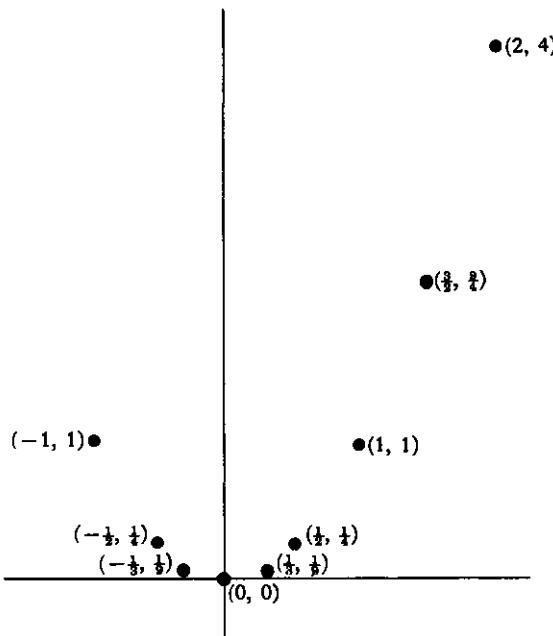


FIGURE 11

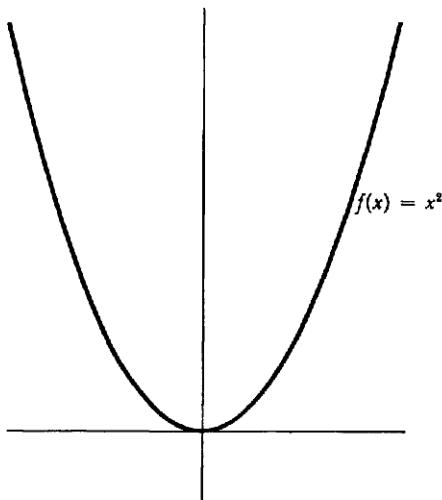
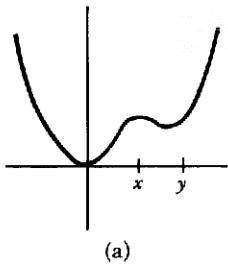
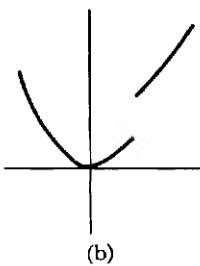


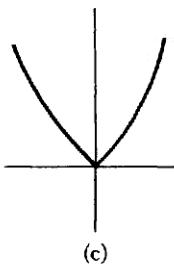
FIGURE 12



(a)



(b)



(c)

FIGURE 13

It is not hard to convince yourself that all the pairs  $(x, x^2)$  lie along a curve like the one shown in Figure 12; this curve is known as a **parabola**.

Since a graph is just a drawing on paper, made (in this case) with printer's ink, the question "Is this what the graph really looks like?" is hard to phrase in any sensible manner. No drawing is ever *really* correct since the line has thickness. Nevertheless, there are some questions which one *can* ask: for example, how can you be sure that the graph does not look like one of the drawings in Figure 13? It is easy to see, and even to prove, that the graph cannot look like (a); for if  $0 < x < y$ , then  $x^2 < y^2$ , so the graph should be higher at  $y$  than at  $x$ , which is not the case in (a). It is also easy to see, simply by drawing a very accurate graph, first plotting many pairs  $(x, x^2)$ , that the graph cannot have a large "jump" as in (b) or a "corner" as in (c). In order to prove these assertions, however, we first need to say, in a mathematical way, what it means for a function not to have a "jump" or "corner"; these ideas already involve some of the fundamental concepts of calculus. Eventually we will be able to define them rigorously, but meanwhile you may amuse yourself by attempting to define these concepts, and then examining your definitions critically. Later these definitions may be compared with the ones mathematicians have agreed upon. If they compare favorably, you are certainly to be congratulated!

The functions  $f(x) = x^n$ , for various natural numbers  $n$ , are sometimes called **power functions**. Their graphs are most easily compared as in Figure 14, by drawing several at once.

The power functions are only special cases of polynomial functions, introduced in the previous chapter. Two particular polynomial functions are graphed in

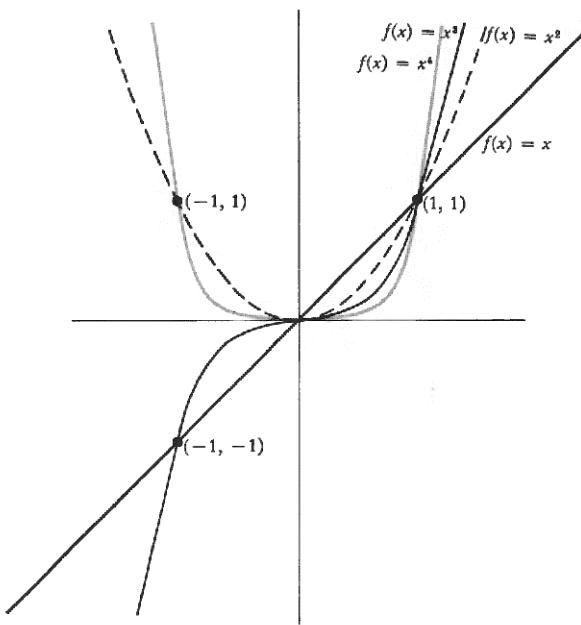


FIGURE 14

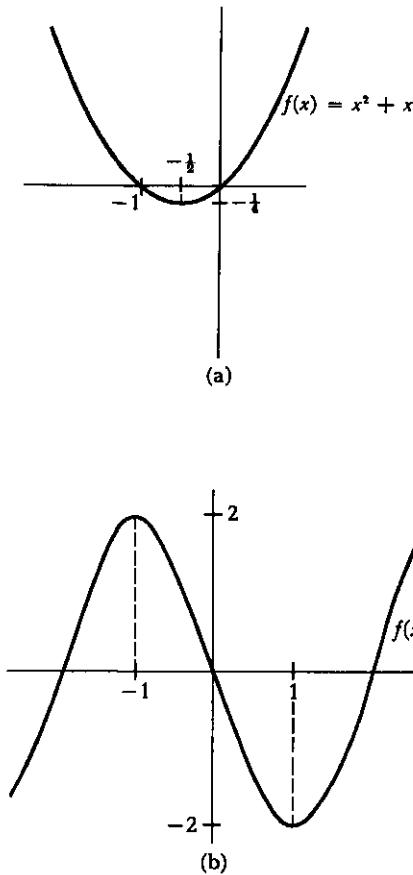


FIGURE 15

Figure 15, while Figure 16 is meant to give a general idea of the graph of the polynomial function

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0,$$

in the case  $a_n > 0$ .

In general, the graph of  $f$  will have at most  $n-1$  “peaks” or “valleys” (a “peak” is a point like  $(x, f(x))$  in Figure 16, while a “valley” is a point like  $(y, f(y))$ ). The number of peaks and valleys may actually be much smaller (the power functions, for example, have at most one valley). Although these assertions are easy to make, we will not even contemplate giving proofs until Part III (once the powerful methods of Part III are available, the proofs will be very easy).

Figure 17 illustrates the graphs of several rational functions. The rational functions exhibit even greater variety than the polynomial functions, but their behavior will also be easy to analyze once we can use the derivative, the basic tool of Part III.

Many interesting graphs can be constructed by “piecing together” the graphs of functions already studied. The graph in Figure 18 is made up entirely of straight lines. The function  $f$  with this graph satisfies

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

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

$$f(x) = 1, \quad |x| \geq 1,$$

and is a linear function on each interval  $[1/(n+1), 1/n]$  and  $[-1/n, -1/(n+1)]$ . (The number 0 is not in the domain of  $f$ .) Of course, one can write out an explicit formula for  $f(x)$ , when  $x$  is in  $[1/(n+1), 1/n]$ ; this is a good exercise in the use of linear functions, and will also convince you that a picture is worth a thousand words.

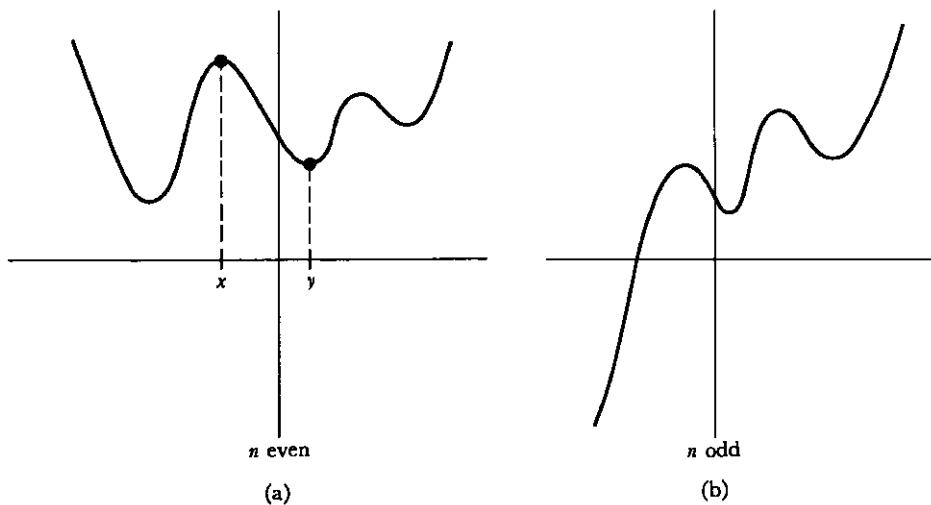


FIGURE 16

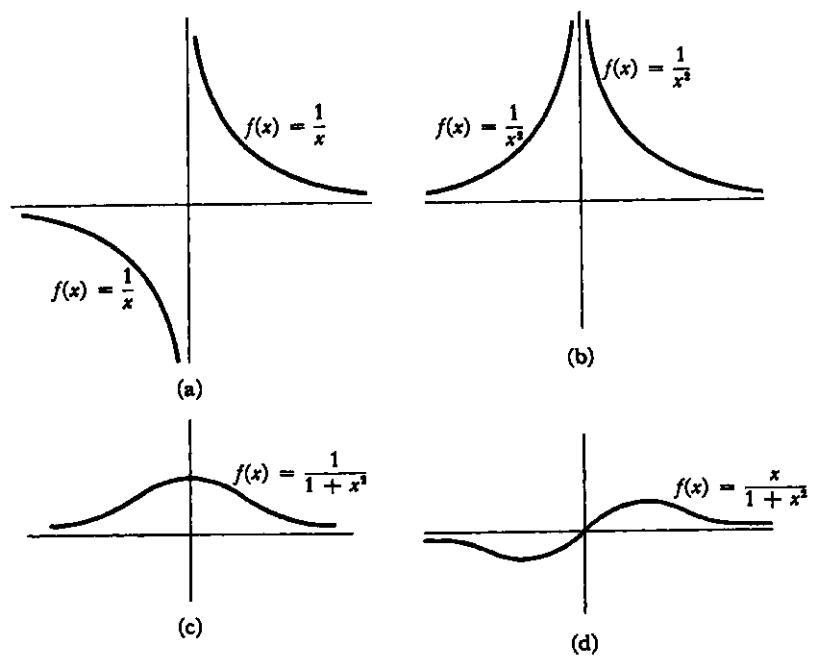


FIGURE 17

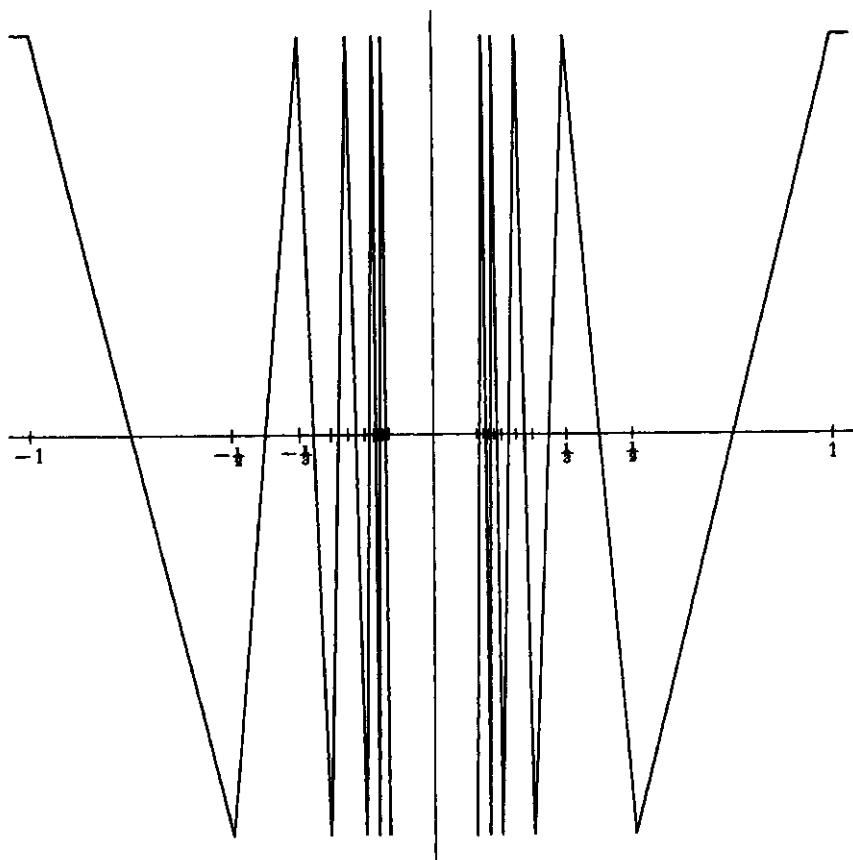


FIGURE 18

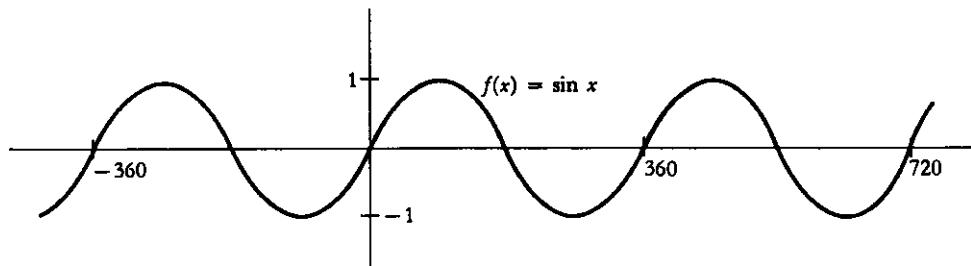


FIGURE 19

It is actually possible to define, in a much simpler way, a function which exhibits this same property of oscillating infinitely often near 0, by using the sine function. In Chapter 15 we will discuss this function in detail, and radian measure in particular; for the time being it will be easiest to use degree measurements for angles. The graph of the sine function is shown in Figure 19 (the scale on the horizontal axis has been altered so that the graph will be clearer; radian measure has, besides important mathematical properties, the additional advantage that such changes are unnecessary).

Now consider the function  $f(x) = \sin 1/x$ . The graph of  $f$  is shown in Figure 20. To draw this graph it helps to first observe that

$$\begin{aligned}f(x) &= 0 && \text{for } x = \frac{1}{180}, \frac{1}{360}, \frac{1}{540}, \dots \\f(x) &= 1 && \text{for } x = \frac{1}{90}, \frac{1}{90+360}, \frac{1}{90+720}, \dots \\f(x) &= -1 && \text{for } x = \frac{1}{270}, \frac{1}{270+360}, \frac{1}{270+720}, \dots\end{aligned}$$

Notice that when  $x$  is large, so that  $1/x$  is small,  $f(x)$  is also small; when  $x$  is

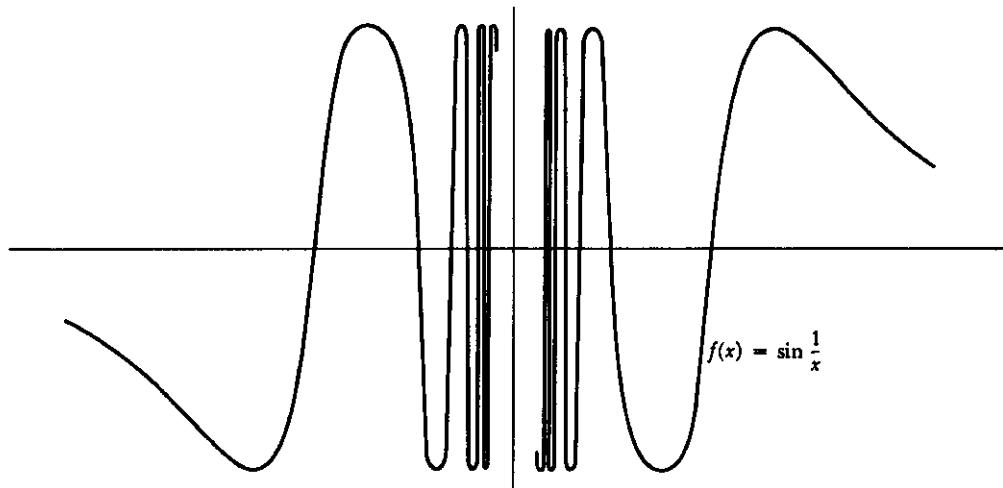


FIGURE 20

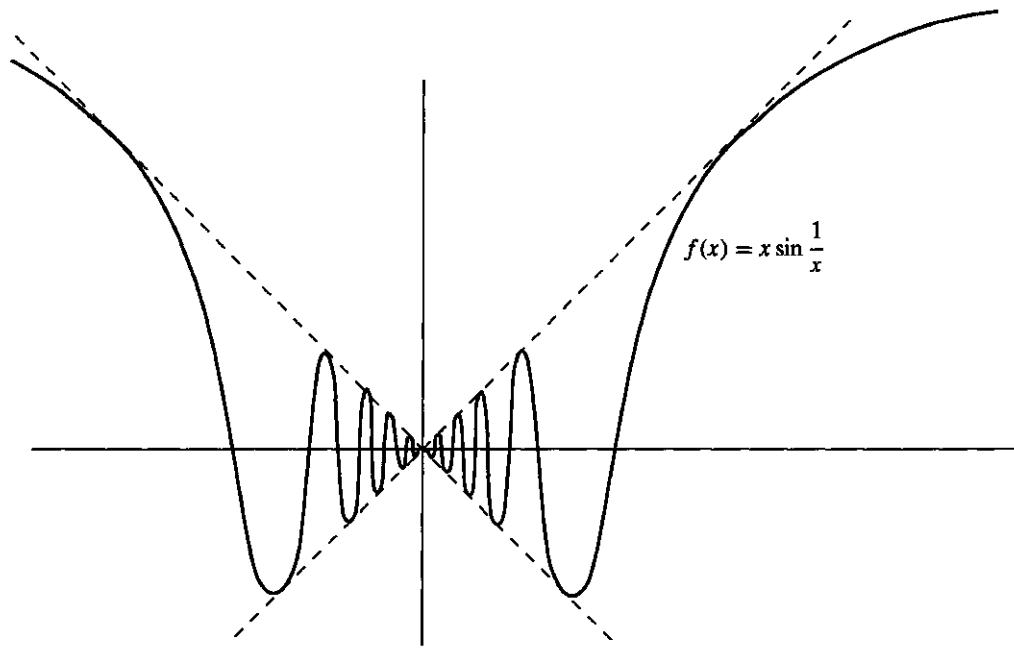


FIGURE 21

"large negative," that is, when  $|x|$  is large for negative  $x$ , again  $f(x)$  is close to 0, although  $f(x) < 0$ .

An interesting modification of this function is  $f(x) = x^2 \sin 1/x$ . The graph of this function is sketched in Figure 22. Since  $\sin 1/x$  oscillates infinitely often near 0 between 1 and -1, the function  $f(x) = x^2 \sin 1/x$  oscillates infinitely often between  $x^2$  and  $-x^2$ . The behavior of the graph for  $x$  large or large negative is harder to

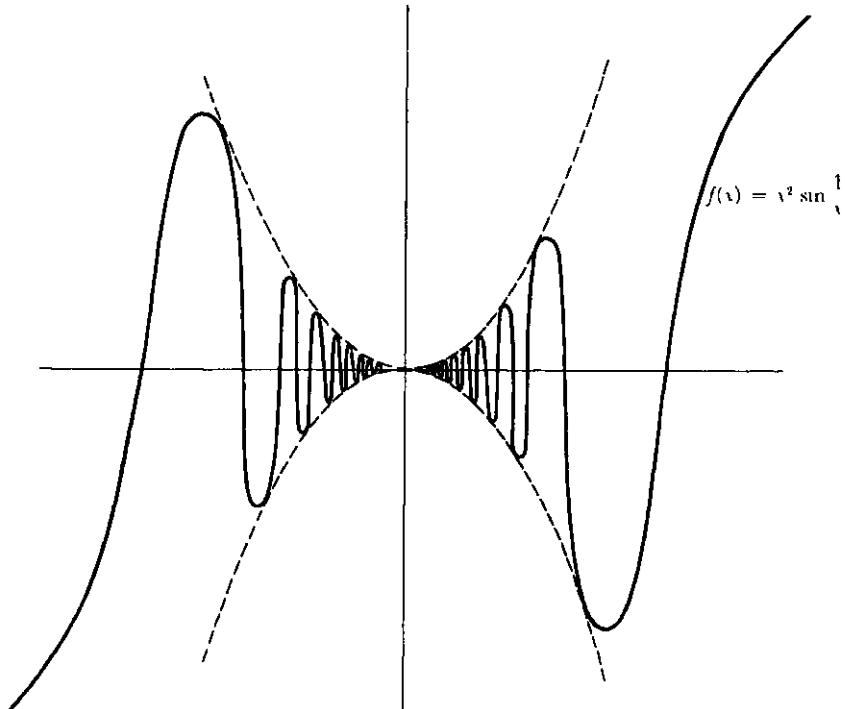


FIGURE 22

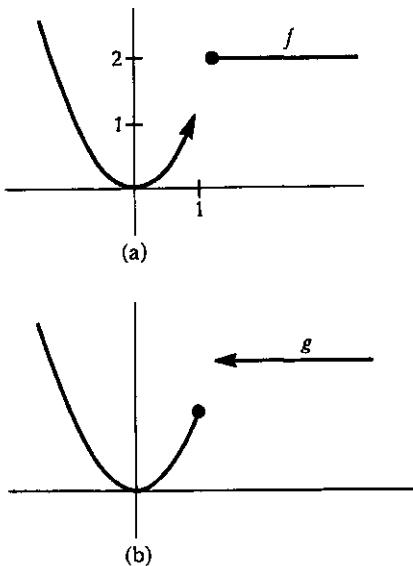


FIGURE 23

analyze. Since  $\sin 1/x$  is getting close to 0, while  $x$  is getting larger and larger, there seems to be no telling what the product will do. It is possible to decide, but this is another question that is best deferred to Part III. The graph of  $f(x) = x^2 \sin 1/x$  has also been illustrated (Figure 22).

For these infinitely oscillating functions, it is clear that the graph cannot hope to be really "accurate." The best we can do is to show part of it, and leave out the part near 0 (which is the interesting part). Actually, it is easy to find much simpler functions whose graphs cannot be "accurately" drawn. The graphs of

$$f(x) = \begin{cases} x^2, & x < 1 \\ 2, & x \geq 1 \end{cases} \quad \text{and} \quad g(x) = \begin{cases} x^2, & x \leq 1 \\ 2, & x > 1 \end{cases}$$

can only be distinguished by some convention similar to that used for open and closed intervals (Figure 23).

Our last example is a function whose graph is spectacularly nondrawable:

$$f(x) = \begin{cases} 0, & x \text{ irrational} \\ 1, & x \text{ rational.} \end{cases}$$

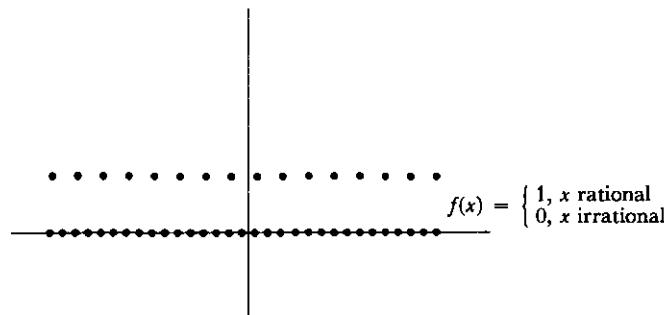


FIGURE 24

The graph of  $f$  must contain infinitely many points on the horizontal axis and also infinitely many points on a line parallel to the horizontal axis, but it must not contain either of these lines entirely. Figure 24 shows the usual textbook picture of the graph. To distinguish the two parts of the graph, the dots are placed closer together on the line corresponding to irrational  $x$ . (There is actually a mathematical reason behind this convention, but it depends on some sophisticated ideas, introduced in Problems 21-5 and 21-6.)

The peculiarities exhibited by some functions are so engrossing that it is easy to forget some of the simplest, and most important, subsets of the plane, which are not the graphs of functions. The most important example of all is the **circle**. A circle with center  $(a, b)$  and radius  $r > 0$  contains, by definition, all the points  $(x, y)$  whose distance from  $(a, b)$  is equal to  $r$ . The circle thus consists (Figure 25) of all points  $(x, y)$  with

$$\sqrt{(x - a)^2 + (y - b)^2} = r$$

or

$$(x - a)^2 + (y - b)^2 = r^2.$$

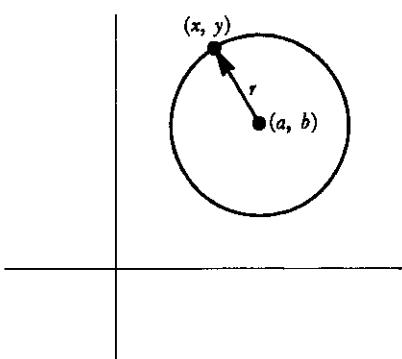


FIGURE 25

The circle with center  $(0,0)$  and radius 1, often regarded as a sort of standard copy, is called the *unit circle*.

A close relative of the circle is the **ellipse**. This is defined as the set of points, the *sum* of whose distances from two “focus” points is a constant. (When the two foci are the same, we obtain a circle.) If, for convenience, the focus points are taken to be  $(-c, 0)$  and  $(c, 0)$ , and the sum of the distances is taken to be  $2a$  (the factor 2 simplifies some algebra), then  $(x, y)$  is on the ellipse if and only if

$$\sqrt{(x - (-c))^2 + y^2} + \sqrt{(x - c)^2 + y^2} = 2a$$

or

$$\sqrt{(x + c)^2 + y^2} = 2a - \sqrt{(x - c)^2 + y^2}$$

or

$$x^2 + 2cx + c^2 + y^2 = 4a^2 - 4a\sqrt{(x - c)^2 + y^2} + x^2 - 2cx + c^2 + y^2$$

or

$$4(cx - a^2) = -4a\sqrt{(x - c)^2 + y^2}$$

or

$$c^2x^2 - 2cxa^2 + a^4 = a^2(x^2 - 2cx + c^2 + y^2)$$

or

$$(c^2 - a^2)x^2 - a^2y^2 = a^2(c^2 - a^2)$$

or

$$\frac{x^2}{a^2} + \frac{y^2}{a^2 - c^2} = 1.$$

This is usually written simply

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

where  $b = \sqrt{a^2 - c^2}$  (since we must clearly choose  $a > c$ , it follows that  $a^2 - c^2 > 0$ ). A picture of an ellipse is shown in Figure 26. The ellipse intersects the horizontal axis when  $y = 0$ , so that

$$\frac{x^2}{a^2} = 1, \quad x = \pm a,$$

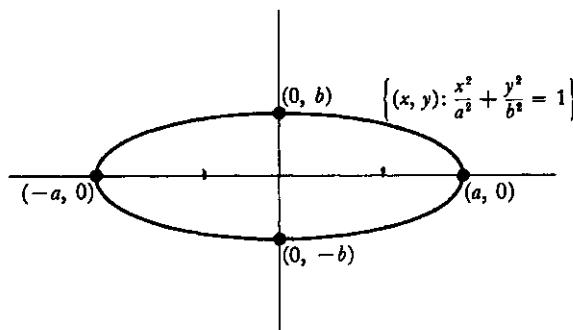


FIGURE 26

and it intersects the vertical axis when  $x = 0$ , so that

$$\frac{y^2}{b^2} = 1, \quad y = \pm b.$$

The **hyperbola** is defined analogously, except that we require the *difference* of the two distances to be constant. Choosing the points  $(-c, 0)$  and  $(c, 0)$  once again, and the constant difference as  $2a$ , we obtain, as the condition that  $(x, y)$  be on the hyperbola,

$$\sqrt{(x+c)^2 + y^2} - \sqrt{(x-c)^2 + y^2} = \pm 2a,$$

which may be simplified to

$$\frac{x^2}{a^2} - \frac{y^2}{a^2 - c^2} = 1.$$

In this case, however, we must clearly choose  $c > a$ , so that  $a^2 - c^2 < 0$ . If  $b = \sqrt{c^2 - a^2}$ , then  $(x, y)$  is on the hyperbola if and only if

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

The picture is shown in Figure 27. It contains two pieces, because the difference between the distances of  $(x, y)$  from  $(-c, 0)$  and  $(c, 0)$  may be taken in two different orders. The hyperbola intersects the horizontal axis when  $y = 0$ , so that  $x = \pm a$ , but it never intersects the vertical axis.

It is interesting to compare (Figure 28) the hyperbola with  $a = b = \sqrt{2}$  and the graph of the function  $f(x) = 1/x$ . The drawings look quite similar, and the two sets are actually identical, except for a rotation through an angle of  $45^\circ$  (Problem 23).

Clearly no rotation of the plane will change circles or ellipses into the graphs of functions. Nevertheless, the study of these important geometric figures can often be reduced to the study of functions. Ellipses, for example, are made up of the

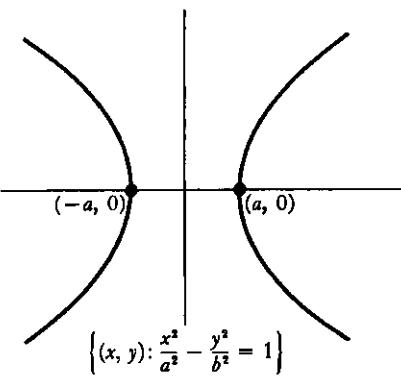


FIGURE 27

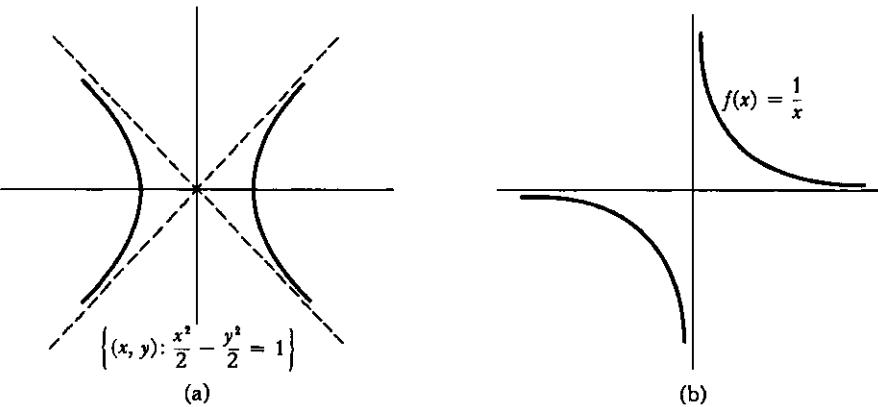


FIGURE 28

graphs of two functions,

$$f(x) = b \sqrt{1 - (x^2/a^2)}, \quad -a \leq x \leq a$$

and

$$g(x) = -b \sqrt{1 - (x^2/a^2)}, \quad -a \leq x \leq a.$$

Of course, there are many other pairs of functions with this same property. For example, we can take

$$f(x) = \begin{cases} b \sqrt{1 - (x^2/a^2)}, & 0 < x \leq a \\ -b \sqrt{1 - (x^2/a^2)}, & -a \leq x \leq 0 \end{cases}$$

and

$$g(x) = \begin{cases} -b \sqrt{1 - (x^2/a^2)}, & 0 < x \leq a \\ b \sqrt{1 - (x^2/a^2)}, & -a \leq x \leq 0. \end{cases}$$

We could also choose

$$f(x) = \begin{cases} b \sqrt{1 - (x^2/a^2)}, & x \text{ rational}, \quad -a \leq x \leq a \\ -b \sqrt{1 - (x^2/a^2)}, & x \text{ irrational}, \quad -a \leq x \leq a \end{cases}$$

and

$$g(x) = \begin{cases} -b \sqrt{1 - (x^2/a^2)}, & x \text{ rational}, \quad -a \leq x \leq a \\ b \sqrt{1 - (x^2/a^2)}, & x \text{ irrational}, \quad -a \leq x \leq a. \end{cases}$$

But all these other pairs necessarily involve unreasonable functions which jump around. A proof, or even a precise statement of this fact, is too difficult at present. Although you have probably already begun to make a distinction between those functions with reasonable graphs, and those with unreasonable graphs, you may find it very difficult to state a reasonable definition of reasonable functions. A mathematical definition of this concept is by no means easy, and a great deal of this book may be viewed as successive attempts to impose more and more conditions that a “reasonable” function must satisfy. As we define some of these conditions, we will take time out to ask if we have really succeeded in isolating the functions which deserve to be called reasonable. The answer, unfortunately, will always be “no,” or at best, a qualified “yes.”

## PROBLEMS

1. Indicate on a straight line the set of all  $x$  satisfying the following conditions. Also name each set, using the notation for intervals (in some cases you will also need the  $\cup$  sign).

- (i)  $|x - 3| < 1$ .
- (ii)  $|x - 3| \leq 1$ .
- (iii)  $|x - a| < \varepsilon$ .
- (iv)  $|x^2 - 1| < \frac{1}{2}$ .

- (v)  $\frac{1}{1+x^2} \geq \frac{1}{5}$ .
- (vi)  $\frac{1}{1+x^2} \leq a$  (give an answer in terms of  $a$ , distinguishing various cases).
- (vii)  $x^2 + 1 \geq 2$ .
- (viii)  $(x+1)(x-1)(x-2) > 0$ .
2. There is a very useful way of describing the points of the closed interval  $[a, b]$  (where we assume, as usual, that  $a < b$ ).
- First consider the interval  $[0, b]$ , for  $b > 0$ . Prove that if  $x$  is in  $[0, b]$ , then  $x = tb$  for some  $t$  with  $0 \leq t \leq 1$ . What is the significance of the number  $t$ ? What is the mid-point of the interval  $[0, b]$ ?
  - Now prove that if  $x$  is in  $[a, b]$ , then  $x = (1-t)a + tb$  for some  $t$  with  $0 \leq t \leq 1$ . Hint: This expression can also be written as  $a + t(b-a)$ . What is the midpoint of the interval  $[a, b]$ ? What is the point  $1/3$  of the way from  $a$  to  $b$ ?
  - Prove, conversely, that if  $0 \leq t \leq 1$ , then  $(1-t)a + tb$  is in  $[a, b]$ .
  - The points of the *open* interval  $(a, b)$  are those of the form  $(1-t)a + tb$  for  $0 < t < 1$ .
3. Draw the set of all points  $(x, y)$  satisfying the following conditions. (In most cases your picture will be a sizable portion of a plane, not just a line or curve.)
- $x > y$ .
  - $x + a > y + b$ .
  - $y < x^2$ .
  - $y \leq x^2$ .
  - $|x - y| < 1$ .
  - $|x + y| < 1$ .
  - $x + y$  is an integer.
  - $\frac{1}{x+y}$  is an integer.
  - $(x-1)^2 + (y-2)^2 < 1$ .
  - $x^2 < y < x^4$ .
4. Draw the set of all points  $(x, y)$  satisfying the following conditions:
- $|x| + |y| = 1$ .
  - $|x| - |y| = 1$ .
  - $|x - 1| = |y - 1|$ .
  - $|1-x| = |y-1|$ .
  - $x^2 + y^2 = 0$ .
  - $xy = 0$ .
  - $x^2 - 2x + y^2 = 4$ .
  - $x^2 = y^2$ .

5. Draw the set of all points  $(x, y)$  satisfying the following conditions:

- (i)  $x = y^2$ .
- (ii)  $\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1$ .
- (iii)  $x = |y|$ .
- (iv)  $x = \sin y$ .

Hint: You already know the answers when  $x$  and  $y$  are interchanged.

6. (a) Show that the straight line through  $(a, b)$  with slope  $m$  is the graph of the function  $f(x) = m(x - a) + b$ . This formula, known as the “point-slope form” is far more convenient than the equivalent expression  $f(x) = mx + (b - ma)$ ; it is immediately clear from the point-slope form that the slope is  $m$ , and that the value of  $f$  at  $a$  is  $b$ .  
 (b) For  $a \neq c$ , show that the straight line through  $(a, b)$  and  $(c, d)$  is the graph of the function

$$f(x) = \frac{d - b}{c - a}(x - a) + b.$$

- (c) When are the graphs of  $f(x) = mx + b$  and  $g(x) = m'x + b'$  parallel straight lines?

7. (a) For any numbers  $A$ ,  $B$ , and  $C$ , with  $A$  and  $B$  not both 0, show that the set of all  $(x, y)$  satisfying  $Ax + By + C = 0$  is a straight line (possibly a vertical one). Hint: First decide when a vertical straight line is described.  
 (b) Show conversely that every straight line, including vertical ones, can be described as the set of all  $(x, y)$  satisfying  $Ax + By + C = 0$ .

8. (a) Prove that the graphs of the functions

$$\begin{aligned} f(x) &= mx + b, \\ g(x) &= nx + c, \end{aligned}$$

are perpendicular if  $mn = -1$ , by computing the squares of the lengths of the sides of the triangle in Figure 29. (Why is this special case, where the lines intersect at the origin, as good as the general case?)

- (b) Prove that the two straight lines consisting of all  $(x, y)$  satisfying the conditions

$$\begin{aligned} Ax + By + C &= 0, \\ A'x + B'y + C' &= 0, \end{aligned}$$

are perpendicular if and only if  $AA' + BB' = 0$ .

9. (a) Prove, using Problem 1-19, that

$$\sqrt{(x_1 + y_1)^2 + (x_2 + y_2)^2} \leq \sqrt{x_1^2 + x_2^2} + \sqrt{y_1^2 + y_2^2}.$$

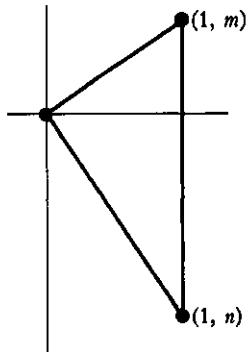


FIGURE 29

(b) Prove that

$$\sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2} \leq \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} + \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2}.$$

Interpret this inequality geometrically (it is called the “triangle inequality”). When does strict inequality hold?

10. Sketch the graphs of the following functions, plotting enough points to get a good idea of the general appearance. (Part of the problem is to make a reasonable decision how many is “enough”; the queries posed below are meant to show that a little thought will often be more valuable than hundreds of individual points.)
- (i)  $f(x) = x + \frac{1}{x}$ . (What happens for  $x$  near 0, and for large  $x$ ? Where does the graph lie in relation to the graph of the identity function? Why does it suffice to consider only positive  $x$  at first?)
  - (ii)  $f(x) = x - \frac{1}{x}$ .
  - (iii)  $f(x) = x^2 + \frac{1}{x^2}$ .
  - (iv)  $f(x) = x^2 - \frac{1}{x^2}$ .
11. Describe the general features of the graph of  $f$  if
- (i)  $f$  is even.
  - (ii)  $f$  is odd.
  - (iii)  $f$  is nonnegative.
  - (iv)  $f(x) = f(x + a)$  for all  $x$  (a function with this property is called **periodic**, with period  $a$ ).
12. Graph the functions  $f(x) = \sqrt[m]{x}$  for  $m = 1, 2, 3, 4$ . (There is an easy way to do this, using Figure 14. Be sure to remember, however, that  $\sqrt[m]{x}$  means the *positive*  $m$ th root of  $x$  when  $m$  is even; you should also note that there will be an important difference between the graphs when  $m$  is even and when  $m$  is odd.)
13. (a) Graph  $f(x) = |x|$  and  $f(x) = x^2$ .  
 (b) Graph  $f(x) = |\sin x|$  and  $f(x) = \sin^2 x$ . (There is an important difference between the graphs, which we cannot yet even describe rigorously. See if you can discover what it is; part (a) is meant to be a clue.)
14. Describe the graph of  $g$  in terms of the graph of  $f$  if

- (i)  $g(x) = f(x) + c.$
- (ii)  $g(x) = f(x + c).$  (It is easy to make a mistake here.)
- (iii)  $g(x) = cf(x).$
- (iv)  $g(x) = f(cx).$  (Distinguish the cases  $c = 0, c > 0, c < 0.$ )
- (v)  $g(x) = f(1/x).$
- (vi)  $g(x) = f(|x|).$
- (vii)  $g(x) = |f(x)|.$
- (viii)  $g(x) = \max(f, 0).$
- (ix)  $g(x) = \min(f, 0).$
- (x)  $g(x) = \max(f, 1).$

15. Draw the graph of  $f(x) = ax^2 + bx + c.$  Hint: Use the methods of Problem 1-18.
16. Suppose that  $A$  and  $C$  are not both 0. Show that the set of all  $(x, y)$  satisfying

$$Ax^2 + Bx + Cy^2 + Dy + E = 0$$

is either a parabola, an ellipse, or an hyperbola (or possibly  $\emptyset$ ). Hint: The case  $C = 0$  is essentially Problem 15, and the case  $A = 0$  is just a minor variant. Now consider separately the cases where  $A$  and  $B$  are both positive or negative, and where one is positive while the other is negative.

17. The symbol  $[x]$  denotes the largest integer which is  $\leq x.$  Thus,  $[2.1] = [2] = 2$  and  $[-0.9] = [-1] = -1.$  Draw the graph of the following functions (they are all quite interesting, and several will reappear frequently in other problems).

- (i)  $f(x) = [x].$
- (ii)  $f(x) = x - [x].$
- (iii)  $f(x) = \sqrt{x - [x]}.$
- (iv)  $f(x) = [x] + \sqrt{x - [x]}.$
- . (v)  $f(x) = \left[ \frac{1}{x} \right].$
- (vi)  $f(x) = \frac{1}{\left[ \frac{1}{x} \right]}.$

18. Graph the following functions.

- (i)  $f(x) = \{x\}$ , where  $\{x\}$  is defined to be the distance from  $x$  to the nearest integer.
- (ii)  $f(x) = \{2x\}.$
- (iii)  $f(x) = \{x\} + \frac{1}{2}\{2x\}.$
- (iv)  $f(x) = \{4x\}.$
- (v)  $f(x) = \{x\} + \frac{1}{2}\{2x\} + \frac{1}{4}\{4x\}.$

Many functions may be described in terms of the decimal expansion of a number. Although we will not be in a position to describe infinite decimals rigorously until Chapter 23, your intuitive notion of infinite decimals should suffice to carry you through the following problem, and others which occur before Chapter 23. There is one ambiguity about infinite decimals which must be eliminated: Every decimal ending in a string of 9's is equal to another ending in a string of 0's (e.g.,  $1.23999\dots = 1.24000\dots$ ). We will always use the one ending in 9's.

- \*19. Describe as best you can the graphs of the following functions (a complete picture is usually out of the question).

- (i)  $f(x) =$  the 1st number in the decimal expansion of  $x$ .
- (ii)  $f(x) =$  the 2nd number in the decimal expansion of  $x$ .
- (iii)  $f(x) =$  the number of 7's in the decimal expansion of  $x$  if this number is finite, and 0 otherwise.
- (iv)  $f(x) = 0$  if the number of 7's in the decimal expansion of  $x$  is finite, and 1 otherwise.
- (v)  $f(x) =$  the number obtained by replacing all digits in the decimal expansion of  $x$  which come after the first 7 (if any) by 0.
- (vi)  $f(x) = 0$  if 1 never appears in the decimal expansion of  $x$ , and  $n$  if 1 first appears in the  $n$ th place.

- \*20. Let

$$f(x) = \begin{cases} 0, & x \text{ irrational} \\ \frac{1}{q}, & x = \frac{p}{q} \text{ rational in lowest terms.} \end{cases}$$

(A number  $p/q$  is in **lowest terms** if  $p$  and  $q$  are integers with no common factor, and  $q > 0$ ). Draw the graph of  $f$  as well as you can (don't sprinkle points randomly on the paper; consider first the rational numbers with  $q = 2$ , then those with  $q = 3$ , etc.).

21. (a) The points on the graph of  $f(x) = x^2$  are the ones of the form  $(x, x^2)$ . Prove that each such point is equidistant from the point  $(0, \frac{1}{4})$  and the graph of  $g(x) = -\frac{1}{4}$ . (See Figure 30.)  
(b) Given a point  $P = (\alpha, \beta)$  and a horizontal line  $L$ , the graph of  $g(x) = \gamma$ , show that the set of all points  $(x, y)$  equidistant from  $P$  and  $L$  is the graph of a function of the form  $f(x) = ax^2 + bx + c$ .

- \*22. (a) Show that the square of the distance from  $(c, d)$  to  $(x, mx)$  is

$$x^2(m^2 + 1) + x(-2md - 2c) + d^2 + c^2.$$

Using Problem 1-18 to find the minimum of these numbers, show that the distance from  $(c, d)$  to the graph of  $f(x) = mx$  is

$$|cm - d|/\sqrt{m^2 + 1}.$$

- (b) Find the distance from  $(c, d)$  to the graph of  $f(x) = mx + b$ . (Reduce this case to part (a).)

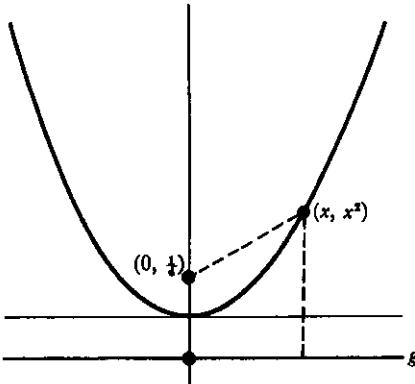


FIGURE 30

- \*23. (a) Using Problem 22, show that the numbers  $x'$  and  $y'$  indicated in Figure 31 are given by

$$x' = \frac{1}{\sqrt{2}} x + \frac{1}{\sqrt{2}} y,$$

$$y' = -\frac{1}{\sqrt{2}} x + \frac{1}{\sqrt{2}} y.$$

- (b) Show that the set of all  $(x, y)$  with  $(x'/\sqrt{2})^2 - (y'/\sqrt{2})^2 = 1$  is the same as the set of all  $(x, y)$  with  $xy = 1$ .

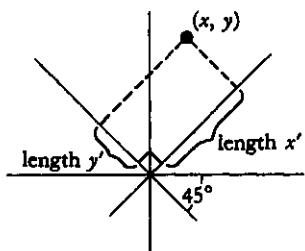


FIGURE 31

## APPENDIX 1. VECTORS

Suppose that  $v$  is a point in the plane; in other words,  $v$  is a pair of numbers

$$v = (v_1, v_2).$$

For convenience, we will use this convention that subscripts indicate the first and second pairs of a point that has been described by a single letter. Thus, if we mention the points  $w$  and  $z$ , it will be understood that  $w$  is the pair  $(w_1, w_2)$ , while  $z$  is the pair  $(z_1, z_2)$ .

Instead of the actual pair of numbers  $(v_1, v_2)$ , we often picture  $v$  as an arrow from the origin  $O$  to this point (Figure 1), and we refer to these arrows as *vectors* in the plane. Of course, we've haven't really said anything new yet, we've simply introduced an alternate term for a point of the plane, and another mental picture. The real point of the new terminology is to emphasize that we are going to do some new things with points in the plane.

For example, suppose that we have two vectors (i.e., points) in the plane,

$$v = (v_1, v_2), \quad w = (w_1, w_2).$$

Then we can define a new vector (a new point of the plane)  $v + w$  by the equation

$$(1) \quad v + w = (v_1 + w_1, v_2 + w_2).$$

Notice that all the letters on the right side of this equation are numbers, and the  $+$  sign is just our usual addition of numbers. On the other hand, the  $+$  sign on the left side is new: previously, the sum of two points in the plane wasn't defined, and we've simply used equation (1) as a *definition*.

A very fussy mathematician might want to use some new symbol for this newly defined operation, like

$$v + w, \quad \text{or perhaps} \quad v \oplus w,$$

but there's really no need to insist on this; since  $v + w$  hasn't been defined before, there's no possibility of confusion, so we might as well keep the notation simple.

Of course, any one can make new notation; for example, since it's our definition, we could just as well have defined  $v + w$  as  $(v_1 + w_1 \cdot w_2, v_2 + w_1^2)$ , or by some other equally weird formula. The real question is, does our new construction have any particular significance?

Figure 2 shows two vectors  $v$  and  $w$ , as well as the point

$$(v_1 + w_1, v_2 + w_2),$$

which, for the moment, we have simply indicated in the usual way, without drawing an arrow. Note that it is easy to compute the slope of the line  $L$  between  $v$  and our new point: as indicated in Figure 2, this slope is just

$$\frac{(v_2 + w_2) - v_2}{(v_1 + w_1) - v_1} = \frac{w_2}{w_1},$$

and this, of course, is the slope of our vector  $w$ , from the origin  $O$  to  $(w_1, w_2)$ . In other words, the line  $L$  is parallel to  $w$ .

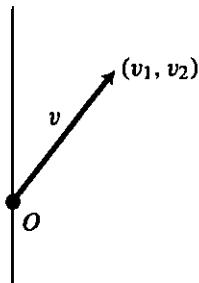


FIGURE 1

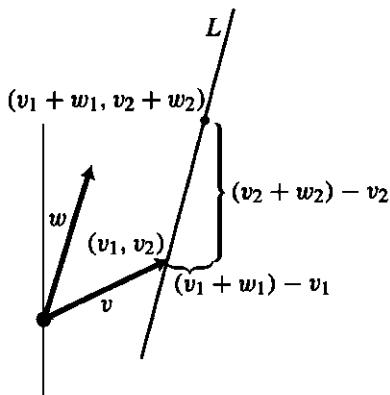


FIGURE 2

Similarly, the slope of the line  $M$  between  $(w_1, w_2)$  and our new point is

$$\frac{(v_2 + w_2) - w_2}{(v_1 + w_1) - v_2} = \frac{v_2}{v_1},$$

which is the slope of the vector  $v$ ; so  $M$  is parallel to  $v$ . In short, the new point  $v + w$  lies on the parallelogram having  $v$  and  $w$  as sides. When we draw  $v + w$  as an arrow (Figure 3), it points along the diagonal of this parallelogram. In physics, vectors are used to symbolize forces, and the sum of two vectors represents the resultant force when two different forces are applied simultaneously to the same object.

Figure 4 shows another way of visualizing the sum  $v + w$ . If we use “ $w$ ” to denote an arrow parallel to  $w$ , and having the same length, but starting at  $v$  instead of at the origin, then  $v + w$  is the vector from  $O$  to the final endpoint; thus we get to  $v + w$  by first following  $v$ , and then following “ $w$ .”

Many of the properties of  $+$  for ordinary numbers also hold for this new  $+$  for vectors. For example, the “commutative law”

$$v + w = w + v,$$

is obvious from the geometric picture, since the parallelogram spanned by  $v$  and  $w$  is the same as the parallelogram spanned by  $w$  and  $v$ . It is also easily checked analytically, since it states that

$$(v_1 + w_1, v_2 + w_2) = (w_1 + v_1, w_2 + v_2),$$

and thus simply depends on the commutative law for numbers:

$$\begin{aligned} v_1 + w_1 &= w_1 + v_1, \\ v_2 + w_2 &= w_2 + v_2. \end{aligned}$$

Similarly, unraveling definitions, we find the “associative law”

$$[v + w] + z = v + [w + z].$$

Figure 5 indicates a method of finding  $v + w + z$ .

The origin  $O = (0, 0)$  is an “additive identity,”

$$O + v = v + O = v,$$

and if we define

$$-v = (-v_1, -v_2),$$

then we also have

$$v + (-v) = -v + v = O.$$

Naturally we can also define

$$w - v = w + (-v),$$

exactly as with numbers; equivalently,

$$w - v = (w_1 - v_1, w_2 - v_2).$$

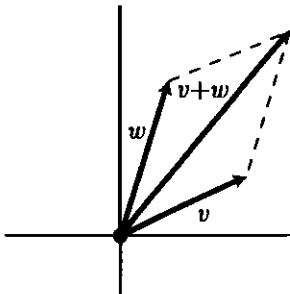


FIGURE 3

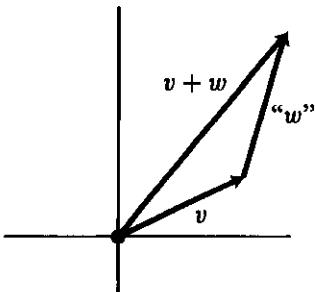


FIGURE 4

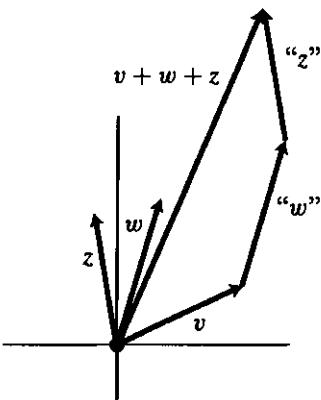


FIGURE 5

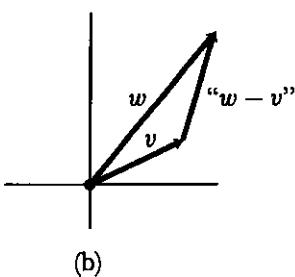
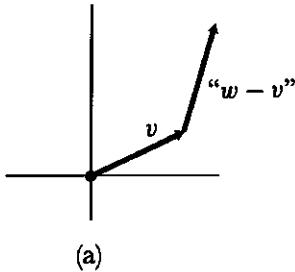


FIGURE 6

Just as with numbers, our definition of  $w - v$  simply means that it satisfies

$$v + (w - v) = w.$$

Figure 6(a) shows  $v$  and an arrow “ $w - v$ ” that is parallel to  $w - v$  but that starts at the endpoint of  $v$ . As we established with Figure 4, the vector from the origin to the endpoint of this arrow is just  $v + (w - v) = w$  (Figure 6(b)). In other words, we can picture  $w - v$  geometrically as the arrow that goes from  $v$  to  $w$  (except that it must then be moved back to the origin).

There is also a way of multiplying a number by a vector: For a number  $a$  and a vector  $v = (v_1, v_2)$ , we define

$$a \cdot v = (av_1, av_2)$$

(We sometimes simply write  $av$  instead of  $a \cdot v$ ; of course, it is then especially important to remember that  $v$  denotes a vector, rather than a number.) The vector  $a \cdot v$  points in the same direction as  $v$  when  $a > 0$  and in the opposite direction when  $a < 0$  (Figure 7).

You can easily check the following formulas:

$$\begin{aligned} a \cdot (b \cdot v) &= (ab) \cdot v, \\ 1 \cdot v &= v, \\ 0 \cdot v &= O, \\ -1 \cdot v &= -v. \end{aligned}$$

Notice that we have only defined a product of a number and a vector, we have not defined a way of ‘multiplying’ two vectors to get another vector.\* However, there are various ways of ‘multiplying’ vectors to get numbers, which are explored in the following problems.

### PROBLEMS

- Given a point  $v$  of the plane, let  $R_\theta(v)$  be the result of rotating  $v$  around the origin through an angle of  $\theta$  (Figure 8). The aim of this problem is to obtain a formula for  $R_\theta$ , with minimal calculation.

- Show that

$$\begin{aligned} R_\theta(1, 0) &= (\cos \theta, \sin \theta), \\ R_\theta(0, 1) &= (-\sin \theta, \cos \theta). \end{aligned}$$

- Explain why we have

$$\begin{aligned} R_\theta(v + w) &= R_\theta(v) + R_\theta(w), \\ R_\theta(a \cdot w) &= a \cdot R_\theta(w). \end{aligned}$$

- Now show that for any point  $(x, y)$  we have

$$R_\theta(x, y) = (x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta).$$

\*If you jump to Chapter 25, you’ll find that there is an important way of defining a product, but this is something very special for the plane—it doesn’t work for vectors in 3-space, for example, even though the other constructions do.

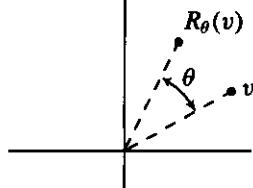


FIGURE 8

(d) Use this result to give another solution to Problem 4-23.

2. Given  $v$  and  $w$ , we define the *number*

$$v \cdot w = v_1 w_1 + v_2 w_2;$$

this is often called the ‘dot product’ or ‘scalar product’ of  $v$  and  $w$  (‘scalar’ being a rather old-fashioned word for a number, as opposed to a vector).

- (a) Given  $v$ , find a vector  $w$  such that  $v \cdot w = 0$ . Now describe the set of all such vectors  $w$ .  
 (b) Show that

$$\begin{aligned} v \cdot w &= w \cdot v \\ v \cdot (w + z) &= v \cdot w + v \cdot z \end{aligned}$$

and that

$$a \cdot (v \cdot w) = (a \cdot v) \cdot w = v \cdot (a \cdot w).$$

Notice that the last of these equations involves *three* products: the dot product  $\cdot$  of two vectors; the product  $\cdot$  of a number and a vector; and the ordinary product  $\cdot$  of two numbers.

- (c) Show that  $v \cdot v \geq 0$ , and that  $v \cdot v = 0$  only when  $v = O$ . Hence we can define the *norm*  $\|v\|$  as

$$\|v\| = \sqrt{v \cdot v},$$

which will be 0 only for  $v = O$ . What is the geometric interpretation of the norm?

- (d) Prove that

$$\|v + w\| \leq \|v\| + \|w\|,$$

and that equality holds if and only if  $v = 0$  or  $w = 0$  or  $w = a \cdot v$  for some number  $a > 0$ .

- (e) Show that

$$v \cdot w = \frac{\|v + w\|^2 - \|v - w\|^2}{4}.$$

3. (a) Let  $R_\theta$  be rotation by an angle of  $\theta$  (Problem 1). Show that

$$R_\theta(v) \cdot R_\theta(w) = v \cdot w.$$

- (b) Let  $e = (1, 0)$  be the vector of length 1 pointing along the first axis, and let  $w = (\cos \theta, \sin \theta)$ ; this is a vector of length 1 that makes an angle of  $\theta$  with the first axis (compare Problem 1). Calculate that

$$e \cdot w = \cos \theta.$$

Conclude that in general

$$v \cdot w = \|v\| \cdot \|w\| \cdot \cos \theta,$$

where  $\theta$  is the angle between  $v$  and  $w$ .

4. Given two vectors  $v$  and  $w$ , we'd expect to have a simple formula, involving the coordinates  $v_1, v_2, w_1, w_2$ , for the area of the parallelogram they span. Figure 9 indicates a strategy for finding such a formula: since the triangle with vertices  $w, A, v + w$  is congruent to the triangle  $OBv$ , we can reduce the problem to an easier one where one side of the parallelogram lies along the horizontal axis:

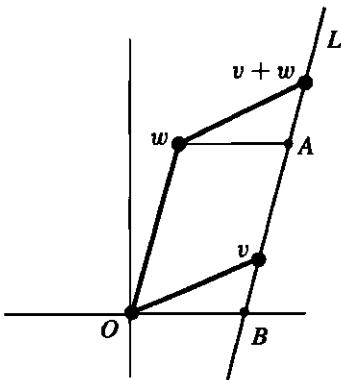


FIGURE 9

- (a) The line  $L$  passes through  $v$  and is parallel to  $w$ , so has slope  $w_2/w_1$ . Conclude that the point  $B$  has coordinate

$$\frac{v_1 w_2 - w_1 v_2}{w_2},$$

and that the parallelogram therefore has area

$$\det(v, w) = v_1 w_2 - w_1 v_2.$$

This formula, which defines the *determinant*  $\det$ , certainly seems to be simple enough, but it can't really be true that  $\det(v, w)$  always gives the area. After all, we clearly have

$$\det(w, v) = -\det(v, w),$$

so sometimes  $\det$  will be negative! Indeed, it is easy to see that our "derivation" made all sorts of assumptions (that  $w_2$  was positive, that  $B$  had a positive coordinate, etc.) Nevertheless, it seems likely that  $\det(v, w)$  is  $\pm$  the area; the next problem gives an independent proof.

5. (a) If  $v$  points along the positive horizontal axis, show that  $\det(v, w)$  is the area of the parallelogram spanned by  $v$  and  $w$  for  $w$  above the horizontal axis ( $w_2 > 0$ ), and the negative of the area for  $w$  below this axis.  
 (b) If  $R_\theta$  is rotation by an angle of  $\theta$  (Problem 1), show that

$$\det(R_\theta v, R_\theta w) = \det(v, w).$$

Conclude that  $\det(v, w)$  is the area of the parallelogram spanned by  $v$  and  $w$  when the rotation from  $v$  to  $w$  is counterclockwise, and the negative of the area when it is clockwise.

6. Show that

$$\begin{aligned}\det(v, w + z) &= \det(v, w) + \det(v, z) \\ \det(v + w, z) &= \det(v, z) + \det(w, z)\end{aligned}$$

and that

$$a \det(v, w) = \det(a \cdot v, w) = \det(v, a \cdot w).$$

7. Using the method of Problem 3, show that

$$\det(v, w) = \|v\| \cdot \|w\| \cdot \sin \theta,$$

which is also obvious from the geometric interpretation (Figure 10).

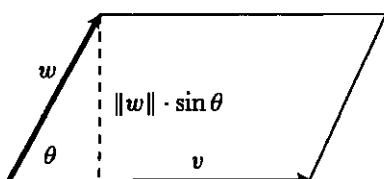


FIGURE 10

## APPENDIX 2. THE CONIC SECTIONS

Although we will be concerned almost exclusively with figures in the plane, defined formally as the set of all pairs of real numbers, in this Appendix we want to consider three-dimensional space, which we can describe in terms of triples of real numbers, using a “three-dimensional coordinate system,” consisting of three straight lines intersecting at right angles (Figure 1). Our *horizontal* and *vertical* axes now mutate to two axes in a horizontal plane, with the third axis perpendicular to both.

One of the simplest subsets of this three-dimensional space is the (infinite) *cone* illustrated in Figure 2; this cone may be produced by rotating a “generating line,” of slope  $C$  say, around the third axis.

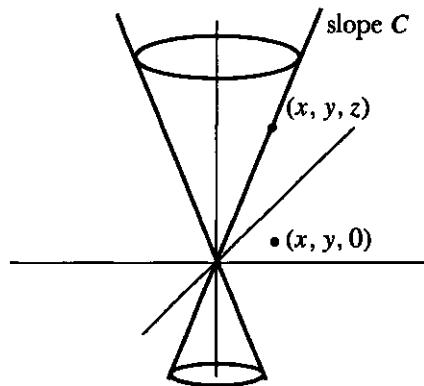


FIGURE 1

FIGURE 2

For any given first two coordinates  $x$  and  $y$ , the point  $(x, y, 0)$  in the horizontal plane has distance  $\sqrt{x^2 + y^2}$  from the origin, and thus

$$(1) \quad (x, y, z) \text{ is in the cone if and only if } z = \pm C\sqrt{x^2 + y^2}.$$

We can descend from these three-dimensional vistas to the more familiar two-dimensional one by asking what happens when we intersect this cone with some plane  $P$  (Figure 3).

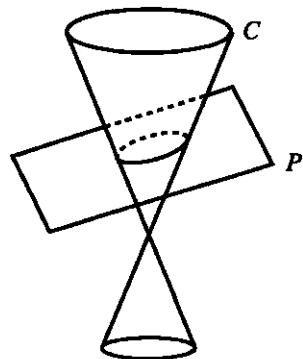


FIGURE 3

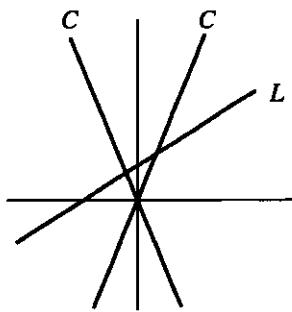


FIGURE 4

If the plane is parallel to the horizontal plane, there's certainly no mystery—the intersection is just a circle. Otherwise, the plane  $P$  intersects the horizontal plane in a straight line. We can make things a lot simpler for ourselves if we rotate everything so that this intersection line points straight out from the plane of the paper, while the first axis is in the usual position that we are familiar with. The plane  $P$  is thus viewed "straight on," so that all we see (Figure 4) is its intersection  $L$  with the plane of the first and third axes; from this view-point the cone itself simply appears as two straight lines.

In the plane of the first and third axes, the line  $L$  can be described as the collection of all points of the form

$$(x, Mx + B),$$

where  $M$  is the slope of  $L$ . For an arbitrary point  $(x, y, z)$  it follows that

$$(2) \quad (x, y, z) \text{ is in the plane } P \text{ if and only if } z = Mx + B.$$

Combining (1) and (2), we see that  $(x, y, z)$  is in the intersection of the cone and the plane if and only if

$$(*) \quad Mx + B = \pm C\sqrt{x^2 + y^2}.$$

Now we have to choose coordinate axes in the plane  $P$ . We can choose  $L$  as the first axis, measuring distances from the intersection  $Q$  with the horizontal plane (Figure 5); for the second axis we just choose the line through  $Q$  parallel to our original second axis. If the first coordinate of a point in  $P$  with respect to these axes is  $x$ , then the first coordinate of this point with respect to the original axes can be written in the form

$$\alpha x + \beta$$

for some  $\alpha$  and  $\beta$ . On the other hand, if the second coordinate of the point with respect to these axes is  $y$ , then  $y$  is also the second coordinate with respect to the original axes.

Consequently,  $(*)$  says that the point lies on the intersection of the plane and the cone if and only if

$$M(\alpha x + \beta) + B = \pm C\sqrt{(\alpha x + \beta)^2 + y^2}.$$

Although this looks fairly complicated, after squaring we can write this as

$$\alpha^2 C^2 y^2 + \alpha^2(M^2 - A^2)x^2 + Ex + F = 0$$

for some  $E$  and  $F$  that we won't bother writing out. Dividing by  $\alpha^2$  simplifies this to

$$C^2 y^2 + (C^2 - M^2)x^2 + Gx + H = 0.$$

Now Problem 4-16 indicates that this is either a parabola, an ellipse, or an hyperbola. In fact, looking a little more closely at the solution (and interchanging

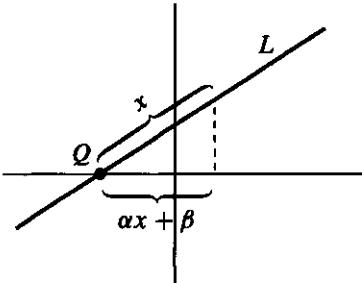


FIGURE 5

the roles of  $x$  and  $y$ ), we see that the values of  $G$  and  $H$  are irrelevant:

- (1) If  $M = \pm C$  we obtain a parabola;
- (2) If  $C^2 > M^2$  we obtain an ellipse;
- (3) If  $C^2 < M^2$  we obtain an hyperbola.

These analytic conditions are easy to interpret geometrically (Figure 6):

- (1) If our plane is parallel to one of the generating lines of the cone we obtain a parabola;
- (2) If our plane slopes less than the generating line of the cone (so that our intersection omits one half of the cone) we obtain an ellipse;
- (3) If our plane slopes more than the generating line of the cone we obtain an hyperbola.

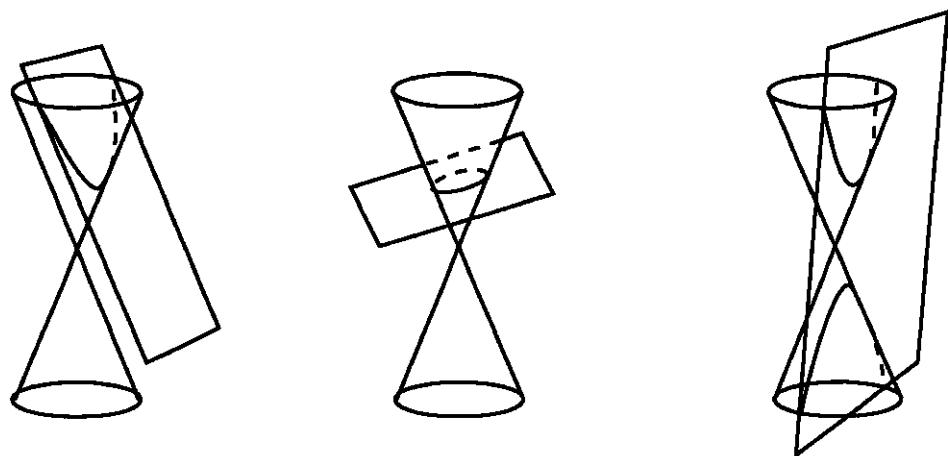


FIGURE 6

In fact, the very names of these “conic sections” are related to this description. The word *parabola* comes from a Greek root meaning ‘alongside,’ the same root that appears in *parable*, not to mention *paradigm*, *paradox*, *paragon*, *paragraph*, *paralegal*, *parallax*, *parallel*, even *parachute*. *Ellipse* comes from a Greek root meaning ‘defect,’ or omission, as in *ellipsis* (an omission, ... or the dots that indicate it). And *hyperbola* comes from a Greek root meaning ‘throwing beyond,’ or excess. With the currency of words like *hyperactive*, *hypersensitive*, and *hyperventilate*, not to mention *hype*, one can probably say, without risk of *hyperbole*, that this root is familiar to almost everyone.\*

### PROBLEMS

1. Consider a cylinder with a generator perpendicular to the horizontal plane (Figure 7); the only requirement for a point  $(x, y, z)$  to lie on this cylinder is

\*Although the correspondence between these roots and the geometric picture correspond so beautifully, for the sake of dull accuracy it has to be reported that the Greeks originally applied the words to describe features of certain equations involving the conic sections.

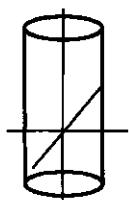


FIGURE 7

that  $(x, y)$  lies on a circle:

$$x^2 + y^2 = C^2.$$

Show that the intersection of a plane with this cylinder can be described by an equation of the form

$$(\alpha x + \beta)^2 + y^2 = C^2.$$

What possibilities are there?

2. In Figure 8, the sphere  $S_1$  has the same diameter as the cylinder, so that its equator  $C_1$  lies along the cylinder; it is also tangent to the plane  $P$  at  $F_1$ . Similarly, the equator  $C_2$  of  $S_2$  lies along the cylinder, and  $S_2$  is tangent to  $P$  at  $F_2$ .
  - (a) Let  $z$  be any point on the intersection of  $P$  and the cylinder. Explain why the length of the line from  $z$  to  $F_1$  is equal to the length of the vertical line  $L$  from  $z$  to  $C_1$ .
  - (b) By proving a similar fact for the length of the line from  $z$  to  $F_2$ , show that the distance from  $z$  to  $F_1$  plus the distance from  $z$  to  $F_2$  is a constant, so that the intersection is an ellipse, with foci  $F_1$  and  $F_2$ .
3. Similarly, use Figure 9 to prove geometrically that the intersection of a plane and a cone is an ellipse.

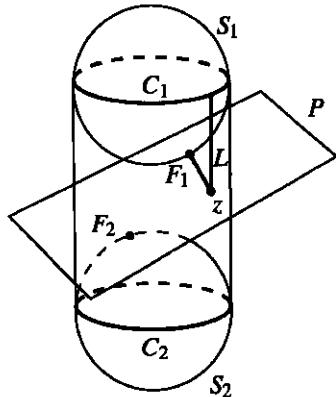


FIGURE 8

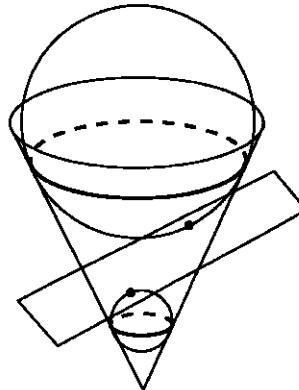


FIGURE 9

## APPENDIX 3. POLAR COORDINATES

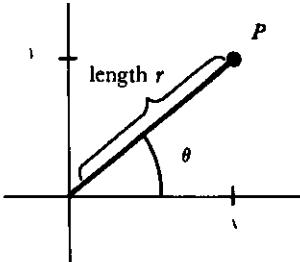
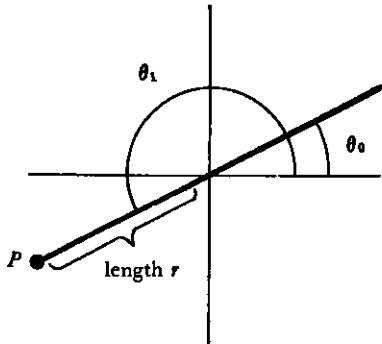


FIGURE 1

In this chapter we've been acting all along as if there's only one way to label points in the plane with pairs of numbers. Actually, there are many different ways, each giving rise to a different "coordinate system." The usual coordinates of a point are called its cartesian coordinates, after the French mathematician and philosopher René Descartes (1596–1650), who first introduced the idea of coordinate systems. In many situations it is more convenient to introduce polar coordinates, which are illustrated in Figure 1. To the point  $P$  we assign the polar coordinates  $(r, \theta)$ , where  $r$  is the distance from the origin  $O$  to  $P$ , and  $\theta$  is the angle between the horizontal axis and the line from  $O$  to  $P$ . This angle can be measured either in degrees or in radians (Chapter 15), but in either case  $\theta$  is not determined unambiguously. For example, with degree measurement, points on the right side of the horizontal axis could have either  $\theta = 0$  or  $\theta = 360$ ; moreover,  $\theta$  is completely ambiguous at the origin  $O$ . So it is necessary to exclude some ray through the origin if we want to assign a unique pair  $(r, \theta)$  to each point under consideration.

On the other hand, there is no problem associating a unique point to any pair  $(r, \theta)$ . In fact, we can even associate a point to  $(r, \theta)$  when  $r < 0$ , according to the scheme indicated in Figure 2. Thus, it always makes sense to talk about "the point with polar coordinates  $(r, \theta)$ ," even though there is some ambiguity when we talk about "the polar coordinates" of a given point.



$P$  is the point with polar coordinates  $(r, \theta_1)$  and also the point with polar coordinates  $(-r, \theta_0)$ .

FIGURE 2

It is clear from Figure 1 (and Figure 2) that the point with polar coordinates  $(r, \theta)$  has cartesian coordinates  $(x, y)$  given by

$$x = r \cos \theta, \quad y = r \sin \theta.$$

Conversely, if a point has cartesian coordinates  $(x, y)$ , then (any of) its polar coordinates  $(r, \theta)$  satisfy

$$r = \pm\sqrt{x^2 + y^2}$$

$$\tan \theta = \frac{y}{x} \quad \text{if } x \neq 0.$$

Now suppose that  $f$  is a function. Then by the **graph of  $f$  in polar coordinates** we mean the collection of all points  $P$  with polar coordinates  $(r, \theta)$  satisfying  $r = f(\theta)$ . In other words, the graph of  $f$  in polar coordinates is the collection of all points with polar coordinates  $(f(\theta), \theta)$ . No special significance should be attached to the fact that we are considering pairs  $(f(\theta), \theta)$ , with  $f(\theta)$  first, as opposed to pairs  $(x, f(x))$  in the usual graph of  $f$ ; it is purely a matter of convention that  $r$  is considered the first polar coordinate and  $\theta$  is considered the second.

The graph of  $f$  in polar coordinates is often described as “the graph of the equation  $r = f(\theta)$ .” For example, suppose that  $f$  is a constant function,  $f(\theta) = a$  for all  $\theta$ . The graph of the equation  $r = a$  is simply a circle with center  $O$  and radius  $a$  (Figure 3). This example illustrates, in a rather blatant way, that polar coordinates are likely to make things simpler in situations that involve symmetry with respect to the origin  $O$ .

The graph of the equation  $r = \theta$  is shown in Figure 4. The solid line corresponds to all values of  $\theta \geq 0$ , while the dashed line corresponds to values of  $\theta \leq 0$ .

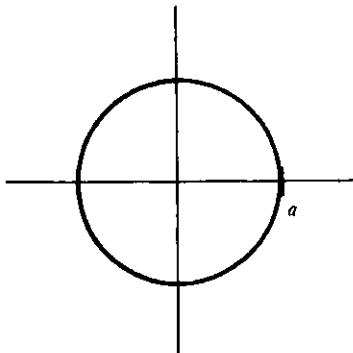


FIGURE 3

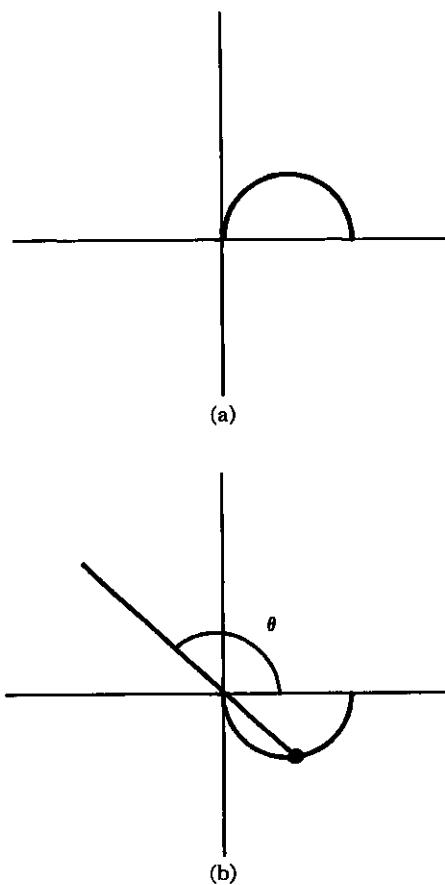


FIGURE 5

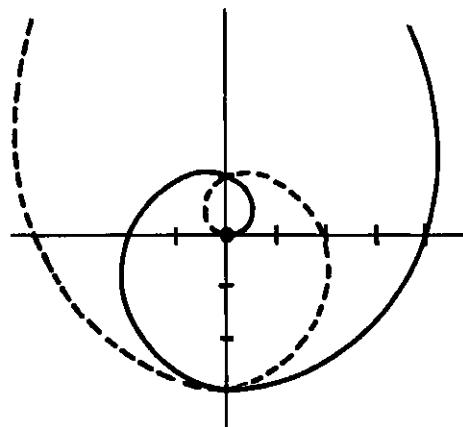


FIGURE 4 Spiral of Archimedes

As another example involving both positive and negative  $r$ , consider the graph of the equation  $r = \cos \theta$ . Figure 5(a) shows the part that corresponds to  $0 \leq \theta \leq 90^\circ$  [with  $\theta$  in degrees]. Figure 5(b) shows the part corresponding to  $90^\circ \leq \theta \leq 180^\circ$ ; here  $r < 0$ . You can check that no new points are added for  $\theta > 180^\circ$  or  $\theta < 0$ . It is easy to describe this same graph in terms of the cartesian coordinates of its points. Since the polar coordinates of any point on the graph satisfy

$$r = \cos \theta,$$

and hence

$$r^2 = r \cos \theta,$$

its cartesian coordinates satisfy the equation

$$x^2 + y^2 = x$$

which describes a circle (Problem 3-16). [Conversely, it is clear that if the cartesian coordinates of a point satisfy  $x^2 + y^2 = x$ , then it lies on the graph of the equation  $r = \cos \theta$ .]

Although we've now gotten a circle in two different ways, we might well be hesitant about trying to find the equation of an ellipse in polar coordinates. But it turns out that we can get a very nice equation if we choose one of the *foci* as the origin. Figure 6 shows an ellipse with one focus at  $O$ , with the sum of the distances of all points from  $O$  and the other focus  $\mathbf{f}$  being  $2a$ . We've chosen  $\mathbf{f}$  to the left of  $O$ , with coordinates written as

$$(-2\epsilon a, 0).$$

(We have  $0 \leq \epsilon < 1$ , since we must have  $2a >$  distance from  $\mathbf{f}$  to  $O$ ).

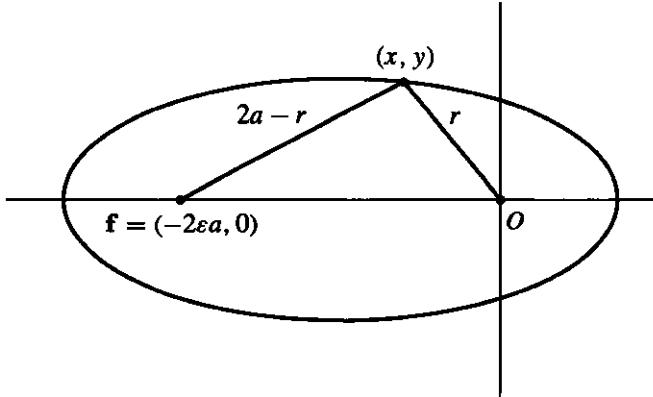


FIGURE 6

The distance  $r$  from  $(x, y)$  to  $O$  is given by

$$(1) \quad r^2 = x^2 + y^2.$$

By assumption, the distance from  $(x, y)$  to  $\mathbf{f}$  is  $2a - r$ , hence

$$(2a - r)^2 = (x - [-2\epsilon a])^2 + y^2,$$

or

$$(2) \quad 4a^2 - 4ar + r^2 = x^2 + 4\epsilon ax + 4\epsilon^2 a^2 + y^2.$$

Subtracting (1) from (2), and dividing by  $4a$ , we get

$$a - r = \epsilon x + \epsilon^2 a,$$

or

$$\begin{aligned} r &= a - \varepsilon x - \varepsilon^2 a \\ &= (1 - \varepsilon^2)a - \varepsilon x, \end{aligned}$$

which we can write as

$$(3) \quad r = \Lambda - \varepsilon x, \quad \text{for } \Lambda = (1 - \varepsilon^2)a.$$

Substituting  $r \cos \theta$  for  $x$ , we have

$$\begin{aligned} r &= \Lambda - \varepsilon r \cos \theta, \\ r(1 + \varepsilon \cos \theta) &= \Lambda, \end{aligned}$$

and thus

$$(4) \quad r = \frac{\Lambda}{1 + \varepsilon \cos \theta}.$$

In Chapter 4 we found that

$$(5) \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

is the equation in cartesian coordinates for an ellipse with  $2a$  as the sum of the distances to the foci, but with the foci at  $(-c, 0)$  and  $(c, 0)$ , where

$$b = \sqrt{a^2 - c^2}.$$

Since the distance between the foci is  $2c$ , this corresponds to the ellipse (4) when we take  $c = \varepsilon a$  or  $\varepsilon = c/a$  (with equation (3) determining  $\Lambda$ ). Conversely, given the ellipse described by (4), for the corresponding equation (5) the value of  $a$  is determined by (3),

$$a = \frac{\Lambda}{1 - \varepsilon^2},$$

and again using  $c = \varepsilon a$ , we get

$$b = \sqrt{a^2 - c^2} = \sqrt{a^2 - \varepsilon^2 a^2} = a\sqrt{1 - \varepsilon^2} = \frac{\Lambda}{\sqrt{1 - \varepsilon^2}}.$$

Thus, we can obtain  $a$  and  $b$ , the lengths of the major and minor axes, immediately from  $\varepsilon$  and  $\Lambda$ .

The number

$$\varepsilon = \frac{c}{a} = \frac{\sqrt{a^2 - b^2}}{a} = \sqrt{1 - \left(\frac{b}{a}\right)^2},$$

the *eccentricity* of the ellipse, determines the “shape” of the ellipse (the ratio of the major and minor axes), while the number  $\Lambda$  determines its “size,” as shown by (4).

## PROBLEMS

1. If two points have polar coordinates  $(r_1, \theta_1)$  and  $(r_2, \theta_2)$ , show that the distance  $d$  between them is given by

$$d^2 = r_1^2 + r_2^2 - 2r_1r_2 \cos(\theta_1 - \theta_2).$$

What does this say geometrically?

2. Describe the general features of the graph of  $f$  in polar coordinates if

- (i)  $f$  is even.
- (ii)  $f$  is odd.
- (iii)  $f(\theta) = f(\theta + 180)$  [when  $\theta$  is measured in degrees].

3. Sketch the graphs of the following equations.

- (i)  $r = a \sin \theta$ .
- (ii)  $r = a \sec \theta$ . Hint: It is a very simple graph!
- (iii)  $r = \cos 2\theta$ . Good luck on this one!
- (iv)  $r = \cos 3\theta$ .
- (v)  $r = |\cos 2\theta|$ .
- (vi)  $r = |\cos 3\theta|$ .

4. Find equations for the cartesian coordinates of points on the graphs (i), (ii) and (iii) in Problem 3.

5. Consider a hyperbola, where the difference of the distance between the two foci is the constant  $2a$ , and choose one focus at  $O$  and the other at  $(-2\epsilon a, 0)$ . (In this case, we must have  $\epsilon > 1$ ). Show that we obtain the exact same equation in polar coordinates

$$r = \frac{\Lambda}{1 + \epsilon \cos \theta}$$

as we obtained for an ellipse.

6. Consider the set of points  $(x, y)$  such that the distance  $(x, y)$  to  $O$  is equal to the distance from  $(x, y)$  to the line  $y = a$  (Figure 7). Show that the distance to the line is  $a - r \cos \theta$ , and conclude that the equation can be written

$$a = r(1 + \cos \theta).$$

Notice that this equation for a parabola is again of the same form as (4).

7. Now, for any  $\Lambda$  and  $\epsilon$ , consider the graph in polar coordinates of the equation (4), which implies (3). Show that the points satisfying this equation satisfy

$$(1 - \epsilon^2)x^2 + y^2 = \Lambda^2 - 2\Lambda\epsilon x.$$

Using Problem 4-16, show that this is an ellipse for  $\epsilon < 1$ , a parabola for  $\epsilon = 1$ , and a hyperbola for  $\epsilon > 1$ .

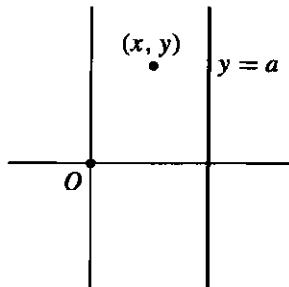


FIGURE 7

8. (a) Sketch the graph of the *cardioid*  $r = 1 - \sin \theta$ .  
 (b) Show that it is also the graph of  $r = -1 - \sin \theta$ .  
 (c) Show that it can be described by the equation

$$x^2 + y^2 = \sqrt{x^2 + y^2} - y,$$

and conclude that it can be described by the equation

$$(x^2 + y^2 + y)^2 = x^2 + y^2$$

9. Sketch the graphs of the following equations.

- (i)  $r = 1 - \frac{1}{2} \sin \theta$ .
- (ii)  $r = 1 - 2 \sin \theta$ .
- (iii)  $r = 2 + \cos \theta$ .

10. (a) Sketch the graph of the *lemniscate*

$$r^2 = 2a^2 \cos 2\theta.$$

- (b) Find an equation for its cartesian coordinates.
- (c) Show that it is the collection of all points  $P$  in Figure 8 satisfying  $d_1 d_2 = a^2$ .
- (d) Make a guess about the shape of the curves formed by the set of all  $P$  satisfying  $d_1 d_2 = b$ , when  $b > a^2$  and when  $b < a^2$ .

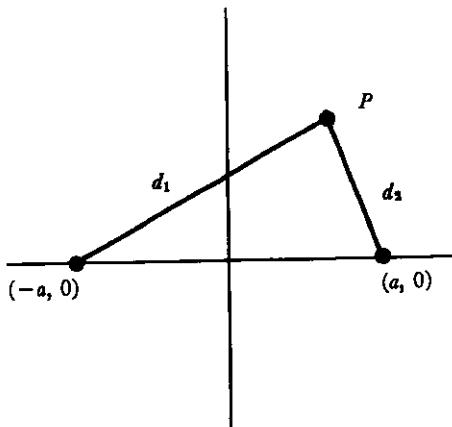


FIGURE 8

# CHAPTER 5

# LIMITS

The concept of a limit is surely the most important, and probably the most difficult one in all of calculus. The goal of this chapter is the definition of limits, but we are, once more, going to begin with a provisional definition; what we shall define is not the word “limit” but the notion of a function approaching a limit.

### PROVISIONAL DEFINITION

The function  $f$  approaches the limit  $l$  near  $a$ , if we can make  $f(x)$  as close as we like to  $l$  by requiring that  $x$  be sufficiently close to, but unequal to,  $a$ .

Of the six functions graphed in Figure 1, only the first three approach  $l$  at  $a$ . Notice that although  $g(a)$  is not defined, and  $h(a)$  is defined “the wrong way,” it is still true that  $g$  and  $h$  approach  $l$  near  $a$ . This is because we explicitly ruled out, in our definition, the necessity of ever considering the value of the function at  $a$ —it is only necessary that  $f(x)$  should be close to  $l$  for  $x$  close to  $a$ , but *unequal to*  $a$ . We are simply not interested in the value of  $f(a)$ , or even in the question of whether  $f(a)$  is defined.

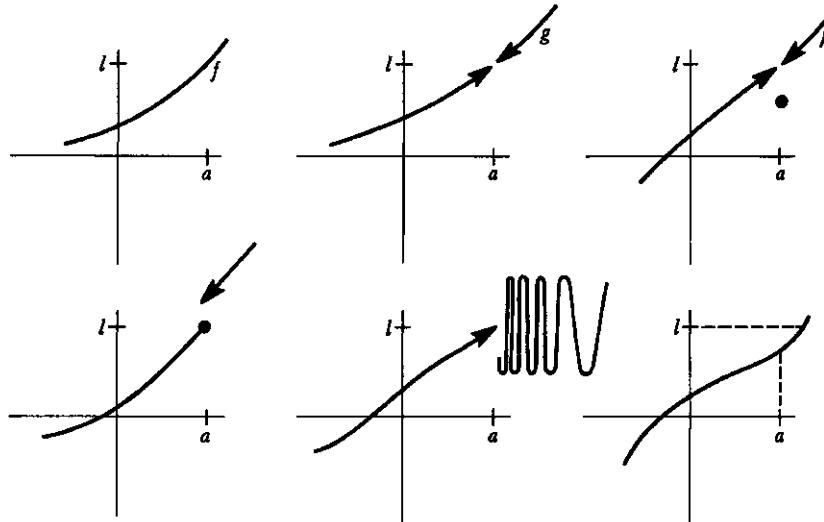


FIGURE 1

One convenient way of picturing the assertion that  $f$  approaches  $l$  near  $a$  is provided by a method of drawing functions that was not mentioned in Chapter 4. In this method, we draw two straight lines, each representing  $\mathbf{R}$ , and arrows from a point  $x$  in one, to  $f(x)$  in the other. Figure 2 illustrates such a picture for two different functions.

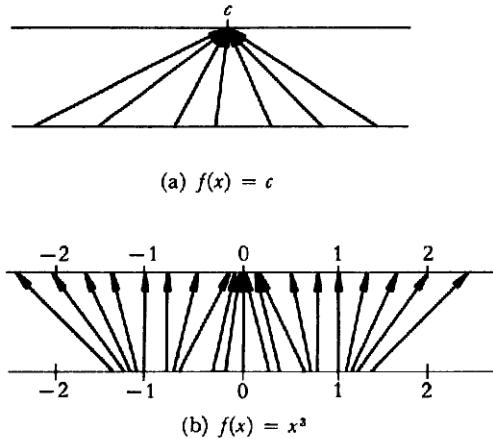


FIGURE 2

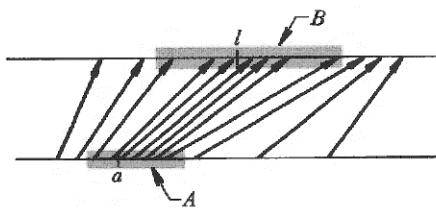


FIGURE 3

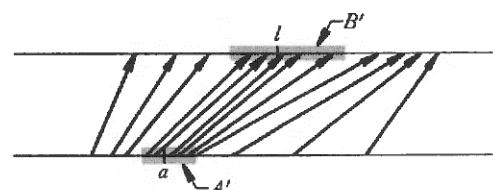


FIGURE 4

Now consider a function  $f$  whose drawing looks like Figure 3. Suppose we ask that  $f(x)$  be close to  $l$ , say within the open interval  $B$  which has been drawn in Figure 3. This can be guaranteed if we consider only the numbers  $x$  in the interval  $A$  of Figure 3. (In this diagram we have chosen the largest interval which will work; any smaller interval containing  $a$  could have been chosen instead.) If we choose a smaller interval  $B'$  (Figure 4) we will, usually, have to choose a smaller  $A'$ , but no matter how small we choose the open interval  $B$ , there is always supposed to be some open interval  $A$  which works.

A similar pictorial interpretation is possible in terms of the graph of  $f$ , but in this case the interval  $B$  must be drawn on the vertical axis, and the set  $A$  on the horizontal axis. The fact that  $f(x)$  is in  $B$  when  $x$  is in  $A$  means that the part of the graph lying over  $A$  is contained in the region which is bounded by the horizontal lines through the end points of  $B$ ; compare Figure 5(a), where a valid interval  $A$  has been chosen, with Figure 5(b), where  $A$  is too large.

In order to apply our definition to a particular function, let us consider  $f(x) = x \sin 1/x$  (Figure 6). Despite the erratic behavior of this function near 0 it is clear, at least intuitively, that  $f$  approaches 0 near 0, and it is certainly to be hoped that our definition will allow us to reach the same conclusion. In the case we are considering, both  $a$  and  $l$  of the definition are 0, so we must ask if we can get  $f(x) = x \sin 1/x$  as close to 0 as desired if we require that  $x$  be sufficiently close to 0, but  $\neq 0$ . To be specific, suppose we wish to get  $x \sin 1/x$  within  $\frac{1}{10}$  of 0. This means we want

$$-\frac{1}{10} < x \sin \frac{1}{x} < \frac{1}{10},$$

or, more succinctly,  $|x \sin 1/x| < \frac{1}{10}$ . Now this is easy. Since

$$\left| \sin \frac{1}{x} \right| \leq 1, \quad \text{for all } x \neq 0,$$

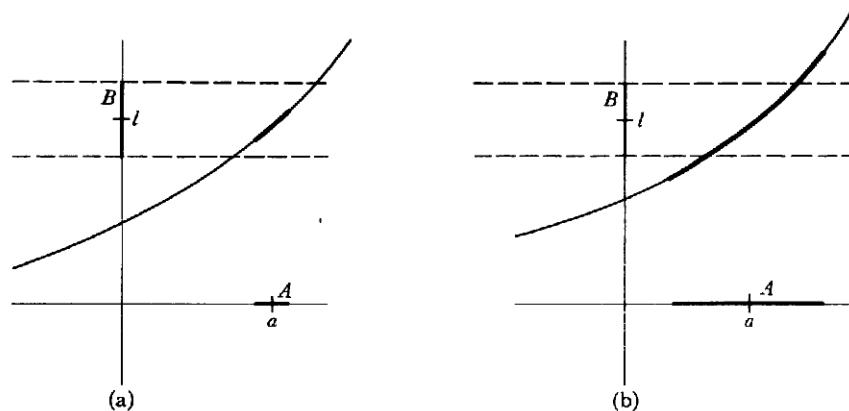


FIGURE 5

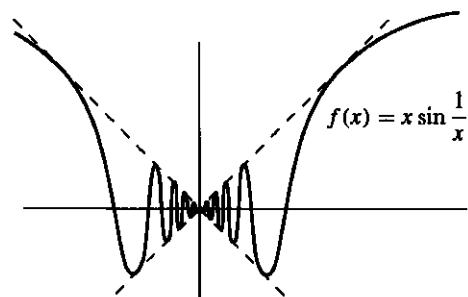


FIGURE 6

we have

$$\left| x \sin \frac{1}{x} \right| \leq |x|, \quad \text{for all } x \neq 0.$$

This means that if  $|x| < \frac{1}{10}$  and  $x \neq 0$ , then  $|x \sin 1/x| < \frac{1}{10}$ ; in other words,  $x \sin 1/x$  is within  $\frac{1}{10}$  of 0 provided that  $x$  is within  $\frac{1}{10}$  of 0, but  $\neq 0$ . There is nothing special about the number  $\frac{1}{10}$ ; it is just as easy to guarantee that  $|f(x) - 0| < \frac{1}{100}$ —simply require that  $|x| < \frac{1}{100}$ , but  $x \neq 0$ . In fact, if we take any positive number  $\varepsilon$  we can make  $|f(x) - 0| < \varepsilon$  simply by requiring that  $|x| < \varepsilon$ , and  $x \neq 0$ .

For the function  $f(x) = x^2 \sin 1/x$  (Figure 7) it seems even clearer that  $f$  approaches 0 near 0. If, for example, we want

$$\left| x^2 \sin \frac{1}{x} \right| < \frac{1}{10},$$

then we certainly need only require that  $|x| < \frac{1}{10}$  and  $x \neq 0$ , since this implies that  $|x^2| < \frac{1}{100}$  and consequently

$$\left| x^2 \sin \frac{1}{x} \right| \leq |x^2| < \frac{1}{100} < \frac{1}{10}.$$

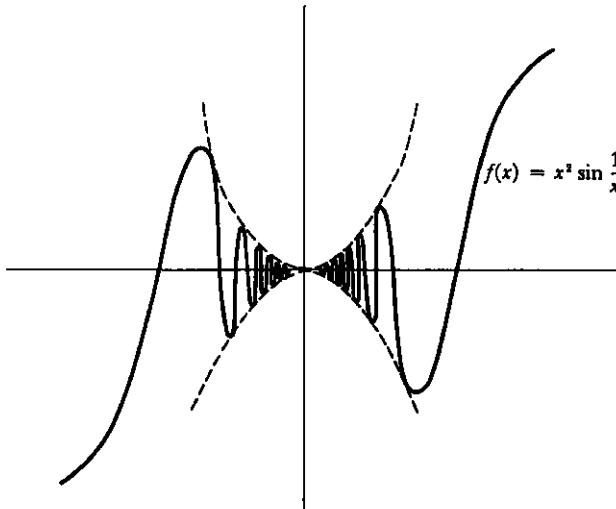


FIGURE 7

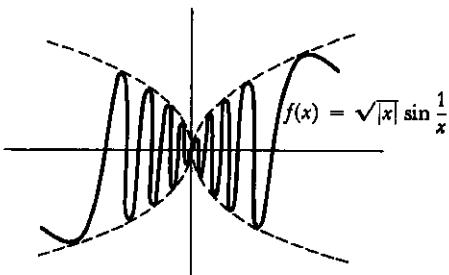


FIGURE 8

we need only require that

$$|x| < \varepsilon \quad \text{and} \quad x \neq 0,$$

(We could do even better, and allow  $|x| < 1/\sqrt{10}$  and  $x \neq 0$ , but there is no particular virtue in being as economical as possible.) In general, if  $\varepsilon > 0$ , to ensure that

$$\left| x^2 \sin \frac{1}{x} \right| < \varepsilon,$$

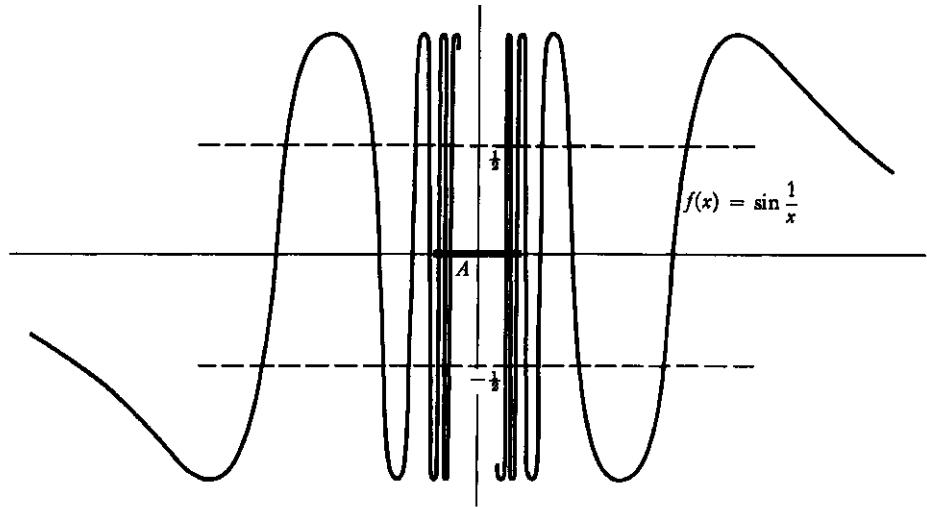


FIGURE 9

provided that  $\varepsilon \leq 1$ . If we are given an  $\varepsilon$  which is greater than 1 (it might be, even thought it is “small”  $\varepsilon$ 's which are of interest), then it does not suffice to require that  $|x| < \varepsilon$ , but it certainly suffices to require that  $|x| < 1$  and  $x \neq 0$ .

As a third example, consider the function  $f(x) = \sqrt{|x|} \sin 1/x$  (Figure 8). In order to make  $|\sqrt{|x|} \sin 1/x| < \varepsilon$  we can require that

$$|x| < \varepsilon^2 \quad \text{and} \quad x \neq 0$$

(the algebra is left to you).

Finally, let us consider the function  $f(x) = \sin 1/x$  (Figure 9). For this function it is *false* that  $f$  approaches 0 near 0. This amounts to saying that it is not true for every number  $\varepsilon > 0$  that we can get  $|f(x) - 0| < \varepsilon$  by choosing  $x$  sufficiently small, and  $\neq 0$ . To show this we simply have to find *one*  $\varepsilon > 0$  for which the condition  $|f(x) - 0| < \varepsilon$  cannot be guaranteed, no matter how small we require  $|x|$  to be. In fact,  $\varepsilon = \frac{1}{2}$  will do: it is impossible to ensure that  $|f(x)| < \frac{1}{2}$  no matter how small we require  $|x|$  to be; for if  $A$  is any interval containing 0, there is some number  $x = 1/(90 + 360n)$  which is in this interval, and for this  $x$  we have  $f(x) = 1$ .

This same argument can be used (Figure 10) to show that  $f$  does not approach *any* number near 0. To show this we must again find, for any particular number  $l$ , some number  $\varepsilon > 0$  so that  $|f(x) - l| < \varepsilon$  is *not* true, no matter how small  $x$  is required to be. The choice  $\varepsilon = \frac{1}{2}$  works for any number  $l$ ; that is, no matter how small we require  $|x|$  to be, we cannot ensure that  $|f(x) - l| < \frac{1}{2}$ . The reason is, that for any interval  $A$  containing 0 there is some  $x_1 = 1/(90 + 360n)$  in this interval, so that

$$f(x_1) = 1,$$

and also some  $x_2 = 1/(270 + 360m)$  in this interval, so that

$$f(x_2) = -1.$$

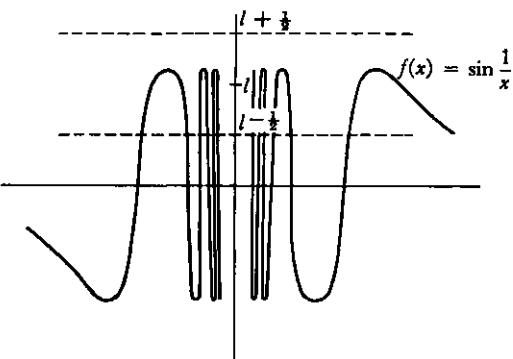


FIGURE 10

But the interval from  $l - \frac{1}{2}$  to  $l + \frac{1}{2}$  cannot contain both  $-1$  and  $1$ , since its total length is only 1; so we cannot have

$$|1 - l| < \frac{1}{2} \text{ and also } |-1 - l| < \frac{1}{2},$$

no matter what  $l$  is.

The phenomenon exhibited by  $f(x) = \sin 1/x$  near 0 can occur in many ways. If we consider the function

$$f(x) = \begin{cases} 0, & x \text{ irrational} \\ 1, & x \text{ rational,} \end{cases}$$

then, no matter what  $a$  is,  $f$  does not approach any number  $l$  near  $a$ . In fact, we cannot make  $|f(x) - l| < \frac{1}{4}$  no matter how close we bring  $x$  to  $a$ , because in any interval around  $a$  there are numbers  $x$  with  $f(x) = 0$ , and also numbers  $x$  with  $f(x) = 1$ , so that we would need  $|0 - l| < \frac{1}{4}$  and also  $|1 - l| < \frac{1}{4}$ .

An amusing variation on this behavior is presented by the function shown in Figure 11:

$$f(x) = \begin{cases} x, & x \text{ rational} \\ 0, & x \text{ irrational.} \end{cases}$$

The behavior of this function is “opposite” to that of  $g(x) = \sin 1/x$ ; it approaches 0 at 0, but does not approach any number at  $a$ , if  $a \neq 0$ . By now you should have no difficulty convincing yourself that this is true.

As a contrast to the functions considered so far, which have been quite pathological, we will now examine some of the simplest functions.

If  $f(x) = c$ , then  $f$  approaches  $c$  near  $a$ , for every number  $a$ . In fact, to ensure that  $|f(x) - c| < \varepsilon$  one does not need to restrict  $x$  to be near  $a$  at all; the condition is automatically satisfied (Figure 12).

As a slight variation, let  $f$  be the function shown in Figure 13:

$$f(x) = \begin{cases} -1, & x < 0 \\ 1, & x > 0. \end{cases}$$

If  $a > 0$ , then  $f$  approaches 1 near  $a$ : indeed, to ensure that  $|f(x) - 1| < \varepsilon$  it certainly suffices to require that  $|x - a| < a$ , since this implies

$$\begin{aligned} -a &< x - a \\ \text{or} \quad 0 &< x \end{aligned}$$

so that  $f(x) = 1$ . Similarly, if  $b < 0$ , then  $f$  approaches  $-1$  near  $b$ : to ensure that  $|f(x) - (-1)| < \varepsilon$  it suffices to require that  $|x - b| < -b$ . Finally, as you may easily check,  $f$  does not approach any number near 0.

The function  $f(x) = x$  is easily dealt with. Clearly  $f$  approaches  $a$  near  $a$ : to ensure that  $|f(x) - a| < \varepsilon$  we just have to require that  $|x - a| < \varepsilon$ .

The function  $f(x) = x^2$  requires a little more work. To show that  $f$  approaches  $a^2$  near  $a$ , we must decide how to ensure that

$$|x^2 - a^2| < \varepsilon.$$

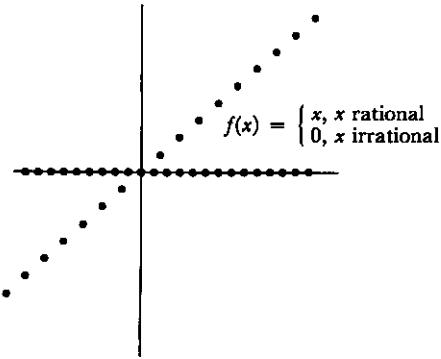


FIGURE 11

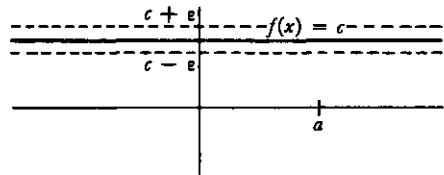


FIGURE 12

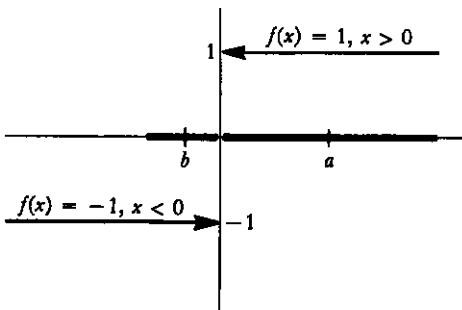


FIGURE 13

Factoring looks like the most promising procedure: we want

$$|x - a| \cdot |x + a| < \varepsilon.$$

Obviously the factor  $|x + a|$  is the one that will cause trouble. On the other hand, there is no need to make  $|x + a|$  particularly small; as long as we know *some* bound on the values of  $|x + a|$  we will be in good shape. For example, if  $|x + a| < 1,000,000$ , then we will just need to require that  $|x - a| < \varepsilon/1,000,000$ . Therefore, to begin with, let us require that  $|x - a| < 1$  (any positive number other than 1 would do just as well); presumably this will ensure that  $x$  is not too large, and consequently that  $|x + a|$  is not too large. As a matter of fact, Problem 1-12 shows that

$$|x| - |a| \leq |x - a| < 1,$$

so

$$|x| < 1 + |a|,$$

and consequently

$$|x + a| \leq |x| + |a| < 2|a| + 1.$$

Now we need only the additional requirement that  $|x - a| < \varepsilon/(2|a| + 1)$ . In other words,

$$\text{if } |x - a| < \min\left(1, \frac{\varepsilon}{2|a| + 1}\right), \text{ then } |x^2 - a^2| < \varepsilon.$$

Naturally,  $\min(1, \varepsilon/(2|a| + 1))$  will just be  $\varepsilon/(2|a| + 1)$  for small  $\varepsilon$ .

Precisely the same sort of trick will show that if  $f(x) = x^3$ , then  $f$  approaches  $a^3$  near  $a$ . In fact,

$$\text{if } |x - a| < \min\left(1, \frac{\varepsilon}{(1 + |a|)^2 + |a|(1 + |a|) + |a|^2}\right), \text{ then } |x^3 - a^3| < \varepsilon.$$

The proof of this assertion will show where the weird denominator comes from: If  $|x - a| < 1$ , then  $|x| < |a| + 1$ , and consequently

$$\begin{aligned} |x^2 + ax + a^2| &\leq |x|^2 + |a| \cdot |x| + |a|^2 \\ &< (1 + |a|)^2 + |a|(1 + |a|) + |a|^2. \end{aligned}$$

Therefore

$$\begin{aligned} |x^3 - a^3| &= |x - a| \cdot |x^2 + ax + a^2| \\ &< \frac{\varepsilon}{(1 + |a|)^2 + |a|(1 + |a|) + |a|^2} \cdot [(1 + |a|)^2 + |a|(1 + |a|) + |a|^2] \\ &= \varepsilon. \end{aligned}$$

The time has now come to point out that of the many demonstrations about limits which we have given, not one has been a real proof. The fault lies not with our reasoning, but with our definition. If our provisional definition of a function was open to criticism, our provisional definition of approaching a limit is even more vulnerable. This definition is simply not sufficiently precise to be used in proofs. It is hardly clear how one “makes”  $f(x)$  close to  $l$  (whatever “close” means) by “requiring”  $x$  to be sufficiently close to  $a$  (however close “suffi-

ciently” close is supposed to be). Despite the criticisms of our definition you may feel (I certainly hope you do) that our arguments were nevertheless quite convincing. In order to present any sort of argument at all, we have been practically forced to invent the real definition. It is possible to arrive at this definition in several steps, each one clarifying some obscure phrase which still remains. Let us begin, once again, with the provisional definition:

The function  $f$  approaches the limit  $l$  near  $a$ , if we can make  $f(x)$  as close as we like to  $l$  by requiring that  $x$  be sufficiently close to, but unequal to,  $a$ .

The very first change which we made in this definition was to note that making  $f(x)$  close to  $l$  meant making  $|f(x) - l|$  small, and similarly for  $x$  and  $a$ :

The function  $f$  approaches the limit  $l$  near  $a$ , if we can make  $|f(x) - l|$  as small as we like by requiring that  $|x - a|$  be sufficiently small, and  $x \neq a$ .

The second, more crucial, change was to note that making  $|f(x) - l|$  “as small as we like” means making  $|f(x) - l| < \varepsilon$  for any  $\varepsilon > 0$  that happens to be given us:

The function  $f$  approaches the limit  $l$  near  $a$ , if for every number  $\varepsilon > 0$  we can make  $|f(x) - l| < \varepsilon$  by requiring that  $|x - a|$  be sufficiently small, and  $x \neq a$ .

There is a common pattern to all the demonstrations about limits which we have given. For each number  $\varepsilon > 0$  we found some other positive number,  $\delta$  say, with the property that if  $x \neq a$  and  $|x - a| < \delta$ , then  $|f(x) - l| < \varepsilon$ . For the function  $f(x) = x \sin 1/x$  (with  $a = 0$ ,  $l = 0$ ), the number  $\delta$  was just the number  $\varepsilon$ ; for  $f(x) = \sqrt{|x|} \sin 1/x$ , it was  $\varepsilon^2$ ; for  $f(x) = x^2$  it was the minimum of 1 and  $\varepsilon/(2|a| + 1)$ . In general, it may not be at all clear how to find the number  $\delta$ , given  $\varepsilon$ , but it is the condition  $|x - a| < \delta$  which expresses how small “sufficiently” small must be:

The function  $f$  approaches the limit  $l$  near  $a$ , if for every  $\varepsilon > 0$  there is some  $\delta > 0$  such that, for all  $x$ , if  $|x - a| < \delta$  and  $x \neq a$ , then  $|f(x) - l| < \varepsilon$ .

This is practically the definition we will adopt. We will make only one trivial change, noting that “ $|x - a| < \delta$  and  $x \neq a$ ” can just as well be expressed “ $0 < |x - a| < \delta$ .”

#### DEFINITION

The function  $f$  approaches the limit  $l$  near  $a$  means: for every  $\varepsilon > 0$  there is some  $\delta > 0$  such that, for all  $x$ , if  $0 < |x - a| < \delta$ , then  $|f(x) - l| < \varepsilon$ .

This definition is so important (*everything* we do from now on depends on it) that proceeding any further without knowing it is hopeless. If necessary memorize it, like a poem! That, at least, is better than stating it incorrectly; if you do this you are doomed to give incorrect proofs. A good exercise in giving correct proofs is to review every fact already demonstrated about functions approaching limits, giving real proofs of each. This requires writing down the correct definition of what you are proving, but not much more—all the algebraic work has been done already. When proving that  $f$  does *not* approach  $l$  at  $a$ , be sure to negate the definition correctly:

If it is *not* true that

for every  $\varepsilon > 0$  there is some  $\delta > 0$  such that, for all  $x$ , if  $0 < |x - a| < \delta$ , then  $|f(x) - l| < \varepsilon$ ,

then

there is *some*  $\varepsilon > 0$  such that for *every*  $\delta > 0$  there is *some*  $x$  which satisfies  $0 < |x - a| < \delta$  but not  $|f(x) - l| < \varepsilon$ .

Thus, to show that the function  $f(x) = \sin 1/x$  does not approach 0 near 0, we consider  $\varepsilon = \frac{1}{2}$  and note that for every  $\delta > 0$  there is some  $x$  with  $0 < |x - 0| < \delta$  but not  $|\sin 1/x - 0| < \frac{1}{2}$ —namely, an  $x$  of the form  $1/(90 + 360n)$ , where  $n$  is so large that  $1/(90 + 360n) < \delta$ .

As an illustration of the use of the definition of a function approaching a limit, we have reserved the function shown in Figure 14, a standard example, but one of the most complicated:

$$f(x) = \begin{cases} 0, & x \text{ irrational, } 0 < x < 1 \\ 1/q, & x = p/q \text{ in lowest terms, } 0 < x < 1. \end{cases}$$

(Recall that  $p/q$  is in lowest terms if  $p$  and  $q$  are integers with no common factor and  $q > 0$ .)

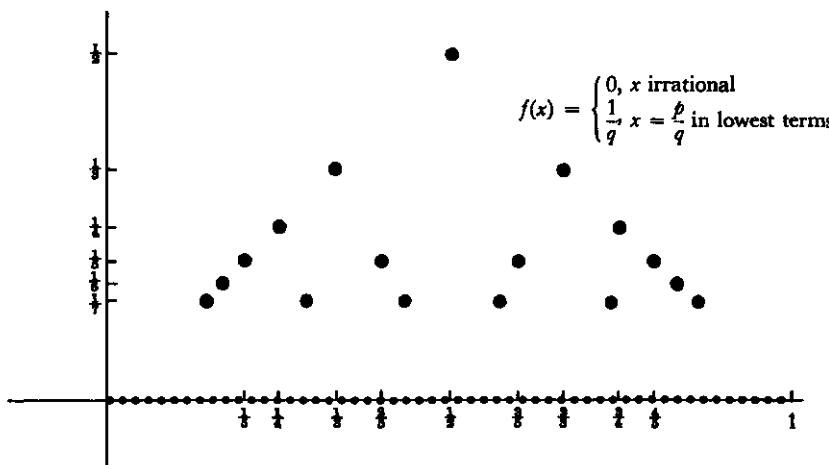


FIGURE 14

For any number  $a$ , with  $0 < a < 1$ , the function  $f$  approaches 0 at  $a$ . To prove this, consider any number  $\varepsilon > 0$ . Let  $n$  be a natural number so large that  $1/n \leq \varepsilon$ . Notice that the only numbers  $x$  for which  $|f(x) - 0| < \varepsilon$  could be false are:

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

(If  $a$  is rational, then  $a$  might be one of these numbers.) However many of these numbers there may be, there are, at any rate, only finitely many. Therefore, of all these numbers, one is closest to  $a$ ; that is,  $|p/q - a|$  is smallest for one  $p/q$  among these numbers. (If  $a$  happens to be one of these numbers, then consider only the values  $|p/q - a|$  for  $p/q \neq a$ .) This closest distance may be chosen as the  $\delta$ . For if  $0 < |x - a| < \delta$ , then  $x$  is *not* one of

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

and therefore  $|f(x) - 0| < \varepsilon$  is true. This completes the proof. Note that our description of the  $\delta$  which works for a given  $\varepsilon$  is completely adequate—there is no reason why we must give a formula for  $\delta$  in terms of  $\varepsilon$ .

Armed with our definition, we are now prepared to prove our first theorem; you have probably assumed the result all along, which is a very reasonable thing to do. This theorem is really a test case for our definition: if the theorem could not be proved, our definition would be useless.

**THEOREM 1** A function cannot approach two different limits near  $a$ . In other words, if  $f$  approaches  $l$  near  $a$ , and  $f$  approaches  $m$  near  $a$ , then  $l = m$ .

**PROOF** Since this is our first theorem about limits it will certainly be necessary to translate the hypotheses according to the definition.

Since  $f$  approaches  $l$  near  $a$ , we know that for any  $\varepsilon > 0$  there is some number  $\delta_1 > 0$  such that, for all  $x$ ,

$$\text{if } 0 < |x - a| < \delta_1, \text{ then } |f(x) - l| < \varepsilon.$$

We also know, since  $f$  approaches  $m$  near  $a$ , that there is some  $\delta_2 > 0$  such that, for all  $x$ ,

$$\text{if } 0 < |x - a| < \delta_2, \text{ then } |f(x) - m| < \varepsilon.$$

We have had to use two numbers,  $\delta_1$  and  $\delta_2$ , since there is no guarantee that the  $\delta$  which works in one definition will work in the other. But, in fact, it is now easy to conclude that for any  $\varepsilon > 0$  there is some  $\delta > 0$  such that, for all  $x$ ,

$$\text{if } 0 < |x - a| < \delta, \text{ then } |f(x) - l| < \varepsilon \text{ and } |f(x) - m| < \varepsilon;$$

we simply choose  $\delta = \min(\delta_1, \delta_2)$ .

To complete the proof we just have to pick a particular  $\varepsilon > 0$  for which the two conditions

$$|f(x) - l| < \varepsilon \quad \text{and} \quad |f(x) - m| < \varepsilon$$

cannot both hold, if  $l \neq m$ . The proper choice is suggested by Figure 15. If  $l \neq m$ , so that  $|l - m| > 0$ , we can choose  $|l - m|/2$  as our  $\varepsilon$ . It follows that there is a  $\delta > 0$  such that, for all  $x$ ,

$$\begin{aligned} \text{if } 0 < |x - a| < \delta, \text{ then } |f(x) - l| &< \frac{|l - m|}{2} \\ \text{and } |f(x) - m| &< \frac{|l - m|}{2}. \end{aligned}$$

This implies that for  $0 < |x - a| < \delta$  we have

$$\begin{aligned} |l - m| &= |l - f(x) + f(x) - m| \leq |l - f(x)| + |f(x) - m| \\ &< \frac{|l - m|}{2} + \frac{|l - m|}{2} \\ &= |l - m|, \end{aligned}$$

a contradiction. ■

The number  $l$  which  $f$  approaches near  $a$  is denoted by  $\lim_{x \rightarrow a} f(x)$  (read: the limit of  $f(x)$  as  $x$  approaches  $a$ ). This definition is possible only because of Theorem 1, which ensures that  $\lim_{x \rightarrow a} f(x)$  never has to stand for two different numbers. The equation

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

has exactly the same meaning as the phrase

$$f \text{ approaches } l \text{ near } a.$$

The possibility still remains that  $f$  does not approach  $l$  near  $a$ , for any  $l$ , so that  $\lim_{x \rightarrow a} f(x) = l$  is false for every number  $l$ . This is usually expressed by saying that “ $\lim_{x \rightarrow a} f(x)$  does not exist.”

Notice that our new notation introduces an extra, utterly irrelevant letter  $x$ , which could be replaced by  $t$ ,  $y$ , or any other letter which does not already appear—the symbols

$$\lim_{x \rightarrow a} f(x), \quad \lim_{t \rightarrow a} f(t), \quad \lim_{y \rightarrow a} f(y),$$

all denote precisely the same number, which depends on  $f$  and  $a$ , and has nothing to do with  $x$ ,  $t$ , or  $y$  (these letters, in fact, do not denote anything at all). A more logical symbol would be something like  $\lim_a f$ , but this notation, despite its brevity, is so infuriatingly rigid that almost no one has seriously tried to use it. The notation  $\lim_{x \rightarrow a} f(x)$  is much more useful because a function  $f$  often has no simple name, even

though it might be possible to express  $f(x)$  by a simple formula involving  $x$ . Thus, the short symbol

$$\lim_{x \rightarrow a} (x^2 + \sin x)$$

could be paraphrased only by the awkward expression

$$\lim_a f, \text{ where } f(x) = x^2 + \sin x.$$

Another advantage of the standard symbolism is illustrated by the expressions

$$\lim_{x \rightarrow a} x + t^3,$$

$$\lim_{t \rightarrow a} x + t^3.$$

The first means the number which  $f$  approaches near  $a$  when

$$f(x) = x + t^3, \quad \text{for all } x;$$

the second means the number which  $f$  approaches near  $a$  when

$$f(t) = x + t^3, \quad \text{for all } t.$$

You should have little difficulty (especially if you consult Theorem 2) proving that

$$\lim_{x \rightarrow a} x + t^3 = a + t^3,$$

$$\lim_{t \rightarrow a} x + t^3 = x + a^3.$$

These examples illustrate the main advantage of our notation, which is its flexibility. In fact, the notation  $\lim_{x \rightarrow a} f(x)$  is so flexible that there is some danger of forgetting what it really means. Here is a simple exercise in the use of this notation, which will be important later: first interpret precisely, and then prove the equality of the expressions

$$\lim_{x \rightarrow a} f(x) \quad \text{and} \quad \lim_{h \rightarrow 0} f(a + h).$$

An important part of this chapter is the proof of a theorem which will make it easy to find many limits. The proof depends upon certain properties of inequalities and absolute values, hardly surprising when one considers the definition of limit. Although these facts have already been stated in Problems 1-20, 1-21, and 1-22, because of their importance they will be presented once again, in the form of a lemma (a lemma is an auxiliary theorem, a result that justifies its existence only by virtue of its prominent role in the proof of another theorem). The lemma says, roughly, that if  $x$  is close to  $x_0$ , and  $y$  is close to  $y_0$ , then  $x + y$  will be close to  $x_0 + y_0$ , and  $xy$  will be close to  $x_0 y_0$ , and  $1/y$  will be close to  $1/y_0$ . This intuitive statement is much easier to remember than the precise estimates of the lemma, and it is not unreasonable to read the proof of Theorem 2 first, in order to see just how these estimates are used.

**LEMMA** (1) If

$$|x - x_0| < \frac{\varepsilon}{2} \text{ and } |y - y_0| < \frac{\varepsilon}{2},$$

then

$$|(x + y) - (x_0 + y_0)| < \varepsilon.$$

(2) If

$$|x - x_0| < \min\left(1, \frac{\varepsilon}{2(|y_0| + 1)}\right) \text{ and } |y - y_0| < \frac{\varepsilon}{2(|x_0| + 1)},$$

then

$$|xy - x_0y_0| < \varepsilon.$$

(3) If  $y_0 \neq 0$  and

$$|y - y_0| < \min\left(\frac{|y_0|}{2}, \frac{\varepsilon|y_0|^2}{2}\right),$$

then  $y \neq 0$  and

$$\left| \frac{1}{y} - \frac{1}{y_0} \right| < \varepsilon.$$

**PROOF**

$$(1) \quad \begin{aligned} |(x + y) - (x_0 + y_0)| &= |(x - x_0) + (y - y_0)| \\ &\leq |x - x_0| + |y - y_0| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

(2) Since  $|x - x_0| < 1$  we have

$$|x| - |x_0| \leq |x - x_0| < 1,$$

so that

$$|x| < 1 + |x_0|.$$

Thus

$$\begin{aligned} |xy - x_0y_0| &= |x(y - y_0) + y_0(x - x_0)| \\ &\leq |x| \cdot |y - y_0| + |y_0| \cdot |x - x_0| \\ &< (1 + |x_0|) \cdot \frac{\varepsilon}{2(|x_0| + 1)} + |y_0| \cdot \frac{\varepsilon}{2(|y_0| + 1)} \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

(3) We have

$$|y_0| - |y| \leq |y - y_0| < \frac{|y_0|}{2},$$

so  $|y| > |y_0|/2$ . In particular,  $y \neq 0$ , and

$$\frac{1}{|y|} < \frac{2}{|y_0|}.$$

Thus

$$\left| \frac{1}{y} - \frac{1}{y_0} \right| = \frac{|y_0 - y|}{|y| \cdot |y_0|} < \frac{2}{|y_0|} \cdot \frac{1}{|y_0|} \cdot \frac{\varepsilon|y_0|^2}{2} = \varepsilon. \blacksquare$$

**THEOREM 2** If  $\lim_{x \rightarrow a} f(x) = l$  and  $\lim_{x \rightarrow a} g(x) = m$ , then

- (1)  $\lim_{x \rightarrow a} (f + g)(x) = l + m;$
- (2)  $\lim_{x \rightarrow a} (f \cdot g)(x) = l \cdot m.$

Moreover, if  $m \neq 0$ , then

$$(3) \quad \lim_{x \rightarrow a} \left( \frac{1}{g} \right)(x) = \frac{1}{m}.$$

**PROOF** The hypothesis means that for every  $\varepsilon > 0$  there are  $\delta_1, \delta_2 > 0$  such that, for all  $x$ ,

$$\begin{aligned} &\text{if } 0 < |x - a| < \delta_1, \text{ then } |f(x) - l| < \varepsilon, \\ &\text{and if } 0 < |x - a| < \delta_2, \text{ then } |g(x) - m| < \varepsilon. \end{aligned}$$

This means (since, after all,  $\varepsilon/2$  is also a positive number) that there are  $\delta_1, \delta_2 > 0$  such that, for all  $x$ ,

$$\begin{aligned} &\text{if } 0 < |x - a| < \delta_1, \text{ then } |f(x) - l| < \frac{\varepsilon}{2}, \\ &\text{and if } 0 < |x - a| < \delta_2, \text{ then } |g(x) - m| < \frac{\varepsilon}{2}. \end{aligned}$$

Now let  $\delta = \min(\delta_1, \delta_2)$ . If  $0 < |x - a| < \delta$ , then  $0 < |x - a| < \delta_1$  and  $0 < |x - a| < \delta_2$  are both true, so both

$$|f(x) - l| < \frac{\varepsilon}{2} \quad \text{and} \quad |g(x) - m| < \frac{\varepsilon}{2}$$

are true. But by part (1) of the lemma this implies that  $|(f + g)(x) - (l + m)| < \varepsilon$ . This proves (1).

To prove (2) we proceed similarly, after consulting part (2) of the lemma. If  $\varepsilon > 0$  there are  $\delta_1, \delta_2 > 0$  such that, for all  $x$ ,

$$\begin{aligned} &\text{if } 0 < |x - a| < \delta_1, \text{ then } |f(x) - l| < \min\left(1, \frac{\varepsilon}{2(|m| + 1)}\right), \\ &\text{and if } 0 < |x - a| < \delta_2, \text{ then } |g(x) - m| < \frac{\varepsilon}{2(|l| + 1)}. \end{aligned}$$

Again let  $\delta = \min(\delta_1, \delta_2)$ . If  $0 < |x - a| < \delta$ , then

$$|f(x) - l| < \min\left(1, \frac{\varepsilon}{2(|m| + 1)}\right) \quad \text{and} \quad |g(x) - m| < \frac{\varepsilon}{2(|l| + 1)}.$$

So, by the lemma,  $|(f \cdot g)(x) - l \cdot m| < \varepsilon$ , and this proves (2).

Finally, if  $\varepsilon > 0$  there is a  $\delta > 0$  such that, for all  $x$ ,

$$\text{if } 0 < |x - a| < \delta, \text{ then } |g(x) - m| < \min\left(\frac{|m|}{2}, \frac{\varepsilon|m|^2}{2}\right).$$

But according to part (3) of the lemma this means, first, that  $g(x) \neq 0$ , so  $(1/g)(x)$  makes sense, and second that

$$\left| \left( \frac{1}{g} \right)(x) - \frac{1}{m} \right| < \varepsilon.$$

This proves (3). ■

Using Theorem 2 we can prove, trivially, such facts as

$$\lim_{x \rightarrow a} \frac{x^3 + 7x^5}{x^2 + 1} = \frac{a^3 + 7a^5}{a^2 + 1},$$

without going through the laborious process of finding a  $\delta$ , given an  $\varepsilon$ . We must begin with

$$\begin{aligned} \lim_{x \rightarrow a} 7 &= 7, \\ \lim_{x \rightarrow a} 1 &= 1, \\ \lim_{x \rightarrow a} x &= a, \end{aligned}$$

but these are easy to prove directly. If we *want* to find the  $\delta$ , however, the proof of Theorem 2 amounts to a prescription for doing this. Suppose, to take a simpler example, that we want to find a  $\delta$  such that, for all  $x$ ,

$$\text{if } 0 < |x - a| < \delta, \text{ then } |x^2 + x - (a^2 + a)| < \varepsilon.$$

Consulting the proof of Theorem 2(1), we see that we must first find  $\delta_1$  and  $\delta_2 > 0$  such that, for all  $x$ ,

$$\begin{aligned} \text{if } 0 < |x - a| < \delta_1, \text{ then } |x^2 - a^2| &< \frac{\varepsilon}{2} \\ \text{and if } 0 < |x - a| < \delta_2, \text{ then } |x - a| &< \frac{\varepsilon}{2}. \end{aligned}$$

Since we have already given proofs that  $\lim_{x \rightarrow a} x^2 = a^2$  and  $\lim_{x \rightarrow a} x = a$ , we know how to do this:

$$\begin{aligned} \delta_1 &= \min \left( 1, \frac{\varepsilon}{2|a| + 1} \right), \\ \delta_2 &= \frac{\varepsilon}{2}. \end{aligned}$$

Thus we can take

$$\delta = \min(\delta_1, \delta_2) = \min \left( \min \left( 1, \frac{\varepsilon}{2|a| + 1} \right), \frac{\varepsilon}{2} \right).$$

If  $a \neq 0$ , the same method can be used to find a  $\delta > 0$  such that, for all  $x$ ,

$$\text{if } 0 < |x - a| < \delta, \text{ then } \left| \frac{1}{x^2} - \frac{1}{a^2} \right| < \varepsilon.$$

The proof of Theorem 2(3) shows that the second condition will follow if we find a  $\delta > 0$  such that, for all  $x$ ,

$$\text{if } 0 < |x - a| < \delta, \text{ then } |x^2 - a^2| < \min \left( \frac{|a|^2}{2}, \frac{\varepsilon |a|^4}{2} \right).$$

Thus we can take

$$\delta = \min \left( 1, \frac{\min \left( \frac{|a|^2}{2}, \frac{\varepsilon |a|^4}{2} \right)}{2|a| + 1} \right).$$

Naturally, these complicated expressions for  $\delta$  can be simplified considerably, after they have been derived.

One technical detail in the proof of Theorem 2 deserves some discussion. In order for  $\lim_{x \rightarrow a} f(x)$  to be defined it is, as we know, not necessary for  $f$  to be defined at  $a$ , nor is it necessary for  $f$  to be defined at all points  $x \neq a$ . However, there must be some  $\delta > 0$  such that  $f(x)$  is defined for  $x$  satisfying  $0 < |x - a| < \delta$ ; otherwise the clause

$$\text{"if } 0 < |x - a| < \delta, \text{ then } |f(x) - l| < \varepsilon\text{"}$$

would make no sense at all, since the symbol  $f(x)$  would make no sense for some  $x$ 's. If  $f$  and  $g$  are two functions for which the definition makes sense, it is easy to see that the same is true for  $f + g$  and  $f \cdot g$ . But this is not so clear for  $1/g$ , since  $1/g$  is undefined for  $x$  with  $g(x) = 0$ . However, this fact gets established in the proof of Theorem 2(3).

There are times when we would like to speak of the limit which  $f$  approaches at  $a$ , even though there is no  $\delta > 0$  such that  $f(x)$  is defined for  $x$  satisfying  $0 < |x - a| < \delta$ . For example, we want to distinguish the behavior of the two functions shown in Figure 16, even though they are not defined for numbers less than  $a$ . For the function of Figure 16(a) we write

$$\lim_{x \rightarrow a^+} f(x) = l \quad \text{or} \quad \lim_{x \downarrow a} f(x) = l.$$

(The symbols on the left are read: the limit of  $f(x)$  as  $x$  approaches  $a$  from above.) These “limits from above” are obviously closely related to ordinary limits, and the definition is very similar:  $\lim_{x \rightarrow a^+} f(x) = l$  means that for every  $\varepsilon > 0$  there is a  $\delta > 0$  such that, for all  $x$ ,

$$\text{if } 0 < x - a < \delta, \text{ then } |f(x) - l| < \varepsilon.$$

(The condition “ $0 < x - a < \delta$ ” is equivalent to “ $0 < |x - a| < \delta$  and  $x > a$ .”)

“Limits from below” (Figure 17) are defined similarly:  $\lim_{x \rightarrow a^-} f(x) = l$  (or

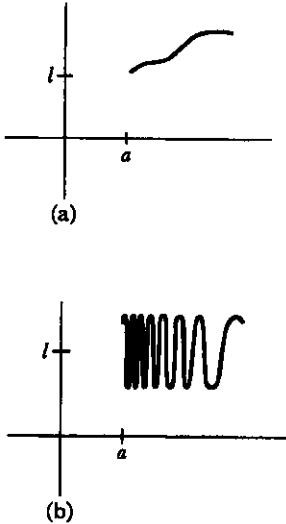


FIGURE 16

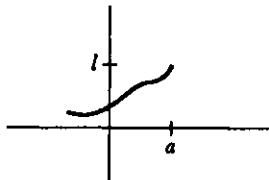


FIGURE 17

$\lim_{x \uparrow a} f(x) = l$  means that for every  $\varepsilon > 0$  there is a  $\delta > 0$  such that, for all  $x$ ,

$$\text{if } 0 < a - x < \delta, \text{ then } |f(x) - l| < \varepsilon.$$

It is quite possible to consider limits from above and below even if  $f$  is defined for numbers both greater and less than  $a$ . Thus, for the function  $f$  of Figure 13, we have

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

It is an easy exercise (Problem 29) to show that  $\lim_{x \rightarrow a} f(x)$  exists if and only if  $\lim_{x \rightarrow a^+} f(x)$  and  $\lim_{x \rightarrow a^-} f(x)$  both exist and are equal.

Like the definitions of limits from above and below, which have been smuggled into the text informally, there are other modifications of the limit concept which will be found useful. In Chapter 4 it was claimed that if  $x$  is large, then  $\sin 1/x$  is close to 0. This assertion is usually written

$$\lim_{x \rightarrow \infty} \sin 1/x = 0.$$

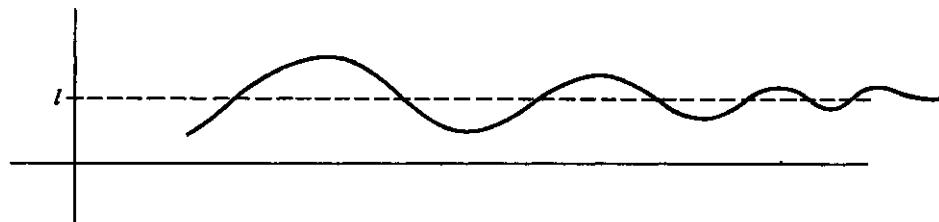


FIGURE 18

The symbol  $\lim_{x \rightarrow \infty} f(x)$  is read “the limit of  $f(x)$  as  $x$  approaches  $\infty$ ,” or “as  $x$  becomes infinite,” and a limit of the form  $\lim_{x \rightarrow \infty} f(x)$  is often called a limit at infinity. Figure 18 illustrates a general situation where  $\lim_{x \rightarrow \infty} f(x) = l$ . Formally,  $\lim_{x \rightarrow \infty} f(x) = l$  means that for every  $\varepsilon > 0$  there is a number  $N$  such that, for all  $x$ ,

$$\text{if } x > N, \text{ then } |f(x) - l| < \varepsilon.$$

The analogy with the definition of ordinary limits should be clear: whereas the condition “ $0 < |x - a| < \delta$ ” expresses the fact that  $x$  is close to  $a$ , the condition “ $x > N$ ” expresses the fact that  $x$  is large.

We have spent so little time on limits from above and below, and at infinity, because the general philosophy behind the definitions should be clear if you understand the definition of ordinary limits (which are by far the most important). Many exercises on these definitions are provided in the Problems, which also contain several other types of limits which are occasionally useful.

## PROBLEMS

1. Find the following limits. (These limits all follow, after some algebraic manipulations, from the various parts of Theorem 2; be sure you know which ones are used in each case, but don't bother listing them.)

$$\begin{aligned} \text{(i)} \quad & \lim_{x \rightarrow 1} \frac{x^2 - 1}{x + 1}. \\ \text{(ii)} \quad & \lim_{x \rightarrow 2} \frac{x^3 - 8}{x - 2}. \\ \text{(iii)} \quad & \lim_{x \rightarrow 3} \frac{x^3 - 8}{x - 2}. \\ \text{(iv)} \quad & \lim_{x \rightarrow y} \frac{x^n - y^n}{x - y}. \\ \text{(v)} \quad & \lim_{y \rightarrow x} \frac{x^n - y^n}{x - y}. \\ \text{(vi)} \quad & \lim_{h \rightarrow 0} \frac{\sqrt{a+h} - \sqrt{a}}{h}. \end{aligned}$$

2. Find the following limits.

$$\begin{aligned} \text{(i)} \quad & \lim_{x \rightarrow 1} \frac{1 - \sqrt{x}}{1 - x}. \\ \text{(ii)} \quad & \lim_{x \rightarrow 0} \frac{1 - \sqrt{1 - x^2}}{x}. \\ \text{(iii)} \quad & \lim_{x \rightarrow 0} \frac{1 - \sqrt{1 - x^2}}{x^2}. \end{aligned}$$

3. In each of the following cases, find a  $\delta$  such that  $|f(x) - l| < \varepsilon$  for all  $x$  satisfying  $0 < |x - a| < \delta$ .

$$\begin{aligned} \text{(i)} \quad & f(x) = x^4; l = a^4. \\ \text{(ii)} \quad & f(x) = \frac{1}{x}; a = 1, l = 1. \\ \text{(iii)} \quad & f(x) = x^4 + \frac{1}{x}; a = 1, l = 2. \\ \text{(iv)} \quad & f(x) = \frac{x}{1 + \sin^2 x}; a = 0, l = 0. \\ \text{(v)} \quad & f(x) = \sqrt{|x|}; a = 0, l = 0. \\ \text{(vi)} \quad & f(x) = \sqrt{x}; a = 1, l = 1. \end{aligned}$$

4. For each of the functions in Problem 4-17, decide for which numbers  $a$  the limit  $\lim_{x \rightarrow a} f(x)$  exists.

- \*5. (a) Do the same for each of the functions in Problem 4-19.  
 (b) Same problem, if we use infinite decimals ending in a string of 0's instead of those ending in a string of 9's.

6. Suppose the functions  $f$  and  $g$  have the following property: for all  $\varepsilon > 0$  and all  $x$ ,

$$\begin{aligned} \text{if } 0 < |x - 2| < \sin^2\left(\frac{\varepsilon^2}{9}\right) + \varepsilon, \text{ then } |f(x) - 2| < \varepsilon, \\ \text{if } 0 < |x - 2| < \varepsilon^2, \text{ then } |g(x) - 4| < \varepsilon. \end{aligned}$$

For each  $\varepsilon > 0$  find a  $\delta > 0$  such that, for all  $x$ ,

- (i) if  $0 < |x - 2| < \delta$ , then  $|f(x) + g(x) - 6| < \varepsilon$ .
- (ii) if  $0 < |x - 2| < \delta$ , then  $|f(x)g(x) - 8| < \varepsilon$ .
- (iii) if  $0 < |x - 2| < \delta$ , then  $\left|\frac{1}{g(x)} - \frac{1}{4}\right| < \varepsilon$ .
- (iv) if  $0 < |x - 2| < \delta$ , then  $\left|\frac{f(x)}{g(x)} - \frac{1}{2}\right| < \varepsilon$ .

7. Give an example of a function  $f$  for which the following assertion is *false*: If  $|f(x) - l| < \varepsilon$  when  $0 < |x - a| < \delta$ , then  $|f(x) - l| < \varepsilon/2$  when  $0 < |x - a| < \delta/2$ .
8. (a) If  $\lim_{x \rightarrow a} f(x)$  and  $\lim_{x \rightarrow a} g(x)$  do not exist, can  $\lim_{x \rightarrow a} [f(x) + g(x)]$  or  $\lim_{x \rightarrow a} f(x)g(x)$  exist?  
 (b) If  $\lim_{x \rightarrow a} f(x)$  exists and  $\lim_{x \rightarrow a} [f(x) + g(x)]$  exists, must  $\lim_{x \rightarrow a} g(x)$  exist?  
 (c) If  $\lim_{x \rightarrow a} f(x)$  exists and  $\lim_{x \rightarrow a} g(x)$  does not exist, can  $\lim_{x \rightarrow a} [f(x) + g(x)]$  exist?  
 (d) If  $\lim_{x \rightarrow a} f(x)$  exists and  $\lim_{x \rightarrow a} f(x)g(x)$  exists, does it follow that  $\lim_{x \rightarrow a} g(x)$  exists?
9. Prove that  $\lim_{x \rightarrow a} f(x) = \lim_{h \rightarrow 0} f(a + h)$ . (This is mainly an exercise in understanding what the terms mean.)
10. (a) Prove that  $\lim_{x \rightarrow a} f(x) = l$  if and only if  $\lim_{x \rightarrow a} [f(x) - l] = 0$ . (First see why the assertion is obvious; then provide a rigorous proof. In this chapter most problems which ask for proofs should be treated in the same way.)  
 (b) Prove that  $\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow a} f(x - a)$ .  
 (c) Prove that  $\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} f(x^3)$ .  
 (d) Give an example where  $\lim_{x \rightarrow 0} f(x^2)$  exists, but  $\lim_{x \rightarrow 0} f(x)$  does not.
11. Suppose there is a  $\delta > 0$  such that  $f(x) = g(x)$  when  $0 < |x - a| < \delta$ . Prove that  $\lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} g(x)$ . In other words,  $\lim_{x \rightarrow a} f(x)$  depends only on the values of  $f(x)$  for  $x$  near  $a$ —this fact is often expressed by saying that limits are a “local property.” (It will clearly help to use  $\delta'$ , or some other letter, instead of  $\delta$ , in the definition of limits.)
12. (a) Suppose that  $f(x) \leq g(x)$  for all  $x$ . Prove that  $\lim_{x \rightarrow a} f(x) \leq \lim_{x \rightarrow a} g(x)$ , provided that these limits exist.

- (b) How can the hypotheses be weakened?
- (c) If  $f(x) < g(x)$  for all  $x$ , does it necessarily follow that  $\lim_{x \rightarrow a} f(x) < \lim_{x \rightarrow a} g(x)$ ?
13. Suppose that  $f(x) \leq g(x) \leq h(x)$  and that  $\lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} h(x)$ . Prove that  $\lim_{x \rightarrow a} g(x)$  exists, and that  $\lim_{x \rightarrow a} g(x) = \lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} h(x)$ . (Draw a picture!)
- \*14. (a) Prove that if  $\lim_{x \rightarrow 0} f(x)/x = l$  and  $b \neq 0$ , then  $\lim_{x \rightarrow 0} f(bx)/x = bl$ . Hint: Write  $f(bx)/x = b[f(bx)/bx]$ .
- (b) What happens if  $b = 0$ ?
- (c) Part (a) enables us to find  $\lim_{x \rightarrow 0} (\sin 2x)/x$  in terms of  $\lim_{x \rightarrow 0} (\sin x)/x$ . Find this limit in another way.
15. Evaluate the following limits in terms of the number  $\alpha = \lim_{x \rightarrow 0} (\sin x)/x$ .
- (i)  $\lim_{x \rightarrow 0} \frac{\sin 2x}{x}$ .
  - (ii)  $\lim_{x \rightarrow 0} \frac{\sin ax}{\sin bx}$ .
  - (iii)  $\lim_{x \rightarrow 0} \frac{\sin^2 2x}{x}$ .
  - (iv)  $\lim_{x \rightarrow 0} \frac{\sin^2 2x}{x^2}$ .
  - (v)  $\lim_{x \rightarrow 0} \frac{1 - \cos x}{x^2}$ .
  - (vi)  $\lim_{x \rightarrow 0} \frac{\tan^2 x + 2x}{x + x^2}$ .
  - (vii)  $\lim_{x \rightarrow 0} \frac{x \sin x}{1 - \cos x}$ .
  - (viii)  $\lim_{h \rightarrow 0} \frac{\sin(x+h) - \sin x}{h}$ .
  - (ix)  $\lim_{x \rightarrow 1} \frac{\sin(x^2 - 1)}{x - 1}$ .
  - (x)  $\lim_{x \rightarrow 0} \frac{x^2(3 + \sin x)}{(x + \sin x)^2}$ .
  - (xi)  $\lim_{x \rightarrow 1} (x^2 - 1)^3 \sin \left( \frac{1}{x-1} \right)^3$ .
16. (a) Prove that if  $\lim_{x \rightarrow a} f(x) = l$ , then  $\lim_{x \rightarrow a} |f|(x) = |l|$ .
- (b) Prove that if  $\lim_{x \rightarrow a} f(x) = l$  and  $\lim_{x \rightarrow a} g(x) = m$ , then  $\lim_{x \rightarrow a} \max(f, g)(x) = \max(l, m)$  and similarly for min.

- 17.** (a) Prove that  $\lim_{x \rightarrow 0} 1/x$  does not exist, i.e., show that  $\lim_{x \rightarrow 0} 1/x = l$  is false for every number  $l$ .  
 (b) Prove that  $\lim_{x \rightarrow 1} 1/(x - 1)$  does not exist.
- 18.** Prove that if  $\lim_{x \rightarrow a} f(x) = l$ , then there is a number  $\delta > 0$  and a number  $M$  such that  $|f(x)| < M$  if  $0 < |x - a| < \delta$ . (What does this mean pictorially?) Hint: Why does it suffice to prove that  $l - 1 < f(x) < l + 1$  for  $0 < |x - a| < \delta$ ?
- 19.** Prove that if  $f(x) = 0$  for irrational  $x$  and  $f(x) = 1$  for rational  $x$ , then  $\lim_{x \rightarrow a} f(x)$  does not exist for any  $a$ .
- \*20.** Prove that if  $f(x) = x$  for rational  $x$ , and  $f(x) = -x$  for irrational  $x$ , then  $\lim_{x \rightarrow a} f(x)$  does not exist if  $a \neq 0$ .
- 21.** (a) Prove that if  $\lim_{x \rightarrow 0} g(x) = 0$ , then  $\lim_{x \rightarrow 0} g(x) \sin 1/x = 0$ .  
 (b) Generalize this fact as follows: If  $\lim_{x \rightarrow 0} g(x) = 0$  and  $|h(x)| \leq M$  for all  $x$ , then  $\lim_{x \rightarrow 0} g(x)h(x) = 0$ . (Naturally it is unnecessary to do part (a) if you succeed in doing part (b); actually the statement of part (b) may make it easier than (a)—that's one of the values of generalization.)
- 22.** Consider a function  $f$  with the following property: if  $g$  is any function for which  $\lim_{x \rightarrow 0} g(x)$  does not exist, then  $\lim_{x \rightarrow 0} [f(x) + g(x)]$  also does not exist. Prove that this happens if and only if  $\lim_{x \rightarrow 0} f(x)$  does exist. Hint: This is actually very easy: the assumption that  $\lim_{x \rightarrow 0} f(x)$  does not exist leads to an immediate contradiction if you consider the right  $g$ .
- \*\*23.** This problem is the analogue of Problem 22 when  $f + g$  is replaced by  $f \cdot g$ . In this case the situation is considerably more complex, and the analysis requires several steps (those in search of an especially challenging problem can attempt an independent solution).  
 (a) Suppose that  $\lim_{x \rightarrow 0} f(x)$  exists and is  $\neq 0$ . Prove that if  $\lim_{x \rightarrow 0} g(x)$  does not exist, then  $\lim_{x \rightarrow 0} f(x)g(x)$  also does not exist.  
 (b) Prove the same result if  $\lim_{x \rightarrow 0} |f(x)| = \infty$ . (The precise definition of this sort of limit is given in Problem 37.)  
 (c) Prove that if neither of these two conditions holds, then there is a function  $g$  such that  $\lim_{x \rightarrow 0} g(x)$  does not exist, but  $\lim_{x \rightarrow 0} f(x)g(x)$  does exist.  
 Hint: Consider separately the following two cases: (1) for some  $\varepsilon > 0$  we have  $|f(x)| > \varepsilon$  for all sufficiently small  $x$ . (2) For every  $\varepsilon > 0$ , there are arbitrarily small  $x$  with  $|f(x)| < \varepsilon$ . In the second case, begin by choosing points  $x_n$  with  $|x_n| < 1/n$  and  $|f(x_n)| < 1/n$ .
- \*24.** Suppose that  $A_n$  is, for each natural number  $n$ , some *finite* set of numbers in  $[0, 1]$ , and that  $A_n$  and  $A_m$  have no members in common if  $m \neq n$ . Define

$f$  as follows:

$$f(x) = \begin{cases} 1/n, & x \text{ in } A_n \\ 0, & x \text{ not in } A_n \text{ for any } n. \end{cases}$$

Prove that  $\lim_{x \rightarrow a} f(x) = 0$  for all  $a$  in  $[0, 1]$ .

25. Explain why the following definitions of  $\lim_{x \rightarrow a} f(x) = l$  are all correct:  
 For every  $\delta > 0$  there is an  $\varepsilon > 0$  such that, for all  $x$ ,
- (i) if  $0 < |x - a| < \varepsilon$ , then  $|f(x) - l| < \delta$ .
  - (ii) if  $0 < |x - a| < \varepsilon$ , then  $|f(x) - l| \leq \delta$ .
  - (iii) if  $0 < |x - a| < \varepsilon$ , then  $|f(x) - l| < 5\delta$ .
  - (iv) if  $0 < |x - a| < \varepsilon/10$ , then  $|f(x) - l| < \delta$ .
- \*26. Give examples to show that the following definitions of  $\lim_{x \rightarrow a} f(x) = l$  are *not* correct.
- (a) For all  $\delta > 0$  there is an  $\varepsilon > 0$  such that if  $0 < |x - a| < \delta$ , then  $|f(x) - l| < \varepsilon$ .
  - (b) For all  $\varepsilon > 0$  there is a  $\delta > 0$  such that if  $|f(x) - l| < \varepsilon$ , then  $0 < |x - a| < \delta$ .
27. For each of the functions in Problem 4-17 indicate for which numbers  $a$  the one-sided limits  $\lim_{x \rightarrow a^+} f(x)$  and  $\lim_{x \rightarrow a^-} f(x)$  exist.
- \*28. (a) Do the same for each of the functions in Problem 4-19.  
 (b) Also consider what happens if decimals ending in 0's are used instead of decimals ending in 9's.
29. Prove that  $\lim_{x \rightarrow a} f(x)$  exists if  $\lim_{x \rightarrow a^+} f(x) = \lim_{x \rightarrow a^-} f(x)$ .
30. Prove that
- (i)  $\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow 0^-} f(-x)$ .
  - (ii)  $\lim_{x \rightarrow 0} f(|x|) = \lim_{x \rightarrow 0^+} f(x)$ .
  - (iii)  $\lim_{x \rightarrow 0} f(x^2) = \lim_{x \rightarrow 0^+} f(x)$ .
- (These equations, and others like them, are open to several interpretations. They might mean only that the two limits are equal if they both exist; or that if a certain one of the limits exists, the other also exists and is equal to it; or that if either limit exists, then the other exists and is equal to it. Decide for yourself which interpretations are suitable.)
- \*31. Suppose that  $\lim_{x \rightarrow a^-} f(x) < \lim_{x \rightarrow a^+} f(x)$ . (Draw a picture to illustrate this assertion.) Prove that there is some  $\delta > 0$  such that  $f(x) < f(y)$  whenever  $x < a < y$  and  $|x - a| < \delta$  and  $|y - a| < \delta$ . Is the converse true?

- \*32. Prove that  $\lim_{x \rightarrow \infty} (a_n x^n + \dots + a_0) / (b_m x^m + \dots + b_0)$  exists if and only if  $m \geq n$ . What is the limit when  $m = n$ ? When  $m > n$ ? Hint: the one easy limit is  $\lim_{x \rightarrow \infty} 1/x^k = 0$ ; do some algebra so that this is the only information you need.

33. Find the following limits.

$$\begin{aligned} \text{(i)} \quad & \lim_{x \rightarrow \infty} \frac{x + \sin^3 x}{5x + 6}. \\ \text{(ii)} \quad & \lim_{x \rightarrow \infty} \frac{x \sin x}{x^2 + 5}. \\ \text{(iii)} \quad & \lim_{x \rightarrow \infty} \sqrt{x^2 + x} - x. \\ \text{(iv)} \quad & \lim_{x \rightarrow \infty} \frac{x^2(1 + \sin^2 x)}{(x + \sin x)^2}. \end{aligned}$$

34. Prove that  $\lim_{x \rightarrow 0^+} f(1/x) = \lim_{x \rightarrow \infty} f(x)$ .

35. Find the following limits in terms of the number  $\alpha = \lim_{x \rightarrow 0} (\sin x)/x$ .

$$\begin{aligned} \text{(i)} \quad & \lim_{x \rightarrow \infty} \frac{\sin x}{x}. \\ \text{(ii)} \quad & \lim_{x \rightarrow \infty} x \sin \frac{1}{x}. \end{aligned}$$

36. Define “ $\lim_{x \rightarrow -\infty} f(x) = l$ .”

- Find  $\lim_{x \rightarrow -\infty} (a_n x^n + \dots + a_0) / (b_m x^m + \dots + b_0)$ .
- Prove that  $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow -\infty} f(-x)$ .
- Prove that  $\lim_{x \rightarrow 0^-} f(1/x) = \lim_{x \rightarrow -\infty} f(x)$ .

37. We define  $\lim_{x \rightarrow a} f(x) = \infty$  to mean that for all  $N$  there is a  $\delta > 0$  such that, for all  $x$ , if  $0 < |x - a| < \delta$ , then  $f(x) > N$ . (Draw an appropriate picture!)

- Show that  $\lim_{x \rightarrow 3} 1/(x - 3)^2 = \infty$ .
- Prove that if  $f(x) > \varepsilon > 0$  for all  $x$ , and  $\lim_{x \rightarrow a} g(x) = 0$ , then

$$\lim_{x \rightarrow a} f(x)/|g(x)| = \infty.$$

38. (a) Define  $\lim_{x \rightarrow a^+} f(x) = \infty$ ,  $\lim_{x \rightarrow a^-} f(x) = \infty$ , and  $\lim_{x \rightarrow a} f(x) = \infty$ . (Or at least convince yourself that you could write down the definitions if you had the energy. How many other such symbols can you define?)  
(b) Prove that  $\lim_{x \rightarrow 0^+} 1/x = \infty$ .  
(c) Prove that  $\lim_{x \rightarrow 0^+} f(x) = \infty$  if and only if  $\lim_{x \rightarrow \infty} f(1/x) = \infty$ .

39. Find the following limits, when they exist.

$$\text{(i)} \quad \lim_{x \rightarrow \infty} \frac{x^3 + 4x - 7}{7x^2 - x + 1}$$

(ii)  $\lim_{x \rightarrow \infty} x(1 + \sin^2 x).$

(iii)  $\lim_{x \rightarrow \infty} x \sin^2 x.$

(iv)  $\lim_{x \rightarrow \infty} x^2 \sin \frac{1}{x}.$

(v)  $\lim_{x \rightarrow \infty} \sqrt{x^2 + 2x} - x.$

(vi)  $\lim_{x \rightarrow \infty} x(\sqrt{x+2} - \sqrt{x}).$

(vii)  $\lim_{x \rightarrow \infty} \frac{\sqrt{|x|}}{x}.$

40. (a) Find the perimeter of a regular  $n$ -gon inscribed in a circle of radius  $r$ ; use radian measure for any trigonometric functions involved. [Answer:  $2rn \sin(\pi/n)$ .]  
 (b) What value does this perimeter approach as  $n$  becomes very large?
41. After sending the manuscript for the first edition of this book off to the printer, I thought of a much simpler way to prove that  $\lim_{x \rightarrow a} x^2 = a^2$  and  $\lim_{x \rightarrow a} x^3 = a^3$ , without going through all the factoring tricks on page 95. Suppose, for example, that we want to prove that  $\lim_{x \rightarrow a} x^2 = a^2$ , where  $a > 0$ . Given  $\epsilon > 0$ , we simply let  $\delta$  be the minimum of  $\sqrt{a^2 + \epsilon} - a$  and  $a - \sqrt{a^2 - \epsilon}$  (see Figure 19); then  $|x - a| < \delta$  implies that  $\sqrt{a^2 - \epsilon} < x < \sqrt{a^2 + \epsilon}$ , so  $a^2 - \epsilon < x^2 < a^2 + \epsilon$ , or  $|x^2 - a^2| < \epsilon$ . It is fortunate that these pages had already been set, so that I couldn't make these changes, because this "proof" is completely fallacious. Wherein lies the fallacy?

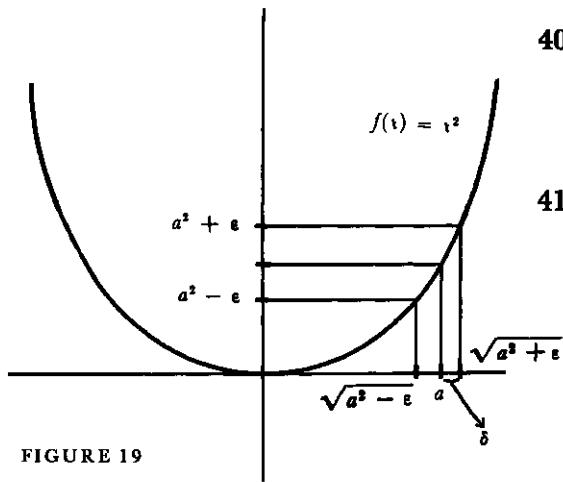


FIGURE 19

# CHAPTER 6

## CONTINUOUS FUNCTIONS

If  $f$  is an arbitrary function, it is not necessarily true that

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

In fact, there are many ways this can fail to be true. For example,  $f$  might not even be defined at  $a$ , in which case the equation makes no sense (Figure 1).

Again,  $\lim_{x \rightarrow a} f(x)$  might not exist (Figure 2). Finally, as illustrated in Figure 3, even if  $f$  is defined at  $a$  and  $\lim_{x \rightarrow a} f(x)$  exists, the limit might not equal  $f(a)$ .

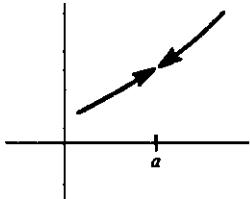


FIGURE 1

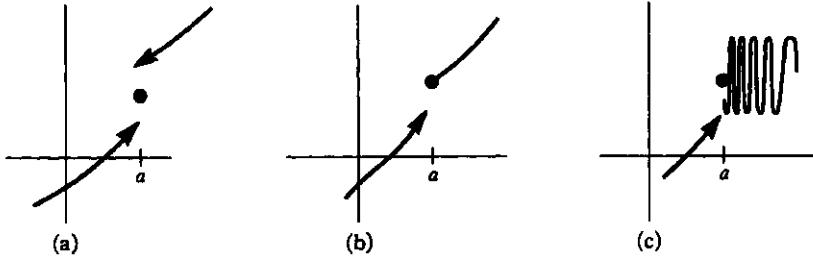


FIGURE 2

We would like to regard all behavior of this type as abnormal and honor, with some complimentary designation, functions which do not exhibit such peculiarities. The term which has been adopted is “continuous.” Intuitively, a function  $f$  is continuous if the graph contains no breaks, jumps, or wild oscillations. Although this description will usually enable you to decide whether a function is continuous simply by looking at its graph (a skill well worth cultivating) it is easy to be fooled, and the precise definition is *very* important.

### DEFINITION

The function  $f$  is **continuous at  $a$**  if

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

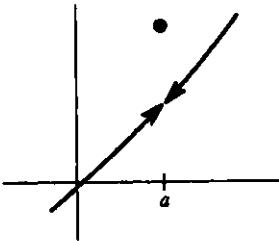


FIGURE 3

We will have no difficulty finding many examples of functions which are, or are not, continuous at some number  $a$ —every example involving limits provides an example about continuity, and Chapter 5 certainly provides enough of these.

The function  $f(x) = \sin 1/x$  is not continuous at 0, because it is not even defined at 0, and the same is true of the function  $g(x) = x \sin 1/x$ . On the other hand, if we are willing to extend the second of these functions, that is, if we wish to define

a new function  $G$  by

$$G(x) = \begin{cases} x \sin 1/x, & x \neq 0 \\ a, & x = 0, \end{cases}$$

then the choice of  $a = G(0)$  can be made in such a way that  $G$  will be continuous at 0—to do this we can (if fact, we must) define  $G(0) = 0$  (Figure 4). This sort of extension is not possible for  $f$ ; if we define

$$F(x) = \begin{cases} \sin 1/x, & x \neq 0 \\ a, & x = 0, \end{cases}$$

then  $F$  will not be continuous at 0, no matter what  $a$  is, because  $\lim_{x \rightarrow 0} f(x)$  does not exist.

The function

$$f(x) = \begin{cases} x, & x \text{ rational} \\ 0, & x \text{ irrational} \end{cases}$$

is not continuous at  $a$ , if  $a \neq 0$ , since  $\lim_{x \rightarrow a} f(x)$  does not exist. However,  $\lim_{x \rightarrow 0} f(x) = 0 = f(0)$ , so  $f$  is continuous at precisely one point, 0.

The functions  $f(x) = c$ ,  $g(x) = x$ , and  $h(x) = x^2$  are continuous at all numbers  $a$ , since

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

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

$$\lim_{x \rightarrow a} h(x) = \lim_{x \rightarrow a} x^2 = a^2 = h(a).$$

Finally, consider the function

$$f(x) = \begin{cases} 0, & x \text{ irrational} \\ 1/q, & x = p/q \text{ in lowest terms.} \end{cases}$$

In Chapter 5 we showed that  $\lim_{x \rightarrow a} f(x) = 0$  for all  $a$ . Since  $0 = f(a)$  only when  $a$  is irrational, this function is continuous at  $a$  if  $a$  is irrational, but not if  $a$  is rational.

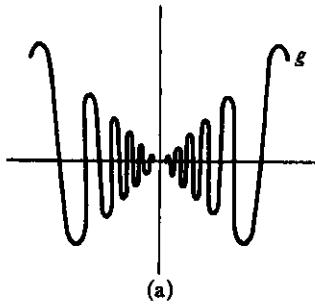
It is even easier to give examples of continuity if we prove two simple theorems.

**THEOREM 1** If  $f$  and  $g$  are continuous at  $a$ , then

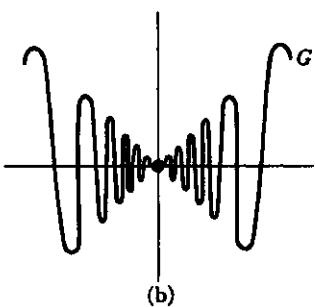
- (1)  $f + g$  is continuous at  $a$ ,
- (2)  $f \cdot g$  is continuous at  $a$ .

Moreover, if  $g(a) \neq 0$ , then

- (3)  $1/g$  is continuous at  $a$ .



(a)



(b)

FIGURE 4

**PROOF** Since  $f$  and  $g$  are continuous at  $a$ ,

$$\lim_{x \rightarrow a} f(x) = f(a) \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = g(a).$$

By Theorem 2(1) of Chapter 5 this implies that

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

which is just the assertion that  $f + g$  is continuous at  $a$ . The proofs of parts (2) and (3) are left to you. ■

Starting with the functions  $f(x) = c$  and  $f(x) = x$ , which are continuous at  $a$ , for every  $a$ , we can use Theorem 1 to conclude that a function

$$f(x) = \frac{b_n x^n + b_{n-1} x^{n-1} + \cdots + b_0}{c_m x^m + c_{m-1} x^{m-1} + \cdots + c_0}$$

is continuous at every point in its domain. But it is harder to get much further than that. When we discuss the sine function in detail it will be easy to prove that  $\sin$  is continuous at  $a$  for all  $a$ ; let us assume this fact meanwhile. A function like

$$f(x) = \frac{\sin^2 x + x^2 + x^4 \sin x}{\sin^2 x + 4x^2 \sin^2 x}$$

can now be proved continuous at every point in its domain. But we are still unable to prove the continuity of a function like  $f(x) = \sin(x^2)$ ; we obviously need a theorem about the composition of continuous functions. Before stating this theorem, the following point about the definition of continuity is worth noting. If we translate the equation  $\lim_{x \rightarrow a} f(x) = f(a)$  according to the definition of limits, we obtain

$$\begin{aligned} \text{for every } \varepsilon > 0 \text{ there is } \delta > 0 \text{ such that, for all } x, \\ \text{if } 0 < |x - a| < \delta, \text{ then } |f(x) - f(a)| < \varepsilon. \end{aligned}$$

But in this case, where the limit is  $f(a)$ , the phrase

$$0 < |x - a| < \delta$$

may be changed to the simpler condition

$$|x - a| < \delta,$$

since if  $x = a$  it is certainly true that  $|f(x) - f(a)| < \varepsilon$ .

**THEOREM 2** If  $g$  is continuous at  $a$ , and  $f$  is continuous at  $g(a)$ , then  $f \circ g$  is continuous at  $a$ . (Notice that  $f$  is required to be continuous at  $g(a)$ , not at  $a$ .)

**PROOF** Let  $\varepsilon > 0$ . We wish to find a  $\delta > 0$  such that for all  $x$ ,

$$\begin{aligned} \text{if } |x - a| < \delta, \text{ then } |(f \circ g)(x) - (f \circ g)(a)| < \varepsilon, \\ \text{i.e., } |f(g(x)) - f(g(a))| < \varepsilon. \end{aligned}$$

We first use continuity of  $f$  to estimate how close  $g(x)$  must be to  $g(a)$  in order for this inequality to hold. Since  $f$  is continuous at  $g(a)$ , there is a  $\delta' > 0$  such that for all  $y$ ,

$$(1) \quad \text{if } |y - g(a)| < \delta', \text{ then } |f(y) - f(g(a))| < \varepsilon.$$

In particular, this means that

$$(2) \quad \text{if } |g(x) - g(a)| < \delta', \text{ then } |f(g(x)) - f(g(a))| < \varepsilon.$$

We now use continuity of  $g$  to estimate how close  $x$  must be to  $a$  in order for the inequality  $|g(x) - g(a)| < \delta'$  to hold. The number  $\delta'$  is a positive number just like any other positive number; we can therefore take  $\delta'$  as the  $\varepsilon$  (!) in the definition of continuity of  $g$  at  $a$ . We conclude that there is a  $\delta > 0$  such that, for all  $x$ ,

$$(3) \quad \text{if } |x - a| < \delta, \text{ then } |g(x) - g(a)| < \delta'.$$

Combining (2) and (3) we see that for all  $x$ ,

$$\text{if } |x - a| < \delta, \text{ then } |f(g(x)) - f(g(a))| < \varepsilon. \blacksquare$$

We can now reconsider the function

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

We have already noted that  $f$  is continuous at 0. A few applications of Theorems 1 and 2, together with the continuity of  $\sin$ , show that  $f$  is also continuous at  $a$ , for  $a \neq 0$ . Functions like  $f(x) = \sin(x^2 + \sin(x + \sin^2(x^3)))$  should be equally easy for you to analyze.

The few theorems of this chapter have all been related to continuity of functions at a single point, but the concept of continuity doesn't begin to be really interesting until we focus our attention on functions which are continuous at all points of some interval. If  $f$  is continuous at  $x$  for all  $x$  in  $(a, b)$ , then  $f$  is called **continuous on  $(a, b)$** . Continuity on a closed interval must be defined a little differently; a function  $f$  is called **continuous on  $[a, b]$**  if

- (1)  $f$  is continuous at  $x$  for all  $x$  in  $(a, b)$ ,
- (2)  $\lim_{x \rightarrow a^+} f(x) = f(a)$  and  $\lim_{x \rightarrow b^-} f(x) = f(b)$ .

Functions which are continuous on an interval are usually regarded as especially well behaved; indeed continuity might be specified as the first condition which a “reasonable” function ought to satisfy. A continuous function is sometimes described, intuitively, as one whose graph can be drawn without lifting your pencil from the paper. Consideration of the function

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

shows that this description is a little too optimistic, but it is nevertheless true that there are many important results involving functions which are continuous on an interval. There theorems are generally much harder than the ones in this chapter,

but there is a simple theorem which forms a bridge between the two kinds of results. The hypothesis of this theorem requires continuity at only a single point, but the conclusion describes the behavior of the function on some interval containing the point. Although this theorem is really a lemma for later arguments, it is included here as a preview of things to come.

**THEOREM 3** Suppose  $f$  is continuous at  $a$ , and  $f(a) > 0$ . Then there is a number  $\delta > 0$  such that  $f(x) > 0$  for all  $x$  satisfying  $|x - a| < \delta$ . Similarly, if  $f(a) < 0$ , then there is a number  $\delta > 0$  such that  $f(x) < 0$  for all  $x$  satisfying  $|x - a| < \delta$ .

**PROOF** Consider the case  $f(a) > 0$ . Since  $f$  is continuous at  $a$ , if  $\varepsilon > 0$  there is a  $\delta > 0$  such that, for all  $x$ ,

$$\text{if } |x - a| < \delta, \text{ then } |f(x) - f(a)| < \varepsilon.$$

Since  $f(a) > 0$  we can take  $f(a)$  as the  $\varepsilon$ . Thus there is  $\delta > 0$  so that for all  $x$ ,

$$\text{if } |x - a| < \delta, \text{ then } |f(x) - f(a)| < f(a),$$

and this last inequality implies  $f(x) > 0$ .

A similar proof can be given in the case  $f(a) < 0$ ; take  $\varepsilon = -f(a)$ . Or one can apply the first case to the function  $-f$ . ■

### PROBLEMS

1. For which of the following functions  $f$  is there a continuous function  $F$  with domain  $\mathbf{R}$  such that  $F(x) = f(x)$  for all  $x$  in the domain of  $f$ ?
  - (i)  $f(x) = \frac{x^2 - 4}{x - 2}$ .
  - (ii)  $f(x) = \frac{|x|}{x}$ .
  - (iii)  $f(x) = 0$ ,  $x$  irrational.
  - (iv)  $f(x) = 1/q$ ,  $x = p/q$  rational in lowest terms.
2. At which points are the functions of Problems 4-17 and 4-19 continuous?
3. (a) Suppose that  $f$  is a function satisfying  $|f(x)| \leq |x|$  for all  $x$ . Show that  $f$  is continuous at 0. (Notice that  $f(0)$  must equal 0.)  
 (b) Give an example of such a function  $f$  which is not continuous at any  $a \neq 0$ .  
 (c) Suppose that  $g$  is continuous at 0 and  $g(0) = 0$ , and  $|f(x)| \leq |g(x)|$ . Prove that  $f$  is continuous at 0.
4. Give an example of a function  $f$  such that  $f$  is continuous nowhere, but  $|f|$  is continuous everywhere.
5. For each number  $a$ , find a function which is continuous at  $a$ , but not at any other points.

6. (a) Find a function  $f$  which is discontinuous at  $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$  but continuous at all other points.  
 (b) Find a function  $f$  which is discontinuous at  $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$ , and at 0, but continuous at all other points.
7. Suppose that  $f$  satisfies  $f(x + y) = f(x) + f(y)$ , and that  $f$  is continuous at 0. Prove that  $f$  is continuous at  $a$  for all  $a$ .
8. Suppose that  $f$  is continuous at  $a$  and  $f(a) = 0$ . Prove that if  $\alpha \neq 0$ , then  $f + \alpha$  is nonzero in some open interval containing  $a$ .
9. (a) Suppose  $f$  is *not* continuous at  $a$ . Prove that for some number  $\varepsilon > 0$  there are numbers  $x$  arbitrarily close to  $a$  with  $|f(x) - f(a)| > \varepsilon$ . Illustrate graphically.  
 (b) Conclude that for some number  $\varepsilon > 0$  *either* there are numbers  $x$  arbitrarily close to  $a$  with  $f(x) < f(a) - \varepsilon$  *or* there are numbers  $x$  arbitrarily close to  $a$  with  $f(x) > f(a) + \varepsilon$ .
10. (a) Prove that if  $f$  is continuous at  $a$ , then so is  $|f|$ .  
 (b) Prove that every continuous  $f$  can be written  $f = E + O$ , where  $E$  is even and continuous and  $O$  is odd and continuous.  
 (c) Prove that if  $f$  and  $g$  are continuous, then so are  $\max(f, g)$  and  $\min(f, g)$ .  
 (d) Prove that every continuous  $f$  can be written  $f = g - h$ , where  $g$  and  $h$  are nonnegative and continuous.
11. Prove Theorem 1(3) by using Theorem 2 and continuity of the function  $f(x) = 1/x$ .
- \*12. (a) Prove that if  $f$  is continuous at  $l$  and  $\lim_{x \rightarrow a} g(x) = l$ , then  $\lim_{x \rightarrow a} f(g(x)) = f(l)$ . (You can go right back to the definitions, but it is easier to consider the function  $G$  with  $G(x) = g(x)$  for  $x \neq a$ , and  $G(a) = l$ .)  
 (b) Show that if continuity of  $f$  at  $l$  is not assumed, then it is not generally true that  $\lim_{x \rightarrow a} f(g(x)) = f(\lim_{x \rightarrow a} g(x))$ . Hint: Try  $f(x) = 0$  for  $x \neq l$ , and  $f(l) = 1$ .
13. (a) Prove that if  $f$  is continuous on  $[a, b]$ , then there is a function  $g$  which is continuous on  $\mathbf{R}$ , and which satisfies  $g(x) = f(x)$  for all  $x$  in  $[a, b]$ . Hint: Since you obviously have a great deal of choice, try making  $g$  constant on  $(-\infty, a]$  and  $[b, \infty)$ .  
 (b) Give an example to show that this assertion is false if  $[a, b]$  is replaced by  $(a, b)$ .
14. (a) Suppose that  $g$  and  $h$  are continuous at  $a$ , and that  $g(a) = h(a)$ . Define  $f(x)$  to be  $g(x)$  if  $x \geq a$  and  $h(x)$  if  $x \leq a$ . Prove that  $f$  is continuous at  $a$ .  
 (b) Suppose  $g$  is continuous on  $[a, b]$  and  $h$  is continuous on  $[b, c]$  and  $g(b) = h(b)$ . Let  $f(x)$  be  $g(x)$  for  $x$  in  $[a, b]$  and  $h(x)$  for  $x$  in  $[b, c]$ .

Show that  $f$  is continuous on  $[a, c]$ . (Thus, continuous functions can be “pasted together”.)

15. (a) Prove the following version of Theorem 3 for “right-hand continuity”: Suppose that  $\lim_{x \rightarrow a^+} f(x) = f(a)$ , and  $f(a) > 0$ . Then there is a number  $\delta > 0$  such that  $f(x) > 0$  for all  $x$  satisfying  $0 \leq x - a < \delta$ . Similarly, if  $f(a) < 0$ , then there is a number  $\delta > 0$  such that  $f(x) < 0$  for all  $x$  satisfying  $0 \leq x - a < \delta$ .
- (b) Prove a version of Theorem 3 when  $\lim_{x \rightarrow b^-} f(x) = f(b)$ .
16. If  $\lim_{x \rightarrow a} f(x)$  exists, but is  $\neq f(a)$ , then  $f$  is said to have a **removable discontinuity** at  $a$ .
- (a) If  $f(x) = \sin 1/x$  for  $x \neq 0$  and  $f(0) = 1$ , does  $f$  have a removable discontinuity at 0? What if  $f(x) = x \sin 1/x$  for  $x \neq 0$ , and  $f(0) = 1$ ?
- (b) Suppose  $f$  has a removable discontinuity at  $a$ . Let  $g(x) = f(x)$  for  $x \neq a$ , and let  $g(a) = \lim_{x \rightarrow a} f(x)$ . Prove that  $g$  is continuous at  $a$ . (Don’t work very hard; this is quite easy.)
- (c) Let  $f(x) = 0$  if  $x$  is irrational, and let  $f(p/q) = 1/q$  if  $p/q$  is in lowest terms. What is the function  $g$  defined by  $g(x) = \lim_{y \rightarrow x} f(y)$ ?
- \*(d) Let  $f$  be a function with the property that every point of discontinuity is a removable discontinuity. This means that  $\lim_{y \rightarrow x} f(y)$  exists for all  $x$ , but  $f$  may be discontinuous at some (even infinitely many) numbers  $x$ . Define  $g(x) = \lim_{y \rightarrow x} f(y)$ . Prove that  $g$  is continuous. (This is not quite so easy as part (b).)
- \*\*(e) Is there a function  $f$  which is discontinuous at every point, and which has only removable discontinuities? (It is worth thinking about this problem now, but mainly as a test of intuition; even if you suspect the correct answer, you will almost certainly be unable to prove it at the present time. See Problem 22-33.)

# CHAPTER 7 THREE HARD THEOREMS

This chapter is devoted to three theorems about continuous functions, and some of their consequences. The proofs of the three theorems themselves will not be given until the next chapter, for reasons which are explained at the end of this chapter.

**THEOREM 1** If  $f$  is continuous on  $[a, b]$  and  $f(a) < 0 < f(b)$ , then there is some  $x$  in  $[a, b]$  such that  $f(x) = 0$ .

(Geometrically, this means that the graph of a continuous function which starts below the horizontal axis and ends above it must cross this axis at some point, as in Figure 1.)

**THEOREM 2** If  $f$  is continuous on  $[a, b]$ , then  $f$  is bounded above on  $[a, b]$ , that is, there is some number  $N$  such that  $f(x) \leq N$  for all  $x$  in  $[a, b]$ .

(Geometrically, this theorem means that the graph of  $f$  lies below some line parallel to the horizontal axis, as in Figure 2.)

**THEOREM 3** If  $f$  is continuous on  $[a, b]$ , then there is some number  $y$  in  $[a, b]$  such that  $f(y) \geq f(x)$  for all  $x$  in  $[a, b]$  (Figure 3).

These three theorems differ markedly from the theorems of Chapter 6. The hypotheses of those theorems always involved continuity at a single point, while

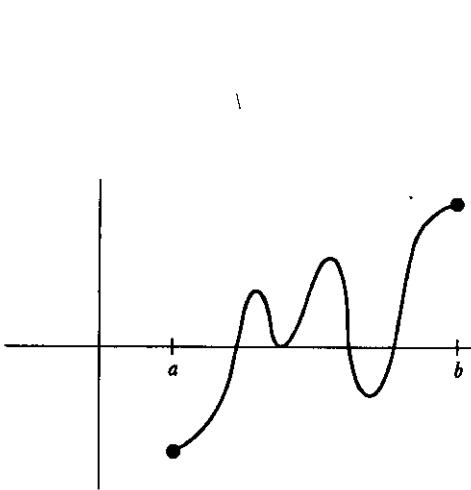


FIGURE 1

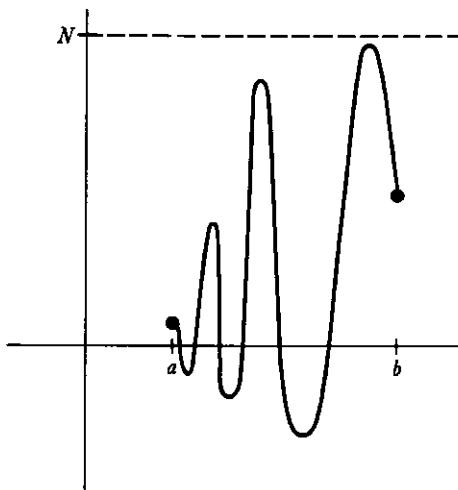


FIGURE 2

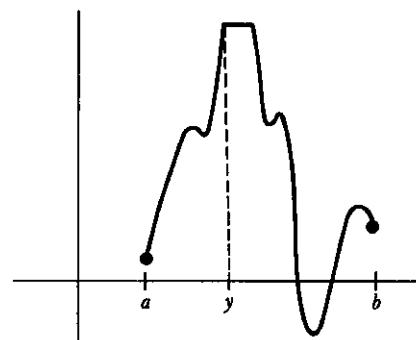


FIGURE 3

the hypotheses of the present theorems require continuity on a whole interval  $[a, b]$ —if continuity fails to hold at a single point, the conclusions may fail. For example, let  $f$  be the function shown in Figure 4,

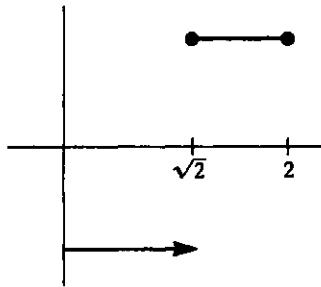


FIGURE 4

$$f(x) = \begin{cases} -1, & 0 \leq x < \sqrt{2} \\ 1, & \sqrt{2} \leq x \leq 2. \end{cases}$$

Then  $f$  is continuous at every point of  $[0, 2]$  except  $\sqrt{2}$ , and  $f(0) < 0 < f(2)$ , but there is no point  $x$  in  $[0, 2]$  such that  $f(x) = 0$ ; the discontinuity at the single point  $\sqrt{2}$  is sufficient to destroy the conclusion of Theorem 1.

Similarly, suppose that  $f$  is the function shown in Figure 5,

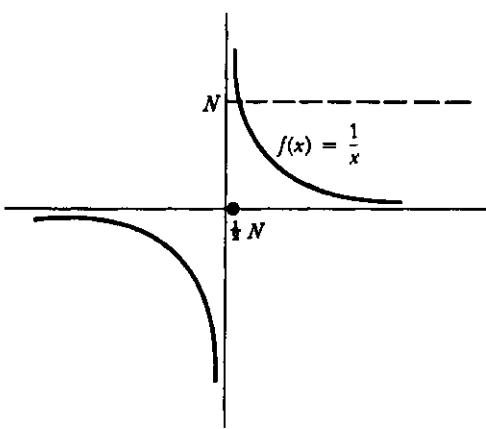


FIGURE 5

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

Then  $f$  is continuous at every point of  $[0, 1]$  except 0, but  $f$  is not bounded above on  $[0, 1]$ . In fact, for any number  $N > 0$  we have  $f(1/2N) = 2N > N$ .

This example also shows that the closed interval  $[a, b]$  in Theorem 2 cannot be replaced by the open interval  $(a, b)$ , for the function  $f$  is continuous on  $(0, 1)$ , but is not bounded there.

Finally, consider the function shown in Figure 6,

$$f(x) = \begin{cases} x^2, & x < 1 \\ 0, & x \geq 1. \end{cases}$$

On the interval  $[0, 1]$  the function  $f$  is bounded above, so  $f$  does satisfy the conclusion of Theorem 2, even though  $f$  is not continuous on  $[0, 1]$ . But  $f$  does not satisfy the conclusion of Theorem 3—there is no  $y$  in  $[0, 1]$  such that  $f(y) \geq f(x)$  for all  $x$  in  $[0, 1]$ ; in fact, it is certainly not true that  $f(1) \geq f(x)$  for all  $x$  in  $[0, 1]$  so we cannot choose  $y = 1$ , nor can we choose  $0 \leq y < 1$  because  $f(y) < f(x)$  if  $x$  is any number with  $y < x < 1$ .

This example shows that Theorem 3 is considerably stronger than Theorem 2. Theorem 3 is often paraphrased by saying that a continuous function on a closed interval “takes on its maximum value” on that interval.

As a compensation for the stringency of the hypotheses of our three theorems, the conclusions are of a totally different order than those of previous theorems. They describe the behavior of a function, not just near a point, but on a whole interval; such “global” properties of a function are always significantly more difficult to prove than “local” properties, and are correspondingly of much greater power. To illustrate the usefulness of Theorems 1, 2, and 3, we will soon deduce some important consequences, but it will help to first mention some simple generalizations of these theorems.

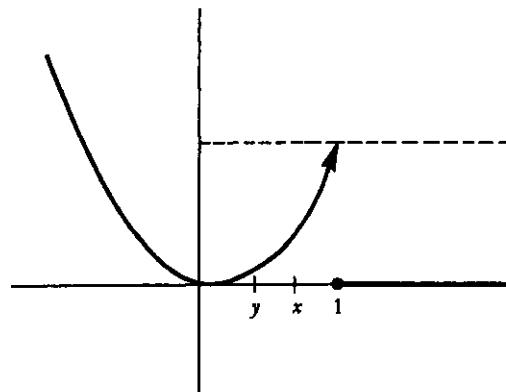


FIGURE 6

**THEOREM 4** If  $f$  is continuous on  $[a, b]$  and  $f(a) < c < f(b)$ , then there is some  $x$  in  $[a, b]$  such that  $f(x) = c$ .

**PROOF** Let  $g = f - c$ . Then  $g$  is continuous, and  $g(a) < 0 < g(b)$ . By Theorem 1, there is some  $x$  in  $[a, b]$  such that  $g(x) = 0$ . But this means that  $f(x) = c$ . ■

**THEOREM 5** If  $f$  is continuous on  $[a, b]$  and  $f(a) > c > f(b)$ , then there is some  $x$  in  $[a, b]$  such that  $f(x) = c$ .

**PROOF** The function  $-f$  is continuous on  $[a, b]$  and  $-f(a) < -c < -f(b)$ . By Theorem 4 there is some  $x$  in  $[a, b]$  such that  $-f(x) = -c$ , which means that  $f(x) = c$ . ■

Theorems 4 and 5 together show that  $f$  takes on any value between  $f(a)$  and  $f(b)$ . We can do even better than this: if  $c$  and  $d$  are in  $[a, b]$ , then  $f$  takes on any value between  $f(c)$  and  $f(d)$ . The proof is simple: if, for example,  $c < d$ , then just apply Theorems 4 and 5 to the interval  $[c, d]$ . Summarizing, if a continuous function on an interval takes on two values, it takes on every value in between; this slight generalization of Theorem 1 is often called the Intermediate Value Theorem.

**THEOREM 6** If  $f$  is continuous on  $[a, b]$ , then  $f$  is bounded below on  $[a, b]$ , that is, there is some number  $N$  such that  $f(x) \geq N$  for all  $x$  in  $[a, b]$ .

**PROOF** The function  $-f$  is continuous on  $[a, b]$ , so by Theorem 2 there is a number  $M$  such that  $-f(x) \leq M$  for all  $x$  in  $[a, b]$ . But this means that  $f(x) \geq -M$  for all  $x$  in  $[a, b]$ , so we can let  $N = -M$ . ■

Theorems 2 and 6 together show that a continuous function  $f$  on  $[a, b]$  is bounded on  $[a, b]$ , that is, there is a number  $N$  such that  $|f(x)| \leq N$  for all  $x$  in  $[a, b]$ . In fact, since Theorem 2 ensures the existence of a number  $N_1$  such that  $f(x) \leq N_1$  for all  $x$  in  $[a, b]$ , and Theorem 6 ensures the existence of a number  $N_2$  such that  $f(x) \geq N_2$  for all  $x$  in  $[a, b]$ , we can take  $N = \max(|N_1|, |N_2|)$ .

**THEOREM 7** If  $f$  is continuous on  $[a, b]$ , then there is some  $y$  in  $[a, b]$  such that  $f(y) \leq f(x)$  for all  $x$  in  $[a, b]$ .

(A continuous function on a closed interval takes on its minimum value on that interval.)

**PROOF** The function  $-f$  is continuous on  $[a, b]$ ; by Theorem 3 there is some  $y$  in  $[a, b]$  such that  $-f(y) \geq -f(x)$  for all  $x$  in  $[a, b]$ , which means that  $f(y) \leq f(x)$  for all  $x$  in  $[a, b]$ . ■

Now that we have derived the trivial consequences of Theorems 1, 2, and 3, we can begin proving a few interesting things.

**THEOREM 8** Every positive number has a square root. In other words, if  $\alpha > 0$ , then there is some number  $x$  such that  $x^2 = \alpha$ .

**PROOF** Consider the function  $f(x) = x^2$ , which is certainly continuous. Notice that the statement of the theorem can be expressed in terms of  $f$ : “the number  $\alpha$  has a square root” means that  $f$  takes on the value  $\alpha$ . The proof of this fact about  $f$  will be an easy consequence of Theorem 4.

There is obviously a number  $b > 0$  such that  $f(b) > \alpha$  (as illustrated in Figure 7); in fact, if  $\alpha > 1$  we can take  $b = \alpha$ , while if  $\alpha < 1$  we can take  $b = 1$ . Since  $f(0) < \alpha < f(b)$ , Theorem 4 applied to  $[0, b]$  implies that for some  $x$  (in  $[0, b]$ ), we have  $f(x) = \alpha$ , i.e.,  $x^2 = \alpha$ . ■

Precisely the same argument can be used to prove that a positive number has an  $n$ th root, for any natural number  $n$ . If  $n$  happens to be odd, one can do better: *every* number has an  $n$ th root. To prove this we just note that if the positive number  $\alpha$  has the  $n$ th root  $x$ , i.e., if  $x^n = \alpha$ , then  $(-x)^n = -\alpha$  (since  $n$  is odd), so  $-\alpha$  has the  $n$ th root  $-x$ . The assertion, that for odd  $n$  any number  $\alpha$  has an  $n$ th root, is equivalent to the statement that the equation

$$x^n - \alpha = 0$$

has a root if  $n$  is odd. Expressed in this way the result is susceptible of great generalization.

**THEOREM 9** If  $n$  is odd, then any equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_0 = 0$$

has a root.

**PROOF** We obviously want to consider the function

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0;$$

we would like to prove that  $f$  is sometimes positive and sometimes negative. The intuitive idea is that for large  $|x|$ , the function is very much like  $g(x) = x^n$  and, since  $n$  is odd, this function is positive for large positive  $x$  and negative for large negative  $x$ . A little algebra is all we need to make this intuitive idea work.

The proper analysis of the function  $f$  depends on writing

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0 = x^n \left( 1 + \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n} \right).$$

Note that

$$\left| \frac{a_{n-1}}{x} + \frac{a_{n-2}}{x^2} + \cdots + \frac{a_0}{x^n} \right| \leq \frac{|a_{n-1}|}{|x|} + \cdots + \frac{|a_0|}{|x^n|}.$$

Consequently, if we choose  $x$  satisfying

$$(*) \quad |x| > 1, 2n|a_{n-1}|, \dots, 2n|a_0|,$$

then  $|x^k| > |x|$  and

$$\frac{|a_{n-k}|}{|x^k|} < \frac{|a_{n-k}|}{|x|} < \frac{|a_{n-k}|}{2n|a_{n-k}|} = \frac{1}{2n},$$

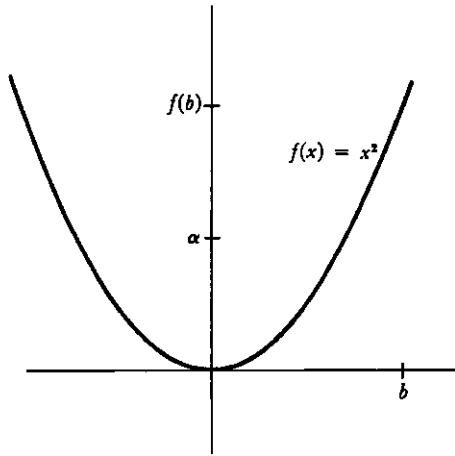


FIGURE 7

so

$$\left| \frac{a_{n-1}}{x} + \frac{a_{n-2}}{x^2} + \cdots + \frac{a_0}{x^n} \right| \leq \underbrace{\frac{1}{2n} + \cdots + \frac{1}{2n}}_{n \text{ terms}} = \frac{1}{2}.$$

In other words,

$$-\frac{1}{2} \leq \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n} \leq \frac{1}{2},$$

which implies that

$$\frac{1}{2} \leq 1 + \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n}.$$

Therefore, if we choose an  $x_1 > 0$  which satisfies (\*), then

$$\frac{(x_1)^n}{2} \leq (x_1)^n \left( 1 + \frac{a_{n-1}}{x_1} + \cdots + \frac{a_0}{(x_1)^n} \right) = f(x_1),$$

so that  $f(x_1) > 0$ . On the other hand, if  $x_2 < 0$  satisfies (\*), then  $(x_2)^n < 0$  and

$$\frac{(x_2)^n}{2} \geq (x_2)^n \left( 1 + \frac{a_{n-1}}{x_2} + \cdots + \frac{a_0}{(x_2)^n} \right) = f(x_2),$$

so that  $f(x_2) < 0$ .

Now applying Theorem 1 to the interval  $[x_2, x_1]$  we conclude that there is an  $x$  in  $[x_2, x_1]$  such that  $f(x) = 0$ . ■

Theorem 9 disposes of the problem of odd degree equations so happily that it would be frustrating to leave the problem of even degree equations completely undiscussed. At first sight, however, the problem seems insuperable. Some equations, like  $x^2 - 1 = 0$ , have a solution, and some, like  $x^2 + 1 = 0$ , do not—what more is there to say? If we are willing to consider a more general question, however, something interesting *can* be said. Instead of trying to solve the equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_0 = 0,$$

let us ask about the possibility of solving the equations

$$x^n + a_{n-1}x^{n-1} + \cdots + a_0 = c$$

for all possible numbers  $c$ . This amounts to allowing the constant term  $a_0$  to vary. The information which can be given concerning the solution of these equations depends on a fact which is illustrated in Figure 8.

The graph of the function  $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ , with  $n$  even, contains, at least the way we have drawn it, a lowest point. In other words, there is a number  $y$  such that  $f(y) \leq f(x)$  for all numbers  $x$ —the function  $f$  takes on a minimum value, not just on each closed interval, but on the whole line. (Notice that this is false if  $n$  is odd.) The proof depends on Theorem 7, but a tricky application will be required. We can apply Theorem 7 to any interval  $[a, b]$ , and obtain a point  $y_0$  such that  $f(y_0)$  is the minimum value of  $f$  on  $[a, b]$ ; but if  $[a, b]$  happens to be the interval shown in Figure 8, for example, then the point  $y_0$  will not be the place where  $f$  has its minimum value for the whole line. In the next

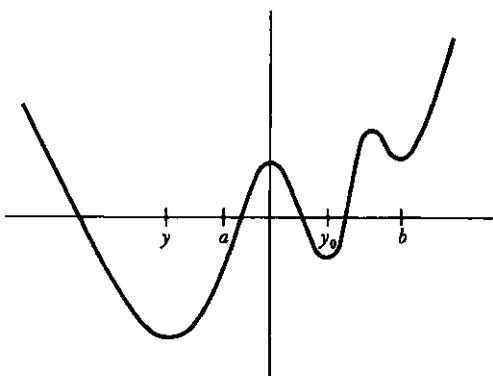


FIGURE 8

theorem the entire point of the proof is to choose an interval  $[a, b]$  in such a way that this cannot happen.

**THEOREM 10** If  $n$  is even and  $f(x) = x^n + a_{n-1}x^{n-1} + \dots + a_0$ , then there is a number  $y$  such that  $f(y) \leq f(x)$  for all  $x$ .

**PROOF** As in the proof of Theorem 9, if

$$M = \max(1, 2n|a_{n-1}|, \dots, 2n|a_0|),$$

then for all  $x$  with  $|x| \geq M$ , we have

$$\frac{1}{2} \leq 1 + \frac{a_{n-1}}{x} + \dots + \frac{a_0}{x^n}.$$

Since  $n$  is even,  $x^n \geq 0$  for all  $x$ , so

$$\frac{x^n}{2} \leq x^n \left(1 + \frac{a_{n-1}}{x} + \dots + \frac{a_0}{x^n}\right) = f(x),$$

provided that  $|x| \geq M$ . Now consider the number  $f(0)$ . Let  $b > 0$  be a number such that  $b^n \geq 2f(0)$  and also  $b > M$ . Then, if  $x \geq b$ , we have (Figure 9)

$$f(x) \geq \frac{x^n}{2} \geq \frac{b^n}{2} \geq f(0).$$

Similarly, if  $x \leq -b$ , then

$$f(x) \geq \frac{x^n}{2} \geq \frac{(-b)^n}{2} = \frac{b^n}{2} \geq f(0).$$

Summarizing:

$$\text{if } x \geq b \text{ or } x \leq -b, \text{ then } f(x) \geq f(0).$$

Now apply Theorem 7 to the function  $f$  on the interval  $[-b, b]$ . We conclude that there is a number  $y$  such that

$$(1) \quad \text{if } -b \leq x \leq b, \text{ then } f(y) \leq f(x).$$

In particular,  $f(y) \leq f(0)$ . Thus

$$(2) \quad \text{if } x \leq -b \text{ or } x \geq b, \text{ then } f(x) \geq f(0) \geq f(y).$$

Combining (1) and (2) we see that  $f(y) \leq f(x)$  for all  $x$ . ■

Theorem 10 now allows us to prove the following result.

**THEOREM 11** Consider the equation

$$(*) \quad x^n + a_{n-1}x^{n-1} + \dots + a_0 = c,$$

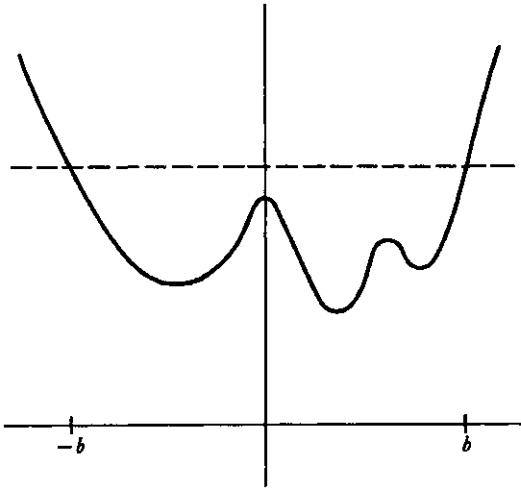


FIGURE 9

and suppose  $n$  is even. Then there is a number  $m$  such that  $(*)$  has a solution for  $c \geq m$  and has no solution for  $c < m$ .

PROOF

Let  $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$  (Figure 10).

According to Theorem 10 there is a number  $y$  such that  $f(y) \leq f(x)$  for all  $x$ . Let  $m = f(y)$ . If  $c < m$ , then the equation  $(*)$  obviously has no solution, since the left side always has a value  $\geq m$ . If  $c = m$ , then  $(*)$  has  $y$  as a solution. Finally, suppose  $c > m$ . Let  $b$  be a number such that  $b > y$  and  $f(b) > c$ . Then  $f(y) = m < c < f(b)$ . Consequently, by Theorem 4, there is some number  $x$  in  $[y, b]$  such that  $f(x) = c$ , so  $x$  is a solution of  $(*)$ . ■

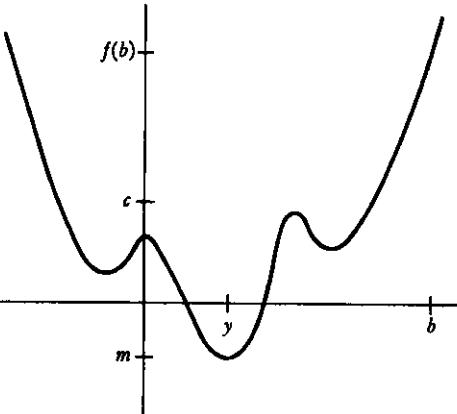


FIGURE 10

These consequences of Theorems 1, 2, and 3 are the only ones we will derive now (these theorems will play a fundamental role in everything we do later, however). Only one task remains—to prove Theorems 1, 2, and 3. Unfortunately, we cannot hope to do this—on the basis of our present knowledge about the real numbers (namely, P1–P12) a proof is *impossible*. There are several ways of convincing ourselves that this gloomy conclusion is actually the case. For example, the proof of Theorem 8 relies only on the proof of Theorem 1; if we could prove Theorem 1, then the proof of Theorem 8 would be complete, and we would have a proof that every positive number has a square root. As pointed out in Part I, it is impossible to prove this on the basis of P1–P12. Again, suppose we consider the function

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

If there were no number  $x$  with  $x^2 = 2$ , then  $f$  would be continuous, since the denominator would never = 0. But  $f$  is not bounded on  $[0, 2]$ . So Theorem 2 depends essentially on the existence of numbers other than rational numbers, and therefore on some property of the real numbers other than P1–P12.

Despite our inability to prove Theorems 1, 2, and 3, they are certainly results which we want to be true. If the pictures we have been drawing have any connection with the mathematics we are doing, if our notion of continuous function corresponds to any degree with our intuitive notion, Theorems 1, 2, and 3 have got to be true. Since a proof of any of these theorems must require some new property of  $\mathbf{R}$  which has so far been overlooked, our present difficulties suggest a way to discover that property: let us try to construct a proof of Theorem 1, for example, and see what goes wrong.

One idea which seems promising is to locate the first point where  $f(x) = 0$ , that is, the smallest  $x$  in  $[a, b]$  such that  $f(x) = 0$ . To find this point, first consider the set  $A$  which contains all numbers  $x$  in  $[a, b]$  such that  $f$  is negative on  $[a, x]$ . In Figure 11,  $x$  is such a point, while  $x'$  is not. The set  $A$  itself is indicated by a heavy line. Since  $f$  is negative at  $a$ , and positive at  $b$ , the set  $A$  contains some points greater than  $a$ , while all points sufficiently close to  $b$  are not in  $A$ . (We are here using the continuity of  $f$  on  $[a, b]$ , as well as Problem 6-15.)

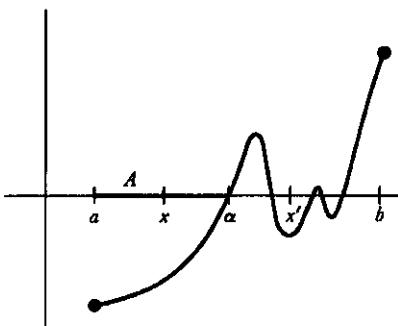


FIGURE 11