

(b) A hard problem: Prove that there is an integral point at a distance of less than  $\frac{1}{1000}$ \* from the straight line  $y = \sqrt{3}x$ .

In solving Problems 5 to 9 make use of the graphs of suitably chosen functions.

5. How many solutions are there for the following equations?

$$(a) -x^2 + x - 1 = |x|;$$

$$(b) |3x^2 + 12x + 9| + x = 0;$$

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

$$(d) |x - 1| + |x - 2| + |x + 1| + |x + 2| = 6;$$

$$(e) x(x + 1)(x + 2) = 0.01;$$

$$(f) |x + 3| = |x + 2|(x^2 - 1);$$

$$(g) [x] = x \text{ in the interval } |x| < 3;$$

$$(h) \frac{1}{x} + \frac{1}{x+1} + \frac{1}{x+2} = 100.$$

6. Solve the equations:

$$(a) 2x^2 - x - 1 = |x|;$$

$$(b) |2x^2 - x - 1| - x = 0;$$

$$(c) |x| = |x - 1| + |x - 2|.$$

7. (a) Determine how many solutions the equation

$$|1 - |x|| = a$$

can have for different values of  $a$ .

(b) The same question for the equation

$$x^2 + \frac{1}{x} = a.\dagger$$

\*The number  $\frac{1}{1000}$  can be replaced by any other number. Then it will be proved that however small a number is taken, there is a point with integral coordinates removed at a distance from the straight line  $y = \sqrt{3}x$  that is less than this number.

†The value of  $a$  separating the different cases can be found approximately from the graph.

8. Solve the inequalities:

- (a)  $\frac{2-x}{x^2+6x+5} > 0;$
- (b)  $x \leq |x^2 - x|;$
- (c)  $|x| + 2|x + 1| > 3.$

9. Find the largest value of the function and determine for which values of  $x$  it is reached:

- (a)  $y = x(a - x);$
- (b)  $y = |x|(a - |x|);$
- (c)  $y = x^2(a - x^2);$
- (d)  $y = \frac{x^2 + 4}{x^2 + x + 1};$
- (e)  $y = 1 - \sqrt{2x}$  in the interval  $|x| \leq \sqrt{2};$
- (f)  $y = -x^2 + 2x - 2$  in the interval  $-5 \leq x \leq 0;$
- (g)  $y = \frac{x+3}{x-1}$  in the interval  $x \geq 2.$

10. Two roads intersect at a right angle. Two cars drive toward the intersection: on the first road at a speed of 60 km/hr, on the second at a speed of 30 km/hr. At noon both cars are 10 km from the intersection.

At what moment will the distance between the cars be least? Where will the cars be at this moment?

11. Among all right triangles with given perimeter  $p$ , find the triangle having the largest area.

12. Suppose  $y = f(x)$  is an even function, and  $y = g(x)$  is an odd function. What can be said about the parity of the following functions?

$$\begin{array}{ll} y = f(x) + g(x); & y = f(x)g(x); \\ y = |g(x)|; & y = f(x) - g(x); \\ y = f(|x|) - g(x); & y = f(x) - g(|x|). \end{array}$$

13. Find all even and all odd functions of the form:

- (a)  $y = kx + b;$

$$(b) \quad y = \frac{px + q}{x + r};$$

$$(c) \quad y = \frac{ax^2 + bx + c}{x^2 + px + q}.$$

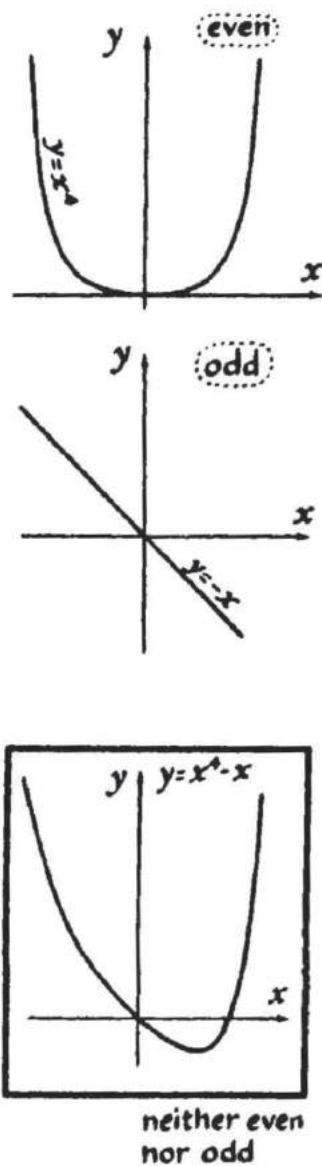


Fig. 2

14. The function  $y = x^4 - x$  is neither even nor odd. However, this function is easily represented in the form of a sum of an even function  $y = x^4$  and of an odd function  $y = -x$  (Fig. 2).

(a) Represent the function  $y = 1/(x^4 - x)$  as a sum of an even and an odd function.

(b) Prove that any function  $f(x)$  can be represented as a sum of an even and an odd function.  $\oplus$

15. Through any two points with different abscissas there passes a straight line (the graph of a linear function  $y = kx + b$ ). Analogously, through any three points with different abscissas and which are not on one straight line, it is possible to draw a parabola, the graph of a function  $y = ax^2 + bx + c$ .

Find the coefficients of the quadratic trinomial  $ax^2 + bx + c$ , whose graph passes through the points:

$$(a) (-1, 0); (0, 2); (1, 0);$$

$$(b) (1, 0); (4, 0); (5, 6);$$

$$(c) (-6, 7); (-4, -1); (-2, 7);$$

$$(d) (0, -4); (1, -3); (2, -1);$$

$$(e) (-1, 9); (3, 1); (6, 16).$$

16. (a) Carry out a similarity transformation of the parabola  $y = x^2$ , choosing the center of similarity at the origin and a ratio of similarity equal to 2. What curve is obtained?

(b) What similarity transformation transforms the curve  $y = x^2$  into the curve  $y = 5x^2$ ?

(c) Using the result of Problem 16b, find the focus and the directrix of the parabola  $y = 4x^2$ . (Definitions of directrix and focus are on p. 43.)

(d) Prove that all parabolas  $y = ax^2 + bx + c$  are geometrically similar.

17. Prove that the point  $F(0, \frac{1}{4})$  is the focus and the straight line  $y = -\frac{1}{4}$  is the directrix of the parabola  $y = x^2$ ; that is, any point of this parabola is equidistant from the point  $F(0, \frac{1}{4})$  and the straight line  $y = -\frac{1}{4}$ .

**Hint.** Take, on the parabola, some point  $M$  with the coordinates  $(a, a^2)$ . Write down the distance of this point from the point  $F(0, \frac{1}{4})$ , using the formula for the distance between two points.\* Then write down the distance from the point  $M(a, a^2)$  to the straight line  $y = -\frac{1}{4}$ .†

Prove the equality of the two resulting expressions.

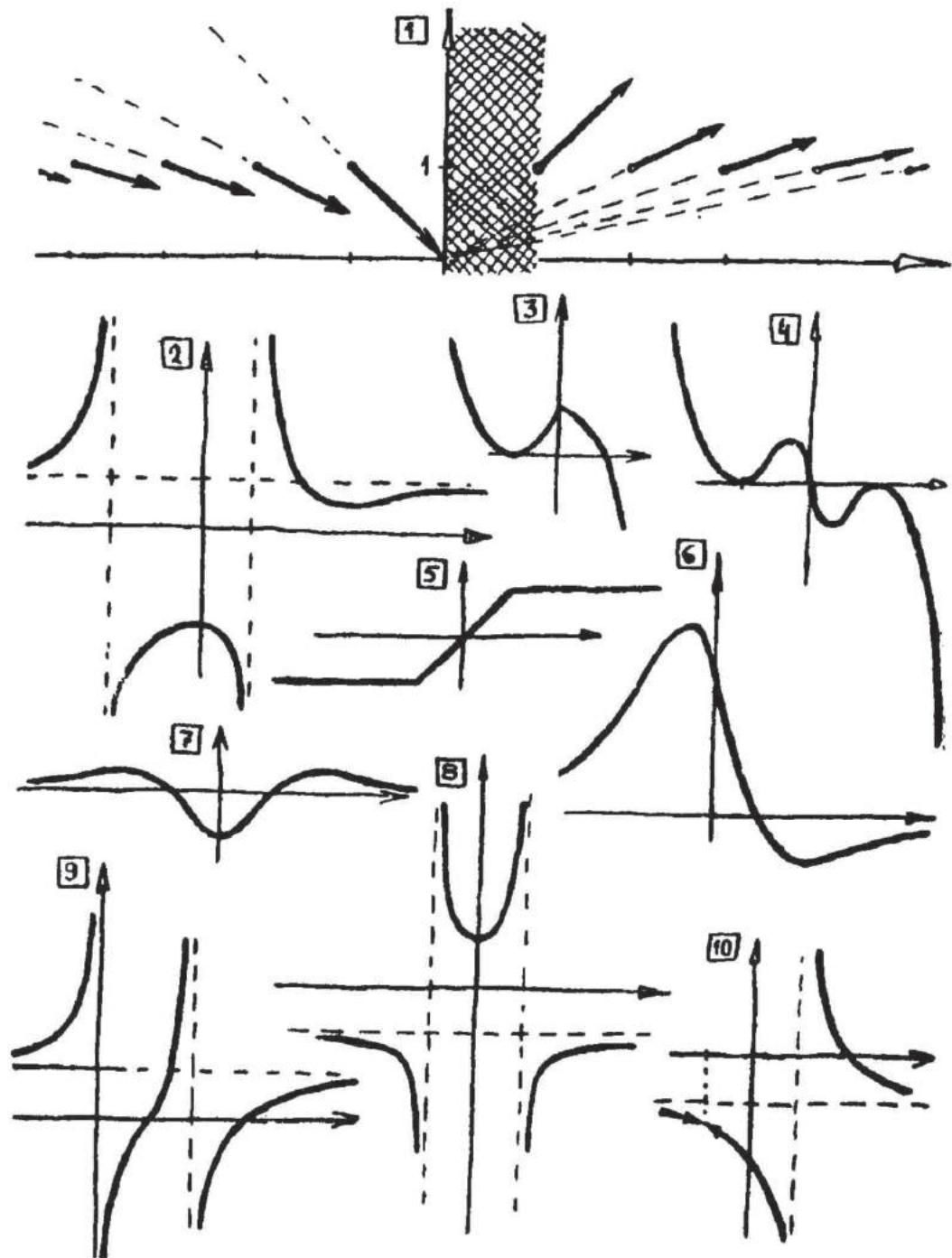
18. Prove that the points  $F_1(\sqrt{2}, \sqrt{2})$  and  $F_2(-\sqrt{2}, -\sqrt{2})$  are the foci of the hyperbola  $y = 1/x$ ; that is, the difference of the distances from any point of this hyperbola to the points  $F_1$  and  $F_2$  is constant in absolute value.

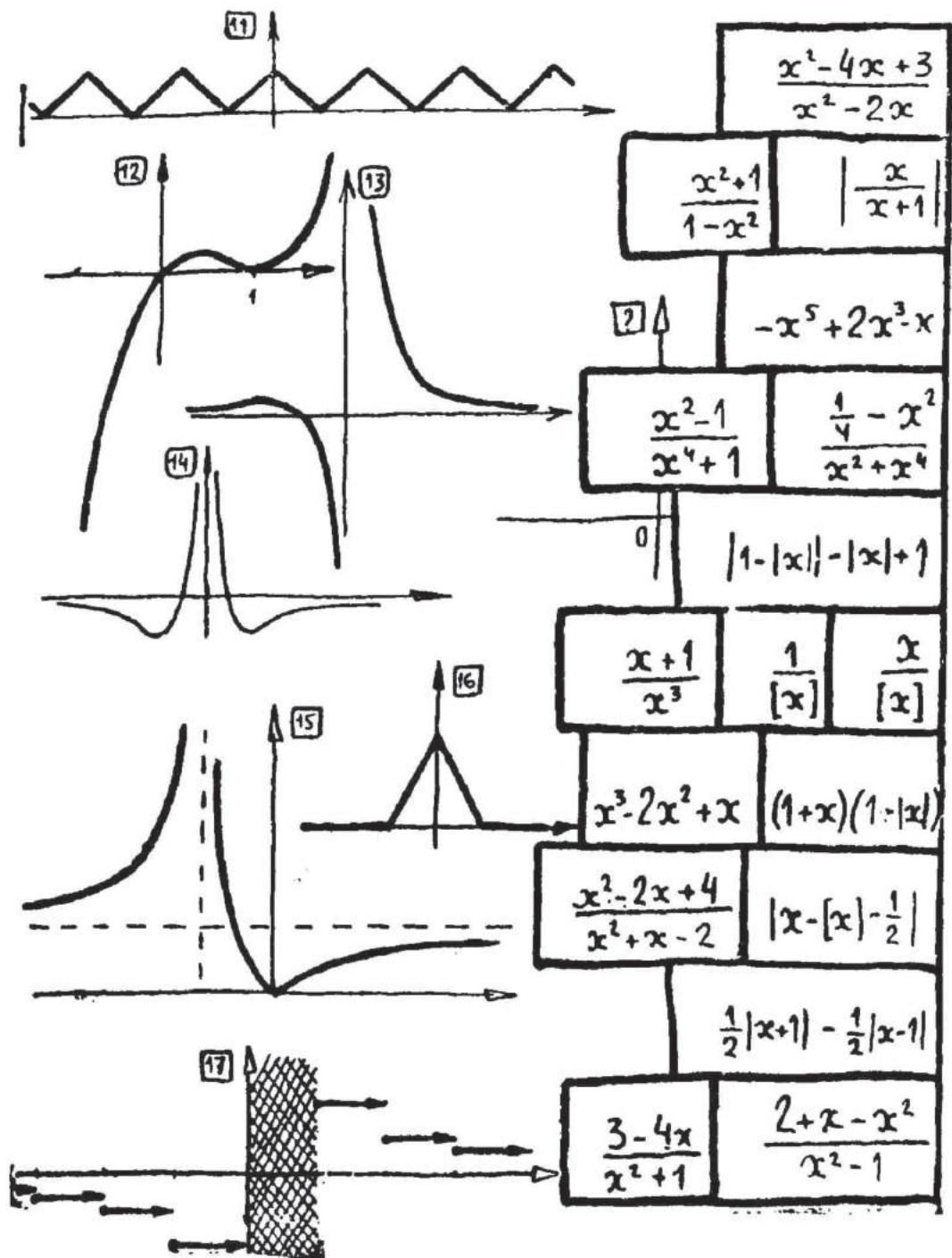
**Hint.** Take an arbitrary point  $M(a, 1/a)$  on the hyperbola  $y = 1/x$ . Express the distances of this point from the point  $F_1(\sqrt{2}, \sqrt{2})$  and from the point  $F_2(-\sqrt{2}, -\sqrt{2})$  in terms of  $a$ . Show that the absolute value of this difference is the same for all values of  $a$  (and, hence, does not depend on the choice of the point on the hyperbola).

19. On pages 96 and 97, seventeen graphs and as many formulas are given. The problem is to determine which formula belongs to which of the numbered graphs. Among these graphs the reader can find the answers to exercises.

\*The distance between the two points  $A(x_1, y_1)$  and  $B(x_2, y_2)$  is given by the formula  $d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ .

†The distance from the point  $A(x_1, y_1)$  to the straight line  $y = c$  equals  $|y_1 - c|$ .





20. Figure 3 represents the graph of the function  $y = f(x)$ .

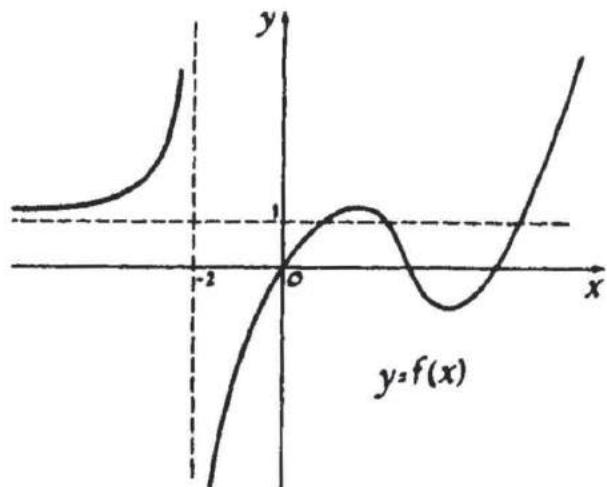


Fig. 3

Sketch the graphs of the following functions:

- (a)  $y = f(x) - 2$ ;      (b)  $y = f(x + 2)$ ;
- (c)  $y = |f(x)|$ ;      (d)  $y = f(|x|)$ ;
- (e)  $y = -3f(x)$ ;      (f)  $y = \frac{1}{f(x)}$ ;
- (g)  $y = (f(x))^2$ ;      (h)  $y = f(-x)$ ;
- (i)  $y = x + f(x)$ ;      (j)  $y = \frac{f(x)}{x}$ .

21. A square with side  $a$  is drawn in the plane (Fig. 4). The curve  $L_s$  is the locus of all points the least distance of which from some point of the square is equal to  $S$ . Let us denote the area bounded by the curve  $L_s$  by  $P(S)$ .

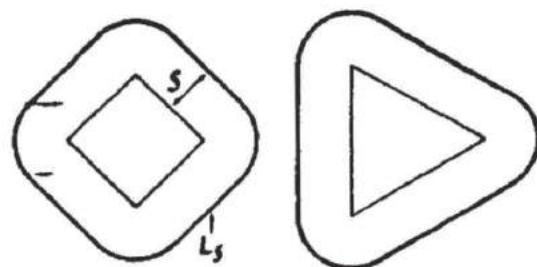


Fig. 4

(a) Find  $P(S)$  as a function of  $S$ .

(b) Solve a problem analogous to Problem 21a, but instead of a square, take a rectangle with sides  $a$  and  $b$ .

(c) The same problem for the triangle with sides  $a, b$ , and  $c$ .

(d) The same problem for the circle of radius  $r$ .

22. Can you find a rule for the resulting expressions for  $P(S)$ ? Write down a general formula for any convex figure. Does this formula hold for non-convex figures?

23. We shall examine quadratic equations of the form  $x^2 + px + q = 0$ ; each such equation is completely determined by the two numbers  $p$  and  $q$ . Let us agree to represent this equation by the point in the plane with coordinates  $(p, q)$ . For example, the equation  $x^2 - 2x + 3 = 0$  is represented by the point  $A(-2, 3)$ ; the equation  $x^2 - 1 = 0$  by the point  $B(0, -1)$ .

(a) What equation corresponds to the origin?

(b) Draw the set of points corresponding to those equations whose roots have a sum equal to zero.

(c) Randomly select a point in the plane. If the equation corresponding to this point has two real roots, mark the point with a green pencil. If the equation does not have any real roots, mark the point with a red pencil. Take a few more points and do the same with them. Can you state which part of the plane is occupied by "green" points and which by "red" ones? What line separates the "green" points from the "red" ones? How many roots have the equations corresponding to the points of this line?

(d) What point set corresponds to those equations whose roots are real and positive?

(e) By what point can an equation be represented if it is known that one of its roots is equal to 1?

24. Up to some moment a car was traveling in uniformly accelerated motion and then started to travel in uniform motion (at the speed attained by it).

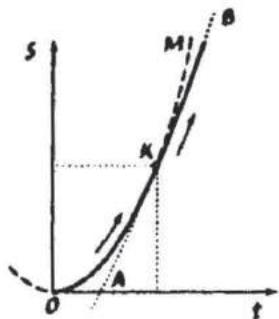


Fig. 5

The graph of the motion of this car is depicted in Fig. 5. Prove that the straight line  $AB$  is a tangent to the parabola  $OKM$ .

25. (a) Using graphs determine the number of solutions of the following cubic equations:

$$(1) 0.01x^3 = x^2 - 1,$$

$$(2) 0.001x^3 = x^2 - 3x + 2.$$

(b) Find approximate values of the roots of these equations.

26. (a) On p. 34 there is a diagram showing that the graph of the polynomial  $y = x^4 - \frac{5}{2}x^2 + \frac{9}{16}$  is obtained from the graph of the polynomial  $y = x^4 - 2x^3 - x^2 + 2x$  by translation along the  $x$ -axis. Find the value of this translation.

(b) Solve the equation of the fourth degree

$$x^4 - 6x^3 + 7x^2 + 6x - 8 = 0.$$

**Hint.** Translate the graph of the polynomial  $x^4 - 6x^3 + 7x^2 + 6x$  along the  $x$ -axis so that it becomes the graph of some biquadratic polynomial.

(c) Under what conditions does the curve

$$y = x^4 + bx^3 + cx^2 + d$$

have an axis of symmetry?  $\oplus$

27. Let us solve Problem 4 on page 38. There are  $19 + 9 + 26 + 8 + 18 + 11 + 14 = 105$  matches in all. Therefore it is necessary to obtain a distribution of  $105 \div 7 = 15$  matches in each box.

Let us denote by  $x$  the number of matches that must be shifted from the first box to the second. (It may, of course, be necessary to shift matches from the second box to the first, in which case  $x$  will be negative.) After we have shifted  $x$  matches from the first box to the second, there will be  $x + 9$  matches in the second box.

$$\begin{aligned} 19+9+26+8+ \\ 18+11+14=105 \\ 105 \div 7=15 \\ 9+x-(x-6)=15 \\ 26+(x-6)=20+x \\ 20+x-(x+5)=15 \end{aligned}$$

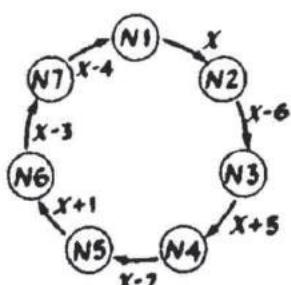


Fig. 6

Therefore it is necessary to move  $x - 6$  matches from the second to the third,  $x + 5$  matches from the third to the fourth. Similarly,  $x - 2$  are shifted from the fourth box to the fifth,  $x + 1$  from the fifth to the sixth,  $x - 3$  from the sixth to the seventh, and finally  $x - 4$  matches from the seventh to the first (Fig. 6).

Let us now denote by  $S$  the total number of shifted matches:

$$S = |x| + |x - 6| + |x + 5| + |x - 2| \\ + |x + 1| + |x - 3| + |x - 4|.$$

In this formula absolute-value signs were used because we are interested only in the number of matches transposed and not in the direction in which they were shifted.

We must now choose  $x$  so that  $S$  has the least value. Here the graph of the function  $S = f(x)$  can be helpful (Fig. 7): The lowest point of the graph is the vertex  $A_4$ ; that is, the function  $S = f(x)$  assumes its least value at  $x = 2$ . Thus  $x$  has been found, and we can say how many matches must be moved, and where they will be moved (Fig. 8). In this manner the problem can also be solved, of course, for an arbitrary number ( $n$ ) of boxes. For this purpose it is necessary, as in our example, to write down an expression for  $S$ . It will be of the form:

$$S = |x| + |x - a_1| + |x - a_2| + \dots \\ + |x - a_{n-1}|.$$

In order to find the required value of  $x$  in the case of an odd number  $n$ , the following simple rule can be used: The numbers  $0, a_1, a_2, \dots, a_{n-1}$  must be written down in increasing order, after which  $x$  is chosen equal to the number exactly in the center of this sequence of numbers (if  $n$  is odd, such a number

$$S = |x+5| + |x+1| \\ + |x| + |x-2| \\ + |x-3| + |x-4| \\ + |x-6|$$

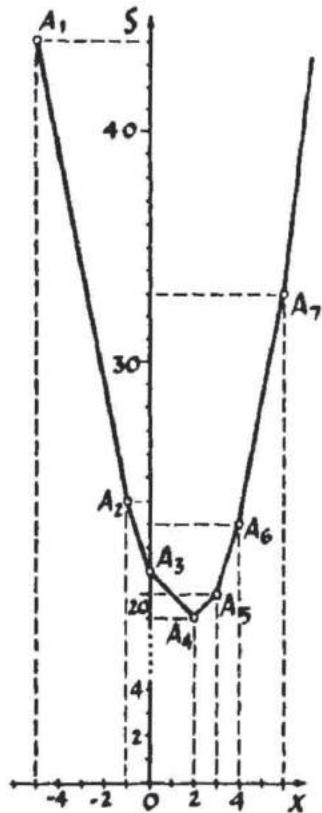


Fig. 7

Answer

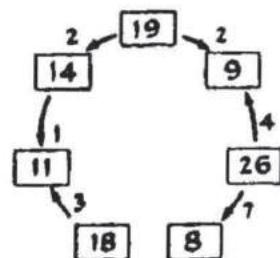


Fig. 8

can always be found). Consider how the graph looks when  $n$  is even.

Then try to state a rule for finding  $x$  in this case.

This problem, which is similar to a game, is connected with the practical problem of transportation along circular routes. Imagine a circular railroad with evenly spaced stations. At some stations there are coal storages; at others there are users of coal who must be furnished with all of this coal. Figure 9 indicates the stocks of coal at the storages and (with a minus sign) the respective needs of the users.

Using the solution of the preceding problem, set up the most economical transportation plan.

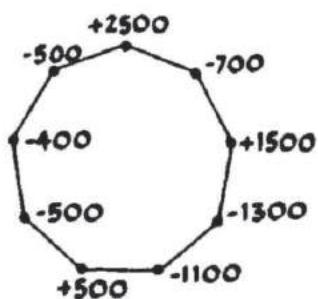


Fig. 9

## **Answers and Hints to Problems and Exercises Marked with the Sign $\oplus$**

**Exercise 2b, p. 17.** Look for the answer among the graphs on pp. 96 and 97.

**Exercise 3b,** on p. 32, for the graph in Fig. 12. Look for the answer among the graphs on pp. 96 and 97.

**Exercise 3,** p. 37. Hint. This function assumes its minimum on an entire segment.

**Problem 4** on p. 38. Solution in Problem 27 on p. 100.

**Exercise 2b** on p. 60. No, it does not. The rigorous proof of this fact is not very simple, and we shall not give it here. It is clear, however, that since the  $x$ -axis and the  $y$ -axis are asymptotes of this curve, the only possible axis of symmetry is the straight line  $y = x$ . It is easy to verify that this straight line is not an axis of symmetry.

**Exercise 2** on p. 79. The equation of the tangent is  $y = 3x - 1$ .

**Exercise 3** on p. 79. Hint. The system

$$y = x + a, \quad y = -x^2 - 1$$

must have two coincident solutions.

**Problem 1h,** p. 89 and **1q,** p. 90. Look for the answer among the graphs on pp. 96 and 97.

**Problem 3b,** p. 91.

Let us take a numerical example. Suppose the roots of the numerator are  $-5$  and  $0$ , and those of the denominator are  $+2$  and  $+4$ . Then our function is of the form  $y = [ax(x + 5)]/[(x - 2)(x - 4)]$ . Let us take some concrete value of  $a$ , e.g.,  $a = 2$ . The function

$$y = \frac{[2x(x + 5)]}{[(x - 2)(x - 4)]}$$

is not defined at  $x = 2$  and  $x = 4$ . As  $x$  approaches these values, the denominator decreases, approaching zero; therefore, the function increases without bound in absolute value — the straight lines  $x = 2$  and  $x = 4$  are vertical asymptotes of the graph.

The function is equal to zero at  $x = 0$  and  $x = -5$ . Let us mark two points of the graph on the  $x$ -axis:  $(0, 0)$  and  $(-5, 0)$ .

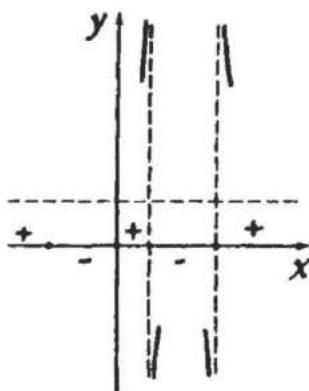


Fig. 10

The four "special" values of the argument,  $x = -5, 0, 2, 4$ , divide the  $x$ -axis into 5 intervals. As  $x$  passes the boundary of any interval, the function changes its sign (vanishing or "going off to infinity") (Fig. 10).

We must explain the behavior of the function when the argument increases without bound in absolute value. Let us try to substitute large numbers for  $x$  (e.g.,  $x = 10,000, x = 1,000,000$ , etc.). Since  $2x^2$  will be considerably larger than  $10x$ , and  $x^2$  considerably larger than  $-6x + 8$ , the fraction

$$\frac{2x(x+5)}{(x-2)(x-4)} = \frac{2x^2 + 10x}{x^2 - 6x + 8}$$

will be approximately equal to the ratio of the highest terms of numerator and denominator,

$$y = \frac{2x^2 + 10x}{x^2 - 6x + 8} \approx \frac{2x^2}{x^2} = 2,$$

and will be the closer to 2, the larger  $|x|$  is. Hence, as  $x$  moves away from the origin, the graph approaches the horizontal straight line  $y = 2$ .

The general form of the graph is given in Fig. 11. In all cases when both roots of the denominator are larger than the roots of the numerator, the graph will be approximately of this form.

**Problem 14, p. 94.** As is often the case in mathematics, the problem is easier to solve in a general form than for a specified concrete function. Therefore, we shall first solve Problem *b*, and obtain the solution of *a* as a particular case.

Thus, suppose we are given some function  $f(x)$ . Let us suppose the problem is solved; that is,  $f(x)$  is represented as a sum of an even function  $g(x)$  and an odd function  $h(x)$ :

$$f(x) = g(x) + h(x). \quad (*)$$

Since this equality holds for all values of  $x$ ,  $-x$  can be substituted for  $x$ , and we obtain

$$f(-x) = g(-x) + h(-x). \quad (**)$$

Since the function  $g(x)$  is even, and  $h(x)$  is odd,  $g(-x) = g(x)$  and  $h(-x) = -h(x)$ . Using this, we first add

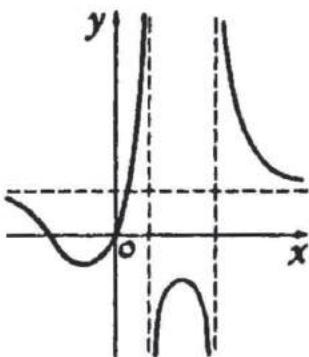


Fig. 11

Eqs. \* and \*\*, and then subtract one from the other; we find

$$f(x) + f(-x) = 2g(x), \quad f(x) - f(-x) = 2h(x).$$

From this the functions  $g(x)$  and  $h(x)$  can be found, and the desired decomposition of the function  $f(x)$  into a sum of an even and an odd function can be obtained:

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

Notice that the formal proof of the result we have obtained is even simpler: Write down the decomposition of Eq. 1, and check that it is identically satisfied for all  $x$  and that the first term on the right-hand side is an even function while the second is odd.

The solution of Problem a is obtained directly by Formula 1:

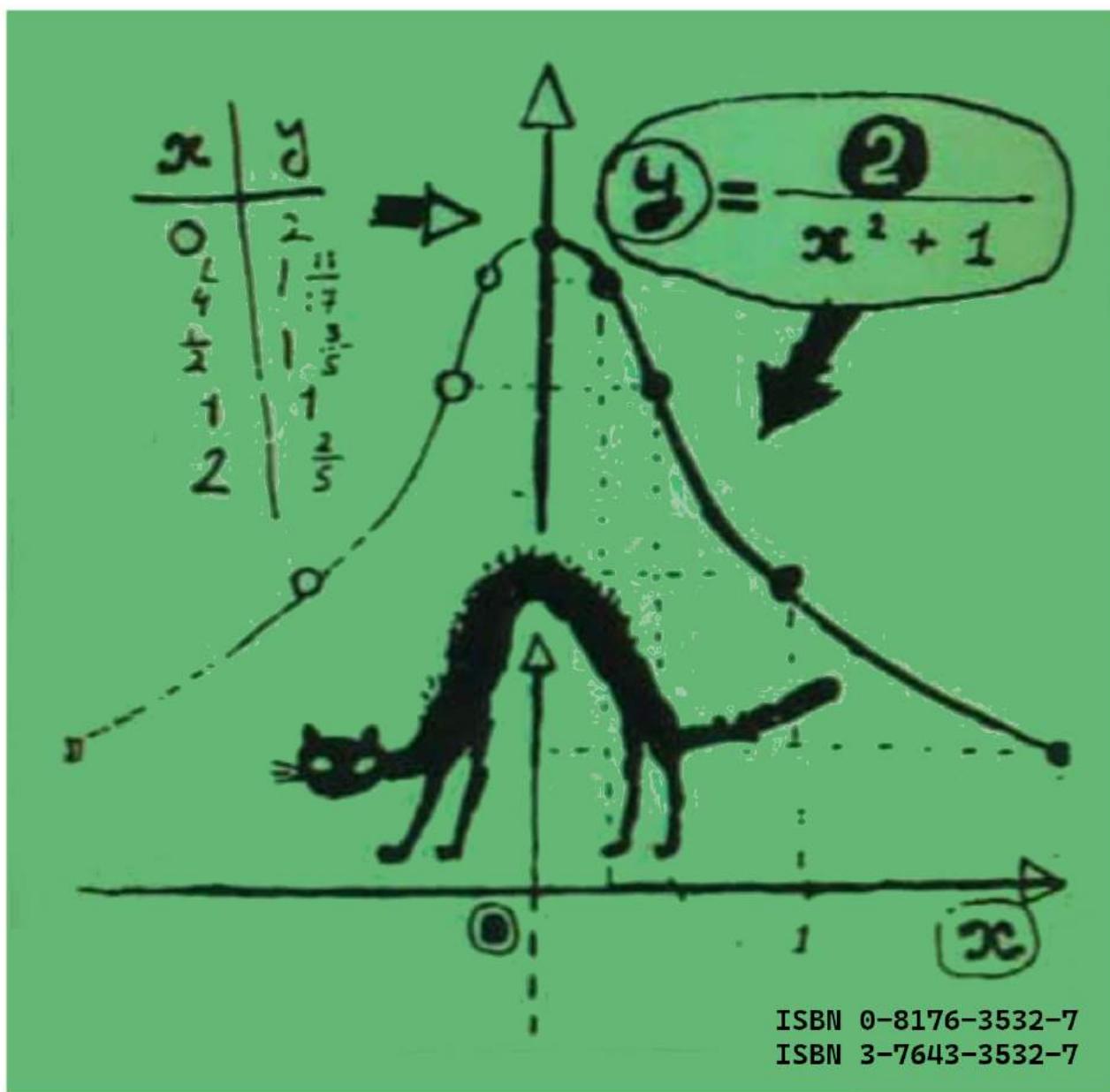
$$\frac{1}{x^4 - x} = \frac{x^2}{x^6 - 1} + \frac{1}{x^7 - x}.$$

**Remark.** If the function  $y = f(x)$  is not defined for some values of  $x$ , then also the functions  $g(x)$  and  $h(x)$  will not be defined for all values of  $x$ . In this case, it may turn out that for some  $x$  the function  $f(x)$  is defined, while  $g(x)$  and  $h(x)$  are not defined.

Exercise 26c, p. 100.  $4d = b^3 + 2bc$ .

The need for improved mathematics education at the high school and college levels has never been more apparent than in the 1990s. As early as the 1960s, I.M. Gel'fand and his colleagues in the USSR thought hard about this same question and developed a style for presenting basic mathematics in a clear and simple form that engaged the curiosity and intellectual interest of thousands of school and college students. These same ideas, this same content, unchanged by over thirty years of experience and mathematical development, are available in the present books to any student who is willing to read, to be stimulated, and to learn.

Functions and Graphs provides instruction in transferring formulas and data into geometrical form. Thus, drawing graphs is shown to be one way to "see" formulas and functions and to observe the ways in which they change. This skill is fundamental to the study of calculus and other mathematical topics. Teachers of mathematics will find here a fresh understanding of the subject and a valuable path to the training of students in mathematical concepts and skills.



ISBN 0-8176-3532-7  
ISBN 3-7643-3532-7

# *BASIC MATHEMATICS*



**SERGE LANG**  
Columbia University

# *BASIC MATHEMATICS*



**ADDISON-WESLEY PUBLISHING COMPANY**  
Reading, Massachusetts  
Menlo Park, California · London · Don Mills, Ontario

This book is in the

**ADDISON-WESLEY SERIES IN INTRODUCTORY MATHEMATICS**

*Consulting Editors:*

Gail S. Young

Richard S. Pieters

Cover photograph by courtesy of Spencer-Phillips and Green, Kentfield, California.

Copyright © 1971 by Addison-Wesley Publishing Company Inc. Philippines copyright 1971 by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada. Library of Congress Catalog Card No. 75-132055.

*To Jerry*



*My publishers, Addison-Wesley, have produced my books for these last eight years. I want it known how much I appreciate their extraordinary performance at all levels. General editorial advice, specific editing of the manuscripts, and essentially flawless typesetting and proof sheets. It is very gratifying to have found such a company to deal with.*

New York, 1970

*Serge Lang*

## *Acknowledgments*

I am grateful to Peter Lerch, Gene Murrow, Dick Pieters, and Gail Young for their careful reading of the manuscript and their useful suggestions.

I am also indebted to Howard Dolinsky, Bernard Duflos, and Arvin Levine for working out the answers to the exercises.

S.L.

# *Foreword*

The present book is intended as a text in basic mathematics. As such, it can have multiple use: for a one-year course in the high schools during the third or fourth year (if possible the third, so that calculus can be taken during the fourth year); for a complementary reference in earlier high school grades (elementary algebra and geometry are covered); for a one-semester course at the college level, to review or to get a firm foundation in the basic mathematics necessary to go ahead in calculus, linear algebra, or other topics.

Years ago, the colleges used to give courses in “college algebra” and other subjects which should have been covered in high school. More recently, such courses have been thought unnecessary, but some experiences I have had show that they are just as necessary as ever. What is happening is that the colleges are getting a wide variety of students from high schools, ranging from exceedingly well-prepared ones who have had a good first course in calculus, down to very poorly prepared ones. This latter group includes both adults who return to college after several years’ absence in order to improve their technical education, and students from the high schools who were not adequately taught. This is the reason why some material properly belonging to the high-school level must still be offered in the colleges.

The topics in this book are covered in such a way as to bring out clearly all the important points which are used afterwards in higher mathematics. I think it is important not to separate arbitrarily in different courses the various topics which involve both algebra and geometry. Analytic geometry and vector geometry should be considered simultaneously with algebra and plane geometry, as natural continuations of these. I think it is much more valuable to go into these topics, especially vector geometry, rather than to go endlessly into more and more refined results concerning triangles or trigonometry, involving more and more complicated technique. A minimum of basic techniques must of course be acquired, but it is better to extend these techniques by applying them to new situations in which they become

motivated, especially when the possible topics are as attractive as vector geometry.

In fact, for many years college courses in physics and engineering have faced serious drawbacks in scheduling because they need simultaneously some calculus and also some vector geometry. It is very unfortunate that the most basic operations on vectors are introduced at present only in college. They should appear at least as early as the second year of high school. I cannot write here a text for elementary geometry (although to some extent the parts on intuitive geometry almost constitute such a text), but I hope that the present book will provide considerable impetus to lower considerably the level at which vectors are introduced. Within some foreseeable future, the topics covered in this book should in fact be the standard topics for the second year of high school, so that the third and fourth years can be devoted to calculus and linear algebra.

If only preparatory material for calculus is needed, many portions of this book can be omitted, and attention should be directed to the rules of arithmetic, linear equations (Chapter 2), quadratic equations (Chapter 4), coordinates (the first three sections of Chapter 8), trigonometry (Chapter 11), some analytic geometry (Chapter 12), a simple discussion of functions (Chapter 13), and induction (Chapter 16, §1). The other parts of the book can be omitted. Of course, the more preparation a student has, the more easily he will go through more advanced topics.

“More preparation”, however, does not mean an accumulation of technical material in which the basic ideas of a subject are completely drowned. I am always disturbed at seeing endless chains of theorems, most of them of no interest, and without any stress on the main points. As a result, students do not remember the essential features of the subject. I am fully aware that because of the pruning I have done, many will accuse me of not going “deeply enough” into some subjects. I am quite ready to confront them on that. Besides, as I prune some technical and inessential parts of one topic, I am able to include the essential parts of another topic which would not otherwise be covered. For instance, what better practice is there with negative numbers than to introduce at once coordinates in the plane as a pair of numbers, and then deal with the addition and subtraction of such pairs, componentwise? This introduction could be made as early as the fourth grade, using maps as a motivation. One could do roughly what I have done here in Chapter 8, §1, Chapter 9, §1, and the beginning of Chapter 9, §2 (addition of pairs of numbers, and the geometric interpretation in terms of a parallelogram). At such a level, one can then leave it at that.

The same remark applies to the study of this book. The above-mentioned sections can be covered very early, at the same time that you study numbers

and operations with numbers. They give a very nice geometric flavor to a slightly dry algebraic theory.

Generally speaking, I hope to induce teachers to leave well enough alone, and to avoid torturing a topic to death. It is easier to advance in one topic by going ahead with the more elementary parts of another topic, where the first one is applied. The brain much prefers to work that way, rather than to concentrate on ugly technical formulas which are obviously unrelated to anything except artificial drilling. Of course, some rote drilling is necessary. The problem is how to strike a balance. Do not regard some lists of exercises as too short. Rather, realize that practice for some notion may come again later in conjunction with another notion. Thus practice with square roots comes not only in the section where they are defined, but also later when the notion of distance between points is discussed, and then in a context where it is more interesting to deal with them. The same principle applies throughout the book.

The Interlude on logic and mathematical expression can be read also as an introduction to the book. Because of various examples I put there, and because we are already going through a Foreword, I have chosen to place it physically somewhat later. Take a look at it now, and go back to it whenever you feel the need for such general discussions. Mainly, I would like to make you feel more relaxed in your contact with mathematics than is usually the case. I want to stimulate thought, and do away with the general uptight feelings which people often have about math. If, for instance, you feel that any chapter gets too involved for you, then skip that part until you feel the need for it, and look at another part of the book. In many cases, you don't necessarily need an earlier part to understand a later one. In most cases, the important thing is to have understood the basic concepts and definitions, to be at ease with the simpler computational aspects of these concepts, and then to go ahead with a more advanced topic.

This advice also applies to the book as a whole. If you find that there is not enough material in this book to occupy you for a whole year, then start studying calculus or possibly linear algebra.

The book deals with mathematics on both the manipulative (or computational) level and the theoretical level. You must realize that a mastery of mathematics involves both levels, although your tastes may direct you more strongly to one or the other, or both. Here again, you may wish to vary the emphasis which you place on them, according to your needs or your taste. Be warned that deficiency at either level can ultimately hinder you in your work. Independently of need, however, it should be a source of pleasure to understand why a mathematical result is true, i.e. to understand its proof as well as to understand how to use the result in concrete circumstances.

Try to rely on yourself, and try to develop a trust in your own judgment. There is no “right” way to do things. Tastes differ, and this book is not meant to suppress yours. It is meant to propose some basic mathematical topics, according to my taste. If I am successful, you will agree with my taste, or you will have developed your own.

*New York  
January 1971*

S.L.

# *Contents*

## **PART I ALGEBRA**

### **Chapter 1 Numbers**

1	The integers . . . . .	5
2	Rules for addition . . . . .	8
3	Rules for multiplication . . . . .	14
4	Even and odd integers; divisibility . . . . .	22
5	Rational numbers . . . . .	26
6	Multiplicative inverses . . . . .	42

### **Chapter 2 Linear Equations**

1	Equations in two unknowns . . . . .	53
2	Equations in three unknowns . . . . .	57

### **Chapter 3 Real Numbers**

1	Addition and multiplication . . . . .	61
2	Real numbers: positivity . . . . .	64
3	Powers and roots . . . . .	70
4	Inequalities . . . . .	75

### **Chapter 4 Quadratic Equations . . . . .**

### **Interlude On Logic and Mathematical Expressions**

1	On reading books . . . . .	93
2	Logic . . . . .	94
3	Sets and elements . . . . .	99
4	Notation . . . . .	100

**PART II INTUITIVE GEOMETRY****Chapter 5 Distance and Angles**

1	Distance . . . . .	107
2	Angles . . . . .	110
3	The Pythagoras theorem . . . . .	120

**Chapter 6 Isometries**

1	Some standard mappings of the plane . . . . .	133
2	Isometries . . . . .	143
3	Composition of isometries . . . . .	150
4	Inverse of isometries . . . . .	155
5	Characterization of isometries . . . . .	163
6	Congruences . . . . .	166

**Chapter 7 Area and Applications**

1	Area of a disc of radius $r$ . . . . .	173
2	Circumference of a circle of radius $r$ . . . . .	180

**PART III COORDINATE GEOMETRY****Chapter 8 Coordinates and Geometry**

1	Coordinate systems . . . . .	191
2	Distance between points . . . . .	197
3	Equation of a circle . . . . .	203
4	Rational points on a circle . . . . .	206

**Chapter 9 Operations on Points**

1	Dilations and reflections . . . . .	213
2	Addition, subtraction, and the parallelogram law . . . . .	218

**Chapter 10 Segments, Rays, and Lines**

1	Segments . . . . .	229
2	Rays . . . . .	231
3	Lines . . . . .	236
4	Ordinary equation for a line . . . . .	246

**Chapter 11 Trigonometry**

1	Radian measure	249
2	Sine and cosine.	252
3	The graphs . . . . .	264
4	The tangent . . . . .	266

5	Addition formulas . . . . .	272
6	Rotations . . . . .	277
<b>Chapter 12 Some Analytic Geometry</b>		
1	The straight line again	281
2	The parabola . . . . .	291
3	The ellipse . . . . .	297
4	The hyperbola . . . . .	300
5	Rotation of hyperbolas	305
 <b>PART IV MISCELLANEOUS</b>		
<b>Chapter 13 Functions</b>		
1	Definition of a function	313
2	Polynomial functions . . . . .	318
3	Graphs of functions . . . . .	330
4	Exponential function . . . . .	333
5	Logarithms . . . . .	338
<b>Chapter 14 Mappings</b>		
1	Definition . . . . .	345
2	Formalism of mappings	351
3	Permutations . . . . .	359
<b>Chapter 15 Complex Numbers</b>		
1	The complex plane	375
2	Polar form . . . . .	380
<b>Chapter 16 Induction and Summations</b>		
1	Induction . . . . .	383
2	Summations . . . . .	388
3	Geometric series . . . . .	396
<b>Chapter 17 Determinants</b>		
1	Matrices . . . . .	401
2	Determinants of order 2 . . . . .	406
3	Properties of $2 \times 2$ determinants	409
4	Determinants of order 3 . . . . .	414
5	Properties of $3 \times 3$ determinants	418
6	Cramer's Rule . . . . .	424
<b>Index</b>		429



*Part One*  
**ALGEBRA**



In this part we develop systematically the rules for operations with numbers, relations among numbers, and properties of these operations and relations: addition, multiplication, inequalities, positivity, square roots,  $n$ -th roots. We find many of them, like commutativity and associativity, which recur frequently in mathematics and apply to other objects. They apply to complex numbers, but also to functions or mappings (in this case, commutativity does not hold in general and it is always an interesting problem to determine when it does hold).

Even when we study geometry afterwards, the rules of algebra are still used, say to compute areas, lengths, etc., which associate numbers with geometric objects. Thus does algebra mix with geometry.

The main point of this chapter is to condition you to have efficient reflexes in handling addition, multiplication, and division of numbers. There are many rules for these operations, and the extent to which we choose to assume some, and prove others from the assumed ones, is determined by several factors. We wish to assume those rules which are most basic, and assume enough of them so that the proofs of the others are simple. It also turns out that those which we do assume occur in many contexts in mathematics, so that whenever we meet a situation where they arise, then we already have the training to apply them and use them. Both historical experience and personal experience have gone into the selection of these rules and the order of the list in which they are given. To some extent, you must trust that it is valuable to have fast reflexes when dealing with associativity, commutativity, distributivity, cross-multiplication, and the like, if you do not have the intuition yourself which makes such trust unnecessary. Furthermore, the long list of the rules governing the above operations should be taken in the spirit of a description of how numbers behave.

It may be that you are already reasonably familiar with the operations between numbers. In that case, omit the first chapter entirely, and go right

ahead to Chapter 2, or start with the geometry or with the study of coordinates in Chapter 7. The whole first part on algebra is much more dry than the rest of the book, and it is good to motivate this algebra through geometry. On the other hand, your brain should also have quick reflexes when faced with a simple problem involving two linear equations or a quadratic equation. Hence it is a good idea to have isolated these topics in special sections in the book for easy reference.

In organizing the properties of numbers, I have found it best to look successively at the integers, rational numbers, and real numbers, at the cost of slight repetitions. There are several reasons for this. First, it is a good way of learning certain rules and their consequences in a special context (e.g. associativity and commutativity in the context of integers), and then observing that they hold in more general contexts. This sort of thing happens very frequently in mathematics. Second, the rational numbers provide a wide class of numbers which are used in computations, and the manipulation of fractions thus deserves special emphasis. Third, to follow the sequence integers–rational numbers–real numbers already plants in your mind a pattern which you will encounter again in mathematics. This pattern is related to the extension of one system of objects to a larger system, in which more equations can be solved than in the smaller system. For instance, the equation  $2x = 3$  can be solved in the rational numbers, but not in the integers. The equations  $x^2 = 2$  or  $10^x = 2$  can be solved in the real numbers but not in the rational numbers. Similarly, the equations  $x^2 = -1$ , or  $x^2 = -2$ , or  $10^x = -3$  can be solved in the complex numbers but not in the real numbers. It will be useful to you to have met the idea of extending mathematical systems at this very basic stage because it exhibits features in common with those in more advanced contexts.

# **I** Numbers

## **§1. THE INTEGERS**

The most common numbers are those used for counting, namely the numbers

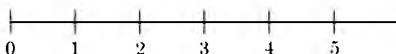
$$1, 2, 3, 4, \dots,$$

which are called the **positive integers**. Even for counting, we need at least one other number, namely,

$$0 \text{ (zero).}$$

For instance, we may wish to count the number of right answers you may get on a test for this course, out of a possible 100. If you get 100, then all your answers were correct. If you get 0, then no answer was correct.

The positive integers and zero can be represented geometrically on a line, in a manner similar to a ruler or a measuring stick:



**Fig. 1-1**

For this we first have to select a unit of distance, say the inch, and then on the line we mark off the inches to the right as in the picture.

For convenience, it is useful to have a name for the positive integers together with zero, and we shall call these the **natural numbers**. Thus 0 is a natural number, so is 2, and so is 124,521. The natural numbers can be used to measure distances, as with the ruler.

By definition, the point represented by 0 is called the **origin**.

The natural numbers can also be used to measure other things. For example, a thermometer is like a ruler which measures temperature. However,

the thermometer shows us that we encounter other types of numbers besides the natural numbers, because there may be temperatures which may go below 0. Thus we encounter naturally what we shall call **negative integers** which we call **minus 1, minus 2, minus 3, . . .**, and which we write as

$$-1, -2, -3, -4, \dots$$

We represent the negative integers on a line as being on the other side of 0 from the positive integers, like this:

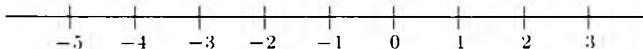


Fig. 1-2

The positive integers, negative integers, and zero all together are called the **integers**. Thus  $-9, 0, 10, -5$  are all integers.

If we view the line as a thermometer, on which a unit of temperature has been selected, say the degree Fahrenheit, then each integer represents a certain temperature. The negative integers represent temperatures below zero.

Our discussion is already typical of many discussions which will occur in this course, concerning mathematical objects and their applicability to physical situations. In the present instance, we have the integers as mathematical objects, which are essentially abstract quantities. We also have different applications for them, for instance measuring distance or temperatures. These are of course not the only applications. Namely, we can use the integers to measure time. We take the origin 0 to represent the year of the birth of Christ. Then the positive integers represent years after the birth of Christ (called **AD** years), while the negative integers can be used to represent **BC** years. With this convention, we can say that the year  $-500$  is the year  $500$  **BC**.

Adding a positive number, say 7, to another number, means that we must move 7 units to the right of the other number. For instance,

$$5 + 7 = 12.$$

Seven units to the right of 5 yields 12. On the thermometer, we would of course be moving upward instead of right. For instance, if the temperature at a given time is  $5^\circ$  and if it goes up by  $7^\circ$ , then the new temperature is  $12^\circ$ .

Observe the very simple rule for addition with 0, namely

N1.

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

for any integer  $a$ .

What about adding negative numbers? Look at the thermometer again. Suppose the temperature at a given time is  $10^\circ$ , and the temperature drops by  $15^\circ$ . The new temperature is then  $-5^\circ$ , and we can write

$$10 - 15 = -5.$$

Thus  $-5$  is the result of subtracting  $15$  from  $10$ , or of adding  $-15$  to  $10$ .

In terms of points on a line, adding a negative number, say  $-3$ , to another number means that we must move 3 units to the left of this other number. For example,

$$5 + (-3) = 2$$

because starting with  $5$  and moving 3 units to the left yields  $2$ . Similarly,

$$7 + (-3) = 4, \quad \text{and} \quad 3 + (-5) = -2.$$

Note that we have

$$3 + (-3) = 0 \quad \text{or} \quad 5 + (-5) = 0.$$

We can also write these equations in the form

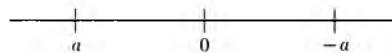
$$(-3) + 3 = 0 \quad \text{or} \quad (-5) + 5 = 0.$$

For instance, if we start 3 units to the left of  $0$  and move 3 units to the right, we get  $0$ . Thus, in general, we have the formulas (by assumption):

N2.

$$a + (-a) = 0 \quad \text{and also} \quad -a + a = 0.$$

In the representation of integers on the line, this means that  $a$  and  $-a$  lie on opposite sides of  $0$  on that line, as shown on the next picture:

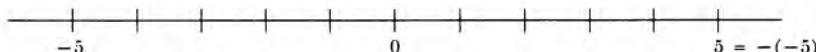
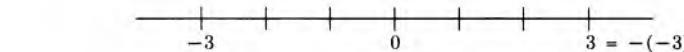


**Fig. 1-3**

Thus according to this representation we can now write

$$3 = -(-3) \quad \text{or} \quad 5 = -(-5).$$

In these special cases, the pictures are:



**Fig. 1-4**

**Remark.** We use the name

$$\text{minus } a \quad \text{for} \quad -a$$

rather than the words “negative  $a$ ” which have found some currency recently. I find the words “negative  $a$ ” confusing, because they suggest that  $-a$  is a negative number. This is not true unless  $a$  itself is positive. For instance,

$$3 = -(-3)$$

is a positive number, but 3 is equal to  $-a$ , where  $a = -3$ , and  $a$  is a negative number.

Because of the property

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

one also calls  $-a$  the **additive inverse** of  $a$ .

The sum and product of integers are also integers, and the next sections are devoted to a description of the rules governing addition and multiplication.

## §2. RULES FOR ADDITION

Integers follow very simple rules for addition. These are:

**Commutativity.** *If  $a, b$  are integers, then*

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

For instance, we have

$$3 + 5 = 5 + 3 = 8,$$

or in an example with negative numbers, we have

$$-2 + 5 = 3 = 5 + (-2).$$

**Associativity.** *If  $a, b, c$  are integers, then*

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

In view of this, it is unnecessary to use parentheses in such a simple context, and we write simply

$$a + b + c.$$

For instance,

$$(3 + 5) + 9 = 8 + 9 = 17,$$

$$3 + (5 + 9) = 3 + 14 = 17.$$

We write simply

$$3 + 5 + 9 = 17.$$

Associativity also holds with negative numbers. For example,

$$(-2 + 5) + 4 = 3 + 4 = 7,$$

$$-2 + (5 + 4) = -2 + 9 = 7.$$

Also,

$$(2 + (-5)) + (-3) = -3 + (-3) = -6,$$

$$2 + (-5 + (-3)) = 2 + (-8) = -6.$$

The rules of addition mentioned above will not be proved, but we shall prove other rules from them.

To begin with, note that:

N3.

*If  $a + b = 0$ , then  $b = -a$  and  $a = -b$ .*

To prove this, add  $-a$  to both sides of the equation  $a + b = 0$ . We get

$$-a + a + b = -a + 0 = -a.$$

Since  $-a + a + b = 0 + b = b$ , we find

$$b = -a$$

as desired. Similarly, we find  $a = -b$ . We could also conclude that

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

As a matter of convention, we shall write

$$a - b$$

instead of

$$a + (-b).$$

Thus a sum involving three terms may be written in many ways, as follows:

$$\begin{aligned}
 (a - b) + c &= (a + (-b)) + c \\
 &= a + (-b + c) && \text{by associativity} \\
 &= a + (c - b) && \text{by commutativity} \\
 &= (a + c) - b && \text{by associativity,}
 \end{aligned}$$

and we can also write this sum as

$$a - b + c = a + c - b,$$

omitting the parentheses. Generally, in taking the sum of integers, we can take the sum in any order by applying associativity and commutativity repeatedly.

As a special case of N3, for any integer  $a$  we have

N4.

$$a = -(-a).$$

This is true because

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

and we can apply N3 with  $b = -a$ . Remark that this formula is true whether  $a$  is positive, negative, or 0. If  $a$  is positive, then  $-a$  is negative. If  $a$  is negative, then  $-a$  is positive. In the geometric representation of numbers on the line,  $a$  and  $-a$  occur symmetrically on the line on opposite sides of 0. Of course, we can pile up minus signs and get other relationships, like

$$-3 = -(-(-3)),$$

or

$$3 = -(-3) = -(-(-(-3))).$$

Thus when we pile up the minus signs in front of  $a$ , we obtain  $a$  or  $-a$  alternatively. For the general formula with the appropriate notation, cf. Exercises 5 and 6 of §4.

From our rules of operation we can now prove:

*For any integers  $a, b$  we have*

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

or, in other words,

N5.

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

*Proof.* Remember that if  $x, y$  are integers, then  $x = -y$  and  $y = -x$  mean that  $x + y = 0$ . Thus to prove our assertion, we must show that

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

But this comes out immediately, namely,

$$\begin{aligned} (a + b) + (-a - b) &= a + b - a - b && \text{by associativity} \\ &= a - a + b - b && \text{by commutativity} \\ &= 0 + 0 \\ &= 0. \end{aligned}$$

This proves our formula.

**Example.** We have

$$\begin{aligned} -(3 + 5) &= -3 - 5 = -8, \\ -(-4 + 5) &= -(-4) - 5 = 4 - 5 = -1, \\ -(3 - 7) &= -3 - (-7) = -3 + 7 = 4. \end{aligned}$$

You should be very careful when you take the negative of a sum which involves itself in negative numbers, taking into account that

$$-(-a) = a.$$

The following rule concerning positive integers is so natural that you probably would not even think it worth while to take special notice of it. We still state it explicitly.

*If  $a, b$  are positive integers, then  $a + b$  is also a positive integer.*

For instance, 17 and 45 are positive integers, and their sum, 62, is also a positive integer.

We assume this rule concerning positivity. We shall see later that it also applies to positive real numbers. From it we can prove:

*If  $a, b$  are negative integers, then  $a + b$  is negative.*

*Proof.* We can write  $a = -n$  and  $b = -m$ , where  $m, n$  are positive. Therefore

$$a + b = -n - m = -(n + m),$$

which shows that  $a + b$  is negative, because  $n + m$  is positive.

**Example.** If we have the relationship between three numbers

$$a + b = c,$$

then we can derive other relationships between them. For instance, add  $-b$  to both sides of this equation. We get

$$a + b - b = c - b,$$

whence  $a + 0 = c - b$ , or in other words,

$$a = c - b.$$

Similarly, we conclude that

$$b = c - a.$$

For instance, if

$$x + 3 = 5,$$

then

$$x = 5 - 3 = 2.$$

If

$$4 - a = 3,$$

then adding  $a$  to both sides yields

$$4 = 3 + a,$$

and subtracting 3 from both sides yields

$$1 = a.$$

If

$$-2 - y = 5,$$

then

$$-7 = y \quad \text{or} \quad y = -7.$$

## EXERCISES

Justify each step, using commutativity and associativity in proving the following identities.

1.  $(a + b) + (c + d) = (a + d) + (b + c)$
2.  $(a + b) + (c + d) = (a + c) + (b + d)$
3.  $(a - b) + (c - d) = (a + c) + (-b - d)$
4.  $(a - b) + (c - d) = (a + c) - (b + d)$
5.  $(a - b) + (c - d) = (a - d) + (c - b)$
6.  $(a - b) + (c - d) = -(b + d) + (a + c)$
7.  $(a - b) + (c - d) = -(b + d) - (-a - c)$
8.  $((x + y) + z) + w = (x + z) + (y + w)$
9.  $(x - y) - (z - w) = (x + w) - y - z$
10.  $(x - y) - (z - w) = (x - z) + (w - y)$
11. Show that  $-(a + b + c) = -a + (-b) + (-c)$ .
12. Show that  $-(a - b - c) = -a + b + c$ .
13. Show that  $-(a - b) = b - a$ .

Solve for  $x$  in the following equations.

- |                    |                   |
|--------------------|-------------------|
| 14. $-2 + x = 4$   | 15. $2 - x = 5$   |
| 16. $x - 3 = 7$    | 17. $-x + 4 = -1$ |
| 18. $4 - x = 8$    | 19. $-5 - x = -2$ |
| 20. $-7 + x = -10$ | 21. $-3 + x = 4$  |

22. Prove the cancellation law for addition:

$$\boxed{\text{If } a + b = a + c, \text{ then } b = c.}$$

23. Prove: If  $a + b = a$ , then  $b = 0$ .

### §3. RULES FOR MULTIPLICATION

We can multiply integers, and the product of two integers is again an integer. We shall list the rules which apply to multiplication and to its relations with addition.

We again have the rules of *commutativity* and *associativity*:

$$\boxed{ab = ba}$$

and

$$\boxed{(ab)c = a(bc).}$$

We emphasize that these apply whether  $a, b, c$  are negative, positive, or zero. Multiplication is also denoted by a dot. For instance

$$3 \cdot 7 = 21,$$

and

$$(3 \cdot 7) \cdot 4 = 21 \cdot 4 = 84,$$

$$3 \cdot (7 \cdot 4) = 3 \cdot 28 = 84.$$

*For any integer  $a$ , the rules of multiplication by 1 and 0 are:*

N6.

$$\boxed{1a = a}$$

and

$$\boxed{0a = 0.}$$

**Example.** We have

$$\begin{aligned} (2a)(3b) &= 2(a(3b)) \\ &= 2(3a)b \\ &= (2 \cdot 3)ab \\ &= 6ab. \end{aligned}$$

In this example we have done something which is frequently useful, namely we have moved to one side all the explicit numbers like 2, 3 and put on the other side those numbers denoted by a letter like  $a$  or  $b$ . Using commutativity and associativity, we can prove similarly

$$(5x)(7y) = 35xy$$

or, with more factors,

$$(2a)(3b)(5x) = 30abx.$$

We suggest that you carry out the proof of this equality completely, using associativity and commutativity for multiplication.

Finally, we have the rule of *distributivity*, namely

$$a(b + c) = ab + ac$$

and also on the other side,

$$(b + c)a = ba + ca.$$

These rules will not be proved, but will be used constantly. We shall, however, make some comments on them, and prove other rules from them.

First observe that if we just assume distributivity on one side, and commutativity, then we can prove distributivity on the other side. Namely, assuming distributivity on the left, we have

$$(b + c)a = a(b + c) = ab + ac = ba + ca,$$

which is the proof of distributivity on the right.

Observe also that our rule  $0a = 0$  can be proved from the other rules concerning multiplication and the properties of addition. We carry out the proof as an example. We have

$$\cdot 0a + a = 0a + 1a = (0 + 1)a = 1a = a.$$

Thus

$$0a + a = a.$$

Adding  $-a$  to both sides, we obtain

$$0a + a - a = a - a = 0.$$

The left-hand side is simply

$$0a + a - a = 0a + 0 = 0a,$$

so that we obtain  $0a = 0$ , as desired.

We can also prove

N7.

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

*Proof.* We have

$$(-1)a + a = (-1)a + 1a = (-1 + 1)a = 0a = 0.$$

By definition,  $(-1)a + a = 0$  means that  $(-1)a = -a$ , as was to be shown.

We have

N8.

$$-(ab) = (-a)b.$$

*Proof.* We must show that  $(-a)b$  is the negative of  $ab$ . This amounts to showing that

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

But we have by distributivity

$$ab + (-a)b = (a + (-a))b = 0b = 0,$$

thus proving what we wanted.

Similarly, we leave to the reader the proof that

N9.

$$-(ab) = a(-b).$$

**Example.** We have

$$-(3a) = (-3)a = 3(-a).$$

Also,

$$4(a - 5b) = 4a - 20b.$$

Also,

$$-3(5a - 7b) = -15a + 21b.$$

In each of the above cases, you should indicate specifically each one of the rules we have used to derive the desired equality. Again, we emphasize that you should be especially careful when working with negative numbers and repeated minus signs. This is one of the most frequent sources of error when we work with multiplication and addition.

**Example.** We have

$$\begin{aligned} (-2a)(3b)(4c) &= (-2) \cdot 3 \cdot 4abc \\ &= -24abc. \end{aligned}$$

Similarly,

$$\begin{aligned} (-4x)(5y)(-3c) &= (-4)5(-3)xyc \\ &= 60xyc. \end{aligned}$$

Note that the product of two minus signs gives a plus sign.

**Example.** We have

$$(-1)(-1) = 1.$$

To see this, all we have to do is apply our rule

$$-(ab) = (-a)b = a(-b).$$

We find

$$(-1)(-1) = -(1(-1)) = -(-1) = 1.$$

**Example.** More generally, for any integers  $a, b$  we have

N10.

$$(-a)(-b) = ab.$$

We leave the proof as an exercise. From this we see that a product of two negative numbers is positive, because if  $a, b$  are positive and  $-a, -b$  are therefore negative, then  $(-a)(-b)$  is the positive number  $ab$ . For instance,  $-3$  and  $-5$  are negative, but

$$(-3)(-5) = -(3(-5)) = -(-(3 \cdot 5)) = 15.$$

**Example.** A product of a negative number and a positive number is negative. For instance,  $-4$  is negative,  $7$  is positive, and

$$(-4) \cdot 7 = -(4 \cdot 7) = -28,$$

so that  $(-4) \cdot 7$  is negative.

When we multiply a number with itself several times, it is convenient to use a notation to abbreviate this operation. Thus we write

$$\begin{aligned}aa &= a^2, \\aaa &= a^3, \\aaaa &= a^4,\end{aligned}$$

and in general if  $n$  is a positive integer,

$$a^n = aa \cdots a \quad (\text{the product is taken } n \text{ times}).$$

We say that  $a^n$  is the  **$n$ -th power** of  $a$ . Thus  $a^2$  is the second power of  $a$ , and  $a^5$  is the fifth power of  $a$ .

If  $m, n$  are positive integers, then

N11.

$$a^{m+n} = a^m a^n.$$

This simply states that if we take the product of  $a$  with itself  $m + n$  times, then this amounts to taking the product of  $a$  with itself  $m$  times and multiplying this with the product of  $a$  with itself  $n$  times.

### Example

$$a^2 a^3 = (aa)(aaa) = a^{2+3} = aaaaa = a^5.$$

### Example

$$(4x)^2 = 4x \cdot 4x = 4 \cdot 4xx = 16x^2.$$

### Example

$$(7x)(2x)(5x) = 7 \cdot 2 \cdot 5xxx = 70x^3.$$

We have another rule for powers, namely

N12.

$$(a^m)^n = a^{mn}.$$

This means that if we take the product of  $a$  with itself  $m$  times, and then take the product of  $a^m$  with itself  $n$  times, then we obtain the product of  $a$  with itself  $mn$  times.

**Example.** We have

$$(a^3)^4 = a^{12}.$$

**Example.** We have

$$(ab)^n = a^n b^n$$

because

$$\begin{aligned} (ab)^n &= abab \cdots ab && (\text{product of } ab \text{ with itself } n \text{ times}) \\ &= \underbrace{aa \cdots a}_{n} \underbrace{bb \cdots b}_{n} \\ &= a^n b^n. \end{aligned}$$

**Example.** We have

$$(2a^3)^5 = 2^5(a^3)^5 = 32a^{15}.$$

**Example.** The population of a city is 300 thousand in 1930, and doubles every 20 years. What will be the population after 60 years?

This is a case of applying powers. After 20 years, the population is  $2 \cdot 300$  thousand. After 40 years, the population is  $2^2 \cdot 300$  thousand. After 60 years, the population is  $2^3 \cdot 300$  thousand, which is a correct answer. Of course, we can also say that the population will be 2 million 400 thousand.

*The following three formulas are used constantly. They are so important that they should be thoroughly memorized by reading them out loud and repeating them like a poem, to get an aural memory of them.*

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

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

$$(a + b)(a - b) = a^2 - b^2.$$

*Proofs.* The proofs are carried out by applying repeatedly the rules for multiplication. We have:

$$\begin{aligned} (a + b)^2 &= (a + b)(a + b) = a(a + b) + b(a + b) \\ &= aa + ab + ba + bb \\ &= a^2 + ab + ab + b^2 \\ &= a^2 + 2ab + b^2, \end{aligned}$$

which proves the first formula.

$$\begin{aligned}(a - b)^2 &= (a - b)(a - b) = a(a - b) - b(a - b) \\&= aa - ab - ba + bb \\&= a^2 - ab - ab + b^2 \\&= a^2 - 2ab + b^2,\end{aligned}$$

which proves the second formula.

$$\begin{aligned}(a + b)(a - b) &= a(a - b) + b(a - b) = aa - ab + ba - bb \\&= a^2 - ab + ab - b^2 \\&= a^2 - b^2,\end{aligned}$$

which proves the third formula.

**Example.** We have

$$\begin{aligned}(2 + 3x)^2 &= 2^2 + 2 \cdot 2 \cdot 3x + (3x)^2 \\&= 4 + 12x + 9x^2.\end{aligned}$$

**Example.** We have

$$\begin{aligned}(3 - 4x)^2 &= 3^2 - 2 \cdot 3 \cdot 4x + (4x)^2 \\&= 9 - 24x + 16x^2.\end{aligned}$$

**Example.** We have

$$\begin{aligned}(-2a + 5b)^2 &= 4a^2 + 2(-2a)(5b) + 25b^2 \\&= 4a^2 - 20ab + 25b^2.\end{aligned}$$

**Example.** We have

$$\begin{aligned}(4a - 6)(4a + 6) &= (4a)^2 - 36 \\&= 16a^2 - 36.\end{aligned}$$

We have discussed so far examples of products of two factors. Of course, we can take products of more factors using associativity.

**Example.** Expand the expression

$$(2x + 1)(x - 2)(x + 5)$$

as a sum of powers of  $x$  multiplied by integers.

We first multiply the first two factors, and obtain

$$\begin{aligned}(2x + 1)(x - 2) &= 2x(x - 2) + 1(x - 2) \\&= 2x^2 - 4x + x - 2 \\&= 2x^2 - 3x - 2.\end{aligned}$$

We now multiply this last expression with  $x + 5$  and obtain

$$\begin{aligned}(2x + 1)(x - 2)(x + 5) &= (2x^2 - 3x - 2)(x + 5) \\&= (2x^2 - 3x - 2)x + (2x^2 - 3x - 2)5 \\&= 2x^3 - 3x^2 - 2x + 10x^2 - 15x - 10 \\&= 2x^3 + 7x^2 - 17x - 10,\end{aligned}$$

which is the desired answer.

## EXERCISES

1. Express each of the following expressions in the form  $2^m 3^n a^r b^s$ , where  $m$ ,  $n$ ,  $r$ ,  $s$  are positive integers.
- |                                   |                              |
|-----------------------------------|------------------------------|
| a) $8a^2b^3(27a^4)(2^5ab)$        | b) $16b^3a^2(6ab^4)(ab)^3$   |
| c) $3^2(2ab)^3(16a^2b^5)(24b^2a)$ | d) $24a^3(2ab^2)^3(3ab)^2$   |
| e) $(3ab)^2(27a^3b)(16ab^5)$      | f) $32a^4b^5a^3b^2(6ab^3)^4$ |

2. Prove:

$$\begin{aligned}(a + b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3, \\(a - b)^3 &= a^3 - 3a^2b + 3ab^2 - b^3.\end{aligned}$$

3. Obtain expansions for  $(a + b)^4$  and  $(a - b)^4$  similar to the expansions for  $(a + b)^3$  and  $(a - b)^3$  of the preceding exercise.

Expand the following expressions as sums of powers of  $x$  multiplied by integers. These are in fact called polynomials. You might want to read, or at least look at, the section on polynomials later in the book (Chapter 13, §2).

- |                 |                 |
|-----------------|-----------------|
| 4. $(2 - 4x)^2$ | 5. $(1 - 2x)^2$ |
| 6. $(2x + 5)^2$ | 7. $(x - 1)^2$  |

8.  $(x + 1)(x - 1)$       9.  $(2x + 1)(x + 5)$   
 10.  $(x^2 + 1)(x^2 - 1)$       11.  $(1 + x^3)(1 - x^3)$   
 12.  $(x^2 + 1)^2$       13.  $(x^2 - 1)^2$   
 14.  $(x^2 + 2)^2$       15.  $(x^2 - 2)^2$   
 16.  $(x^3 - 4)^2$       17.  $(x^3 - 4)(x^3 + 4)$   
 18.  $(2x^2 + 1)(2x^2 - 1)$       19.  $(-2 + 3x)(-2 - 3x)$   
 20.  $(x + 1)(2x + 5)(x - 2)$       21.  $(2x + 1)(1 - x)(3x + 2)$   
 22.  $(3x - 1)(2x + 1)(x + 4)$       23.  $(-1 - x)(-2 + x)(1 - 2x)$   
 24.  $(-4x + 1)(2 - x)(3 + x)$       25.  $(1 - x)(1 + x)(2 - x)$   
 26.  $(x - 1)^2(3 - x)$       27.  $(1 - x)^2(2 - x)$   
 28.  $(1 - 2x)^2(3 + 4x)$       29.  $(2x + 1)^2(2 - 3x)$
30. The population of a city in 1910 was 50,000, and it doubles every 10 years. What will it be (a) in 1970 (b) in 1990 (c) in 2,000?
31. The population of a city in 1905 was 100,000, and it doubles every 25 years. What will it be after (a) 50 years (b) 100 years (c) 150 years?
32. The population of a city was 200 thousand in 1915, and it triples every 50 years. What will be the population  
 a) in the year 2215?      b) in the year 2165?
33. The population of a city was 25,000 in 1870, and it triples every 40 years. What will it be  
 a) in 1990?      b) in 2030?

#### §4. EVEN AND ODD INTEGERS; DIVISIBILITY

We consider the positive integers 1, 2, 3, 4, 5, . . . , and we shall distinguish between two kinds of integers. We call

$$1, 3, 5, 7, 9, 11, 13, \dots$$

the **odd integers**, and we call

$$2, 4, 6, 8, 10, 12, 14, \dots$$

the **even integers**. Thus the odd integers go up by 2 and the even integers go up by 2. The odd integers start with 1, and the even integers start with 2. Another way of describing an even integer is to say that it is a positive integer which can be written in the form  $2n$  for some positive integer  $n$ . For instance, we can write

$$2 = 2 \cdot 1,$$

$$4 = 2 \cdot 2,$$

$$6 = 2 \cdot 3,$$

$$8 = 2 \cdot 4,$$

and so on. Similarly, an odd integer is an integer which differs from an even integer by 1, and thus can be written in the form  $2m - 1$  for some positive integer  $m$ . For instance,

$$1 = 2 \cdot 1 - 1,$$

$$3 = 2 \cdot 2 - 1,$$

$$5 = 2 \cdot 3 - 1,$$

$$7 = 2 \cdot 4 - 1,$$

$$9 = 2 \cdot 5 - 1,$$

and so on. Note that we can also write an odd integer in the form

$$2n + 1$$

if we allow  $n$  to be a natural number, i.e., allowing  $n = 0$ . For instance, we have

$$1 = 2 \cdot 0 + 1,$$

$$3 = 2 \cdot 1 + 1,$$

$$5 = 2 \cdot 2 + 1,$$

$$7 = 2 \cdot 3 + 1,$$

$$9 = 2 \cdot 4 + 1,$$

and so on.

**Theorem 1.** *Let  $a, b$  be positive integers.*

*If  $a$  is even and  $b$  is even, then  $a + b$  is even.*

*If  $a$  is even and  $b$  is odd, then  $a + b$  is odd.*

*If  $a$  is odd and  $b$  is even, then  $a + b$  is odd.*

*If  $a$  is odd and  $b$  is odd, then  $a + b$  is even.*

*Proof.* We shall prove the second statement, and leave the others as exercises. Assume that  $a$  is even and that  $b$  is odd. Then we can write

$$a = 2n \quad \text{and} \quad b = 2k + 1$$

for some positive integer  $n$  and some natural number  $k$ . Then

$$\begin{aligned} a + b &= 2n + 2k + 1 \\ &= 2(n + k) + 1 \\ &= 2m + 1 \end{aligned} \quad (\text{letting } m = n + k).$$

This proves that  $a + b$  is odd.

**Theorem 2.** *Let  $a$  be a positive integer. If  $a$  is even, then  $a^2$  is even. If  $a$  is odd, then  $a^2$  is odd.*

*Proof.* Assume that  $a$  is even. This means that  $a = 2n$  for some positive integer  $n$ . Then

$$a^2 = 2n \cdot 2n = 2(2n^2) = 2m,$$

where  $m = 2n^2$  is a positive integer. Thus  $a^2$  is even.

Next, assume that  $a$  is odd, and write  $a = 2n + 1$  for some natural number  $n$ . Then

$$\begin{aligned} a^2 &= (2n + 1)^2 = (2n)^2 + 2(2n)1 + 1^2 \\ &= 4n^2 + 4n + 1 \\ &= 2(2n^2 + 2n) + 1 \\ &= 2k + 1, \end{aligned} \quad \text{where } k = 2n^2 + 2n.$$

Hence  $a^2$  is odd, thus proving our theorem.

**Corollary.** *Let  $a$  be a positive integer. If  $a^2$  is even, then  $a$  is even. If  $a^2$  is odd, then  $a$  is odd.*

*Proof.* This is really only a reformulation of the theorem, taking into account ordinary logic. If  $a^2$  is even, then  $a$  cannot be odd because the square of an odd number is odd. If  $a^2$  is odd, then  $a$  cannot be even because the square of an even number is even.

We can generalize the property used to define an even integer. Let  $d$  be a positive integer and let  $n$  be an integer. We shall say that  $d$  divides  $n$ , or that  $n$  is divisible by  $d$  if we can write

$$n = dk$$

for some integer  $k$ . Thus an even integer is a positive integer which is divisible by 2. According to our definition, the number 9 is divisible by 3 because

$$9 = 3 \cdot 3.$$

Also, 15 is divisible by 3 because

$$15 = 3 \cdot 5.$$

Also,  $-30$  is divisible by 5 because

$$-30 = 5(-6).$$

Note that every integer is divisible by 1, because we can always write

$$n = 1 \cdot n.$$

Furthermore, every positive integer is divisible by itself.

### EXERCISES

1. Give the proofs for the cases of Theorem 1 which were not proved in the text.
2. Prove: If  $a$  is even and  $b$  is any positive integer, then  $ab$  is even.
3. Prove: If  $a$  is even, then  $a^3$  is even.
4. Prove: If  $a$  is odd, then  $a^3$  is odd.
5. Prove: If  $n$  is even, then  $(-1)^n = 1$ .
6. Prove: If  $n$  is odd, then  $(-1)^n = -1$ .
7. Prove: If  $m, n$  are odd, then the product  $mn$  is odd.

Find the largest power of 2 which divides the following integers.

- |        |        |         |        |
|--------|--------|---------|--------|
| 8. 16  | 9. 24  | 10. 32  | 11. 20 |
| 12. 50 | 13. 64 | 14. 100 | 15. 36 |

Find the largest power of 3 which divides the following integers.

- |        |        |        |        |
|--------|--------|--------|--------|
| 16. 30 | 17. 27 | 18. 63 | 19. 99 |
| 20. 60 | 21. 50 | 22. 42 | 23. 45 |

24. Let  $a, b$  be integers. Define  $a \equiv b \pmod{5}$ , which we read “ $a$  is congruent to  $b$  modulo 5”, to mean that  $a - b$  is divisible by 5. Prove: If  $a \equiv b \pmod{5}$  and  $x \equiv y \pmod{5}$ , then

$$a + x \equiv b + y \pmod{5}$$

and

$$ax \equiv by \pmod{5}.$$

25. Let  $d$  be a positive integer. Let  $a, b$  be integers. Define

$$a \equiv b \pmod{d}$$

to mean that  $a - b$  is divisible by  $d$ . Prove that if  $a \equiv b \pmod{d}$  and  $x \equiv y \pmod{d}$ , then

$$a + x \equiv b + y \pmod{d}$$

and

$$ax \equiv by \pmod{d}.$$

26. Assume that every positive integer can be written in one of the forms  $3k$ ,  $3k + 1$ ,  $3k + 2$  for some integer  $k$ . Show that if the square of a positive integer is divisible by 3, then so is the integer.

## §5. RATIONAL NUMBERS

By a **rational number** we shall mean simply an ordinary fraction, that is a quotient

$$\frac{m}{n} \quad \text{also written} \quad m/n,$$

where  $m, n$  are integers and  $n \neq 0$ . In taking such a quotient  $m/n$ , we emphasize that we cannot divide by 0, and thus we must always be sure that  $n \neq 0$ . For instance,

$$\frac{1}{4}, \frac{2}{3}, -\frac{3}{4}, -\frac{5}{7}$$

are rational numbers. Finite decimals also give us examples of rational numbers. For instance,

$$1.4 = \frac{14}{10} \quad \text{and} \quad 1.41 = \frac{141}{100}.$$

Just as we did with the integers, we can represent the rational numbers on the line. For instance,  $\frac{1}{2}$  lies one-half of the way between 0 and 1, while

$\frac{2}{3}$  lies two-thirds of the way between 0 and 1, as shown on the following picture.



Fig. 1-5

The negative rational number  $-\frac{3}{4}$  lies on the opposite side of 0 at a distance  $\frac{3}{4}$  from 0. On the next picture, we have drawn  $-\frac{3}{4}$  and  $-\frac{5}{4}$ .

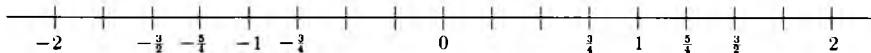


Fig. 1-6

There is no unique representation of a rational number as a quotient of two integers. For instance, we have

$$\frac{1}{2} = \frac{2}{4}.$$

We can interpret this geometrically on the line. If we cut up the segment between 0 and 1 into four equal pieces, and we take two-fourths of them, then this is the same as taking one-half of the segment. Picture:



Fig. 1-7

We need a general rule to determine when two expressions of quotients of integers give the same rational numbers. We assume this rule without proof. It is stated as follows.

**Rule for cross-multiplying.** Let  $m, n, r, s$  be integers and assume that  $n \neq 0$  and  $s \neq 0$ . Then

$$\frac{m}{n} = \frac{r}{s} \quad \text{if and only if} \quad ms = rn.$$

The name “cross-multiplying” comes from our visualization of the rule in the following diagram:

$$\frac{m}{n} \times \frac{r}{s}.$$

**Example.** We have

$$\frac{1}{2} = \frac{2}{4}$$

because

$$1 \cdot 4 = 2 \cdot 2.$$

Also, we have

$$\frac{3}{7} = \frac{9}{21}$$

because

$$3 \cdot 21 = 9 \cdot 7$$

(both sides are equal to 63).

We shall make no distinction between an integer  $m$  and the rational number  $m/1$ . Thus we write

$$m = m/1 = \frac{m}{1}.$$

With this convention, we see that every integer is also a rational number. For instance,  $3 = 3/1$  and  $-4 = -4/1$ .

Observe the special case of cross-multiplying when one side is an integer. For instance:

$$\frac{2n}{5} = \frac{6}{1}, \quad \frac{2n}{5} = 6, \quad 2n = 30, \quad n = \frac{30}{2} = 15$$

are all equivalent formulations of a relation involving  $n$ .

Of course, cross-multiplying also works with negative numbers. For instance,

$$\frac{-4}{5} = \frac{8}{-10}$$

because

$$(-4)(-10) = 8 \cdot 5$$

(both sides are equal to 40).

**Remark.** For the moment, we are dealing with quotients of integers and describing how they behave. In the next section we shall deal with multiplicative inverses. There, you can see how the rule for cross-multiplication can in fact be proved from properties of such an inverse. Some people view this proof as the reason why cross-multiplication “works”. However, in some contexts, one wants to *define* the multiplicative inverse by using the rule for cross-multiplication. This is the reason for emphasizing it here independently.

**Cancellation rule for fractions.** Let  $a$  be a non-zero integer. Let  $m, n$  be integers,  $n \neq 0$ . Then

$$\frac{am}{an} = \frac{m}{n}.$$

*Proof.* To test equality, we apply the rule for cross-multiplying. We must verify that

$$(am)n = m(an),$$

which we see is true by associativity and commutativity.

The examples which we gave are special cases of this cancellation rule. For instance

$$\frac{-4}{5} = \frac{(-2)(-4)}{(-2)5} = \frac{8}{-10}.$$

In dealing with quotients of integers which may be negative, it is useful to observe that

$$\frac{-m}{n} = \frac{m}{-n}.$$

This is proved by cross-multiplying, namely we must verify that

$$(-m)(-n) = mn,$$

which we already know is true.

The cancellation rule leads us to use the notion of divisibility already mentioned in §4. Indeed, suppose that  $d$  is a positive integer and  $m, n$  are divisible by  $d$  (or as we also say, that  $d$  is a **common divisor** of  $m$  and  $n$ ). Then we can write

$$m = dr \quad \text{and} \quad n = ds$$

for some integers  $r$  and  $s$ , so that

$$\frac{m}{n} = \frac{dr}{ds} = \frac{r}{s}.$$

We see that our cancellation rule is applicable.

**Example.** We have

$$\frac{10}{15} = \frac{2 \cdot 5}{3 \cdot 5} = \frac{2}{3}$$

because 10 and 15 are both divisible by 5.

We say that a rational number is **positive** if it can be written in the form  $m/n$ , where  $m, n$  are positive integers. Let  $a$  be a positive rational number. We shall say that  $a$  is **expressed in lowest form** as a fraction

$$a = \frac{r}{s}$$

where  $r, s$  are positive integers if the only common divisor of  $r$  and  $s$  is 1.

**Theorem 3.** *Any positive rational number has an expression as a fraction in lowest form.*

*Proof.* First write a given positive rational number as a quotient of positive integers  $m/n$ . We know that 1 is a common divisor of  $m$  and  $n$ . Furthermore, any common divisor is at most equal to  $m$  or  $n$ . Thus among all common divisors there is a greatest one, which we denote by  $d$ . Thus we can write

$$m = dr \quad \text{and} \quad n = ds$$

with positive integers  $r$  and  $s$ . Our rational number is equal to

$$\frac{m}{n} = \frac{dr}{ds} = \frac{r}{s}.$$

All we have to do now is to show that the only common divisor of  $r$  and  $s$  is 1. Suppose that  $e$  is a common divisor which is greater than 1. Then we can write

$$r = ex \quad \text{and} \quad s = ey$$

with positive integers  $x$  and  $y$ . Hence

$$m = dr = dex \quad \text{and} \quad n = ds = dey.$$

Therefore  $de$  is a common divisor for  $m$  and  $n$ , and is greater than  $d$  since  $e$  is greater than 1. This is impossible because we assumed that  $d$  was the greatest common divisor of  $m$  and  $n$ . Therefore 1 is the only common divisor of  $r$  and  $s$ , and our theorem is proved.

**Example.** Any positive rational number can be expressed as a quotient  $m/n$ , where  $m, n$  are positive integers which are not both even, because if  $m/n$  is the expression of this rational number in lowest form, then 2 cannot divide both  $m$  and  $n$ , and therefore at least one of them must be odd.

Let

$$\frac{m}{n} \quad \text{and} \quad \frac{r}{s}$$

be rational numbers, expressed as quotients of integers. We can put these rational numbers over a common denominator  $ns$  by writing

$$\frac{m}{n} = \frac{ms}{ns} \quad \text{and} \quad \frac{r}{s} = \frac{nr}{ns}.$$

For instance, to put  $\frac{3}{5}$  and  $\frac{5}{7}$  over the common denominator  $5 \cdot 7 = 35$ , we write

$$\frac{3}{5} = \frac{3 \cdot 7}{5 \cdot 7} = \frac{21}{35} \quad \text{and} \quad \frac{5}{7} = \frac{5 \cdot 5}{7 \cdot 5} = \frac{25}{35}.$$

This leads us to the formula for the addition of rational numbers. Consider first a special case, when the rational numbers have a common denominator, for instance,

$$\frac{3}{5} + \frac{8}{5} = \frac{11}{5}.$$

This is reasonable just from the interpretation of rational numbers: If we have three-fifths of something, and add eight-fifths of that same thing, then we get eleven-fifths of that thing. In general, we can write the rule for addition when the rational numbers have a common denominator as

$$\frac{a}{d} + \frac{b}{d} = \frac{a+b}{d}.$$

**Example.** We have

$$\frac{-5}{8} + \frac{2}{8} = \frac{-3}{8}.$$

When the rational numbers do not have a common denominator, we get the formula for their addition by putting them over a common denominator.

Namely, let  $\frac{m}{n}$  and  $\frac{r}{s}$  be rational numbers, expressed as quotients of integers  $m, n$  and  $r, s$  with  $n \neq 0$  and  $s \neq 0$ . Then we have seen that

$$\frac{m}{n} = \frac{sm}{sn} \quad \text{and} \quad \frac{r}{s} = \frac{nr}{ns}.$$

Thus our rational numbers now have the common denominator  $sn$ , and thus the formula for addition in this general case is

$$\frac{m}{n} + \frac{r}{s} = \frac{ms+rn}{ns}.$$

**Example.** We have

$$\frac{3}{5} + \frac{4}{7} = \frac{3 \cdot 7 + 4 \cdot 5}{35} = \frac{21 + 20}{35} = \frac{41}{35}.$$

**Example.** We have

$$\frac{-5}{2} + \frac{3}{7} = \frac{(-5) \cdot 7 + 2 \cdot 3}{14} = \frac{-29}{14}.$$

**Example.** We have

$$\frac{3}{-4} + \frac{5}{7} = \frac{21 - 20}{-28} = \frac{1}{-28}.$$

Using our rule for adding rational numbers, we conclude at once:

*The sum of positive rational numbers is also positive.*

Observe that our number 0 has the property that

$$\frac{0}{n} = \frac{0}{1} = 0$$

for any integer  $n \neq 0$ . Indeed, applying our test for the equality of two fractions, we must verify that

$$0 \cdot 1 = 0 \cdot n,$$

and this is true because both sides are equal to 0.

*For any rational number  $a$ , we have*

$$0 + a = a + 0 = a.$$

This is easily seen using the analogous property for integers. Namely, write  $a = m/n$ , where  $m, n$  are integers, and  $n \neq 0$ . Then

$$0 + a = \frac{0}{n} + \frac{m}{n} = \frac{0 + m}{n} = \frac{m}{n} = a,$$

and similarly on the other side.

Let  $a = m/n$  be a rational number, where  $m, n$  are integers and  $n \neq 0$ . Then we have

$$\frac{-m}{n} + \frac{m}{n} = \frac{-m + m}{n} = 0.$$

For this reason, we shall write

$$\boxed{\frac{-m}{n} = -\frac{m}{n}}.$$

By a previous remark, we also see that

$$\boxed{-\frac{m}{n} = \frac{m}{-n}}.$$

This shows how a minus sign can be moved around the various terms of a fraction without changing the value of the fraction.

A rational number which can be written as a fraction

$$-\frac{m}{n} = \frac{-m}{n} = \frac{m}{-n}$$

where  $m, n$  are positive integers will be called **negative**. For example, the number

$$\frac{3}{-5} = \frac{-3}{5} = -\frac{3}{5}$$

is negative. Using the definition of addition of rational numbers, you can easily verify for yourselves that a sum of negative rational numbers is negative.

*Addition of rational numbers satisfies the properties of commutativity and associativity.*

Just as we did for integers, the above statement will be accepted without proof. It is in fact a general property of much more general numbers, which will be restated again for these numbers in the next section.

*In §2, we proved a number of properties of addition using only commutativity and associativity, together with the rules*

$$0 + a = a \quad \text{and} \quad a + (-a) = 0.$$

*These properties therefore remain valid for rational numbers. Similarly, all the exercises of §2 remain valid for rational numbers.*

This remark will again be made later whenever we meet a similar situation. For instance, we see as before that

$$\text{if } a + b = 0, \text{ then } b = -a.$$

We just add  $-a$  to both sides of the equation  $a + b = 0$ . In words, we can say: To test whether a given rational number is equal to minus another, all we need to verify is that the sum of the numbers is equal to 0.

We shall now give the formula for **multiplication** of rational numbers. This formula is:

$$\frac{m}{n} \cdot \frac{r}{s} = \frac{mr}{ns}.$$

Thus to take the product of two rational numbers, we multiply their numerators and multiply their denominators. More precisely, the numerator of the product is the product of the numerators, and the denominator of the product is the product of the denominators.

**Example.** We have

$$\frac{3}{5} \cdot \frac{7}{8} = \frac{21}{40}.$$

Also,

$$\frac{2}{7} \cdot \frac{11}{16} = \frac{22}{112}.$$

We can write this last fraction in simpler form, namely

$$\frac{2}{7} \cdot \frac{11}{16} = \frac{2 \cdot 11}{7 \cdot 2 \cdot 8}.$$

We can then cancel 2 and get

$$\frac{2}{7} \cdot \frac{11}{16} = \frac{11}{56}.$$

This shows that sometimes it is best not to carry out a multiplication before looking at the possibility of cancellations.

**Example.** We have

$$\frac{-4}{5} \cdot \frac{7}{-3} = \frac{(-4)7}{5(-3)} = \frac{-28}{-15} = \frac{28}{15}.$$

**Example.** Let  $a = m/n$  be a rational number expressed as a quotient of integers. Then

$$a^2 = \left(\frac{m}{n}\right)^2 = \frac{m}{n} \cdot \frac{m}{n} = \frac{m^2}{n^2}.$$

Similarly,

$$a^3 = \frac{m}{n} \cdot \frac{m}{n} \cdot \frac{m}{n} = \frac{m^3}{n^3}.$$

In general, for any positive integer  $k$ , we have

$$a^k = \left(\frac{m}{n}\right)^k = \frac{m^k}{n^k}.$$

**Example.** We have

$$\left(\frac{1}{2}\right)^3 = \frac{1}{2^3} = \frac{1}{8}.$$

Also,

$$\left(\frac{3}{5}\right)^4 = \frac{3^4}{5^4} = \frac{81}{525}.$$

**Example.** A chemical substance disintegrates in such a way that it gets halved every 10 min. If there are 20 grams (g) of the substance present at a given time, how much will be left after 50 min?

This is easily done. At the end of 10 min, we have  $\frac{1}{2} \cdot 20$  g left. At the end of 20 min, we have  $\frac{1}{2^2} \cdot 20$  g left, and so on; at the end of 50 min, we have

$$\frac{1}{2^5} \cdot 20 = \frac{20}{32}$$

grams left. This is a correct answer. If you want to put the fraction in lowest form, you may do so, and then you get the answer in the form  $\frac{5}{8}$  g. You can also put it in approximate decimals, which we don't do here.

We ask: Is there a positive rational number  $a$  whose square is 2? The answer is at first not obvious. Such a number would be a square root of 2. Note that  $1^2 = 1 \cdot 1 = 1$  and  $2^2 = 4$ . Thus the square of 1 is smaller than 2 and the square of 2 is bigger than 2. Any positive square root of 2 will therefore lie between 1 and 2 if it exists. We could experiment with various decimals to see whether they yield a square root of 2. For instance, let us try the decimal just in the middle between 1 and 2. We have

$$(1.5)^2 = 2.25,$$

which is bigger than 2. Thus 1.5 is not a square root of 2, and is too big to be one.

We could try more systematically, namely:

$$\begin{aligned}(1.1)^2 &= 1.21 \quad (\text{too small}), \\ (1.2)^2 &= 1.44 \quad (\text{too small}), \\ (1.3)^2 &= 1.69 \quad (\text{too small}), \\ (1.4)^2 &= 1.96 \quad (\text{too small but coming closer}).\end{aligned}$$

We know that 1.5 is too big, and hence we must go to the next decimal place to try out further.

$$\begin{aligned}(1.41)^2 &= 1.9881 \quad (\text{too small}), \\ (1.42)^2 &= 2.0164 \quad (\text{too big}).\end{aligned}$$

Thus we must go to the next decimal place for further experimentation. We try successively  $(1.411)^2$ ,  $(1.412)^2$ ,  $(1.413)^2$ ,  $(1.414)^2$  and find that they are too small. Computing  $(1.415)^2$  we see that it is too big. We could keep on going like this. There are several things to be said about our procedure.

- (1) It is very systematic, and could be programmed on a computer.
- (2) It gives us increasingly good approximations to a square root of 2, namely it gives us rational numbers whose squares come closer and closer to 2.

However, to find a rational number whose square is 2, the procedure is a bummer because of the following theorem.

**Theorem 4.** *There is no positive rational number whose square is 2.*

*Proof.* Suppose that such a rational number exists. We can write it in lowest form  $m/n$  by Theorem 3. In particular, not both  $m$  and  $n$  can be even. We have

$$\left(\frac{m}{n}\right)^2 = \frac{m^2}{n^2} = 2.$$

Consequently, we obtain

$$m^2 = 2n^2,$$

and therefore  $m^2$  is even. By the Corollary of Theorem 2 of §4, we conclude that  $m$  must be even, and we can therefore write

$$m = 2k$$

for some positive integer  $k$ . Thus we obtain

$$m^2 = (2k)^2 = 4k^2 = 2n^2.$$

We can cancel 2 from both sides of the equation

$$4k^2 = 2n^2,$$

and obtain

$$n^2 = 2k^2.$$

This means that  $n^2$  is even, and as before, we conclude that  $n$  itself must be even. Thus from our original assumption that  $(m/n)^2 = 2$  and  $m/n$  is in lowest form, we have obtained the impossible fact that both  $m, n$  are even. This means that our original assumption  $(m/n)^2 = 2$  cannot be true, and concludes the proof of our theorem.

A number which is not rational is called **irrational**. From Theorem 4, we see that if a positive number  $a$  exists such that  $a^2 = 2$ , then  $a$  must be irrational. We shall discuss this further in the next section dealing with real numbers in general.

Multiplication of rational numbers satisfies the same basic rules as multiplication of integers. We state these once more:

*For any rational number  $a$  we have  $1a = a$  and  $0a = 0$ . Furthermore, multiplication is associative, commutative, and distributive with respect to addition.*

As before, we *assume* these as properties of numbers. Moreover, we have the same remark for multiplication that we did for addition. All the properties of §3 which were proved using only the basic ones are therefore also valid for rational numbers. Thus the formulas which we had, like

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

are now seen to be valid for rational numbers as well. All the exercises at the end of §3 are valid for rational numbers.

**Example.** Solve for  $a$  in the equation

$$3a - 1 = 7.$$

We add 1 to both sides of the equation, and thus obtain

$$3a = 7 + 1 = 8.$$

We then divide by 3 and get

$$a = \frac{8}{3}.$$

**Example.** Solve for  $x$  in the equation

$$2(x - 3) = 7.$$

To do this, we use distributivity first, and get the equivalent equation

$$2x - 6 = 7.$$

Next we find

$$2x = 7 + 6 = 13,$$

whence

$$x = \frac{13}{2}.$$

Of course we could have given other arguments to find the answer. For instance, we could first get

$$x - 3 = \frac{7}{2},$$

whence

$$x = \frac{7}{2} + 3.$$

This is a perfectly correct answer. However, we can also give the answer in fraction form. We write  $3 = \frac{6}{2}$ , and find that

$$x = \frac{7}{2} + \frac{6}{2} = \frac{13}{2}.$$

**Example.** Solve for  $x$  in the equation

$$\frac{3x - 7}{2} + 4 = 2x.$$

We multiply both sides of the equation by 2 and obtain

$$3x - 7 + 8 = 4x.$$

We then add  $-3x$  to both sides, to get

$$1 = 4x - 3x = x.$$

This solves our problem.

**EXERCISES**

1. Solve for  $a$  in the following equations.

a)  $2a = \frac{3}{4}$

b)  $\frac{3a}{5} = -7$

c)  $\frac{-5a}{2} = \frac{3}{8}$

2. Solve for  $x$  in the following equations.

a)  $3x - 5 = 0$

b)  $-2x + 6 = 1$

c)  $-7x = 2$

3. Put the following fractions in lowest form.

a)  $\frac{10}{25}$

b)  $\frac{3}{9}$

c)  $\frac{30}{25}$

d)  $\frac{50}{15}$

e)  $\frac{45}{9}$

f)  $\frac{62}{4}$

g)  $\frac{23}{46}$

h)  $\frac{16}{40}$

4. Let  $a = m/n$  be a rational number expressed as a quotient of integers  $m, n$  with  $m \neq 0$  and  $n \neq 0$ . Show that there is a rational number  $b$  such that  $ab = ba = 1$ .

5. Solve for  $x$  in the following equations.

a)  $2x - 7 = 21$       b)  $3(2x - 5) = 7$       c)  $(4x - 1)2 = \frac{1}{4}$

d)  $-4x + 3 = 5x$       e)  $3x - 2 = -5x + 8$       f)  $3x + 2 = -3x + 4$

g)  $\frac{4x}{3} + 1 = 3x$       h)  $-\frac{3x}{2} + \frac{4}{3} = 5x$       i)  $\frac{2x - 1}{3} + 4x = 10$

6. Solve for  $x$  in the following equations.

a)  $2x - \frac{3}{7} = \frac{x}{5} + 1$       b)  $\frac{3}{4}x + 5 = -7x$       c)  $\frac{-2}{13}x = 3x - 1$

d)  $\frac{4x}{3} + \frac{3}{4} = 2x - 5$       e)  $\frac{4(1 - 3x)}{7} = 2x - 1$       f)  $\frac{2 - x}{3} = \frac{7}{8}x$

7. Let  $n$  be a positive integer. By  $n$  factorial, written  $n!$ , we mean the product

$$1 \cdot 2 \cdot 3 \cdots n$$

of the first  $n$  positive integers. For instance,

$$2! = 2,$$

$$3! = 2 \cdot 3 = 6,$$

$$4! = 2 \cdot 3 \cdot 4 = 24.$$

a) Find the value of  $5!$ ,  $6!$ ,  $7!$ , and  $8!$ .

- b) Define  $0! = 1$ . Define the **binomial coefficient**

$$\binom{m}{n} = \frac{m!}{n!(m-n)!}$$

for any natural numbers  $m, n$  such that  $n$  lies between 0 and  $m$ . Compute the binomial coefficients

$$\begin{aligned} & \binom{3}{0}, \binom{3}{1}, \binom{3}{2}, \binom{3}{3}, \binom{4}{0}, \binom{4}{1}, \binom{4}{2}, \binom{4}{3}, \binom{4}{4}, \\ & \binom{5}{0}, \binom{5}{1}, \binom{5}{2}, \binom{5}{3}, \binom{5}{4}, \binom{5}{5}. \end{aligned}$$

The binomial coefficient  $\binom{m}{n}$  is equal to the number of ways  $n$  things can be selected out of  $m$  things. You may want to look at the discussion of Chapter 16, §1 at this time to see why this is so.

- c) Show that

$$\binom{m}{n} = \binom{m}{m-n}.$$

- d) Show that if  $n$  is a positive integer at most equal to  $m$ , then

$$\binom{m}{n} + \binom{m}{n-1} = \binom{m+1}{n}.$$

8. Prove that there is no positive rational number  $a$  such that  $a^3 = 2$ .
9. Prove that there is no positive rational number  $a$  such that  $a^4 = 2$ .
10. Prove that there is no positive rational number  $a$  such that  $a^2 = 3$ . You may assume that a positive integer can be written in one of the forms  $3k, 3k+1, 3k+2$  for some integer  $k$ . Prove that if the square of a positive integer is divisible by 3, then so is the integer. Then use a similar proof as for  $\sqrt{2}$ .
11. a) Find a positive rational number, expressed as a decimal, whose square approximates 2 up to 3 decimals.  
 b) Same question, but with 4 decimals accuracy instead.
12. a) Find a positive rational number, expressed as a decimal, whose square approximates 3 up to 2 decimals.  
 b) Same question but with 3 decimals instead.
13. Find a positive rational number, expressed as a decimal, whose square approximates 5 up to  
 a) 2 decimals,  
 b) 3 decimals.

14. Find a positive rational number whose cube approximates 2 up to  
a) 2 decimals, b) 3 decimals.

15. Find a positive rational number whose cube approximates 3 to  
a) 2 decimals, b) 3 decimals.

16. A chemical substance decomposes in such a way that it halves every 3 min. If there are 6 grams (g) of the substance present at the beginning, how much will be left  
a) after 3 min? b) after 27 min? c) after 36 min?

17. A chemical substance reacts in such a way that one third of the remaining substances decomposes every 15 min. If there are 15 g of the substance present at the beginning, how much will be left  
a) after 30 min? b) after 45 min? c) after 165 min?

18. A substance reacts in water in such a way that one-fourth of the undissolved part dissolves every 10 min. If you put 25 g of the substance in water at a given time, how much will be left after  
a) 10 min? b) 30 min? c) 50 min?

19. You are testing the effect of a noxious substance on bacteria. Every 10 min, one-tenth of the bacteria which are still alive are killed. If the population of bacteria starts with  $10^6$ , how many bacteria are left after  
a) 10 min? b) 30 min? c) 50 min?  
d) Within which period of 10 min will half the bacteria be killed?  
e) Within which period of 10 min will 70% of the bacteria be killed?  
f) Within which period of 10 min will 80% of the bacteria be killed?  
[Note: If one-tenth of those alive are killed, then nine-tenths remain.]

20. A chemical pollutant is being emptied in a lake with 50,000 fishes. Every month, one-third of the fish still alive die from this pollutant. How many fish will be alive after  
a) 1 month? b) 2 months?  
c) 4 months? d) 6 months?  
(Give your answer to the nearest 100.)  
e) What is the first month when more than half the fish will be dead?  
f) During which month will 80% of the fish be dead?  
[Note: If one-third die, then two thirds remain.]

21. Every 10 years the population of a city is five-fourths of what it was 10 years before. How many years does it take  
a) before the population doubles? b) before it triples?

## §6. MULTIPLICATIVE INVERSES

Rational numbers satisfy one property which is not satisfied by integers, namely:

*If  $a$  is a rational number  $\neq 0$ , then there exists a rational number, denoted by  $a^{-1}$ , such that*

$$a^{-1}a = aa^{-1} = 1.$$

Indeed, if  $a = m/n$  where  $m, n$  are integers  $\neq 0$ , then  $a^{-1} = n/m$  because

$$\frac{m}{n} \cdot \frac{n}{m} = \frac{mn}{mn} = 1.$$

We call  $a^{-1}$  the **multiplicative inverse of  $a$** .

**Example.** The multiplicative inverse of  $\frac{1}{2}$  is  $\frac{2}{1}$ , or simply 2, because

$$2 \cdot \frac{1}{2} = 1.$$

The multiplicative inverse of  $\frac{2}{3}$  is  $\frac{3}{2}$ . The multiplicative inverse of  $-\frac{5}{7}$  is  $-\frac{7}{5}$ .

*Observe that if  $a$  and  $b$  are rational numbers such that*

$$ab = 1,$$

*then*

$$b = a^{-1}.$$

*Proof.* We multiply both sides of the relation  $ab = 1$  by  $a^{-1}$ , and get

$$a^{-1}ab = a^{-1} \cdot 1 = a^{-1}.$$

Using associativity on the left, we find

$$a^{-1}ab = 1b = b,$$

so that we do find  $b = a^{-1}$  as desired.

From the existence of an inverse for non-zero rational numbers, we deduce:

*If  $ab = 0$ , then  $a = 0$  or  $b = 0$ .*

*Proof.* Suppose  $a \neq 0$ . Multiply both sides of the equation  $ab = 0$  by  $a^{-1}$ . We get:

$$a^{-1}ab = 0a^{-1} = 0.$$

On the other hand,  $a^{-1}ab = 1b = b$ , so that we find  $b = 0$ , as desired.

We shall use the same notation as for quotients of integers in taking quotients of rational numbers. We write

$$\frac{a}{b} \quad \text{or} \quad a/b \quad \text{instead of} \quad b^{-1}a \quad \text{or} \quad ab^{-1}.$$

**Example.** Let  $a = \frac{3}{4}$  and  $b = \frac{5}{7}$ . Then

$$\frac{3/4}{5/7} = \frac{3}{4} \left( \frac{5}{7} \right)^{-1} = \frac{3}{4} \cdot \frac{7}{5} = \frac{21}{20}.$$

**Example.** We have

$$\begin{aligned} \frac{1 + \frac{1}{2}}{2 - \frac{4}{3}} &= \left( 1 + \frac{1}{2} \right) \cdot \left( 2 - \frac{4}{3} \right)^{-1} \\ &= \frac{2 + 1}{2} \cdot \left( \frac{6 - 4}{3} \right)^{-1} \\ &= \frac{3}{2} \left( \frac{2}{3} \right)^{-1} = \frac{3}{2} \cdot \frac{3}{2} = \frac{9}{4}. \end{aligned}$$

Our rule for cross-multiplication which applied to quotients of integers applies as well when we want to cross-multiply rational numbers. We state it, and prove it using only the basic properties of addition, multiplication, and inverses.

**Cross-multiplication.** Let  $a, b, c, d$  be rational numbers, and assume that  $b \neq 0$  and  $d \neq 0$ .

If  $\frac{a}{b} = \frac{c}{d}$ , then  $ad = bc$ .

If  $ad = bc$ , then  $\frac{a}{b} = \frac{c}{d}$ .

*Proof.* Assume that  $a/b = c/d$ . We can rewrite this relation in the form

$$b^{-1}a = d^{-1}c.$$

Multiply both sides by  $db$  (which is the same as  $bd$ ). We obtain

$$dbb^{-1}a = bdd^{-1}c,$$

so that

$$da = bc$$

because  $bb^{-1}a = 1a = a$ , and similarly,  $dd^{-1}c = 1c = c$ .

Conversely, assume that  $ad = bc$ . Multiply both sides by  $b^{-1}d^{-1}$ , which is equal to  $d^{-1}b^{-1}$ . We find:

$$add^{-1}b^{-1} = d^{-1}b^{-1}bc,$$

whence

$$ab^{-1} = d^{-1}c.$$

This means that  $a/b = c/d$ , as desired.

**Example.** By cross-multiplying, we have

$$\frac{3}{x - 1} = 2$$

if and only if

$$3 = 2(x - 1) = 2x - 2,$$

which is equivalent to

$$3 + 2 = 2x.$$

Thus we can solve for  $x$ , and get  $x = \frac{5}{2}$ .

**Example.** By cross-multiplying we have

$$\frac{4 + x}{\frac{1}{2}x} = 5$$

if and only if

$$4 + x = 5 \cdot \frac{1}{2}x = \frac{5x}{2}.$$

Again by cross-multiplication this is equivalent to

$$2(4 + x) = 5x,$$

or

$$8 + 2x = 5x.$$

Subtracting  $2x$  from both sides of this equation, we solve for  $x$ , and get

$$x = \frac{8}{3}.$$

**Cancellation law for multiplication.** Let  $a$  be a rational number  $\neq 0$ .

If  $ab = ac$ , then  $b = c$ .

*Proof.* Multiply both sides of the equation  $ab = ac$  by  $a^{-1}$ . We get

$$a^{-1}ab = a^{-1}ac,$$

whence  $b = c$ .

We also have a **cancellation law** similar to that for quotients of integers.

If  $a, b, c, d$  are rational numbers and  $a \neq 0, c \neq 0$ , then

$$\frac{ab}{ac} = \frac{b}{c}.$$

This can be verified, for instance, by cross-multiplication, because we have

$$abc = bac$$

(using commutativity and associativity).

Thus we can operate with fractions formed with rational numbers much as we could operate with fractions formed with integers.

**Example.** If  $a/b$  and  $c/d$  are two quotients of rational numbers (and  $b \neq 0, d \neq 0$ ), then we can put them over a “common denominator” and write

$$\frac{a}{b} = \frac{ad}{bd}, \quad \frac{c}{d} = \frac{bc}{bd}.$$

**Example.** If  $x, y, b$  are rational numbers and  $b \neq 0$ , then we can add quotients in a manner similar to the addition for quotients of integers, namely

$$\begin{aligned} \frac{x}{b} + \frac{y}{b} &= b^{-1}x + b^{-1}y \\ &= b^{-1}(x + y) && \text{by distributivity} \\ &= \frac{x + y}{b} && \text{by definition.} \end{aligned}$$

Combining this with the “common denominator” procedure of the preceding example, we find

$$\boxed{\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}.}$$

This formula is entirely analogous to the formula expressing the sum of two rational numbers.

**Example.** Show that

$$\frac{1}{x-y} + \frac{1}{x+y} = \frac{2x}{x^2 - y^2}.$$

To do this, we add the two quotients on the left by our general formula which we just derived, and get:

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

as was to be shown.

**Remark.** In the preceding example, the quotients  $1/(x-y)$  and  $1/(x+y)$  make no sense if  $x-y=0$  or  $x+y=0$ . In such instances, we assume tacitly that  $x$  and  $y$  are such that  $x-y \neq 0$  and  $x+y \neq 0$ . In the sequel we shall sometimes omit the explicit mention of such conditions if there is no danger of confusion.

**Example.** Solve for  $x$  in the equation

$$\frac{3x+1}{2x-5} = 4.$$

We cross-multiply. For  $2x-5 \neq 0$ , i.e.  $x \neq \frac{5}{2}$ , we find the equivalent equation

$$3x+1 = 4(2x-5) = 8x-20.$$

Hence

$$8x-3x = 1 - (-20) = 1 + 20 = 21.$$

This yields finally

$$5x = 21,$$

whence

$$x = \frac{21}{5}.$$

**Example.** We give an example from the physical world. Suppose that an object is moving along a straight line at constant speed. Let  $s$  denote the speed, let  $d$  denote the distance traveled by the object, and let  $t$  denote the time taken to travel the distance  $d$ . Then in physics one verifies the formula

$$d = st.$$

Of course, we must select units of time and distance before we can associate numbers with these. For instance, suppose that the distance traveled is 5 mi, and the time taken is  $\frac{1}{2}$  hr. Then the speed is

$$s = d/t = \frac{5 \text{ mi}}{\frac{1}{2} \text{ hr}} = 2 \cdot 5 \text{ mi/hr} = 10 \text{ mi/hr.}$$

**Example.** A person takes a trip and drives 8 hr, a distance of 400 mi. His average speed is 60 mph on the freeway, and 30 mph when he drives through a town. How long did the person drive through towns during his trip?

To solve this, let  $x$  be the length of time the person drives through towns. Then the length of time the person is on the freeway is  $8 - x$ . The distance driven through towns is therefore equal to  $30x$ , and the distance driven on freeways is  $60(8 - x)$ . Since the total distance driven is 400 mi, we have

$$30x + 60(8 - x) = 400.$$

This is equivalent to the equations

$$30x + 480 - 60x = 400$$

and

$$80 = 30x.$$

Thus we find

$$x = \frac{80}{30} = \frac{8}{3}.$$

Hence the person spent  $\frac{8}{3}$  hrs driving through towns.

**Example.** The radiator of a car contains 8 qt of liquid, consisting of water and 40% antifreeze. How much should be drained and replaced by antifreeze if the resultant mixture should have 90% antifreeze?

Let  $x$  be the number of quarts which must be drained. After draining this amount, we are left with  $(8 - x)$  qt of liquid, of which 40% is antifreeze. Thus we are left with

$$\frac{40}{100}(8 - x) \text{ qt}$$

of antifreeze. Since we now add  $x$  qt of antifreeze, we see that  $x$  satisfies

$$x + \frac{40}{100}(8 - x) = \frac{90}{100} \cdot 8.$$

From this we can solve for  $x$ , transforming this equation into equivalent equations as follows:

$$x + \frac{40}{100} \cdot 8 - \frac{40}{100} x = \frac{90}{100} \cdot 8,$$

which amounts to

$$\frac{60}{100} x = \frac{50}{100} \cdot 8,$$

whence

$$x = \frac{400}{60} = \frac{20}{3}.$$

This is a correct answer, but if you insist on putting the fraction in lowest form, then we can say that  $6\frac{2}{3}$  qt should be replaced by antifreeze.

**Remark.** The above examples, and the exercises, can also be worked using two unknowns. Cf. the end of Chapter 2, §1.

## EXERCISES

1. Solve for  $x$  in the following equations.

a)  $\frac{2x - 1}{3x + 2} = 7$

b)  $\frac{2 - 4x}{x + 1} = \frac{3}{4}$

c)  $\frac{x}{x + 5} = \frac{5}{7}$

d)  $2x + 5 = \frac{3x - 2}{7}$

e)  $\frac{1 - 2x}{3x + 4} = -3$

f)  $\frac{-2 - 5x}{-3x - 4} = \frac{4}{-3}$

g)  $\frac{-2 - 7x}{4} + 1 = \frac{1 - x}{5}$

h)  $\frac{3x + 1}{4 - 2x} + \frac{7}{3} = 0$

i)  $\frac{-2 - 4x}{3} = \frac{x - 1}{4} + 5$

2. Prove the following relations. It is assumed that all values of  $x$  and  $y$  which occur are such that the denominators in the indicated fractions are not equal to 0.

$$\text{a) } \frac{1}{x+y} - \frac{1}{x-y} = \frac{-2y}{x^2 - y^2} \quad \text{b) } \frac{x^3 - 1}{x - 1} = 1 + x + x^2$$

$$\text{c) } \frac{x^4 - 1}{x - 1} = 1 + x + x^2 + x^3$$

$$\text{d) } \frac{x^n - 1}{x - 1} = x^{n-1} + x^{n-2} + \cdots + x + 1. \quad [\text{Hint: Cross-multiply and cancel as much as possible.}]$$

3. Prove the following relations.

$$\text{a) } \frac{1}{2x+y} + \frac{1}{2x-y} = \frac{4x}{4x^2 - y^2}$$

$$\text{b) } \frac{2x}{x+5} - \frac{3x+1}{2x+1} = \frac{x^2 - 14x - 5}{2x^2 + 11x + 5}$$

$$\text{c) } \frac{1}{x+3y} + \frac{1}{x-3y} = \frac{2x}{x^2 - 9y^2}$$

$$\text{d) } \frac{1}{3x-2y} + \frac{x}{x+y} = \frac{x+y+3x^2-2xy}{3x^2+xy-2y^2}$$

For more exercises of this type, see Chapter 13, §2

4. Prove the following relations.

$$\text{a) } \frac{x^3 - y^3}{x - y} = x^2 + xy + y^2$$

$$\text{b) } \frac{x^4 - y^4}{x - y} = x^3 + x^2y + xy^2 + y^3$$

c) Let

$$x = \frac{1 - t^2}{1 + t^2} \quad \text{and} \quad y = \frac{2t}{1 + t^2}.$$

Show that  $x^2 + y^2 = 1$ .

5. Prove the following relations.

$$\text{a) } \frac{x^3 + 1}{x + 1} = x^2 - x + 1$$

b)  $\frac{x^5 + 1}{x + 1} = x^4 - x^3 + x^2 - x + 1$

c) If  $n$  is an odd integer, prove that

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

[Hint: Cross-multiply.]

6. Assume that a particle moving with uniform speed on a straight line travels a distance of  $\frac{5}{4}$  ft at a speed of  $\frac{2}{3}$  ft/sec. What time did it take the particle to do that?
7. If a solid has uniform density  $d$ , occupies a volume  $v$ , and has mass  $m$ , then we have the formula

$$m = vd.$$

Find the density if

- a)  $m = \frac{3}{10}$  lb and  $v = \frac{2}{3}$  in $^3$ ,      b)  $m = 6$  lb and  $v = \frac{4}{3}$  in $^3$ .  
 c) Find the volume if the mass is 15 lb and the density is  $\frac{2}{3}$  lb/in $^3$ .

8. Let  $F$  denote temperature in degrees Fahrenheit, and  $C$  the temperature in degrees centigrade. Then  $F$  and  $C$  are related by the formula

$$C = \frac{5}{9}(F - 32).$$

Find  $C$  when  $F$  is

- a) 32,      b) 50,      c) 99,      d) 100,      e) -40.

9. Let  $F$  and  $C$  be as in Exercise 8. Find  $F$  when  $C$  is:

- a) 0,      b) -10,      c) -40,      d) 37,      e) 40,      f) 100.

10. In electricity theory, one denotes the current by  $I$ , the resistance by  $R$ , and the voltage by  $E$ . These are related by the formula

$$E = IR$$

(with appropriate units). Find the resistance when the voltage and current are:

- a)  $E = 10$ ,  $I = 3$ ;      b)  $E = 220$ ,  $I = 10$ .

11. A solution contains 35% alcohol and 65% water. If you start with 12 cm $^3$  (cubic centimeters) of solution, how much water must be added to make the percentage of alcohol equal to

- a) 20%?      b) 10%?      c) 5%?

12. A plane travels 3,000 mi in 4 hr. When the wind is favorable, the plane averages 900 mph. When the wind is unfavorable, the plane averages 500 mph. During how many hours was the wind favorable?
13. Tickets for a performance sell at \$5.00 and \$2.00. The total amount collected was \$4,100, and there are 1,300 tickets in all. How many tickets of each price were sold?
14. A salt solution contains 10% salt and weighs 80 g. How much pure water must be added so that the percentage of salt drops to  
a) 4%?                    b) 6%?                    c) 8%?
15. How many quarts of water must you add to 6 qt of pure alcohol to get a mixture containing  
a) 25% alcohol?        b) 20% alcohol?        c) 15% alcohol?
16. A boat travels a distance of 500 mi, along two rivers, for 50 hr. The current goes in the same direction as the boat along one river, and then the boat averages 20 mph. The current goes in the opposite direction along the other river, and then the boat averages 8 mph. During how many hours was the boat on the first river?
17. How much water must evaporate from a salt solution weighing 2 lb and containing 25% salt, if the remaining mixture must contain  
a) 40% salt?                    b) 60% salt?
18. The radiator of a car can contain 10 gal of liquid. If it is half full with a mixture having 60% antifreeze and 40% water, how much more water must be added so that the resulting mixture has only  
a) 40% antifreeze?                    b) 10% antifreeze?  
Will it fit in the radiator?



# 2 *Linear Equations*

## §1. EQUATIONS IN TWO UNKNOWNNS

Suppose that we are given two equations like

$$\begin{aligned} 1) \quad & 2x + y = 1, \\ 2) \quad & 3x - 2y = 4. \end{aligned}$$

We wish to solve these equations for  $x$  and  $y$ . We follow what is known as the elimination method. We try to get rid of  $x$ , say, so as to obtain only one equation in  $y$ . We observe that  $x$  is multiplied by 2 in the first equation and by 3 in the second. We want to multiply each one of these equations by a suitable number so that the coefficients of  $x$  become the same. Thus we multiply the first equation by 3 and the second by 2. We obtain

$$\begin{aligned} 6x + 3y &= 3, \\ 6x - 4y &= 8. \end{aligned}$$

If we now subtract the second equation from the first, i.e. subtract each side of the second equation from the corresponding side of the first, we see that the  $6x$  cancels, and we find:

$$3y - (-4y) = 3 - 8,$$

whence

$$3y + 4y = 7y = -5.$$

This yields

$$y = \frac{-5}{7}.$$

We can then solve for  $x$ , using (1), which gives  $2x = 1 - y$ . Thus

$$2x = 1 - \frac{-5}{7} = \frac{7 + 5}{7} = \frac{12}{7}.$$

Hence

$$x = \frac{12}{2 \cdot 7} = \frac{12}{14}.$$

Our answer is therefore:

$$y = \frac{-5}{7} \quad \text{and} \quad x = \frac{12}{14}.$$

If we want  $x$  in lowest form, we can always write  $x = \frac{6}{7}$ , but  $\frac{12}{14}$  is quite correct.

As a variation, we could also have eliminated  $y$  first. Thus we multiply the first equation by 2, leave the second unchanged, and add the equations. We get:

$$\begin{aligned} 4x + 2y &= 2, \\ 3x - 2y &= 4. \end{aligned}$$

Adding yields

$$4x + 3x = 6.$$

Thus  $7x = 6$  and  $x = \frac{6}{7}$ , which is of course the same answer that we found above. We could then solve for  $y$  using the first equation, namely,

$$y = 1 - 2x,$$

so that

$$y = 1 - \frac{12}{7} = \frac{7 - 12}{7} = \frac{-5}{7}.$$

It may happen that a system of linear equations has no solutions. For instance the system

$$(3) \quad \begin{aligned} 2x - y &= 5, \\ 2x - y &= 7 \end{aligned}$$

obviously has no solution. The system

$$(4) \quad \begin{aligned} 2x - y &= 5, \\ 6x - 3y &= 7 \end{aligned}$$

has no solution either. Indeed, any solution of  $6x - 3y = 7$  is also a solution of

$$2x - y = \frac{7}{3},$$

divide the equation by 3), and it is again obvious that no simultaneous solution exists for the system of equations (4).

We do not want to overemphasize here the theory determining precisely the cases when a solution exists and when it does not. It "usually" exists, unless one has a case essentially like the examples above. Our purposes here are mainly to put you at ease with two simple equations in two unknowns, so that you have some simple approach to them. We don't intend to overburden you or give you any worries about them. On the other hand, you may wish to do Exercises 9 and 10 to get the general criterion indicating when a solution exists. These exercises make precise our meaning of "usually".

When you learn about coordinates, then you will see that the simultaneous equations we have been considering represent straight lines, and that finding their simultaneous solution gives the coordinates of the point of intersection of these lines. If you wish, you may look up coordinates right away, and the first section of Chapter 12 to see about this.

One final remark. Observe that our elimination procedure actually proves that if  $x, y$  are numbers satisfying the simultaneous equations, then they must have the value obtained by the method indicated. Conversely these values for  $x, y$  actually are solutions of the equations. This can be checked each time explicitly. To prove it in general is easy but requires setting up convenient notation and using general letters for the coefficients of the equation. See Exercise 11. Here we don't want to get bogged down in abstraction. Our purpose in this section was simply to teach you a simple and efficient way of finding the solutions of a simple system of equations.

Simultaneous equations like the above can be used to solve problems which we gave in the context of one variable at the end of Chapter 1. We give an example of this.

**Example.** A person takes a trip and drives 8 hr, a distance of 400 mi. His average speed is 60 mph on the freeway, and 30 mph when he drives through a town. How long did the person drive through towns during his trip?

To solve this, let  $x$  be the length of time driven on freeways, and let  $y$  be the length of time driven through towns. Then

$$x + y = 8.$$

This gives us a first equation. Furthermore, the distance driven on freeways is equal to  $60x$ , and the distance driven through towns is equal to  $30y$ . Hence we get a second equation

$$60x + 30y = 400.$$

We can now solve our pair of equations, by multiplying the first by 60 and subtracting the second. We get

$$60y - 30y = 480 - 400,$$

or more simply,

$$30y = 80.$$

Therefore

$$y = \frac{80}{30} = \frac{8}{3}$$

is our numerical answer, and the person drove  $\frac{8}{3}$  hr through towns. This is of course the same answer that we found when working with only one variable.

You may now wish to work out the exercises at the end of Chapter 1, §6 by means of two unknowns, which may be easier to handle the problems.

## EXERCISES

Solve the following systems of equations for  $x$  and  $y$ .

- |                   |                    |
|-------------------|--------------------|
| 1. $2x - y = 3$   | 2. $-4x + 7y = -1$ |
| $x + y = 2$       | $x - 2y = -4$      |
| 3. $3x + 4y = -2$ | 4. $-3x + 2y = -1$ |
| $-2x - 3y = 1$    | $x - y = 2$        |
| 5. $-3x + y = 0$  | 6. $3x + 7y = 0$   |
| $x - y = 1$       | $x - y = 0$        |
| 7. $7x - y = 2$   | 8. $-4x - 7y = 5$  |
| $2x + 2y = 4$     | $2x + y = 6$       |

9. Let  $a, b, c, d$  be numbers such that  $ad - bc \neq 0$ . Solve the following systems of equations for  $x$  and  $y$  in terms of  $a, b, c, d$ .
- |                   |                  |
|-------------------|------------------|
| a) $ax + by = 1$  | b) $ax + by = 3$ |
| $cx + dy = 2$     | $cx + dy = -4$   |
| c) $ax + by = -2$ | d) $ax + by = 5$ |
| $cx + dy = 3$     | $cx + dy = 7$    |
10. Making the same assumptions as in Exercise 9, show that the solution of the system

$$\begin{aligned} ax + by &= 0, \\ cx + dy &= 0 \end{aligned}$$

must be  $x = 0$  and  $y = 0$ .

11. Let  $a, b, c, d, u, v$  be numbers and assume that  $ad - bc \neq 0$ . Solve the following system of equations for  $x$  and  $y$  in terms of  $a, b, c, d, u, v$ :

$$\begin{aligned} ax + by &= u, \\ cx + dy &= v. \end{aligned}$$

Verify that the answer you get is actually a solution.

## §2. EQUATIONS IN THREE UNKNOWNS

We now want to solve a system of equations like

$$\begin{aligned} 1) \quad 3x + 2y + 4z &= 1, \\ -x + y + 2z &= 2, \\ x - 3y + z &= -1 \end{aligned}$$

for  $x, y, z$ . We follow the same pattern as before, eliminating successively  $x, y$ , and then solving for  $z$ . We choose the order of elimination so as to make it easier on ourselves. Thus adding the second and third equations already gets rid of  $x$ , so we do this, and get

$$y - 3y + 3z = 2 - 1$$

or

$$(2) \quad -2y + 3z = 1.$$

We go back to (1), and eliminate  $x$  from the first two equations. We multiply the second by 3 and add it to the first. This yields

$$2y + 3y + 4z + 6z = 1 + 6$$

or

$$(3) \quad 5y + 10z = 7.$$

Then equations (2) and (3) form a pair of equations in two unknowns which can be solved as in the first section of this chapter. We multiply (2) by 5, we multiply (3) by 2, and add. This gets rid of  $y$ , and we obtain

$$15z + 20z = 5 + 14$$

which yields

$$35z = 19,$$

whence

$$z = \frac{19}{35}.$$

Having found the value for  $z$ , we can go back to (2) or (3) to find the value for  $y$ . Suppose we use (2). We get

$$\begin{aligned} 2y &= 3z - 1 = 3 \cdot \frac{19}{35} - 1 \\ &= \frac{57 - 35}{35} \\ &= \frac{22}{35}. \end{aligned}$$

Hence dividing by 2, we find the value for  $y$ , namely

$$y = \frac{11}{35}.$$

Finally we can solve for  $x$  using any one of the first three equations in (1). Say we use the third equation. We have

$$\begin{aligned} x &= -1 + 3y - z \\ &= -\frac{35}{35} + \frac{33}{35} - \frac{19}{35} \\ &= -\frac{91}{35}. \end{aligned}$$

Thus the solution to our problem is:

$$x = \frac{-21}{35},$$

$$y = \frac{11}{35},$$

$$z = \frac{19}{35}.$$

)

### EXERCISES

Solve the following equations for  $x, y, z$ .

- |                       |                       |
|-----------------------|-----------------------|
| 1. $2x - 3y + z = 0$  | 2. $2x - y + z = 1$   |
| $x + y + z = 1$       | $4x + y + z = 2$      |
| $x - 2y - 4z = 2$     | $x - y - 2z = 0$      |
| 3. $x + 4y - 4z = 1$  | 4. $x + y + z = 0$    |
| $x + 2y + z = 2$      | $x - y + z = 0$       |
| $4x - 3y - 2z = 1$    | $2x - y - z = 0$      |
| 5. $5x + 3y - z = 0$  | 6. $2x + 2y - 3z = 0$ |
| $x + 2y + 2z = 1$     | $x - 3y + z = 3$      |
| $x - 2y - 2z = 0$     | $2x + y - 4z = 0$     |
| 7. $4x - 2y + 5z = 1$ | 8. $x + y + z = 0$    |
| $x + y + z = 0$       | $x - y - z = 1$       |
| $-x + y - 2z = 2$     | $x + y - z = 1$       |

In the next exercises, you will find it easiest to clear denominators before solving.

- |  |   |
|--|---|
| 9. $\frac{1}{2}x + y - \frac{3}{4}z = 1$ | 10. $\frac{1}{2}x - y + z = 1$            |
| $\frac{2}{3}x - \frac{1}{3}y + z = 2$    | $x + \frac{1}{3}y - \frac{2}{3}z = 2$     |
| $x - \frac{1}{5}y + 2z = 1$              | $x + y - z = 3$                           |
| 11. $\frac{3}{4}x - y + z = 1$           | 12. $\frac{1}{2}x - \frac{2}{3}y + z = 1$ |
| $x - \frac{1}{2}y + z = 0$               | $x - \frac{1}{5}y + z = 0$                |
| $x + y - \frac{1}{3}z = 1$               | $2x - \frac{1}{3}y + \frac{2}{5}z = 1$    |



# 3 Real Numbers

## §1. ADDITION AND MULTIPLICATION

The integers and rational numbers are part of a larger system of numbers. As we know, the integers and rational numbers correspond to some points on the line. The real numbers are those numbers which correspond to all points on the line. Another way of describing them is to say that they consist of all numbers which have a decimal expansion, possibly infinite. For instance

$$9.123145 \dots$$

is a real number. It is a rather long and tedious process to develop the theory of the real numbers systematically. Hence we shall first summarize some of the algebraic properties which they satisfy, and then, as we go along, if the need arises, we shall mention other properties. Unless otherwise specified, “number” will mean “real number”.

The sum and product of two numbers are numbers, and they satisfy the following properties, similar to those of rational numbers.

**Properties of addition.** Addition is commutative and associative, meaning that for all real numbers  $a, b, c$  we have

$$a + b = b + a \quad \text{and} \quad a + (b + c) = (a + b) + c.$$

Furthermore, we have

$$0 + a = a.$$

To each real number  $a$  there is associated a number denoted by  $-a$  such that

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

As with integers, or rational numbers, we represent  $a$  and  $-a$  on opposite sides of 0 on the line.

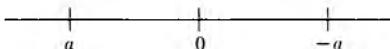


Fig. 3-1

The number  $-a$  is called the **additive inverse** of  $a$ , as before. We also read it as **minus  $a$** . If  $b$  is a number such that  $a + b = 0$ , then we must have  $b = -a$ , as we see by adding  $-a$  to both sides of the equation  $a + b = 0$ . This is called the **uniqueness of the additive inverse**.

**Properties of multiplication.** *Multiplication is commutative and associative, meaning that for all real numbers  $a, b, c$  we have*

$$ab = ba \quad \text{and} \quad a(bc) = (ab)c.$$

*Furthermore, we have*

$$1a = a \quad \text{and} \quad 0a = 0.$$

*Multiplication is distributive with respect to addition, meaning that*

$$a(b + c) = ab + ac \quad \text{and} \quad (b + c)a = ba + ca.$$

So far, these properties are the same as those satisfied by the integers and rational numbers. *In particular, further properties which were proved using only these basic ones are now valid for the real numbers.* For instance, we recall the important formulas:

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

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

$$(a + b)(a - b) = a^2 - b^2.$$

These are true if  $a, b$  are real numbers because their proofs used only commutativity and associativity.

**Existence of the multiplicative inverse.** *If  $a$  is a real number  $\neq 0$ , then there exists a real number denoted by  $a^{-1}$  such that*

$$a^{-1}a = aa^{-1} = 1.$$

As with rational numbers, this number  $a^{-1}$  is called the **multiplicative inverse of  $a$** . Instead of writing  $a^{-1}$  we write  $1/a$ , and

we write  $a/b$  instead of  $b^{-1}a$  or  $ab^{-1}$ .

The proofs of properties concerning inverses before depended only on the basic ones we have mentioned so far, and are thus applicable to the real numbers. Thus we have cross-multiplication, cancellation rules, etc. We

have the **uniqueness of the multiplicative inverse**. Namely, if

$$ab = 1,$$

then multiplying both sides by  $a^{-1}$  shows that  $b = a^{-1}$ .

### EXERCISES

1. Let  $E$  be an abbreviation for even, and let  $I$  be an abbreviation for odd. We know that:

$$\begin{aligned} E + E &= E, \\ E + I &= I + E = I, \\ I + I &= E, \\ EE &= E, \\ II &= I, \\ IE = EI &= E. \end{aligned}$$

- a) Show that addition for  $E$  and  $I$  is associative and commutative. Show that  $E$  plays the role of a zero element for addition. What is the additive inverse of  $E$ ? What is the additive inverse of  $I$ ?
- b) Show that multiplication for  $E$  and  $I$  is commutative and associative. Which of  $E$  or  $I$  behaves like 1? Which behaves like 0 for multiplication? Show that multiplication is distributive with respect to addition.

**Remark.** The system consisting of  $E$  and  $I$  gives an example of a system with only two objects satisfying the basic properties of addition and multiplication. Thus real numbers are not the only system to satisfy these properties.

## §2. REAL NUMBERS: POSITIVITY

We have the positive numbers, represented geometrically on the straight line by those numbers unequal to 0 and lying to the right of 0. If  $a$  is a positive number, we write  $a > 0$ . We shall list the basic properties of positivity from which others will be proved.

**POS 1.** *If  $a, b$  are positive, so are the product  $ab$  and the sum  $a + b$ .*

**POS 2.** *If  $a$  is a real number, then either  $a$  is positive, or  $a = 0$ , or  $-a$  is positive, and these possibilities are mutually exclusive.*

If a number is not positive and not 0, then we say that this number is **negative**. By **POS 2**, if  $a$  is negative, then  $-a$  is positive.

We know already that the number 1 is positive, but this could be proved from our two properties, and the basic rules for addition and multiplication. It may interest you to see the proof, which runs as follows and is very simple. By **POS 2**, we know that either 1 or  $-1$  is positive. If 1 is not positive, then  $-1$  is positive. By **POS 1**, it must follow that  $(-1)(-1)$  is positive. But this product is equal to 1. Consequently, it must be 1, which is positive and not  $-1$ .

Using property **POS 1**, we can now conclude that  $1 + 1 = 2$  is positive, that  $2 + 1 = 3$  is positive, and so forth. Thus our calling

$$1, 2, 3, \dots$$

the positive integers is compatible with our two rules **POS 1** and **POS 2**.

Other basic properties of positivity are easily proved from the two basic ones, and are left as exercises (Exercises 1, 2, 3), namely:

*If  $a$  is positive and  $b$  is negative, then  $ab$  is negative.*

*If  $a$  is negative and  $b$  is negative, then  $ab$  is positive.*

*If  $a$  is positive, then  $1/a$  is positive.*

*If  $a$  is negative, then  $1/a$  is negative.*

One of the properties of real numbers which we assume without proof is that every positive real number has a square root. This means:

*If  $a > 0$ , then there exists a number  $b$  such that  $b^2 = a$ .*

Because of this, and Theorem 4, §5 of Chapter 1, we now see that a number whose square is 2 is *irrational*, but exists as a real number.

It is a reasonable question to ask right away how many numbers there are whose squares are equal to a given positive number. For instance, what are all the real numbers  $x$  such that  $x^2 = 2$ ?

This is easily answered. *There are precisely two such numbers.* One is positive, and the other is negative. Let us prove this. Let  $b^2 = 2$ , and let  $x$  be any real number such that  $x^2 = 2$  also. We have

$$x^2 - b^2 = 0.$$

However, the left-hand side factors, and we find

$$(x + b)(x - b) = 0.$$

Hence we must have

$$x + b = 0 \quad \text{or} \quad x - b = 0$$

so that

$$x = -b \quad \text{or} \quad x = b.$$

On the other hand, the square of  $-b$  is equal to 2, because

$$(-b)^2 = (-b)(-b) = b^2 = 2.$$

Thus we have proved our assertion.

Of the two numbers whose square is 2, we conclude from POS 2 that precisely one of them is positive. We now adopt a convention, in force throughout all of mathematics. **We agree to call the square root of 2 only the positive number  $b$  whose square is 2.** This positive number will be denoted by

$$\sqrt{2}.$$

Therefore the two numbers whose square is 2 are

$$\sqrt{2} \quad \text{and} \quad -\sqrt{2},$$

and we have

$$\sqrt{2} > 0.$$

Exactly the same arguments show that given any positive number  $a$ , there exist precisely two numbers whose square is  $a$ . If  $b$  is one of them, then  $-b$  is the other. Just replace 2 by  $a$  in the preceding arguments. Again by convention, we let

$$\sqrt{a}$$

denote the *unique positive number whose square is  $a$ .* The other number whose square is  $a$  is therefore  $-\sqrt{a}$ . We shall express this also by saying

that the solutions of the equation  $x^2 = a$  are

$$x = \pm\sqrt{a}.$$

We read this as “ $x$  equals plus or minus square root of  $a$ ”.

Another way of putting this is:

*If  $x, y$  are numbers such that  $x^2 = y^2$ , then  $x = y$  or  $x = -y$ .*

But we cannot conclude that  $x = y$ . Furthermore, for any number  $x$ , the number

$$\sqrt{x^2}$$

is  $\geq 0$ . Thus

$$\sqrt{(-3)^2} = \sqrt{9} = 3.$$

There is a special notation for this. We call  $\sqrt{x^2}$  the **absolute value** of  $x$ , and denote it by

$$|x| = \sqrt{x^2}.$$

Thus we have

$$|-3| = 3 \quad \text{and also} \quad |-5| = 5.$$

Of course, for any positive number  $a$ , we have

$$|a| = a.$$

Thus

$$|3| = 3 \quad \text{and} \quad |5| = 5.$$

We won't work too much with absolute values in this book, and we do not want to overemphasize them here. Occasionally, we need the notion, and we need to know that the absolute value of  $-3$  is  $3$ . In that spirit, we give an example showing how to solve an equation with an absolute value in it, just to drive the definition home, but not to belabor the point.

**Example.** Find all values of  $x$  such that  $|x + 5| = 2$ .

To do this, we note that  $|x + 5| = 2$  if and only if  $x + 5 = 2$  or  $x + 5 = -2$ . Thus we have two possibilities, namely

$$x = 2 - 5 = -3 \quad \text{and} \quad x = -5 - 2 = -7.$$

This solves our problem.

Observe that

$$\frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}.$$

This is because

$$2 = \sqrt{2} \sqrt{2},$$

and so our assertion is true because of cross-multiplication. It is a tradition in elementary schools to transform a quotient like

$$\frac{1}{\sqrt{2}}$$

into another one in which the square root sign does not appear in the denominator. As far as we are concerned, doing this is not particularly useful in general. It may be useful in special cases, but neither more nor less than other manipulations with quotients, to be determined *ad hoc* as the need arises. Actually, in many cases it is useful to have the square root in the denominator. We shall give two examples of how to transform an expression involving square roots in the numerator or denominator. The manipulations of these examples will be based on the old rule

$$(x + y)(x - y) = x^2 - y^2.$$

**Example.** Consider a quotient

$$\frac{3}{2 + \sqrt{5}}.$$

We wish to express it as a quotient where the denominator is a rational number. We multiply both numerator and denominator by

$$2 - \sqrt{5}.$$

This yields

$$\frac{3}{(2 + \sqrt{5})} \cdot \frac{(2 - \sqrt{5})}{(2 - \sqrt{5})} = \frac{6 - 3\sqrt{5}}{2^2 - (\sqrt{5})^2} = \frac{6 - 3\sqrt{5}}{-1} = -6 + 3\sqrt{5}.$$

**Example.** This example has the same notation as an actual case which arises in more advanced courses of calculus. Let  $x$  and  $h$  be numbers such that  $x$  and  $x + h$  are positive. We wish to write the quotient

$$\frac{\sqrt{x+h} - \sqrt{x}}{h}$$

in such a way that the square root signs occur only in the denominator. We multiply numerator and denominator by  $(\sqrt{x+h} + \sqrt{x})$ . We obtain:

$$\begin{aligned}\frac{(\sqrt{x+h} - \sqrt{x})}{h} \cdot \frac{(\sqrt{x+h} + \sqrt{x})}{(\sqrt{x+h} + \sqrt{x})} &= \frac{(\sqrt{x+h})^2 - (\sqrt{x})^2}{h(\sqrt{x+h} + \sqrt{x})} \\ &= \frac{x+h-x}{h(\sqrt{x+h} + \sqrt{x})} \\ &= \frac{h}{h(\sqrt{x+h} + \sqrt{x})} \\ &= \frac{1}{\sqrt{x+h} + \sqrt{x}}.\end{aligned}$$

Thus we find finally:

$$\frac{\sqrt{x+h} - \sqrt{x}}{h} = \frac{1}{\sqrt{x+h} + \sqrt{x}}.$$

In the first example, the procedure we have followed is called **rationalizing the denominator**. In the second example, the procedure is called **rationalizing the numerator**. In a quotient involving square roots, rationalizing the numerator means that we transform this quotient into another one, equal to the first, but such that no square root sign appears in the numerator. Similarly, rationalizing the denominator means that we transform this quotient into another one, equal to the first, but such that no square root appears in the denominator. Both procedures are useful in practice.

Square roots will be used when we discuss the Pythagoras Theorem, and the distance between points in Chapter 8, §2. You could very well look up these sections right now to see these applications, especially the section on distance.

## EXERCISES

1. Prove:

- a) If  $a$  is a real number, then  $a^2$  is positive.
- b) If  $a$  is positive and  $b$  is negative, then  $ab$  is negative.
- c) If  $a$  is negative and  $b$  is negative, then  $ab$  is positive.

2. Prove: If  $a$  is positive, then  $a^{-1}$  is positive.
3. Prove: If  $a$  is negative, then  $a^{-1}$  is negative.
4. Prove: If  $a, b$  are positive numbers, then

$$\sqrt{\frac{a}{b}} = \frac{\sqrt{a}}{\sqrt{b}}.$$

5. Prove that

$$\frac{1}{1 - \sqrt{2}} = -(1 + \sqrt{2}).$$

6. Prove that the multiplicative inverse of  $2 + \sqrt{3}$  can be expressed in the form  $c + d\sqrt{3}$ , where  $c, d$  are rational numbers.
7. Prove that the multiplicative inverse of  $3 + \sqrt{5}$  can be expressed in the form  $c + d\sqrt{5}$ , where  $c, d$  are rational numbers.
8. Let  $a, b$  be rational numbers. Prove that the multiplicative inverse of  $a + b\sqrt{2}$  can be expressed in the form  $c + d\sqrt{2}$ , where  $c, d$  are rational numbers.
9. Same question as in Exercise 8, but replace  $\sqrt{2}$  by  $\sqrt{3}$ .
10. Let  $x, y, z, w$  be rational numbers. Show that a product

$$(x + y\sqrt{5})(z + w\sqrt{5})$$

can be expressed in the form  $c + d\sqrt{5}$ , where  $c, d$  are rational numbers.

11. Generalize Exercise 10, replacing  $\sqrt{5}$  by  $\sqrt{a}$  for any positive integer  $a$ .
12. Rationalize the numerator in the following expressions.

a)  $\frac{\sqrt{2x+3}+1}{4}$

b)  $\frac{\sqrt{1+x}-3}{2}$

c)  $\frac{\sqrt{x-h}-\sqrt{x}}{h}$

d)  $\frac{\sqrt{x-h}+\sqrt{x}}{h}$

e)  $\frac{\sqrt{x+h}+\sqrt{x}}{h}$

f)  $\frac{\sqrt{x+2h}-\sqrt{x}}{h}$

13. Find all possible numbers  $x$  such that

a) $ x - 1  = 2$ ,	b) $ x  = 5$ ,	c) $ x - 3  = 4$ ,
d) $ x + 1  = 6$ ,	e) $ x + 4  = 3$ ,	f) $ x - 2  = 1$ .

14. Find all possible numbers  $x$  such that

a) $ 2x - 1  = 3$ ,	b) $ 3x + 1  = 2$ ,	c) $ 2x + 1  = 4$ ,
d) $ 3x - 1  = 1$ ,	e) $ 4x - 5  = 6$ .	

15. Rationalize the numerator in the following expressions.

a)  $\frac{\sqrt{x} + \sqrt{y}}{\sqrt{x} - \sqrt{y}}$

b)  $\frac{\sqrt{x+y} - \sqrt{y}}{\sqrt{x+y} + \sqrt{y}}$

c)  $\frac{\sqrt{x+1} + \sqrt{x-1}}{\sqrt{x+1} - \sqrt{x-1}}$

d)  $\frac{\sqrt{x-3} + \sqrt{x}}{\sqrt{x-3} - \sqrt{x}}$

e)  $\frac{\sqrt{x+y} - 1}{3 + \sqrt{x+y}}$

f)  $\frac{\sqrt{x+y} + x}{\sqrt{x+y}}$

16. Rationalize the denominator in each one of the cases of Exercise 15.

17. Prove that there is no real number  $x$  such that

$$\sqrt{x-1} = 3 + \sqrt{x}.$$

[Hint: Start by squaring both sides.]

18. If  $\sqrt{x-1} = 3 - \sqrt{x}$ , prove that  $x = \frac{25}{9}$ .

19. Determine in each of the following cases whether there exists a real number  $x$  satisfying the indicated relation, and if there is, determine this number.

a)  $\sqrt{x-2} = 3 + 2\sqrt{x}$

b)  $\sqrt{x-2} = 3 - 2\sqrt{x}$

c)  $\sqrt{x+3} = 1 + \sqrt{x}$

d)  $\sqrt{x+3} = 1 - \sqrt{x}$

e)  $\sqrt{x-4} = 3 + \sqrt{x}$

f)  $\sqrt{x-4} = 3 - \sqrt{x}$

20. If  $a, b$  are two numbers, prove that  $|a - b| = |b - a|$ .

### §3. POWERS AND ROOTS

Let  $n$  be a positive integer and let  $a$  be a real number. As before, we let

$$a^n$$

be the product of  $a$  with itself  $n$  times. The rule

$$a^{m+n} = a^m a^n$$

holds as before, if  $m, n$  are positive integers.

Let  $a$  be a positive number and let  $n$  be a positive integer. As part of the properties of real numbers, we assume, but do not prove, that *there exists a unique positive real number  $r$  such that*

$$r^n = a.$$

This number  $r$  is called the  **$n$ -th root** of  $a$ , and is denoted by

$$a^{1/n} \quad \text{or} \quad \sqrt[n]{a}.$$

The  $n$ -th root generalizes the existence and uniqueness of the square root discussed in the preceding section.

**Theorem 1.** *Let  $a, b$  be positive real numbers. Then*

$$(ab)^{1/n} = a^{1/n}b^{1/n}.$$

*Proof.* Let  $r = a^{1/n}$  and  $s = b^{1/n}$ . This means that  $r^n = a$  and  $s^n = b$ . Therefore

$$(rs)^n = r^n s^n = ab.$$

This means that  $rs$  is the  $n$ -th root of  $ab$ , and proves our theorem.

The  $n$ -th root can be further generalized to fractional powers. Let  $a$  be a positive real number. We shall assume without proof the following property of numbers.

**Fractional powers.** *Let  $a$  be a positive number. To each rational number  $x$  we can associate a positive number denoted by  $a^x$ , which is the  $n$ -th power of  $a$  when  $x$  is a positive integer  $n$ , the  $n$ -th root of  $a$  when  $x = 1/n$ , and satisfying the following conditions:*

**POW 1.** *For all rational numbers  $x, y$  we have*

$$a^{x+y} = a^x a^y.$$

**POW 2.** *For all rational numbers  $x, y$  we have*

$$(a^x)^y = a^{xy}.$$

**POW 3.** *If  $a, b$  are positive, then*

$$(ab)^x = a^x b^x.$$

We shall now derive consequences from these conditions.

First, we compute  $a^0$ . Let  $b = a^0$ . We have:

$$a = a^1 = a^{0+1} = a^0 a^1 = a^0 a.$$

Thus  $a = ba$ . Multiply both sides by  $a^{-1}$ . We get

$$1 = aa^{-1} = baa^{-1} = b.$$

Thus we find the important formula

$$a^0 = 1.$$

Next we shall see what a negative power is like. Let  $x$  be a positive rational number. Then

$$1 = a^0 = a^{x+(-x)} = a^x a^{-x}.$$

Thus the product of  $a^x$  and  $a^{-x}$  is 1. This means that  $a^{-x}$  is the multiplicative inverse of  $a^x$ , and thus that

$$a^{-x} = \frac{1}{a^x}.$$

This also justifies our notation, writing  $a^{-1}$  for the multiplicative inverse of  $a$ .

**Example.** Let  $x = 3$ . Then

$$a^{-3} = (a^3)^{-1} = \frac{1}{a^3}.$$

Similarly,

$$2^{-4} = (2^4)^{-1} = \frac{1}{2^4}.$$

Thus very roughly speaking, taking negative powers corresponds to taking a quotient.

Next, let  $x = m/n$  be a positive rational number, expressed as a quotient of positive integers  $m, n$ . Then by **POW 2**, we find:

$$a^{m/n} = (a^m)^{1/n} = (a^{1/n})^m.$$

The fractional power can be decomposed into an ordinary power with an integer and an  $n$ -th root.

**Example.** We have

$$\begin{aligned} 8^{2/3} &= (8^{1/3})^2 \\ &= 2^2 = 4. \end{aligned}$$

**Example.** We have

$$\begin{aligned} (\sqrt{2})^{3/4} &= (\sqrt{2}^{1/4})^3 \\ &= (2^{1/8})^3 = 2^{3/8}. \end{aligned}$$

**Example.** We have

$$\begin{aligned} (\sqrt{2})^3 &= \sqrt{2} \sqrt{2} \sqrt{2} \\ &= 2\sqrt{2} = 2^{3/2}. \end{aligned}$$

**Example.** We have

$$\left(\frac{25}{9}\right)^{3/2} = \frac{25^{3/2}}{9^{3/2}} = \frac{125}{27}.$$

We would also like to take powers with irrational exponents, i.e. we would like to define numbers like

$$2^{\sqrt{2}}.$$

This is much more difficult, but it can be done in such a way that the two conditions **POW 1** and **POW 2** are satisfied. We shall not need this, and therefore shall postpone a systematic development for a more advanced course, although we shall make some further comments on the situation in the chapter on functions. However, we are led to make a final comment concerning the real numbers, as distinguished from the rational numbers.

Note that the properties of addition, multiplication, and positivity hold for rational numbers. What distinguishes the real numbers from the rationals is the existence of more numbers, like square roots,  $n$ -th roots, general exponents, etc. To make this “etc.” precise is a more complicated undertaking. We can ask: Is there a neat way (besides stating that the real numbers consist of all infinite decimals) of expressing a property of the reals which guarantees that any number which we want to exist intuitively can be proved to exist using just this property? The answer is yes, but belongs to a much more advanced course. Thus throughout this course and throughout elementary calculus, whenever we wish a real number to exist so that we can carry out a certain discussion, our policy is to assume its existence and to postpone the proof to more advanced courses.

## EXERCISES

1. Express each one of the following in the form  $2^k 3^m a^r b^s$ , where  $k, m, r, s$  are integers.
  - a)  $\frac{1}{8} a^3 b^{-4} 2^5 a^{-2}$
  - b)  $3^{-4} 2^5 a^3 b^6 \cdot \frac{1}{2^3} \cdot \frac{1}{a^4} \cdot b^{-1} \cdot \frac{1}{9}$
  - c)  $\frac{3a^3 b^4}{2a^5 b^6}$
  - d)  $\frac{16a^{-3} b^{-5}}{9b^4 a^7 2^{-3}}$
2. What integer is  $81^{1/4}$  equal to?
3. What integer is  $(\sqrt{2})^6$  equal to?
4. Is  $(\sqrt{2})^5$  an integer?
5. Is  $(\sqrt{2})^{-5}$  a rational number? Is  $(\sqrt{2})^5$  a rational number?
6. In each case, the expression is equal to an integer. Which one?
  - a)  $16^{1/4}$
  - b)  $8^{1/3}$
  - c)  $9^{3/2}$
  - d)  $1^{5/4}$
  - e)  $8^{4/3}$
  - f)  $64^{2/3}$
  - g)  $25^{3/2}$
7. Express each of the following expressions as a simple decimal.
  - a)  $(.09)^{1/2}$
  - b)  $(.027)^{1/3}$
  - c)  $(.125)^{2/3}$
  - d)  $(1.21)^{1/2}$
8. Express each of the following expressions as a quotient  $m/n$ , where  $m, n$  are integers  $> 0$ .
  - a)  $\left(\frac{8}{27}\right)^{2/3}$
  - b)  $\left(\frac{4}{9}\right)^{1/2}$
  - c)  $\left(\frac{25}{16}\right)^{3/2}$
  - d)  $\left(\frac{49}{4}\right)^{3/2}$
9. Solve each of the following equations for  $x$ .
  - a)  $(x - 2)^3 = 5$
  - b)  $(x + 3)^2 = 4$
  - c)  $(x - 5)^{-2} = 9$
  - d)  $(x + 3)^3 = 27$
  - e)  $(2x - 1)^{-3} = 27$
  - f)  $(3x + 5)^{-4} = 64$

[Warning: Be careful with possible minus signs when extracting roots.]

#### §4. INEQUALITIES

We recall that we write

$$a > 0$$

if  $a$  is positive. If  $a, b$  are two real numbers, we shall write

$$a > b \quad \text{instead of} \quad a - b > 0.$$

We shall write

$$a < 0 \quad \text{instead of} \quad -a > 0$$

and also

$$b < a \quad \text{instead of} \quad a > b.$$

**Example.** We have  $3 > 2$  because  $3 - 2 = 1 > 0$ . We have

$$-1 > -2$$

because

$$-1 + 2 = 1 > 0.$$

In the geometric representation of numbers on the line, the relation  $a > b$  means that  $a$  lies to the right of  $b$ . We see that  $-1$  lies to the right of  $-2$  in our example.

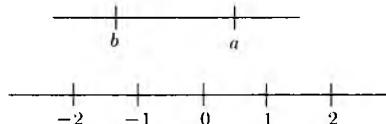


Fig. 3-2

We shall write

$$a \geq b$$

to mean  $a$  is greater than or equal to  $b$ . Thus

$$3 \geq 2 \quad \text{and} \quad 3 \geq 3$$

are both true inequalities.

Using only our two properties **POS 1** and **POS 2**, we shall prove rules for dealing with inequalities. In what follows, we let  $a, b, c$  be real numbers.

**IN 1.** If  $a > b$  and  $b > c$ , then  $a > c$ .

**IN 2.** If  $a > b$  and  $c > 0$ , then  $ac > bc$ .

**IN 3.** If  $a > b$  and  $c < 0$ , then  $ac < bc$ .

Rule IN 2 expresses the fact that an inequality which is multiplied by a positive number is preserved. Rule IN 3 tells us that if we multiply both sides of an inequality by a negative number, then the inequality gets *reversed*. For instance, we have the inequality

$$1 < 3.$$

But  $-2$  is negative, and if we multiply both sides by  $-2$  we get

$$-2 > -6.$$

This is represented geometrically by the fact that  $-2$  lies to the right of  $-6$  on the line.

Let us now prove the rules for inequalities.

To prove IN 1, suppose that  $a > b$  and  $b > c$ . By definition, this means that

$$a - b > 0$$

and

$$b - c > 0.$$

Using property POS 1, we conclude that

$$a - b + b - c > 0.$$

Canceling  $b$  gives us

$$a - c > 0,$$

which means that  $a > c$ , as was to be shown.

To prove IN 2, suppose that  $a > b$  and  $c > 0$ . By definition,

$$a - b > 0.$$

Hence using POS 1 concerning the product of positive numbers, we conclude that

$$(a - b)c > 0.$$

The left-hand side of this inequality is equal to  $ac - bc$  by distributivity. Therefore

$$ac - bc > 0,$$

which means that

$$ac > bc,$$

thus proving IN 2.

We shall leave the proof of IN 3 as an exercise.

Other properties which can easily be proved from the three basic ones will be left as exercises (see Exercises 2 through 5). They will be used constantly without further reference. In particular, we use some of them in the next examples.

**Example.** We wish to show that the inequality

$$2x - 4 > 5$$

is equivalent to an inequality of type  $x > a$  or  $x < b$ . Indeed, it is equivalent to

$$2x > 5 + 4 = 9,$$

which is equivalent to

$$x > \frac{9}{2}.$$

**Example.** Suppose that  $x$  is a number such that

$$(1) \quad \frac{3x + 5}{x - 4} < 2.$$

We wish to find equivalent conditions under which this is true, expressed by simpler inequalities like  $x > a$  or  $x < b$ . Note that the quotient on the left makes no sense if  $x = 4$ . Thus it is natural to consider the two cases separately,  $x > 4$  and  $x < 4$ . Suppose that  $x > 4$ . Then  $x - 4 > 0$  and hence, in this case, our inequality (1) is equivalent to

$$3x + 5 < 2(x - 4) = 2x - 8.$$

This in turn is equivalent to

$$3x - 2x < -8 - 5$$

or, in other words,

$$x < -13.$$

However, in our case  $x > 4$ , so that  $x < -13$  is impossible. Hence there is no number  $x > 4$  satisfying (1).

Now assume that  $x < 4$ . Then  $x - 4 < 0$  and  $x - 4$  is negative. We multiply both sides of our inequality (1) by  $x - 4$  and reverse the inequality. Thus inequality (1) is equivalent in the present case to

$$2) \quad 3x + 5 > 2(x - 4) = 2x - 8.$$

Furthermore, this inequality is equivalent to

$$(3) \quad 3x - 2x > -8 - 5$$

or, in other words,

$$(4) \quad x > -13.$$

However, in our case,  $x < 4$ . Thus in this case, we find that the numbers  $x$  such that  $x < 4$  and  $x > -13$  are precisely those satisfying inequality (1). This achieves what we wanted to do. Note that the preceding two inequalities holding simultaneously can be written in the form

$$-13 < x < 4.$$

The set of numbers  $x$  satisfying such inequalities is called an **interval**. The numbers  $-13$  and  $4$  are called the **endpoints** of the interval. We can represent the interval as in the following figure.



Fig. 3-3

**Example.** The set of numbers  $x$  such that  $3 < x < 7$  is an interval, shown in the next figure.

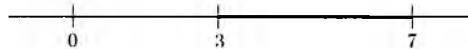


Fig. 3-4

**Example.** The set of numbers  $x$  such that  $3 \leq x \leq 7$  is also called an interval. In this case, we include the endpoints,  $3$  and  $7$ , in the interval. The word “interval” applies to both cases, whether or not we admit the endpoints. We represent the interval with the endpoints in the next figure.



Fig. 3-5

In general, let  $a, b$  be numbers with  $a \leq b$ . Any one of the following sets of numbers is called an **interval**.

The set of numbers  $x$  such that  $a < x < b$ , called an **open interval**.

The set of numbers  $x$  such that  $a \leq x \leq b$ , called a **closed interval**.

The set of numbers  $x$  such that  $a \leq x < b$ .

The set of numbers  $x$  such that  $a < x \leq b$ .

The last two intervals are called **half open** or **half closed**.

**Example.** Again by convention, it is customary to say that the set of all numbers  $x$  such that  $x > 7$  is an **infinite interval**. Similarly, the set of all numbers  $x$  such that  $x < -3$  is an infinite interval. In general, if  $a$  is a number, the set of numbers  $x$  such that  $x > a$  is an infinite interval, and so is the set of numbers  $x$  such that  $x < a$ . Again by convention, we may wish to include the endpoint. For instance, the set of numbers  $x$  such that  $x \geq 7$  is also called an infinite interval. The set of numbers  $x$  such that  $x \leq -3$  is called an infinite interval. We illustrate some of these intervals in the next figure.

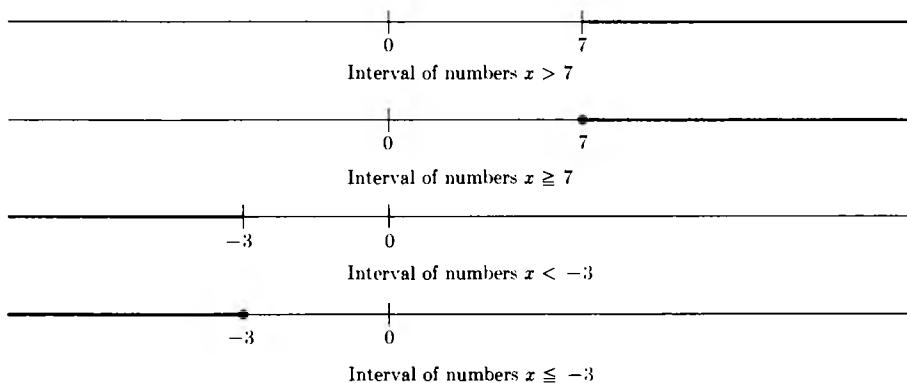


Fig. 3-6

## EXERCISES

1. Prove IN 3.
2. Prove: If  $0 < a < b$ , if  $c < d$ , and  $c > 0$ , then

$$ac < bd.$$

3. Prove: If  $a < b < 0$ , if  $c < d < 0$ , then

$$ac > bd.$$

4. a) If  $x < y$  and  $x > 0$ , prove that  $1/y < 1/x$ .  
 b) Prove a rule of **cross-multiplication of inequalities**: If  $a, b, c, d$  are numbers and  $b > 0, d > 0$ , and if

$$\frac{a}{b} < \frac{c}{d},$$

prove that

$$ad < bc.$$

Also prove the converse, that if  $ad < bc$ , then  $a/b < c/d$ .

5. Prove: If  $a < b$  and  $c$  is any real number, then

$$a + c < b + c.$$

Also,

$$a - c < b - c.$$

Thus a number may be subtracted from each side of an inequality without changing the validity of the inequality.

6. Prove: If  $a < b$  and  $a > 0$ , then

$$a^2 < b^2.$$

More generally, prove successively that

$$a^3 < b^3,$$

$$a^4 < b^4,$$

$$a^5 < b^5.$$

Proceeding stepwise, we conclude that

$$a^n < b^n$$

for every positive integer  $n$ . To make this stepwise argument formal, one must state explicitly a property of integers which is called induction, and is discussed later in the book.

7. Prove: If  $0 < a < b$ , then  $a^{1/n} < b^{1/n}$ . [Hint: Use Exercise 6.]  
 8. Let  $a, b, c, d$  be numbers and assume  $b > 0$  and  $d > 0$ . Assume that

$$\frac{a}{b} < \frac{c}{d}.$$

a) Prove that

$$\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}.$$

(There are two inequalities to be proved here, the one on the left and the one on the right.)

b) Let  $r$  be a number  $> 0$ . Prove that

$$\frac{a}{b} < \frac{a+rc}{b+rd} < \frac{c}{d}.$$

c) If  $0 < r < s$ , prove that

$$\frac{a+rc}{b+rd} < \frac{a+sc}{b+sd}.$$

9. If  $3x - 1 > 0$ , prove that  $x > \frac{1}{3}$ .

10. If  $4x + 5 < 0$ , prove that  $x < -\frac{5}{4}$ .

In each of the following cases, find the intervals of numbers  $x$  satisfying the stated inequalities.

11.  $5x + 2 > -3$

12.  $-2x + 1 < 4$

13.  $3x + 2 < 1$

14.  $-3x - 2 > 5$

15.  $3x - 1 < 4x + 5$

16.  $2x + 7 > -x + 3$

17.  $-3x - 1 > 5x$

18.  $2x + 1 < -3x - 2$

19.  $\frac{3x - 1}{x - 2} < 1$

20.  $\frac{-2x + 5}{x + 3} < 1$

21.  $\frac{2 - x}{2x + 1} > 2$

22.  $\frac{3 - x}{x - 5} > 4$

23.  $\frac{3 - 4x}{3x - 1} > 2$

24.  $\frac{3x + 1}{2x - 6} < 3$

25.  $x^2 < 1$

26.  $x^2 < 2$

27.  $x^2 < 3$

28.  $x^2 < 4$

29.  $x^2 > 1$

30.  $x^2 > 2$

31.  $x^2 > 3$

32.  $x^2 > 4$



# 4 Quadratic Equations

We know how to solve an equation like

$$3x - 2 = 0.$$

In such an equation,  $x$  appears only in the first power. We shall now consider the next most difficult case, when  $x$  appears to the second power. We first deal with examples.

**Example 1.** Consider the equation

$$(1) \quad x^2 - 3x + 1 = 0.$$

We wish to solve for  $x$ , that is, determine all values for  $x$  which satisfy this equation. We shall ultimately derive a general formula for this. Before deriving the formula, we carry out on this special example the method used to derive the general formula.

Solving our equation amounts to solving

$$(2) \quad x^2 - 3x = -1.$$

We wish to add a number to both sides of this equation so that the left-hand side becomes a square, of the form  $(x - s)^2$ . We know that

$$(x - s)^2 = x^2 - 2sx + s^2.$$

Thus we need  $2s = 3$ , or  $s = \frac{3}{2}$ . Consequently, adding  $(\frac{3}{2})^2$  to each side of equation (2), we find

$$x^2 - 3x + \frac{9}{4} = -1 + \frac{9}{4} = \frac{5}{4}.$$

The left-hand side has been adjusted so that it is a square, namely

$$x^2 - 3x + \frac{9}{4} = \left(x - \frac{3}{2}\right)^2,$$

and hence solving this equation amounts to solving

$$\left(x - \frac{3}{2}\right)^2 = \frac{5}{4}.$$

We can now take the square root, and we find that  $x$  is a solution if and only if

$$x - \frac{3}{2} = \pm \sqrt{\frac{5}{4}}.$$

Therefore finally we find two possible values for  $x$ , namely

$$x = \frac{3}{2} \pm \sqrt{\frac{5}{4}}.$$

This is an abbreviation for the two values

$$x = \frac{3}{2} + \sqrt{\frac{5}{4}} \quad \text{and} \quad x = \frac{3}{2} - \sqrt{\frac{5}{4}}.$$

**Example 2.** We wish to solve the equation

$$(3) \quad x^2 + 2x + 2 = 0.$$

We apply the same method as before. We must solve

$$x^2 + 2x = -2.$$

We add 1 to both sides, so that we are able to express the left-hand side in the form

$$x^2 + 2x + 1 = (x + 1)^2.$$

Solving equation (3) is equivalent to solving

$$(x + 1)^2 = -2 + 1 = -1.$$

But a negative real number cannot possibly be a square of a real number and we conclude that our equation does not have a solution in real numbers.

**Example 3.** We wish to solve the equation

$$(4) \quad 2x^2 - 3x - 5 = 0.$$

This amounts to solving

$$2x^2 - 3x = 5.$$

This time, we see that  $x^2$  is multiplied by 2. To reduce our problem to one similar to those already considered, we divide the whole equation by 2, and

solving (4) is equivalent to solving

$$(5) \quad x^2 - \frac{3}{2}x = \frac{5}{2}.$$

We can now complete the square on the left as we did before. We need to find a number  $s$  such that

$$x^2 - \frac{3}{2}x = x^2 - 2sx.$$

This means that  $s = \frac{3}{4}$ . Adding  $s^2$  to both sides of (5), we find

$$x^2 - \frac{3}{2}x + \frac{9}{16} = \frac{5}{2} + \frac{9}{16} = \frac{49}{16}.$$

Expressing the left-hand side as a square, this is equivalent to

$$\left(x - \frac{3}{4}\right)^2 = \frac{49}{16}.$$

We can now solve for  $x$ , getting

$$x - \frac{3}{4} = \pm \sqrt{\frac{49}{16}}$$

or equivalently,

$$x = \frac{3}{4} \pm \sqrt{\frac{49}{16}},$$

which is our answer. Although this answer is correct, it is sometimes convenient to watch for possible simplifications. In the present case, we note that

$$\sqrt{\frac{49}{16}} = \frac{7}{4},$$

and hence

$$x = \frac{3}{4} \pm \frac{7}{4}.$$

Therefore

$$x = \frac{10}{4} \quad \text{and} \quad x = \frac{-4}{4} = -1$$

are the two possible solutions of our equation.

We are now ready to deal with the general case.

**Theorem.** *Let  $a$ ,  $b$ ,  $c$  be real numbers and  $a \neq 0$ . The solutions of the quadratic equation*

$$6) \quad ax^2 + bx + c = 0$$

are given by the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

provided that  $b^2 - 4ac$  is positive, or 0. If  $b^2 - 4ac$  is negative, then the equation has no solution in the real numbers.

*Proof.* Solving our equation amounts to solving

$$ax^2 + bx = -c.$$

Dividing by  $a$ , we see that this is equivalent to solving

$$(7) \quad x^2 + \frac{b}{a}x = -\frac{c}{a}.$$

To complete the square on the left, we need

$$x^2 + \frac{b}{a}x = x^2 + 2sx,$$

and hence  $s = b/2a$ . We therefore add  $s^2 = b^2/4a^2$  to both sides of (7), and find the equivalent equation

$$\begin{aligned} \left(x + \frac{b}{2a}\right)^2 &= -\frac{c}{a} + \frac{b^2}{4a^2} \\ &= \frac{b^2 - 4ac}{4a^2}. \end{aligned}$$

If  $b^2 - 4ac$  is negative, then the right-hand side

$$\frac{b^2 - 4ac}{4a^2}$$

is negative, and hence cannot be the square of a real number. Thus our equation has no real solution. If  $b^2 - 4ac$  is positive, or 0, then we can take the square root, and we find

$$x + \frac{b}{2a} = \pm \frac{\sqrt{b^2 - 4ac}}{2a}.$$

Solving for  $x$  now yields

$$x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a},$$

which can be rewritten as

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

This proves our theorem.

**Remark.** If  $b^2 - 4ac = 0$ , then we get precisely one solution for the quadratic equation, namely

$$x = \frac{-b}{2a}.$$

If  $b^2 - 4ac > 0$ , then we get precisely two solutions, namely

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

and

$$x = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

The quadratic formula is so important that it should be memorized. Read it out loud like a poem, to get an aural memory of it:

*“x equals minus b plus or minus square root of b squared minus four ac over two a.”*

**Example 4.** Solve the equation

$$3x^2 - 2x + 1 = 0.$$

We use the formula this time, and get

$$\begin{aligned} x &= \frac{-(-2) \pm \sqrt{(-2)^2 - 4 \cdot 3}}{2 \cdot 3} \\ &= \frac{4 \pm \sqrt{-8}}{6}. \end{aligned}$$

In this case, we see that the expression  $b^2 - 4ac$  under the square root sign is negative, and thus our equation has no solution in the real numbers.

**Example 5.** Solve the equation

$$2x^2 + 3x - 4 = 0.$$

Again, use the formula, to get

$$\begin{aligned}x &= \frac{-3 \pm \sqrt{9 - 4 \cdot 2 \cdot (-4)}}{2 \cdot 2} \\&= \frac{-3 \pm \sqrt{9 + 32}}{4} \\&= \frac{-3 \pm \sqrt{41}}{4}.\end{aligned}$$

This is our answer, and we get the usual two values for  $x$ , namely

$$x = \frac{-3 + \sqrt{41}}{4} \quad \text{and} \quad x = \frac{-3 - \sqrt{41}}{4}.$$

**Remark.** In the proof of our theorem concerning the solutions of the quadratic equation, we needed to operate with addition, multiplication, and square roots. If we knew that the real numbers could be extended to a larger system of numbers in which these operations were valid, *including the possibility of taking square roots of negative real numbers*, then our formula would be valid in this bigger system of numbers, and would again give the solutions of the equation in all cases. We shall see in the chapter on complex numbers how to get such a system.

## EXERCISES

Solve the following equations. If there is no solution in the real numbers, say so, and give your reasons why. In each case, however, give the values for  $x$  which would solve the equation in a larger system of numbers where negative numbers have a “square root”. Use the formula.

- |                        |                        |
|------------------------|------------------------|
| 1. $x^2 + 3x - 2 = 0$  | 2. $x^2 - 3x - 2 = 0$  |
| 3. $x^2 - 4x + 5 = 0$  | 4. $x^2 - 4x - 5 = 0$  |
| 5. $3x^2 + 2x - 1 = 0$ | 6. $3x^2 - 4x + 1 = 0$ |
| 7. $3x^2 + 3x - 4 = 0$ | 8. $-2x^2 - 5x = 7$    |
| 9. $-2x^2 - 5x = -7$   | 10. $4x^2 + 5x = 6$    |

11.  $x^2 - \sqrt{2}x + 1 = 0$

12.  $x^2 + \sqrt{2}x - 1 = 0$

13.  $x^2 + 3x - \sqrt{2} = 0$

14.  $x^2 - 3x - \sqrt{5} = 0$

15.  $x^2 - 3x + \sqrt{5} = 0$

16.  $x^2 - 2x - \sqrt{3} = 0$

You will solve more quadratic equations when you do the exercises in Chapter 12, finding the intersection of a straight line with a circle, parabola, ellipse, or hyperbola.



*Interlude On Logic  
and Mathematical Expressions*



## **§1. ON READING BOOKS**

This part of the book can really be read at any time. We put it in the middle because that's as good as any place to start reading a book. Very few books are meant to be read from beginning to end, and there are many ways of reading a book. One of them is to start in the middle, and go simultaneously backwards and forward, looking back for the definitions of any terms you don't understand, while going ahead to see applications and motivation, which are very hard to put coherently in a systematic development. For instance, although we must do algebra first, it is quite appealing to look simultaneously at the geometry, in which we use algebraic tools to systematize our geometric intuition.

In writing the book, the whole subject has to be organized in a totally ordered way, along lines and pages, which is not the way our brain works naturally. But it is unavoidable that some topics have to be placed before others, even though our brain would like to perceive them simultaneously. This simultaneity cannot be achieved in writing, which thus gives a distortion of the subject. It is clear, however, that I cannot substitute for you in perceiving various sections of this book together. You must do that yourself. The book can only help you, and must be organized so that any theorem or definition which you need can be easily found.

Another way of reading this book is to start at the beginning, and then skip what you find obvious or skip what you find boring, while going ahead to further sections which appeal to you more. If you meet some term you don't understand, or if you need some previous theorem to push through the logical development of that section, you can look back to the proper reference, which now becomes more appealing to you because you need it for something which you already find appealing.

Finally, you may want to skim through the book rapidly from beginning to end, looking just at the statements of theorems, or at the discussions between theorems, to get an overall impression of the whole subject. Then you can go back to cover the material more systematically.

Any of these ways is quite valid, and which one you follow depends on your taste. When you take a course, the material will usually be covered in the same order as the book, because that is the safest way to keep going logically. Don't let that prevent you from experimenting with other ways.

## §2. LOGIC

We always try to keep clearly in mind what we assume and what we prove. By a "proof" we mean a sequence of statements each of which is either assumed, or follows from the preceding statements by a rule of deduction, which is itself assumed. These rules of deduction are essentially rules of common sense.

We use "If . . . , then" sentences when one statement implies another. For instance, we use sentences like:

$$(1) \quad \text{If } 2x = 5, \text{ then } x = \frac{5}{2}.$$

This is a true statement, patterned after the general sentence structure:

If  $A$ , then  $B$ .

The **converse** of this statement is given by:

If  $B$ , then  $A$ .

Thus the converse of our assertion (1) is:

$$(2) \quad \text{If } x = \frac{5}{2}, \text{ then } 2x = 5.$$

We see that the converse is also true.

Whenever we meet such a situation, we can save ourselves space, and simply say:

$$(3) \quad 2x = 5 \text{ if and only if } x = \frac{5}{2}.$$

Thus

“A only if B” means “If A, then B”.

However, using “only if” by itself rather than in the context of “if and only if” always sounds a little awkward. Because of the structure of the English language, one has a tendency to interpret “A only if B” to mean “if B, then A”. Consequently, we shall never use the phrase “only if” by itself, only as part of the full phrase “A if and only if B”.

**Example.** The assertion: “If  $x = -3$ , then  $x^2 = 9$ ” is a true statement. Its converse:

“If  $x^2 = 9$ , then  $x = -3$ ”

is a false statement, because  $x$  may be equal to 3. Thus the statement:

“ $x^2 = 9$  if and only if  $x = -3$ ”

is a false statement.

**Example.** The statement:

“If two lines are perpendicular, then they have a point in common”

is a true statement. Its converse:

“If two lines have a point in common, then they are perpendicular”

is a false statement.

**Example.** The statement:

“Two circles are congruent if and only if they have the same radius”

is a true statement.

We often give proofs by what is called the “method of contradiction”. We want to prove that a certain statement  $A$  is true. To do this, we suppose that  $A$  is false, and then by logical reasoning starting from the supposition that  $A$  is false, we arrive at an absurdity, or at a contradiction of a true statement. We then conclude that our supposition “ $A$  is false” cannot hold, whence  $A$  must be true. An example of this occurred when we proved that

$\sqrt{2}$  is not a rational number. We did this by assuming that  $\sqrt{2}$  is rational, then expressing it as a fraction in lowest form, and then showing that in fact, both numerator and denominator of this fraction must be even. This contradicted the hypothesis that  $\sqrt{2}$  could be a fraction, in lowest form, whence we concluded that  $\sqrt{2}$  is not a rational number.

Some assertions are true, some are false, and some are meaningless. Sometimes a set of symbols is meaningless because some letters, like  $x$ , or  $a$ , appear without being properly qualified. We give examples of this. When we write an equation like  $2x = 5$ , as in (1), the context is supposed to make it clear that  $x$  denotes a number. However, if there is any chance of doubt, this should always be specified. Thus a more adequate formulation of (1) would be:

$$(4) \quad \text{If } x \text{ is a real number and } 2x = 5, \text{ then } x = \frac{5}{2}.$$

Similarly, a more adequate formulation of (2) would be:

$$(5) \quad \text{Let } x \text{ be a real number. Then } 2x = 5 \text{ if and only if } x = \frac{5}{2}.$$

The symbols

$$2x = 5$$

by themselves are called an equation. As it stands, this equation simply indicates a possible relationship, but to give it meaning we must say something more about  $x$ . For instance:

- a) There exists a number  $x$  such that  $2x = 5$ .
- b) For all numbers  $x$ , we have  $2x = 5$ .
- c) There is no number  $x$  such that  $2x = 5$ .
- d) If  $x$  is a real number and  $2x = 5$ , then  $x < 7$ .

Of these statements, (a) is true, (b) is false, (c) is false, and (d) is true. We can also use the symbols “ $2x = 5$ ” in a context like:

- e) Determine all numbers  $x$  such that  $2x = 5$ .

This sentence is actually a little ambiguous, because of the word “determine”. In a sense, the equation itself,  $2x = 5$ , determines such numbers  $x$ . We have tried to avoid such ambiguities in this book. However, the context of a chapter can make the meaning of this sentence clear to us as follows:

- f) Express all rational numbers  $x$  such that  $2x = 5$  in the form  $m/n$ , where  $m, n$  are integers,  $n \neq 0$ .

This is what we would understand when faced with sentence (e), or with a similar sentence like:

- g) Solve for  $x$  in the equation  $2x = 5$ .

In writing mathematics, it is essential that complete sentences be used. Many mistakes occur because you allow incomplete symbols like

$$2x = 5$$

to occur, without the proper qualifications, as in sentences (a), (b), (c), (d), (e), (f), (g).

**Example.** The symbols " $x^2 = 2$ " by themselves are merely an equation. The sentence:

“There exists a rational number  $x$  such that  $x^2 = 2$ .”

is false. The sentence:

“There exists a real number  $x$  such that  $x^2 = 2$ .”

is true.

### Equality

We shall use the word “equality” between objects to mean that they are the same object. Thus when we write

$$2 + 3 = 6 - 1,$$

we mean that the number obtained by adding 2 and 3 is the same number as that obtained by subtracting 1 from 6. It is the number 5.

We use the word “equivalent” in several contexts. First, if  $A$  and  $B$  are assertions (which may be true or false), we say that they are equivalent to mean:

$A$  is true if and only if  $B$  is true.

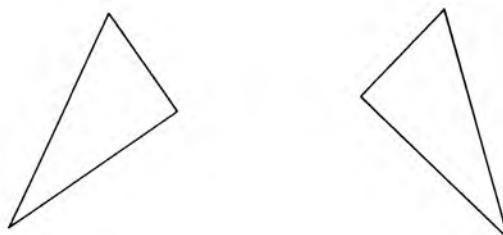
For instance, the following two assertions are equivalent in this sense:

*The number  $x$  satisfies the equation  $2x + 5 = 3$ .*

*The number  $x$  is equal to  $-1$ .*

We shall use the word “equivalent” in other contexts, but will explain these as the need arises.

We DO NOT USE THE WORD “EQUALITY” AS IT IS SOMETIMES USED, for instance in elementary geometry. The following two triangles are not equal:



**Fig. I-1**

They are, however, congruent, and under a suitable definition of equivalence for triangles, we might even say that they are equivalent. Note that the *areas* of these triangles are *equal*. In the same vein, the following line segments are not equal:



**Fig. I-2**

However, their lengths are equal.

The mathematics which we discuss in this book, like most mathematics, has many applications and counterparts in the physical world. For instance, numbers can be used to measure length, area, speed, density, etc. For clarity, we try to use language in such a way that the mathematical notions are not usually identified with their physical counterparts. Thus we use words like “*correspond*”, or “*represent*”, when we wish to associate a physical quantity with a mathematical one. In line with this, we can deal with mathematical objects on two levels: the purely logical level of axioms, deductions, and proofs; and the mixed physical level. Often, it is quite tedious and not necessarily illuminating to insist that we follow only the strictly logical procedures. It is useful and perhaps more pleasant to follow our physical intuition for certain arguments. We shall see examples of both types of arguments when we discuss geometry in its intuitive setting and its analytical setting.

### §3. SETS AND ELEMENTS

Following mathematical terminology, a collection of objects is called a **set**. The objects in this set are called the **elements** of the set.

The set of all real numbers is denoted by **R**. To say:

“ $x$  is an element of **R**”

means the same thing as to say

“ $x$  is a real number”.

Let  $S$  and  $T$  be sets. We say that  $S$  is a **subset** of  $T$  if every element of  $S$  is also an element of  $T$ . For instance:

The set of rational numbers is a subset of the set of real numbers.

The set of integers is a subset of the set of rational numbers.

The set of integers is a subset of the set of rational numbers. It is also a subset of **R** (i.e. a subset of the real numbers).

The set of boys is a subset of the set of all children.

The set of all real numbers  $x$  such that  $2x + 3 < 5$  is a subset of the real numbers.

As a matter of convention, we allow a subset of a set  $S$  to be all of  $S$ . Thus any set is a subset of itself. The sentence:

“For any set  $S$ ,  $S$  is a subset of  $S$ ”

is a true sentence.

A set is often described by stating the conditions under which something is an element of the set. Sometimes we state such conditions so that there are no elements in the set.

**Example.** There is no element in the set of all numbers  $x$  which satisfy the conditions

$$x < 0 \quad \text{and} \quad x > 0.$$

There is no element in the set of all positive numbers  $x$  which satifsy the conditions

$$\frac{2x}{x+1} > 1 \quad \text{and} \quad x < \frac{1}{2}.$$

Whenever this happens, that a set has no elements, we say that the set is empty. Thus the set of numbers  $x$  such that  $2x > 1$  and  $x < -3$  is empty.

Let  $S, S'$  be sets. Often, to prove that  $S = S'$ , we prove that  $S$  is a subset of  $S'$  and that  $S'$  is a subset of  $S$ .

**Example.** Let  $S$  be the set of numbers  $x$  such that  $1 \leq x \leq 2$ . Let  $T$  be the set of all numbers  $5x$  with all  $x$  in  $S$ . We contend that  $T$  is the set of numbers  $y$  with  $5 \leq y \leq 10$ . First note that if  $x$  is in  $S$ , then  $5x$  satisfies the inequalities

$$5 \leq 5x \leq 10.$$

Hence if  $T'$  is the set of all numbers  $y$  satisfying  $5 \leq y \leq 10$ , we see that  $T$  is contained in  $T'$ . Conversely, let  $y$  be a point of  $T'$ , i.e. assume that

$$5 \leq y \leq 10.$$

Let  $x = y/5$ . Then  $x$  is in  $S$  and  $y = 5x$ . Hence  $T'$  is contained in  $T$ . This proves that  $T = T'$ .

#### §4. INDICES

In a sentence like

“Let  $x, y$  be numbers”

it is a convention of mathematical language to allow the possibility that  $x = y$ . Similarly, if we say

“Let  $P, Q$  be points in the plane”

we do not exclude the possibility that  $P = Q$ . If we wish to exclude this possibility then we say so explicitly. For instance, we would say:

“Let  $x, y$  be distinct numbers”

or

“Let  $x, y$  be numbers,  $x \neq y$ ”

or

“Let  $P, Q$  be points, such that  $P \neq Q$ ”.

Similarly, we may wish to speak of several numbers instead of two numbers like  $x, y$ . Thus we might say

“Let  $x, y, z$  be numbers”,

without excluding the possibility that some of these numbers may be equal to each other. It is clear that we would soon run out of letters of the alphabet in enumerating numbers just with letters, and hence we use a notation with subscripts, as exemplified in the following sentences.

“Let  $x_1, x_2$  be numbers.”

“Let  $x_1, x_2, x_3$  be numbers.”

“Let  $x_1, x_2, x_3, x_4$  be numbers.”

Finally, in the most general case we have the corresponding sentence:

“Let  $x_1, \dots, x_n$  be numbers.”

We repeat that in such a sentence, it is possible that  $x_i = x_j$  for some pair of subscripts  $i, j$  such that  $i \neq j$ . Such subscripts are also called **indices**.

Objects indexed by integers from 1 to  $n$  (or sometimes from 0 to  $n$ ) are called a **sequence** of objects or, more precisely, a finite sequence. Thus in a finite sequence of numbers, denoted by

$$\{x_1, \dots, x_n\}$$

we associate a number  $x_j$  to each integer  $j$  satisfying  $1 \leq j \leq n$ . Thus, considering a sequence as above amounts to considering a first number  $x_1$ , a second number  $x_2$ , and so forth, up to an  $n$ -th number  $x_n$ .

**Example.** For each integer  $j$  we let  $x_j = (-1)^j$ . Then

$$x_1 = -1, \quad x_2 = 1, \quad x_3 = -1, \quad x_4 = 1, \quad \dots, \quad x_n = (-1)^n.$$

Observe how in this sequence the numbers  $x_j$  take on the values 1 or  $-1$ .

**Example.** We shall study polynomials later, and we shall write a polynomial in the form

$$a_n x^n + a_{n-1} x^{n-1} + \dots + a_0.$$

The sequence of coefficients is the sequence

$$\{a_0, a_1, \dots, a_n\}.$$

For instance, the sequence of coefficients of the polynomial

$$4x^3 - 2x^2 + 4x - 5$$

is the sequence  $\{-5, 4, -2, 4\}$ . We have

$$a_0 = -5, \quad a_1 = 4, \quad a_2 = -2, \quad a_3 = 4.$$

## §5. NOTATION

The notation used in giving an account of a mathematical theory is important. It is very useful that the printed page should look appealing visually as well as mathematically.

It is also important that notation be fairly consistent, namely that certain symbols be used only to denote certain objects. For instance, we use lower case letters like  $a, b, c, x, y, z$  to denote numbers. We use capital letters like  $A, B, P, Q, X, Y$  to denote mostly points, although we also use  $A, B$ , to denote angles. When we do that, we reserve  $P, Q$ , for points. Within any given section, we try not to mix the two. Although one should of course always specify what a letter stands for, it is convenient if the use of letters follows a pattern, so that one knows at one glance what certain letters represent.

Notation can be slick. You will see in the chapters on coordinates that the notation of points and vectors is fairly slick. Sometimes, notation can be too slick. I hope that this is not the case in those chapters.

We usually reserve letters like  $f, g, F, G$  for functions or mappings. We cannot observe complete uniformity in this respect; otherwise, we would run out of letters very soon. For instance, we use (a), (b), (c), ... to denote a sequence of exercises, not numbers.

We use  $m, n$  to denote integers, except in cases where they are used as abbreviations for words. For instance,  $m$  is sometimes used as an abbreviation to denote the measure of an angle  $A$ , which we write  $m(A)$ .

In any book, it is impossible to avoid some mistakes, some confusion, some incorrectness of language, and some misuse of notation. If you find any such things in the present book, then correct them or improve them for yourself, or write your own book. This is still the best way to learn a subject aside from teaching it.

*Part Two*

**INTUITIVE GEOMETRY**



This part is concerned with the geometry of the plane. We assume some basic properties, which we always try to state explicitly, including properties of straight lines, segments, angles, distance, etc. We then prove other facts from these.

There are two basic aspects of geometry: the Pythagoras theorem, and the notion of congruence. Classical treatments obscure these because of undue emphasis on “constructions”, and ever more complicated diagrams involving triangles, parallelograms, etc. I am trying to combat this by returning to an exposition based directly on these two aspects. You will see how easy some of the usual properties of triangles are to prove, if one starts with the Pythagoras theorem. On the other hand, the briefest glance at other expositions will convince you how unnatural and complicated plane geometry can be otherwise. For instance, the theorem that the shortest distance between a point  $P$  and points on a line is given by the perpendicular segment becomes obvious from Pythagoras, so obvious that we leave it as an exercise.

Taking the Pythagoras theorem and the general properties of isometries as our starting point has other advantages: it is precisely this approach which fits more advanced mathematics best. It provides an exceedingly nice introduction to mappings. It provides the intuitive basis for the study of perpendicularity (and ultimately of orthogonality in vector spaces in subsequent courses in linear algebra).

In the next part, we shall indicate systematically how we can give algebraic definitions for most of the concepts handled in intuitive geometry, and how we can prove the results of geometry using only properties of numbers, but in a way which most often parallels exactly the geometric arguments. This program involves giving algebraic definitions for points, lines, angles, triangles, reflections, dilations, translations, congruence, etc.

A proof which is based only on properties of real numbers is called an **analytic** proof. A proof which is based on properties of the plane involving our intuition as in Part II is called a **geometric** proof.

For the logical development of the subject, it is not absolutely necessary to read all of this part. The reader could skip it, and refer to it as needed later. But I don't think it is desirable to inhibit your geometric intuition (as a recent trend among some educational schools would do). Don't be afraid of arguing geometrically.

Giving analytic foundations for geometry, however, does not serve solely, or even principally, the purpose of making such foundations more "rigorous". It provides computational means which are not available in the intuitive context. For instance, given two lines which intersect, how do you compute their point of intersection? There is little problem if you have an analytic definition for a line. If you don't, the question doesn't even make that much sense. For another example, try to see geometrically why reflection through an arbitrary point can be obtained as a composite of reflection through the origin and a translation (to understand this, read the appropriate sections). It isn't nearly as clear as when you have the analytic definitions for translations and reflections given in Chapters 8 and 9. Thus analytic foundations for the subject should not be viewed as an invention to make the subject arid. Feel free to mix the two approaches—geometric and analytic—to get the best feeling for the subject. One reason why the Greeks did not get further in their mathematics is that they suffered from the inhibition of using numbers to deal with geometric objects. Essentially, mathematics had to wait till Descartes to overcome this inhibition. It is equally pointless to fall now into the opposite inhibition.

# 5 *Distance and Angles*

## §1. DISTANCE

The notion of distance is perhaps the most basic one concerning the plane.

We shall assume basic properties of distance without proof. We denote the distance between points  $P, Q$  in the plane by  $d(P, Q)$ . It is a number, which satisfies the following properties.

**DIST 1.** *For any points  $P, Q$ , we have  $d(P, Q) \geq 0$ . Furthermore,*

$$d(P, Q) = 0 \quad \text{if and only if} \quad P = Q.$$

**DIST 2.** *For any points  $P, Q$  we have*

$$d(P, Q) = d(Q, P).$$

**DIST 3.** *Let  $P, Q, M$  be points. Then*

$$d(P, M) \leq d(P, Q) + d(Q, M).$$

This third property is called the **triangle inequality**. The reason is that it expresses the geometric fact that the length of one side of a triangle is at most equal to the sum of the lengths of the other two sides, as illustrated in Fig. 5-1.

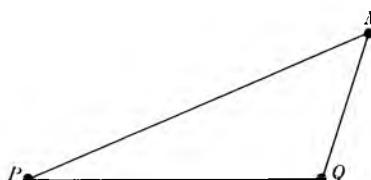


Fig. 5-1

We assume the basic fact that two distinct points  $P, Q$  lie on one and only one line, denoted by  $L_{PQ}$ . The portion of this line lying between  $P$  and  $Q$  is called the line **segment** between  $P$  and  $Q$ , and is denoted by  $\overline{PQ}$ . If units of measurement are selected, then the length of this segment is equal to the distance  $d(P, Q)$ . The straight line and the segment determined by  $P$  and  $Q$  are illustrated in Fig. 5-2.

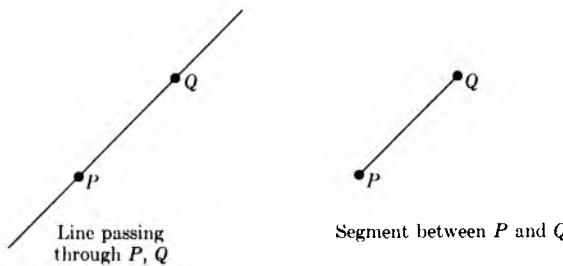


Fig. 5-2

We shall assume two facts relating line segments with the notion of distance. The first one is:

**SEG 1.** *Let  $P, Q, M$  be points. We have*

$$d(P, M) = d(P, Q) + d(Q, M)$$

*if and only if  $Q$  lies on the segment between  $P$  and  $M$ .*

This property **SEG 1** certainly fits our intuition of line segments, and is illustrated in Fig. 5-3, where  $Q$  lies on the segment  $\overline{PM}$ .

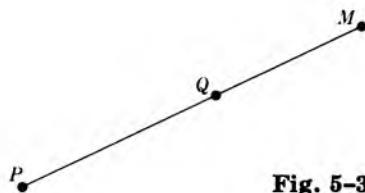


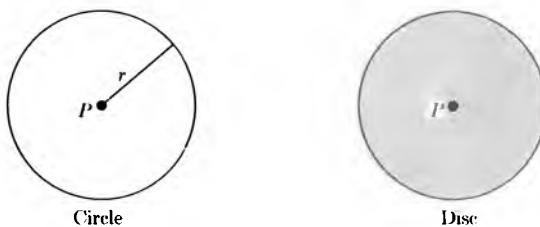
Fig. 5-3

The second fact which we assume is:

**SEG 2.** *Let  $P, M$  be points in the plane, and let  $d = d(P, M)$ . If  $c$  is a number such that  $0 \leq c \leq d$ , then there exists a unique point  $Q$  on the segment  $\overline{PM}$  such that  $d(P, Q) = c$ .*

Again, this property fits our intuition perfectly.

Let  $r$  be a positive number, and let  $P$  be a point in the plane. We define the **circle** of center  $P$  and radius  $r$  to be the set of all points  $Q$  whose distance from  $P$  is  $r$ . We define the **disc** of center  $P$  and radius  $r$  to be the set of all points  $Q$  whose distance from  $P$  is  $\leq r$ . The circle and the disc are drawn on Fig. 5-4.



**Fig. 5-4**

**Note.** In many books you will find that the word “circle” is used to denote both what we call the circle and also the disc. This is not good terminology and leads to confusion, because it is always best if a single word is not used to denote two different objects or concepts. Geometrically speaking, we see that the circle is the boundary of the disc.

**Remark.** In our preceding discussion, we have made the usual tacit convention that a unit of length has been fixed. For instance, just to speak of the distance between two points as a number, we must have agreed already on a unit of distance. Interpreting the distance between  $P$  and  $Q$  as the length of the segment between  $P$  and  $Q$  again presupposes that such a unit of measurement has been fixed.

A similar remark applies to later discussions, for instance concerning area. When a unit of distance is selected, then it determines a unit of area. For instance, if our unit of distance is the inch, then the unit area is the square inch. We then commit a simplification of language by referring to distance or area as numbers. We say that the area of a square of side  $a$  is  $a^2$ . Having fixed the unit of measurement as the inch, this means that the length of each side is  $a$  in., and that the area is  $a^2$  in<sup>2</sup>. For simplicity of language, we agree once for all that a unit of measurement is fixed throughout our discussion, and then omit the units when speaking of length (distance) or area. Sometimes we also speak of the “numerical value” of the length, or area, with respect to such a choice of units, to emphasize that we are dealing with a number. For instance, the numerical value of the area of a square whose area is 9 in<sup>2</sup> is the number 9.

## §2. ANGLES

We base our discussion of geometry on our concept of the plane. We are willing to assume some standard geometric facts, but for the convenience of the reader, we have reproved many facts from more basic ones. We assume the following facts about straight lines.

*Two distinct points  $P$ ,  $Q$  lie on one and only one line, denoted by  $L_{PQ}$ . Two lines which are not parallel meet in exactly one point. Given a line  $L$  and a point  $P$ , there exists a unique line through  $P$  parallel to  $L$ . If  $L_1$ ,  $L_2$ ,  $L_3$  are lines, if  $L_1$  is parallel to  $L_2$  and  $L_2$  is parallel to  $L_3$ , then  $L_1$  is parallel to  $L_3$ .*

*Given a line  $L$  and a point  $P$ , there exists a unique line through  $P$  perpendicular to  $L$ . If  $L_1$  is perpendicular to  $L_2$  and  $L_2$  is parallel to  $L_3$ , then  $L_1$  is perpendicular to  $L_3$ . If  $L_1$  is perpendicular to  $L_2$  and  $L_2$  is perpendicular to  $L_3$ , then  $L_1$  is parallel to  $L_3$ .*

Two points  $P$  and  $Q$  also determine two rays, one starting from  $P$  and the other starting from  $Q$ , as shown in Fig. 5-5. Each of these rays stops at  $P$ , but extends infinitely in one direction.

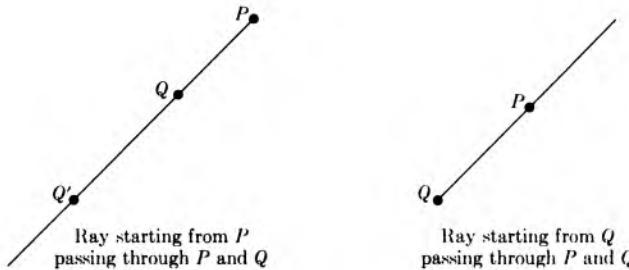


Fig. 5-5

A ray starting from  $P$  is simply a half line, consisting of all points on a line through  $P$  lying to one side of  $P$ . The ray starting from  $P$  and passing through another point  $Q$  will be denoted by  $R_{PQ}$ . If  $Q'$  is another point on this ray, distinct from  $P$ , then of course we have

$$R_{PQ} = R_{PQ'}.$$

In other words, a ray is determined by its starting point and by any other point on it.

If a ray starts at a point  $P$ , we also call  $P$  the **vertex** of the ray.

Consider two rays  $R_{PQ}$  and  $R_{PM}$  starting from the same point  $P$ . These rays separate the plane into two regions, as shown in Fig. 5-6.

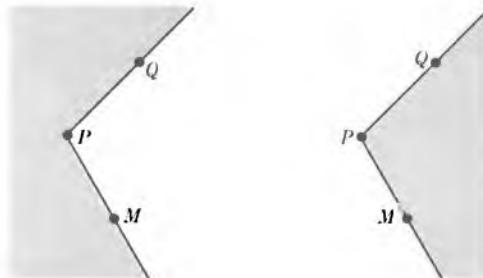


Fig. 5-6

Each one of these regions will be called an **angle** determined by the rays. Thus the rays  $R_{PQ}$  and  $R_{PM}$  determine two angles.

**Remark on terminology.** There is some divergence in the way an angle is defined in other books. For instance, an angle is sometimes defined as the union of two rays having a common vertex, rather than the way we have defined it. I have chosen a different convention for several reasons. First, people do tend to think of one or the other side of the rays when they meet two rays like this:

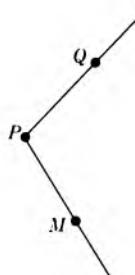


Fig. 5-7

They do not think neutrally. Second, and more importantly, when we want to measure angles later, and assign a number to an angle, as when we shall say that an angle has 30 degrees, or 270 degrees, adopting the definition of an angle as the union of two rays would provide insufficient information for such purposes, and we would need to give additional information to determine the associated measure. Thus it is just as well to incorporate this information

in our definition of an angle. Finally, the manner in which the measure of an angle will be found will rely on area, and is therefore natural, starting with our definition.

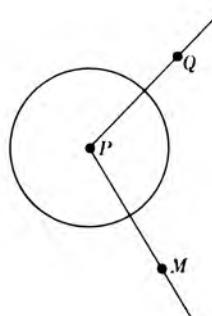
If we just draw these rays as shown in Fig. 5-8, like this, without any other indication, then we cannot tell which angle is meant, and thus we need some additional notation to distinguish one angle from the other, which we now describe.



**Fig. 5-8**

Recall that given a point  $P$  and a positive number  $r$ , the circle of radius  $r$  and center  $P$  is the collection of all points whose distance from  $P$  is equal to  $r$ .

Let  $R_{PQ}$  and  $R_{PM}$  be rays with vertex  $P$ . If  $C$  is a circle centered at  $P$  (of positive radius), then our two rays separate the circle into two arcs, as shown in Fig. 5-9.



**Fig. 5-9**

Each arc lies within one of the angles, and thus to characterize each angle it suffices to draw the corresponding arc. The two parts of Fig. 5-10 thus show the usual way in which we draw the two angles formed by the rays.

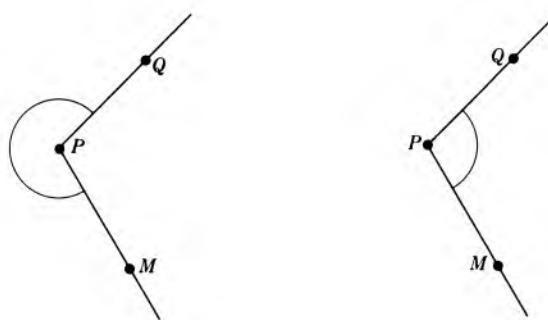


Fig. 5-10

Just knowing the two rays is not enough information to be able to distinguish one angle from the other. However, if the rays are given in an ordered fashion, selecting one of them as the first and the other as the second, then we do have enough information to determine one specific angle. This is done as follows.

Let  $R_{PQ}$  be the first ray and  $R_{PM}$  the second one. Then one of the angles determined by  $R_{PQ}$  and  $R_{PM}$  contains the arc going from the first ray to the second in the **counterclockwise** direction. We denote that angle by  $\angle QPM$ . The other angle contains the arc from  $R_{PM}$  to  $R_{PQ}$  in the **counterclockwise** direction, and we therefore denote this other angle by  $\angle MPQ$ . Thus the order in which  $Q, M$  occur is very important in this notation. We represent the angles  $\angle QPM$  and  $\angle MPQ$  by putting a little arrow on the arc, to indicate the counterclockwise direction, as in Fig. 5-11.

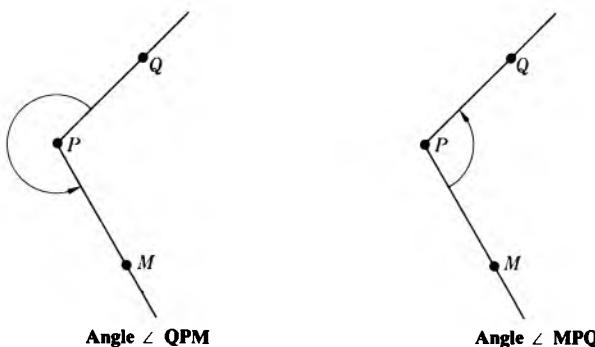


Fig. 5-11

**Example.** If  $Q, P, M$  lie on the same straight line, and  $Q, M$  lie on the same ray starting at  $P$ , then the angle  $\angle QPM$  looks like this:

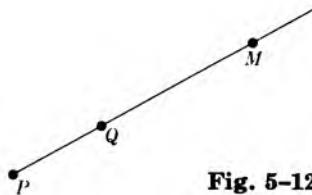


Fig. 5-12

In this case, the arc of a circle between the two rays is just a point, and we say that the angle  $\angle QPM$  is the **zero angle**.

Note that when we deal with this degenerate case in which the two rays coincide, one of the angles is the zero angle but the other angle is the whole plane, and is called the **full angle**. However, with our conventions, we do not write this full angle with the notation  $\angle QPM$ . We do, however, represent it by an arrow going all the way around as follows:

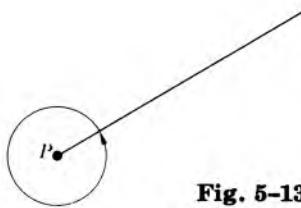


Fig. 5-13

**Example.** Suppose that  $Q, P, M$  lie on the same straight line but that  $Q$  and  $M$  do not lie on the same ray, that is,  $Q$  and  $M$  lie on opposite sides of  $P$  on the line. Our angle  $\angle QPM$  looks like this:

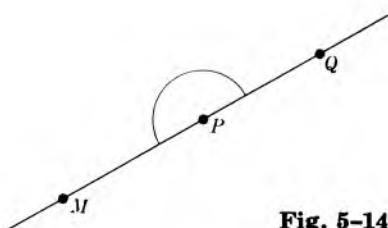


Fig. 5-14

In this case, we say that the angle  $\angle QPM$  is a **straight angle**. Observe that we draw the angle  $\angle MPQ$  with a different arc, namely:

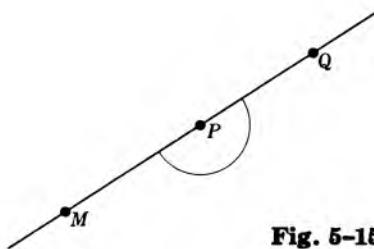


Fig. 5-15

Thus in this case,  $\angle MPQ$  is different from  $\angle QPM$ , and both are straight angles because the three points  $P, Q, M$  lie on the same straight line.

Given an angle  $A$  with vertex  $P$ , let  $D$  be a disc centered at  $P$ . That part of the angle which also lies in the disc is called the **sector** of the disc determined by the angle. Picture:

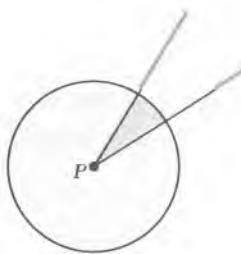


Fig. 5-16

The shaded part represents the sector  $S$ .

Just as we used numbers to measure distance, we can now use them to measure angles, provided that we select a unit of measurement first. This can be done in several ways. Here we discuss the most elementary way (but we shall return later to this question, and discuss another unit, which turns out to be more convenient in most mathematics).

The unit of measurement which we select here is the degree, such that the full angle has 360 degrees. Let  $A$  be an angle centered at  $P$  and let  $S$  be the sector determined by  $A$  in the disc  $D$  centered at  $P$ . Let  $x$  be a number between 0 and 360. We shall say that

$A$  has  $x$  degrees

to mean that

$$\frac{\text{area of } S}{\text{area of } D} = \frac{x}{360}.$$

Thus

$$x = 360 \cdot \frac{\text{area of } S}{\text{area of } D}.$$

In computing the number of degrees in an angle, we do not have to determine the area of  $S$  or even that of  $D$ , only the ratio between the two. We shall now give examples.

**Example.** The straight angle has 180 degrees because it separates the disc into two sectors of equal area, as shown in Fig. 5-17.

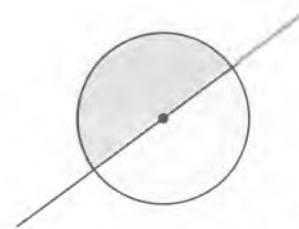


Fig. 5-17

**Example.** An angle whose measure is half that of the straight angle is called a **right angle**, and has 90 degrees, as in the Fig. 5-18.

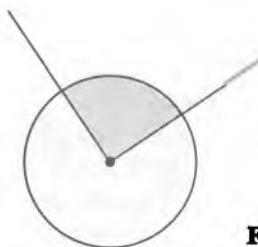


Fig. 5-18

**Example.** In Fig. 5-19, we have drawn the sectors determined by angles of 45 degrees and 30 degrees. The one with 30 degrees has one-third the measure of a right angle. In the picture of an angle of 45 degrees, we have drawn a dotted line to suggest the angle of 90 degrees. In the picture of an angle of 30 degrees, we have drawn two dotted lines to suggest the angles of 90 degrees.

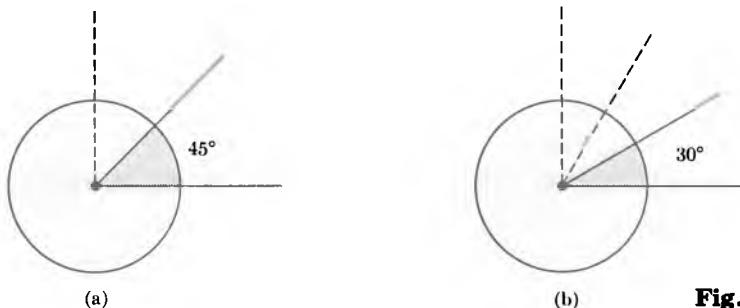


Fig. 5-19

and  $60$  degrees, respectively, showing how the angle of  $90$  degrees gets divided into three parts having equal measures.

**Example.** In Fig. 5-19(c) we have drawn the sector lying between the angles of  $30^\circ$  and  $45^\circ$ , and inside the circle of radius  $2$ .

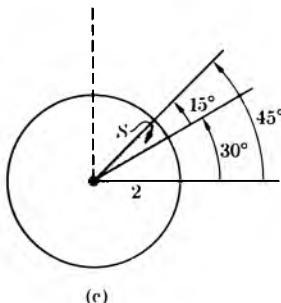


Fig. 5-19 (cont.)

We can compute the area of this sector using the definition of degrees. Let us assume the fact that the area of the disc of radius  $r$  is  $\pi r^2$ , where  $\pi$  is approximately equal to  $3.14159 \dots$ . (The decimal for  $\pi$  can be determined as accurately as you wish, but we don't go into this here.) The area of the disc of radius  $2$  is therefore equal to  $4\pi$ . The sector  $S$  in Fig. 5-19(c) has  $15^\circ$  (because  $15 = 45 - 30$ ), and hence

$$\text{area of } S = \frac{15}{360} \cdot 4\pi = \frac{\pi}{6}.$$

This is the numerical value of the area, in whatever units we are dealing with. You can put this answer into decimal form, using tables for  $\pi$ , and a computer, but we prefer to leave it as  $\pi/6$ .

Similarly, the area of the sector lying between the angles of  $30^\circ$  and  $45^\circ$ , and inside the circle of radius  $5$ , is given by

$$\frac{15}{360} \cdot 25\pi = \frac{25}{24}\pi.$$

We shall abbreviate "degree" by deg, and also use a small upper circle to the right of a number to denote degrees. Thus we write

$$\begin{aligned} 30 \text{ degrees} &= 30 \text{ deg} \\ &= 30^\circ, \end{aligned}$$

or more generally with any number  $x$  between 0 and 360, we write

$$\begin{aligned} x \text{ degrees} &= x \text{ deg} \\ &= x^\circ. \end{aligned}$$

The measure of an angle  $A$  will be denoted by  $m(A)$ . For the moment we shall deal with the measure in degrees. Thus to say that an angle  $A$  has  $50^\circ$  means the same thing as

$$m(A) = 50^\circ.$$

**Remark addressed to those who like to ask questions.** In defining the number of degrees of an angle, we used a disc  $D$ . We did not specify the radius. It should be intuitively clear that when we change the disc, and hence the sector at the same time, the ratio of their areas remains the same. We shall assume this for now, and return to a more thorough discussion of this question later when we discuss area, and similar figures.

It is convenient to write inequalities between angles. Let  $A$  and  $B$  be angles. Suppose that  $A$  has  $x$  degrees and  $B$  has  $y$  degrees, where  $x, y$  are numbers satisfying

$$0 \leq x \leq 360$$

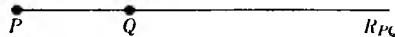
and

$$0 \leq y \leq 360.$$

We shall say that  $A$  is **smaller than or equal to**  $B$  if  $x \leq y$ . For instance, an angle of  $37^\circ$  is smaller than an angle of  $52^\circ$ .

## EXERCISES

1. Let  $R_{PQ}$  be a ray as drawn, horizontally.



Draw a second ray  $R_{PM}$  such that the angle  $\angle QPM$  has:

- a)  $60^\circ$
- b)  $120^\circ$
- c)  $135^\circ$
- d)  $160^\circ$
- e)  $210^\circ$
- f)  $225^\circ$
- g)  $240^\circ$
- h)  $270^\circ$

Let  $D$  be a disc centered at  $P$  and assume that its area is  $60 \text{ in}^2$ . In each one of the above cases find the area of the sector in the disc cut out by the two rays.

2. Assume that the area of a disc of radius 1 is equal to the number  $\pi$  (approximately equal to  $3.14159\dots$ ) and that the area of a disc of radius  $r$  is equal to  $\pi r^2$ .

- a) What is the area of a sector in the disc of radius  $r$  lying between angles of  $\theta_1$  and  $\theta_2$  degrees, as shown in Fig. 5-20(a)?
- b) What is the area of the band lying between two circles of radii  $r_1$  and  $r_2$  as shown in Fig. 5-20(b)?
- c) What is the area in the region bounded by angles of  $\theta_1$  and  $\theta_2$  degrees and lying between circles of radii  $r_1$  and  $r_2$  as shown in Fig. 5-20(c)?

Give your answers in terms of  $\pi$ ,  $\theta_1$ ,  $\theta_2$ ,  $r_1$ ,  $r_2$ .

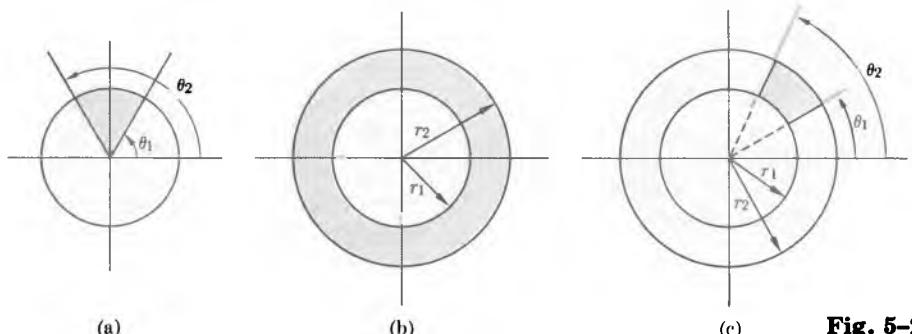


Fig. 5-20

Work out numerical examples of Exercise 2 as follows. In each case, express your answer in terms of rational multiples of  $\pi$ .

3. What is the area of a sector in the disc of radius 2 lying between angles of:
- a)  $35^\circ$  and  $75^\circ$ ,
  - b)  $15^\circ$  and  $60^\circ$ ,
  - c)  $80^\circ$  and  $110^\circ$ ,
  - d)  $130^\circ$  and  $250^\circ$ ?

4. What is the area of the band lying between two circles having the same center, and having the following radii?
- 2 and 5
  - 3 and 4
  - 2 and 6
  - 1 and 5
5. What is the area of the region bounded by angles of  $\theta_1$  and  $\theta_2$  degrees, and lying between circles of radii 3 and 5 when  $\theta_1$  and  $\theta_2$  have the values of Exercise 3(a), (b), (c), and (d)?
6. What is the area of the region lying between two circles having the same center, of radii 3 and 4, and bounded by angles of:
- $60^\circ$  and  $70^\circ$ ,
  - $110^\circ$  and  $270^\circ$ ,
  - $65^\circ$  and  $120^\circ$ ,
  - $240^\circ$  and  $310^\circ$ ?

### §3. THE PYTHAGORAS THEOREM

Let  $P, Q, M$  be three points in the plane, not on the same line. These points determine three line segments, namely

$$\overline{PQ}, \quad \overline{QM}, \quad \overline{PM}.$$

The set consisting of these three line segments is called the **triangle** determined by  $P, Q, M$ , and is denoted by

$$\triangle PQM.$$

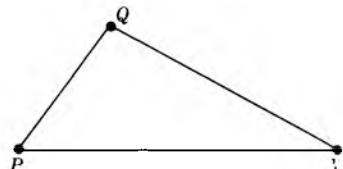


Fig. 5-21

**Remark on terminology.** We adopt here a convention which seems the most widespread. However, there is some pervasive ambiguity about the notion of a triangle, similar to the ambiguity which we have already mentioned about "circles". The word "triangle" is also used to denote the region

bounded by the three line segments. There is no convenient word like "disc" which I could think of here to serve in a similar capacity. Nobody will accept "trisc". (Mathematicians have a good time thinking up words like that.) "Triangular region" is the expression that seems the most natural to use. There is a mathematical word, "simplex", which is used for triangular regions and their analogs in higher dimensions (pyramids, tetrahedrons, etc.). For this book, we shall be satisfied with "triangle" as we defined it. On the other hand, we shall commit a slight abuse of language, and speak of the "area of a triangle", when we mean the "area of the triangular region bounded by a triangle". This is current usage, and although slightly incorrect, it does not really lead to serious misunderstandings.

Each pair of sides of the triangle determines an angle. We shall say that a triangle is a **right triangle** if one of these angles is a right angle. The sides of the triangle which determine this angle are then called the **legs** of the right triangle. A right triangle looks like this. (Fig. 5-22.)

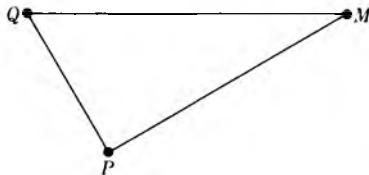


Fig. 5-22

The legs of the right triangle in Fig. 5-22 are the sides  $\overline{PQ}$  and  $\overline{PM}$ .

In our development of geometry, we adopt the following attitude. We take for granted certain basic properties about lines (mentioned before), perpendicularity, and figures like right triangles and rectangles whose main features have to do with perpendicularity. These will be stated as explicitly as possible, to make the situation psychologically satisfactory, I hope. We then prove properties about other geometric figures from these.

In subsequent sections and chapters of the book, we shall show how such foundational materials can be further understood (e.g. by our discussion of congruences, and by coordinates).

One basic fact which we take for granted about right triangles is:

**RT.** *If two right triangles  $\triangle PQM$  and  $\triangle P'Q'M'$  have legs  $\overline{PQ}$ ,  $\overline{PM}$  and  $\overline{P'Q'}$ ,  $\overline{P'M'}$ , respectively of equal lengths, that is,*

$$\begin{aligned} \text{length } \overline{PQ} &= \text{length } \overline{P'Q'} \\ \text{length } \overline{PM} &= \text{length } \overline{P'M'}, \end{aligned}$$

*then: (a) the corresponding angles of the triangles have equal measure, (b) their areas are equal, and (c) the length of  $\overline{QM}$  is equal to the length of  $\overline{Q'M'}$ .*

Figure 5–23 illustrates such right triangles as in our axiom RT.

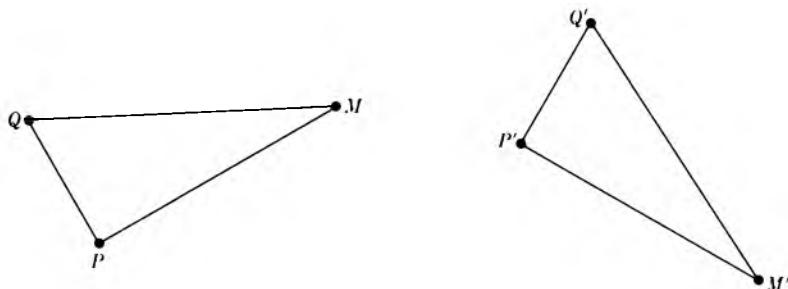


Fig. 5-23

Some of you may already know about the notion of congruence, and if you do then you will immediately realize that under the hypotheses of RT. the two triangles are congruent. Roughly speaking, this means that you can “move” one over the other so that the corresponding legs lie over each other. Later we shall deal formally with the notion of congruence, and develop the theory systematically. At this point, we are only concerned with getting one basic theorem, the Pythagoras Theorem, and we don’t want to burden ourselves with a whole theory just to get to it, especially since our axiom RT is psychologically very satisfactory (to me, and I hope to you). We summarized in RT just what we need for our immediate purposes.

The choice of what is assumed in a theory and what is “proved” depends on many requirements. In a subject like geometry, what we assume must in some sense be intuitively obvious. We do not wish to assume so much that we feel uncomfortable about it, and feel that we have really cheated on the theory. On the other hand, we do not wish to assume so little that it becomes very difficult or involved to prove statements which our mind perceives at once as “obvious”. We wish to minimize the basic assumptions, and to maximize what we can deduce easily from them. If our system of assumptions is very small, and corresponds to geometric properties which we regard as “obvious”, and if we can then deduce easily and fast many properties which we do not regard as “obvious”, then we have gone a long way towards finding a satisfactory set of assumptions. If I didn’t think that the choice I have made about this was reasonably successful, I wouldn’t have written a book . . . . In addition to that, however, experience with more advanced topics shows that the notion of perpendicularity is all-pervasive in mathematics, and that it turns out to be always worth while to have taken it as a fundamental notion, and to have taken as axioms some of its basic properties.

We shall see how some properties of triangles can be reduced to properties of rectangles.

Before defining rectangles, we mention explicitly a property relating parallel lines and distance. Let  $L, L'$  be parallel lines, and let  $P, Q$  be points on  $L$ . Let  $K_P$  be the line perpendicular to  $L$  passing through  $P$ , and let  $P'$  be the intersection of  $K_P$  with  $L'$ . Similarly, let  $K_Q$  be the line through  $Q$  perpendicular to  $L$ , and let  $Q'$  be the intersection of  $K_Q$  with  $L'$ . We shall assume:

**PD.** *The lengths of the segments  $\overline{PP'}$  and  $\overline{QQ'}$  are the same. In other words,*

$$d(P, P') = d(Q, Q').$$

This is illustrated by Fig. 5-24(a). We may call this length the **distance** between the two lines.

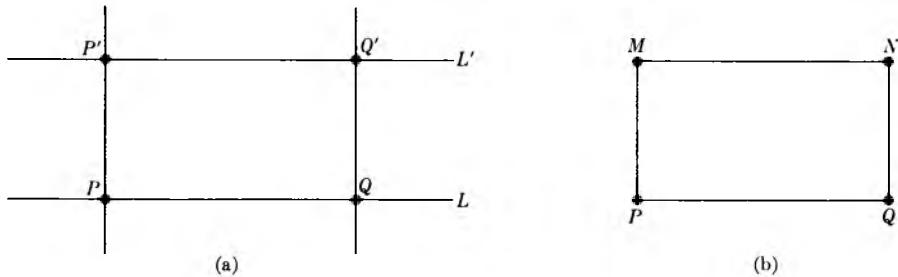


Fig. 5-24

Suppose now that  $P, Q, M, N$  are four points, such that the segments  $\overline{PQ}, \overline{QN}, \overline{NM}$ , and  $\overline{MP}$  form a four-sided figure. Suppose that the opposite sides  $\overline{PQ}, \overline{NM}$  are parallel, and also the opposite sides  $\overline{QN}, \overline{MP}$  are parallel; suppose also that the adjacent sides are perpendicular, that is:  $\overline{PQ}, \overline{QN}$  are perpendicular and  $\overline{NM}, \overline{MP}$  are perpendicular. Then we shall call the set consisting of the four segments

$$\overline{PQ}, \overline{QN}, \overline{NM}, \overline{MP}$$

the **rectangle** determined by  $P, Q, N, M$ . This rectangle is illustrated in Fig. 5-24(b). Observe that according to our property PD, it follows that the opposite sides of the rectangle have the same length.

If  $a, b$  are the lengths of the sides of the rectangle, then we assume that the area of the rectangle is equal to  $ab$ . (Comment: We are committing here the same abuse of language by speaking of the area of the rectangle that we did with triangles. We mean, of course, the area of the region bounded by the rectangle.) As usual, a **square** is a rectangle all of whose sides have the same length. If this length is equal to  $a$ , then the area of the square is  $a^2$ .

We are interested in the area of a right triangle. Consider a right triangle  $\triangle QPM$  such that  $\angle MPQ$  is its right angle, as shown in Fig. 5-25. Then the segments  $\overline{PQ}$  and  $\overline{PM}$  are perpendicular. We let  $N$  be the point of intersection of the line through  $M$  parallel to  $\overline{PQ}$ , and the line through  $Q$  perpendicular to  $\overline{PQ}$ . Then  $N$  is the fourth corner of the rectangle whose three other corners are  $Q, P, M$ . Then the sides  $\overline{QN}$  and  $\overline{PM}$  have equal lengths. The sides  $\overline{QP}$  and  $\overline{NM}$  have equal lengths.

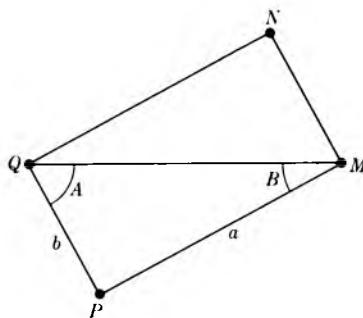


Fig. 5-25

Let  $A, B$  be the angles of the right triangle, other than the right angle, as shown in Fig. 5-25. It follows from RT that  $\angle NQM$  has the same measure as  $B$ . Since  $\angle NQP$  is a right angle, and since  $A$  and  $\angle NQM$  together form this right angle  $\angle NQP$ , it follows that:

**Theorem 1.** *If  $A, B$  are the angles of a right triangle other than the right angle, then*

$$m(A) + m(B) = 90^\circ.$$

Let  $a, b$  be the lengths of the sides of the rectangle. Then  $a, b$  are also the lengths of the legs of the right triangle  $\triangle MPQ$ . We assumed that the area of the rectangle is equal to  $ab$ . Again by RT, we conclude that the two triangles  $\triangle QPM$  and  $\triangle QNM$  which form this rectangle have the same area.

Hence we find:

**Theorem 2.** *The area of a right triangle whose legs have lengths  $a, b$ , is equal to*

$$\frac{ab}{2}.$$

The third side of a right triangle, which is not one of the legs, is called the **hypotenuse**. The next theorem gives us the relation between the length of the hypotenuse and the lengths of the other two sides.

**Pythagoras Theorem.** *Let  $a, b$  be the lengths of the two legs of a right triangle, and let  $c$  be the length of the hypotenuse. Then*

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

**Proof.** Let us draw the right triangle with the leg of length  $a$  horizontally as shown in Fig. 5-26. Let the triangle be  $\triangle PQM$ , with right angle at  $P$ , as shown. Let  $P_1$  be the point on the line through  $P, M$  at a distance  $b$  from  $M$ .

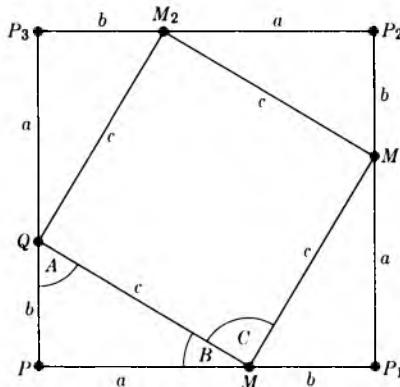


Fig. 5-26

and distance  $a + b$  from  $P$ . We draw the segment  $\overline{P_1P_2}$ , perpendicular to  $\overline{PP_1}$ , on the same side of  $\overline{PP_1}$  as the triangle, and of length  $a + b$ . We then draw the other two sides of the square whose sides have length  $a + b$ , as shown. The point  $P_1$  is the vertex of a right angle, and we can form a right triangle one of whose legs is the segment  $\overline{MP_1}$ , of length  $b$ , and the other leg is the vertical segment  $\overline{P_1M_1}$  of length  $a$ . We can now repeat this construction, forming a third right triangle  $\triangle M_1P_2M_2$ , and then a fourth right triangle  $\triangle M_2P_3Q$ . Each one of these right triangles has legs of lengths  $a$  and  $b$ , respectively. Consequently, by RT, the sides of the four-sided figure inside

the big square have the same length, equal to  $c$ . Let  $A, B$  be the angles of our right triangle other than the right angle. Let  $C$  be any one of the angles of the four-sided figure, say  $\angle M_1MQ$ . By RT we know that  $\angle M_1MP_1$  has the same measure as  $A$ , and therefore

$$m(B) + m(C) + m(A) = 180^\circ.$$

But we have by Theorem 1,

$$m(A) + m(B) = 90^\circ,$$

so that  $m(C) = 90^\circ$ . Hence the four-sided figure inside the big square is a square, whose sides have length  $c$ .

We now compute areas. The area of the big square is

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

This area is equal to the sum of the areas of the four triangles, and the area of the square whose sides have length  $c$ . Thus it is also equal to

$$\frac{ab}{2} + \frac{ab}{2} + \frac{ab}{2} + \frac{ab}{2} + c^2 = 2ab + c^2.$$

This yields

$$a^2 + 2ab + b^2 = 2ab + c^2,$$

whence

$$a^2 + b^2 = c^2,$$

and the theorem is proved.

**Example.** The length of the diagonal of a square whose sides have length 1, as in Fig. 5-27(a), is equal to

$$\sqrt{1^2 + 1^2} = \sqrt{2}.$$

The length of the diagonal of a rectangle whose sides have lengths 3 and 4, as in Fig. 5-27(b), is equal to

$$\begin{aligned}\sqrt{3^2 + 4^2} &= \sqrt{9 + 16} \\ &= \sqrt{25} = 5.\end{aligned}$$

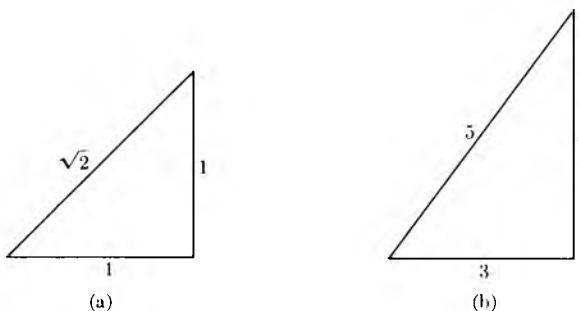


Fig. 5-27

**Example.** One leg of a right triangle has length 10 in., and the hypotenuse has length 15 in. What is the length of the other side?

This is easily done. Let  $b$  be this length. Then by Pythagoras, we get

$$10^2 + b^2 = 15^2,$$

whence

$$\begin{aligned} b^2 &= 15^2 - 10^2 \\ &= 225 - 100 = 125. \end{aligned}$$

Hence  $b = \sqrt{125}$ .

Let  $P, Q$  be distinct points in the plane. We recall that the **perpendicular bisector** of the segment  $\overline{PQ}$  is the line perpendicular to  $\overline{PQ}$  passing through the point which lies on  $\overline{PQ}$ , halfway between  $P$  and  $Q$ . Observe that if  $O$  is any point on the line passing through  $P$  and  $Q$  which is such that

$$d(O, P) = d(O, Q),$$

then  $O$  is necessarily on the segment between  $P$  and  $Q$ . Proof: If this were not the case, then either  $P$  would be on the segment  $\overline{OQ}$  or  $Q$  would be on the segment  $\overline{OP}$  (draw the picture). Say  $P$  is on the segment  $\overline{OQ}$ . Then

$$d(O, P) + d(P, Q) = d(O, Q),$$

whence  $d(P, Q) = 0$  and  $P = Q$ , contrary to our assumption that  $P$  and  $Q$  are distinct. The case when  $Q$  might be on the segment  $\overline{OP}$  is proved similarly.

The next result is an important consequence of the Pythagoras theorem.

**Corollary.** Let  $P, Q$  be distinct points in the plane. Let  $M$  be also a point in the plane. We have

$$d(P, M) = d(Q, M)$$

if and only if  $M$  lies on the perpendicular bisector of  $\overline{PQ}$ .

**Proof.** Let  $L$  be the line passing through  $P, Q$  and let  $K$  be the line perpendicular to  $L$ , passing through  $M$ . Let  $O$  be the point of intersection of  $K$  and  $L$ . In Fig. 5-28(a), we show the case when  $K$  is the perpendicular bisector of  $\overline{PQ}$ , and in Fig. 5-28(b) we show the case when it is not.

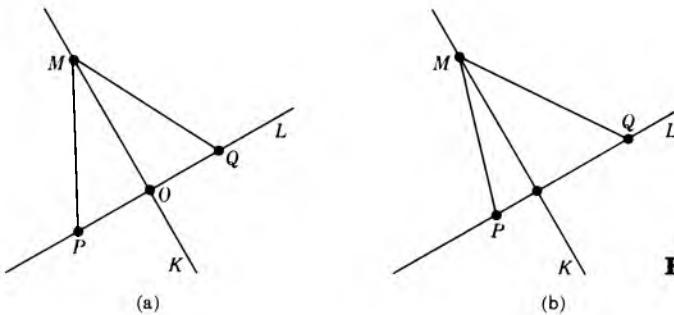


Fig. 5-28

Assume first that  $d(P, M) = d(Q, M)$ . By Pythagoras, we have

$$\begin{aligned} d(P, O)^2 + d(O, M)^2 &= d(P, M)^2 \\ &= d(Q, M)^2 \\ &= d(Q, O)^2 + d(O, M)^2. \end{aligned}$$

It follows that

$$d(P, O)^2 = d(Q, O)^2,$$

whence  $d(P, O) = d(Q, O)$ , and  $M$  is on the perpendicular bisector of  $\overline{PQ}$ .

Conversely, assume that  $d(P, O) = d(Q, O)$ . Similar steps show that

$$d(P, M) = d(Q, M),$$

thus proving our corollary.

## EXERCISES

1. What is the length of the diagonal of a square whose sides have length  
 a) 2,              b) 3,              c) 4,              d) 5,              e)  $r$ ?
2. What is the length of the diagonal of a rectangle whose sides have lengths  
 a) 1 and 2,              b) 3 and 5,              c) 4 and 7,  
 d)  $r$  and  $2r$ ,              e)  $3r$  and  $5r$ ,              f)  $4r$  and  $7r$ ?
3. What is the length of the diagonal of a cube whose sides have length  
 a) 1,              b) 2,              c) 3,              d) 4,              e)  $r$ ?

[Hint: First compute the square of the length. Consider the diagonal of the base square of the cube, and apply the Pythagoras theorem twice.]

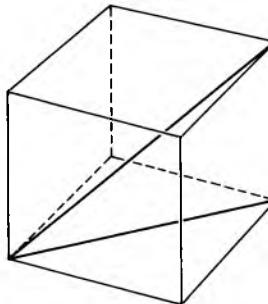


Fig. 5-29

4. What is the length of the diagonal of a rectangular solid whose sides have lengths  
 a) 3, 4, 5;      b) 1, 2, 4;      c) 2, 3, 5;      d) 1, 3, 4;      e) 1, 3, 5?
5. What is the length of the diagonal of a rectangular solid whose sides have lengths  $a$ ,  $b$ ,  $c$ ? What if the sides have lengths  $ra$ ,  $rb$ ,  $rc$ ?

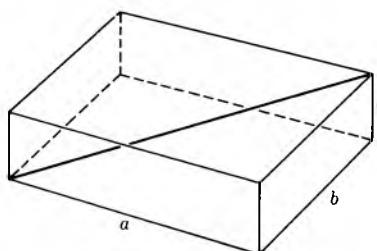


Fig. 5-30

6. You stand at a distance of 500 ft from a tower, and the tower is 100 ft high. What is the distance between you and the top of the tower?
7. a) In a right triangle, one side has length 7 ft and the hypotenuse has length 10 ft. What is the length of the other side?  
 b) Same question if one side has length 11 ft and the hypotenuse has length 17 ft.  
 c) Same question if one side has length 6 ft and the hypotenuse has length 13 ft.
8. a) You are flying a kite. Assume that the string between you and the kite forms a straight line segment. Suppose that the string has length 70 ft. A friend of yours stands exactly below the kite, and is at a distance of 30 ft from you. How high is the kite?  
 b) Same question if the length of the string is 50 ft, and if the distance between you and your friend is 30 ft.  
 c) Same question if the length of the string is 110 ft, and the distance between you and your friend is 40 ft.
9. Write down in detail the “similar steps” left to the reader in the proof of the corollary to the Pythagoras theorem.
10. Prove that if  $A$ ,  $B$ ,  $C$  are the angles of an arbitrary triangle, then

$$m(A) + m(B) + m(C) = 180^\circ$$

by the following method: From any vertex draw the perpendicular to the line of the opposite side. Then use the result already known for right triangles. Distinguish the two pictures in Fig. 5-31.

11. Show that the area of an arbitrary triangle of height  $h$  whose base has length  $b$  is  $bh/2$ . [Hint: Decompose the triangle into two right triangles. Distinguish between the two pictures in Fig. 5-31. In one case the area of the triangle is the difference of the areas of two right triangles, and in the other case, it is the sum.]

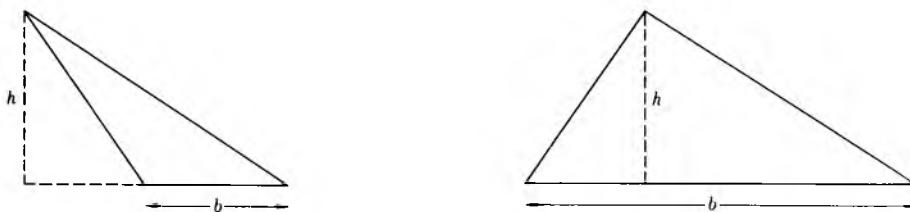


Fig. 5-31

12. a) Show that the length of the hypotenuse of a right triangle is  $\geq$  the length of a leg.
- b) Let  $P$  be a point and  $L$  a line. Show that the smallest value for the distances  $d(P, M)$  between  $P$  and points  $M$  on the line is the distance  $d(P, Q)$ , where  $Q$  is the point of intersection between  $L$  and the line through  $P$ , perpendicular to  $L$ .

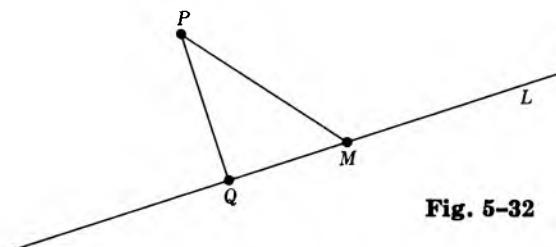


Fig. 5-32

13. This exercise asks you to derive some standard properties of angles from elementary geometry. They are used very commonly. We refer to the following figures.

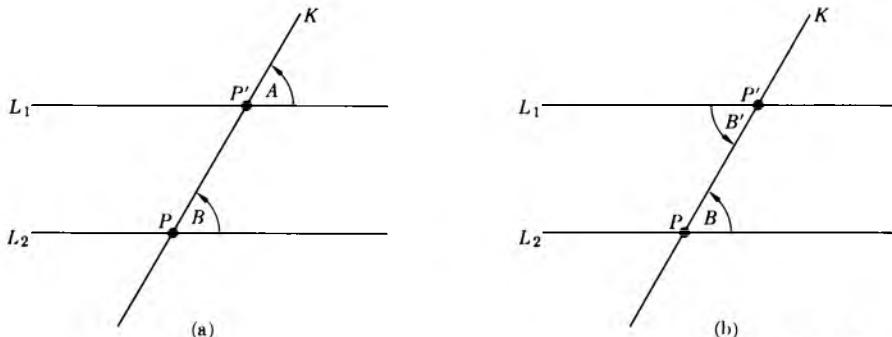


Fig. 5-33

- a) In Fig. 5-33(a), you are given two parallel lines  $L_1, L_2$  and a line  $K$  which intersects them at points  $P$  and  $P'$  as shown. Let  $A$  and  $B$  then be angles which  $K$  makes with  $L_1$  and  $L_2$  respectively, as shown. Prove that

$$m(A) = m(B).$$

[Hint: Draw a line from a point of  $K$  above  $L_1$  perpendicular to  $L_1$  and  $L_2$ . Then use the fact that the sum of the angles of a right triangle has  $180^\circ$ .]

- b) In Fig. 5-33(b), you are given  $L_1$ ,  $L_2$  and  $K$  again. Let  $B$  and  $B'$  be the alternate angles formed by  $K$  and  $L_1$ ,  $L_2$  respectively, as shown. Prove that  $m(B) = m(B')$ . (Actually, all you need to do here is refer to the appropriate portion of the text. Which is it?)
- c) Let  $K, L$  be two lines as shown on Fig. 5-33(c). Prove that the opposite angles  $A$  and  $A'$  as shown have equal measure.

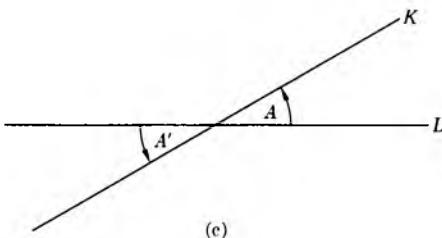


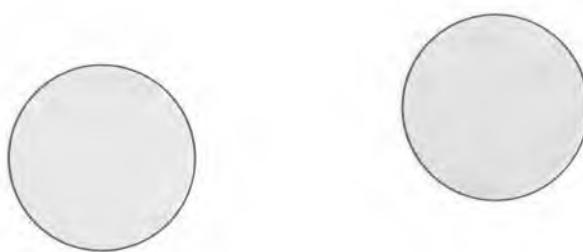
Fig. 5-33 (cont.)

14. Let  $\triangle PQM$  be a triangle. Let  $L_1$  be the perpendicular bisector of  $\overline{PQ}$  and let  $L_2$  be the perpendicular bisector of  $\overline{QM}$ . Let  $O$  be the point of intersection of  $L_1$  and  $L_2$ . Show that  $d(O, P) = d(O, M)$ , and hence that  $O$  lies on the perpendicular bisector of  $\overline{PM}$ . Thus the perpendicular bisectors of the sides of the triangle meet in a point.

# **6 Isometries**

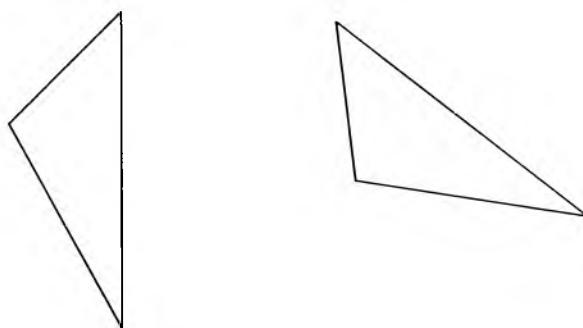
## **§1. SOME STANDARD MAPPINGS OF THE PLANE**

We need to define the notion of congruence. For instance, given two discs of the same radius as in Fig. 6–1, we want to say that they are congruent.



**Fig. 6–1**

Similarly, given two triangles as in Fig. 6–2, we also want to say that they are congruent.



**Fig. 6–2**

Roughly speaking, this means that one figure can be laid over the other. To discuss the notion of congruence properly, it is convenient to define first a slightly more general notion, namely isometry. To do that, we must define still a more general notion, namely mapping. All these notions are quite common, and you will see that they include, as special cases, things which you can easily visualize, like reflections, rotations, stretching, etc., given as examples. We discuss these first, and then take up congruences in the last section.

By a **mapping** (or a **map**) of the plane into itself, we shall mean an association, which to each point of the plane associates another point of the plane. If  $P$  is a point and  $P'$  is the point associated with  $P$  by the mapping, then we denote this by the special arrow

$$P \mapsto P'.$$

The point  $P'$  associated with  $P$  is called the **value** of the mapping at  $P$ . We also say that  $P'$  **corresponds** to  $P$  under the mapping, or that  $P$  is **mapped** on  $P'$ .

Just as we used letters to denote numbers, it is useful to use letters to denote mappings. Thus if  $F$  is a mapping of the plane into itself, we denote the value of  $F$  at  $P$  by the symbols

$$F(P).$$

We shall also say that the value  $F(P)$  of  $F$  at  $P$  is the **image** of  $P$  under  $F$ . If  $F(P) = P'$ , then we also say that  $F$  **maps**  $P$  on  $P'$ .

*By definition, if  $F, G$  are mappings of the plane into itself, we have*

$$F = G$$

*if and only if, for every point  $P$ ,*

$$F(P) = G(P).$$

In other words, a map  $F$  is equal to a map  $G$  if and only if  $F$  and  $G$  have the same value at every point  $P$ .

### Constant mapping

Let  $O$  be a given point in the plane. To each point  $P$  we associate this given point  $O$ . Then we obtain a mapping, and  $O$  is the value of the mapping at every point  $P$ . We say that this mapping is **constant**, and that  $O$  is its constant value.

### Identity

To each point  $P$  we associate  $P$  itself. This is a rather simple mapping, which is called the **identity**, and is denoted by  $I$ . Thus we have

$$I(P) = P$$

for every point  $P$ .

### Reflection through a line

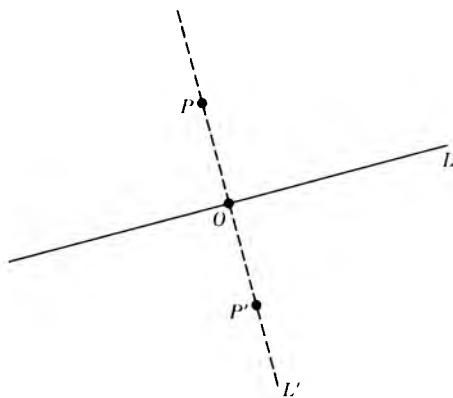
Let  $L$  be a line. If  $P$  is any point, let

$$L' = L'_P$$

be the line through  $P$  perpendicular to  $L$ . Let  $O$  be the point of intersection of  $L$  and  $L'$ . Let  $P'$  be the point on  $L'$  which is at the same distance from  $O$  as  $P$ , but in the opposite direction. The association

$$P \mapsto P'$$

is called **reflection through  $L$** , and could be denoted by  $R_L$ . Picture:



**Fig. 6-3**

### Reflection through a point

Let  $O$  be a given point of the plane. To each point  $P$  of the plane we associate the point  $P'$  lying on the line passing through  $P$  and  $O$ , on the other side of  $O$  from  $P$ , and at the same distance from  $O$  as  $P$ . This mapping is

called the **reflection through  $O$** . Picture:

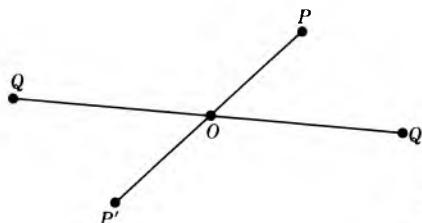


Fig. 6-4

We have drawn a point  $P$  and its value  $P'$  under the mapping, and also a point  $Q$  and its value  $Q'$  under the mapping.

For instance, we may reflect the four corners of a rectangle through the midpoint  $O$  of the rectangle. Each corner is mapped on the opposite corner, as in Fig. 6-5.

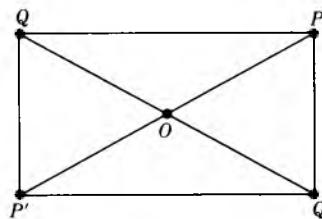


Fig. 6-5

What happens to the sides of the rectangle when we reflect them through  $O$ ? Figure it out, and then look ahead at Theorem 1 of §2.

### Dilations, or stretching

Let  $O$  be a given point of the plane. To each point  $P$  of the plane we associate the point  $P'$  lying on the ray  $RO_P$ , with vertex  $O$ , passing through  $P$ , at a distance from  $O$  equal to twice that of  $P$  from  $O$ . The point  $P'$  in this case is also denoted by  $2P$ . Picture:

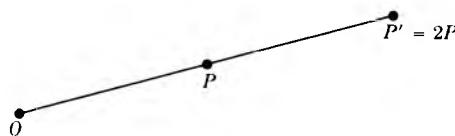


Fig. 6-6

This particular mapping is called **dilation** by 2, or **stretching** by 2, **relative to**  $O$ . According to our notation, if  $F$  is dilation by 2, relative to  $O$ , then we have

$$F(P) = 2P.$$

Let  $r$  be any positive number. We can of course define the dilation by  $r$  just as we define dilation by 2, relative to a given point  $O$ . Namely, we define **dilation by  $r$**  to be the mapping  $F_r$ , such that for any point  $P$  the value  $F_r(P)$  is the point on the ray with vertex  $O$ , passing through  $P$ , at a distance from  $O$  equal to  $r$  times the distance between  $O$  and  $P$ . It is convenient to write  $rP$  instead of  $F_r(P)$ .

In Fig. 6-7(a) we have drawn the points

$$Q \text{ and } F_3(Q).$$

In Fig. 6-7(b) we have drawn the points

$$P, F_{1/3}(P), F_{2/3}(P).$$

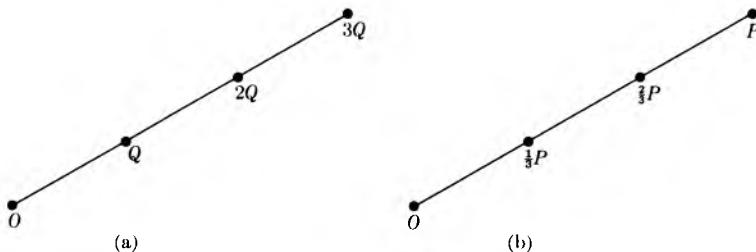


Fig. 6-7

A dilation is also sometimes called a **similarity transformation**, but the word dilation is the shortest and best term to be used to denote the concept.

### Rotation

Let  $O$  be a given point in the plane, and let  $A$  be an angle. Let  $P$  be a point at distance  $d$  from  $O$ . Let  $C$  be the circle of radius  $d$  centered at  $O$ . Let  $P'$  be the point on this circle such that the angle  $\angle POP'$  has the same number of degrees as  $A$ . The mapping which associates  $P'$  to  $P$  is called **rotation** (counterclockwise) by  $A$ , with respect to  $O$ , or relative to  $O$ . If

we denote this mapping by  $G_A$ , then we can illustrate it on Fig. 6-8(a) where we have drawn  $O$ ,  $P$ , and  $G_A(P)$ .

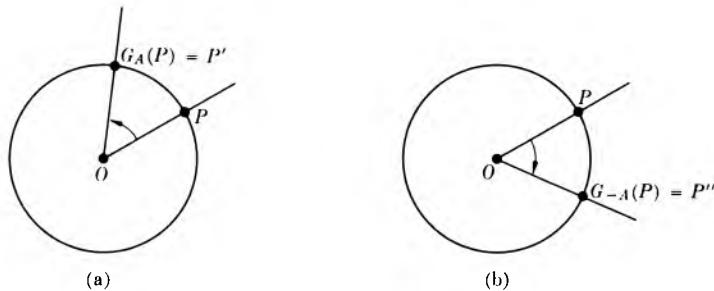


Fig. 6-8

Unless otherwise specified, a rotation by an angle  $A$  will always mean counter-clockwise rotation. Of course, we can also define clockwise rotation by  $A$ , which we denote by  $G_{-A}$ . It associates with each point  $P$  the point  $P''$  at the same distance from  $O$  as  $P$ , and such that the angle

$$\angle P''OP$$

has the same measure as  $A$ , as on Fig. 6-8(b).

Observe that a given angle  $A$  is not necessarily the same angle as that formed by

$$\angle POP'$$

(cf. Fig. 6-9) even though they have the same measure. However, experience shows that it is harmless for this type of discussion to indicate  $\angle POP'$  as the

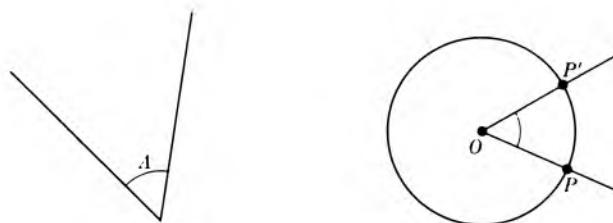


Fig. 6-9

angle  $A$ , and it is often convenient to do this to suggest what's going on. Still, it is not always harmless to do this. For instance, in a triangle like that in

Fig. 6-10 where the two bottom angles  $A$  and  $B$  have equal measures, we would not think of denoting them by the same letter.

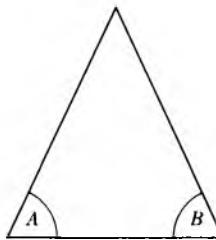


Fig. 6-10

Observe that rotation by  $180^\circ$  with respect to  $O$  is none other than reflection through  $O$ . Thus if  $R$  denotes a reflection through  $O$ , we have

$$G_{180^\circ} = R.$$

We also have

$$G_{-180^\circ} = R,$$

because the values of each one of these mappings at a point  $P$  is the same point  $P'$ . Even though these mappings are described by conditions which appear different, the mappings are nevertheless equal. Recall that, by definition, mappings  $F, G$  are equal if and only if

$$F(P) = G(P)$$

for all points  $P$ .

It is convenient to associate a rotation with a number rather than an angle. If  $x$  is a number between 0 and  $360$ , we let

$$G_x$$

be the rotation by an angle of  $x$  degrees. Observe that

$$G_0 = G_{360} = I$$

is none other than the identity. (To be absolutely correct, we should also indicate the point  $O$  in our notation, but throughout our discussion we deal with the same given point, and thus omit it. If we wanted to indicate it explicitly, we could write for instance

$$G_{O,x}$$

for the rotation by an angle of  $x$  degrees relative to the given point  $O$ .)

Let  $x$  be an arbitrary number. We write  $x$  in the form

$$x = 360n + w,$$

where  $n$  is an integer, and  $w$  is a number such that  $0 \leq w < 360$ . We define  $G_x$  to be

$$G_x = G_w.$$

**Example.** Let  $x = 500$ . We write

$$500 = 360 + 140.$$

Then by definition

$$G_x = G_{140}$$

is rotation by  $140^\circ$ .

**Example.** Let  $x = -210$ . We write

$$-210 = -360 + 150.$$

Then

$$G_{-210} = G_{150}$$

is rotation by  $150^\circ$ .

**Example.** If  $x$  and  $y$  are numbers such that

$$x = y + 360m$$

for some integer  $m$ , then

$$G_x = G_y.$$

Namely, if  $x = 360n + w$  with some integer  $n$ , and  $0 \leq w < 360$ , then

$$y + 360m = 360n + w,$$

and hence

$$y = 360(n - m) + w.$$

According to our definition, we have

$$G_x = G_w = G_y.$$

We can interpret counterclockwise rotation by a negative number as clockwise rotation by a positive number.

**Example.** Let  $x = -90$ . Then

$$-90 = -360 + 270.$$

Thus

$$G_{-90} = G_{270}.$$

We visualize this as saying that clockwise rotation by  $90^\circ$  is the same as counterclockwise rotation by  $270^\circ$ .

We shall use the convention of saying that the rotation  $G_x$  is rotation by  $x$  degrees, or even by an angle of  $x$  degrees, even though  $x$  may be greater than  $360$ , or may be negative. This is convenient language, and reflects our geometric intuition without leading to great confusion.

### Translations

Let us select a direction in the plane, and a distance  $d$ . We can represent these by an arrow as in Fig. 6-11. The arrow is simply an ordered pair of points  $(O, M)$ , where  $O$  is its beginning point and  $M$  is its end point.

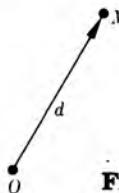


Fig. 6-11

The arrow points in the given direction, and the length of the arrow is equal to our given distance. To each point  $P$ , we associate the point  $P'$  which is at a distance  $d$  from  $P$  in the given direction. This is a mapping, which is called the **translation** (determined by the given direction and the given distance). In Fig. 6-12, letting  $T$  be this translation, we have drawn two points  $P, Q$  and their images under  $T$ .

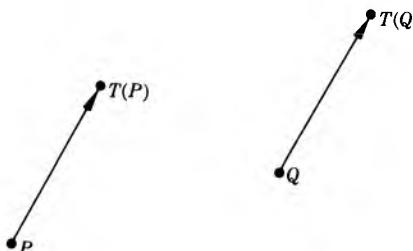


Fig. 6-12

The translation determined by an ordered pair of points  $(O, M)$  will be denoted by

$$T_{OM}.$$

Observe that if  $T = T_{OM}$ , then  $T(O) = M$ , and that  $T_{OO}$  is the identity.

**Example.** Let  $T_l$  be the translation by one inch to the left,  $T_r$  the translation by one inch to the right, and similarly  $T_u$  and  $T_d$  the translations by one inch upward and one inch downward respectively.

In Fig. 6-13 we have drawn a flower  $\mathfrak{F}$  and its translations by  $T_l$ ,  $T_r$ ,  $T_d$  and  $T_u$ .

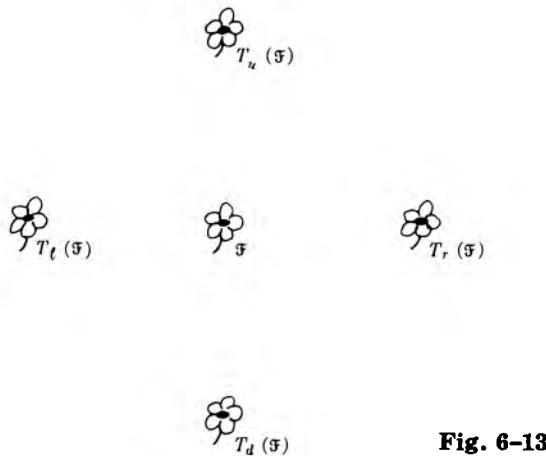


Fig. 6-13

## EXERCISES

Let  $F$  be a mapping of the plane into itself. We define a **fixed point** for  $F$  to be a point  $P$  such that  $F(P) = P$ .

1. Describe the fixed points of the following mappings.
  - a) The identity.
  - b) Reflection through a given point  $O$ .
  - c) Reflection through a line.
  - d) A rotation not equal to the identity, with respect to a given point  $O$ .
  - e) A translation not equal to the identity.
  - f) Dilation by a number  $r > 0$ , relative to a given point  $O$ .

2. Write each one of the following numbers in the form

$$360n + w$$

with an integer  $n$ , and  $0 \leq w < 360$ .

- |         |         |         |         |
|---------|---------|---------|---------|
| a) -30  | b) -90  | c) -180 | d) -270 |
| e) -45  | f) -225 | g) 120  | h) 540  |
| i) -400 | j) 600  | k) 720  | l) 450  |

3. Let the point  $P$  be as illustrated in Fig. 6-14. In each one of the cases of Exercise 2, draw the image of  $P$  under the corresponding rotation, by the number of degrees given in (a) through (l). You can use a protractor, or just make an approximate estimate of the position of this image.



Fig. 6-14

If you want to make things look better, draw the image of a flower instead of a point.

## §2. ISOMETRIES

Let  $F$  be a mapping of the plane into itself. We say that  $F$  **preserves distances**, or is **distance preserving**, if and only if for every pair of points  $P, Q$  in the plane, the distance between  $P$  and  $Q$  is the same as the distance between  $F(P)$  and  $F(Q)$ . Such a mapping is also called an **isometry**. ("Iso" means same, and "metry" means measure. It is useful to have one word instead of two for this notion.)

**Example.** The constant mapping which to each point  $P$  associates a given point  $O$  is not an isometry. Dilation by 2 is not an isometry. Why?

**Example.** Let  $F$  be any one of the following maps.

*Reflection through a point*

*Reflection through a line*

*Rotation*

*Translation*

*Then  $F$  is an isometry.*

This will be assumed without proof. Later when we give definitions for these mappings depending on coordinates, we shall be able to prove that these maps are isometries very simply. In §5 we shall prove that any isometry can be obtained by simple combinations of the examples given above, in a sense which will be made precise.

**Remark.** Let  $F$  be an isometry. If  $P, Q$  are distinct points, then  $F(P)$  and  $F(Q)$  must be distinct, because the distance between  $P$  and  $Q$  is not 0, and hence the distance between  $F(P)$  and  $F(Q)$  cannot be 0 either. Cf. property **DIST 1** of distances.

Let  $S$  be a set of points in the plane and let  $F$  be a mapping of the plane into itself. The set of points consisting of all points  $F(P)$ , for all  $P$  in  $S$ , is called the **image of  $S$  under  $F$** , and is denoted by  $F(S)$ .

**Example.** Let  $F$  be the reflection through a line  $L$ . Let  $S$  be the line segment between two points  $P$  and  $Q$ . Then the image of  $S$  under  $F$  is the line segment between  $F(P)$  and  $F(Q)$ . Picture:

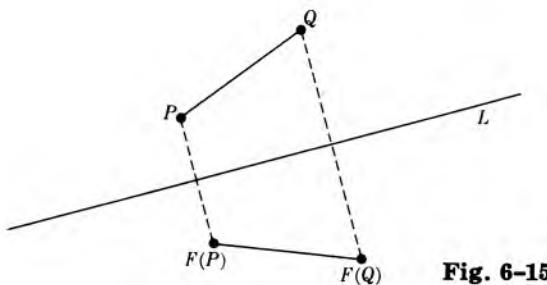


Fig. 6-15

With our notation, we have

$$F(\overline{PQ}) = \overline{F(P)F(Q)}.$$

The preceding example is but a special case of a general property of isometries, which we state in the next theorem.

**Theorem 1.** Let  $F$  be an isometry. The image of a line segment under  $F$  is a line segment. In fact, the image of the line segment  $\overline{PQ}$  under  $F$  is the line segment between  $F(P)$  and  $F(Q)$ .

*Proof.* (See Fig. 6-16.) Let  $X$  be a point on  $\overline{PQ}$ . For simplicity we denote  $F(X)$  by  $X'$ . Since  $F$  preserves distances, we know that

$$d(P, X) = d(P', X'), \quad d(X, Q) = d(X', Q').$$

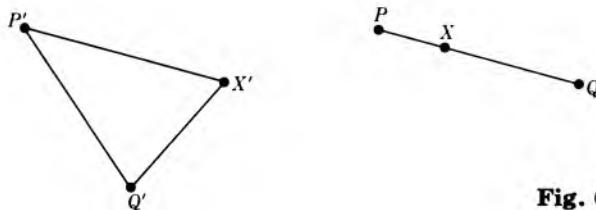


Fig. 6-16

By **SEG 1**, we have

$$d(P, Q) = d(P, X) + d(X, Q).$$

By assumption on  $F$ , we have

$$d(P', Q') = d(P', X') + d(X', Q').$$

Again by **SEG 1**, we conclude that  $X'$  must lie on the segment between  $P'$  and  $Q'$ , thus proving that the image of  $\overline{PQ}$  is contained in the segment  $\overline{P'Q'}$ .

We must still prove that every point of the segment  $\overline{P'Q'}$  can be expressed as the image under  $F$  of a point on  $\overline{QP}$ . Let  $X'$  be a point on  $\overline{P'Q'}$  at distance  $r$  from  $P'$ . Let  $X$  be the point on  $\overline{PQ}$  at distance  $r$  from  $P$ . Then  $F(X)$  is at distance  $r$  from  $F(P) = P'$ . It follows that  $F(X) = X'$ .

**Remark.** In this proof, we want to show that two sets of points are equal. We have followed a standard pattern, namely we have proved that each one is part of the other. This pattern will be repeated later.

**Corollary.** *An isometry preserves straight lines. In other words, if  $L$  is a straight line in the plane and  $F$  is an isometry, then  $F(L)$  (the image of  $L$  under  $F$ ) is also a straight line. If  $L$  is the line passing through two distinct points  $P, Q$ , then  $F(L)$  is the line passing through  $F(P)$  and  $F(Q)$ .*

*Proof.* We leave the proof as an exercise.

**Example.** Let  $O$  be the point of intersection of the diagonals of a rectangle. If we reflect through  $O$ , then the opposite corners are mapped on each other,

and hence the opposite sides are mapped on each other, as illustrated in Fig. 6-17.

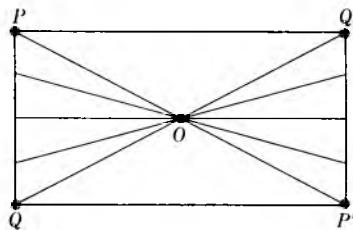


Fig. 6-17

Let  $F$  be a mapping of the plane into itself. We recall that a **fixed point**  $P$  of  $F$  is a point such that  $F(P) = P$ . Fixed points of isometries will now play a very important role in describing all isometries. You should definitely do Exercise 2 of the preceding section if you have not already done it. We shall investigate systematically isometries with no fixed point, one fixed point, two fixed points, and three fixed points, but in reverse order. In this last case, we shall see that the isometry must be the identity. Then we consider each case with one less fixed point, and analyze it by composing the given isometry with a reflection, a rotation, or a translation to get the ultimate result that any isometry must be a composite of these. (For the definition of a composite isometry, read ahead in §3.)

**Theorem 2.** *Let  $F$  be an isometry. Let  $P, Q$  be two distinct points in the plane. Assume that they are fixed points, in other words*

$$F(P) = P \quad \text{and} \quad F(Q) = Q.$$

*Then every point on the line through  $P, Q$  is a fixed point of  $F$ .*

**Proof.** We shall distinguish cases. Let  $M$  be a point on the line passing through  $P$  and  $Q$ . We wish to show that  $F(M) = M$ .

**Case 1.** The point  $M$  lies on the segment  $\overline{PQ}$ . Let  $M' = F(M)$ . Since  $F$  preserves distances, we have

$$\begin{aligned} d(P, M) &= d(P, M'), \\ d(M', Q) &= d(M, Q). \end{aligned}$$

Hence

$$d(P, M') + d(M', Q) = d(P, Q).$$

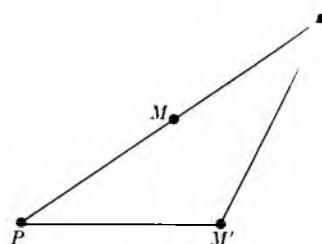


Fig. 6-18

By **SEG 1**, this means that  $M'$  lies on the segment between  $P$  and  $Q$ . Since

$$d(P, M) = d(P, M'),$$

it follows that  $M = M'$ .

**Case 2.** Suppose that  $M$  does not lie on the segment  $\overline{PQ}$ . Suppose that  $M$  lies on the ray having vertex  $P$  and passing through  $Q$ , but at a distance from  $P$  greater than that of  $Q$ , as in Fig. 6-19.

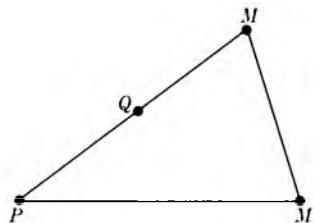


Fig. 6-19

Then

$$\begin{aligned} d(P, M') &= d(P, M) = d(P, Q) + d(Q, M) \\ &= d(P, Q) + d(Q, M'). \end{aligned}$$

By **SEG 1**, this means that  $Q$  lies on the segment between  $P$  and  $M'$ . Hence  $P, Q, M'$  lie on the same straight line, and therefore  $M'$  lies on the straight line passing through  $P, Q$ . Since  $Q$  lies on the segment between  $P$  and  $M'$ , we conclude that  $M'$  lies on the ray having vertex  $P$  passing through  $Q$ . Since

$$d(P, M) = d(P, M'),$$

it follows that  $M = M'$ .

**Case 3.** This case is similar to Case 2, when  $M$  lies on the other side of  $P$  from  $Q$ . In this case, the role of  $P$  and  $Q$  is reversed, and the proof goes on as in Case 2, interchanging  $P$  and  $Q$ . This concludes the proof of Theorem 2.

**Theorem 3.** Let  $F$  be an isometry. Let  $P, Q, M$  be three distinct points which do not lie on a straight line. Assume that  $P, Q, M$  are fixed points of  $F$ ; that is,

$$F(P) = P, \quad F(Q) = Q, \quad F(M) = M.$$

Then  $F$  is the identity.

*Proof.* Let  $L_{PQ}$  and  $L_{QM}$  be the lines passing through  $P, Q$  and  $Q, M$ , respectively. Let  $X$  be a point. We must show that  $F(X) = X$ . We can find a line  $L$  passing through  $X$  which intersects  $L_{PQ}$  in a point  $Z$ , and intersects  $L_{QM}$  in a point  $Y$  such that  $Y \neq Z$ . (For instance, pick a point  $Z$  on  $L_{QM}$  which is distinct from  $M, Q, X$ , and such that the line  $L_{XZ}$  is not parallel to  $L_{PQ}$ . Let  $L = L_{XZ}$ , and let  $Y$  be the point of intersection of  $L_{XZ}$  and  $L_{PQ}$ .) The situation is illustrated in Fig. 6-20.

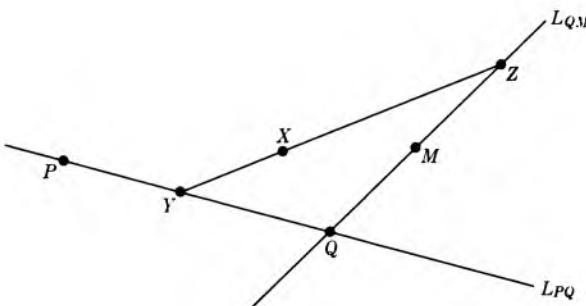


Fig. 6-20

By Theorem 2, every point on the lines  $L_{PQ}$  and  $L_{QM}$  are fixed. Therefore we have

$$F(Y) = Y \quad \text{and} \quad F(Z) = Z.$$

Again by Theorem 2, every point on the line  $L_{YZ}$  is fixed, and we conclude that  $F(X) = X$ . This proves our theorem.

**Remark.** A very important corollary of this theorem will be stated when we have the notion of inverse of an isometry, in §4.

## EXERCISES

1. Draw the image of a line segment under
  - a) reflection through a point,
  - b) reflection through a line,
  - c) rotation by  $90^\circ$ ,

- d) translation,  
e) rotation by  $180^\circ$ .
2. For which values of  $r$  is dilation by  $r$  an isometry?
3. Draw the image of a circle of radius  $r$ , center  $P$  under
- reflection through its center,
  - reflection through a line  $L$  outside of the circle as illustrated in Fig. 6-21,
  - rotation by  $90^\circ$  with respect to a point  $O$  outside of the circle,
  - rotation by  $180^\circ$  with respect to  $O$ ,
  - rotation by  $270^\circ$  with respect to  $O$ ,
  - translation.

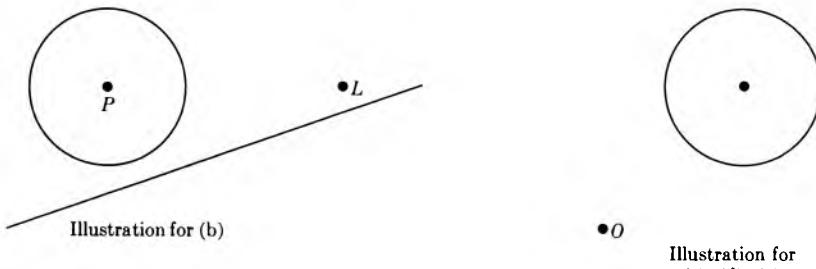


Fig. 6-21

4. Let  $L, K$  be two parallel lines, and let  $F$  be an isometry. Prove that  $F(L)$  and  $F(K)$  are parallel.
5. Let  $K, L$  be perpendicular lines, and let  $F$  be an isometry. Prove that  $F(K)$  and  $F(L)$  are perpendicular. [Hint: Use the corollary of the Pythagoras theorem.]
6. Visualize 3-dimensional space. We also have the notion of distance in space, satisfying the same basic properties as in a plane. We can therefore define an isometry of 3-space in the same way that we defined an isometry of the plane. It is a mapping of 3-space into itself which is distance preserving. Are Theorems 1 and 2 valid in 3-space? How would you formulate Theorem 3? (Consider the plane in which the three points lie.) Now formulate a theorem in 3-space about an isometry being the identity provided that it leaves enough points fixed. Describe a proof for such a theorem, similar to the proof of Theorem 3. Make a list of

what you need to assume to make such a proof go through. Write all of this up as if you were writing a book. Aside from learning mathematical substance, you will also learn how to think more clearly, and how to write mathematics in the process.

### §3. COMPOSITION OF ISOMETRIES

We can take isometries in succession. For instance, we could first rotate the plane through an angle of  $30^\circ$  relative to a given point  $O$ ; then reflect through a given line  $L$ ; then rotate again through an angle of  $45^\circ$ ; finally make a translation. When we take such isometries in succession like that, we say that we **compose** them.

In general, let  $F, G$  be isometries. To each point  $P$  let us associate the point  $F(G(P))$ , obtained by first taking the image of  $P$  under  $G$ , and then the image of this latter point under  $F$ . Then we obtain an association

$$P \mapsto F(G(P)),$$

which is a mapping. In fact, this mapping is an isometry, because the distance between two points  $P, Q$  is the same as the distance between  $G(P), G(Q)$  (because  $G$  is an isometry), and is the same as the distance between  $F(G(P))$ ,  $F(G(Q))$  (because  $F$  is an isometry). The association

$$P \mapsto F(G(P))$$

is called the **composite** of  $G$  and  $F$ , and is denoted by the symbols

$$F \circ G.$$

Thus we have

$$(F \circ G)(P) = F(G(P)).$$

**Example.** Let  $O$  be a given point. Let  $F$  be rotation by  $90^\circ$  with respect to  $O$ , and let  $G$  be reflection through  $O$ . Then  $G \circ F$  is rotation by  $270^\circ$ . We also see that  $F \circ F$  is rotation by  $180^\circ$ , and thus we may write

$$F \circ F = G.$$

**Example.** Let  $F$  be any isometry and let  $I$  be the identity. Then

$$\begin{aligned} F \circ I &= I \circ F \\ &= F. \end{aligned}$$

Thus  $I$  behaves like multiplication by 1.

**Composition of rotations.** If  $F, G$  are rotations, relative to the same point  $O$ , then  $F \circ G$  is also a rotation, relative to  $O$ .

Let  $O$  be a given point and let  $r$  be a number  $> 0$ . Let  $P, Q$  be points different from  $O$ , at the same distance  $r$  from  $O$ . Then there exists a unique rotation  $G_{PQ}$  relative to  $O$ , which maps  $P$  on  $Q$  (i.e. such that the value of  $G_{PQ}$  at  $P$  is  $Q$ ).

We shall assume these statements without proof. They are both intuitively clear. Using them, we can write down a nice formula for the composite. Let  $P, Q, M$  be points at distance  $r$  from  $O$ . Then

$$G_{QM} \circ G_{PQ} = G_{PM} \quad \text{and} \quad G_{PP} = I.$$

*Proof.* The image of  $P$  under the composite  $G_{QM} \circ G_{PQ}$  is

$$\begin{aligned} G_{QM}(G_{PQ}(P)) &= G_{QM}(Q) \\ &= M, \end{aligned}$$

which is the same as the image of  $P$  under  $G_{PM}$ . By assumption there is only one rotation having this effect on  $P$ . Hence we get the first formula. The second is proved similarly, and even more simply.

In terms of numbers, we shall also assume without proof the following fact, which is intuitively clear.

Let  $x, y$  be numbers. Let  $G_x$  be the rotation associated with  $x$  as in §1. Then

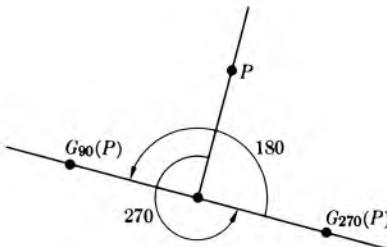
$$G_x \circ G_y = G_{x+y}.$$

For example,

$$G_{45} \circ G_{45} = G_{90}.$$

This means that a rotation by  $45^\circ$  followed by a rotation by  $45^\circ$  is the same as a rotation by  $90^\circ$ . Also, as in Fig. 6-22, we have

$$G_{180} \circ G_{270} = G_{450} = G_{90}.$$



**Fig. 6-22**

**Composition of translations.** If  $F, G$  are translations, then the composite  $F \circ G$  is also a translation.

Given points  $P, Q$ , there exists a unique translation  $T_{PQ}$  such that the image of  $P$  under  $T_{PQ}$  is  $Q$ .

Again, we assume these two statements without proof. As an exercise, prove the following formulas.

$$T_{QM} \circ T_{PQ} = T_{PM} \quad \text{and} \quad T_{PP} = I.$$

**Associativity of isometries.** Let  $F, G, H$  be isometries. Then we have

$$(F \circ G) \circ H = F \circ (G \circ H).$$

*Proof.* For any point  $P$ , we have

$$\begin{aligned} ((F \circ G) \circ H)(P) &= (F \circ G)(H(P)) = F(G(H(P))) \\ (F \circ (G \circ H))(P) &= F((G \circ H)(P)) = F(G(H(P))). \end{aligned}$$

This proves our remark, because the two maps  $(F \circ G) \circ H$  and  $F \circ (G \circ H)$  have the same value at  $P$ , and this is true for every point  $P$ .

We shall use the same notation with isometries that we used with numbers for multiplication. If  $F$  is an isometry,

$$\begin{aligned} \text{we denote } F \circ F \text{ by } F^2, \\ \text{we denote } F \circ F \circ F \text{ by } F^3, \end{aligned}$$

and so on. We denote by  $F^n$  the isometry obtained by iterating  $F$  with itself  $n$  times.

**Example.** Let  $G$  be the reflection through a given point  $O$ . Then we see that

$$G^2 = \text{identity} = I.$$

This is like the relation  $(-1)^2 = 1$ . Note that we have:

$$\begin{aligned} G^3 &= G \\ G^4 &= I \\ &\vdots \end{aligned}$$

again in analogy with powers of  $-1$ .

**Example.** Let  $F$  be rotation by  $90^\circ$ . Then:

$$\begin{aligned} F^2 &= \text{rotation by } 180^\circ, \\ F^3 &= \text{rotation by } 270^\circ, \\ F^4 &= \text{rotation by } 360^\circ = \text{identity}, \\ F^5 &= \text{rotation by } 90^\circ = F. \end{aligned}$$

Note this interesting cyclical nature of  $F$ , that  $F^5 = F$ .

If  $F$  is an isometry, we define

$$F^0 = I.$$

Then for any natural numbers  $m, n$  we have the old relation

$$F^{m+n} = F^m \circ F^n.$$

Thus composition behaves like a multiplication.

**Example.** Let  $T$  be translation by 1 in. to the right, and let  $P$  be a point. Then  $P, T(P), T^2(P), T^3(P), \dots$  are points on a horizontal line, and  $T^{n+1}(P)$  is 1 in. to the right of  $T^n(P)$ . Draw one picture, with  $0 \leq n \leq 5$ .

## EXERCISES

1. Let  $F$  be reflection through a line  $L$ . What is the smallest positive integer  $n$  such that  $F^n = I$ ?

In Exercises 2 and 3, let  $O$  be a given point in the plane. Let  $K$  be a vertical line and  $L$  a horizontal line intersecting at  $O$ . Let  $H$  be reflection through  $L$  and  $V$  reflection through  $K$ . Let  $G_x$  be rotation with respect to  $O$ , by an angle of  $x$  degrees.

2. Let  $P$  be the point shown in Fig. 6-23. Draw the image of  $P$  under the following isometries.

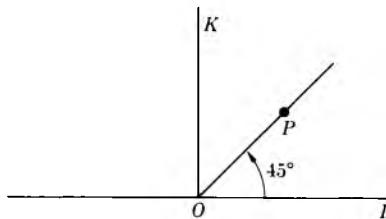


Fig. 6-23

- a)  $H \circ G_{90}$
- b)  $G_{90} \circ H$
- c)  $V \circ G_{90}$
- d)  $G_{90} \circ V$
- e)  $H \circ V \circ G_{90}$
- f)  $V \circ H \circ G_{90}$
- g)  $H \circ G_{180}$
- h)  $G_{180} \circ H$
- i)  $G_{180} \circ V$
- j)  $V \circ G_{180}$
- k)  $H \circ V \circ G_{180}$
- l)  $V \circ H \circ G_{180}$

3. Let  $Q$  be the point shown in Fig. 6-24. Draw the image of  $Q$  under each one of the mappings (a) through (l) of the preceding exercise.

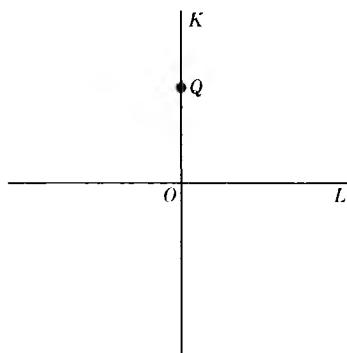


Fig. 6-24

4. Give an example of two isometries  $F_1, F_2$  such that

$$F_1 \circ F_2 \neq F_2 \circ F_1.$$

5. Let  $G$  be rotation by

- a)  $90^\circ$ ,
- b)  $60^\circ$ ,
- c)  $45^\circ$ ,
- d)  $30^\circ$ ,
- e)  $15^\circ$ .

In each case, determine the smallest positive integer  $k$  such that  $G^k = I$ .

6. Draw a small flower. Let  $T$  be translation by 1 in. to the right, and let  $U$  be translation by 1 in. vertically upward. Draw the image of the flower under  $T, T^2, T^3, T^4, U, U^2, U^3, U^4, T \circ U, T^2 \circ U, T^3 \circ U, T \circ U^2, T \circ U^3, T^2 \circ U^2$ . Admire your pattern. Draw other images of the flower under isometries to make up other beautiful patterns.

#### §4. INVERSE OF ISOMETRIES

Let  $F$  be an isometry. By an **inverse** (isometry) for  $F$  we shall mean an isometry  $G$  such that

$$F \circ G = G \circ F = I.$$

Suppose that  $G$  and  $H$  are inverses for  $F$ . Then

$$H \circ F \circ G = H \circ I = H.$$

By associativity, the left-hand side is equal to

$$H \circ F \circ G = I \circ G = G.$$

Thus we find

$$G = H.$$

This is the same type of proof which we used before to prove the uniqueness of the inverse, and we see that it applies to our present setting with isometries. We denote the inverse of  $F$  by  $F^{-1}$  if it exists. Assume that the inverse exists. If  $P, Q$  are points, then the relations

$$P = F(Q) \quad \text{and} \quad Q = F^{-1}(P)$$

are equivalent. Indeed, if  $P = F(Q)$ , then applying  $F^{-1}$  we obtain

$$F^{-1}(P) = F^{-1}(F(Q)) = Q,$$

and similarly for the converse. If  $P$  is the image of  $Q$  under  $F$ , then we also say that  $Q$  is the **inverse image** of  $P$  under  $F$ .

**Example.** Let  $F$  be reflection through a given line  $L$ . Since

$$F^2 = F \circ F = I,$$

we conclude that

$$F^{-1} = F.$$

**Example.** For each number  $x$  let  $G_x$  be the associated rotation through  $x$  degrees. Then

$$G_x^{-1} = G_{-x},$$

because  $G_{-x} \circ G_x = G_0 = I$ . For instance,

$$G_{90}^{-1} = G_{270} = G_{-90}.$$

Also observe that if  $G = G_{90}$ , then

$$G^{-1} = G^3.$$

**Example.** Let  $T_{OM}$  denote the translation determined by the ordered pair of points  $(O, M)$ . This is the translation in the direction of the ray with vertex  $O$ , passing through  $M$ , and such that the image of a point  $P$  lies at distance from  $P$  equal to the length of the segment  $\overline{OM}$ . Then  $T_{OM}$  has an inverse, which is none other than  $T_{MO}$ , namely the translation going in the opposite direction, but the same distance, because we have

$$T_{MO} \circ T_{OM} = I.$$

**Example.** Let  $F, G$  be isometries, having inverses  $F^{-1}$  and  $G^{-1}$ , respectively. Then the composite  $F \circ G$  has an inverse, namely

$$(F \circ G)^{-1} = G^{-1} \circ F^{-1}.$$

This is easily seen. All we have to do is to verify that the right-hand side composed with  $F \circ G$  on either side yields the identity. But we have

$$G^{-1} \circ F^{-1} \circ F \circ G = G^{-1} \circ I \circ G = G^{-1} \circ G = I,$$

and similarly on the other side. This proves our assertion.

Let  $n$  be a negative integer, say  $n = -k$ , where  $k$  is positive. We define  $F$  to be the composite of  $F^{-1}$  with itself  $k$  times, i.e.

$$F^{-k} = (F^{-1})^k.$$

Also we define  $F^0 = I$  (identity). Then we have the formula

$$F^{m+n} = F^m \circ F^n$$

valid for any values of  $m, n$  as integers. This relation is analogous to that holding for powers of numbers. We omit the proof, which is in any case easy.

**Example.** If  $T$  is the translation by 1 in. to the right, then  $T^{-1}$  is the translation by 1 in. to the left. If  $U$  is the translation by 1 in. upward, then  $U^{-1}$  is the translation by 1 in. downward. Also,  $T^{-5}$  is the translation by 5 in. to the left, and  $U^{-6}$  is the translation by 6 in. downward.

Using inverses, we can now prove a very useful corollary of Theorem 3 which tells us when two isometries are equal.

**Corollary of Theorem 3.** Let  $P, Q, M$  be three distinct points which do not lie on the same line. Let  $F, G$  be isometries such that

$$F(P) = G(P), \quad F(Q) = G(Q), \quad F(M) = G(M).$$

Assume that  $F^{-1}$  exists. Then  $F = G$ .

*Proof.* The proof is very easy and will be left as an exercise.

**Example.** Let  $K$  be a vertical line and let  $L$  be a horizontal line. Let  $H$  be reflection with respect to  $L$  and let  $V$  be reflection with respect to  $K$ . Then  $H \circ V = V \circ H$ . To see this, we have only to verify that  $H \circ V$  and  $V \circ H$  have the same effect on three corners of a square centered at the point of intersection of the lines, and this is clear.

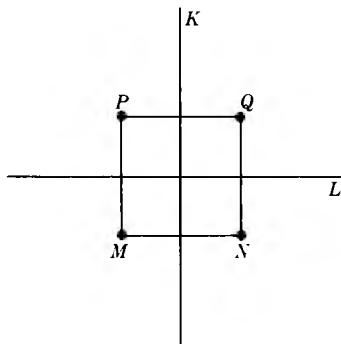


Fig. 6-25

**Remark.** It will be proved later that every isometry has an inverse.

**EXERCISES**

1. a) Let  $F$  be an isometry which has an inverse  $F^{-1}$ . Let  $S$  be a circle of radius  $r$ , and center  $P$ . Show that the image of  $S$  under  $F$  is a circle.  
[Hint: Let  $S'$  be the circle of center  $F(P)$  and radius  $r$ . Show that  $F(S)$  is contained in  $S'$  and that every point of  $S'$  is the image under  $F$  of a point in  $S$ .]  
b) Let  $F$  be an isometry which has an inverse  $F^{-1}$ . Let  $D$  be a disc of radius  $r$  and center  $P$ . Show that the image of  $D$  under  $F$  is a disc.
2. Let  $P, Q, P', Q'$  be points such that

$$d(P, Q) = d(P', Q').$$

Prove that there exists an isometry  $F$  such that  $F(P) = P'$  and  $F(Q) = Q'$ . You may assume the same statements we have assumed in this section.

3. Let  $F, G, H$  be isometries and assume that  $F$  has an inverse. If

$$F \circ G = F \circ H,$$

prove that  $G = H$  (**cancellation law** for isometries).

4. a) Let  $F$  be an isometry such that  $F^2 = I$  and  $F^3 = I$ . Prove that  $F = I$ .  
b) Let  $F$  be an isometry such that  $F^4 = I$  and  $F^7 = I$ . Prove that  $F = I$ .  
c) Let  $F$  be an isometry such that  $F^5 = I$  and  $F^8 = I$ . Prove that  $F = I$ .
5. Write out the proof of the corollary to Theorem 3. (Consider  $F^{-1} \circ G$ .)
6. Let  $F \circ G \circ H$  be the composite of three isometries. Assume that  $F^{-1}, G^{-1}, H^{-1}$  exist. Prove that  $(F \circ G \circ H)^{-1}$  exists, and express this inverse in terms of the inverses for  $F, G, H$ .
7. Let  $F$  be an isometry such that  $F^7 = I$ . Express  $F^{-1}$  as a positive power of  $F$ .
8. Let  $n$  be a positive integer and let  $F$  be an isometry such that  $F^n = I$ . Express  $F^{-1}$  as a positive power of  $F$ .

**For the rest of the exercises, we let  $H$  denote reflection through the horizontal axis, and we let  $V$  denote reflection through the vertical axis.**

9. Consider the corners of a square centered at the origin. For convenience of notation, number these corners 1, 2, 3, 4 as in Fig. 6-26.

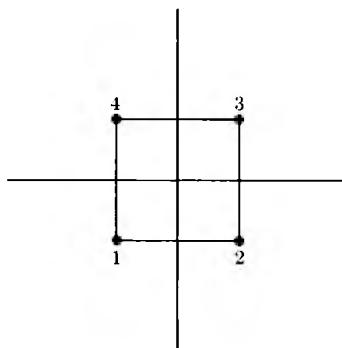


Fig. 6-26

Write the image of each one of these corners under the isometries  $H$ ,  $V$ ,  $H \circ V$  and  $V \circ H$ . Just to show you an easy notation to do this, we write down the images of these corners under rotation by  $90^\circ$  in the following form:

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{bmatrix}.$$

This notation means that if  $G$  is rotation by  $90^\circ$ , then  $G(1) = 2$ ,  $G(2) = 3$ ,  $G(3) = 4$ , and  $G(4) = 1$ .

10. Let  $G$  be rotation by  $90^\circ$  so that  $G^4 = I$ . Express  $H \circ G \circ H$  as a power of  $G$ . For what positive integer  $n$  do we have

$$H \circ G = G^n \circ H?$$

Write down the images of the corner of the square as in the preceding exercise, under the maps  $I$ ,  $G$ ,  $G^2$ ,  $G^3$ ,  $H$ ,  $H \circ G$ ,  $H \circ G^2$ ,  $H \circ G^3$ ,  $G \circ H$ ,  $G^2 \circ H$ ,  $G^3 \circ H$ .

Compare with the section on permutations in Chapter 14, §3.

### Multiplication tables

Let us simplify the notation and write  $FG$  instead of  $F \circ G$ , to make the analogy with multiplication more striking. If  $H$ ,  $V$  are the reflections along the horizontal line and vertical line, respectively, as above, then we can make

a “multiplication table” for the products of the four elements  $I$ ,  $H$ ,  $V$ ,  $HV$ , as follows.

	$I$	$H$	$V$	$HV$
$I$	$I$	$H$	$V$	$HV$
$H$	$H$	$I$	$HV$	$V$
$V$	$V$	$HV$	$I$	$H$
$HV$	$HV$	$V$	$H$	$I$

This multiplication table is to be read like a multiplication table for numbers. Where a row intersects a column, we have the value of the product of an element on the far left in the row, multiplied by the element on the top of each column. For instance, the product of  $H$  and  $HV$  is

$$HHV = V,$$

because

$$H^2 = I.$$

Similarly, the product of  $HV$  and  $HV$  is

$$\begin{aligned} HVHV &= HHVV \\ &= H^2V^2 \\ &= I. \end{aligned}$$

A multiplication table for numbers would look like this.

	1	3	5	17
1	1	3	5	17
3	3	9	15	51
5	5	15	25	85
17	17	51	85	289

11. Let  $G$  be rotation by  $90^\circ$ , so that  $G^4 = I$ . Fill out the multiplication table given below. Write each entry in the form  $HG^k$  or  $G^m$  for suitable integers  $k, m$  between 0 and 3.

	$I$	$G$	$G^2$	$G^3$	$H$	$HG$	$HG^2$	$HG^3$
$I$								
$G$								
$G^2$								
$G^3$								
$H$								
$HG$								
$HG^2$								
$HG^3$								

12. Let  $G$  be rotation by  $90^\circ$ . Make up and fill out the multiplication table for the elements  $I, G, G^2, G^3, V, VG, VG^2, VG^3$ . Again, express each entry of the table as one of these elements.
13. Consider a triangle whose three sides have equal length and whose three angles have the same measure,  $60^\circ$ , as in Fig. 6-27.

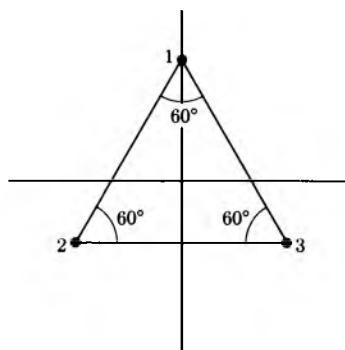


Fig. 6-27

The vertices of the triangle are numbered 1, 2, 3. Let  $G$  be rotation by  $120^\circ$  and let  $V$ , as usual, be reflection through the vertical axis.

- Give the effect of the six isometries  $I, G, G^2, V, VG, VG^2$  on the vertices, using the same notation as in Exercise 9.
- Make up the multiplication table for these six isometries.

14. Let  $G$  be rotation by  $60^\circ$ . Find a positive integer  $k$  such that  $HG = G^k H$ . What is the smallest positive integer  $m$  such that  $G^m = I$ ?
15. Consider a hexagon, i.e., a six-sided figure, whose sides all have the same length and whose angles have the same measure, as shown in Fig. 6-28.

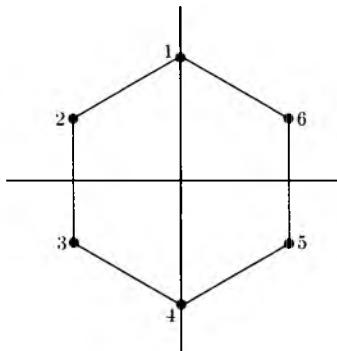


Fig. 6-28

- a) What is the measure of these angles?  
 b) Let  $G$  be rotation by  $60^\circ$  and let  $H, V$  be as before, reflection through the horizontal and vertical axes, respectively. Give the effect of the 12 isometries

$$I, G, G^2, G^3, G^4, G^5, H, HG, HG^2, HG^3, HG^4, HG^5$$

- on the six vertices, using the same notation as in Exercise 9.  
 c) Give the effect of the isometries

$$V, VG, VG^2, VG^3, VG^4, VG^5$$

- on the six vertices, using the same notation as above.  
 d) Make up the multiplication table for the twelve elements

$$I, G, G^2, G^3, G^4, G^5, H, HG, HG^2, HG^3, HG^4, HG^5.$$

16. Using a pentagon instead of a hexagon, answer the same types of questions that were raised in the preceding exercises. Draw the picture, so that the pentagon has one vertex on the vertical axis and admits reflection through the vertical axis as a symmetry. Your picture should be similar to that of Exercise 13, but with a 5-sided figure.
17. Let  $G$  be rotation by  $72^\circ$ . What is the smallest positive integer  $k$  such that  $G^k = I$ ? Express  $G^{-1}$  as a positive power of  $G$ .

## §5. CHARACTERIZATION OF ISOMETRIES

The main result of this section is that an isometry can be expressed as a composite of a translation, a rotation, and possibly a reflection. We first prove an intermediate result.

**Theorem 4.** *Let  $P, Q$  be distinct points. Let  $F$  be an isometry which leaves  $P$  and  $Q$  fixed. Then either  $F$  is the identity, or  $F$  is a reflection through the line  $L_{PQ}$  passing through  $P$  and  $Q$ .*

*Proof.* Let  $M$  be a point on the perpendicular bisector of the segment  $\overline{PQ}$ , but not lying on this segment, as in Fig. 6-29.

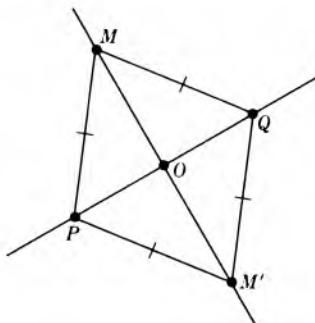


Fig. 6-29

Then

$$d(P, M) = d(Q, M).$$

If  $M$  is fixed by  $F$ , i.e. if  $F(M) = M$ , then we can apply Theorem 3 to conclude that  $F$  is the identity. Suppose that  $F(M) \neq M$ . Let  $M' = F(M)$  as shown in Fig. 6-33. Since  $F$  preserves distances, we have

$$d(P, M') = d(M', Q).$$

Hence by the corollary of the Pythagoras theorem, the point  $M'$  lies on the perpendicular bisector of the segment  $\overline{PQ}$ .

Let  $O$  be the point of intersection of  $\overline{PQ}$  and  $\overline{MM'}$ , i.e. the midpoint between  $P$  and  $Q$ . Again since  $F$  preserves distances, and since  $O$  is fixed under  $F$  (by Theorem 2, §2) we have

$$d(O, M) = d(O, M').$$

Hence  $M'$  is the reflection of  $M$  through the straight line  $L_{PQ}$ . Let  $R$  denote the reflection through this line, so that we have  $R = R^{-1}$ . We also have

$$R(M) = M' \quad \text{and} \quad R(M') = M.$$

Consider the composite isometry

$$R \circ F.$$

It leaves  $P$  and  $Q$  fixed. Furthermore,

$$R(F(M)) = R(M') = M.$$

Hence  $R \circ F$  leaves  $M$  fixed. By Theorem 3, we conclude that  $R \circ F = I$ . Composing with  $R^{-1} = R$  on the left, we find

$$R \circ R \circ F = R \circ I,$$

whence

$$F = R,$$

and our theorem is proved.

**Theorem 5.** *Let  $F$  be an isometry which leaves one point  $O$  fixed. Then either  $F$  is a rotation, or  $F$  is a rotation composed with a reflection through a line.*

*Proof.* Let  $P$  be any point  $\neq O$ . If  $F(P) = P$ , then we are in the case of Theorem 4 and we are done. Suppose that  $F(P) \neq P$ , and let  $F(P) = P'$ . Since  $F$  preserves distances, we have

$$d(O, P) = d(O, P').$$



Fig. 6-30

There exists a rotation with respect to  $O$  which maps  $P$  on  $P'$ . Let us denote this rotation by  $G$ . We know that  $G^{-1}$  is a rotation, and

$$G^{-1}(P') = P.$$

Therefore

$$G^{-1}(F(P)) = G^{-1}(P') = P.$$

This means that  $G^{-1} \circ F$  leaves  $P$  fixed. But  $G^{-1} \circ F$  also leaves  $O$  fixed. Hence we can apply Theorem 4, and we conclude that  $G^{-1} \circ F$  is either the identity or a reflection  $R$ . In the first case, we have

$$G^{-1} \circ F = I,$$

whence composing with  $G$  on the left we find

$$G \circ G^{-1} \circ F = G \circ I = G,$$

and hence

$$F = G$$

is a rotation.

In the second case, we find

$$G^{-1} \circ F = R,$$

whence

$$G \circ G^{-1} \circ F = G \circ R$$

and

$$F = G \circ R.$$

This proves our theorem.

**Theorem 6.** *Let  $F$  be an arbitrary isometry of the plane. If  $F$  does not leave any point fixed, then  $F$  is either a translation, or the composite of a translation and a rotation, or the composite of a translation, a rotation, and a reflection through a line.*

*Proof.* Suppose that  $F$  does not leave any point fixed. Let  $O$  be any point and let  $P = F(O)$ . Let  $T$  be the translation such that  $T(O) = P$ . Then  $T^{-1}$  is a translation, and

$$T^{-1}(P) = O.$$

Hence

$$T^{-1}(F(O)) = T^{-1}(P) = O.$$

This means that  $T^{-1} \circ F$  leaves  $O$  fixed. But  $T^{-1} \circ F$  is an isometry. We can therefore apply Theorem 5, and we see that

$$T^{-1} \circ F = G \quad \text{or} \quad T^{-1} \circ F = G \circ R,$$

where  $G$  is a rotation and  $R$  is a reflection through a line. In the first case, we find

$$F = T \circ G$$

and in the second case we find

$$F = T \circ G \circ R.$$

This proves our theorem.

### EXERCISES

1. Prove that every isometry has an inverse.
2. If  $P$  is a fixed point for an isometry  $F$ , prove that  $P$  is also a fixed point for  $F^{-1}$ .
3. Let  $T$  be the translation by 1 in. to the right and let  $U$  be the translation by 1 in. upward. Draw the image of a point  $P$  under  $T^{-1}$ ,  $T^{-2}$ ,  $T^{-3}$ ,  $U^{-1}$ ,  $U^{-2}$ ,  $U^{-3}$ ,  $T^{-1} \circ U^{-1}$ .

### §6. CONGRUENCES

Let  $S, S'$  be sets of points in the plane. We shall say that  $S$  is **congruent** to  $S'$  if there exists an isometry  $F$  such that the image  $F(S)$  is equal to  $S'$ .

**Theorem 7.** *Two circles of the same radius are congruent.*

*Proof.* Let the first circle be  $C(r, O)$ , or radius  $r$ , centered at  $O$ , and let the other circle be  $C(r, O')$ , centered at  $O'$ . Let  $T$  be the translation which maps  $O$  on  $O'$ . We know that  $T$  preserves distances. Hence if  $P$  is at distance  $r$  from  $O$ , then  $T(P)$  is at distance  $r$  from  $T(O) = O'$ . Hence the image of the circle  $C(r, O)$  is contained in the circle  $C(r, O')$ . We must still show that every point on  $C(r, O')$  is the image of a point on  $C(r, O)$  under  $T$ . Let  $Q$  be a point at distance  $r$  from  $O'$ . Note that the point

$$P = T^{-1}(Q)$$

is at distance  $r$  from  $O$ , and that  $T(P) = T(T^{-1}(Q)) = Q$ . This proves our assertion.

To prove that two figures are congruent, it is often useful to use Exercise 2 at the end of this section. We can then change one figure by any number of isometries. It suffices to prove that its image under these isometries is congruent to the other figure. We shall illustrate this by an example from classical geometry, after we prove the next theorem.

**Theorem 8.** *Any two segments of the same length are congruent.*

*Proof.* Let  $\overline{PQ}$  and  $\overline{MN}$  be segments of the same length. Let  $T$  be the translation which maps  $M$  on  $P$ . Then  $T(N)$  is at the same distance from  $T(M) = P$  as  $Q$ , because  $T$  is an isometry. Hence there exists a rotation  $G$  with respect to  $P$  such that  $G(T(N)) = Q$ . By Theorem 1 of §2, we conclude that  $G \circ T$  maps  $\overline{MN}$  on  $\overline{PQ}$ , thus proving our theorem.

The two steps of the proof in Theorem 8 corresponding to  $T$  and  $G$  are illustrated in Fig. 6-31.

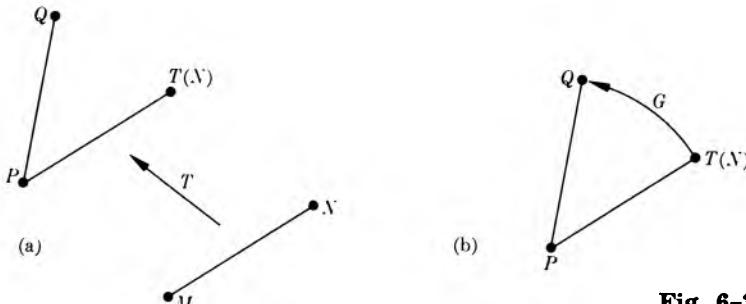


Fig. 6-31

Look at Exercise 2. In the light of this exercise, we could also have phrased the proof of Theorem 8 as follows. We let  $T$  be the translation such that  $T(M) = P$ . Since the image of  $\overline{MN}$  under  $T$  is congruent to  $\overline{MN}$ , we are reduced to the case when  $P = M$ , which we now assume. By assumption,

$$d(P, Q) = d(P, N).$$

Hence there exists a rotation  $G$  with respect to  $P$  such that  $G(N) = Q$ . By Theorem 1 of §2, we conclude that  $\overline{PN}$  is congruent to  $\overline{PQ}$ . This concludes the proof.

Using language as we did, reducing the proof to the case when  $P = M$ , has the slight advantage whereby we avoid having to write the composite

$G \circ T$  explicitly. We shall phrase the proof of Theorem 10 that way also. Note that Theorem 10 is a classical congruence case of elementary courses in plane geometry, finding its natural place within our present system.

**Theorem 9.** *Let  $\triangle PQM$  and  $\triangle P'Q'M'$  be right triangles whose right angles are at  $Q$  and  $Q'$ , respectively. Assume that the corresponding legs have the same lengths, that is:*

$$d(P, Q) = d(P', Q')$$

and

$$d(Q, M) = d(Q', M').$$

*Then the triangles are congruent.*

*Proof.* We leave the proof as an exercise. Observe that this theorem puts our old assumption **RT** in the context of congruences, as we anticipated in Chapter 5, §3.

Actually, Theorem 9 is a special case of a more general result, stated in the next theorem, and whose proof we shall give in full.

**Theorem 10.** *Let  $\triangle PQM$  and  $\triangle P'Q'M'$  be triangles whose corresponding sides have equal lengths, that is*

$$d(P, Q) = d(P', Q'),$$

$$d(P, M) = d(P', M'),$$

$$d(Q, M) = d(Q', M').$$

*These triangles are congruent.*

*Proof.* There exists a translation which maps  $P$  on  $P'$ . Hence it suffices to prove our assertion when  $P = P'$  (cf. Exercise 2). We now assume this, i.e.  $P = P'$ . Since  $d(P, Q) = d(P, Q')$ , there exists a rotation relative to  $P$  which maps  $Q$  on  $Q'$ . This rotation leaves  $P$  fixed. Again by Exercise 2, we are reduced to the case when

$$P = P'$$

and

$$Q = Q'.$$

We assume that this is the case. Now either  $M = M'$ , or  $M \neq M'$ . Suppose  $M \neq M'$ . We illustrate this by Fig. 6-32.

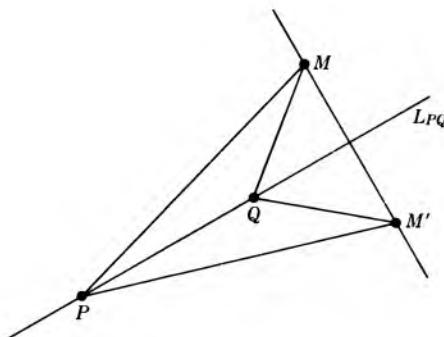


Fig. 6-32

Let  $L_{PQ}$  be the line passing through  $P$  and  $Q$ . By the Corollary of the Pythagoras theorem, and the fact that

$$d(P, M) = d(P, M')$$

and

$$d(Q, M) = d(Q, M'),$$

we conclude that  $L_{PQ}$  is the perpendicular bisector of  $\overline{MM'}$ . In particular,  $M'$  is the reflection of  $M$  through  $L_{PQ}$ . Hence if we reflect  $M'$  through  $L_{PQ}$ , we get  $M$ . Thus we have found a composite of isometries which map  $P$  on  $P'$ ,  $Q$  on  $Q'$ , and  $M$  on  $M'$ . By Theorem 1 of §2 we conclude that our triangles are congruent, as was to be shown.

**Remark.** In Theorems 7 through 10 we have dealt with figures which consist of line segments. Of course, we may also want to deal with other types of figures, for instance discs (cf. Exercise 1), or, say, the triangular region bounded by a triangle, or the region bounded by a rectangle. Because of this, it is useful to have a description of these regions in terms of line segments. We treat the triangle as an example.

Let  $\triangle PQM$  be a triangle, and let  $S$  be the region bounded by the triangle. We represent  $S$  as the shaded region in Fig. 6-33(a).

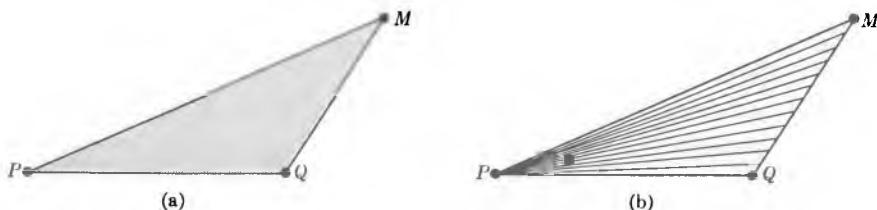


Fig. 6-33

We can give a definition of  $S$  using only the concept of a line segment by saying that  $S$  consists of all the points on all line segments  $\overline{PX}$ , where  $X$  ranges over all points of  $\overline{QM}$ . Looking at Fig. 6-33(b) convinces you that this indeed coincides with your geometric intuition of the triangular region. Using this definition, it is then very easy to see that if  $F$  is an isometry, the image  $F(S)$  is the triangular region bounded by the triangle whose vertices are  $F(P)$ ,  $F(Q)$ ,  $F(M)$ . Carry out the proof in detail as an exercise. This definition is also the one that is used both in pure mathematics and applied mathematics (e.g. economics).

Similarly, let  $R_{PQ}$  and  $R_{PM}$  be two rays which define an angle whose measure is less than  $180^\circ$ . The angle can be described as the set of all points on all segments  $\overline{XY}$ , where  $X$  is a point on  $R_{PQ}$  and  $Y$  is a point on  $R_{PM}$ , as in Fig. 6-34.

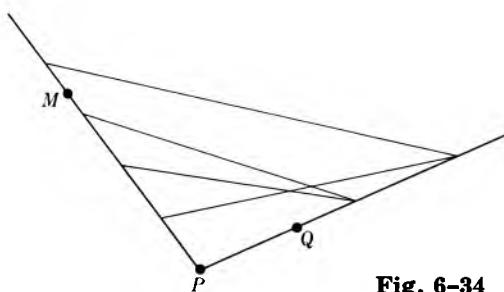


Fig. 6-34

**Isometries and area.** In the next chapter, we shall discuss the notion of area. We may be interested as to how the area of a region behaves under an isometry. It is natural to take the following statement as a basic axiom.

*Let  $S$  be a region of the plane, whose area is equal to  $a$ . Let  $F$  be an isometry. Then the area of  $F(S)$  is also equal to  $a$ .*

To convince ourselves that this is a reasonable statement, we can use the characterization of isometries. If we visualize rotations, reflections, or translations, then our intuition tells us that, in each case, the area of a region is preserved under each one of these mappings. Since any isometry is a composite of such mappings, we see that the area of a region is preserved under an arbitrary isometry.

Let  $A$  be an angle and let  $F$  be an isometry. Then  $F(A)$  is an angle, whose measure is the same as that of  $A$ . We see this from the definition of an angle, looking at the portion of the angle  $A$  lying in a disc centered at the vertex of  $A$ , and using the fact that isometries preserve area. We have drawn the case

when the isometry is the translation  $T_{PP'}$  in Fig. 6-35. Note, however, that a reflection reverses the order of the rays which are used to compute the measure of the angle in counterclockwise direction. Draw the picture.

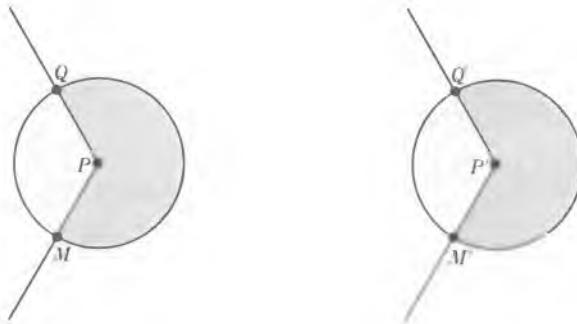


Fig. 6-35

### EXERCISES

1. Prove that two discs of the same radius are congruent.
2. Let  $S, S', S''$  be sets in the plane. Prove that if  $S$  is congruent to  $S'$ , and  $S'$  is congruent to  $S''$ , then  $S$  is congruent to  $S''$ . Prove that if  $S$  is congruent to  $S'$ , then  $S'$  is congruent to  $S$ .
3. Prove that two squares whose sides have the same length are congruent.
4. Prove that any two lines are congruent.
5. Let  $\triangle PQM$  be a triangle whose three angles all have  $60^\circ$ . Prove that the sides have equal length. [Hint: From any vertex draw the perpendicular to the other side, and reflect through this perpendicular.]
6. Prove Theorem 9. At first you are not allowed to use Theorem 10. If you were allowed to use Theorem 10, how could you deduce Theorem 9 from it?
7. Let  $\triangle PQM$  and  $\triangle P'Q'M'$  be triangles having one corresponding angle of the same measure, say,  $\angle PQM$  and  $\angle P'Q'M'$  have the same measure, and having adjacent sides of the same length, i.e.

$$d(P, Q) = d(P', Q') \quad \text{and} \quad d(Q, M) = d(Q', M').$$

Prove that the triangles are congruent.

8. Prove that two rectangles having corresponding sides of equal lengths are congruent.
9. Give a definition of the region bounded by a square in terms of line segments. Same thing for a rectangle.
10. Let  $A$  be the angle shown on Fig. 6-36. Draw the image of  $A$  under
  - a) rotation by  $60^\circ$ ,
  - b) rotation by  $90^\circ$ ,
  - c) rotation by  $120^\circ$ ,
  - d) reflection through  $O$ ,
  - e) reflection through the indicated line  $L$ ,
  - f) reflection through one side of the angle.

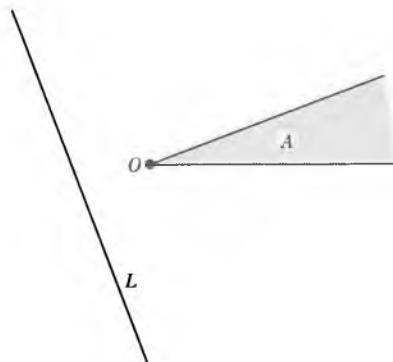


Fig. 6-36

11. Let  $\triangle PQM$  and  $\triangle P'Q'M'$  be triangles whose corresponding angles have the same measures (i.e. the angle with vertex at  $P$  has the same measure as the angle with vertex at  $P'$ , and similarly for the angles with vertices at  $Q, Q'$  and  $M, M'$ ). Assume that  $d(P, Q) = d(P', Q')$ . Prove that the triangles are congruent.
12. Let  $\triangle PQM$  be a triangle. Let  $L_1, L_2, L_3$  be the three lines which bisect the three angles of the triangle, respectively. Let  $O$  be the point of intersection of  $L_1$  and  $L_2$ . Prove that  $O$  lies on  $L_3$ . [Hint: From  $O$ , draw the perpendicular segments to the corresponding sides. Prove that their lengths are equal.]

# 7 Area and Applications

## §1. AREA OF A DISC OF RADIUS $r$

We assume that the notion of area and its basic properties, corresponding to our simple intuition of area, are known. In particular, the area of a square of side  $a$  is  $a^2$ , and the area of a rectangle whose sides have lengths  $a, b$  is  $ab$ . (Remember, a unit of length is fixed throughout, and determines a unit of area.)

Let  $r$  be a positive number, and consider dilation by  $r$ . We wish to analyze what happens to area under such a dilation. We start with the simplest case, that of a rectangle. Consider a rectangle whose sides have lengths  $a$  and  $b$  as on Fig. 7-1(a). Suppose that we multiply the lengths of the sides by 2, and obtain the rectangle illustrated on Fig. 7-1(b).

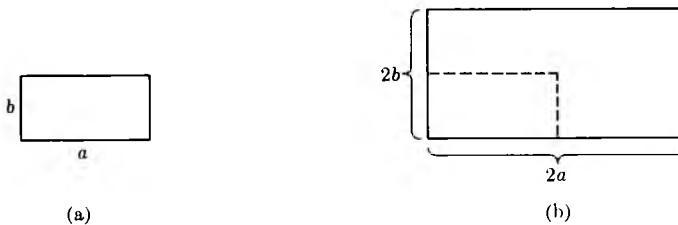


Fig. 7-1

Then the sides of this dilated rectangle have lengths  $2a$  and  $2b$ . Hence the area of the dilated rectangle is equal to  $2a2b = 4ab = 2^2ab$ . Similarly, suppose that we dilate the sides by 3, as illustrated on Fig. 7-1(c).

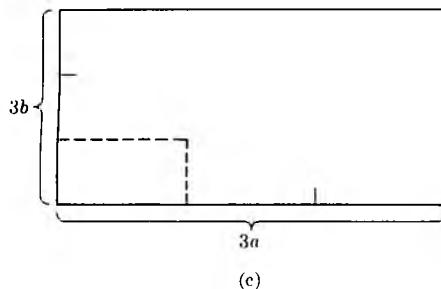


Fig. 7-1 (cont.)

Then the sides of the dilated rectangle have lengths  $3a$  and  $3b$ , whence the area of the dilated rectangle is equal to  $3a3b = 9ab = 3^2ab$ .

In general, let  $S$  be a rectangle whose sides have lengths  $a, b$  respectively. Let  $rS$  be the dilation of  $S$  by  $r$ . Then the sides of  $rS$  have lengths  $ra$  and  $rb$  respectively, so that the area of the dilated rectangle  $rS$  is equal to

$$(ra)(rb) = r^2ab.$$

Thus the area of rectangles changes by  $r^2$  under dilation by  $r$ .

This makes it very plausible that if  $S$  is an arbitrary region of the plane, whose area can be approximated by the area of a finite number of rectangles, then the area of  $S$  itself changes by  $r^2$  under dilation by  $r$ . In other words,

$$\text{area of } rS = r^2(\text{area of } S).$$

For instance, let  $D_r$  be the disc of radius  $r$ , so that  $D_1 = D$  is the disc of radius 1, both centered at the origin (see Fig. 7-2). Then  $D_r = rD_1$ .

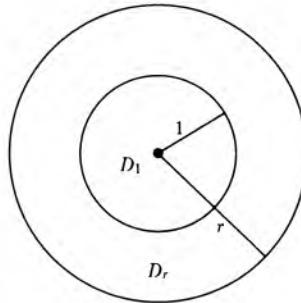


Fig. 7-2

Let  $\pi$  denote the numerical value of the area of  $D_1$ . Then it is plausible that

$$\text{area of } D_r = \pi r^2.$$

We used the symbol  $\pi$  to denote the area of  $D_1$ . It is of course a problem to determine its numerical value. Various devices allow us to do this, and we find the familiar decimal,  $\pi = 3.14159 \dots$ . There is the possibility that you have heard of  $\pi$  only as the ratio of the circumference to the diameter of a circle of radius 1. In the next section, we shall indicate how to prove that this ratio has the same value as the area of the disc of radius 1. Thus the  $\pi$  we are using now is the same one that you may know already. This relationship then gives us a method for computing  $\pi$ . For instance, get a circular pan and a soft measuring tape, measure the circumference of the pan, measure its diameter, and take the ratio. This will give you a value for  $\pi$ , good at least to one decimal place. There are more sophisticated ways of

finding more decimal places for  $\pi$ , and you will learn some of these in a subsequent course in calculus.

One of the methods used to compute  $\pi$  is also the method which convinces us that the area of a disc of radius  $r$  is  $\pi r^2$ . Namely, we approximate the disc by rectangles, or even squares, as in Fig. 7-3.

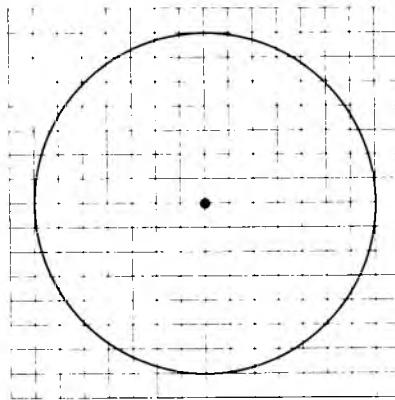


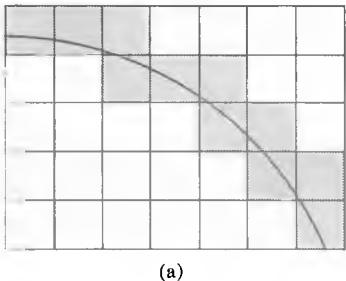
Fig. 7-3

We draw a grid consisting of vertical and horizontal lines at the same distance from each other, thus determining a decomposition of the plane into squares. If the grid is fine enough, that is, if the sides of the square are sufficiently small, then the area of the disc is approximately equal to the sum of the areas of the squares which are contained in the disc. The difference between the area of the disc and this sum will be, at most, the sum of the areas of the squares which intersect the circle. To determine the area of the disc approximately, you just count all the squares that lie inside the circle, measure their sides, add up their areas, and get the desired approximation. Using fine graph paper, you can do this yourself and arrive at your own approximation of the area of the disc.

Of course, you want to estimate how good your approximation is. The difference between the sum of the areas of all the little squares contained in the disc and the area of the disc itself is determined by all the small portions of squares which touch the boundary of the disc, i.e. which touch the circle. We have a very strong intuition that the sum of such little squares will be quite small if our grid is fine enough, and in fact, we give an estimate for this smallness in the following discussion.

Suppose that we make the grid so that the squares have sides of length  $c$ . Then the diagonal of such a square has length  $c\sqrt{2}$ . If a square intersects

the circle, then any point on the square is at distance at most  $c\sqrt{2}$  from the circle. Look at Fig. 7-4(a).



(a)

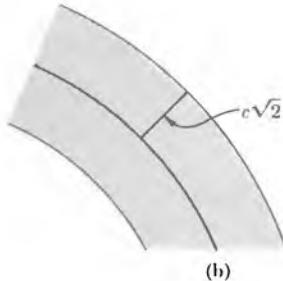


Fig. 7-4

This is because the distance between any two points of the square is at most  $c\sqrt{2}$ . Let us draw a band of width  $c\sqrt{2}$  on each side of the circle, as shown in Fig. 7-4(b). Then all the squares which intersect the circle must lie within that band. It is very plausible that the area of the band is at most equal to

$$2c\sqrt{2} \text{ times the length of the circle.}$$

Thus if we take  $c$  to be very small, i.e. if we take the grid to be a very fine grid, then we see that the area of the disc is approximated by the area covered by the square lying entirely inside the disc.

The same type of argument also works for more general regions. For instance, we have drawn a region  $S$  inside a curve in Fig. 7-5(a), and we have drawn the dilation of  $S$  by 2 in Fig. 7-5(b). Then the area of  $2S$  is equal to  $4A$ , where  $A$  is the area of  $S$ .

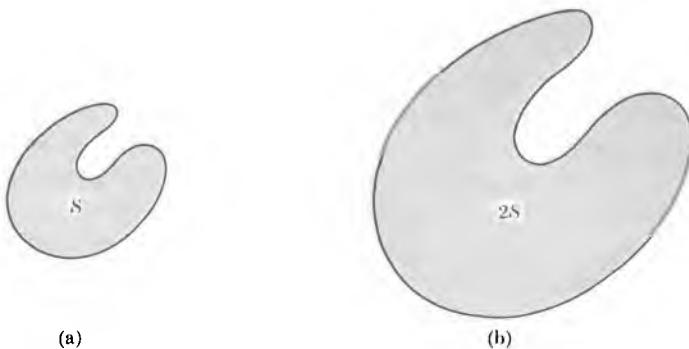
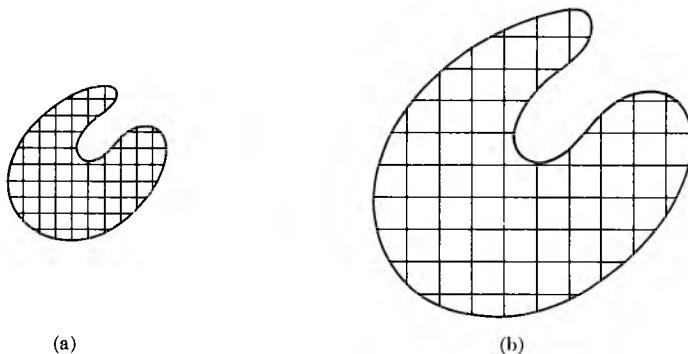


Fig. 7-5

We also illustrate the fact that these areas are approximated by squares in Fig. 7-6(a) and (b).



**Fig. 7-6**

## **EXERCISES**

1. a) Draw a rectangle whose sides have lengths 2 in. and  $\frac{1}{2}$  in., respectively. What is the area of this rectangle?  
b) Draw the rectangle whose sides have lengths equal to twice the length of the sides of the rectangle in part (a). What is the area of this rectangle?  
c) Same question for the rectangle whose sides have lengths equal to three times the lengths of the sides of the rectangle in part (a).  
d) Same question for the rectangle whose sides have lengths equal to one-half the lengths of the sides of the rectangle in part (a).
  2. a) Draw a rectangle whose sides have lengths  $\frac{2}{3}$  in. and 2 in. respectively. What is the area of this rectangle? Draw the rectangle whose sides have lengths equal to:  
b) twice,  
c) three times,  
d) half the lengths of the sides of the rectangle in part (a). In each case, what is the area of the rectangle?

- e) Suppose that a rectangle has an area equal to  $15 \text{ in}^2$ . Dilate the rectangle by 2. What is the area of the dilated rectangle?
- f) If a rectangle has an area equal to  $25 \text{ in}^2$ , what is the area of the rectangle whose sides have lengths equal to one-fifth the lengths of the original rectangle?

Read about coordinates and dilations in terms of coordinates, that is, Chapter 8, §1, §2, §3, and Chapter 9, §1. Then consider the following generalization of a dilation. Let  $a > 0, b > 0$ . To each point  $(x, y)$  of the plane, associate the point

$$(ax, by).$$

Thus we stretch the  $x$ -coordinate by  $a$  and the  $y$ -coordinate by  $b$ . This association is a mapping which we may denote by  $F_{a,b}$ .

3. a) Suppose that the sides of a rectangle  $S$  have lengths  $r$  and  $s$ . What are the lengths of the sides of the rectangle  $F_{a,b}(S)$ , i.e. of the rectangle obtained by the mixed dilation  $F_{a,b}$ ?  
 b) What is the volume of  $F_{a,b}(S)$ ?  
 c) If  $S$  is a bounded region in the plane with volume  $V$ , what is the volume of  $F_{a,b}(S)$ ?
4. a) Show that the set of points  $(u, v)$  satisfying the equation

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

- is the image of the circle of radius 1 centered at  $O$  under the map  $F_{a,b}$ .  
 b) Let  $a = 3$  and  $b = 2$ . Sketch this set, which is called an **ellipse**.  
 c) Can you guess and motivate your guess as to what the area of the region bounded by the ellipse in (a) should be?

5. What is the area of the region bounded by the following ellipses:
- |  |  |
|--|--|
| a) $\left(\frac{x}{7}\right)^2 + \left(\frac{y}{4}\right)^2 = 1$ ? | b) $\left(\frac{x}{3}\right)^2 + \left(\frac{y}{7}\right)^2 = 1$ ? |
| c) $\frac{x^2}{6} + \frac{y^2}{3} = 1$ ?                           | d) $\frac{x^2}{5} + \frac{y^2}{7} = 1$ ?                           |
6. What is the area of the region bounded by the following ellipses:
- |                        |                         |
|------------------------|-------------------------|
| a) $3x^2 + 4y^2 = 1$ ? | b) $2x^2 + 5y^2 = 1$ ?  |
| c) $4x^2 + 9y^2 = 1$ ? | d) $4x^2 + 25y^2 = 1$ ? |

7. Write up a discussion of how to give coordinates  $(x, y, z)$  to a point in 3-space. In terms of these coordinates, what would be the effect of dilation by  $r$ ?
8. Generalize the discussion of this section to the 3-dimensional case. Specifically:
  - a) Under dilation by  $r$ , how does the volume of a cube change?
  - b) How does the volume of a rectangular box with sides  $a, b, c$  change? Draw a picture, say for  $r = \frac{1}{2}, r = 2, r = 3$ , arbitrary  $r$ .
  - c) How would the volume of a 3-dimensional solid change under dilation by  $r$ ?
  - d) The volume of the solid ball of radius 1 in 3-space is equal to  $\frac{4}{3}\pi$ . What is the volume of the ball of radius  $r$  in 3-space?
9. Write down the equation of a sphere of radius  $r$  centered at the origin in 3-space.
10. How would you define the volume of a rectangular solid whose sides have lengths  $a, b, c$ ?
11. Let  $a, b, c$  be positive numbers. Let  $\mathbf{R}^3$  be 3-space, that is, the set of all triples of numbers  $(x, y, z)$ . Let

$$F_{a,b,c}: \mathbf{R}^3 \rightarrow \mathbf{R}^3$$

be the mapping

$$(x, y, z) \mapsto (ax, by, cz).$$

Thus  $F_{a,b,c}$  is a generalization to 3-space of our mixed dilation  $F_{a,b}$ .

- a) What is the image of a cube whose sides have length 1 under  $F_{a,b,c}$ ?
- b) A rectangular box  $S$  has sides of lengths  $r, s, t$  respectively. What are the lengths of the sides of the image  $F_{a,b,c}(S)$ ? What is the volume of  $F_{a,b,c}(S)$ ?
- c) Let  $S$  be a solid in 3-space, and let  $V$  be its volume. In terms of  $V, a, b, c$ , what is the volume of the image of  $S$  under  $F_{a,b,c}$ ?
12. What is the volume of the solid in 3-space consisting of all points  $(x, y, z)$  satisfying the inequality

$$\left(\frac{x}{3}\right)^2 + \left(\frac{y}{2}\right)^2 + \left(\frac{z}{7}\right)^2 \leq 1?$$

13. What is the volume of the solid in 3-space consisting of all points  $(x, y, z)$  satisfying the inequality

$$\frac{x^2}{5} + \frac{y^2}{3} + \frac{z^2}{10} \leq 1?$$

14. Let  $a, b, c$  be numbers  $> 0$ . What is the volume of the solid in 3-space consisting of all points  $(x, y, z)$  satisfying the inequality

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 \leq 1?$$

15. What about 4-space?  $n$ -space for arbitrary  $n$ ?

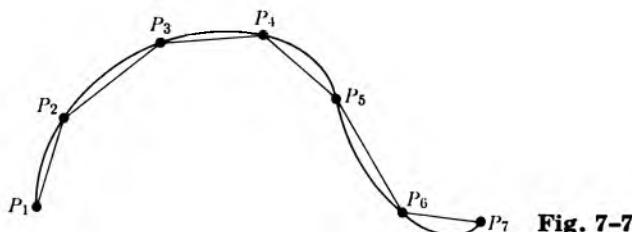
## §2. CIRCUMFERENCE OF A CIRCLE OF RADIUS $r$

Let  $C$  be the circle of radius 1, and let  $C_r$  be the circle of radius  $r$ . We wish to convince ourselves that

length of  $C_1 = 2\pi$ ,  
 length of  $C_r = 2\pi r$ .

We studied area in the preceding section by approximating a region  $S$  by means of squares. To study length, we have to approximate a curve by means of straight line segments.

On the following picture, we show how to approximate a curve by 6 segments.

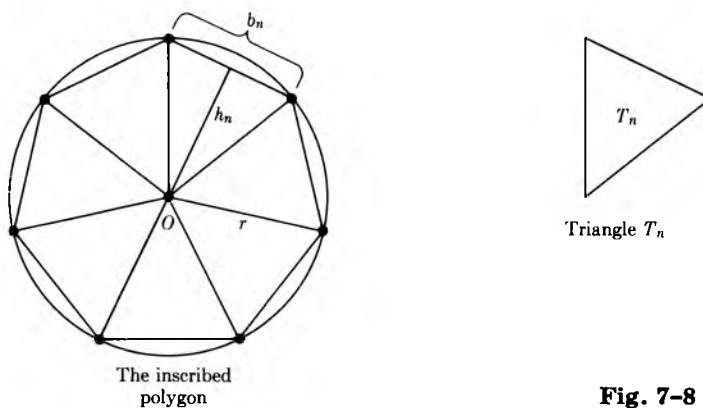


**Fig. 7-7**

To approximate a circle by segments, we select a special kind of segment. We decompose the disc of radius  $r$  into  $n$  sectors whose angles have

$$\frac{360}{n} \text{ degrees.}$$

Here,  $n$  is an integer. The picture is that of Fig. 7-8, drawn with  $n = 7$ .



**Fig. 7-8**

You should now do Exercise 1, taking special values for  $n$ , namely  $n = 4, n = 5, n = 6, n = 8$  (more if you wish). In each case, determine the number of degrees

$$\frac{360}{4}, \quad \frac{360}{5}, \quad \frac{360}{6}, \quad \frac{360}{8}, \quad \text{etc.},$$

and draw the corresponding sectors. In general we then join the end points of the sectors by line segments, thus obtaining a polygon inscribed in the circle. The region bounded by this polygon consists of  $n$  triangles congruent to the same triangle  $T_n$ , lying between the segments from the origin  $O$  to the vertices of the polygon. Thus in the case of 4 sides, we call the triangle  $T_4$ . In the figure with 5 sides, we call the triangle  $T_5$ . In the figure with 6 sides, we call the triangle  $T_6$ . In the figure with 8 sides, we call the triangle  $T_8$ . And so on, to the polygon having  $n$  sides, when we call the triangle  $T_n$ .

We denote the (length of the) base of  $T_4$  by  $b_4$  and its height by  $h_4$ . We denote the base of  $T_5$  by  $b_5$  and its height by  $h_5$ . In general, we denote the base of  $T_n$  by  $b_n$  and its height by  $h_n$ , as indicated on Fig. 7-8. Since the area of a triangle whose base has length  $b$  and whose height has length  $h$

is  $\frac{1}{2}bh$ , we see that the area of our triangle  $T_n$  is given by

$$\text{area of } T_n = \frac{1}{2}b_n h_n.$$

Let  $A_n$  be the area of the region surrounded by the polygon, and let  $P_n$  be the perimeter of the polygon. Since the polygonal region consists of  $n$  triangular regions congruent to the same triangle  $T_n$ , we find that the area of  $A_n$  is equal to  $n$  times the area of  $T_n$ , or in symbols,

$$A_n = n \cdot \text{area of } T_n = \frac{1}{2}n b_n h_n.$$

On the other hand, the perimeter of the polygon consists of  $n$  segments whose lengths are all equal to  $b_n$ . Hence  $P_n = nb_n$ . Substituting  $P_n$  for  $nb_n$  in the value for  $A_n$  which we just found, we get

$$A_n = \frac{1}{2}P_n h_n.$$

As  $n$  becomes arbitrarily large,

$A_n$  approaches the area of the disc  $D_r$ ,

$P_n$  approaches the circumference of the circle  $C_r$ ,

and

$h_n$  approaches the radius  $r$  of the disc.

For instance, if we double the number of sides of the polygon successively, the picture looks like Fig. 7-9.

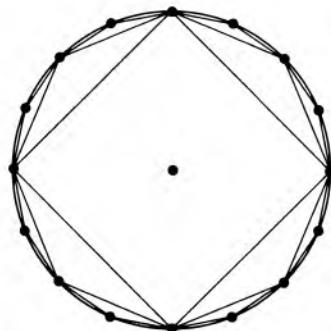


Fig. 7-9

Let  $c$  denote the circumference of the circle of radius  $r$ . Then  $A_n$  approaches  $\pi r^2$ . Since  $A_n = \frac{1}{2}P_n h_n$ , it follows that  $A_n$  also approaches  $\frac{1}{2}cr$ . Thus we obtain

$$\pi r^2 = \frac{1}{2}cr.$$

We cancel  $r$  from each side of this equation, and multiply both sides by 2. We conclude that

$$c = 2\pi r,$$

as was to be shown.

We have discussed above the behavior of area under dilation. We conclude this section by a discussion of the behavior of length under dilation.

Let  $r$  be a positive number. Under dilation by  $r$ , the distance between two points is multiplied by  $r$ ; in other words,

$$d(rP, rQ) = r \cdot d(P, Q).$$

When we have coordinates later, this will be proved. We can already justify it to some extent by using the Pythagoras theorem. Consider a right triangle  $\triangle POQ$  with sides  $a, b$  as shown in Fig. 7-10, such that the right angle is at the origin  $O$ .

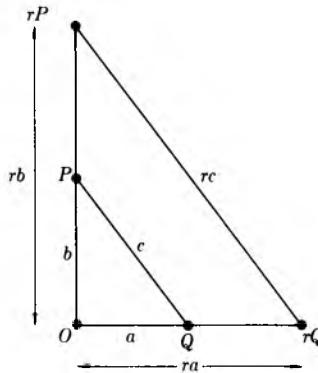


Fig. 7-10

Under dilation by  $r$ , the two sides are dilated by  $r$ , so that the three points

$$rP, \quad O, \quad rQ,$$

form a right triangle whose legs have lengths  $ra$  and  $rb$ , respectively. By the Pythagoras theorem, the hypotenuse of the dilated triangle has length

$$\sqrt{r^2a^2 + r^2b^2} = \sqrt{r^2(a^2 + b^2)}.$$

If we let

$$c = \sqrt{a^2 + b^2},$$

we see that the hypotenuse has length  $rc$ . Thus the length of the hypotenuse also gets multiplied by  $r$ . As we shall see later when we have coordinates, this is also true even if the vertex of the right angle is not necessarily at the origin.

To investigate what happens to the length of an arbitrary curve  $S$  under dilation, we approximate the curve by segments as on Fig. 7-7. In general, suppose that we approximate the curve by  $n$  segments where  $n$  is a positive integer. Let these segments be  $S_1, \dots, S_n$ . Let  $l(S_i)$  denote the length of  $S_i$ . If we dilate the curve by  $r$ , then  $rS$  is approximated by the segments  $rS_1, \dots, rS_n$ . Hence the length of  $rS$  is approximated by

$$l(rS_1) + \dots + l(rS_n) = r[l(S_1) + \dots + l(S_n)].$$

Thus we see that whenever we can approximate  $S$  by line segments, we have the formula

$$\text{length of } rS = r(\text{length of } S).$$

In particular, the length of the circle of radius 1 is  $2\pi$ , and the length of the circle of radius  $r$  is  $2\pi r$ . We proved this above, and we now see that it is compatible with the general behavior of length under dilation.

Our arguments are based on the idea of taking a limit as  $n$  becomes arbitrarily large, and on the notion of approximation. These are the basic ideas of the calculus, which is devoted to systematizing these notions and giving an analytic basis for their logical development. However, it is always useful to have the intuitive ideas first. Thus you may view this section as a good introduction to the calculus.

## EXERCISES

1. Draw the picture of a polygon with sides of equal length, inscribed in a circle of radius 2 inches, in the cases when the polygon has:
  - a) 4 sides,
  - b) 5 sides,
  - c) 6 sides,
  - d) 8 sides,
  - e) 9 sides.

Draw the radii from the center of the circle to the vertices of the polygon. Use a protractor for the angles of the sectors. Using a ruler, measure

(approximately) the base of each triangle and measure its height. From your measurements, compute the area inside the polygon, and the circumference of the polygon. Compare these values with the area of the disc and its circumference, given as  $\pi r^2$  and  $2\pi r$  respectively, and  $r = 2$ . Use the value  $\pi = 3.14$ .

2. Get a tin can with as big a circular bottom as possible. Take a tape, measure the circumference of the bottom, measure the diameter, take the ratio and get a value for  $\pi$ , probably good to one decimal place. Do the same thing to another circular object, say a frying pan, and verify that you get the same value for  $\pi$ .
3. Read the definitions of coordinates at the beginning of Chapter 8. Then read the sections of Chapter 16 on induction and summation, and do the exercises at the end of these sections. The material just mentioned is logically self-contained, so you should have no trouble. It provides a direct continuation of the topics which were discussed concerning area and volumes. You will see how to compute the volume of a sphere and a cone, or other similar figures, in 3-space.



*Part Three*

**COORDINATE GEOMETRY**



As stated before, we shall see how to use coordinates to give definitions for geometric terms using only properties of numbers.

The chapters in this part are logically independent of each other to a large extent. For instance, the chapter on trigonometry could be read immediately after the introduction of coordinates in Chapter 8, §1, and the discussion of the distance formula between points. Many readers may want to do that, instead of reading first about translations, addition of points, etc. On the other hand, the chapter on segments, rays, and lines is also independent of the trigonometry and the analytic geometry. These three chapters can be read in any order. Take your pick as to what approach you like most.

Giving coordinates to points not only allows us to give analytic proofs. It allows us also to compute in a way that the “intuitive” geometry did not. For instance, given a coordinatized definition of straight lines, we can compute the point of intersection of two lines explicitly.

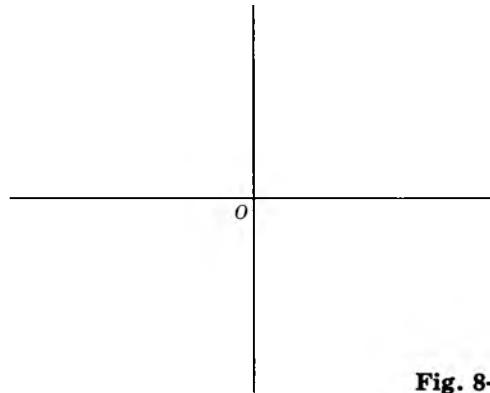


# *8 Coordinates and Geometry*

## **§1. COORDINATE SYSTEMS**

Once a unit length is selected, we can represent numbers as points on a line. We shall now extend this procedure to the plane, and to pairs of numbers.

In Fig. 8-1, we visualize a horizontal line and a vertical line intersecting at a point  $O$ , called the **origin**.



**Fig. 8-1**

These lines will be called **coordinate axes**, or simply **axes**.

We select a unit length and cut the horizontal line into segments of lengths 1, 2, 3, . . . to the left and to the right. We do the same to the vertical line, but up and down, as indicated on Fig. 8-2.

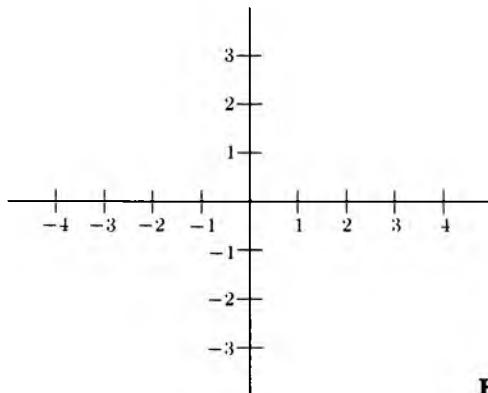


Fig. 8-2

On the vertical line, we visualize the points going below  $O$  as corresponding to the negative integers, just as we visualized points on the left of the horizontal line as corresponding to negative integers. We follow the same idea as that used in grading a thermometer, where the numbers below zero are regarded as negative.

We can now cut the plane into squares whose sides have length 1.

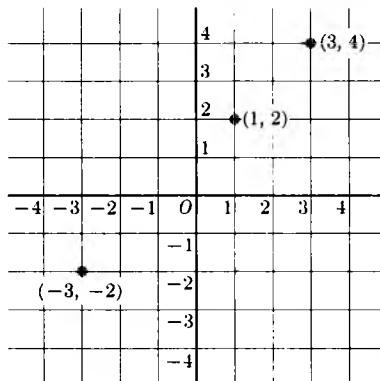


Fig. 8-3

We can describe each point where two lines intersect by a pair of integers. Suppose that we are given a pair of integers, like  $(1, 2)$ . We go 1 unit to the right of the origin and up 2 units vertically to get the point  $(1, 2)$  which has been indicated in Fig. 8-3. We have also indicated the point  $(3, 4)$ . The diagram is just like a map.

Furthermore, we could also use negative numbers. For instance, to describe the point  $(-3, -2)$ , we go 3 units to the left of the origin and 2 units vertically downward.

There is actually no reason why we should limit ourselves to points which are described by integers. For instance we can also describe the point  $(\frac{1}{2}, -1)$  and the point  $(-\sqrt{2}, 3)$  as on Fig. 8-4.

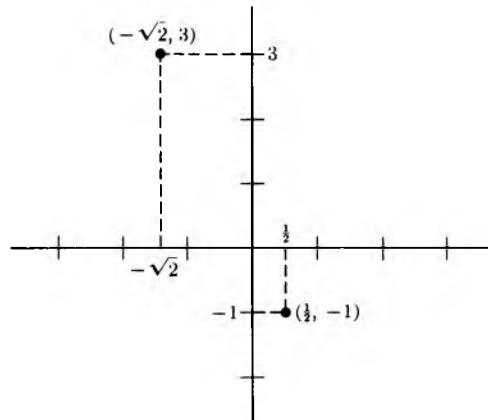


Fig. 8-4

In general, if we take any point  $P$  in the plane and draw the perpendicular lines to the horizontal axis and to the vertical axis, we obtain two numbers  $x, y$  as on Fig. 8-5.

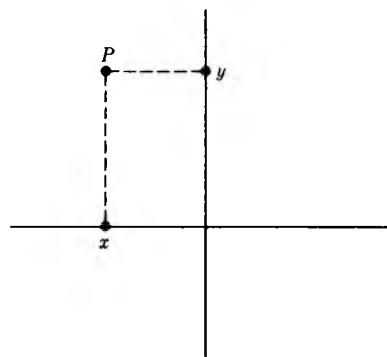


Fig. 8-5

We then say that the numbers  $x, y$  are the **coordinates** of the point  $P$ , and we write

$$P = (x, y).$$

Conversely, every pair of numbers  $(x, y)$  determines a point of the plane. If  $x$  is positive, then this point lies to the right of the vertical axis. If  $x$  is negative, then this point lies to the left of the vertical axis. If  $y$  is positive, then this point lies above the horizontal axis. If  $y$  is negative, then this point lies below the horizontal axis.

The coordinates of the origin are

$$O = (0, 0).$$

We usually call the horizontal axis the  $x$ -axis, and the vertical axis the  $y$ -axis (exceptions will always be noted explicitly). Thus if

$$P = (5, -10),$$

then we say that 5 is the  $x$ -coordinate and  $-10$  is the  $y$ -coordinate. Of course, if we don't want to fix the use of  $x$  and  $y$ , then we say that 5 is the first coordinate, and  $-10$  is the second coordinate. What matters here is the ordering of the coordinates, so that we can distinguish between a first and a second.

We can, and sometimes do, use other letters besides  $x$  and  $y$  for coordinates, for instance  $t$  and  $s$ , or  $u$  and  $v$ .

Our two axes separate the plane into four **quadrants**, which are numbered as indicated in Fig. 8-6.

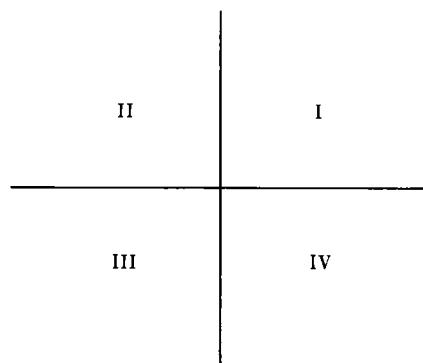


Fig. 8-6

A point  $(x, y)$  lies in the **first quadrant** if and only if both  $x$  and  $y$  are  $> 0$ . A point  $(x, y)$  lies in the **fourth quadrant** if and only if  $x > 0$ , but  $y < 0$ .

Finally, we note that we placed our coordinates horizontally and vertically for convenience. We could also place the coordinates in a slanted way, as shown on Fig. 8-7.

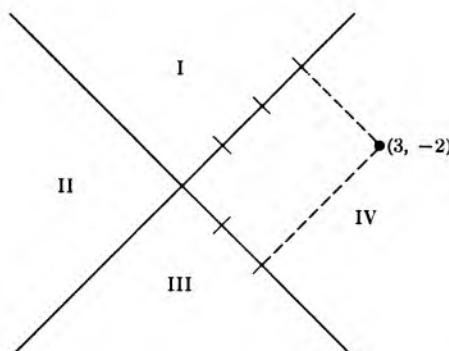


Fig. 8-7

In Fig. 8-7, we have indicated the quadrants corresponding to this coordinate system, and we have indicated the point  $(3, -2)$  having coordinates 3 and  $-2$  with respect to this coordinate system. Of course, when we change the coordinate system, we also change the coordinates of a point.

**Remark.** Throughout this book, when we select a coordinate system, the positive direction of the second axis will always be determined by rotating counterclockwise the positive direction of the first axis through a right angle.

We observe that the selection of a coordinate system amounts to the same procedure that is used in constructing a map. For instance, on the following (slightly distorted) map, the coordinates of Los Angeles are  $(-6, -2)$ , those of Chicago are  $(3, 2)$ , and those of New York are  $(7.2, 3)$ . (View the distortion in the same spirit as you view modern art.)

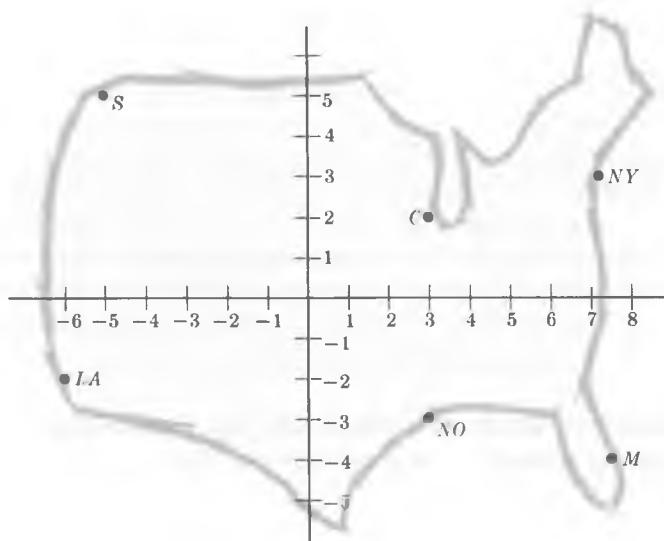


Fig. 8-8

## EXERCISES

In these exercises, draw the coordinate axes horizontally and vertically.

1. Plot the following points:  $(-1, 1)$ ,  $(0, 5)$ ,  $(-5, -2)$ ,  $(1, 0)$ .
2. Plot the following points:  $(\frac{1}{2}, 3)$ ,  $(-\frac{1}{3}, -\frac{1}{2})$ ,  $(\frac{4}{3}, -2)$ ,  $(-\frac{1}{4}, -\frac{1}{2})$ .
3. Let  $(x, y)$  be the coordinates of a point in the second quadrant. Is  $x$  positive or negative? Is  $y$  positive or negative?
4. Let  $(x, y)$  be the coordinates of a point in the third quadrant. Is  $x$  positive or negative? Is  $y$  positive or negative?
5. Plot the following points:  $(1.2, -2.3)$ ,  $(1.7, 3)$ .
6. Plot the following points:  $(-2.5, \frac{1}{3})$ ,  $(-3.5, \frac{5}{4})$ .
7. Plot the following points:  $(1.5, -1)$ ,  $(-1.5, -1)$ .
8. What are the coordinates of Seattle ( $S$ ), Miami ( $M$ ) and New Orleans ( $NO$ ) on our map?

## Big exercise

To be carried through this chapter and the next two chapters: Define a point in 3-space to be a triple of numbers

$$(x_1, x_2, x_3).$$

Try to formulate the same results in this 3-dimensional case that we have done in the book for the 2-dimensional case. In particular, define distance, define addition of points, dilation, translations, straight lines in 3-space, the whole lot. Write it all up as if you were writing a book. This will make you really learn the subject. Point out the similarities with the 2-dimensional case, and point out the differences if any. You will find practically no difference! To give you some guidelines, we shall often state explicitly what you should do.

So to get you started, we draw in Fig. 8-9 a system of perpendicular coordinate axes in 3-space in a manner quite similar to 2-space.

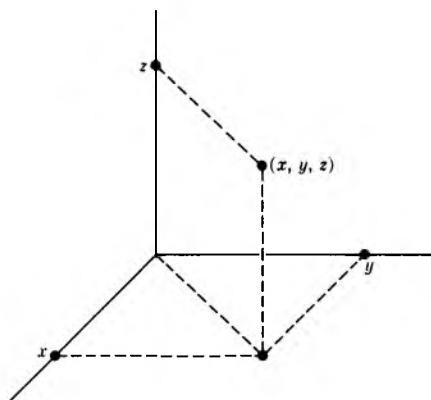


Fig. 8-9

Thus a point in 3-space is represented by three numbers, and from the analytic point of view, we define such a point to be a triple of numbers  $(x, y, z)$ . For instance,  $(1, -2, 7)$  is a point in 3-space. We denote 3-space by  $\mathbf{R}^3$ .

## §2. DISTANCE BETWEEN POINTS

First let us consider the distance between two points on a line. For instance, the distance between the points 1 and 4 on the line is  $4 - 1 = 3$ .

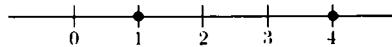


Fig. 8-10

Observe that if we take the difference between 1 and 4 in the other direction, namely  $1 - 4$ , then we find  $-3$ , which is negative. However, if we then take the square, we find the same number, namely

$$(-3)^2 = 3^2 = 9.$$

Thus when we take the square, it does not matter in which order we took the difference.

**Example.** Find the square of the distance between the points  $-2$  and  $3$  on the line.

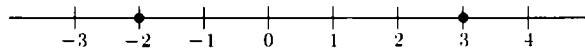


Fig. 8-11

The square of this distance is

$$(3 - (-2))^2 = (3 + 2)^2 = 25;$$

or computing the other way,

$$(-2 - 3)^2 = (-5)^2 = 25.$$

Note that again we take the difference between the coordinates of the points, and that we can deal with points having negative coordinates. If we want the distance rather than its square, then we take the square root, and we find

$$\sqrt{(-5)^2} = \sqrt{5^2} = \sqrt{25} = 5.$$

Because of our universal convention that the square root of a positive number is taken to be positive, we see that we can express the general formula for the distance between points on a line as follows.

*Let  $x_1, x_2$  be points on a line. Then the distance between  $x_1$  and  $x_2$  is equal to*

$$\sqrt{(x_1 - x_2)^2}.$$

Next we discuss distance between points in the plane, given a coordinate system, which we draw horizontally and vertically for convenience. We recall the Pythagoras theorem from plane geometry.

*In a right triangle, let  $a, b$  be the lengths of the legs (i.e. the sides forming the right angle), and let  $c$  be the length of the third side (i.e. the hypotenuse). Then*

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

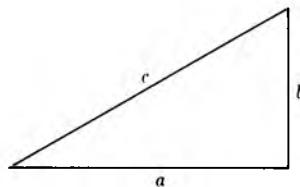


Fig. 8-12

**Example.** Let  $(1, 2)$  and  $(3, 5)$  be two points in the plane. Using the Pythagoras theorem, we wish to find the distance between them. First we draw the picture of the right triangle obtained from these two points, as on Fig. 8-13.

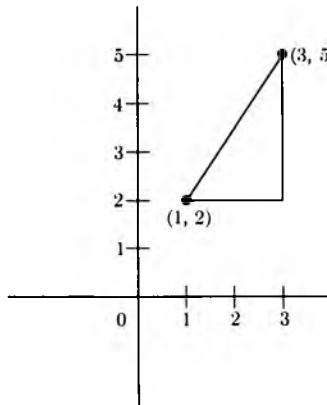


Fig. 8-13

We see that the square of the length of one side is equal to

$$(3 - 1)^2 = 4.$$

The square of the length of the other side is

$$(5 - 2)^2 = (3)^2 = 9.$$

By the Pythagoras theorem, we conclude that the square of the length between the points is  $4 + 9 = 13$ . Hence the distance itself is  $\sqrt{13}$ .

Now in general, let  $(x_1, y_1)$  and  $(x_2, y_2)$  be two points in the plane. We can again make up a right triangle, as shown in Fig. 8-14.

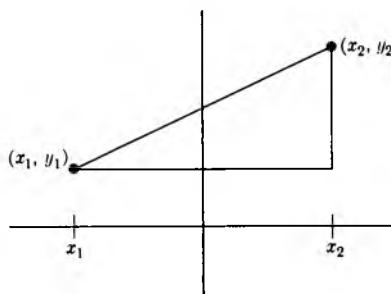


Fig. 8-14

The square of the bottom side is  $(x_2 - x_1)^2$ , which is also equal to  $(x_1 - x_2)^2$ . The square of the vertical side is  $(y_2 - y_1)^2$ , which is also equal to  $(y_1 - y_2)^2$ . If  $d$  denotes the distance between the two points, then

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2,$$

and therefore we get the formula for the distance between the points, namely

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

**Example.** Let the two points be  $(1, 2)$  and  $(1, 3)$ . Then the distance between them is equal to

$$\sqrt{(1 - 1)^2 + (3 - 2)^2} = 1.$$

**Example.** Find the distance between the points  $(-1, 5)$  and  $(4, -3)$ . This distance is equal to

$$\sqrt{(4 - (-1))^2 + (-3 - 5)^2} = \sqrt{89}.$$

**Example.** Find the distance between the points  $(2, 4)$  and  $(1, -1)$ . The square of the distance is equal to

$$(1 - 2)^2 + (-1 - 4)^2 = 26.$$

Hence the distance is equal to  $\sqrt{26}$ .

*Warning:* Always be careful when you meet minus signs. Place the parentheses correctly and remember the rules of algebra.

As it should be, we can compute the distance between points in any order. In the last example, the square of the distance is equal to

$$(2 - 1)^2 + (4 - (-1))^2 = 26.$$

If we let  $d(P, Q)$  denote the distance between points  $P$  and  $Q$ , then our last remark can also be expressed by saying that

$$d(P, Q) = d(Q, P).$$

Note that this is the basic property **DIST 2** of Chapter 5.

We have the general program of giving foundations for geometry assuming only properties of numbers. In line with this program, we must make clear what we take as definitions. The rules of the game are that only properties of numbers can be assumed as known.

We therefore define the **plane** to be the set of all pairs  $(x, y)$  of real numbers. We denote the plane by  $\mathbf{R}^2$ .

If  $X = (x_1, x_2)$  and  $Y = (y_1, y_2)$  are two points of the plane, then we define the **distance** between them to be

$$d(X, Y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Thus we see that we have defined geometric objects using only numbers.

### EXERCISES

Find the distance between the following points  $P$  and  $Q$ . Draw these points on a sheet of graph paper.

1.  $P = (2, 1)$  and  $Q = (1, 5)$
2.  $P = (-3, -1)$  and  $Q = (-4, -6)$
3.  $P = (-2, 1)$  and  $Q = (3, 7)$
4.  $P = (-3, -4)$  and  $Q = (-2, -3)$
5.  $P = (3, -2)$  and  $Q = (-6, -7)$
6.  $P = (-3, 2)$  and  $Q = (6, 7)$
7.  $P = (-3, -4)$  and  $Q = (-1, -2)$
8.  $P = (-1, 5)$  and  $Q = (-4, -2)$
9.  $P = (2, 7)$  and  $Q = (-2, -7)$
10.  $P = (3, 1)$  and  $Q = (4, -1)$
11. Prove that if  $d(P, Q) = 0$ , then  $P = Q$ . Thus we have now proved two of the basic properties of distance.
12. Let  $A = (a_1, a_2)$  and  $B = (b_1, b_2)$ . Let  $r$  be a positive number. Write down the formula for  $d(A, B)$ . Define the **dilation**  $rA$  to be

$$rA = (ra_1, ra_2).$$

For instance, if  $A = (-3, 5)$  and  $r = 7$ , then  $rA = (-21, 35)$ . If  $B = (4, -3)$  and  $r = 8$ , then  $rB = (32, -24)$ . Prove in general that

$$d(rA, rB) = r \cdot d(A, B).$$

We shall investigate dilations more thoroughly in the next chapter.

### Exercise on 3-space

Let

$$A = (a_1, a_2, a_3) \quad \text{and} \quad B = (b_1, b_2, b_3)$$

be points in 3-space. Define the distance between them to be

$$d(A, B) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2}.$$

This generalizes the Pythagoras theorem. Draw a picture. Draw the segment between the origin  $O = (0, 0, 0)$  and a point  $X = (x, y, z)$ . Write down the simple formula for the distance between  $(0, 0, 0)$  and  $(x, y, z)$ . Draw right triangles, showing that geometrically, our formula for the distance between  $O$  and  $X$  in 3-space can be justified in terms of an iterated application of the ordinary Pythagoras theorem in a horizontal plane and a vertical plane. Look back at the exercises of Chapter 5, §3, and note their relation with the present considerations.

13. Give the value for the distance between the following.

- a)  $P = (1, 2, 4)$  and  $Q = (-1, 3, -2)$
- b)  $P = (1, -2, 1)$  and  $Q = (-1, 1, 1)$
- c)  $P = (-2, -1, -3)$  and  $Q = (3, 2, 1)$
- d)  $P = (-4, 1, 1)$  and  $Q = (1, -2, -5)$

14. Let  $r$  be a positive number. Define  $rA$  in a manner similar to the definition of Exercise 12. Prove that

$$d(rA, rB) = r \cdot d(A, B).$$

### §3. EQUATION OF A CIRCLE

Let  $P$  be a given point and  $r$  a number  $> 0$ . The **circle of radius  $r$  centered at  $P$**  is by definition the set of all points whose distance from  $P$  is equal to  $r$ . We can now express this condition in terms of coordinates.

**Example.** Let  $P = (1, 4)$  and let  $r = 3$ . A point whose coordinates are  $(x, y)$  lies on the circle of radius 3 centered at  $(1, 4)$  if and only if the distance between  $(x, y)$  and  $(1, 4)$  is 3. This condition can be written as

$$(1) \quad \sqrt{(x - 1)^2 + (y - 4)^2} = 3.$$

This relationship is called the equation of the circle of center  $(1, 4)$  and radius 3. Note that both sides are positive. Thus this equation holds if and only if

$$(2) \quad (x - 1)^2 + (y - 4)^2 = 9.$$

Indeed, if (1) is true, then (2) is true because we can square each side of (1) and obtain (2). On the other hand, if (2) is true and we take the square root of each side, we obtain (1), because the numbers on each side of (1) are positive. It is often convenient to leave the equation of the circle in the form (2), to avoid writing the messy square root sign. We also call (2) the equation of the circle of radius 3 centered at  $(1, 4)$ .

**Example.** The equation

$$(x - 2)^2 + (y + 5)^2 = 16$$

is the equation of a circle of radius 4 centered at  $(2, -5)$ . Indeed, the square of the distance between a point  $(x, y)$  and  $(2, -5)$  is

$$(x - 2)^2 + (y - (-5))^2 = (x - 2)^2 + (y + 5)^2.$$

Thus a point  $(x, y)$  lies on the prescribed circle if and only if

$$(x - 2)^2 + (y + 5)^2 = 4^2 = 16.$$

Note especially the  $y + 5$  in this equation.

**Example.** The equation

$$(x + 2)^2 + (x + 3)^2 = 7$$

is the equation of a circle of radius  $\sqrt{7}$  centered at  $(-2, -3)$ .

**Example.** The equation

$$x^2 + y^2 = 1$$

is the equation of a circle of radius 1 centered at the origin. More generally, let  $r$  be a number  $> 0$ . The equation

$$x^2 + y^2 = r^2$$

is the equation of a circle of radius  $r$  centered at the origin. We can draw this circle as in Fig. 8-15.

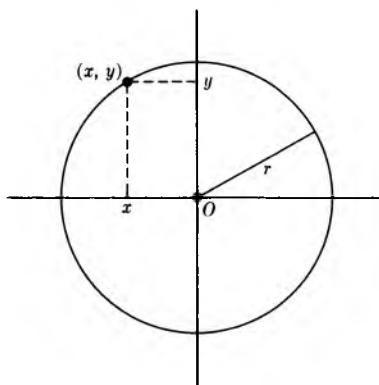


Fig. 8-15

In general, let  $a, b$  be two numbers, and  $r$  a number  $> 0$ . Then the equation of the circle of radius  $r$ , centered at  $(a, b)$ , is the equation

$$(x - a)^2 + (y - b)^2 = r^2.$$

This means that the circle is the set of all points satisfying this equation.

## EXERCISES

Write down the equation of a circle centered at the indicated point  $P$ , with radius  $r$ .

1.  $P = (-3, 1), r = 2$

2.  $P = (1, 5), r = 3$

3.  $P = (-1, -2), r = \frac{1}{3}$

4.  $P = (-1, 4), r = \frac{2}{5}$

5.  $P = (3, 3), r = \sqrt{2}$

6.  $P = (0, 0), r = \sqrt{8}$

Give the coordinates of the center of the circle defined by the following equations, and also give the radius.

7.  $(x - 1)^2 + (y - 2)^2 = 25$

8.  $(x + 7)^2 + (y - 3)^2 = 2$

9.  $(x + 1)^2 + (y - 9)^2 = 8$

10.  $(x + 1)^2 + y^2 = \frac{5}{3}$

11.  $(x - 5)^2 + y^2 = 10$

12.  $x^2 + (y - 2)^2 = \frac{3}{2}$

In each one of the following cases, we give an equation for a pair of numbers  $(x, y)$ . Show that the set of all points  $(x, y)$  satisfying each equation is a circle. Give the center of the circle and its radius. [Hint: Complete the square to transform the equation into an equivalent one of the type studied above.]

13.  $x^2 + 2x + y^2 = 5$

14.  $x^2 + y^2 - 3y - 7 = 0$

15.  $x^2 + 4x + y^2 - 4y = 20$

16.  $x^2 - 4x + y^2 - 2y + 1 = 0$

17.  $x^2 - 2x + y^2 + 5y = 26$

18.  $x^2 + x + y^2 - 3y - 4 = 0$

## The case of 3-space.

19. a) Write down the equation for a sphere of radius 1 centered at the origin in 3-space, in terms of the coordinates  $(x, y, z)$ .  
b) Same question for a sphere of radius 3.  
c) Same question for a sphere of radius  $r$ .
20. Write down the equation of a sphere centered at the given point  $P$  in 3-space, with the given radius  $r$ .
  - a)  $P = (1, -3, 2)$  and  $r = 1$
  - b)  $P = (-1, 5, 3)$  and  $r = 2$
  - c)  $P = (-1, 1, 4)$  and  $r = 3$
  - d)  $P = (1, 2, -5)$  and  $r = 1$
  - e)  $P = (-2, -1, -3)$  and  $r = 2$
  - f)  $P = (1, 3, 1)$  and  $r = 7$
21. In each of the following cases, write down the center of the sphere with the given equation, and write down its radius.
  - a)  $(x - 2)^2 + y^2 + z^2 = 25$
  - b)  $x^2 + y^2 + z^2 = 1$

- c)  $x^2 + (y - 3)^2 + (z - 10)^2 = 3$   
d)  $(x + 3)^2 + y^2 + (z + 2)^2 = 8$   
e)  $(x - 6)^2 + (y + 4)^2 + (z + 7)^2 = 2$   
f)  $(x - 4)^2 + (y - 5)^2 + z^2 = 11$

#### §4. RATIONAL POINTS ON A CIRCLE

The result proved in this section is not essential for what follows, and may be skipped. It is, however, quite beautiful.

Let us go back to Pythagoras. We ask whether we can describe all right triangles with sides having lengths  $a, b, c$  such that

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

and  $a, b, c$  are integers. For instance, we have a right triangle with sides 3, 4, 5, because

$$3^2 + 4^2 = 25 = 5^2.$$

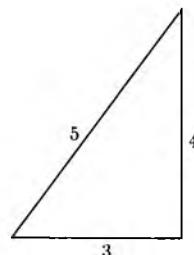


Fig. 8-16

First, it is not clear what we mean by "describe all such right triangles". Is there even another one? The answer to that is yes; for instance, a right triangle with sides (8, 15, 17). By experimenting, we might find still another, and the first question that arises is: Are there infinitely many? The answer to that is again yes, but is not immediately clear, although after we get through with our discussion, we shall see at once that there are infinitely many.

We shall transform our problem to an equivalent one. Observe that our equation

$$(*) \quad a^2 + b^2 = c^2$$

is true if and only if

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

All we need to do is to divide by  $c^2$  on both sides. If  $a, b, c$  are integers, then the quotients  $a/c$  and  $b/c$  are rational numbers. Thus every solution of equation (\*) yields a solution of the equation

$$(**) \quad x^2 + y^2 = 1$$

with rational numbers  $x$  and  $y$ .

Conversely, suppose that  $x, y$  are rational numbers satisfying equation (\*\*). Express  $x, y$  as fractions over a common denominator, say  $c$ . Thus write

$$x = \frac{a}{c} \quad \text{and} \quad y = \frac{b}{c}$$

with integers  $a, b, c$  such that  $c \neq 0$ . Then

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

If we multiply both sides of this equation by  $c^2$ , then we obtain (\*), namely,

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

Thus to solve (\*) in integers, it suffices to solve (\*\*) in rational numbers, and this is what we shall do.

A point  $(x, y)$  satisfying the equation  $x^2 + y^2 = 1$  may be viewed as a point on the circle of radius 1, centered at the origin. Thus to solve (\*\*) in rational numbers, we may say that we want to find all rational points on the circle. (A rational point  $(x, y)$  is by definition a point such that its coordinates  $x, y$  are rational numbers.)

Our next step is to give examples of such points.

For any number  $t$ , let

$$x = \frac{1 - t^2}{1 + t^2} \quad \text{and} \quad y = \frac{2t}{1 + t^2}.$$

Simple algebraic manipulations will show you that

$$x^2 + y^2 = 1.$$

If we give  $t$  special values which are rational numbers, or integers, then both  $x$  and  $y$  will be rational numbers, and this gives us our desired examples.

**Example.** Let  $t = 2$ . Then

$$x = \frac{1 - 4}{1 + 4} = -\frac{3}{5} \quad \text{and} \quad y = \frac{2 \cdot 2}{1 + 4} = \frac{4}{5}.$$

This yields an example which we already had, namely

$$\left(\frac{-3}{5}\right)^2 + \left(\frac{4}{5}\right)^2 = 1.$$

Multiplying by 25 on both sides yields the relation

$$3^2 + 4^2 = 5^2,$$

and thus we recover the (3, 4, 5) right triangle.

**Example.** Let  $t = 4$ . Then computing the values of  $x$  and  $y$  will show you that we recover the right triangle with sides (8, 15, 17).

**Example.** Let  $t = 5$ . Then

$$x = \frac{1 - 25}{1 + 25} = \frac{-24}{26} \quad \text{and} \quad y = \frac{2 \cdot 5}{1 + 25} = \frac{10}{26}.$$

Simplifying the fractions, we find that  $x = -\frac{12}{13}$  and  $y = \frac{5}{13}$ . This corresponds to a right triangle with sides 12, 5, and 13. Observe that

$$5^2 + 12^2 = 13^2.$$

It is clear that we have found a way of getting lots of rational points on the circle, or equivalently, lots of right triangles with integral sides. It is not difficult to show that two different values of  $t$  yield different points  $(x, y)$ . (Can you prove this as an exercise?)

You can ask: How did we guess the formulas expressing  $x$  and  $y$  in terms of  $t$  in the first place? Answer: These formulas have been known for a long time. As far as I know, history does not tell us who discovered them first, but he was a good mathematician. What distinguishes someone with talent for mathematics from someone without talent is that the first person will be able to discover such beautiful formulas and the second person will not.

However, everybody is able to plug numbers in the formula once it is written down. That does not take much talent.

We still have not solved our problem of rational points on the circle completely, namely we can ask: Are the points described by our two formulas

$$x = \frac{1 - t^2}{1 + t^2} \quad \text{and} \quad y = \frac{2t}{1 + t^2}$$

with rational values for  $t$ , the only rational points on the circle? In other words, does plugging rational numbers for  $t$  in these formulas yield all points

$(x, y)$  with rational  $x, y$  lying on the circle? The answer is yes, with only one exception. It is based on the following result.

**Theorem 1.** Let  $(x, y)$  be a point satisfying the equation

$$x^2 + y^2 = 1$$

and such that  $x \neq -1$ . (See Fig. 8-17.)

Let

$$t = \frac{y}{x+1}.$$

Then

$$x = \frac{1-t^2}{1+t^2} \quad \text{and} \quad y = \frac{2t}{1+t^2}.$$

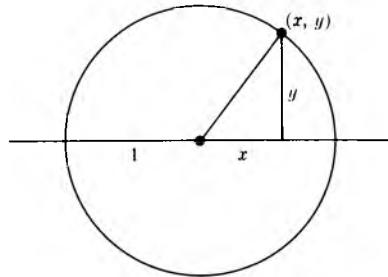


Fig. 8-17

*Proof.* Multiplying both sides of the equation  $t = y/(x+1)$  by  $x+1$ , we find that

$$t(x+1) = y.$$

Squaring yields

$$t^2(x+1)^2 = y^2,$$

which gives

$$t^2(x+1)^2 = 1 - x^2 = (1+x)(1-x).$$

We cancel  $(x+1)$  from both sides, and find

$$t^2(x+1) = 1 - x.$$

Expanding the left-hand side yields

$$t^2x + t^2 = 1 - x,$$

and also

$$(t^2 + 1)x = 1 - t^2.$$

Dividing by  $t^2 + 1$  gives us our expression for  $x$ . Using

$$y = t(x+1)$$

and an easy algebraic manipulation which we leave to you, we find the expression for  $y$ , namely

$$y = \frac{2t}{1 + t^2}.$$

This proves our theorem.

In Theorem 1, suppose that  $x, y$  are rational numbers. Then

$$t = \frac{y}{x + 1}$$

is also a rational number. From the expression for  $x$  and  $y$  in terms of  $t$  we conclude:

*Corollary.* *Let  $x, y$  be rational numbers such that*

$$x^2 + y^2 = 1.$$

*If  $x \neq -1$ , then there exists a rational number  $t$  such that*

$$x = \frac{1 - t^2}{1 + t^2} \quad \text{and} \quad y = \frac{2t}{1 + t^2}.$$

So we have completely described the rational points on the circle, or equivalently, the right triangles with integral sides.

We could now ask further questions, like:

*Determine all pairs of rational numbers  $(x, y)$  such that*

$$x^3 + y^3 = 1.$$

This is harder, but it can be shown that the only solutions are

$$x = 0 \quad \text{or} \quad x = 1,$$

with the obvious corresponding value for  $y$ . In general, given a positive integer  $k$ , the problem is to find all solutions of Fermat's equation:

$$x^k + y^k = 1;$$

say, with positive  $x$  and  $y$ . It is known for many values of  $k$  that there is no solution other than  $x = 0$  or  $x = 1$ , but a solution in general is unknown. This is the famous Fermat problem.

**EXERCISES**

1. Write down explicitly five examples of positive integers  $(a, b, c)$  such that

$$a^2 + b^2 = c^2,$$

which have not already been listed in the text and which are not multiples of those listed in the text.

2. Prove that if  $s, t$  are real numbers such that  $0 \leq s < t$ , then

$$\frac{1 - s^2}{1 + s^2} > \frac{1 - t^2}{1 + t^2}.$$

[Hint: Prove appropriate inequalities for the numerators and denominators, before taking the quotient.] This proves that different values for  $t > 0$  already give different values for  $x$ .

3. Using the formulas of this section, give explicitly the values of  $x$  and  $y$  as quotients of integers, when  $t$  has the following values:

a)  $t = \frac{1}{2}$ ,      b)  $t = \frac{1}{3}$ ,      c)  $t = \frac{1}{4}$ ,      d)  $t = \frac{1}{5}$ .

4. When  $t$  becomes very large positive, what happens to

$$\frac{1 - t^2}{1 + t^2}?$$

When  $t$  becomes very large negative, what happens to

$$\frac{1 - t^2}{1 + t^2}?$$

Substitute large values of  $t$ , like 10,000 or  $-10,000$ , to get a feeling for what happens.

5. Analyze what happens to

$$\frac{2t}{1 + t^2}$$

when  $t \leq 0$  and when  $t$  becomes very large negative. Next analyze what happens when  $t \geq 0$  and  $t$  becomes very large positive.



# 9 Operations on Points

## §1. DILATIONS AND REFLECTIONS

From now on, unless otherwise specified, we deal with a fixed coordinate system. Thus we make no distinction between a point and its associated coordinates.

We use  $\mathbf{R}$  to denote the set of all real numbers, and  $\mathbf{R}^2$  to denote the set of all pairs  $(x, y)$ , where  $x, y$  are real numbers. Thus a point of the plane is simply an element of  $\mathbf{R}^2$ .

Let  $A$  be a point in the plane, with coordinates

$$A = (a_1, a_2).$$

If  $c$  is any real number, we define the product  $cA$  to be the point

$$cA = (ca_1, ca_2).$$

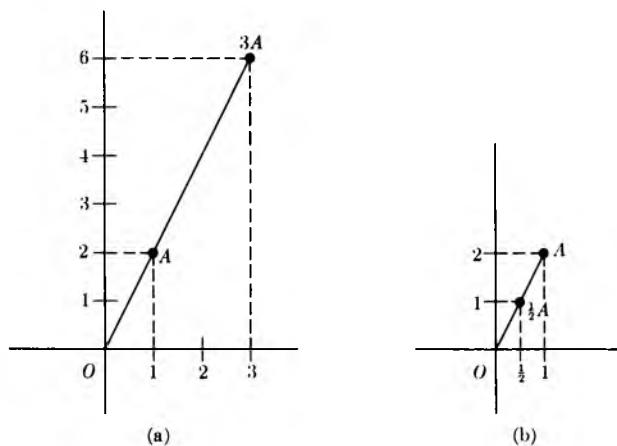
Thus we multiply each coordinate of  $A$  by  $c$  to get the coordinates of  $cA$ .

**Example.** Let  $A = (2, 5)$  and  $c = 6$ . Then  $cA = (12, 30)$ .

**Example.** Let  $A = (-3, 7)$  and  $c = -4$ . Then  $cA = (12, -28)$ .

We shall now interpret this multiplication geometrically.

**Example.** Suppose that  $c$  is positive, and let us draw the picture, with the point  $A = (1, 2)$  and  $c = 3$ . Then  $3A = (3, 6)$ , as on Fig. 9-1(a).



**Fig. 9-1**

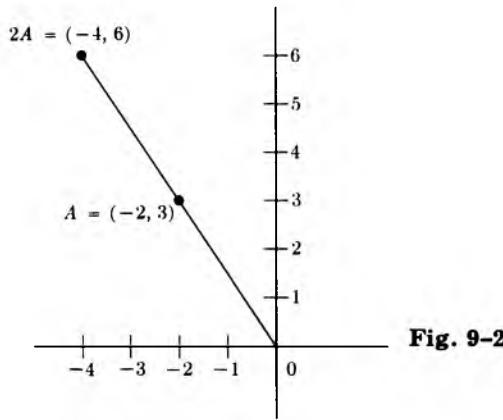
Geometrically, we see that multiplication by 3 stretches the coordinates by 3. Similarly, if  $r$  is a positive number, then

$$rA = (r, 2r),$$

and we see that multiplication by  $r$  stretches the coordinates by  $r$ . For instance, stretching by  $\frac{1}{2}$  amounts to halving, e.g. in Fig. 9-1(b),

$$\frac{1}{2}A = (\frac{1}{2}, 1).$$

**Example.** Let  $A = (-2, 3)$  and  $r = 2$ . Then  $rA = (-4, 6)$ . Picture:



If  $r$  is a positive number, we call  $rA$  the **dilation** of  $A$  by  $r$ . The association

$$A \mapsto rA$$

which to each point  $A$  associates  $rA$  is called **dilation** by  $r$ , and gives us an analytic definition for the concept introduced in Chapter 6. It is dilation with respect to  $O$ , and leaves  $O$  fixed.

Next, we consider the case when  $c$  is negative. If  $A = (a_1, a_2)$ , we define  $-A$  to be  $(-1)A$ , so that

$$-A = (-a_1, -a_2).$$

**Example.** Let  $A = (1, 2)$ . Then  $-A = (-1, -2)$ . We represent  $-A$  in Fig. 9-3.

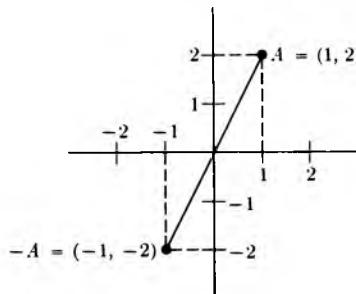


Fig. 9-3

**Example.** Let  $A = (-2, 3)$ . Then  $-A = (2, -3)$ . We draw  $A$  and  $-A$  in Fig. 9-4.

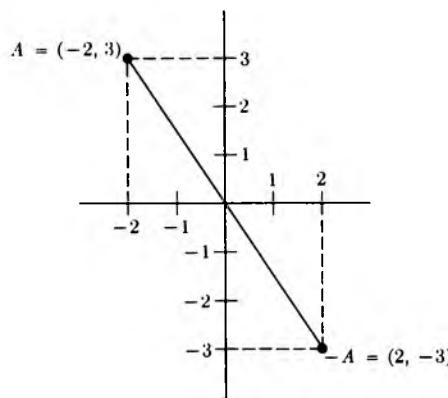


Fig. 9-4

We see that  $-A$  is obtained by a certain symmetry, which justifies the next definition.

We define **reflection through our given origin  $O$**  to be the association which to each point  $A$  associates the point  $-A$ . As usual, this association is denoted by

$$A \mapsto -A.$$

If  $R$  denotes reflection through  $O$ , then

$$R(A) = -A.$$

Thus we have been able to give a definition of reflection using only numbers and their properties, i.e. an analytic definition.

If  $c$  is negative, we write  $c = -r$ , where  $r$  is positive, and we see that multiplication of  $A$  by  $c$  can be obtained by first multiplying  $A$  by  $r$  and then taking the reflection  $-rA$ . Thus we can say that  $-rA$  points in the opposite direction from  $A$ , with a stretch of  $r$ . We have drawn an example with  $A = (1, 2)$  and  $c = -3$  in Fig. 9-5.

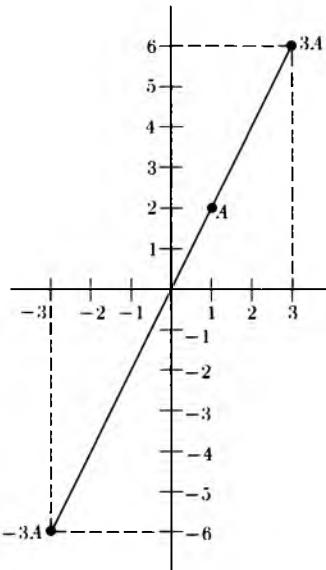


Fig. 9-5

We now consider the effect of a dilation on distances.

**Theorem 1.** Let  $r$  be a positive number. If  $A, B$  are points, then

$$d(rA, rB) = r \cdot d(A, B).$$

*Proof.* This was assigned as an exercise in the preceding chapter. We work it out here, so that you see how simple it is. Let  $A = (a_1, a_2)$  and  $B = (b_1, b_2)$  as usual. Then  $rA = (ra_1, ra_2)$  and  $rB = (rb_1, rb_2)$ . Hence

$$\begin{aligned} d(rA, rB)^2 &= (rb_1 - ra_1)^2 + (rb_2 - ra_2)^2 \\ &= (r(b_1 - a_1))^2 + (r(b_2 - a_2))^2 \\ &= r^2(b_1 - a_1)^2 + r^2(b_2 - a_2)^2 \\ &= r^2 \cdot d(A, B)^2. \end{aligned}$$

Taking the square root proves our theorem.

What happens to the distance under multiplication of points by a negative number  $c$ ? Recall that the absolute value of a number  $c$  is defined to be

$$|c| = \sqrt{c^2}.$$

Thus  $|-3| = \sqrt{(-3)^2} = \sqrt{9} = 3$ .

**Theorem 2.** *Let  $c$  be a number. Then*

$$d(cA, cB) = |c| \cdot d(A, B).$$

*Proof.* The proof follows exactly the same pattern as the proof of Theorem 1, except that, at the very end when we take  $\sqrt{c^2}$  instead of  $r^2$ , we find  $|c|$  instead of  $r$ . Write this proof out in full.

## EXERCISES

1. Write the coordinates for  $cA$  for the following values of  $c$  and  $A$ . In each case, draw  $A$  and  $cA$ .
  - $A = (-3, 5)$  and  $c = 4$
  - $A = (4, -2)$  and  $c = 3$
  - $A = (-4, -5)$  and  $c = 2$
  - $A = (2, -3)$  and  $c = -2$
  - $A = (-4, -5)$  and  $c = -1$
  - $A = (2, -3)$  and  $c = 2$
2. Let  $A$  be a point,  $A \neq O$ . If  $b, c$  are numbers such that  $bA = cA$ , prove that  $b = c$ .

3. Prove that reflection through  $O$  preserves distances. In other words, prove that

$$d(A, B) = d(-A, -B).$$

4. **The 3-dimensional case**

- a) Define the multiplication (dilation) of a point  $A = (a_1, a_2, a_3)$  by a number  $c$ . Write out interpretations for this similar to those we did in the plane. Draw pictures.
- b) Define reflection of  $A$  through  $O = (0, 0, 0)$ .
- c) State and prove the analogs of Theorems 1 and 2.

## §2. ADDITION, SUBTRACTION, AND THE PARALLELOGRAM LAW

Let  $A$  and  $B$  be points in the plane. We write their coordinates,

$$A = (a_1, a_2) \quad \text{and} \quad B = (b_1, b_2).$$

We define their sum  $A + B$  to be

$$A + B = (a_1 + b_1, a_2 + b_2).$$

Thus we define their sum componentwise.

**Example.** Let  $A = (1, 4)$  and  $B = (-1, 5)$ . Then

$$A + B = (1 - 1, 4 + 5) = (0, 9).$$

**Example.** Let  $A = (-3, 6)$  and  $B = (-2, -7)$ . Then

$$A + B = (-3 - 2, 6 - 7) = (-5, -1).$$

This addition satisfies properties similar to the addition of numbers — and no wonder, since the coordinates of a point are numbers. Thus we have for any points  $A, B, C$ :

**Commutativity.**  $A + B = B + A$ .

**Associativity.**  $A + (B + C) = (A + B) + C$ .

**Zero element.** Let  $O = (0, 0)$ . Then  $A + O = O + A = A$ .

**Additive inverse.** If  $A = (a_1, a_2)$  then the point

$$-A = (-a_1, -a_2)$$

is such that

$$A + (-A) = O.$$

These properties are immediately proved from the definitions. For instance, let us prove the first one. We have:

$$A + B = (a_1 + b_1, a_2 + b_2) = (b_1 + a_1, b_2 + a_2) = B + A.$$

Our proof simply reduces the property concerning points to the analogous property concerning numbers. The same principle applies to the other properties. Note especially the additive inverse. For instance,

$$\text{if } A = (2, -5), \text{ then } -A = (-2, 5).$$

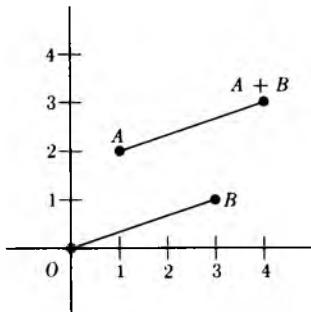
As with numbers, we shall write  $A - B$  instead of  $A + (-B)$ .

**Example.** If  $A = (2, -5)$  and  $B = (3, -4)$ , then

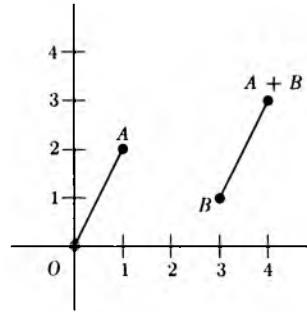
$$A - B = (2 - 3, -5 - (-4)) = (-1, -1).$$

We shall now interpret this addition and subtraction geometrically. We consider examples.

**Example.** Let  $A = (1, 2)$  and  $B = (3, 1)$ . To find  $A + B$ , we start at  $A$ , go 3 units to the right, and 1 unit up, as shown in Fig. 9-6(a) and (b).



(a)



(b)

Fig. 9-6

Thus we see geometrically that  $A + B$  is obtained from  $A$  by the same procedure as  $B$  is obtained from  $O$ . In this geometric representation, we also see that the line segment between  $A$  and  $A + B$  is parallel to the line segment between  $O$  and  $B$ , as shown in Fig. 9-6(a). Similarly, the line segment between  $O$  and  $A$  is parallel to the line segment between  $B$  and  $A + B$ . Thus the four points

$$O, A, B, A + B$$

form the four corners of a parallelogram, which we draw in Fig. 9-6(c).

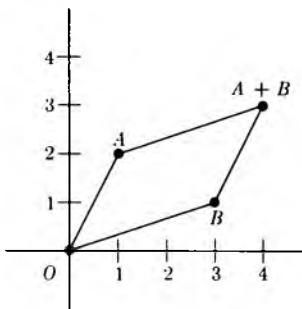


Fig. 9-6 (cont.)

(c)

This gives us a geometric interpretation of addition.

Next, we consider subtraction.

**Example.** Let  $A = (1, 2)$  and  $B = (3, 1)$ . Then  $A - B = (-2, 1)$ . By definition,  $A - B = A + (-B)$ . Thus we can represent this subtraction as in Fig. 9-7.

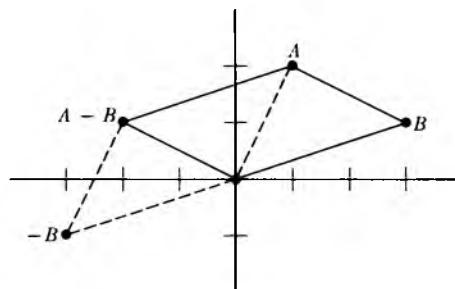


Fig. 9-7

The four points  $O, A, B, A - B$  are still the four corners of a parallelogram, but starting from  $A$  we have to move in the opposite direction to go from  $A$  to  $A - B$  than when we moved from  $A$  to  $A + B$ .

Observe that the point  $A$  can be written in the form

$$A = (A - B) + B.$$

Thus we also obtain a parallelogram whose corners are  $O, A - B, A, B$ .

Let  $A$  be a fixed element of  $\mathbf{R}^2$ . We define the **translation** by  $A$  to be the association which to each point  $P$  of the plane associates the point  $P + A$ . In Fig. 9-8, we have drawn the effect of this translation on several points.

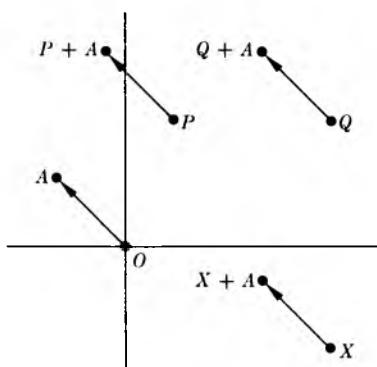


Fig. 9-8

The association

$$P \mapsto P + A$$

has been represented by arrows in Fig. 9-8.

It is useful to abbreviate “translation by  $A$ ” by the symbol  $T_A$ . You should read Chapter 6, §1, which now applies to the present situation. By definition, the value of  $T_A$  at a point  $P$  is

$$T_A(P) = P + A.$$

**Example.** Let  $A = (-2, 3)$  and  $P = (1, 5)$ . Then

$$T_A(P) = P + A = (-1, 8).$$

We see that we have been able to define one more of the intuitive geometric notions within our system of coordinates, based only on properties of numbers.

So far, we have given analytic definitions (i.e. definitions based only on properties of numbers) for points, distance, reflection through  $O$ , and translation. We recall that a **mapping** of the plane into itself is an association which to each point  $P$  associates another point. If the mapping is denoted by  $F$ , then this other point is denoted by  $F(P)$ , and is called the **value** of  $F$  at  $P$ . If  $F$  is translation by  $A$ , i.e. if  $F = T_A$ , then  $P + A$  is the value of  $F$  at  $P$ .

Our previous definition of isometry now makes sense analytically: An **isometry** is a mapping of the plane into itself which preserves distances. In other words,  $F$  is an isometry if and only if, for every pair of points  $P, Q$ , we have

$$d(P, Q) = d(F(P), F(Q)).$$

Since we have introduced addition and subtraction for points, we shall now describe a way of expressing the distance between points by using subtraction.

We recall that the distance between two points  $P, Q$  is denoted by  $d(P, Q)$ . We shall use a special symbol for the distance between a point and the origin, namely the absolute value sign. We let

$$d(A, O) = |A|.$$

Thus we use two vertical bars on the sides of  $A$ . If  $A = (a_1, a_2)$ , then

$$|A| = \sqrt{a_1^2 + a_2^2},$$

and therefore

$$|A|^2 = a_1^2 + a_2^2.$$

We call  $|A|$  the **norm** of  $A$ . The norm generalizes the absolute value of a number. We can represent the norm of  $A$  as in Fig. 9-9. Geometrically, it is the length of the line segment from  $O$  to  $A$ .

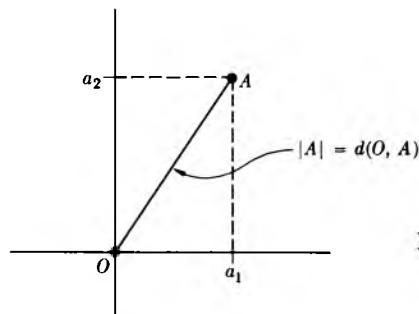


Fig. 9-9

Note that  $|A| = |-A|$ . Proof?

Using our addition of points, or rather subtraction, we can express the distance between two points  $A, B$  by

$$d(A, B) = |A - B| = |B - A|.$$

Indeed, if  $A = (a_1, a_2)$  and  $B = (b_1, b_2)$ , then

$$\begin{aligned} d(A, B) &= \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2} = |B - A| \\ &= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} = |A - B|. \end{aligned}$$

With this notation, we see that a mapping  $F$  of the plane into itself is an isometry if and only if for every pair of points  $P, Q$  we have

$$|F(P) - F(Q)| = |P - Q|.$$

With this notation, a special case of Theorem 2 of §1 can now be written:

$$|cA| = |c| |A|.$$

If  $r$  is a positive number, then

$$|rA| = r|A|.$$

We suggest that you read the section on mappings in Chapter 6 if you have not already done so. We recall some definitions. If  $F$  is a mapping of the plane into itself, and  $P$  is a point, then  $F(P)$  is also called the **image** of  $P$  under  $F$ . If  $S$  is a subset of the plane, then we denote by  $F(S)$  the set of all points  $F(P)$  with  $P$  in  $S$ , and call  $F(S)$  the **image** of  $S$  under  $F$ .

We can now prove analytically a result which was intuitively clear.

**Theorem 3.** *The circle of radius  $r$  and center  $A$  is the translation by  $A$  of the circle of radius  $r$  and center  $O$ .*

*Proof.* Let  $X$  be a point on the circle of radius  $r$  and center  $O$ . This means that

$$|X| = r.$$

The translation of  $X$  by  $A$ , which is  $X + A$ , satisfies the condition

$$|X + A - A| = r.$$

Thus we see that  $X + A$  is at distance  $r$  from  $A$ , and hence lies on the circle of radius  $r$  centered at  $A$ . Conversely, given a point  $Y$  on this circle, so that

$$|Y - A| = r,$$

let  $X = Y - A$ . Then  $Y = X + A$  is the translation of  $X$  by  $A$ , and  $|X| = r$ . Therefore every point on the circle of radius  $r$ , centered at  $A$ , is the

image under  $T_A$  of a point on the circle of radius  $r$  centered at  $O$ . This proves our theorem.

Actually, you can also do Exercise 11, and then the proof given for Theorem 7 of Chapter 6, §6 is seen to be essentially the same proof as that given above.

Finally, we make a remark concerning the relation between addition and the multiplication of points by numbers as in §1. We ask whether the ordinary rules which we had for numbers also apply, and the answer is yes. Namely, we have:

**Associativity.** *If  $b, c$  are numbers, then  $b(cA) = (bc)A$ .*

**Distributivity.** *If  $b, c$  are numbers, and  $A, B$  are points, then*

$$(b + c)A = bA + cA \quad \text{and} \quad c(A + B) = cA + cB.$$

*Also,*

$$1A = A \quad \text{and} \quad 0A = O.$$

The proofs are easy, since we can reduce each statement to the analogous property for numbers. For instance, to prove one of the distributivities, we have:

$$\begin{aligned} (b + c)A &= ((b + c)a_1, (b + c)a_2) && \text{by definition} \\ &= (ba_1 + ca_1, ba_2 + ca_2) && \text{by distributivity for numbers} \\ &= (ba_1, ba_2) + (ca_1, ca_2) && \text{by definition} \\ &= bA + cA. \end{aligned}$$

We leave the proofs of the other properties to you. They are just as easy, or easier.

## EXERCISES

Plot the points  $A, B, A + B$ , drawing appropriate parallelograms.

- |                               |                                 |
|-------------------------------|---------------------------------|
| 1. $A = (1, 4), B = (3, 2)$   | 2. $A = (1, 5), B = (1, 1)$     |
| 3. $A = (-1, 2), B = (3, 1)$  | 4. $A = (-2, 1), B = (1, 2)$    |
| 5. $A = (-1, 1), B = (-1, 2)$ | 6. $A = (-3, -2), B = (-1, -1)$ |

7.  $A = (-2, -1)$ ,  $B = (-3, 5)$       8.  $A = (-4, -1)$ ,  $B = (1, -3)$

9.  $A = (2, -3)$ ,  $B = (-1, -2)$       10.  $A = (2, -3)$ ,  $B = (-1, 5)$

In each of the preceding exercises, plot the points  $A$ ,  $-B$ , and  $A - B$ .

11. Let  $T_A$  be a translation by  $A$ . Prove that it is an isometry, in other words, that for any pair of points  $P, Q$ , we have

$$d(P, Q) = d(T_A(P), T_A(Q)).$$

12. Let  $D(r, A)$  denote the disc of radius  $r$  centered at  $A$ . Show that  $D(r, A)$  is the translation by  $A$  of the disc  $D(r, O)$  of radius  $r$  centered at  $O$ . [For the definition of the disc, cf. Chapter 5, and observe that this definition is now analytic since all terms entering in it have been defined analytically.]

13. Let  $S(r, P)$  denote the circle of radius  $r$  centered at  $P$ .

- a) Show that the reflection of this circle through  $O$  is again a circle.  
What is the center of the reflected circle?  
b) Show that the reflection of the disc  $D(r, P)$  through  $O$  is a disc.  
What is the center of this reflected disc?

14. Let  $P, Q$  be points. Write  $P = Q + A$ , where  $A = P - Q$ . Define the **reflection of  $P$  through  $Q$**  to be the point  $Q - A$ . If  $R_Q$  denotes reflection through  $Q$ , then we have  $R_Q(P) = 2Q - P$ . (Why?) Draw the picture, showing  $P, Q, A$ , and  $Q - A$  to convince yourself that this definition corresponds to our geometric intuition.

15. a) Prove that reflection through a point  $Q$  can be expressed in terms of reflection through  $O$ , followed by a translation.

- b) Let  $T_A$  be translation by  $A$ , and  $R_O$  reflection with respect to the origin. Prove that the composite  $T_A \circ R_O$  is equal to  $R_Q$  for some point  $Q$ . Which one?

16. a) Let  $r$  be a positive number. Give an analytic definition of **dilation by  $r$  with respect to a point  $Q$** , and denote this dilation by  $F_{r,Q}$ . To give this definition, look at Exercise 14. You may also want to look at the discussion about line segments in §4. If  $P$  is a point, draw the picture with  $O, P, Q, P - Q$ , and  $F_{r,Q}(P)$ .

- b) From your definition, it should be clear that  $F_{r,Q}$  can be obtained as a composite of dilation with respect to  $O$ , and a translation. Translation by what point?

17. Let  $S(r, A)$  be the circle of radius  $r$  and center  $A$ . Show that the reflection of this circle through a point  $Q$  is a circle. What is the center of this reflected circle? What is its radius? Draw a picture.

Let  $F$  be a mapping of the plane into itself. Recall that the inverse mapping  $G$  of  $F$  (if it exists) is the mapping such that  $F \circ G = G \circ F = I$  (the identity mapping). This inverse mapping is denoted by  $F^{-1}$ . This is all you need to know for Exercises 18 through 21.

18. The inverse of the translation  $T_A$  is also a translation. By what? Prove your assertion.
19. Let  $F_r$  be dilation by a positive number  $r$ , with respect to  $O$ , and let  $T_A$  be translation by  $A$ .
  - a) Show that  $F_r^{-1}$  is also a dilation. By what number?
  - b) Show that  $F_r \circ T_A \circ F_r^{-1}$  is a translation.
20. Show that the composite of two translations is a translation. If  $T_A \circ T_B = T_C$ , how would you express  $C$  in terms of  $A$  and  $B$ ?
21. Let  $R$  be reflection through the origin.
  - a) Show that  $R^{-1}$  exists.
  - b) Show that  $R \circ T_A \circ R^{-1}$  is a translation. By what?
22. Let  $A = (a_1, a_2)$  be a point. Define its **reflection through the  $x$ -axis** to be the point  $(a_1, -a_2)$ . Draw  $A$  and its reflection through the  $x$ -axis in the following cases.
  - a)  $A = (1, 2)$
  - b)  $A = (-1, 3)$
  - c)  $A = (-2, -4)$
  - d)  $A = (5, -2)$
23. Prove that reflection through the  $x$ -axis is an isometry.
24. Define reflection through the  $y$ -axis in a similar way, and prove that it is an isometry. Draw the points of Exercise 22, and their reflections through the  $y$ -axis.
25. Recall that a **fixed point** of a mapping  $F$  is a point  $P$  such that  $F(P) = P$ . Using the coordinate definition, determine the fixed points of
  - a) a translation,
  - b) reflection through  $O$ ,
  - c) reflection through an arbitrary point  $P$ ,
  - d) reflection through the  $x$ -axis and through the  $y$ -axis.
 Prove your assertions.
26. a) Let
 
$$E_1 = (1, 0) \quad \text{and} \quad E_2 = (0, 1).$$

We call  $E_1$  and  $E_2$  the **basic unit points** of the plane. Plot these points. If  $A = (a_1, a_2)$ , prove that

$$A = a_1 E_1 + a_2 E_2.$$

- b) If  $c$  is a number, what are the coordinates of  $cE_1, cE_2$ ?

27. Let  $A = (2, 3)$ . Draw the points

$$A, \quad A + E_1, \quad A + E_2, \quad A + E_1 + E_2.$$

28. Let  $A = (2, 3)$ . Draw the points

$$A, \quad A + 3E_1, \quad A + 3E_2, \quad A + 3E_1 + 3E_2.$$

29. Given a number  $r > 0$  and a point  $A$ , we can define the corners of a square, having sides of length  $r$  parallel to the axes, and  $A$  as its lower left-hand corner, to be the points

$$A, \quad A + rE_1, \quad A + rE_2, \quad A + rE_1 + rE_2.$$

Let  $s$  be a positive number. Show that if these four points are dilated by multiplication with  $s$ , they again form the corners of a square. What are the corners of this dilated square?

30. Let the notation be as in Exercise 29. What is the area of the dilated square? How does it compare with the area of the original square?
31. Let  $A$  be a point and  $r, s$  positive numbers. How would you define the corners of a rectangle whose sides are parallel to the axes, with  $A$  as the lower left-hand corner, and such that the vertical side has length  $r$  and the horizontal side has length  $s$ ?
32. Let  $t$  be a positive number. What is the effect of dilation by  $t$  on the sides and on the area of the rectangle in Exercise 31?
33. Let  $A$  be a point and let  $r = |A|$ . Assume that  $r \neq 0$ . What is the norm of  $(1/r)A$ ? Prove your assertion.
34. Do the exercise at the end of Chapter 7, §1.

### The 3-dimensional case

Define addition of points in 3-space componentwise. Verify the basic properties of commutativity, associativity, zero element, and additive inverse. Define translations. Verify that the sphere of radius  $r$  centered at a point  $A$  is the translation of the sphere of radius  $r$  centered at the origin  $O$ . The analog of the disc in 3-space is called the **ball**, for obvious reasons.

Define reflection of a point  $P$  through a point  $Q$ . Does the definition differ from the 2-dimensional case? Note that using notation  $P, Q, A, B$  without coordinates allows us to generalize at once certain notations from 2-space to 3-space. And higher. Why not?

Define an isometry of 3-space. Prove that translations and reflections are isometries.

Show that Exercises 18, 19, 20, and 21 apply to the 3-dimensional case.

Write all this up as if you were writing a book. Part of your mathematical training should consist of making you write mathematics in full English sentences. This forces you to think clearly, and is antidote to slapping down answers to routine plugging problems. As you will notice, carrying out the theory in 3-space amounts practically to copying the theory in 2-space. There is nothing wrong or harmful in copying mathematics. Do you know one of the means Bach used to learn how to compose? He copied practically the entire works of Vivaldi. Some 300 years later rock musicians still use approximately the same technique on each other.

# 10 Segments, Rays and Lines

## §1. SEGMENTS

### Line segments

Let  $P, Q$  be points in the plane. We can write  $Q = P + A$  for some  $A$ , namely  $A = Q - P$ . We wish to give an analytic definition of the line segment between  $P$  and  $Q$ , as shown in Fig. 10-1(a).

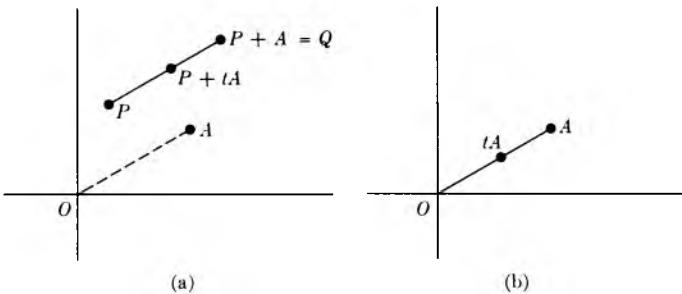


Fig. 10-1

This is easy. We define this segment to be the set of all points

$$P + tA, \quad \text{with } 0 \leq t \leq 1.$$

**Example.** The point halfway between  $P$  and  $P + A$  is the point

$$P + \frac{1}{2}A.$$

**Example.** The line segment between  $O$  and  $A$  consists of all dilations  $tA$  with  $0 \leq t \leq 1$ , as in Fig. 10-1(b). Thus we see that the line segment between  $P$  and  $P + A$  is simply the translation by  $P$  of the line segment between  $O$  and  $A$ .

The line segment between points  $P$  and  $Q$  is denoted by  $\overline{PQ}$ . Let  $Q = P + A$ . Then in the notation of mappings, we have

$$T_P(\overline{OA}) = \overline{PQ}.$$

The line segment  $\overline{PQ}$  is the image under  $T_P$  of the line segment  $\overline{OA}$ . The length of a line segment  $\overline{PQ}$  is simply the distance between  $P$  and  $Q$ .

The line segment between  $P$  and  $Q$  consists of all points

$$P + t(Q - P), \quad 0 \leq t \leq 1.$$

The above expression can be written in the form  $P + tQ - tP$ , or, in other words,

$$(1 - t)P + tQ.$$

Let  $s = 1 - t$  and  $t = 1 - s$ . When  $t$  takes on all values from 0 to 1, we see that  $s$  takes on all values from 1 to 0. The points of the line segment between  $P$  and  $Q$  can be written in the form

$$sP + (1 - s)Q, \quad 0 \leq s \leq 1.$$

Thus we see that the segment between  $P$  and  $Q$  consists of the same points as the segment between  $Q$  and  $P$ . Of course, we had a right to expect this. Thus we have

$$\overline{PQ} = \overline{QP}.$$

It does not matter which is written first,  $P$  or  $Q$ .

We can define another concept, that of **directed segment**, or **located vector**, in which the order *does* matter. Thus we define a **located vector** to be an ordered pair of points, which we denote by the symbols

$$\overrightarrow{PQ}.$$

The arrow on top means that  $P$  is the first point and  $Q$  is the second. We draw a located vector as an arrow, shown in Fig. 10-2.

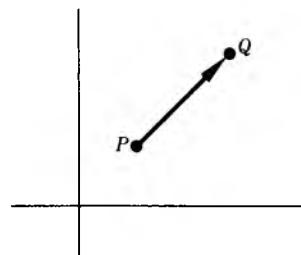


Fig. 10-2

We say that  $P$  is the **beginning point** and  $Q$  is the **end point** of the located vector. We say that the located vector is **located at  $P$** . Having ordered our points, we see that

$$\overrightarrow{PQ} \neq \overrightarrow{QP}.$$

## §2. RAYS

Let  $\overrightarrow{OA}$  be a located vector, located at the origin, such that  $A \neq O$ . Let  $P$  be a point. We define the **ray with vertex  $P$ , in the direction of  $\overrightarrow{OA}$** , to be the set of all points

$$(1) \quad P + tA, \quad t \geq 0.$$

Observe that the set of all points  $tA$ , with  $t \geq 0$  is a ray with vertex at the origin, in the direction of  $\overrightarrow{OA}$ , as on Fig. 10-3. Thus the ray with vertex  $P$  in the direction of  $\overrightarrow{OA}$  is the translation by  $P$  of the ray consisting of all points  $tA$  with  $t \geq 0$ .

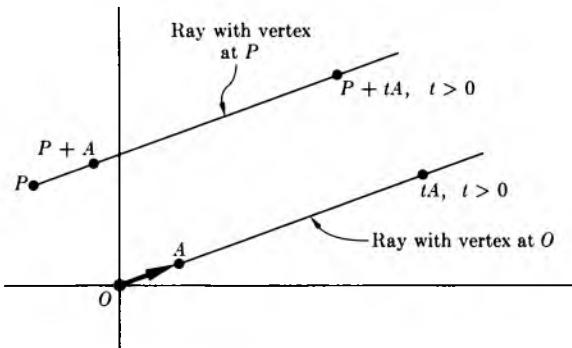


Fig. 10-3

Because the origin  $O$  has been fixed throughout our discussion, the mention of  $O$  when we speak of the ray in the direction of  $\overrightarrow{OA}$  is superfluous, and we shall also say that this ray has the **direction of  $A$** . Such a ray, with vertex  $P$ , is completely determined by the expression in (1), which involves simply  $P$  and  $A$  (and all numbers  $t \geq 0$ ).

**Example.** Let

$$P = (-1, 3)$$

and

$$A = (2, 1).$$

Letting  $t = 5$ , we see that the point

$$(-1, 3) + 5(2, 1) = (9, 8)$$

lies on the ray with vertex  $P$  in the direction of  $A$ . Similarly, letting  $t = \frac{1}{3}$ , we see that the point

$$(-1, 3) + \frac{1}{3}(2, 1) = \left(-\frac{1}{3}, \frac{10}{3}\right)$$

lies on this ray.

Given two points  $P, Q$  such that  $P \neq Q$ , we can define the **ray with vertex  $P$ , passing through  $Q$**  to be the ray with vertex  $P$  in the direction of  $Q - P$ . This ray consists therefore of all points

$$P + t(Q - P), \quad t \geq 0.$$

**Example.** Let  $P = (-1, 3)$  and  $Q = (2, 5)$ . Then  $Q - P = (3, 2)$ . The ray with vertex  $P$  passing through  $Q$  is shown in Fig. 10-4. It consists of all points

$$(-1, 3) + t(2, 5) = (-1 + 2t, 3 + 5t)$$

with  $t \geq 0$ .

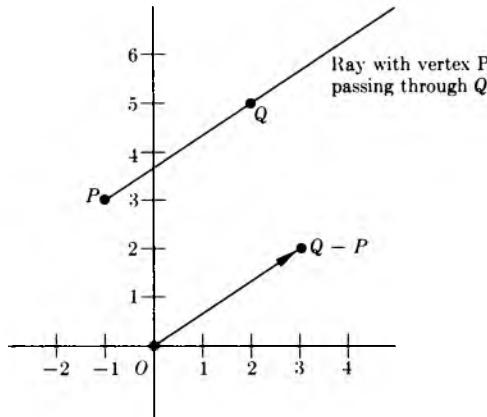


Fig. 10-4

**Remark.** Let  $A$  be a point  $\neq O$ . Let  $c$  be a positive number. Then the ray having a given vertex  $P$  in the direction of  $A$  is the same as the ray having this same vertex  $P$  in the direction of  $cA$ .

*Proof.* The first ray consists of all points

$$P + tA, \quad t \geq 0.$$

The second ray consists of all points

$$P + scA, \quad s \geq 0.$$

The point  $P$  is common to both of them, as we see by taking  $t = s = 0$ . Suppose that we have a point  $P + tA$  on the first with a given value of  $t$ . Let  $s = t/c$ . Then this point can be written in the form  $P + scA$ , with  $s \geq 0$ , and hence is also a point of the second ray. Conversely given a point on the second ray, we let  $t = sc$ , and therefore see that it is also a point on the first ray. Thus the two rays are equal.

Our remark, combined with our geometric intuition, leads us to make a definition. Let  $A \neq O$  and  $B \neq O$ . We say that  $\overrightarrow{OA}$  and  $\overrightarrow{OB}$  have the **same direction** (or also that  $A$  and  $B$  have the **same direction**) if there exists a number  $c > 0$  such that

$$B = cA.$$

Since we can then write  $A = (1/c)B$ , we see that  $A$  and  $B$  have the same direction if and only if each is a positive multiple of the other. We could also say that each is the dilation of the other by a positive number.

Similarly, let  $\overrightarrow{PQ}$  and  $\overrightarrow{MN}$  be located vectors. We say that they have the **same direction** if there exists a number  $c > 0$  such that

$$Q - P = c(N - M).$$

We draw located vectors having the same direction in Fig. 10-5.

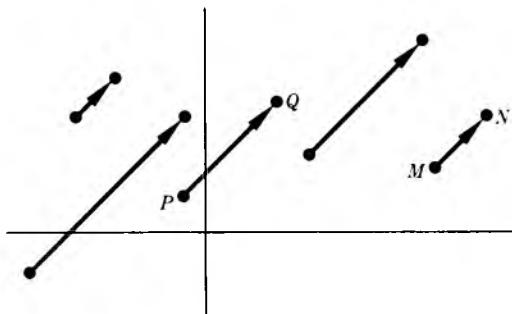


Fig. 10-5

**Example.** The two points  $A = (3, 5)$  and  $B = (9, 15)$  define the same direction. According to our convention, this means that  $\overrightarrow{OA}$  and  $\overrightarrow{OB}$  have the same direction. We see that  $B$  is just the dilation of  $A$  by 3.

**Example.** The located vectors  $\overrightarrow{PQ}$  and  $\overrightarrow{MN}$  have the same direction if

$$P = (1, 3), \quad Q = (-4, 2), \quad M = (7, 1), \quad N = (-3, -1)$$

because

$$Q - P = (-5, -1)$$

and

$$N - M = (-10, -2)$$

so that  $N - M = 2(Q - P)$ .

**Example.** Let

$$P = (-3, 5), \quad Q = (1, 2), \quad M = (4, 1).$$

Find the point  $N$  such that  $\overrightarrow{MN}$  has the same direction as  $\overrightarrow{PQ}$ , and such that the length of  $\overrightarrow{PQ}$  is the same as the length of  $\overrightarrow{MN}$ .

We need to find  $N$  such that

$$N - M = Q - P,$$

and hence

$$N = M + Q - P.$$

We can solve this easily by adding and subtracting components, and we get

$$\begin{aligned} N &= (4, 1) + (1, 2) - (-3, 5) \\ &= (8, -2). \end{aligned}$$

This solves our problem.

**Example.** In physics, located vectors are very useful to represent physical forces. For instance, suppose that a particle is at a point  $P$ , and that a force is acting on the particle with a certain magnitude and direction. We represent this direction by a located vector, and the magnitude by the length of this located vector. Thus when we draw the picture



Fig. 10-6

we can interpret it as a force acting on the particle at  $P$ .

Similarly, suppose that an airplane is located at  $O$  as in Fig. 10-7. We can interpret  $\overrightarrow{OA}$  as the force of the wind acting on the plane. If the pilot runs his engines with a certain force, and gives direction to the airplane by placing his rudders a certain way, we can also represent this force and direction by a located vector  $\overrightarrow{OB}$ .

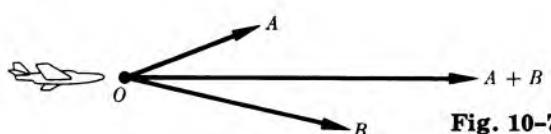


Fig. 10-7

It can be verified experimentally that under these two forces, the airplane moves in a manner described by addition as we defined it; namely, in the direction of  $A + B$ . Furthermore, the resultant of the two forces acting on the airplane has a magnitude equal to the norm of  $A + B$ , i.e.  $|A + B|$ . Thus our concepts of distance, located vectors, and vector sums are used constantly in the sciences.

### EXERCISES

Let  $P, Q$  be the indicated points. Give the coordinates of the point

- a) halfway,
  - b) one-third of the way,
  - c) two-thirds of the way
- between  $P$  and  $Q$ .

1.  $P = (1, 5), Q = (3, -1)$
2.  $P = (2, 4), Q = (3, 7)$
3.  $P = (-3, -2), Q = (-4, 5)$
4.  $P = (-5, -1), Q = (4, 6)$

5. Prove that the image of a line segment  $\overline{PQ}$  under translation  $T_A$  is also a line segment. What are the end points of this image?

Let  $P, Q, M$  be the indicated points. In Exercises 6 through 9, find the point  $N$  such that  $\overrightarrow{PQ}$  has the same direction as  $\overrightarrow{MN}$  and such that the length of  $\overrightarrow{MN}$  is

- a) 3 times the length of  $\overrightarrow{PQ}$ ,
- b) one-third the length of  $\overrightarrow{PQ}$ .
6.  $P = (1, 4), Q = (1, -5), M = (-2, 3)$
7.  $P = (-1, -1), Q = (3, -2), M = (4, 4)$
8.  $P = (1, -2), Q = (5, 2), M = (-4, 3)$
9.  $P = (-1, 3), Q = (\frac{1}{2}, 4), M = (\frac{1}{3}, -1)$
10. Let  $F$  be
  - a) translation  $T_A$ ,
  - b) reflection through  $O$ ,
  - c) reflection through the  $x$ -axis,
  - d) reflection through the  $y$ -axis,
  - e) dilation by a number  $r > 0$ .

In each one of these cases, prove that the image under  $F$  of (i) a segment, (ii) a ray, is again (i) a segment, (ii) a ray, respectively. Thus you really have 10 cases to consider ( $10 = 5 \times 2$ ), but they are all easy.

11. After you have read the definition of a straight line in the next section, prove that the image under  $F$  of a straight line is again a straight line. [Here  $F$  is any one of the mappings of Exercise 10.]
12. Give a definition for two located vectors to have **opposite direction**. Similarly, if  $A \neq O$  and  $B \neq O$ , give a definition for  $A$  and  $B$  to have opposite direction. Draw the corresponding pictures.

### The 3-dimensional case

Give the definition for a segment, a ray in 3-space. Does the discussion about  $\overrightarrow{PQ}$  being equal to  $\overrightarrow{QP}$  apply? What about the definitions for having the same direction, opposite direction, etc?

13. Give the coordinates of the point

- a) one-third of the distance,
- b) two-thirds of the distance,
- c) one-half of the distance  
between the points

$$P = (3, 1, 5) \quad \text{and} \quad Q = (-1, 4, 3)$$

14. Same question for  $P = (6, -1, -2)$  and  $Q = (-4, 2, -3)$ .

Let  $P, Q, M$  be the indicated points. In Exercises 15 through 17, find the point  $N$  such that  $\overrightarrow{PQ}$  has the same direction as  $\overrightarrow{MN}$  and such that the length of  $\overrightarrow{MN}$  is

- a) 3 times the length of  $\overrightarrow{PQ}$ ,
- b) one-third the length of  $PQ$ .

15.  $P = (1, 2, 3)$ ,  $Q = (-1, 4, 5)$ ,  $M = (-1, 5, 4)$

16.  $P = (-2, 4, 1)$ ,  $Q = (-3, 5, -1)$ ,  $M = (2, 3, -1)$

17.  $P = (3, -2, -2)$ ,  $Q = (-1, -3, -4)$ ,  $M = (3, 1, 1)$

### §3. LINES

We first discuss the notion of parallelism, and give an analytic definition for it.

Let  $\overrightarrow{PQ}$  and  $\overrightarrow{MN}$  be located vectors such that  $P \neq Q$  and  $M \neq N$ . We shall say that they are **parallel** if there exists a number  $c$  such that

$$\overrightarrow{Q} - \overrightarrow{P} = c(\overrightarrow{M} - \overrightarrow{N}).$$

Observe that this time we may take  $c$  to be positive or negative. Our definition applies equally well to segments  $\overrightarrow{PQ}$  or  $\overrightarrow{MN}$  instead of located vectors, i.e. we can take  $P, Q$  in any order, and we can take  $M, N$  in any order.

As an exercise, prove:

If  $\overrightarrow{P_1Q_1}$  is parallel to  $\overrightarrow{P_2Q_2}$ , and if  $\overrightarrow{P_2Q_2}$  is parallel to  $\overrightarrow{P_3Q_3}$ , then  $\overrightarrow{P_1Q_1}$  is parallel to  $\overrightarrow{P_3Q_3}$ .

If  $\overrightarrow{PQ}$  is parallel to  $\overrightarrow{MN}$ , then  $\overrightarrow{MN}$  is parallel to  $\overrightarrow{PQ}$ .

Similarly, let  $A \neq O$  and  $B \neq O$ . We define  $A$  to be **parallel** to  $B$  if there exists a number  $c \neq 0$  such that  $A = cB$ . This amounts to saying that  $\overrightarrow{OA}$  is parallel to  $\overrightarrow{OB}$ . Having fixed our origin  $O$ , it suffices to give the end point of the segment  $\overrightarrow{OA}$  to determine parallelism with respect to  $\overrightarrow{OA}$ .

In Fig. 10-8(a) we illustrate parallel located vectors. In Fig. 10-8(b), we see that the end points of located vectors  $\overrightarrow{OA}$  and  $\overrightarrow{OB}$  lie on the same line, when  $A, B$  are parallel.

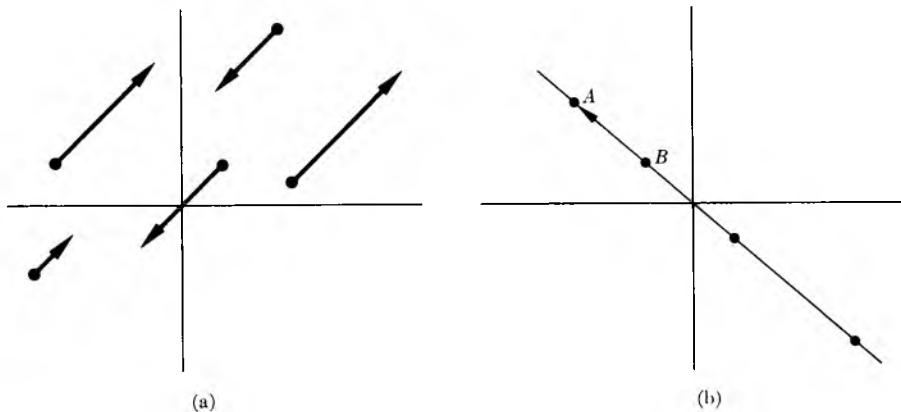


Fig. 10-8

Figure 10-8(b) suggests to us how to give an analytic definition for a line. Let  $A$  be a point  $\neq O$ . We define the **straight line** (or simply the line) parallel to  $\overrightarrow{OA}$  (or to  $A$ ) passing through the origin to be the set of all points  $tA$ , with all real numbers  $t$ .

For an arbitrary line, not necessarily passing through the origin, we just take a translation.

We define the **straight line** (or simply the line) passing through a given point  $P$ , parallel to  $\overrightarrow{OA}$  (or more simply, parallel to  $A$ ) to be the set of all points

$$P + tA,$$

with all real numbers  $t$ , positive, negative, or 0. Picture:

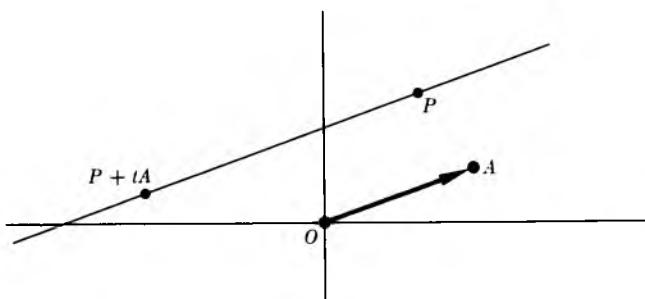


Fig. 10-9

In Fig. 10-9, we have drawn the point  $P + tA$  corresponding to a negative value of  $t$ .

We shall also use the symbols

$$\{P + tA\}_{t \text{ in } \mathbf{R}}$$

to denote this straight line.

**Example.** The line passing through the point  $(-3, 4)$ , parallel to  $(1, -5)$  is the set of all points

$$(-3, 4) + t(1, -1), \quad t \text{ in } \mathbf{R}.$$

It is easy to draw a picture of this line. We merely find two points on it and draw the line through these two points. For instance, giving  $t$  the value 0 we see that  $(-3, 4)$  is on the line. Giving  $t$  the value 1 we see that the point

$$(-3, 4) + (1, -1) = (-2, 3)$$

is on the line. This line is illustrated on Fig. 10-10.

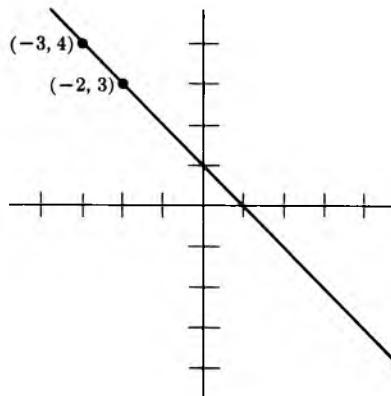


Fig. 10-10

The representation of a line in the form  $P + tA$  is called a **parametric representation**, and we call  $t$  the **parameter**. One sometimes interprets the point  $P + tA$  as describing the position of a bug, or a particle, moving along the line with uniform speed, and we interpret  $t$  as the time. Thus at time  $t = 0$ , the bug is at the point  $P$ . The coordinates  $(x, y)$  of a point on the line then depend on  $t$ , and it is customary to write them in the form

$$(x(t), y(t))$$

to indicate this dependence on  $t$ . This parametric representation of a straight line is advantageous for at least two reasons. First, it generalizes easily to 3-space. Second, it represents our physical intuition of the moving bug, and allows us to give a simple coordinate representation for the position of the bug at a given time.

**Example.** Let

$$P = (-1, 4) \quad \text{and} \quad A = (2, 3).$$

Then the coordinates for an arbitrary point on the line passing through  $P$  parallel to  $A$  are given by

$$x(t) = -1 + 2t$$

and

$$y(t) = 4 + 3t.$$

**Example.** Find a parametric representation of a line passing through two points  $P = (1, 5)$  and  $Q = (-2, 3)$ ; see Fig. 10-11.

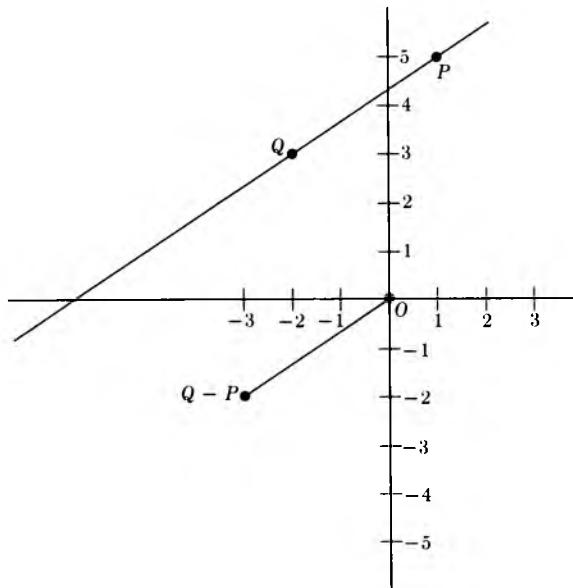


Fig. 10-11

Let  $A = Q - P$ . Then the parametric representation is

$$P + tA = P + t(Q - P).$$

In terms of the individual coordinates,  $Q - P = (-3, -2)$ , and hence

$$x(t) = 1 - 3t, \quad y(t) = 5 - 2t.$$

Observe that when  $t = 0$ , we obtain the point  $P$ , and when  $t = 1$ , we obtain the point  $Q$ . Thus we have found a parametric representation of a line passing through  $P$  and  $Q$ . We can also write this representation in the form

$$\{(1, 5) - t(3, 2)\}_{t \in \mathbb{R}}.$$

**Example.** Find the point at which the line of the preceding example crosses the  $x$ -axis.

The second coordinate of a point on this line is

$$5 - 2t.$$

Thus we must find the value for  $t$  such that  $5 - 2t = 0$ . This is easily solved, and gives  $t = \frac{5}{2}$ . Hence the point at which the line of the preceding example crosses the  $x$ -axis is

$$(1, 5) - \frac{5}{2}(3, 2) = \left(-\frac{13}{2}, 0\right).$$

**Remark.** Suppose that a bug starts from a point  $P$  and moves along a straight line with a certain speed. We can use a located vector  $\overrightarrow{OA}$ , or simply  $A$ , to represent the velocity of the bug. We interpret  $\overrightarrow{OA}$  as representing the direction in which the bug is moving, and we interpret  $|A|$  (that is, the length of  $\overrightarrow{OA}$ ) as the speed with which the bug is moving. Then at time  $t$  the position of the bug is given by

$$P + tA.$$

If another bug moves in the same direction, but with three times the speed, then at time  $t$  the position of the other bug is given by

$$P + t \cdot 3A = P + 3tA.$$

Both bugs will cover the same ground, but the second bug will do so three times as fast. Observe that the set of points

$$\{P + 3tA\}_{t \in \mathbb{R}}$$

is the same set as the set of points

$$\{P + tA\}_{t \in \mathbb{R}}.$$

In intuitive geometry, we assume that two lines which are not parallel have exactly one point in common. We now have a means of determining this point.

**Example.** Let two lines be represented parametrically by

$$(1, 2) + t(3, 4) \quad \text{and} \quad (-1, 1) + s(2, -1),$$

where  $t, s$  are the respective parameters. Find the point of intersection of these two lines.

We must find the values of  $s$  and  $t$  such that

$$\begin{aligned} 1 + 3t &= -1 + 2s, \\ 2 + 4t &= 1 - s, \end{aligned}$$

or, in other words,

$$\begin{aligned} 3t - 2s &= -2, \\ 4t + s &= -1. \end{aligned}$$

This is a system of two equations in two unknowns which we know how to solve from algebra. The solutions are

$$t = -\frac{4}{11} \quad \text{and} \quad s = \frac{5}{11}.$$

Hence the common point of the two lines is the point

$$(1, 2) - \frac{4}{11}(3, 4) = \left(\frac{-1}{11}, \frac{6}{11}\right).$$

We can also find the point of intersection of a line and other geometric figures given by equations.

**Example.** Find the points of intersection of the line given parametrically by

$$(-1, 2) + t(3, -4), \quad t \text{ in } \mathbf{R},$$

and the circle

$$x^2 + y^2 = 4.$$

The first and second coordinates of points on the line are given by

$$x(t) = -1 + 3t \quad \text{and} \quad y(t) = 2 - 4t.$$

We must find those values of  $t$  which are such that  $x(t)$  and  $y(t)$  satisfy the equation of the circle. This means that we must find those values of  $t$  such that

$$(-1 + 3t)^2 + (2 - 4t)^2 = 4.$$

Expanding out, this amounts to solving for  $t$  the equation

$$1 - 6t + 9t^2 + 4 - 16t + 16t^2 = 4,$$

or in other words

$$25t^2 - 22t + 1 = 0.$$

This is a quadratic equation, which we know how to solve. We obtain:

$$\begin{aligned} t &= \frac{22 \pm \sqrt{(22)^2 - 4 \cdot 25 \cdot 1}}{50} \\ &= \frac{14 \pm \sqrt{384}}{50}. \end{aligned}$$

Thus we obtain two values for  $t$ . We can simplify slightly, writing  $384 = 4 \cdot 96$ . The two values for  $t$  are then

$$t = \frac{7 + \sqrt{96}}{25}$$

and

$$t = \frac{7 - \sqrt{96}}{25}.$$

The points of intersection of the line and the circle are then given by  $(x_1, y_1)$  and  $(x_2, y_2)$  where  $x_1, y_1, x_2, y_2$  have the following values:

$$x_1 = -1 + 3 \cdot \frac{7 + \sqrt{96}}{25},$$

$$y_1 = 2 - 4 \cdot \frac{7 + \sqrt{96}}{25},$$

$$x_2 = -1 + 3 \cdot \frac{7 - \sqrt{96}}{25},$$

$$y_2 = 2 - 4 \cdot \frac{7 - \sqrt{96}}{25}.$$

The intersection of the line and the circle can be illustrated as follows. Giving  $t$  the special values  $t = 0$  and  $t = 1$ , we find that the two points

$$(-1, 2) \quad \text{and} \quad (2, -2)$$

lie on the line. Thus we draw the line passing through these two points, and we see it intersect the circle in Fig. 10-12.

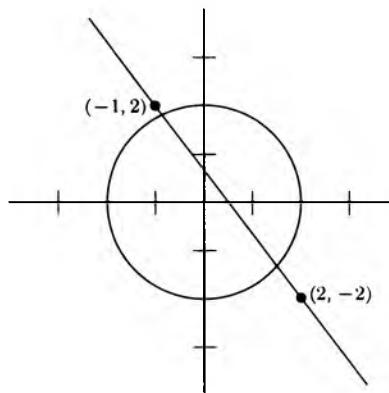


Fig. 10-12

**Example.** On the other hand, there may be cases when the line and the circle do not intersect. For instance, consider the line given by

$$(-5, 0) + t(1, 1), \quad t \text{ in } \mathbf{R}.$$

We wish to determine the points of intersection of this line, and the circle having the equation

$$x^2 + y^2 = 4.$$

Proceeding as before, we note that the coordinates of the line are given by

$$x(t) = -5 + t \quad \text{and} \quad y(t) = t.$$

Substituting these in the equation for the circle, we get an equation for  $t$ , namely

$$(-5 + t)^2 + t^2 = 4,$$

which expanded out yields

$$25 - 10t + t^2 + t^2 = 4,$$

or in other words,

$$2t^2 - 10t + 21 = 0.$$

Using the quadratic formula to solve for  $t$  gives us

$$t = \frac{10 \pm \sqrt{100 - 168}}{4} = \frac{10 \pm \sqrt{-68}}{4}.$$

We see that the expression under the square root sign is negative, and hence

there is no real value of  $t$  satisfying our equation. This means that the circle and the line do not intersect. The situation is illustrated in Fig. 10-13.

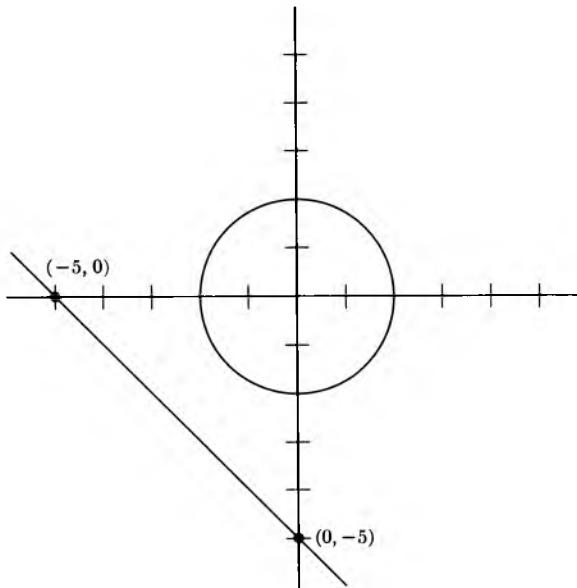


Fig. 10-13

To draw the line in this figure, all we need are two points on it. Using the special values  $t = 0$ ,  $t = 5$  shows that the points

$$(-5, 0) \quad \text{and} \quad (0, -5)$$

are on the line. These points have also been indicated on the figure.

## EXERCISES

For Exercises 1 through 6: (a) write down parametric representations of the lines passing through the indicated points  $P$  and  $Q$ , (b) find the point of intersection of the line and the  $x$ -axis, (c) find the point of intersection of the line and the  $y$ -axis.

1.  $P = (1, -1)$ ,  $Q = (3, 5)$

2.  $P = (2, 1)$ ,  $Q = (4, -1)$

3.  $P = (-4, -2), Q = (-1, 1)$       4.  $P = (3, -1), Q = (1, 1)$   
 5.  $P = (3, -5), Q = (-2, 6)$       6.  $P = (-1, 4), Q = (-1, -3)$

One airplane moves along a straight line in the plane, starting at a point  $P$  in the direction of  $A$ . Another plane also moves along a straight line, starting at a point  $Q$  in the direction of  $B$ . Find the point at which they may collide if  $P, Q, A, B$  are given by the following values. Draw the two lines.

7.  $P = (1, -1), Q = (3, 5), A = (-3, 1), B = (2, -1)$   
 8.  $P = (-4, -2), Q = (-1, 1), A = (5, 1), B = (2, -5)$   
 9.  $P = (1, -1), Q = (4, 1), A = (-1, -2), B = (-3, -1)$   
 10.  $P = (1, 1), Q = (2, -1), A = (3, 3), B = (-4, 3)$   
 11. Find the point of intersection of the lines of Exercises  
     a) 1 and 2,      b) 2 and 3,      c) 1 and 3,      d) 5 and 6.  
 12. Let  $A = (a_1, a_2)$  and  $B = (b_1, b_2)$ . Assume that  $A \neq O$  and  $B \neq O$ .  
     Prove that  $A$  is parallel to  $B$  if and only if

$$a_1 b_2 - a_2 b_1 = 0.$$

13. Prove: If two lines are not parallel, then they have exactly one point in common. [Hint: Let the two lines be represented parametrically by

$$\begin{aligned} \{P + tA\}_{t \in \mathbb{R}} &= \{(p_1, p_2) + t(a_1, a_2)\}_{t \in \mathbb{R}} \\ \{Q + sB\}_{s \in \mathbb{R}} &= \{(q_1, q_2) + s(b_1, b_2)\}_{s \in \mathbb{R}}. \end{aligned}$$

Write down the general system of two equations for  $s$  and  $t$  and show that it can be solved.]

14. Find the points of intersection of the given line  $\{P + tA\}_{t \in \mathbb{R}}$  and the circle of radius 8 centered at the origin, when  $P, A$  are as indicated. If there are no such points, say so, and why. Draw the line and circle.  
     a)  $P = (1, -1), A = (-3, 1)$       b)  $P = (2, -3), A = (1, 1)$   
     c)  $P = (-2, 1), A = (3, 5)$       d)  $P = (3, 1), A = (-2, -4)$   
     [Hint: Substitute the values for  $x(t), y(t)$  in the equation of the circle, and solve for  $t$ .]  
 15. Given (a), (b), (c), (d) as in Exercise 14. Find the points of intersection of the lines and the circle of radius 2 centered at the origin.  
 16. Given (a), (b), (c), (d) as in Exercise 14. Find the points of intersection of the lines and the circle of radius 1 centered at  $(1, 1)$ .  
 17. Given (a), (b), (c), (d) as in Exercise 14. Find the points of intersection of the lines and the circle of radius 2 centered at  $(-1, -1)$ .

18. (Slightly harder.) Let  $S$  be the circle of radius  $r > 0$  centered at the origin. Let  $P = (p, q)$  be a point such that

$$p^2 + q^2 \leq r^2.$$

In other words,  $P$  is a point in the disc of radius  $r$  centered at  $O$ . Show that any line passing through  $P$  must intersect the circle, and find the points of intersection. [Hint: Write the line in the form

$$P + tA,$$

where  $A = (a, b)$ , substitute in the equation of the circle, and find the coordinates of the points of intersection in terms of  $p, q, a, b$ . Show that the quantity you get under the square root sign is  $\geq 0$ .]

#### §4. ORDINARY EQUATION FOR A LINE

So far we have described a line in terms of a parameter  $t$ . We can eliminate this parameter  $t$  and get another type of equation for the line. We show this by an example.

**Example.** Let  $L = \{P + tA\}_{t \in \mathbb{R}}$ , where  $P = (3, 5)$  and  $A = (-2, 7)$ . Thus a point  $(x(t), y(t))$  on the line is given by

$$x = 3 - 2t, \quad y = 5 + 7t.$$

Multiply the expression for  $x$  by 7, multiply the expression for  $y$  by 2, and add. We get

$$\begin{aligned} 7x + 2y &= 3 \cdot 7 - 14t + 2 \cdot 5 + 14t \\ &= 31. \end{aligned}$$

Thus any point  $(x, y)$  on the line  $L$  satisfies the equation

$$7x + 2y = 31.$$

Conversely, let  $(x, y)$  be a point satisfying this equation. We want to solve for  $t$  such that

$$x = 3 - 2t \quad \text{and} \quad y = 5 + 7t.$$

Let

$$t = -\frac{x - 3}{2}.$$

Then certainly  $x = 3 - 2t$ . But  $y = (31 - 7x)/2$ . We must verify that this is equal to

$$5 - 7 \frac{(x - 3)}{2}.$$

This is obvious by a simple algebraic manipulation, and we have solved for the desired  $t$ .

We shall call

$$7x + 2y = 31$$

the **ordinary equation** of the line.

Similarly, any equation of the form

$$ax + by = c$$

where  $a, b, c$  are numbers, is the equation of a straight line.

## EXERCISES

Find the ordinary equation of the line  $\{P + tA\}_{t \in \mathbb{R}}$  in each one of the following cases.

- |                                  |                                   |
|----------------------------------|-----------------------------------|
| 1. $P = (3, 1)$ , $A = (7, -2)$  | 2. $P = (-2, 5)$ , $A = (5, 3)$   |
| 3. $P = (-4, -2)$ , $A = (7, 1)$ | 4. $P = (-2, -5)$ , $A = (5, 4)$  |
| 5. $P = (-1, 5)$ , $A = (2, 4)$  | 6. $P = (-3, 2)$ , $A = (-3, -2)$ |
| 7. $P = (1, 1)$ , $A = (1, 1)$   | 8. $P = (-5, -6)$ , $A = (-4, 3)$ |



# 11 Trigonometry

*This chapter can be read immediately after the definition of coordinates and distance. We don't need anything about segments, lines, etc. covered in the other chapters of this part.*

## §1. RADIAN MEASURE

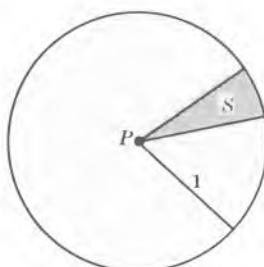
In a sense, the measurement of angles by degrees is not a natural measurement. It is much more reasonable to take another measure which we now describe.

Let  $\pi$  be the (numerical value of the) area of a disc of radius 1. The approximate value for  $\pi$  is 3.14159 . . . . (Look at our comments about  $\pi$  in Chapter 7, §1.) Let  $A$  be an angle with vertex at  $P$  and let  $S$  be the sector determined by  $A$  in the disc  $D$  of radius 1 centered at  $P$ . Let  $x$  be a number between 0 and  $2\pi$ . We shall say that

**$A$  has  $x$  radians**

to mean that

$$\frac{\text{area of } S}{\text{area of } D} = \frac{x}{2\pi}.$$



**Fig. 11-1**

Thus if  $D$  is the disc of radius 1 centered at  $P$ , our definition is adjusted so that the area of the sector  $S$  determined by an angle of  $x$  radians is  $x/2$ ; see Fig. 11-1. We draw various angles and indicate their radian measure, in Fig. 11-2.

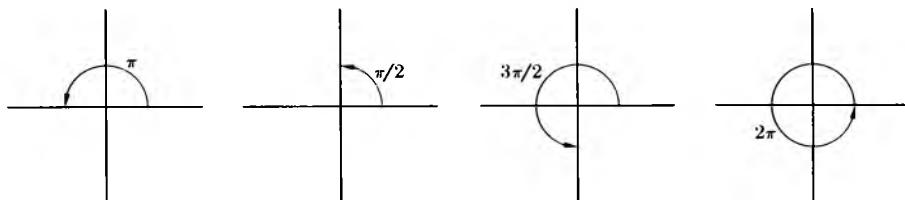


Fig. 11-2

Of course, degrees are related to radians. For instance:

$$360 \text{ degrees} = 2\pi \text{ radians}$$

$$180 \text{ degrees} = \pi \text{ radians}$$

$$60 \text{ degrees} = \pi/3 \text{ radians}$$

$$45 \text{ degrees} = \pi/4 \text{ radians}$$

$$30 \text{ degrees} = \pi/6 \text{ radians.}$$

In general,

$$x \text{ degrees} = \frac{\pi}{180} x \text{ radians.}$$

However, from now on, unless otherwise specified, we *always* deal with radian measure.

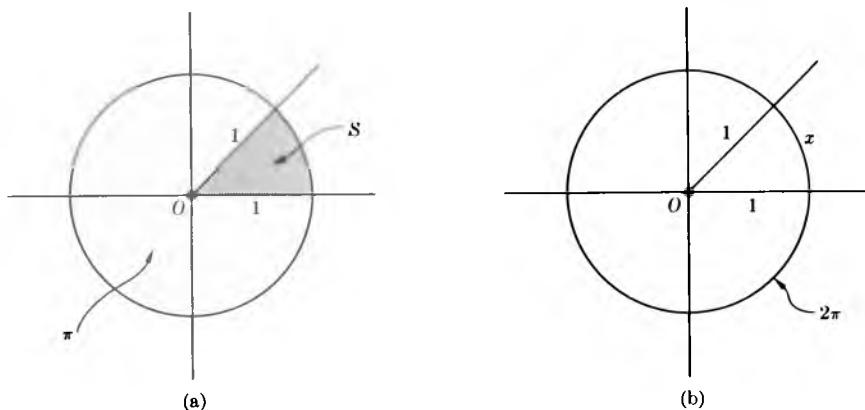


Fig. 11-3

Observe that the length of the circle of radius 1 is  $2\pi$ . (Cf. Chapter 7.) The angle  $A$  determines an arc on this circle, and radian measure is so adjusted that

$$\frac{\text{length of this arc}}{\text{total length of circle}} = \frac{x}{2\pi}.$$

Thus  $x$  is the length of arc determined by  $A$ ; this is illustrated in Fig. 11-3.

The Greeks realized that they had to choose a constant which appears very frequently in mathematics, relating the radius and circumference of a circle or its diameter and circumference. The way they chose  $\pi$ , however, is somewhat inconvenient, because it introduces a factor of 2 in front of  $\pi$  in most of mathematics. It would have been more useful to use the constant  $c$  such that

$$\frac{\text{area of } S}{\text{area of } D} = \frac{x}{c}.$$

Too late to change, however.

For convenience of language, we shall sometimes speak, incorrectly but usefully, of an angle of  $x$  radians even if  $x$  does not lie between 0 and  $2\pi$ . To do this, we write

$$x = 2n\pi + w,$$

with a number  $w$  such that

$$0 \leq w < 2\pi.$$

Then, by an angle of  $x$  radians, we mean the angle of  $w$  radians.

Also for convenience of language, if  $x$  is negative, say

$$x = -z$$

where  $z$  is positive, we shall speak of an angle of  $x$  radians to mean an angle of  $z$  radians in the clockwise direction. We draw such an angle in Fig. 11-4.

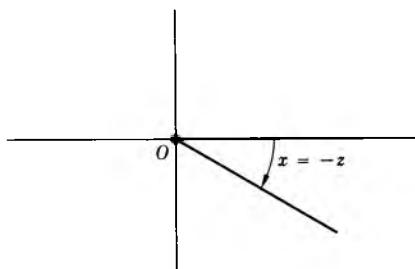


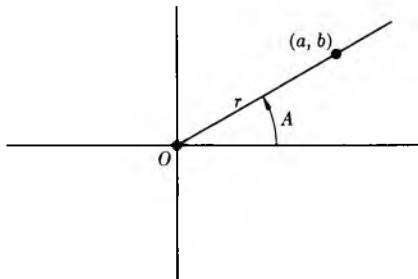
Fig. 11-4

**EXERCISES**

1. Give the following values of angle in radians, as a fractional multiple of  $\pi$ .
  - a)  $15^\circ$
  - b)  $75^\circ$
  - c)  $105^\circ$
  - d)  $120^\circ$
  - e)  $135^\circ$
  - f)  $150^\circ$
  - g)  $165^\circ$
2. Same question in the following cases.
  - a)  $20^\circ$
  - b)  $40^\circ$
  - c)  $140^\circ$
  - d)  $310^\circ$
3. Find the measure in degrees (between  $0^\circ$  and  $360^\circ$ ) for the following angles given in radians.
  - a)  $-\frac{\pi}{4}$
  - b)  $\frac{8\pi}{9}$
  - c)  $\frac{5\pi}{9}$
  - d)  $\frac{7\pi}{4}$
  - e)  $\frac{14\pi}{3}$
  - f)  $\frac{22\pi}{3}$
  - g)  $-\frac{\pi}{3}$

**§2. SINE AND COSINE**

Suppose that we have given a coordinate system and an angle  $A$  with vertex at the origin  $O$  as shown in Fig. 11-5.

**Fig. 11-5**

The positive  $x$ -axis is one side of our angle, and the other side is a ray with vertex at  $O$ . We select a point  $(a, b)$  not equal to  $O$  on this ray, and we let

$$r = \sqrt{a^2 + b^2}.$$

Then  $r$  is the distance from  $(0, 0)$  to the point  $(a, b)$ . We define

$$\text{sine } A = \frac{b}{r} = \frac{b}{\sqrt{a^2 + b^2}}$$

$$\text{cosine } A = \frac{a}{r} = \frac{a}{\sqrt{a^2 + b^2}}.$$

If we select another point  $(a_1, b_1)$  on the ray determining our angle  $A$ , and use its coordinates to get the sine and cosine, then we obtain the same values as with  $(a, b)$ . Indeed, there is a positive number  $c$  such that

$$a_1 = ca \quad \text{and} \quad b_1 = cb.$$

Hence

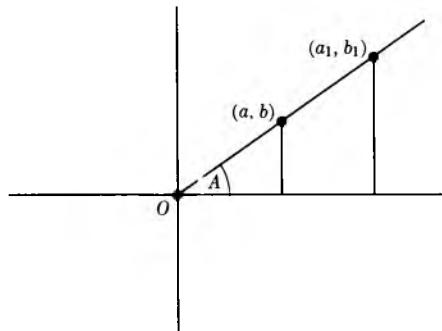
$$\frac{b_1}{\sqrt{a_1^2 + b_1^2}} = \frac{cb}{\sqrt{c^2a^2 + c^2b^2}}.$$

We can factor  $c$  from the denominator, and then cancel  $c$  in both the numerator and denominator on the right-hand side, to get

$$\frac{b}{\sqrt{a^2 + b^2}}.$$

This proves that our definition of sine  $A$  does not depend on the choice of coordinates  $(a, b)$  on the ray. The proof for the cosine is similar.

The geometric interpretation of the above argument simply states that the triangles in Fig. 11-6 are similar, i.e. are obtained by a dilation of each other.



**Fig. 11-6**

In particular, we can select the point  $(a, b)$  at any distance from the origin that we find convenient. For many purposes, it is convenient to select  $(a, b)$

on the circle of radius 1, so that  $r = 1$ . In that case,

$$\sin A = b \quad \text{and} \quad \cos A = a.$$

Consequently, by definition, the coordinates of a point on the circle of radius 1 are

$$(\cos \theta, \sin \theta)$$

if  $\theta$  is the angle in radians; see Fig. 11-7.

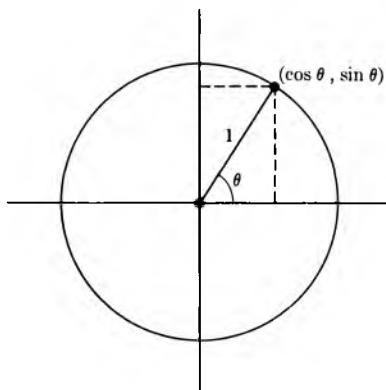


Fig. 11-7

An angle  $A$  can be determined by a ray in any one of the four quadrants. Fig. 11-8 depicts both the case when the ray is in the second quadrant and the case when it is in the third.

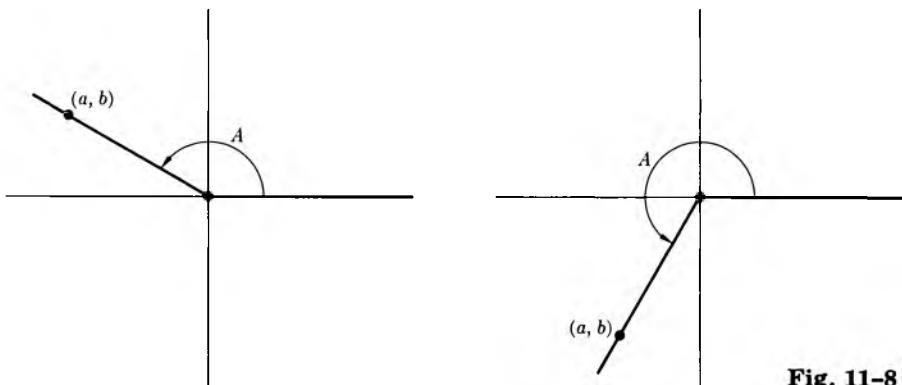


Fig. 11-8

When the ray is in the first quadrant, then both the sine and cosine are positive because both  $a$  and  $b$  are positive. When the ray is in the second quadrant, then the sine is positive because  $b$  is positive. The cosine is negative

because  $a$  is negative. When the ray is in the third quadrant, then sine  $A$  is negative and cosine  $A$  is also negative.

It is also convenient to remember the sine in the context of a right triangle. Let  $A$  be one of the angles in a right triangle, other than the right angle, as shown in Fig. 11-9. Let  $a$  be the length of the opposite side of  $A$ , let  $b$  be the length of the adjacent side, and let  $c$  be the length of the hypotenuse. Then we have

$$\sin A = \frac{a}{c} = \frac{\text{opposite side}}{\text{hypotenuse}}.$$

Similarly,

$$\cos A = \frac{b}{c} = \frac{\text{adjacent side}}{\text{hypotenuse}}.$$

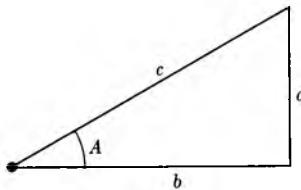


Fig. 11-9

We want to define the sine and cosine of a number. Let  $x$  be a number. We write

$$x = 2n\pi + w$$

with some integer  $n$  and some number  $w$  such that

$$0 \leq w < 2\pi.$$

We then *define*:

$$\sin x = \text{sine of } w \text{ radians}$$

$$\cos x = \text{cosine of } w \text{ radians.}$$

Thus we have for all numbers  $x$ :

$$\sin(x + 2\pi) = \sin x$$

$$\cos(x + 2\pi) = \cos x.$$

Instead of adding  $2\pi$ , we could also add any integral multiple of  $2\pi$ , and find the similar property, namely

$$\sin(x + 2n\pi) = \sin x$$

$$\cos(x + 2n\pi) = \cos x.$$

We can compute a few values of the sine and cosine, as shown in the following table.

Number	Sine	Cosine
$\pi/6$	$\frac{1}{2}$	$\sqrt{3}/2$
$\pi/4$	$1/\sqrt{2}$	$1/\sqrt{2}$
$\pi/3$	$\sqrt{3}/2$	$\frac{1}{2}$
$\pi/2$	1	0
$\pi$	0	-1
$2\pi$	0	1

These can be determined by plane geometry and the Pythagoras theorem. For instance, we get the sine of the angle  $\pi/4$  radians from a right triangle with two equal legs:

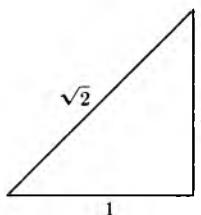


Fig. 11-10

We see that the sine is equal to  $1/\sqrt{2}$ .

Similarly, consider a right triangle whose angles other than the right angle have  $\pi/6$  and  $\pi/3$  radians (in other words  $30^\circ$  and  $60^\circ$ , respectively).

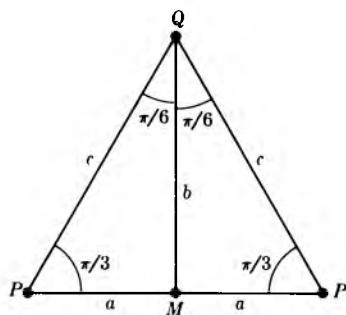


Fig. 11-11

In Fig. 11-11, we reflect our triangle  $\triangle PQM$  through the side  $\overline{QM}$  and let  $P'$  be the reflection of  $P$  through  $M$ . Then  $\triangle P'QP$  is a triangle all of whose angles have the same measure, so that the three sides have equal length. Hence  $c = 2a$  and

$$a^2 + b^2 = c^2 = 4a^2.$$

Therefore

$$b^2 = 3a^2$$

and

$$b = \sqrt{3}a.$$

Hence

$$\sin \frac{\pi}{6} = \frac{1}{2} \quad \text{and} \quad \cos \frac{\pi}{6} = \frac{\sqrt{3}}{2}.$$

Of course we could have taken our triangle to be normalized so that  $b = 1$ . In this case, we have a triangle with sides of lengths 1, 2,  $\sqrt{3}$ .

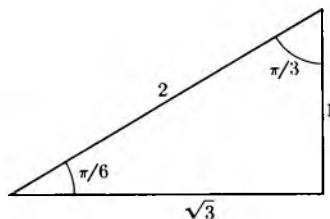


Fig. 11-12

The sine and cosine satisfy a basic relation, namely for all numbers  $x$ , we have

$$\sin^2 x + \cos^2 x = 1.$$

[Notation:  $\sin^2 x$  means  $(\sin x)^2$ , and similarly for the cosine.] This is immediate because

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

Figure 11-13 illustrates this argument.

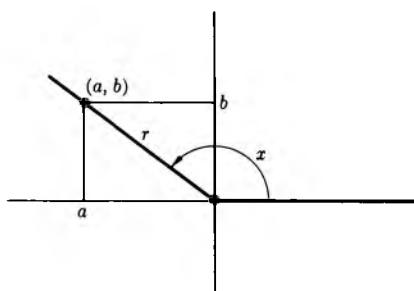


Fig. 11-13

**Theorem 1.** For any number  $x$ , we have

$$\cos x = \sin\left(x + \frac{\pi}{2}\right) \quad \text{and} \quad \sin x = \cos\left(x - \frac{\pi}{2}\right).$$

*Proof.* We may assume that  $0 \leq x < 2\pi$ . Let  $A$  be an angle of  $x$  radians, let  $P = (a, b)$  be a point on the ray which forms one side of  $A$  as shown in Fig. 11-14 and such that  $P$  lies on the circle of radius 1.

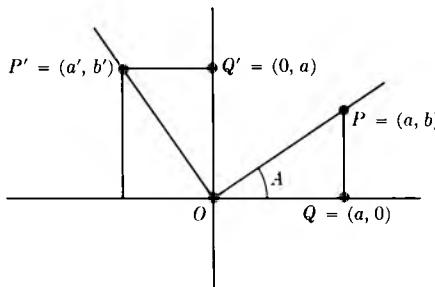


Fig. 11-14

Let  $Q = (a, 0)$ . Let  $G$  be rotation by an angle of  $\pi/2$ . Let  $P' = G(P)$  and  $Q' = G(Q)$ . Then  $Q' = (0, a)$ . If  $P = Q$ , then the first formula is clear. If  $P \neq Q$ , then the segment  $\overline{PQ}$  is perpendicular to the first axis, and hence  $\overline{P'Q'}$  is perpendicular to the second axis. Hence the second coordinate of  $Q'$  is  $a$ , in other words, we have  $b' = a$ . Since we took  $P$  on the circle of radius 1, it follows that

$$\cos x = a \quad \text{and} \quad \sin\left(x + \frac{\pi}{2}\right) = b'.$$

This proves the first formula. The second is proved similarly, and we leave it to you.

**Theorem 2.** For all numbers  $x$ , we have

$$\sin(-x) = -\sin x \quad \text{and} \quad \cos(-x) = \cos x.$$

*Proof.* This comes from looking at Fig. 11-15 and using the definition of the sine and cosine. If  $(a, b)$  is a point on the ray corresponding to an angle of  $x$  radians, then  $(a, -b)$  is a point on the ray corresponding to an angle of  $-x$  radians.

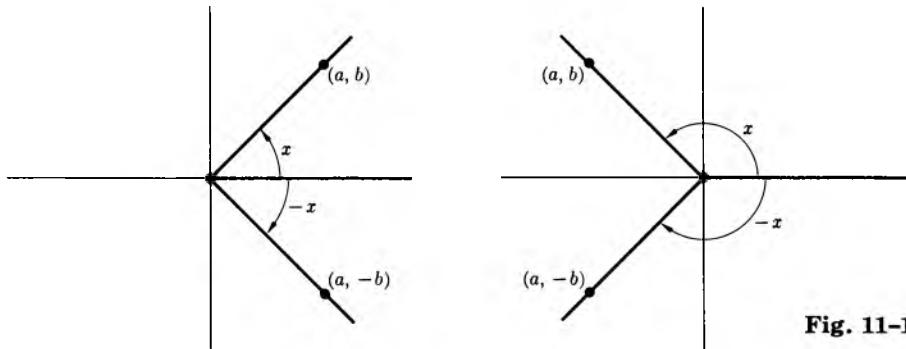


Fig. 11-15

We can take  $(a, b)$  on the circle of radius 1, in which case

$$\sin x = b \quad \text{and} \quad \cos x = a.$$

Our assertion is then clear.

The sine and cosine have various applications. We give one example.

**Example.** A boat  $B$  starts from a point  $P$  on a straight river and moves down the river. An observer  $O$  stands at a distance of 1,000 ft from  $P$ , on the line perpendicular to the river passing through  $P$ . After 10 min, the observer finds that the angle  $\theta$  formed by  $P$ , himself, and the boat, is such that  $\cos \theta = 0.7$ . What is the distance between the observer and the boat at that time?

Picture first:

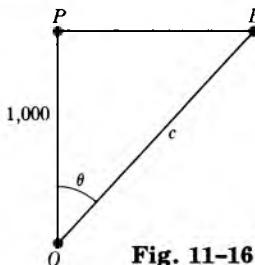


Fig. 11-16

We want to find  $c$ . We have

$$\frac{1,000}{c} = \cos \theta.$$

Hence

$$\frac{1,000}{c} = 0.7$$

and

$$c = \frac{1,000}{0.7} = \frac{10,000}{7}.$$

We don't care here if you compute the decimal or not.

**Remark.** The whole point in this example is that the angle  $\theta$  can be determined with a very small-sized instrument. The distance  $d(O, P)$  of course must be measured, but both  $O, P$  are fixed so there is no difficulty in that.

### Polar coordinates

Referring to Fig. 11-17, let  $(x, y)$  be a point in the plane. We can describe this point by using other coordinates. Let

$$r = \sqrt{x^2 + y^2}.$$

Thus  $(x, y)$  lies at distance  $r$  from the origin. Then by definition we have

$$\frac{x}{r} = \cos \theta \quad \text{and} \quad \frac{y}{r} = \sin \theta$$

for some number  $\theta$ , provided that  $r \neq 0$ . We can rewrite these in the form

$$x = r \cos \theta \quad \text{and} \quad y = r \sin \theta,$$

and since we don't divide by  $r$ , these are valid whether  $r = 0$  or not.

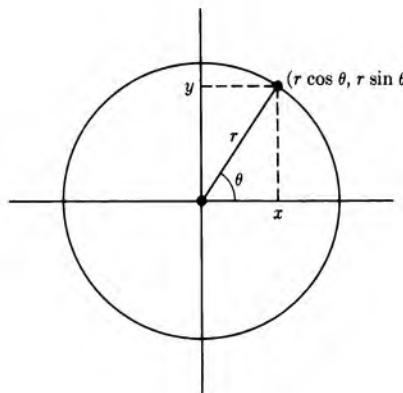


Fig. 11-17

We call  $(r, \theta)$  **polar coordinates** for the point  $(x, y)$ . When we deal simultaneously with polar coordinates and the other coordinates  $(x, y)$ , then we call  $(x, y)$  the **Cartesian, or rectangular coordinates**.

**Example.** Find polar coordinates for the point whose rectangular coordinates are  $(1, \sqrt{3})$ .

We have  $x = 1$  and  $y = \sqrt{3}$ , so that

$$r = \sqrt{1 + 3} = 2.$$

Also

$$\cos \theta = \frac{1}{2}$$

and

$$\sin \theta = \frac{\sqrt{3}}{2}.$$

We see that  $\theta = \pi/3$ , and that the polar coordinates for the point are  $(2, \pi/3)$ . According to our convention, we note that if  $(r, \theta)$  are polar coordinates for a point, then  $(r, \theta + 2\pi)$  are also polar coordinates. Thus in our example, our given point also has polar coordinates given by

$$\left(2, \frac{\pi}{3} + 2\pi\right).$$

In practice, we usually select the value of  $\theta$  such that  $0 \leq \theta < 2\pi$ .

**Example.** Given numbers  $a, b$  such that  $a^2 + b^2 = 1$ , we can always find a number  $\theta$  such that  $a = \cos \theta$  and  $b = \sin \theta$ . Therefore the point whose rectangular coordinates are  $(a, b)$  has polar coordinates  $(1, \theta)$  for such  $\theta$ . If  $a$  is positive, then  $-\pi/2 < \theta < \pi/2$ , and our choice of  $\theta$  is restricted to two possibilities as shown in Fig. 11-18.

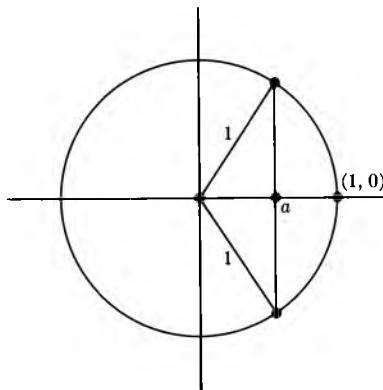


Fig. 11-18

The number  $a$  must lie between 0 and 1 because  $a^2 \leq 1$ . The angle of  $\theta$  radians is determined by the point of intersection of the circle of radius 1 centered at the origin, and the line through the point  $(a, 0)$ , perpendicular to the  $x$ -axis. Now depending on whether  $b$  is positive or negative, we eliminate one of the two possibilities for  $\theta$ ; namely, we take  $\theta$  positive if  $b$  is positive, and  $\theta$  negative if  $b$  is negative, such that  $\sin \theta = b$ .

When  $a$  is negative, we argue in a similar way.

**Example.** Let  $a = \frac{1}{2}$  and  $b = -\sqrt{3}/2$ . Then  $a^2 + b^2 = 1$ . In this case, illustrated in Fig. 11-19, we take  $\theta = -\pi/3$ .

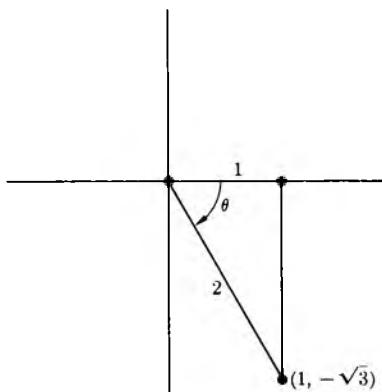


Fig. 11-19

## EXERCISES

1. Make a table of values of  $\sin x$  and  $\cos x$  when  $x$  is equal to:

- a)  $n\pi/6$ , and  $n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ ;
- b)  $n\pi/4$ , and  $n = 1, 2, 3, 4, 5, 6, 7, 8$ .

In each case, draw the corresponding angle. In your table, also make a column giving the measure of the angle in degrees.

2. Make a table as you did in Exercise 1 when  $n$  ranges over the negative values of the numbers given.

3. When the angle  $A$  has its defining ray in the fourth quadrant, determine whether the sine is positive or negative. Repeat for the cosine.
4. A boat  $B$  starts from a point  $P$  and moves along a straight river. An observer  $O$  stands at a distance of 600 ft from  $P$ , on the line perpendicular to the river passing through  $P$ . Find the distance from the boat to the observer when the angle  $\theta$  formed by  $B$ ,  $O$ , and  $P$  has
- $\pi/6$  radians,
  - $\pi/4$  radians,
  - $\pi/3$  radians.
5. A balloon  $B$  starts from a point  $P$  on earth and goes straight up. A man  $M$  is at a distance of  $\frac{1}{2}$  mi from  $P$ . After 2 min, the angle  $\theta$  formed by  $P$ ,  $M$ ,  $B$  has a cosine equal to
- 0.3,
  - 0.4,
  - 0.2.
- Find the distance between the man and the balloon at that time.
6. Repeat Exercise 5 if, after 10 min,  $\theta$  itself has
- $\pi/3$ ,
  - $\pi/4$ ,
  - $\pi/6$  radians.
7. Plot the following points with polar coordinates  $(r, \theta)$ .
- $(2, \pi/4)$
  - $(3, \pi/6)$
  - $(1, -\pi/4)$
  - $(2, -5\pi/6)$
8. Find polar coordinates for the following points given in the usual rectangular coordinates.
- $(1, 1)$
  - $(-1, -1)$
  - $(3, 3\sqrt{3})$
  - $(-1, 0)$
9. Let  $a$ ,  $b$  be the lengths of the legs of a right triangle. Let  $\theta$  be the angle between these legs, as shown in Fig. 11-20.

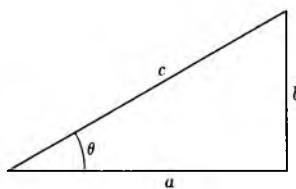


Fig. 11-20

Find the length of the hypotenuse  $c$  under the following conditions.

- $a = 4$  and  $\cos \theta = 0.3$
  - $a = 5$  and  $\cos \theta = 0.5$
  - $a = 2$  and  $\cos \theta = 0.8$
  - $b = 1$  and  $\sin \theta = 0.2$
  - $b = 4$  and  $\sin \theta = 0.3$
  - $b = 3$  and  $\sin \theta = 0.2$
10. Find the value of  $b$  in the preceding exercise under the following conditions.
- $c = 4$  and  $\sin \theta = 0.2$
  - $c = 6$  and  $\sin \theta = 0.3$
  - $c = 5$  and  $\sin \theta = 0.25$
  - $c = 3$  and  $\sin \theta = 0.6$

11. Find the value of  $\alpha$  in the preceding exercise under the following conditions.

- a)  $c = 3$  and  $\cos \theta = 0.1$
- b)  $c = 6$  and  $\cos \theta = 0.2$
- c)  $c = 4$  and  $\cos \theta = 0.6$
- d)  $c = 5$  and  $\cos \theta = 0.25$

12. Find the number  $\theta$  such that  $0 \leq \theta \leq \pi/2$  and satisfying the following conditions:

- a)  $\sin \theta = \frac{1}{2}$ ,
- b)  $\cos \theta = \frac{1}{\sqrt{2}}$ ,
- c)  $\sin \theta = \frac{\sqrt{3}}{2}$ ,
- d)  $\sin \theta = \frac{1}{\sqrt{2}}$ ,
- e)  $\cos \theta = \frac{1}{2}$ .

13. Find a number  $\theta$  satisfying the following conditions. In case of the sine, your number should satisfy  $-\pi/2 \leq \theta \leq \pi/2$ . In case of the cosine, it should satisfy  $0 \leq \theta \leq \pi/2$ .

- a)  $\sin \theta = -\frac{1}{2}$ ,
- b)  $\cos \theta = -\frac{1}{2}$ ,
- c)  $\sin \theta = -\frac{\sqrt{3}}{2}$ ,
- d)  $\cos \theta = -\frac{\sqrt{3}}{2}$ ,
- e)  $\sin \theta = \frac{-1}{\sqrt{2}}$ ,
- f)  $\cos \theta = \frac{-1}{\sqrt{2}}$ ,
- g)  $\sin \theta = \frac{\sqrt{3}}{2}$ ,
- h)  $\cos \theta = \frac{\sqrt{3}}{2}$ ,
- i)  $\sin \theta = \frac{1}{2}$ ,
- j)  $\cos \theta = \frac{1}{2}$ .

### §3. THE GRAPHS

We consider the values of  $\sin x$  when  $x$  goes from 0 to  $2\pi$ . We take a circle of radius 1 centered at the origin, and determine  $\sin x$  from a point  $(a, b)$  on this circle, as in Fig. 11-21.

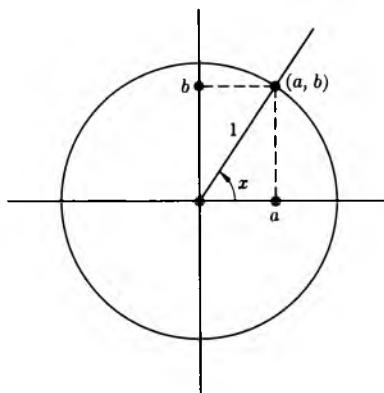


Fig. 11-21

Then

$$\sin x = b.$$

We start with  $\sin 0 = 0$ . As  $x$  goes from 0 to  $\pi/2$ , the sine of  $x$  increases (namely  $b$  increases) until  $x$  reaches  $\pi/2$ , at which point the sine is equal to 1.

As  $x$  ranges from  $\pi/2$  to  $\pi$ , the sine decreases until it becomes  $\sin \pi = 0$ .

As  $x$  ranges from  $\pi$  to  $3\pi/2$ , the sine becomes negative, but otherwise behaves in a way similar to the sine in the first quadrant. The sine in this range decreases until it reaches

$$\sin 3\pi/2 = -1.$$

Finally, as  $x$  goes from  $3\pi/2$  to  $2\pi$ , the sine of  $x$  goes from  $-1$  to 0 and increases.

At this point we are ready to start all over again.

It is interesting to plot the points whose coordinates are

$$(x, \sin x).$$

The set of points in the plane having these coordinates is called the **graph** of the sine. According to the preceding remarks, we see that the graph of the sine looks approximately as in Fig. 11-22.

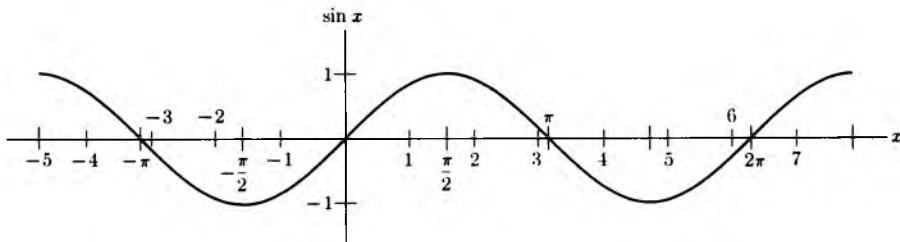


Fig. 11-22

**EXERCISES**

1. Make a similar study of the cosine, and draw its graph.
2. Sketch the graph of  $\sin 2x$ , i.e. all points whose coordinates are  $(x, \sin 2x)$ .
3. Sketch the graphs, i.e. all pairs  $(x, y)$  when  $y$  is given by:
 

a) $y = \sin(-x)$ ,	b) $y = \sin 3x$ ,	c) $y = \cos 2x$ ,
d) $y = \cos 3x$ ,	e) $y = \sin 4x$ ,	f) $y = \cos 4x$ ,
g) $y = \cos(-x)$ ,	h) $y = \sin 5x$ ,	i) $y = \cos 5x$ .

**§4. THE TANGENT**

We define the **tangent** of an angle  $A$  to be

$$\tan A = \frac{\sin A}{\cos A}.$$

This is defined only when  $\cos A \neq 0$ , and therefore is defined for angles other than  $\pi/2$  radians or  $3\pi/2$  radians.

Similarly, we define

$$\tan x = \frac{\sin x}{\cos x},$$

whenever  $x$  is a number which is not of the form

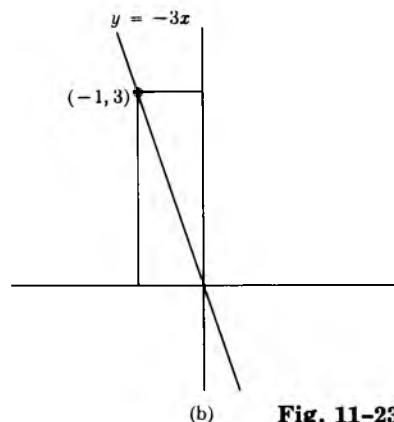
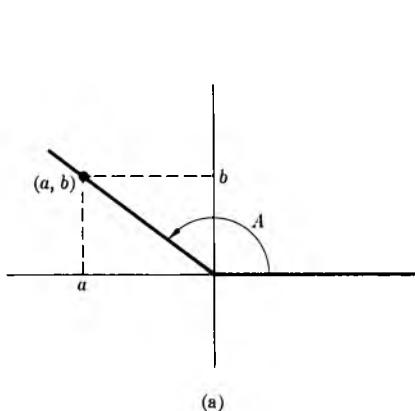
$$\frac{\pi}{2} + n\pi,$$

and  $n$  is an integer. We let you make a table of values for  $\tan x$  when  $x$  ranges over the usual simple numbers, multiples of  $\pi/4$  or multiples of  $\pi/6$ .

Suppose that the angle  $A$  is defined by the positive  $x$ -axis and a ray as before. We select a point  $(a, b)$  on the ray as before. Then

$$\tan A = \frac{b}{a}.$$

This is seen at once by taking the quotient of  $b/r$  and  $a/r$ : the  $r$  cancels.

**Fig. 11-23**

**Remark.** If you read about the slope of a straight line in the next chapter, you will observe that the tangent of the angle which the line makes with the  $x$ -axis is precisely its slope. For instance, if the line is given by the equation

$$y = -3x,$$

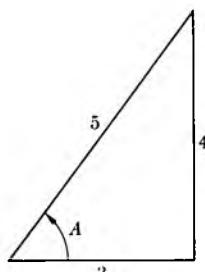
then the tangent of this angle is  $-3$ , as shown in Fig. 11-23(b). We see this by picking a point on the line, say  $(1, -3)$  and then using the definition of the tangent.

**Example.** If  $A$  is the angle as indicated in Fig. 11-24(a), in a  $3, 4, 5$  right triangle, then

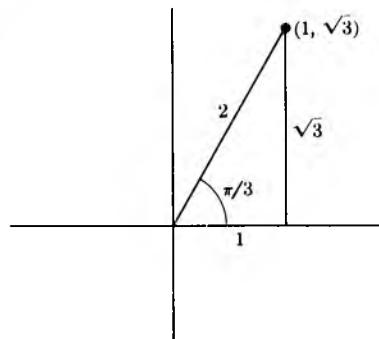
$$\tan A = \frac{4}{3}.$$

We also have

$$\tan \frac{\pi}{3} = \sqrt{3}.$$



(a)



(b)

**Fig. 11-24**

**Example.** Determine all possible values of  $\cos x$  given that  $\tan x = 2$ .

To do this, first note that  $\tan x$  is negative when  $x$  lies between  $\pi/2$  and  $\pi$ , and also when  $x$  lies between  $3\pi/2$  and  $2\pi$ . On the other hand, the tangent is positive for other values of  $x$ , and there will be two possible values of  $x$  such that  $\tan x = 2$ , as shown in Fig. 11-25(a) and (b).

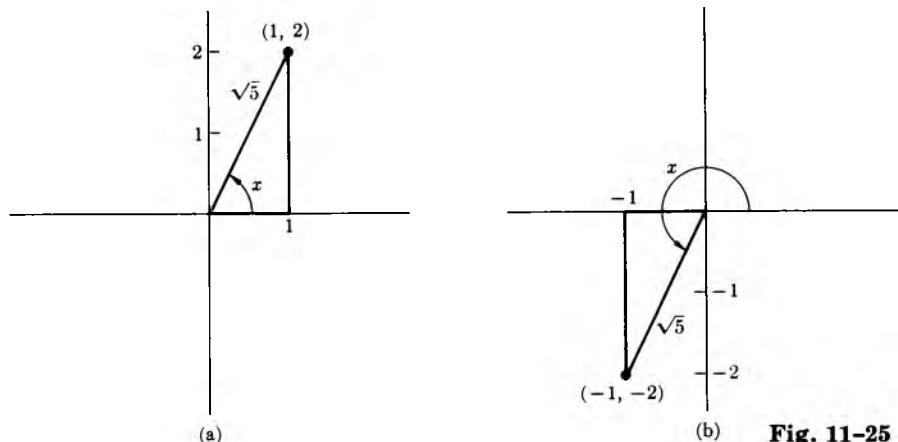


Fig. 11-25

These values correspond to right triangles whose legs have lengths 1 and 2, respectively. Therefore the hypotenuse has length  $\sqrt{5}$ . In the case of Fig. 11-25(a), it follows that

$$\cos x = \frac{1}{\sqrt{5}},$$

and in the case of Fig. 11-25(b), it follows that

$$\cos x = -\frac{1}{\sqrt{5}}.$$

These are the desired values.

We can draw the graph of the tangent just as easily as the graph of sine or cosine. We use a table of values for a few points, and also take into account the manner in which the tangent increases or decreases. For instance, when  $x$  goes from 0 to  $\pi/2$ , we note that  $\sin x$  increases from 0 to 1, while  $\cos x$  decreases from 1 to 0. Hence  $1/\cos x$  increases, starting with the value 1 when  $x = 0$  and becoming arbitrarily large. Thus finally

$$\tan x = \sin x \cdot \frac{1}{\cos x}$$

increases, starting with the value 0 when  $x = 0$ , and becomes arbitrarily large. A similar discussion for other intervals shows us that the graph of the tangent looks like this.

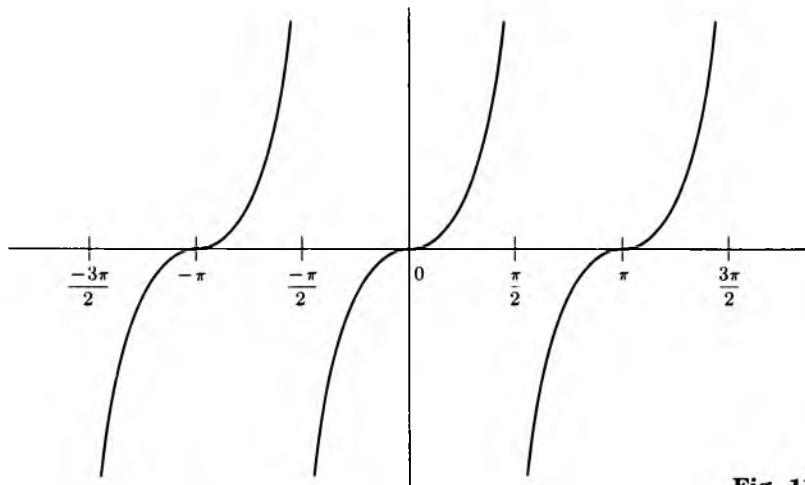


Fig. 11-26

The tangent is more practical than the sine or cosine for many purposes.

**Example.** Suppose we want to determine the height of a tower without climbing the tower. We go a distance  $a$  from the tower, as shown on Fig. 11-27.

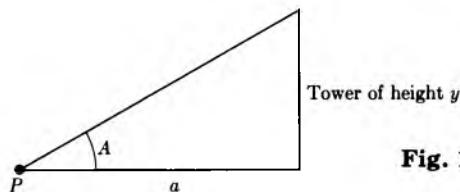


Fig. 11-27

The distance  $a$  is known, and we wish to determine the height  $y$ . We can determine the angle  $A$  easily with any mechanical device available for that purpose. We can then look up the tangent,  $\tan A$ , in tables. Since

$$\frac{y}{a} = \tan A,$$

we can then solve for  $y$ , namely  $y = a \cdot \tan A$ .

For instance, suppose that the distance  $a$  is equal to 100 ft and that the angle  $A$  has  $\pi/3$  radians. Then

$$\tan \frac{\pi}{3} = \sqrt{3},$$

and hence the height of the tower is equal to

$$100\sqrt{3} \text{ ft.}$$

## EXERCISES

1. Make a table for the values of the tangent at all points

$$\frac{n\pi}{4} \quad \text{and} \quad \frac{n\pi}{6},$$

with integers  $n$  having the same values as in Exercise 1 of §1, except for those  $n$  where the cosine is zero. In your table, also make a column giving the values of the angle in degrees.

2. Discuss how the tangent is increasing or decreasing for  $-\pi/2 < x \leq 0$ . Also for  $\pi/2 < x < 3\pi/2$  and  $-3\pi/2 < x < -\pi/2$ .
3. Define the cotangent  $\cot x = 1/\tan x$ . Draw an approximate graph for the cotangent, i.e. for the points  $(x, \cot x)$ .
4. Define the secant and cosecant by

$$\sec x = \frac{1}{\cos x} \quad \text{and} \quad \operatorname{cosec} x = \frac{1}{\sin x}$$

for values of  $x$  where  $\cos x \neq 0$  and  $\sin x \neq 0$ , respectively. Find enough values of the secant and cosecant until you feel that you have the hang of things. Draw their graphs.

5. Prove that

$$1 + \tan^2 x = \sec^2 x.$$

6. State and prove a similar formula relating the cotangent and the cosecant.
7. Determine all possible values of  $\cos x$  if  $\tan x$  has the following values.
- a)  $\tan x = 1$ ,      b)  $\tan x = -1$ ,      c)  $\tan x = \sqrt{3}$ ,  
 d)  $\tan x = 1/\sqrt{3}$ ,      e)  $\tan x = 0$ .

8. Determine all possible values of  $\sin x$  in each one of the cases of Exercise 7.
9. You are looking at a tall building from a distance of 500 ft. The angle formed by the base of the building, your eyes, and the top of the building has
- a)  $\pi/4$  radians,      b)  $\pi/3$  radians,      c)  $\pi/6$  radians.  
 Find the height of the building.
10. A balloon  $B$  starts from a point  $P$  on earth and goes straight up. An observer  $O$  stands at a distance of 500 ft from  $P$ . After 20 min, the angle  $\theta$  formed by  $P, O, B$  has
- a)  $\pi/3$  radians,      b)  $\pi/4$  radians,      c)  $\pi/6$  radians.  
 Find the height of the balloon at that time.
11. A boat  $B$  starting from a point  $P$  moves along a straight river. A man  $M$  stands  $\frac{1}{2}$  mi from  $P$  on the line through  $P$  perpendicular to the river.

After 5 hr, the boat has traveled 10 mi. Let  $\theta$  be the angle  $\angle BMP$ . Find  
 a)  $\cos \theta$ ,                          b)  $\sin \theta$ ,                          c)  $\tan \theta$   
 at that time.

A billiard ball table is rectangular, and its sides have 10 ft and 7 ft, respectively. A billiard ball is hit starting at a point  $P$  on one side as drawn on the picture. It hits the next side at  $Q$ , bounces off, hits the third side at  $M$ , bounces off, and hits the fourth side at  $N$ . Each time it bounces off a side, the angle of approach to this side has the same measure as the angle of departure. Let  $\theta$  be the first angle of departure as drawn in Fig. 11-28.

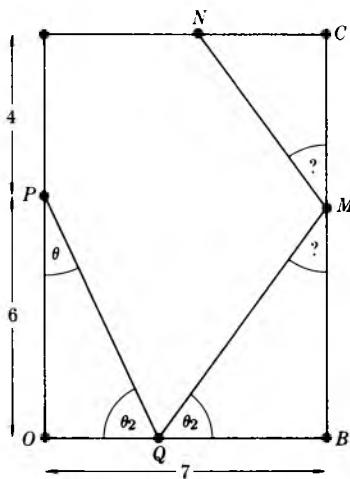


Fig. 11-28

Assume that  $P$  is at a distance of 6 ft from the corner  $O$ .

12. Assume that  $\theta$  has  $30^\circ$ . Find:

  - a)  $d(O, Q)$ ,
  - b)  $d(P, Q)$ ,
  - c)  $d(Q, B)$ ,
  - d)  $d(B, M)$ ,
  - e)  $d(Q, M)$ ,
  - f)  $d(M, C)$ ,
  - g)  $d(M, N)$ .

13. Repeat the problem for (a) through (g) of Exercise 12 if  $\tan \theta = \frac{1}{2}$ .

14. Find the general formula for

  - a)  $d(O, Q)$ ,
  - b)  $d(P, Q)$ ,
  - c)  $d(Q, M)$ ,
  - d)  $d(Q, B)$ ,
  - e)  $d(B, M)$

in terms of  $\tan \theta$ .

15. Find the general formula for the distances of Exercise 12 (a) through (e) in terms of  $\sin \theta$ .

### §5. ADDITION FORMULAS

Our main results are the addition formulas for sine and cosine.

**Theorem 3.** *For any angles  $A, B$ , we have*

$$\begin{aligned}\sin(A + B) &= \sin A \cos B + \cos A \sin B \\ \cos(A + B) &= \cos A \cos B - \sin A \sin B.\end{aligned}$$

*Proof.* We shall prove the second formula first.

We consider two angles  $A, B$  and their sum:

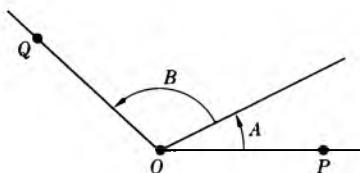


Fig. 11-29

We take two points  $P, Q$  as indicated in Fig. 11-29, at a distance 1 from the origin  $O$ . We shall compute the distance from  $P$  to  $Q$ , using two different coordinate systems. First we take a coordinate system as usual, illustrated in Fig. 11-30.

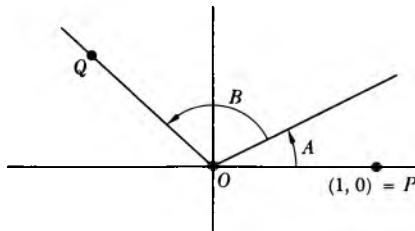


Fig. 11-30

Then the coordinates of  $P$  are  $(1, 0)$  and those of  $Q$  are

$$(\cos(A + B), \sin(A + B)).$$

The square of the distance between  $P$  and  $Q$  is

$$\sin^2(A + B) + (\cos(A + B) - 1)^2,$$

which is equal to  $\sin^2(A + B) + \cos^2(A + B) - 2 \cos(A + B) + 1$ , and hence equal to

$$-2 \cos(A + B) + 2.$$

Next we place the coordinate system as shown in Fig. 11-31.

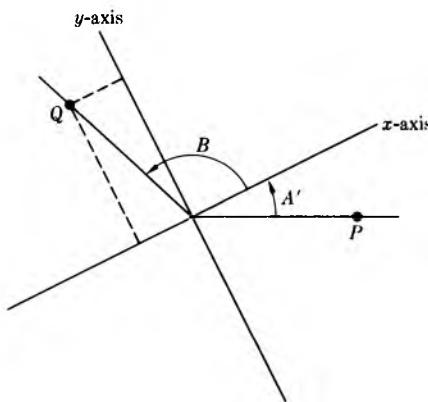


Fig. 11-31

Then the coordinates of  $P$  become

$$(\cos A, \sin(-A)) = (\cos A, -\sin A).$$

Those of  $Q$  are simply  $(\cos B, \sin B)$ . The square of the distance between  $P$  and  $Q$  is equal to

$$(\sin B + \sin A)^2 + (\cos B - \cos A)^2,$$

which is equal to

$$\begin{aligned} \sin^2 B + 2 \sin B \sin A + \sin^2 A + \cos^2 B - 2 \cos B \cos A + \cos^2 A \\ = 2 + 2 \sin A \sin B - 2 \cos A \cos B. \end{aligned}$$

If we set the squares of the two distances equal to each other, then we get the desired addition formula for the cosine.

The addition formula for the sine can be obtained by the following device, using Theorem 2:

$$\begin{aligned} \sin(A + B) &= \cos\left(A + B - \frac{\pi}{2}\right) \\ &= \cos A \cos\left(B - \frac{\pi}{2}\right) - \sin A \sin\left(B - \frac{\pi}{2}\right) \\ &= \cos A \sin B + \sin A \sin\left(\frac{\pi}{2} - B\right) \\ &= \cos A \sin B + \sin A \cos B. \end{aligned}$$

This proves the addition formula for the sine.

**Corollary.** For any numbers  $x, y$ , we have:

$$\begin{aligned}\sin(x - y) &= \sin x \cos y - \cos x \sin y \\ \cos(x - y) &= \cos x \cos y + \sin x \sin y.\end{aligned}$$

*Proof.* This follows from the theorem by using Theorem 2, §2. Write out the details in full.

**Remark.** Among all the formulas for sine and cosine, you should remember those of Theorem 3, and the following ones:

**SC 1.**  $\sin^2 x + \cos^2 x = 1$ ,

**SC 2.**  $\cos(-x) = \cos x$ ,

**SC 3.**  $\sin(-x) = -\sin x$ ,

**SC 4.**  $\sin\left(x + \frac{\pi}{2}\right) = \cos x$ ,

**SC 5.**  $\cos\left(x - \frac{\pi}{2}\right) = \sin x$ .

Read these out loud and get an aural memory of them. All other formulas are immediate consequences of these, and can be derived each time you need them. Experience shows that this is the most economical use of your brain in dealing with these formulas.

**Example.** Find  $\sin(\pi/12)$ .

We write

$$\frac{\pi}{12} = \frac{\pi}{3} - \frac{\pi}{4}.$$

Then

$$\sin\left(\frac{\pi}{12}\right) = \sin\left(\frac{\pi}{3}\right)\cos\left(\frac{\pi}{4}\right) - \cos\left(\frac{\pi}{3}\right)\sin\left(\frac{\pi}{4}\right).$$

We know all the values on the right-hand side, and a simple computation shows that

$$\sin\left(\frac{\pi}{12}\right) = \frac{\sqrt{3} - 1}{2\sqrt{2}}.$$

From Theorem 3 we can also easily prove four other formulas which are used quite frequently and which should be memorized. They are:

**SC 6.**  $\sin 2x = 2 \sin x \cos x$ ,

**SC 7.**  $\cos 2x = \cos^2 x - \sin^2 x$ ,

$$\text{SC 8. } \cos^2 x = \frac{1 + \cos 2x}{2},$$

$$\text{SC 9. } \sin^2 x = \frac{1 - \cos 2x}{2}.$$

You should have fun carrying out the easy proofs, and we leave them to you as Exercise 4. We shall now see how to use these formulas in examples to compute new values for the sine and cosine.

**Example.** Suppose that  $x$  is a number such that

$$\sin x = 0.8 \quad \text{and} \quad 0 < x < \frac{\pi}{2}.$$

To find  $\sin 2x$  we use formula SC 6. Note that

$$\cos x = \sqrt{1 - \sin^2 x} = \sqrt{1 - 0.64} = \sqrt{0.36} = 0.6.$$

We took the positive square root because we prescribed that  $x$  should lie between 0 and  $\pi/2$ , so that  $\cos x$  must be positive. Applying SC 6 now yields

$$\sin 2x = 2 \sin x \cos x = 2(0.8)(0.6) = 0.96.$$

**Example.** We wish to compute  $\cos \pi/8$ . Let us use SC 8, with  $x = \pi/8$ . We know  $\cos 2x$ , namely

$$\cos \frac{\pi}{4} = \frac{1}{\sqrt{2}}.$$

Hence

$$\cos^2 \frac{\pi}{8} = \frac{1 + \cos \pi/4}{2} = \frac{1}{2} + \frac{1}{2\sqrt{2}}.$$

Taking the square root yields the desired answer, namely

$$\cos \frac{\pi}{8} = \sqrt{\frac{1}{2} + \frac{1}{2\sqrt{2}}}.$$

To put this into decimal form, it is easier to use a computing machine than your brain, and we leave the answer in the correct form above.

**Example.** To find  $\sin \pi/8$  we use SC 1, and the value for  $\cos \pi/8$  which we have just determined. Thus we obtain

$$\sin \frac{\pi}{8} = \sqrt{1 - \cos^2 \frac{\pi}{8}} = \sqrt{\frac{1}{2} - \frac{1}{2\sqrt{2}}}.$$

**EXERCISES**

1. Find  $\sin 7\pi/12$ . [Hint: Write  $7\pi/12 = 4\pi/12 + 3\pi/12$ .]

2. Find  $\cos 7\pi/12$ .

3. Find the following values:

a)  $\sin \pi/12$ ,

b)  $\cos \pi/12$ ,

c)  $\sin 5\pi/12$ ,

d)  $\cos 5\pi/12$ ,

e)  $\sin 11\pi/12$ ,

f)  $\cos 11\pi/12$ .

4. Prove the following formulas. They should be memorized.

a)  $\sin 2x = 2 \sin x \cos x$

b)  $\cos 2x = \cos^2 x - \sin^2 x$ .

c)  $\cos^2 x = \frac{1 + \cos 2x}{2}$

d)  $\sin^2 x = \frac{1 - \cos 2x}{2}$

Of course, you may assume Theorem 3 in proving these formulas. For formula (c), start with (b) and substitute  $1 - \cos^2 x$  for  $\sin^2 x$ . Use a similar idea for (d).

5. In each one of the following cases give a numerical value for  $\sin 2x$ , when  $\sin x$  has the indicated value.

a)  $\sin x = 0.7$

b)  $\sin x = 0.6$

c)  $\sin x = 0.4$

d)  $\sin x = 0.3$

e)  $\sin x = 0.2$

6. In each one of the following cases give a numerical value for  $\cos 2x$  when  $\sin x$  has the indicated value.

a)  $\sin x = 0.7$

b)  $\sin x = 0.6$

c)  $\sin x = 0.4$

d)  $\sin x = 0.3$

e)  $\sin x = 0.2$

7. In each case give a numerical value for  $\cos x/2$  when  $\cos x$  has the following value, and  $0 \leq x \leq \pi/2$ .

a)  $\cos x = 0.7$

b)  $\cos x = 0.6$

c)  $\cos x = 0.4$

d)  $\cos x = 0.3$

e)  $\cos x = 0.2$

8. In each of the cases of Exercise 7, what is the value for  $\cos x/2$  if we assume that  $-\pi/2 \leq x \leq 0$ ?

9. In each case give a numerical value for  $\sin x/2$  when  $\cos x$  has the following value, and  $0 \leq x \leq \pi/2$ .

a)  $\cos x = 0.7$

b)  $\cos x = 0.6$

c)  $\cos x = 0.4$

d)  $\cos x = 0.3$

e)  $\cos x = 0.2$

10. Find a formula for  $\sin 3x$  in terms of  $\sin x$  and  $\cos x$ . Similarly, for  $\sin 4x$  and  $\sin 5x$ .

11. Find a formula for  $\sin x/2$  if  $0 \leq x \leq \pi/2$  in terms of  $\cos x$  and possible square root signs.
12. a) A person throws a heavy ball at an angle  $\theta$  from the ground. Let  $d$  be the distance from the person to the point where the ball strikes the ground. Then  $d$  is given by

$$d = \frac{2v^2}{g} \sin \theta \cos \theta,$$

where  $v, g$  are constants. For what value of  $\theta$  is the distance a maximum? [Hint: Give another expression for  $2 \sin \theta \cos \theta$ .]

- b) You are watering the lawn, and point the watering hose at an angle of  $\theta$  degrees from the ground. The distance from the nozzle at which the water strikes the ground is given by

$$d = 2c \sin \theta \cos \theta,$$

where  $c$  is a constant. For what value of  $\theta$  is the distance a maximum?

13. Prove the following formulas for any integers  $m, n$ :

$$\sin mx \sin nx = \frac{1}{2}[\cos(m - n)x - \cos(m + n)x],$$

$$\sin mx \cos nx = \frac{1}{2}[\sin(m + n)x + \sin(m - n)x],$$

$$\cos mx \cos nx = \frac{1}{2}[\cos(m + n)x + \cos(m - n)x].$$

[Hint: Expand the right-hand side using the addition formulas.]

## §6. ROTATIONS

We have not yet investigated rotations from the point of view of coordinates, and we now fill this gap. We ask the basic question: Given a point  $P$  with coordinates  $(x, y)$ , let  $G_\varphi$  be the rotation with respect to the origin  $O = (0, 0)$  by an angle of  $\varphi$  radians. Let

$$G_\varphi(P) = P' = (x', y').$$

How do we describe the coordinates  $(x', y')$  of  $P'$  in terms of those of  $P$ ? The answer is quite simple, and we shall use polar coordinates, as well as the addition formula, to get this answer.

Let  $(r, \theta)$  be the polar coordinates of  $P$ . Then the polar coordinates of the point  $P'$  are simple  $(r, \theta + \varphi)$ .

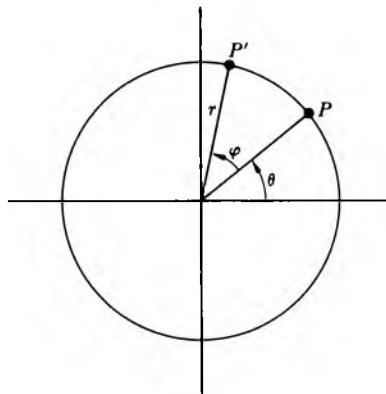


Fig. 11-32

On the other hand, we know that

$$\begin{aligned}x &= r \cos \theta, & x' &= r \cos(\theta + \varphi), \\y &= r \sin \theta, & y' &= r \sin(\theta + \varphi).\end{aligned}$$

We use the addition formula for the sine and cosine on the expressions on the right, and find

$$\begin{aligned}x' &= r[\cos \theta \cos \varphi - \sin \theta \sin \varphi], \\y' &= r[\sin \theta \cos \varphi + \sin \varphi \cos \theta].\end{aligned}$$

Expanding this out, and using the expression for  $x, y$  in terms of  $r, \cos \theta$  and  $\sin \theta$ , we find:

$$\begin{aligned}x' &= (\cos \varphi)x - (\sin \varphi)y, \\y' &= (\sin \varphi)x + (\cos \varphi)y.\end{aligned}$$

Thus rotation by an angle of  $\varphi$  radians is described by the four numbers

$$\begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}.$$

An array of numbers like

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

has a technical name: It is called a **matrix** (in fact a  $2 \times 2$  matrix).

**Example.** For instance,

$$\begin{pmatrix} 2 & -1 \\ 3 & 7 \end{pmatrix}$$

is a  $2 \times 2$  matrix.

**Example.** The matrix associated with the rotation  $G_{\pi/2}$  is the matrix

$$\begin{pmatrix} \cos \pi/2 & -\sin \pi/2 \\ \sin \pi/2 & \cos \pi/2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

If  $(x, y)$  are the coordinates of a point  $P$ , then the coordinates of the point  $P'$  obtained by rotating  $P$  by an angle of  $\pi/2$  are

$$\begin{aligned} x' &= -y, \\ y' &= x. \end{aligned}$$

The boxed formula for the coordinates  $(x', y')$  in terms of  $(x, y)$  is sometimes written in the form of a “product”,

$$\begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x' \\ y' \end{pmatrix}.$$

This is something like a multiplication.

In general, if

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is a  $2 \times 2$  matrix, we write the product

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}.$$

For example,

$$\begin{pmatrix} 3 & 2 \\ -1 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3x + 2y \\ -x + 5y \end{pmatrix}.$$

The theory of matrices can be considerably generalized, and we refer you to texts in linear algebra for this.

**EXERCISES**

In each one of the following cases, write the matrix associated with the rotation  $G_\varphi$  when  $\varphi$  has the indicated value. Write explicitly the coordinates  $(x', y')$  of  $G_\varphi(P)$  if  $P$  has coordinates  $(x, y)$ .

- |                        |                       |                       |
|------------------------|-----------------------|-----------------------|
| 1. $\varphi = \pi$     | 2. $\varphi = \pi/4$  | 3. $\varphi = \pi/3$  |
| 4. $\varphi = \pi/6$   | 5. $\varphi = 3\pi/4$ | 6. $\varphi = 3\pi/2$ |
| 7. $\varphi = -\pi/2$  | 8. $\varphi = -\pi/3$ | 9. $\varphi = -\pi/4$ |
| 10. $\varphi = 5\pi/4$ |                       |                       |

When  $(x, y)$  have the following numerical values, give the numerical values for  $(x', y')$  in each one of the rotations of Exercises 1 through 10. Draw the picture in each case.

- |                   |                    |
|-------------------|--------------------|
| 11. $P = (3, 1)$  | 12. $P = (5, -2)$  |
| 13. $P = (-2, 4)$ | 14. $P = (2, -1)$  |
| 15. $P = (0, 5)$  | 16. $P = (3, 0)$   |
| 17. $P = (-1, 1)$ | 18. $P = (-2, -1)$ |
| 19. $P = (2, 1)$  | 20. $P = (-2, -2)$ |
21. Associate a matrix with a dilation by  $r$ . Interpret dilation by  $r$  in terms of the matrix multiplication. Do the same for mixed dilations of type which we have written  $F_{a,b}$ .
22. Write down the matrices for the rotations  $G_\varphi$ ,  $G_\psi$ ,  $G_{\varphi+\psi}$ .

# 12 Some Analytic Geometry

## §1. THE STRAIGHT LINE AGAIN

Let  $F(x, y)$  be an expression involving a pair of numbers  $(x, y)$ . Let  $c$  be a number. We consider the equation

$$F(x, y) = c.$$

The set of points  $(x, y)$  for which this equation holds is called the **graph** of the equation. In this section we study the simplest case of equations like

$$3x - 2y = 5.$$

If you have read Chapter 10, you already know this equation, but we shall not assume any knowledge from this chapter. We develop everything from scratch.

Consider first a simple example, namely

$$y = 3x.$$

The set of points  $(x, 3x)$  is the graph of this equation, or equivalently of the equation

$$y - 3x = 0.$$

We can give  $x$  an arbitrary value, and thus we see that the graph looks like Fig. 12-1(a).

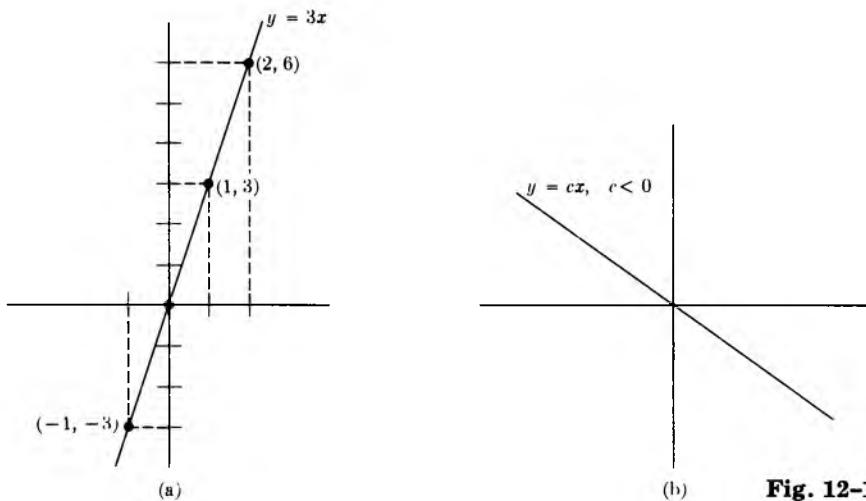


Fig. 12-1

If we consider the graph of

$$y = 4x,$$

we see that it is a line which slants more steeply. In general, the graph of the equation

$$y = ax,$$

where  $a$  is a number, represents a straight line. An arbitrary point on this line is of type

$$(x, ax) = x(1, a).$$

Thus the  $(x, y)$  coordinates of a point on the line are obtained by making the dilation of  $(1, a)$  by  $x$ .

If  $a$  is positive, then the line slants to the right. If  $a$  is negative, then the line slants to the left, as shown on Fig. 12-1(b). For instance, the graph of

$$y = -x$$

consists of all points  $(x, -x)$ , and looks like this (Fig. 12-2).

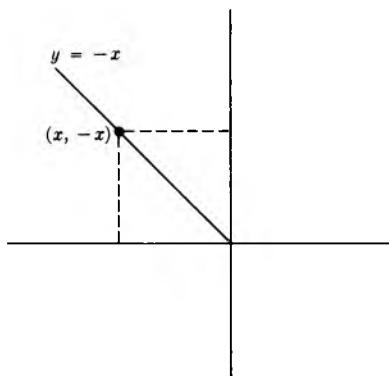


Fig. 12-2

If we drop the perpendiculars from the point  $(x, -x)$  to the axes, we obtain a right triangle, which in this case has legs of equal length.

Let  $a, b$  be numbers. The graph of the equation

$$(1) \quad y = ax + b$$

is also a straight line, which is parallel to the graph of the equation

$$(2) \quad y = ax.$$

To convince ourselves of this, we observe the following. Let

$$y' = y - b.$$

The equation

$$(3) \quad y' = ax$$

is of the type just discussed. If we have a point  $(x, y')$  on the graph of (3), then we get a point  $(x, y' + b)$  on the graph of (1), simply by adding  $b$  to the second coordinate. This means that the graph of the equation

$$y = ax + b$$

is the straight line parallel to the line determined by the equation

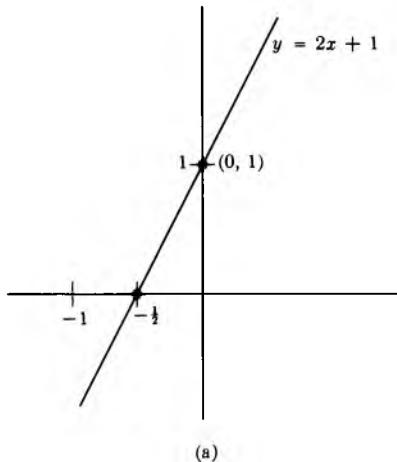
$$y = ax,$$

and passing through the point  $(0, b)$ .

**Example.** We want to draw the graph of the equation

$$y = 2x + 1.$$

When  $x = 0$ , then  $y = 1$ . When  $y = 0$ , then  $x = -\frac{1}{2}$ . Hence the graph looks like Fig. 12-3(a).



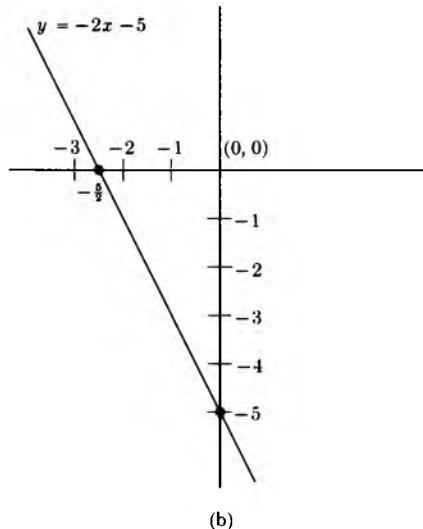
(a)

**Fig. 12-3**

**Example.** We want to draw the graph of the equation

$$y = -2x - 5.$$

When  $x = 0$ , then  $y = -5$ . When  $y = 0$ , then  $x = -\frac{5}{2}$ . Hence the graph looks like Fig. 12-3(b).



(b)

**Fig. 12-3 (cont.)**

If a line  $L$  is the graph of the equation

$$y = ax + b,$$

then we call the number  $a$  the **slope** of the line. For instance, the slope of the line whose equation is

$$y = -4x + 7$$

is  $-4$ . The slope determines how the line is slanted.

Let  $y = ax + b$  be the equation of a line.

The slope of this line can be obtained from two distinct points on the line. Let  $(x_1, y_1)$  and  $(x_2, y_2)$  be the two points. By definition, we know that

$$y_1 = ax_1 + b,$$

$$y_2 = ax_2 + b.$$

Subtracting, we find that

$$y_2 - y_1 = a(x_2 - x_1).$$

Consequently, if the two points are distinct,  $x_2 \neq x_1$ , and we can divide by  $(x_2 - x_1)$  to find

$$a = \frac{y_2 - y_1}{x_2 - x_1}.$$

This formula gives us the slope in terms of the coordinates of two distinct points.

**Example.** Consider the line defined by the equation

$$y = 2x + 5.$$

The two points  $(1, 7)$  and  $(-1, 3)$  lie on the line. The slope is equal to 2, and, in fact,

$$2 = \frac{7 - 3}{1 - (-1)}.$$

as it should be.

Geometrically, our quotient

$$\frac{y_2 - y_1}{x_2 - x_1}$$

is the ratio of the vertical side and horizontal side of the triangle in Fig. 12-4.

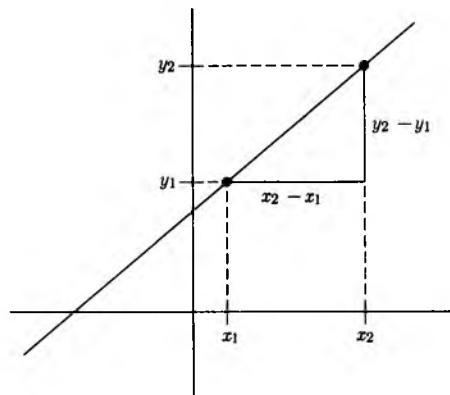


Fig. 12-4

Observe that it does not matter which point we call  $(x_1, y_1)$  and which one we call  $(x_2, y_2)$ . We would get the same value for the slope. This is because if we invert their order, then in the quotient expressing the slope both the numerator and denominator will change by a sign, so that their quotient does not change.

Conversely, given two points, it is easy to determine the equation of the line passing through the two points.

**Example.** Find the equation of the line passing through the points  $(1, 2)$  and  $(2, -1)$ .

We first find the slope. It must be the quotient

$$\frac{y_2 - y_1}{x_2 - x_1}$$

which in this case is equal to

$$\frac{-1 - 2}{2 - 1} = -3.$$

Thus we know that our line is the graph of the equation

$$y = -3x + b$$

for some number  $b$ , which we must determine. We know that the line passes through the point  $(1, 2)$ . Hence

$$2 = -3 \cdot 1 + b,$$

and we get the value of  $b$ , namely

$$b = 5.$$

Thus the equation of the line is

$$y = -3x + 5.$$

We can also determine the equation of a line given the slope and one point.

**Example.** Find the equation of the line having slope  $-7$  and passing through the point  $(-1, 2)$ .

The equation must be of the form

$$y = -7x + b$$

with some number  $b$ . Furthermore, when  $x = -1$  the corresponding value of  $y$  must be  $2$ . Hence

$$2 = (-7)(-1) + b,$$

and

$$b = -5.$$

Hence the equation of the line is

$$y = -7x - 5.$$

In general, the equation of the line passing through the point  $(x_1, y_1)$  and having slope  $a$  is

$$y - y_1 = a(x - x_1).$$

For points such that  $x \neq x_1$ , we can also write this in the form

$$\frac{y - y_1}{x - x_1} = a.$$

The equation of the line passing through two points  $(x_1, y_1)$  and  $(x_2, y_2)$  is

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}.$$

for all points such that  $x \neq x_1$ .

We should also mention vertical lines. These cannot be represented by equations of type

$$y = ax + b.$$

Suppose that we have a vertical line intersecting the  $x$ -axis at the point  $(2, 0)$ . The  $y$ -coordinate of any point on the line can be an arbitrary number, while the  $x$ -coordinate is always 2. Hence the equation of this line is simply

$$x = 2.$$

Similarly, the equation of a vertical line intersecting the  $x$ -axis at the point  $(c, 0)$  is

$$x = c.$$

When a line is given in the form  $y = ax + b$ , it is easy to find its intersection with a circle. We give an example.

**Example.** Find the points of intersection of the line and the circle given by the following equations:

$$y = 3x + 2 \quad \text{and} \quad x^2 + y^2 = 1.$$

To do this, note that a point  $(x, y)$  lies on the intersection if and only if  $y = 3x + 2$  and

$$x^2 + (3x + 2)^2 = 1.$$

Thus we must solve for  $x$  in this last equation, which is equivalent to

$$x^2 + 9x^2 + 12x + 4 = 1.$$

Again, this equation is equivalent to

$$10x^2 + 12x + 3 = 0,$$

which we can solve by the quadratic formula. We find

$$x = \frac{-12 \pm \sqrt{144 - 4 \cdot 30}}{20},$$

or in other words, we find the two values

$$x = \frac{-6 \pm \sqrt{6}}{10}.$$

These two possible values for  $x$  give us the two points of intersection, as illustrated in Fig. 12-5(a). The  $y$ -coordinates can be found by the expression  $y = 3x + 2$ , and hence the points of intersection are

$$\left( \frac{-6 + \sqrt{6}}{10}, \frac{2 + 3\sqrt{6}}{10} \right) \quad \text{and} \quad \left( \frac{-6 - \sqrt{6}}{10}, \frac{2 - 3\sqrt{6}}{10} \right).$$

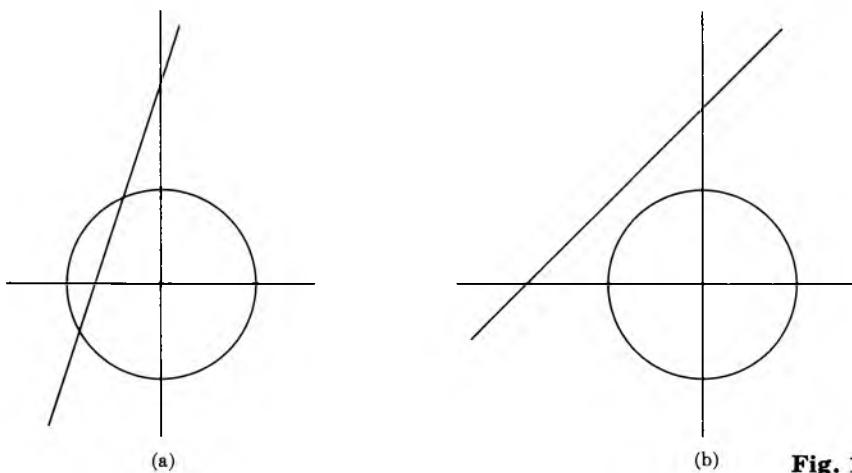


Fig. 12-5

Of course, it may happen that a circle and a line do not intersect, as shown in Fig. 12-5(b). If we followed the same procedure to solve for  $x$  as before, we would find in this case that the quadratic equation has no real solution because the number under the square root sign is negative.

**Example.** Find the points of intersection (if any) of the line and the circle given by the following equations:

$$y = x + 5 \quad \text{and} \quad x^2 + y^2 = 1.$$

We follow the same procedure as before, and we see that the  $x$ -coordinate of a point of intersection must satisfy the equation

$$x^2 + (x + 5)^2 = 1,$$

or equivalently

$$x^2 + x^2 + 10x + 25 = 1.$$

This amounts to solving

$$2x^2 + 10x + 24 = 0.$$

The quadratic formula gives

$$x = \frac{-10 \pm \sqrt{100 - 192}}{4} = \frac{-10 \pm \sqrt{-92}}{4}.$$

In this case, the number under the square root sign is negative, and hence there is no real solution. Hence the circle and the line do not intersect.

## **EXERCISES**

Sketch the graphs of the following lines.

1. a)  $y = -2x + 5$     b)  $y = 5x - 3$   
 2. a)  $y = \frac{x}{2} + 7$     b)  $y = -\frac{x}{3} + 1$   
 3. a) Find the point of intersection of the two lines of Exercise 1.  
     b) Find the point of intersection of the two lines of Exercise 2.

What is the equation of the line passing through the following points?

4. a)  $(-1, 1)$  and  $(2, -7)$       b)  $(3, \frac{1}{2})$  and  $(4, -1)$   
 5. a)  $(\sqrt{2}, -1)$  and  $(\sqrt{2}, 1)$       b)  $(-3, -5)$  and  $(\sqrt{3}, 4)$   
 6. Find the point of intersection of the lines in Exercise 4.  
 7. Find the point of intersection of the lines in Exercise 5.  
 8. Find the point of intersection of the following pairs of lines.

- a)  $3x - 2y = 1$  and  $4x - y = 2$
  - b)  $x - 5y = 1$  and  $2x + y = 0$
  - c)  $3x + 2y = 2$  and  $4x - y = 1$
  - d)  $-2x + 3y = 4$  and  $-5x - 3y = -2$

What is the equation of the line having the given slope and passing through the given point?

9. slope 4 and point  $(1, 1)$       10. slope  $-2$  and point  $(\frac{1}{2}, 1)$   
 11. slope  $-\frac{1}{2}$  and point  $(\sqrt{2}, 3)$       12. slope 3 and point  $(-1, 5)$

Sketch the graphs of the following lines:

13.  $x = 5$       14.  $x = -1$       15.  $x = -3$   
16.  $y = -4$       17.  $y = 2$       18.  $y = 0$

What is the slope of the line passing through the following points?

19.  $(1, \frac{1}{2})$  and  $(-1, 1)$       20.  $(\frac{1}{4}, 1)$  and  $(\frac{1}{2}, -1)$   
 21.  $(2, 1)$  and  $(\sqrt{2}, 1)$       22.  $(\sqrt{3}, 1)$  and  $(3, 2)$

What is the equation of the line passing through the following points?

23.  $(\pi, 1)$  and  $(\sqrt{2}, 3)$       24.  $(\sqrt{2}, 2)$  and  $(1, \pi)$   
 25.  $(-1, 2)$  and  $(\sqrt{2}, -1)$       26.  $(-1, \sqrt{2})$  and  $(-2, -3)$

27. Sketch the graphs of the following lines.

- a)  $y = 2x$       b)  $y = 2x + 1$       c)  $y = 2x + 5$   
d)  $y = 2x - 1$       e)  $y = 2x - 5$

28. For our purposes here, define two straight lines to be parallel if they have the same slope. Let

$$y = ax + b \quad \text{and} \quad y = cx + d$$

be the equations of two lines with  $b \neq d$ .

- a) If they are parallel, show that they have no point in common.
- b) If they are not parallel, show that they have exactly one point in common.

29. Find the common point of the following pairs of lines.

- a)  $y = 3x + 5$  and  $y = 2x + 1$
- b)  $y = 3x - 2$  and  $y = -x + 4$
- c)  $y = 2x + 3$  and  $y = -x + 2$
- d)  $y = x + 1$  and  $y = 2x + 7$

30. If a straight line is expressed in parametric form,

$$\{P + tA\}_{t \in \mathbb{R}}$$

and  $A = (a_1, a_2)$ , what is the slope of the line in terms of the coordinates of  $A$ ? Does this slope depend on the coordinates of  $P$ ?

31. Find the points of intersection, if any, of the indicated line and circle. Draw the pictures.

- a)  $y = 2x - 1$  and  $x^2 + y^2 = 1$
- b)  $y = 3x$  and  $x^2 + y^2 = 4$
- c)  $y = 3x - 2$  and  $x^2 + y^2 = 2$
- d)  $y = x - 1$  and  $(x - 1)^2 + (y - 2)^2 = 4$
- e)  $y = 4x + 1$  and  $(x - 3)^2 + (y - 4)^2 = 1$
- f)  $y = -2x - 1$  and  $(x - 3)^2 + y^2 = 1$
- g)  $y = -2x + 3$  and  $(x - 3)^2 + y^2 = 1$

## §2. THE PARABOLA

Next we consider the graph of the equation

$$y = x^2.$$

In this case, we see that for any value of  $x$ , the corresponding value of  $y$  is positive. We make a small table of values.

$x$	$y = x^2$
1	1
2	4
3	9
4	16

We see that as  $x$  increases, so does  $x^2$ . Furthermore  $(-x)^2 = x^2$ . Thus our graph  $(x, x^2)$  is symmetric with respect to the  $y$ -axis. It looks like this.

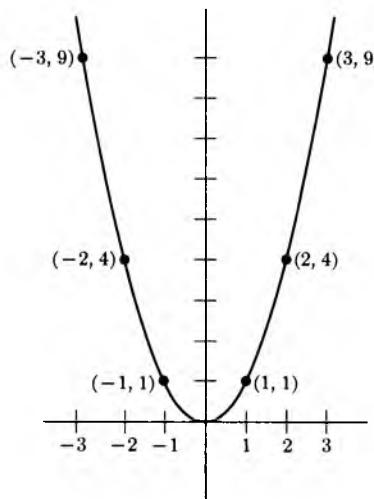


Fig. 12-6

Suppose that we want to draw the graph of the equation

$$(1) \quad y = (x - 1)^2.$$

We shall find that it looks exactly the same, but as if the origin were placed at the point  $(1, 0)$ . (See Fig. 12-7.) Namely, let

$$x' = x - 1.$$

Then our equation (1) is equivalent to

$$(2) \quad y = x'^2.$$

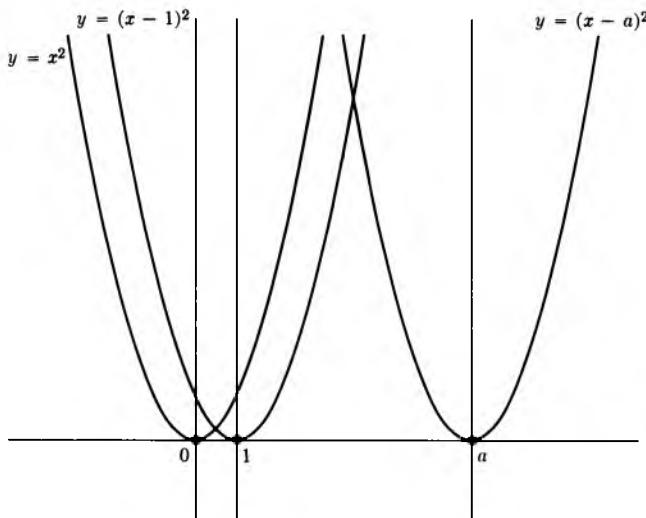


Fig. 12-7

Thus if we place our coordinate system with origin at the point  $(1, 0)$ , we see that equation (1) becomes equation (2) in terms of the new coordinates  $(x', y)$ . Similarly, the equation

$$y = (x - a)^2$$

has a graph which looks like the others, but translated to  $(a, 0)$ .

We can also perform a translation on  $y$ . Let  $(a, b)$  be a given point. We let

$$x' = x - a \quad \text{and} \quad y' = y - b.$$

Thus when  $x = a$ , we have  $x' = 0$  and when  $y = b$ , we have  $y' = 0$ . As we can see in Fig. 12-8, the graph of the equation

$$y' = x'^2$$

looks the same as the graph of the equation  $y = x^2$ , but translated to the point  $(a, b)$ .

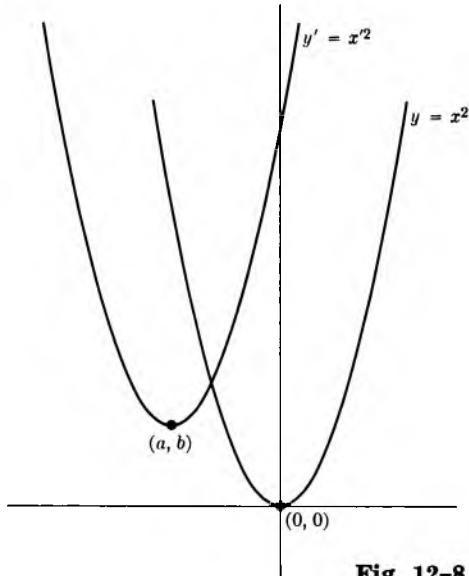


Fig. 12-8

Note that the equation  $y' = x'^2$  is the same as the equation

$$(y - b) = (x - a)^2$$

in terms of the coordinates  $(x, y)$ .

A curve which has an equation

$$(y - b) = c(x - a)^2$$

in some coordinate system is called a **parabola**.

**Example.** Describe the graph of the equation

$$2y - x^2 - 4x + 6 = 0.$$

We complete the square in  $x$ . We can write

$$x^2 + 4x = (x + 2)^2 - 4.$$

Hence our equation can be rewritten

$$2y = (x + 2)^2 - 10,$$

or

$$2(y + 5) = (x + 2)^2.$$

Choosing the new coordinates

$$x' = x + 2 \quad \text{and} \quad y' = y + 5,$$

our graph is defined in terms of these coordinates by the equation

$$y' = \frac{1}{2}x'^2,$$

which is easily drawn as in Fig. 12-9.

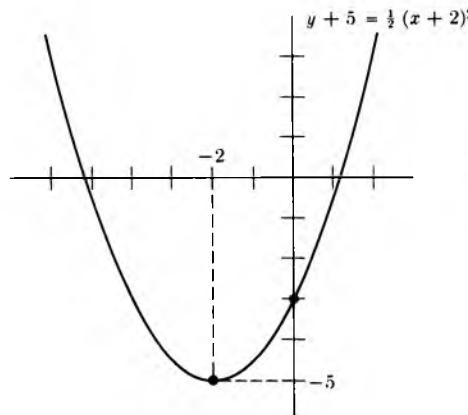


Fig. 12-9

Finally, we remark that if we have an equation

$$x - y^2 = 0$$

or

$$x = y^2,$$

then we get the graph of a curve which is also called a parabola, and is tilted horizontally as in Fig. 12-10.

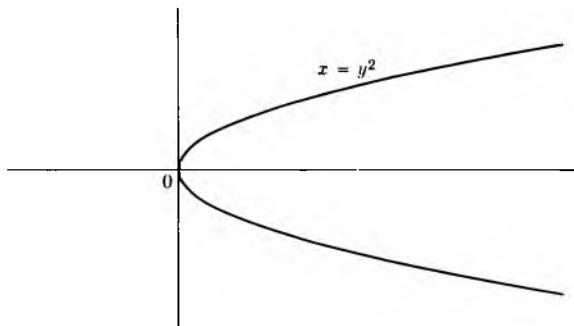


Fig. 12-10

In this case, all values of  $x$  are positive when  $y \neq 0$  because then  $y^2 > 0$ .

We can apply the same technique of changing the coordinate system to see what the graph of a more general equation looks like.

**Example.** Sketch the graph of the equation

$$x - y^2 + 2y + 5 = 0.$$

After completing the square in  $y$ , we can write this equation in the form

$$(x + 6) = (y - 1)^2,$$

and hence its graph looks like this (Fig. 12-11).

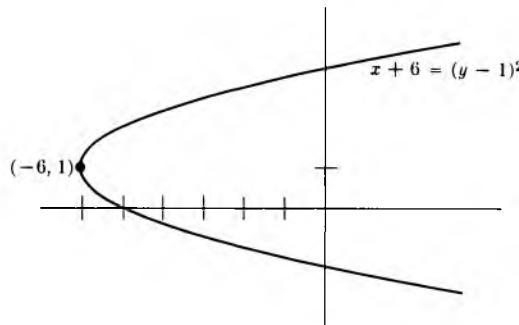


Fig. 12-11

We don't go into further properties of the parabola. However, we mention one of them here which shows how parabolas arise in optics. If one makes a mirror in the shape of a parabola, then any horizontal ray coming into the mirror gets reflected, and all these reflections meet at one point, called the **focus**  $F$  of the parabola. We have drawn this in Fig. 12-12.

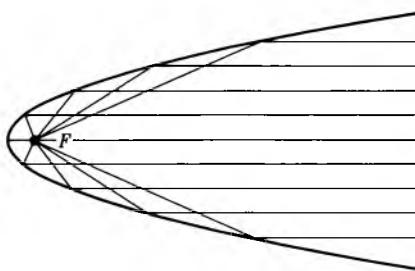


Fig. 12-12

Of course, if the parabola is given by an equation, we want a means of determining the coordinates of the focus. This belongs in a special course in analytic geometry, or optics, and we don't cover it here in our general survey of basic mathematics. A similar remark holds for other properties of the various curves discussed in this chapter (ellipse and hyperbola, in addition to the parabola). They have a number of interesting properties which are traditionally covered at this elementary level. However, I feel that it is best to postpone the discussion of such properties to the time when they are needed. All we are trying to do here is to get a quick acquaintance with the graphs of simple curves defined by simple equations. This is more than sufficient background to do calculus and many other applications (in physics, engineering, or economics).

## EXERCISES

Sketch the graphs of the following equations.

- |                             |                         |
|-----------------------------|-------------------------|
| 1. a) $y = 3x^2$            | b) $y = \frac{1}{2}x^2$ |
| 2. a) $y = 2x^2 + x - 3$    | b) $x - 4y^2 = 0$       |
| 3. a) $x - y^2 + y + 1 = 0$ | b) $y - 5 = (x + 3)^2$  |

4. a)  $(y + 4) = (x + 2)^2$       b)  $(y - 1)^2 = x + 7$

Complete the square in the following equations and change the coordinate system to put them into the form

$$x'^2 + y'^2 = r^2 \quad \text{or} \quad y' = cx'^2, \quad \text{or} \quad x' = cy'^2.$$

with a suitable constant  $c$ . Sketch the graphs.

- |  |                                |
|--|--------------------------------|
| 5. $x^2 + y^2 - 4x + 2y - 20 = 0$  | 6. $x^2 + y^2 - 2y - 8 = 0$    |
| 7. $x^2 + y^2 + 2x - 2 = 0$  | 8. $y - 2x^2 - x + 3 = 0$      |
| 9. $y - x^2 - 4x - 5 = 0$  | 10. $y - x^2 + 2x + 3 = 0$     |
| 11. $x^2 + y^2 + 2x - 4y = -3$   | 12. $x^2 + y^2 - 4x - 2y = -3$ |
| 13. $x - 2y^2 - y + 3 = 0$   | 14. $x - y^2 - 4y = 5$         |
| 15. $y = -x^2$   | 16. $y = -2x^2$                |
| 17. $y = -3x^2$  | 18. $y = -(x - 1)^2$           |
| 19. $y = -(x + 2)^2$   | 20. $y = -(x + 3)^2$           |
| 21. $y = -(x - 3)^2$   | 22. $y - 1 = -(x + 5)^2$       |
| 23. $y + 2 = -(x + 2)^2$   | 24. $y + 2 = -(x - 3)^2$       |
| 25. Find the point of intersection of the parabola and the straight line given by the following equations. |                                |
| a) $3x + y = 1$ and $y = 3x^2$   |                                |
| b) $4x - 2y = 3$ and $y = -2x^2$   |                                |
| c) $y = 4x - 1$ and $y = -(x - 2)^2$   |                                |
| d) $y = -3x + 5$ and $y = (x + 3)^2$   |                                |
| e) $y = 4x - 1$ and $(y - 2) = (x - 3)^2$  |                                |

### §3. THE ELLIPSE

Let  $a$  be a positive number. If  $(x, y)$  is a point in the plane, then we call the point

$$(ax, ay)$$

the dilation of  $(x, y)$  by  $a$ . Multiplication of each coordinate by  $a$  amounts to stretching by  $a$ .

We can generalize this slightly. Let  $a, b$  be positive numbers. To each point  $(x, y)$  we associate the point  $(ax, by)$ . Thus we stretch the first coordinate by  $a$  and the second coordinate by  $b$ . The association

$$(x, y) \mapsto (ax, by)$$

will be denoted by  $F_{a,b}$ .

Consider the circle consisting of all points  $(x, y)$  such that

$$(1) \quad x^2 + y^2 = 1.$$

It is the unit circle centered at the origin. What happens to this circle when we impose on it this mixed type of dilation? Let

$$u = ax$$

and

$$v = by.$$

Then  $u, v$  satisfy the equation

$$(2) \quad \frac{u^2}{a^2} + \frac{v^2}{b^2} = 1.$$

Conversely, if  $(u, v)$  is a point satisfying this equation, letting

$$x = \frac{u}{a}$$

and

$$y = \frac{v}{b}$$

shows that  $(u, v)$  is the image of  $(x, y)$  under  $F_{a,b}$ . Hence the image of the circle defined by equation (1) is the set of points  $(u, v)$  satisfying equation (2). This image is called an **ellipse**. In general, an **ellipse** is a curve in the plane which is the graph of an equation of type (2) in some coordinate system.

**Example.** The set of points  $(u, v)$  satisfying the equation

$$\frac{u^2}{3^2} + \frac{v^2}{2^2} = 1$$

is an ellipse. We wish to draw its graph. Call the horizontal axis the  $u$ -axis, and the vertical axis the  $v$ -axis. When  $u = 0$ , we see that  $v^2 = 2^2$  so that  $v = \pm 2$ . When  $v = 0$ , we see that  $u^2 = 3^2$  so that  $u = \pm 3$ . If we visualize the circle of radius 1 undergoing the mixed stretching  $F_{3,2}$ , then we see that the graph of the ellipse looks like this (Fig. 12-13).

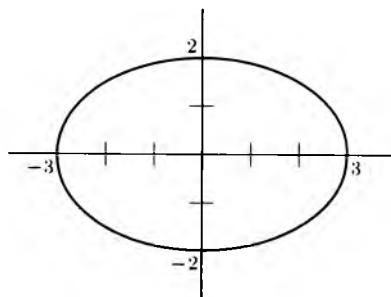


Fig. 12-13

We can also perform a translation on an ellipse just as we did for circles and parabolas.

**Example.** Draw the graph of the ellipse defined by the equation

$$\frac{(x - 1)^2}{25} + \frac{(y - 4)^2}{16} = 1.$$

Let  $x' = x - 1$  and  $y' = y - 4$ . Then  $(x', y')$  satisfy the equation

$$\frac{x'^2}{5^2} + \frac{y'^2}{4^2} = 1.$$

Thus we have the equation of an ellipse centered at the point  $(1, 4)$ , as shown in Fig. 12-14.

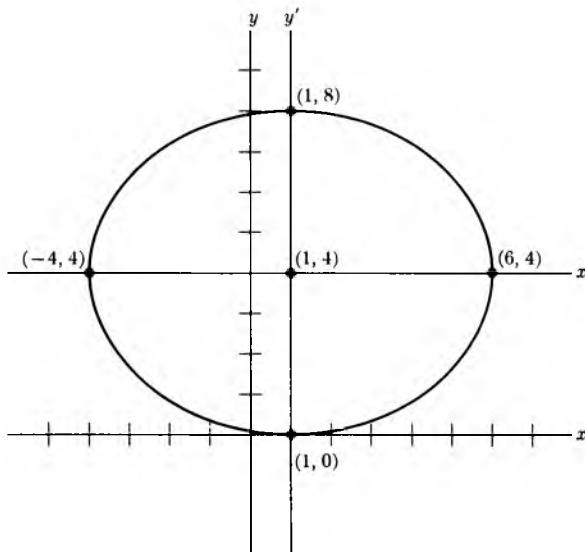


Fig. 12-14

**EXERCISES**

Sketch the graphs of the following equations. In each case, indicate the center of the ellipse, and its extremities.

1.  $\frac{x^2}{16} + \frac{y^2}{4} = 1$

2.  $\frac{(x - 3)^2}{4} + \frac{(y - 1)^2}{9} = 1$

3.  $\frac{(x + 2)^2}{4} + \frac{(y - 1)^2}{16} = 1$

4.  $\frac{(x - 1)^2}{25} + \frac{(y + 3)^2}{9} = 1$

5.  $\frac{x^2}{16} + \frac{(y + 1)^2}{9} = 1$

6.  $\frac{(x + 3)^2}{25} + \frac{y^2}{4} = 1$

7.  $4x^2 + 9y^2 = 1$

8.  $25x^2 + 16y^2 = 1$

9.  $9x^2 + 16y^2 = 1$

10.  $4x^2 + 25y^2 = 1$

11.  $4(x - 1)^2 + 16(y + 2)^2 = 1$

12.  $16(x + 3)^2 + 9(y + 1)^2 = 1$

13. Find the points of intersection of the ellipses given in Exercises 1 through 12 with the straight line given by the equation  $y = 2x - 1$ .

**§4. THE HYPERBOLA**

We wish to graph the equation

$$xy = 1.$$

First observe that if  $x, y$  satisfy this equation, they cannot be equal to 0.

We can solve for  $y$  in terms of  $x$ , namely

$$y = \frac{1}{x}.$$

We make a small table of values.

$x$	$y$	$x$	$y$
1	1	$\frac{1}{2}$	2
2	$\frac{1}{2}$	$\frac{1}{3}$	3
3	$\frac{1}{3}$	$\frac{1}{4}$	4
100	$\frac{1}{100}$		

We see that as  $x$  grows larger,  $y$  grows smaller. Also we have an obvious symmetry when  $x$  takes on negative values. Thus the graph looks like this (Fig. 12-15).

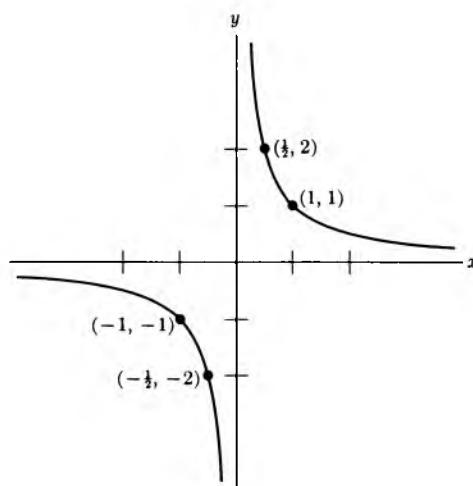


Fig. 12-15

Similarly, we can perform a translation as we did before.

**Example.** Sketch the graph of the equation

$$y - 2 = \frac{1}{x + 3}.$$

Let  $x' = x + 3$  and  $y' = y - 2$ . Then the coordinates  $x'$ ,  $y'$  satisfy the equation

$$x'y' = 1.$$

Thus our graph looks like this (Fig. 12-16).

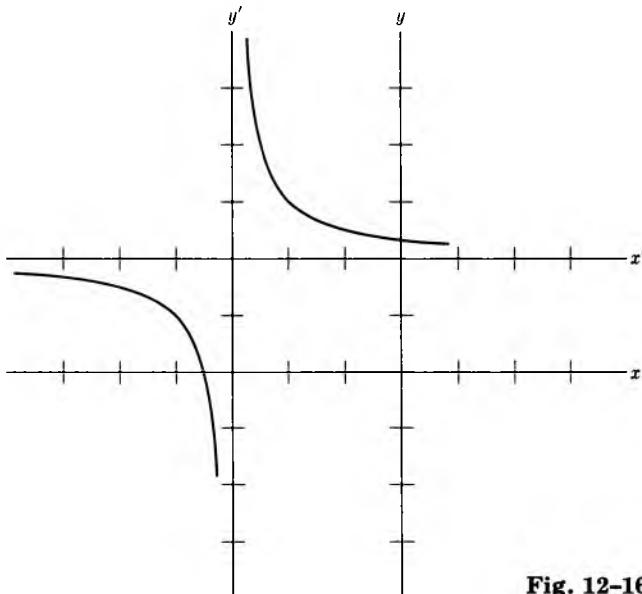


Fig. 12-16

If  $c$  is a number, we can graph in a similar way the equation  $xy = c$ . A curve which has an equation of type  $xy = c$  in some coordinate system is called a **hyperbola**.

**Example.** If we are given an equation like

$$xy - 2x + 3y + 4 = 5,$$

then we can factor the left-hand side and rewrite the equation as

$$(x + 3)(y - 2) + 6 + 4 = 5,$$

or

$$(x + 3)(y - 2) = -5.$$

In terms of the new coordinates  $x' = x + 3$  and  $y' = y - 2$ , we can rewrite our equation in the form

$$x'y' = -5.$$

Using a table of values, and a similar analysis of what happens to  $y'$  when  $x'$  increases or decreases, we see that the graph looks like this (Fig. 12-17).

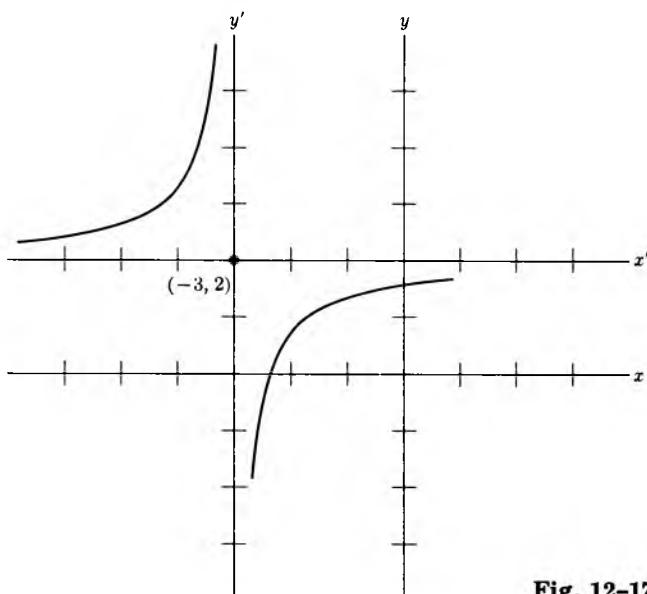


Fig. 12-17

**Example.** In physics we encounter cases where an object is repelled from the origin along a straight line, with a force whose magnitude is inversely proportional to the distance from the origin. Suppose this straight line is the  $x$ -axis. Let  $y$  be the magnitude of the force acting on the particle. Then we can write

$$y = \frac{k}{x},$$

for some constant  $k$ . For instance, we can take  $k = 2$ . We can then draw the graph of the force, which is similar to a hyperbola. We make a small table of values to get an idea of the behavior of the curve.

$x$	$y$	$x$	$y$
1	2	$\frac{1}{2}$	4
2	1	$\frac{1}{3}$	6
3	$\frac{2}{3}$	$\frac{1}{4}$	8
4	$\frac{1}{2}$	$\frac{1}{5}$	10
5	$\frac{2}{5}$		

We see that the graph looks like this (Fig. 12-18).

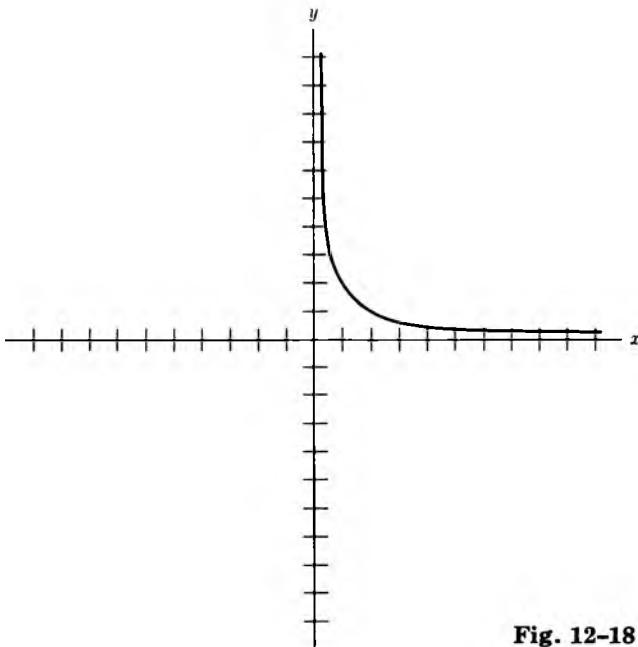


Fig. 12-18

### EXERCISES

Sketch the graphs of the following curves, defined by the given equations. If need be, make small tables of values. Try to rewrite the equation in a form that makes it obvious what the graph looks like.

1. a)  $y = \frac{3}{x}$       b)  $y = \frac{1}{2x}$       c)  $y = \frac{1}{3x}$

2. a)  $y = \frac{1}{x - 1}$       b)  $y - 4 = \frac{2}{x - 2}$       c)  $y = \frac{1}{2(x - 3)}$

3. a)  $(x - 1)(y - 2) = 1$       b)  $x(y + 1) = 1$

4. a)  $(x - 1)(y - 2) = 2$       b)  $x(y + 1) = 3$

5.  $xy - 4 = 0$

6.  $y = \frac{2}{1-x}$

7.  $y = \frac{3}{x+1}$

8.  $(x+2)(y-1) = 2$

9.  $(x-1)(y-1) = 2$

10.  $(x-1)(y-1) = 1$

11.  $y = \frac{1}{x-2} + 4$

12.  $y = \frac{1}{x+1} - 2$

13.  $y = \frac{4x-7}{x-2}$

14.  $y = \frac{-2x-1}{x+1}$

15.  $y = \frac{x+1}{x-1}$

16.  $y = \frac{x-1}{x+1}$

17.  $xy = -1$

18.  $xy = -2$

19.  $xy = -3$

20.  $xy = -4$

21.  $(x-1)y = -2$

22.  $(x+1)y = -2$

23.  $(x-1)(y+2) = -1$

24.  $(x-1)(y-2) = -1$

25.  $(x+1)(y-3) = -4$

26.  $(x-1)(y+3) = -4$

27. Find the point of intersection of the hyperbolas given in Exercises 1 through 26 with the straight line given by the equation  $y = x - 3$ .

*The next section is slightly less important than the preceding one, and may be omitted.*

## §5. ROTATION OF HYPERBOLAS

There is another standard equation for a hyperbola which occurs frequently, and which it is useful to recognize, namely, the equation

(1) 
$$y^2 - x^2 = 1,$$

or more generally, the equation

$$y^2 - x^2 = c,$$

with some constant  $c$ . We shall now see that these equations represent

hyperbolas, obtained by rotating the hyperbola defined by our standard equation

$$xy = 1,$$

and then perhaps performing a dilation.

**Theorem 1.** *Let  $H$  be the set of points  $(x, y)$  satisfying the equation*

$$xy = 1.$$

*Let  $G$  be rotation by  $\pi/4$ , counterclockwise, as usual. Then the image of  $H$  under  $G$  is the curve  $H'$  consisting of all points  $(u, v)$  satisfying the equation*

$$(2) \quad v^2 - u^2 = 2.$$

We may illustrate Theorem 1 in Fig. 12-19.

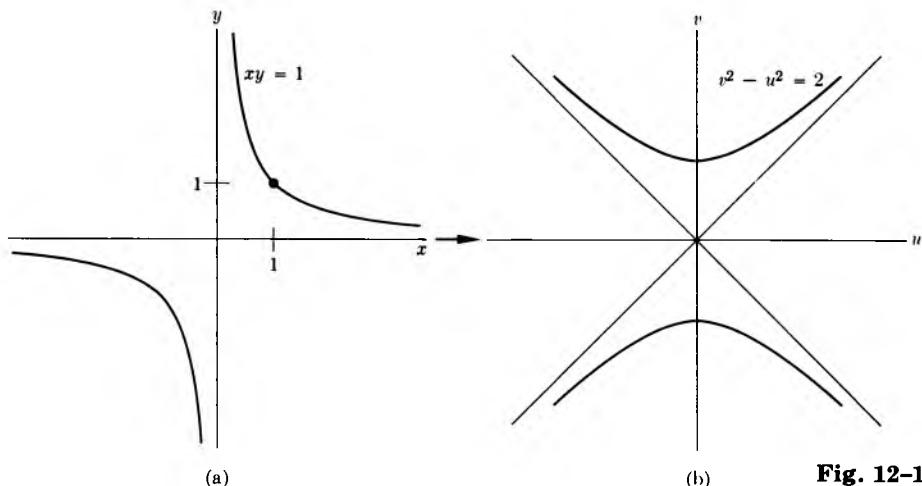


Fig. 12-19

Fig. 12-19(b) represents the hyperbola of Fig. 12-19(a) rotated counterclockwise by  $\pi/4$ . We have also drawn the axes of Fig. 12-19(a), rotated by  $\pi/4$ .

*Proof of Theorem 1.* We recall that under rotation by an angle,  $\theta$ , the coordinates of a point  $(x, y)$  change by the matrix

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

(Cf. Chapter 11, §6.) This means that if  $(u, v)$  are the coordinates of the rotated point, then

$$\begin{aligned} u &= (\cos \theta)x - (\sin \theta)y \\ v &= (\sin \theta)x + (\cos \theta)y. \end{aligned}$$

We apply this when  $\theta = \pi/4$ , in which case we have

$$\cos \theta = \sin \theta = \frac{1}{\sqrt{2}}.$$

We then see that the coordinates of the rotated point  $(u, v)$  are given by

$$\begin{aligned} (3) \quad u &= \frac{1}{\sqrt{2}}x - \frac{1}{\sqrt{2}}y = \frac{1}{\sqrt{2}}(x - y) \\ v &= \frac{1}{\sqrt{2}}x + \frac{1}{\sqrt{2}}y = \frac{1}{\sqrt{2}}(x + y). \end{aligned}$$

Furthermore, we get:

$$\begin{aligned} v^2 - u^2 &= \frac{1}{2}(x + y)^2 - \frac{1}{2}(x - y)^2 \\ &= \frac{1}{2}[x^2 + 2xy + y^2 - (x^2 - 2xy + y^2)] \\ &= \frac{1}{2} \cdot 4xy \\ &= 2xy. \end{aligned}$$

If a point  $(x, y)$  satisfies the condition  $xy = 1$ , we conclude that its image  $(u, v)$  satisfies the equation

$$v^2 - u^2 = 2.$$

Hence we have proved that the image of  $H$  is contained in the curve  $H'$ . In other words, we have proved that if  $(x, y)$  satisfies (1), then its rotation by  $\pi/4$  satisfies (2).

Conversely, let  $P' = (u, v)$  be a point satisfying (2), i.e. such that

$$v^2 - u^2 = 2.$$

We must show that  $P' = G(P)$  for some point  $P = (x, y)$  satisfying (1). Our intuition immediately tells us how to find  $(x, y)$ : these coordinates are obtained by rotating  $(u, v)$  through an angle of  $-\pi/4$ . Thus we let

$$\begin{aligned} x &= \frac{1}{\sqrt{2}}(u + v) \\ y &= \frac{1}{\sqrt{2}}(-u + v). \end{aligned}$$

Taking the product, we find that

$$xy = \frac{1}{2}(u + v)(v - u) = \frac{1}{2}(v^2 - u^2) = 1.$$

If you now use formulas (3) applied to this point  $(x, y)$ , you will find precisely that they yield  $(u, v)$ . This proves what we wanted, and concludes the proof of Theorem 1.

**Remark.** Instead of rotating the hyperbola defined by the equation  $xy = 1$ , suppose that we rotate the hyperbola represented by the equation

$$xy = 6.$$

Denote this hyperbola by  $H_{\sqrt{6}}$ . It is obtained from  $H$  by a dilation, namely, we can rewrite our equation in the form

$$\frac{x}{\sqrt{6}} \cdot \frac{y}{\sqrt{6}} = 1.$$

Let  $x' = x/\sqrt{6}$  and  $y' = y/\sqrt{6}$ . Then  $(x', y')$  satisfy the equation

$$x'y' = 1.$$

We also have

$$x = \sqrt{6} \cdot x' \quad \text{and} \quad y = \sqrt{6} \cdot y'.$$

Thus  $H_{\sqrt{6}}$  is the dilation by  $\sqrt{6}$  of the hyperbola  $H$ .

In general, let  $r$  be a positive number. The same argument shows that the hyperbola  $H_r$  represented by the equation

$$xy = r^2$$

is the dilation by  $r$  of the hyperbola  $H = H_1$ .

Let  $G$  be any rotation. For any positive number  $r$ , and any point  $P$ , we note that

$$G(rP) = rG(P).$$

In other words, dilating by  $r$  followed by the rotation  $G$  is the same as first taking the rotation  $G$  and then dilating by  $r$ . We could say that a rotation commutes with a dilation. We then see that rotating the hyperbola  $H_r$  by  $\pi/4$  yields the hyperbola defined by the equation

$$v^2 - u^2 = 2r^2.$$

**Example.** The hyperbola

$$v^2 - u^2 = 18$$

is the rotation by  $\pi/4$  of the hyperbola defined by the equation

$$xy = 9.$$

**Example.** The hyperbola

$$v^2 - u^2 = 10$$

is the rotation by  $\pi/4$  of the hyperbola represented by the equation

$$xy = 5.$$

(Watch out: We have  $10 = 2 \cdot 5$ , so that we use  $r = \sqrt{5}$ .)

**Remark.** When sketching the graph of a hyperbola with an equation like

$$v^2 - u^2 = 10,$$

it is useful to note the point at which  $u = 0$ , i.e. the points where the hyperbola meets the vertical axis. The  $v$ -coordinates of these points satisfy

$$v^2 = 10,$$

or in other words,

$$v = \sqrt{10} \quad \text{or} \quad v = -\sqrt{10}.$$

Thus the graph of this equation can be sketched as in Fig. 12-20.

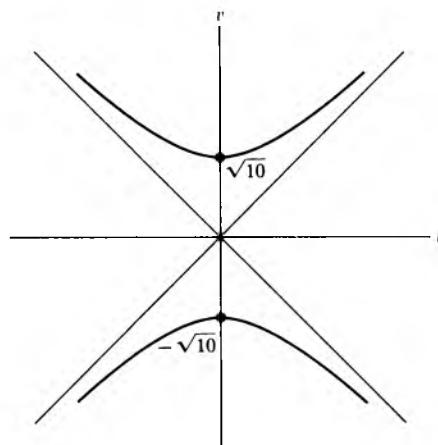


Fig. 12-20

**EXERCISES**

1. Sketch the graphs of the following hyperbolas.
  - a)  $v^2 - u^2 = 8$
  - b)  $v^2 - u^2 = 1$
  - c)  $v^2 - u^2 = 4$
2. Rotate the hyperbola  $H$  defined by the equation  $xy = 1$  by  $-\pi/4$  (i.e. clockwise by an angle of  $\pi/4$ ). What is the equation satisfied by the image of  $H$ ?
3. Rotate the hyperbola  $H$  defined by the equation  $xy = 1$  by
  - a)  $3\pi/4$ , and
  - b)  $-3\pi/4$ .In each case give the equation satisfied by the image of  $H$  under the prescribed rotation, and sketch this image.
4. Sketch the graph of the hyperbolas:
  - a)  $xy = -1$ ,
  - b)  $xy = -2$ ,
  - c)  $xy = -3$ ,
  - d)  $xy = -4$ .
5. In each one of the cases of Exercise 4, rotate the given hyperbola by  $\pi/4$ . Give the equation of the image.
6. In each one of the cases of Exercise 4, rotate the given hyperbola by  $-\pi/4$ . Give the equation of the image.
7. Sketch the graph of the following hyperbolas.
  - a)  $u^2 - v^2 = 2$
  - b)  $u^2 - v^2 = 1$
  - c)  $u^2 - v^2 = 18$
  - d)  $u^2 - v^2 = 4$
8. Prove the statement made in the text: If  $G$  is a rotation and  $F_r$  is dilation by  $r$ , then  $G \circ F_r = F_r \circ G$ .

*Part Four*

**MISCELLANEOUS**



# 13 Functions

## §1. DEFINITION OF A FUNCTION

A **function**, defined for all numbers, is an association which to each number associates another number. If we denote a function by  $f$ , then this association is denoted by

$$x \mapsto f(x).$$

We call  $f(x)$  the **value** of the function at  $x$ , or the **image** of  $x$  under  $f$ .

**Example.** The association

$$f: x \mapsto x^2$$

is a function called the **square**. Similarly, the association

$$g: x \mapsto x + 1$$

is a function. We have  $f(2) = 4$ ,  $f(3) = 9$ ,  $f(-1) = 1$  and  $f(-2) = 4$ . We have  $g(1) = 2$ ,  $g(2) = 3$ ,  $g(50) = 51$ .

**Example.** If you have read the chapter on trigonometry, then you now see that the association

$$x \mapsto \sin x$$

is a function, which we already called the **sine**.

**Example.** The association which to each number  $x$  associates the number 4 is called the **constant function, with constant value 4**. Similarly, if  $c$  is a given number, the association

$$x \mapsto c,$$

for all numbers  $x$

is called the **constant function with value  $c$** .

**Remark on terminology.** We have adopted the convention that the values of a function are numbers. Although this is not a universal convention, I find it useful. For an analogous notion when other types of values are allowed, see mappings in the next chapter.

We would like to say that the square root is also a function, but we know that negative numbers do not have real square roots. Thus we extend our definition of function as follows.

Let  $S$  be a set of numbers. A **function defined on  $S$**  is an association which to each element of  $S$  associates a number.

**Example.** Let  $S$  be the set of numbers  $\geq 0$ . Then the square root is a function defined on  $S$ . Its value at 4 is 2. Its value at 10 is  $\sqrt{10}$ .

**Example.** Let  $S$  be the set of numbers  $\neq 0$ . Let  $f$  be the function defined on  $S$  such that

$$f(x) = \frac{1}{x}$$

for all  $x$  in  $S$ . Then  $f(1) = 1$ ,  $f(2) = \frac{1}{2}$ ,  $f(100) = \frac{1}{100}$ , and  $f(\frac{3}{2}) = \frac{2}{3}$ .

**Example.** Let  $n$  be a positive integer. Then the association

$$x \mapsto x^n$$

is a function, which we called the  $n$ -th power.

**Warning.** It is very convenient sometimes to use slightly incorrect language. Often you will find written a sentence like:

“Let  $f(x)$  be such and such a function.”

The trouble here is that there is no quantification of  $x$ . What one means in that case, is the function whose value at a number  $x$  is such and such. We shall try to avoid incorrect language at least at the beginning of our discussion.

The preceding examples of functions have been given by formulas. Functions can, of course, be defined quite arbitrarily.

**Example.** Let  $G$  be the function such that

$$G(x) = 0 \text{ if } x \text{ is a rational number,}$$

$$G(x) = 1 \text{ if } x \text{ is not a rational number.}$$

Then in particular,  $G(2) = G(\frac{2}{3}) = G(-\frac{3}{4}) = 0$ , but

$$G(\sqrt{2}) = 1.$$

To describe a function amounts to giving its values at all numbers for which it is defined.

Let  $f, g$  be functions defined on the same set  $S$ . We can define the **sum**  $f + g$  of the two functions to be the function whose value at an element  $x$  of  $S$  is

$$f(x) + g(x).$$

Then associativity for addition of numbers gives us associativity for addition of functions; namely, for any functions  $f, g, h$  defined on  $S$ , we have

$$(f + g) + h = f + (g + h).$$

Similarly, for commutativity, we have

$$f + g = g + f.$$

**Example.** Let  $f(x) = x^2$  and  $g(x) = \sin x$ . Then

$$(f + g)(x) = x^2 + \sin x.$$

Also,

$$(f + g)(\pi) = \pi^2 + \sin \pi = \pi^2,$$

because  $\sin \pi = 0$ .

We have the **zero function**, whose value at every element  $x$  of  $S$  is 0. We denote this function also by 0. Thus for any function  $f$  defined on  $S$ , we have

$$f + 0 = 0 + f = f.$$

If  $f$  is a function defined on  $S$ , then we can define minus  $f$ , written  $-f$ , to be the function whose value at an element  $x$  of  $S$  is

$$-f(x).$$

**Example.** If  $f(x) = x^2$ , then  $(-f)(x) = -x^2$ , and  $(-f)(5) = -25$ . We see that

$$f + (-f) = 0 \quad (\text{the zero function}).$$

Thus we see that functions satisfy the same basic rules for addition as numbers.

The same is true for multiplication. If  $f, g$  are functions defined on the same set  $S$ , we define their **product**  $fg$  to be the function whose value at an element  $x$  of  $S$  is the product

$$f(x)g(x).$$

Thus we have by definition

$$(fg)(x) = f(x)g(x).$$

This product is commutative and associative. Furthermore, if 1 denotes the constant function having the value 1 for all  $x$  in  $S$ , then we have the usual rules

$$1f = f \quad \text{and} \quad 0f = 0.$$

Finally, our multiplication is distributive with respect to addition, because

$$\begin{aligned} ((f + g)h)(x) &= (f + g)(x) \cdot h(x) \\ &= (f(x) + g(x))h(x) \\ &= f(x)h(x) + g(x)h(x) \\ &= (fh)(x) + (gh)(x) \\ &= (fh + fg)(x). \end{aligned}$$

Thus

$$(f + g)h = fh + gh.$$

In physical life, functions of numbers occur when we describe one quantity in terms of another.

**Example.** To each year we associate the population of the United States. Then the function is defined only for those years which are  $\geq 1776$ . If  $P$  denotes this function ( $P$  for population), then

$$\begin{aligned} P(1800) &= 7.2 \cdot 10^6, \\ P(1900) &= 76.0 \cdot 10^6, \\ P(1940) &= 140 \cdot 10^6, \\ P(1970) &= 200 \cdot 10^6. \end{aligned}$$

**Example.** To each year we associate the price (in cents) of subway fare in New York City. Let  $S$  denote this function ( $S$  for subway). Then

$$\begin{aligned} S(1950) &= 15, \\ S(1969) &= 20, \\ S(1970) &= 30. \end{aligned}$$

We suggest that you look at the section on arbitrary mappings for a more general notion.

**EXERCISES**

1. Let  $f(x) = 1/x$ . What is  $f(\frac{3}{4})$ ,  $f(-\frac{2}{3})$ ?
2. For what numbers could you define a function  $f$  by the formula

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

What is the value of this function for  $x = 5$ ?

3. For what numbers could you define a function  $f$  by the formula

$$f(x) = \sqrt[3]{x} \quad (\text{cube root of } x) ?$$

What is  $f(27)$ ?

4. Let  $x \mapsto |x|$  be the function which we already met earlier, namely

$$|x| = \sqrt{x^2}.$$

What is a)  $f(1)$ ? b)  $f(-3)$ ? c)  $f(-\frac{4}{3})$ ?

5. Let  $f(x) = x + |x|$ . What is

a) $f(\frac{1}{2})$ ?	b) $f(2)$ ?
c) $f(-4)$ ?	d) $f(-5)$ ?

6. Let  $f(x) = 2x + x^2 - 5$ . What is

a) $f(1)$ ?	b) $f(-1)$ ?
-------------	--------------

7. For what numbers could you define a function  $f$  by the formula  $f(x) = \sqrt[4]{x}$  (fourth root of  $x$ )? What is  $f(16)$ ?

8. A function (defined for all numbers) is said to be an **even** function if  $f(x) = f(-x)$  for all numbers  $x$ . It is said to be an **odd** function if  $f(x) = -f(-x)$  for all  $x$ . Determine which of the following functions are odd or even.

a) $f(x) = x$	b) $f(x) = x^2$	c) $f(x) = x^3$
d) $f(x) = 1/x$ if $x \neq 0$ and $f(0) = 0$		

9. Show that any function defined for all numbers can be written as a sum of an even function and an odd function. [Hint: The term

$$\frac{f(x) + f(-x)}{2} .$$

will be the even function.]

10. Which of the following functions is odd or even, or neither?

- |               |               |
|---------------|---------------|
| a) $\sin x$   | b) $\cos x$   |
| c) $\tan x$   | d) $\cot x$   |
| e) $\sin^2 x$ | f) $\cos^2 x$ |
| g) $\tan^2 x$ |               |

11. a) Show that the sum of odd functions is odd.

b) Show that the sum of even functions is even.

12. Determine whether the product of the following types of functions is odd, even, or neither. Prove your assertions.

- a) Product of odd function with odd function.
- b) Product of even function with odd function.
- c) Product of even function with even function.

## §2. POLYNOMIAL FUNCTIONS

A function  $f$  defined for all numbers is called a **polynomial** if there exists numbers  $a_0, a_1, \dots, a_n$  such that for all numbers  $x$  we have

$$(1) \quad f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0.$$

**Example.** The function  $f$  such that

$$f(x) = 3x^3 - 2x + 1$$

is a polynomial function. We have  $f(1) = 3 - 2 + 1 = 2$ .

**Example.** The function

$$g : x \mapsto \frac{1}{2}x^4 + 3x^2 - x + 5$$

is a polynomial. We have

$$g(2) = \frac{1}{2}2^4 + 3 \cdot 2^2 - 2 + 5 = 23.$$

When a polynomial can be written as in (1) above, we shall say that it is of **degree**  $\leq n$ . If  $a_n \neq 0$ , then we want to say that the polynomial has

degree  $n$ . However, we must be careful. Is it possible that there are other numbers  $b_0, \dots, b_m$  such that for all  $x$  we have

$$f(x) = b_m x^m + \cdots + b_0?$$

For instance, can we have

$$7x^5 - 5x^4 + 2x + 1 = x^6 - 17x^3 + x + 1$$

for all numbers  $x$ ? The answer is not immediately clear just by looking. Suppose that the coefficients were big complicated numbers. We would have no simple test for the equality of values. If the answer were YES, then it would be hopeless to define the degree, because in the example just written down, for instance, we would not know whether the degree is 5 or 6. That the answer is NO will be proved in the corollary to Theorem 2 below.

Let  $f$  be a polynomial. If  $c$  is a number such that  $f(c) = 0$ , then we call  $c$  a **root** of  $f$ . We shall see in a moment that a non-zero polynomial can have only a finite number of roots, and we shall give a bound for the number of these roots.

**Example.** Let  $f(x) = x^2 - 3x + 2$ . Then  $f(1) = 0$ . Hence 1 is a root of  $f$ . Also,  $f(2) = 0$ . Hence 2 is also a root of  $f$ .

**Example.** Let  $f(x) = ax^2 + bx + c$ . If  $b^2 - 4ac = 0$ , then the polynomial has one real root, which is

$$-\frac{b}{2a}.$$

If  $b^2 - 4ac > 0$ , then the polynomial has two distinct real roots which are

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

These assertions are merely reformulations of the theory of quadratic equations in Chapter 4.

**Theorem 1.** *Let  $f$  be a polynomial of degree  $\leq n$  and let  $c$  be a root. Then there exists a polynomial  $g$  of degree  $\leq n - 1$  such that for all numbers  $x$  we have*

$$f(x) = (x - c)g(x).$$

*Proof.* Write

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

Substitute the value

$$x = (x - c) + c$$

for  $x$ . Each  $k$ -th power, for  $k = 0, \dots, n$ , of the form

$$((x - c) + c)^k$$

can be expanded out as a sum of powers of  $(x - c)$  times a number. Hence there exist numbers  $b_0, b_1, \dots, b_n$  such that

$$f(x) = b_0 + b_1(x - c) + b_2(x - c)^2 + \cdots + b_n(x - c)^n,$$

for all  $x$ . But  $f(c) = 0$ . Hence

$$0 = f(c) = b_0,$$

and all the other terms on the right have the value 0 for  $x = c$ . This proves that  $b_0 = 0$ . But then we can factor

$$f(x) = (x - c)(b_1 + b_2(x - c) + \cdots + b_n(x - c)^{n-1}).$$

We let

$$g(x) = b_1 + b_2(x - c) + \cdots + b_n(x - c)^{n-1},$$

and we see that our theorem is proved.

**Remark.** Let us look more carefully at the polynomial  $g$  which we obtained in the proof. When we expand out a power

$$((x - c) + c)^k,$$

we get a term  $(x - c)^k$ , and all other terms in the sum involve a lower power of  $(x - c)$ . Hence the highest power of  $(x - c)$  which we get is the  $n$ -th power, and this power comes from the expansion of

$$((x - c) + c)^n.$$

Thus the term with the highest power of  $(x - c)$  can be determined explicitly, and it is precisely

$$a_n(x - c)^n.$$

In other words, we have

$$b_n = a_n.$$

Hence the polynomial  $g$  which we obtain has an expansion of the form

$$g(x) = a_n x^{n-1} + \text{lower terms.}$$

This remark will be useful later.

**Theorem 2.** *Let  $f$  be a polynomial. Let  $a_0, \dots, a_n$  be numbers such that  $a_n \neq 0$ , and such that we have*

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$$

*for all  $x$ . Then  $f$  has at most  $n$  roots.*

*Proof.* Let  $c_1, c_2, \dots, c_r$  be distinct roots of  $f$ . Suppose that  $r \geq n$ . Write

$$f(x) = (x - c_1)g_1(x)$$

for some polynomial  $g_1$  of degree  $\leq n - 1$ . Then

$$0 = f(c_2) = (c_2 - c_1)g_1(c_2).$$

Since  $c_2 \neq c_1$ , it follows that  $g_1(c_2) = 0$ ; in other words,  $c_2$  is a root of  $g_1$ , which has degree  $\leq n - 1$ . We can now give the same argument, factoring

$$g_1(x) = (x - c_2)g_2(x).$$

We see that

$$g_2(c_3) = 0,$$

and that  $g_2$  has degree  $\leq n - 2$ . We keep on going, until we find  $g_n$  to be constant. Thus we can write

$$f(x) = (x - c_1)(x - c_2) \cdots (x - c_n)c$$

for all  $x$ . In fact, the remark following Theorem 1 shows that

$$c = a_n \neq 0.$$

Thus if  $x \neq c_i$  for any  $i = 1, \dots, n$ , we see that  $f(x) \neq 0$ . This means that  $f$  has at most  $n$  roots, and our theorem is proved.

**Corollary.** *Let  $f$  be a polynomial, which can be written in the form*

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$$

*and also in the form*

$$f(x) = b_n x^n + b_{n-1} x^{n-1} + \cdots + b_0.$$

*Then*

$$a_i = b_i \quad \text{for every } i = 0, \dots, n.$$

*Proof.* Consider the polynomial

$$0 = f(x) - f(x) = (a_n - b_n)x^n + (a_{n-1} - b_{n-1})x^{n-1} + (a_0 - b_0).$$

We have to show that

$$a_i - b_i = 0 \quad \text{for all } i = 1, \dots, n.$$

Let  $d_i = a_i - b_i$ . Suppose that there exists some index  $i$  such that  $d_i \neq 0$ . Let  $m$  be the largest of these indices, so that we can write

$$0 = d_m x^m + \cdots + d_0$$

for all  $x$  and  $d_m \neq 0$ . This contradicts Theorem 2. Therefore we conclude that  $d_i = 0$  for all  $i = 1, \dots, n$ , thus proving our corollary.

The corollary shows that if  $f$  is a polynomial, then there is a unique way of writing  $f$  in the form

$$f(x) = a_n x^n + \cdots + a_0$$

for all  $x$ . In other words, the numbers  $a_n, \dots, a_0$  are uniquely determined. They will be called the **coefficients** of  $f$ , and we call  $a_n$  the **leading coefficient** if  $a_n \neq 0$ . We call  $a_0$  the **constant term**.

**Example.** Let  $f$  be the polynomial such that for all  $x$  we have

$$f(x) = 4x^5 - 7x^3 + x - 20.$$

Then the coefficients of  $f$  are 4, 0, 7, 0, 1,  $-20$ . (We have included the coefficients of all powers of  $x$  up to the fifth. Note that the fourth and second power have coefficients equal to 0.) The leading coefficient is 4. The constant term is  $-20$ .

If a polynomial  $f$  is such that

$$f(x) = a_n x^n + \cdots + a_0,$$

and  $a_n \neq 0$ , then we say that  $f$  has **degree**  $n$ . The polynomial in the preceding example has degree 5.

**Remark.** It often happens that if  $f$  is a polynomial, then

$$f(x) = 0$$

for some  $x$ . In other words, there may exist a number  $c$  such that

$$f(c) = 0.$$

This does not mean that  $f$  is the zero polynomial. By definition, we call  $f$  the **zero polynomial** if and only if

$$f(x) = 0 \quad \text{for all numbers } x.$$

Thus the zero polynomial is the polynomial all of whose coefficients are 0. If some coefficient of a polynomial is not equal to 0, then  $f$  is not the zero polynomial. If  $f$  is the zero polynomial, we also sometimes say that  $f$  is **identically zero** (to distinguish this case from the one in which  $f$  may take on the value 0 at some number).

In Chapter 4 we found a way of determining all roots of a polynomial of degree 2. For polynomials of higher degree, it is much more difficult to determine the roots, except in very special cases. For polynomials of degree 3 and 4, one can give formulas involving radicals, but it is a classical result that such formulas cannot be given in general for polynomials of degree at least 5.

For polynomials whose coefficients are integers, one may ask for the rational roots, and much time is often spent in elementary classes finding such roots by factoring the polynomial. It is very unusual that a polynomial can be factored with integral or rational roots, and I think much too much emphasis is placed on this kind of accident. One result of this kind of emphasis (which I have found among students) is that they try to factor even in the quadratic case, when the systematic answer is available from the formula. My experience indicates that such training in factoring is not worth the time spent, and therefore we do not emphasize it here. However, we give two important examples of factoring.

**Example.** Let  $f(x) = 3x^2 - 5x + 1$ . Then the roots of  $f$  are given by the quadratic formula of Chapter 4, namely

$$c_1 = \frac{5 + \sqrt{25 - 12}}{6} = \frac{5 + \sqrt{13}}{6}$$

$$c_2 = \frac{5 - \sqrt{25 - 12}}{6} = \frac{5 - \sqrt{13}}{6}.$$

If we look back either at Theorem 1 or at the beginning of the proof of Theorem 2, we see that

$$f(x) = 3 \left( x - \frac{5 + \sqrt{13}}{6} \right) \left( x - \frac{5 - \sqrt{13}}{6} \right).$$

We have therefore factored  $f$  into factors of degree 1.

**Example.** The general quadratic case follows the same pattern. Let

$$f(x) = ax^2 + bx + c,$$

and  $a \neq 0$ . Assume that  $b^2 - 4ac > 0$ . Then there are two distinct roots  $c_1$  and  $c_2$  of  $f$ , and therefore we have the factorization

$$f(x) = a(x - c_1)(x - c_2).$$

**Example.** The other important case of factoring occurs for the polynomial

$$f(x) = x^n - 1.$$

We see that 1 is a root, because  $f(1) = 1^n - 1 = 1 - 1 = 0$ . Hence we know that  $f$  must have  $(x - 1)$  as a factor, i.e. we have

$$f(x) = (x - 1)g(x)$$

for some polynomial  $g(x)$ . What is  $g(x)$ ? We suggest that you look at Exercise 2 of Chapter 1, §6, and do Exercise 6 of this chapter.

We shall now discuss what is sometimes called “long division”. For polynomials it is the analog of the division of two positive integers, with a remainder. Therefore we recall this division process for positive integers.

First, we carry out an example.

**Example.** We want to divide 327 by 17, with a possible remainder. As you know from elementary school, this division can be represented schematically as follows.

$$\begin{array}{r} 19 \\ 17 \overline{)327} \\ 17 \\ \hline 157 \\ 153 \\ \hline 4 \end{array}$$

This procedure tells us that

$$327 = 19 \cdot 17 + 4.$$

We call 4 the remainder. We can describe the preceding steps as follows. We determined the first digit 1 of 19 as being the largest positive integer whose product with 17 would still be  $\leq 32$ . We then multiply 1 by 17, write it under 32, subtract, get 15, and bring down the 7. We then repeat the process. We determine 9 as the largest positive integer whose product with 17 is  $\leq 157$ . We multiply 17 by 9, get 153, write it under 157, subtract, and get 4. This number 4 is now less than 17, so we stop.

We can summarize what we have done in general by the following statement.

*Let  $n, d$  be positive integers. Then there exists an integer  $r$  such that  $0 \leq r < d$ , and an integer  $q \geq 0$  such that*

$$n = qd + r.$$

Note that even though the standard procedure of the example, which gives us  $q, r$ , is called “long division”, in fact our procedure uses only multiplication and subtraction.

We shall now describe an analogous procedure for polynomials, which gives us the

**Euclidean algorithm.** *Let  $f$  and  $g$  be non-zero polynomials. Then there exist polynomials  $q, r$  such that  $\deg r < \deg g$  and such that*

$$f(x) = q(x)g(x) + r(x).$$

**Example.** Let

$$f(x) = 4x^3 - 3x^2 + x + 2 \quad \text{and} \quad g(x) = x^2 + 1.$$

We want to find  $q(x)$  and  $r(x)$ . We first lay out what we do in a diagram

similar to that of long division for integers, and then explain how each step is obtained.

$$\begin{array}{r} 4x - 3 \\ x^2 + 1 \overline{)4x^3 - 3x^2 + x + 2} \\ 4x^3 + \quad \quad \quad 4x \\ \hline -3x^2 - 3x + 2 \\ -3x^2 \quad \quad \quad -3 \\ \hline -3x + 5 \end{array}$$

We have

$$q(x) = 4x - 3 \quad \text{and} \quad r(x) = -3x + 5,$$

so that

$$(1) \quad 4x^3 - 3x^2 + x + 2 = (4x - 3)(x^2 + 1) + (-3x + 5).$$

Now we describe each step in the computation. We first determine  $4x$  because

$$4x \cdot x^2$$

is equal to the term of highest degree in  $f(x)$ ; that is, equal to  $4x^3$ . We then multiply  $4x$  by  $x^2 + 1$ , we obtain  $4x^3 + 4x$ , which we write under  $f(x)$ , placing corresponding powers of  $x$  under each other. We then subtract  $4x^3 + 4x$  from  $4x^3 - 3x^2 + x + 2$ , and obtain  $-3x^2 - 3x + 2$ . We then repeat our procedure, and determine  $-3$  because

$$(-3) \cdot x^2$$

is equal to the term of highest degree in  $-3x^2 - 3x + 2$ . We multiply  $-3$  by  $x^2 + 1$ , obtain  $-3x^2 - 3$ , which we write under  $-3x^2 - 3x + 2$ . We subtract, and obtain  $-3x + 5$ . We note that the polynomial  $-3x + 5$  has degree 1, which is smaller than the degree of  $g(x) = x^2 + 2$ . Our computation is therefore finished.

Why does the above procedure actually provide us with polynomials  $q(x)$  and  $r(x)$  satisfying relation (1)? This is easily seen. Write  $f(x) = f_3(x)$  to indicate the fact that  $f$  has degree 3. We determined  $4x$  in such a way that

$$f_3(x) - 4x(x^2 + 1) = f_3(x) - 4x^3 - 4x$$

has degree 2; i.e. in such a way that the term  $4x^3$  would cancel. Write

$$f_2(x) = f_3(x) - 4x^3 - 4x = -3x^2 - 3x + 2.$$

Then  $f_2$  has degree 2. We determine  $-3$  so that

$$f_2(x) - (-3)(x^2 + 1)$$

has degree 1; i.e. in such a way that the term  $-3x^2$  cancels. Then

$$f_2(x) - (-3)(x^2 + 1) = f_1(x)$$

has degree 1. Now we see that

$$f_3(x) - 4x \cdot g(x) - (-3) \cdot g(x) = -3x + 5$$

has degree 1. Thus

$$f_3(x) - (4x - 3)g(x) = -3x + 5.$$

This shows why

$$f_3(x) = q(x)g(x) + r(x).$$

**Example.** Let

$$f(x) = 2x^4 - 3x^2 + 1 \quad \text{and} \quad g(x) = x^2 - x + 3.$$

We wish to find  $q(x)$  and  $r(x)$  as in the Euclidean algorithm. We write down our computation in the standard pattern.

$$\begin{array}{r} 2x^2 + 2x - 7 \\ x^2 - x + 3 \overline{)2x^4 - 3x^2 + 1} \\ 2x^4 - 2x^3 + 6x^2 \\ \hline +2x^3 - 9x^2 + 1 \\ 2x^3 - 2x^2 + 6x \\ \hline -7x^2 - 6x + 1 \\ -7x^2 + 7x - 21 \\ \hline -13x + 22 \end{array}$$

Hence we get

$$q(x) = 2x^2 + 2x - 7 \quad \text{and} \quad r(x) = -13x + 22.$$

As a matter of terminology, we call  $r(x)$  the **remainder** in the Euclidean algorithm.

**Remark.** The Euclidean algorithm allows us once more to prove that if  $f(x)$  has a root  $c$ , then we can write

$$f(x) = (x - c)q(x)$$

for some polynomial  $q$ . Indeed, in the Euclidean algorithm, we have

$$f(x) = q(x)(x - c) + r(x),$$

where  $\deg r < 1$ . Hence  $r$  must be constant, say equal to a number  $a$ . Thus

$$f(x) = q(x)(x - c) + a.$$

Now evaluate the left-hand side and the right-hand side at  $x = c$ . We get

$$0 = 0 + a,$$

whence  $a = 0$ . The remainder is equal to 0, and this gives what we wanted.

We do not prove the Euclidean algorithm. The proof would consist of carrying out the procedure of the example with general coefficients.

## EXERCISES

1. What is the degree of the following polynomials?

- a)  $3x^2 - 4x + 5$
- b)  $-5x^5 + x$
- c)  $-38x^4 + x^3 - x - 1$
- d)  $(3x^2 - 4x + 5)(-5x^5 + x)$
- e)  $(-5x^5 + x)(-7x + 3)$
- f)  $(-4x^2 + 5x - 4)(3x^3 + x - 1)$
- g)  $(6x^7 - x^3 + 5)(7x^4 - 3x^2 + x - 1)$
- h) Let  $f, g$  be polynomials which are not the zero polynomials. Show that

$$\deg(fg) = \deg f + \deg g.$$

2. Factor the quadratic polynomials of the exercises in Chapter 4 into factors of degree 1.
3. Let  $f$  be a polynomial of degree 3. If there exist polynomials  $g, h$  of degree  $\geq 1$  such that  $f = gh$ , show that  $f$  has a root.
4. a) Give an example of a polynomial of degree 2 which has no root in the real numbers.  
 b) Give an example of a polynomial of degree 3 which has only one root in the real numbers.  
 c) Give an example of a polynomial of degree 4 which has no root in the real numbers.

5. Let

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

be a polynomial whose coefficients are integers and whose leading coefficient is 1. If  $c$  is an integer and is a root of  $f$ , show that  $c$  divides  $a_0$ .

6. What are all the roots in the real numbers of the following polynomials?

- |              |              |              |              |
|--------------|--------------|--------------|--------------|
| a) $x^3 - 1$ | b) $x^4 - 1$ | c) $x^5 - 1$ | d) $x^n - 1$ |
| e) $x^3 + 1$ | f) $x^4 + 1$ | g) $x^5 + 1$ | h) $x^n + 1$ |

7. Find the polynomials  $q(x)$  and  $r(x)$  of the Euclidean algorithm when  $f(x) = 4x^3 - x + 2$ , and:

- |                           |                           |
|---------------------------|---------------------------|
| a) $g(x) = x - 2$ ,       | b) $g(x) = x^2 - 1$ ,     |
| c) $g(x) = x^2 + 1$ ,     | d) $g(x) = x^2 - x$ ,     |
| e) $g(x) = x^2 - x + 1$ , | f) $g(x) = x^2 + x - 1$ , |
| g) $g(x) = x^3 + 2$ ,     | h) $g(x) = x^3 - x + 1$ . |

8. Repeat Exercise 7 for the case in which  $f(x) = 6x^4 - x^3 + x^2 - 2x + 5$  and:

- |                        |                             |
|------------------------|-----------------------------|
| a) $g(x) = x^3 - 1$ ,  | b) $g(x) = x^2 - 5$ ,       |
| c) $g(x) = x + 2$ ,    | d) $g(x) = 3x + 1$ ,        |
| e) $g(x) = 4x + 6$ ,   | f) $g(x) = x^4 - x^2 + 1$ , |
| g) $g(x) = x^3 - 5x$ , | h) $g(x) = x^3 - 2x^2$ .    |

9. **Rational functions.** A rational function is a function which can be expressed as a quotient of polynomials, i.e. a function whose value at a number  $x$  is

$$\frac{f(x)}{g(x)},$$

where  $f, g$  are polynomials, and  $g$  is not the zero polynomial. Thus a rational function is defined only for those numbers  $x$  such that  $g(x) \neq 0$ . In an expression as above, we call  $f$  the numerator of the rational function, and  $g$  its denominator. We can then work with rational functions just as we did with rational numbers. In particular, we can put two rational functions over a common (polynomial) denominator, and take their sum in a manner analogous to taking the sum of rational numbers. We give an example of this.

**Example.** Put the two rational functions

$$R(x) = \frac{3x^2 - 2x + 6}{x - 1} \quad \text{and} \quad S(x) = \frac{4x + 3}{x - 5}$$

over a common denominator.

This denominator will be  $(x - 1)(x - 5)$ , and the two rational functions over this common denominator are

$$R(x) = \frac{(3x^2 - 2x + 6)(x - 5)}{(x - 1)(x - 5)} \quad \text{and} \quad S(x) = \frac{(4x + 3)(x - 1)}{(x - 1)(x - 5)}.$$

Their sum is then

$$R(x) + S(x) = \frac{(3x^2 - 2x + 6)(x - 5) + (4x + 3)(x - 1)}{(x - 1)(x - 5)},$$

which we see is again a rational function, defined for all numbers  $\neq 1$  or 5. We can expand the expressions in the numerator and denominator into the standard form for polynomials, and we see that

$$R(x) + S(x) = \frac{3x^3 - 13x^2 + 15x - 33}{x^2 - 6x + 5}.$$

Now you do the work, and express the sums of the following rational functions  $R(x) + S(x)$  as quotients of polynomials just as we did in this example.

a)  $R(x) = \frac{x - 4}{x + 3}$  and  $S(x) = \frac{2x + 1}{x - 5}$

b)  $R(x) = \frac{3x - 1}{2x + 2}$  and  $S(x) = \frac{x - 4}{3x + 2}$

c)  $R(x) = \frac{x^2 - 1}{x + 5}$  and  $S(x) = \frac{3x^3 + 2}{x + 1}$

d)  $R(x) = \frac{x^2 - x + 1}{3x - 4}$  and  $S(x) = \frac{x + 3}{x^2 + 2}$

e)  $R(x) = \frac{x^3 + 1}{x + 4}$  and  $S(x) = \frac{x^4 - 2}{3x + 1}$

f)  $R(x) = \frac{x^4 - 1}{x}$  and  $S(x) = \frac{x - 1}{x^2 + 2}$

g)  $R(x) = \frac{x^3 - 2x}{x^2}$  and  $S(x) = \frac{x^2 + 1}{x - 1}$

h)  $R(x) = \frac{2x^3 - 1}{x^2 + 2}$  and  $S(x) = \frac{x - 1}{x + 1}$

### §3. GRAPHS OF FUNCTIONS

Let  $f$  be a function, defined on a set of numbers  $S$ . By the **graph** of the function  $f$ , we shall mean the set of all points

$$(x, f(x)),$$

i.e. the set of all points whose first coordinate is  $x$ , and whose second coordinate is  $f(x)$ .

**Example.** Let  $f(x) = x^2$ . Then the graph of  $f$  is the graph of the equation  $y = x^2$ , and is a parabola, as discussed in Chapter 7.

**Example.** An object moves along the positive  $x$ -axis, subjected to a force inversely proportional to the square of the distance from the origin. We can then write this force as

$$y = \frac{k}{x^2},$$

where  $k$  is some constant. Suppose that  $k = 3$ , so that  $y = 3/x^2$ . It is easy to draw the graph of this equation, which is nothing but the graph of the function  $f$  such that  $f(x) = 3/x^2$ ; see Fig. 13-1. We make a table of values, and observe that as  $x$  increases,  $f(x)$  decreases.

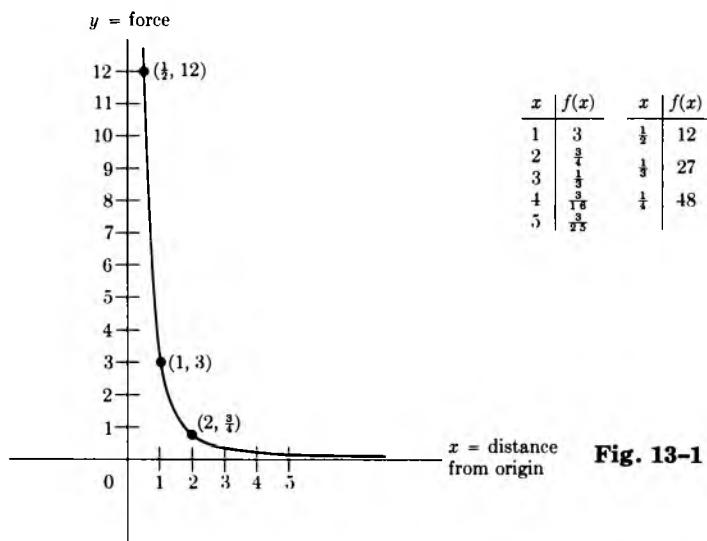


Fig. 13-1

**Example.** We know that the sine is a function, associating with each number  $\theta$  the value  $\sin \theta$ . In fact, in the chapter on trigonometry, we had already drawn its graph which looked like this (Fig. 13-2.)

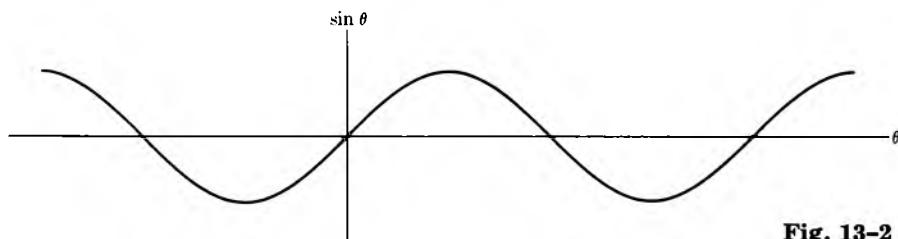


Fig. 13-2

**Example.** We want to sketch the graph of the function

$$y = (x - 1)(x - 2).$$

Observe that when  $x = 1$  or when  $x = 2$ , then  $y = 0$ . Furthermore, there are no other values of  $x$  for which  $y = 0$ . Thus the graph crosses the  $x$ -axis precisely at the points  $x = 1$  and  $x = 2$ . When  $x < 1$ , then  $x - 1 < 0$  and  $x - 2 < 0$ . Hence when  $x < 1$ , we see that  $y > 0$ . Similarly, when  $x > 2$ , we see that  $y > 0$ . Finally, when

$$1 < x < 2,$$

then  $x - 1 > 0$  and  $x - 2 < 0$ . Hence when  $1 < x < 2$ , we see that  $y$  is negative. Finally, when  $x = 0$ , we see that

$$y = (-1)(-2) = 2.$$

The graph of our function looks like this (Fig. 13-3):

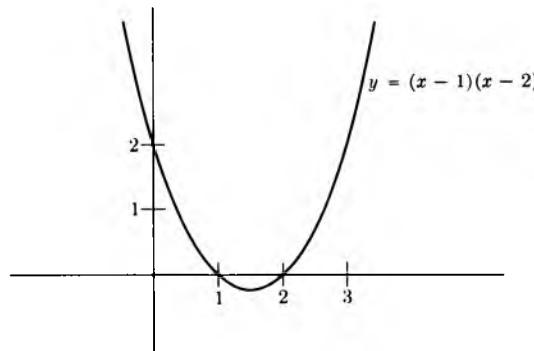


Fig. 13-3

This graph is of course the graph of a parabola, since it is the graph of

$$y = x^2 - 3x + 2.$$

We can complete the square, and use the same method as in Chapter 12 to sketch the graph. However, the method used here gives us a quicker insight into the rough way the graph looks.

**Example.** Let  $[x]$  be the largest integer  $\leq x$ . Let  $f(x) = [x]$ . Then  $f(2) = 2$  and  $f(\frac{3}{2}) = 1$ . The graph of  $f$  is shown in Fig. 13-4. It looks like a staircase.

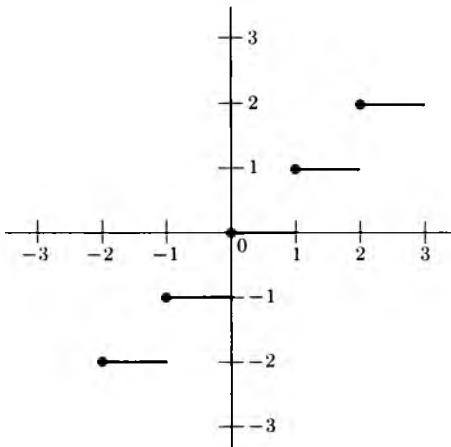


Fig. 13-4

## EXERCISES

Sketch the graphs of the following functions. Make small tables of values to get an idea of what's happening.

- |                           |                           |
|---------------------------|---------------------------|
| 1. $f(x) = 1/x^3$         | 2. $f(x) = -1/x^3$        |
| 3. $f(x) = -1/x^2$        | 4. $f(x) = -1/(x - 2)^2$  |
| 5. $f(x) = \sin(x - \pi)$ | 6. $f(x) = \cos(x - \pi)$ |

- |  |  |
|--|--|
| 7. $f(x) = \sin\left(x - \frac{\pi}{4}\right)$ | 8. $f(x) = \cos\left(x - \frac{\pi}{4}\right)$ |
| 9. $f(x) =  x $                                | 10. $f(x) = x +  x $                           |
| 11. $f(x) = x -  x $                           | 12. $f(x) = 2 +  x $                           |
| 13. $f(x) =  x  - x$                           | 14. $f(x) =  x - 1 $                           |
| 15. $f(x) =  x + 1 $                           | 16. $f(x) =  x - 2 $                           |
| 17. $f(x) =  x + 2 $                           | 18. $f(x) =  x + 3 $                           |
| 19. $f(x) = (x + 1)(x - 3)$                    | 20. $f(x) = (x - 2)(x - 5)$                    |
| 21. $f(x) = (x + 2)(x - 1)$                    | 22. $f(x) = (x + 3)x$                          |
| 23. $f(x) = (x + 2)(x + 1)$                    | 24. $f(x) = (x + 4)(x - 3)$                    |
| 25. $f(x) = [x] + 1$                           | 26. $f(x) = [x + 1]$                           |
| 27. $f(x) = [x + 2]$                           | 28. $f(x) = [x] - x$                           |
| 29. $f(x) = x - [x]$                           | 30. $f(x) = [x] + x$                           |

#### §4. EXPONENTIAL FUNCTION

We have already discussed powers like

$$a^{m/n},$$

where  $a$  is a positive number, and  $m, n$  are integers, i.e. fractional powers.

It is difficult to give an analytic development of the theory of powers

$$a^x$$

when  $x$  is not a fraction, but at least we can state the basic properties, which are intuitively clear, and then use them in applications.

**Let  $a$  be a number  $> 0$ .** To each number  $x$  we can associate a number denoted by  $a^x$ , such that when  $x = m/n$  is a quotient of integers ( $n \neq 0$ ) then  $a^{m/n}$  is the ordinary fractional power discussed in Chapter 3, §3, and such that the function

$$x \mapsto a^x$$

has the following properties.

**EXP 1.** *For any numbers  $x, y$ , we have*

$$a^{x+y} = a^x a^y.$$

**EXP 2.** *For all numbers  $x, y$ , we have*

$$(a^x)^y = a^{xy}.$$

**EXP 3.** *If  $a, b$  are positive, then*

$$(ab)^x = a^x b^x.$$

**EXP 4.** *Assume that  $a > 1$ . If  $x < y$ , then  $a^x < a^y$ .*

The function

$$x \mapsto a^x$$

is called the **exponential function to the base  $a$** . The proof given in Chapter 3, §3 that  $a_0 = 1$  is valid here. From this and **EXP 1** we conclude that

$$a^{-x} = \frac{1}{a^x}.$$

This is because

$$a^{-x} a^x = a^{x-x} = a^0 = 1.$$

**Remark.** The values of the exponential function are always positive. In fact, if  $x > 0$ , and  $a > 1$ , then  $a^x > 1$  because  $a^0 = 1$  and the exponential function is increasing according to **EXP 4**. If  $x < 0$ , say  $x = -z$  where  $z$  is positive, then

$$a^x = a^{-z} = \frac{1}{a^z},$$

and we see again that  $a^x$  is positive. However, in this case, we have  $a^x < 1$ .

**Question.** If we allow  $0 < a < 1$  instead of  $a > 1$ , how do you have to adjust property EXP 4 to make it valid?

It is now easy to sketch the graph of an exponential function.

**Example.** Let  $f(x) = 2^x$ . We make a small table of values, with rational numbers and integers for  $x$ .

$x$	$f(x)$	$x$	$f(x)$
0	1	-1	$\frac{1}{2}$
1	2	-2	$\frac{1}{4}$
2	4	-3	$\frac{1}{8}$
3	8	-4	$\frac{1}{16}$
4	16	-5	$\frac{1}{32}$
5	32		

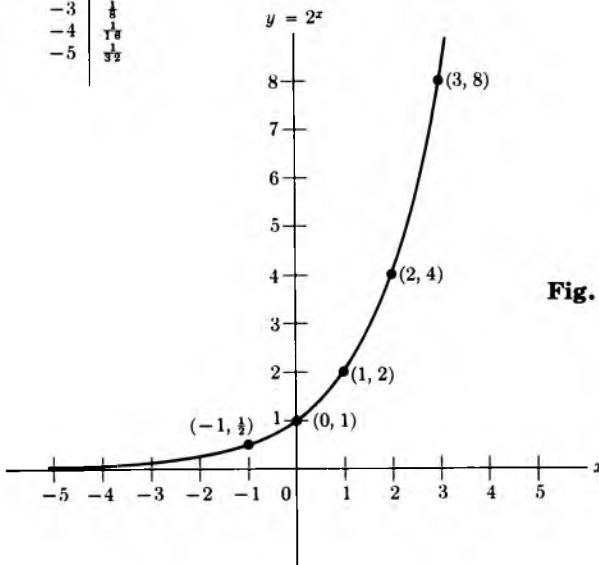


Fig. 13-5

Note that the graph climbs very steeply to the right of 0, and that it becomes very flat very fast to the left of 0.

**Example.** The population of a city doubles every year, and at time  $t = 0$  it is equal to 100 persons. We can then express this population in the form

$$P(t) = 100 \cdot 2^t,$$

when  $t$  is an integer. Namely, when  $t = 0$ , we get

$$P(0) = 100.$$

Each time that  $t$  increases by 1, we see that  $P(t)$  is multiplied by 2. Thus the population  $P$  is a function of time  $t$ , and again we draw the graph by making a table of values as shown in Fig. 13-6, and noticing that  $P(t)$  increases when  $t$  increases.

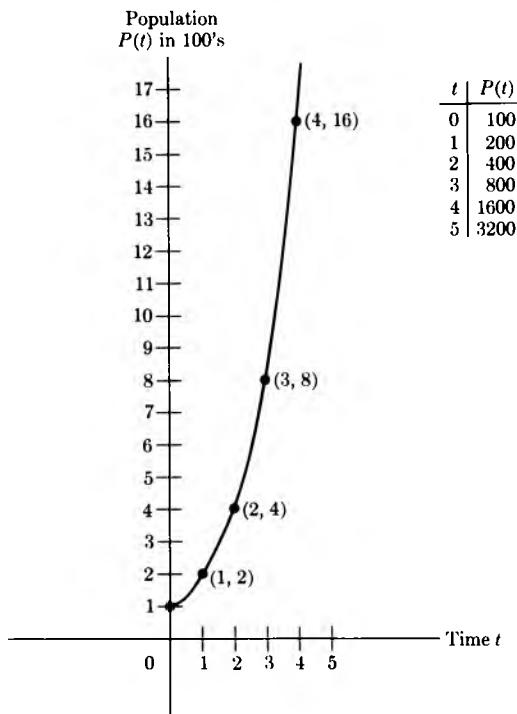


Fig. 13-6

**Example.** Let  $f$  be a function such that  $f(t) = Ca^{5t}$ , where  $C$  is a number. Such a function is said to have exponential growth. Note that when  $t = 0$ , we get

$$f(0) = C,$$

so that  $C$  is the value of  $f$  at  $t = 0$ . If  $f$  represents a population, given as a function of time, then  $C$  is the original number of persons at time  $t = 0$ .

**Example.** Certain substances disintegrate at a rate proportional to the amount of substance present. If  $f$  denotes this amount, then it is known that  $f$  as a function of time  $t$  is given by the formula

$$f(t) = Ca^{Kt},$$

where  $C, K$  are constants. Again note that  $C$  is the amount of substance

present at time  $t = 0$ . Since the amount of substance decreases, the constant  $K$  is negative. For instance, if

$$f(t) = 3a^{-2t},$$

if  $a = 2$ , and if  $f$  gives the number of grams of radium at a certain place as a function of time, then there are 3 g of radium present at the beginning when  $t = 0$ . If the time is measured in years, then after 4 years, there are

$$f(4) = 3 \cdot a^{-8} = \frac{3}{256}$$

grams of radium left.

## EXERCISES

1. Sketch the graph of the function  $f$  such that:
  - a)  $f(t) = 3^t$ ,
  - b)  $f(t) = 4^t$
  - c)  $f(t) = 5^t$
2. Sketch the graph of the function  $f$  such that:
  - a)  $f(t) = 2^{-t}$
  - b)  $f(t) = 3^{-t}$
  - c)  $f(t) = 5^{-t}$
3. If  $1 < a < b$ , which is steeper: the graph of  $a^x$  or the graph of  $b^x$ ?
4. Let  $a = 5$ . At time  $t = 0$ , there are 4 g of radium in a cave. The amount of radium given as a function of time is

$$f(t) = 5a^{-3t}.$$

How much radium is left

- a) at  $t = 1$ ?      b) at  $t = 2$ ?      c) at  $t = 3$ ?
5. The population of a city triples every 50 years. At time  $t = 0$ , this population is 100,000. Give a formula for the population  $P(t)$  as a function of  $t$ . What is the population after
  - a) 100 years?
  - b) 150 years?
  - c) 200 years?
6. Bacteria in a solution double every 3 min. If there are  $10^4$  bacteria at the beginning, give a formula for the number of bacteria at time  $t$ . How many bacteria are there after
  - a) 3 min?
  - b) 9 min?
  - c) 27 min?
  - d) one hour?

7. The function  $f(t) = 4 \cdot 16^t$  describes the growth of bacteria.
- How many bacteria are present at the beginning, when  $t = 0$ ?
  - After  $\frac{1}{2}$  hr, how many bacteria are there?
  - Same question, after  $\frac{1}{4}$  hr.
  - Same question after 1 hr.
8. A radioactive element decays so that the amount  $f(t)$  left at time  $t$  satisfies the formula

$$f(t) = 60 \cdot 2^{-0.02t}.$$

- What is the initial quantity of this element at  $t = 0$ ?
- How much is left after 500 years?
- How much is left after 1,000 years?
- How much is left after 2,000 years?

## §5. LOGARITHMS

**Let  $a$  be a number  $> 1$ .** If  $y = a^x$ , then we shall say that

$x$  is the log of  $y$  to the base  $a$ , and write  $x = \log_a y$ .

**Example.** Since  $8 = 2^3$ , we see that 3 is the log of 8 to the base 2. We write  $3 = \log_2 8$ .

**Example.** Since  $27 = 3^3$ , we see that 3 is the log of 27 to the base 3. We write  $3 = \log_3 27$ .

The log is a function, defined for all positive numbers. We shall not prove this, but assume that it is a basic property of numbers. In other words, we assume that:

*Given a number  $y > 0$ , there exists a number  $x$  such that*

$$a^x = y.$$

We can now prove properties of the log which are similar to those of the exponential function.

**LOG 1.** *For any numbers  $x, y$ , we have*

$$\log_a(xy) = \log_a x + \log_a y.$$

**LOG 2.** *We have*

$$\log_a 1 = 0.$$

**LOG 3.** *If  $x < y$ , then  $\log_a x < \log_a y$ .*

These properties can be *proved* from the corresponding properties of the exponential function. We now do this.

*Proof of LOG 1.* Let  $u = \log_a x$  and  $v = \log_a y$ . This means that

$$x = a^u \quad \text{and} \quad y = a^v.$$

Hence

$$xy = a^u a^v = a^{u+v}.$$

By definition,

$$\log_a(xy) = u + v = \log_a x + \log_a y.$$

*Proof of LOG 2.* By definition, since  $a^0 = 1$ , this means that

$$0 = \log_a 1.$$

*Proof of LOG 3.* Let  $x < y$ . Let  $u = \log_a x$  and  $v = \log_a y$ . Then

$$a^u = x \quad \text{and} \quad a^v = y.$$

If  $u = v$ , then  $x = y$ , which is impossible. If  $v < u$ , then by EXP 4 we find  $y < x$ , which is also impossible. Hence we must have  $u < v$ , thereby proving our property **LOG 3**.

It is easy to draw a graph for the log. We leave this as an exercise. We also let you go through the properties of the log when  $a < 1$ , but  $a > 0$ . Which ones are still valid, and which type should be changed?

We can use the log to solve equations of the exponential type.

**Example.** Let  $f(t) = 10 \cdot 2^{kt}$ , where  $k$  is constant. Suppose that  $f(\frac{1}{2}) = 3$ . Find  $k$ .

We have

$$3 = 10 \cdot 2^{k/2}.$$

Taking the log to the base 2 on both sides, we find that

$$\log_2 3 = \log_2 10 + \frac{k}{2}.$$

Hence

$$k = 2(\log_2 3 - \log_2 10).$$

This argument can be used in a laboratory. For instance, suppose that we leave 10 g of a substance to decompose at time  $t = 0$ . We know that the amount of substance is given by the formula in the example above, with an unspecified constant  $k$ , and we wish to determine the value of  $k$ . If after half an hour we have 3 g of the substance left, then we get our value of  $k$  as in the example. Tables of logarithms or a computing machine then give us a decimal value for  $k$ .

**Example.** A radioactive substance disintegrates according to the formula

$$r(t) = C3^{-5t},$$

where  $C$  is a constant  $\neq 0$ . At what time will there be exactly one-third of the original amount left?

At time  $t = 0$  we have  $r(0) = C$  amount of substance. We must find that value of  $t$  such that

$$r(t) = \frac{1}{3}C,$$

or in other words,

$$\frac{1}{3}C = C3^{-5t}.$$

Note that we can cancel  $C$ .

Take the log to the base 3. Then the previous equation is equivalent with

$$\log_3\left(\frac{1}{3}\right) = -5t,$$

or equivalently,

$$-1 = -5t.$$

Thus

$$t = \frac{1}{5}.$$

Observe how in this example the unspecified constant  $C$  does not appear in the final answer.

## EXERCISES

1. Sketch the graph of the function  $g$  such that:

a)  $g(x) = \log_2 x$       b)  $g(x) = \log_3 x.$

Make a table of values to help you draw these graphs. For instance, in (a) use  $x = 2, 4, 8, 16, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}.$

2. Find the following values.

a)  $\log_2 64$       b)  $\log_3 \left(\frac{1}{27}\right)$

c)  $\log_5 25$       d)  $\log_5 \left(\frac{1}{25}\right)$

e)  $\log_2 \left(\frac{1}{64}\right)$       f)  $\log_3 \left(\frac{1}{81}\right)$

3. Let  $e$  be a fixed number  $> 1$  and abbreviate  $\log_e$  by  $\log$ . If  $a$  is  $> 1$  and  $x$  is an arbitrary number, prove that

$$\log a^x = x \log a.$$

[Hint: Consider  $e^{(\log a)x}$ , and use EXP 2.] For instance,

$$\log 10^{2/3} = \frac{2}{3} \log 10.$$

For the next exercises, assume that the number  $e$  has been fixed, and that  $\log = \log_e$ . Thus your answers will be expressed in terms of  $\log_e$ . There are tables for various values of  $e$  which can be used if you want decimal answers, but we are not concerned with this here. Remember that  $\log_e e = 1$ . Among all numbers  $e$  which we may take as the base for logarithms, there is one which is the most useful, and you will understand why when you take calculus. For our purposes here, we want you mainly to learn to operate with the formalism of the logarithm. However, work out the next computations. Suppose that  $e$  is chosen so that  $\log 2 = 0.6$ . Find a decimal for:

a)  $\log 2^{3/5}$       b)  $\log 2^{5/2}$       c)  $\log 2^{1/6}$

d)  $\log 2^{2/3}$       e)  $\log 8$       f)  $\log \frac{1}{2}$

g)  $\log \frac{1}{16}$

4. Let  $f(t) = Ce^{2t}$ . Suppose that you know that  $f(2) = 5$ . Determine the constant  $C$ .

5. Radium disintegrates according to the formula

$$f(t) = Ce^{-5t},$$

where  $C$  is a constant. At what time will there be exactly one-half of the original amount left?

6. Bacteria increase according to the formula

$$B(t) = Ce^{kt},$$

where  $C$  and  $k$  are constants, and  $B(t)$  gives the number of bacteria as a function of time  $t$  in min. At time  $t = 0$ , there are  $10^6$  bacteria. How long will it take before they increase to  $10^7$  if it takes 12 min to increase to  $2 \times 10^6$ ?

7. A radioactive substance disintegrates according to the formula

$$r(t) = Ce^{-7t},$$

where  $C$  is a constant. At what time will there be exactly one-third of the original amount left?

8. A substance decomposes according to the formula

$$S(t) = Ce^{-kt},$$

where  $C, k$  are constants. At the end of 3 min, 10% of the original substance has decomposed. When will one-half of the original amount have decomposed?

9. In 1900 the population of a city was 50,000. In 1950, it was 100,000. Assume that the population as a function of time is given by the formula

$$P(t) = Ce^{kt},$$

where  $C, k$  are constants. What will be the population in 1984? In what year will it be 200,000?

10. The atmospheric pressure as a function of height is given by the formula

$$p = Ce^{-kh},$$

where  $C, k$  are constants,  $p$  is the pressure, and  $h$  is the height. If the barometer reads 30 at sea level, and 24 at 6,000 ft above sea level, find the barometric reading 10,000 ft above sea level.

11. Sugar in water decomposes according to the formula

$$S = Ce^{-kt},$$

where  $C, k$  are constants. If 30 lb of sugar reduces to 10 lb in 4 hr, when will 95% of the sugar be decomposed?

12. A particle moves with speed given by

$$s(t) = Ce^{-kt},$$

where  $C, k$  are constants. If the initial speed at  $t = 0$  is 16 units/min, and if the speed is halved in 2 min, find the value of  $t$  when the speed is 10 units/min.

13. Assume that the difference  $d$  between the temperature of a body and that of surrounding air is given by the formula

$$d(t) = Ce^{-kt},$$

where  $C, k$  are constants. Let  $d = 100^\circ$ , when  $t = 0$ , and  $d = 40^\circ$  when  $t = 40$  min. Find  $t$ :

- a) when  $d = 70^\circ$ ,
- b) when  $d = 16^\circ$ .
- c) Find the value of  $d$  when  $t = 20$  min.

14. In 1800 the population of a city was 100,000. In 1900 it was 500,000. Assume that the population as a function of time is given by the formula

$$P(t) = Ce^{kt},$$

where  $C, k$  are constants. What will be the population in the year 2,000? In what year will it be 1,000,000?



# 14 Mappings

## §1. DEFINITION

We note that a function is an association. We have already seen other types of associations, namely mappings of the plane into itself. It is therefore convenient to define a general notion which covers both cases, and any other case like them.

We have also seen functions which are not defined for all numbers, and for which it was necessary to specify those numbers where it is defined.

Similarly, a function like  $x \mapsto x^2$  does not take on all real numbers as values, only numbers  $\geq 0$ . Thus it is convenient to specify which values a function might take. We therefore make the following definition.

Let  $S, S'$  be sets. A **mapping** from  $S$  into  $S'$  is an association

$$f: S \rightarrow S'$$

which to each element  $x$  in  $S$  associates an element  $f(x)$  of  $S'$ . We call  $f(x)$  the **value** of the mapping  $f$  at  $x$ , or the **image** of  $x$  under  $f$ .

Functions are special cases of mappings, and so are mappings of the plane into itself. In the latter case,  $S = \mathbf{R}^2 = S'$ . We have other examples. Disregard those for which you have not read the corresponding section in the book.

**Example.** Let  $P = (1, 2)$  and  $A = (-3, 5)$ . The association

$$t \mapsto P + tA = (1 - 3t, 2 + 5t)$$

is a mapping from the real numbers into the plane. It is our old parametrization of a line.

**Example.** The association

$$\theta \mapsto (\cos \theta, \sin \theta), \quad \theta \text{ in } \mathbf{R},$$

is a mapping from the real numbers into the plane. In fact, it is a mapping into the circle of radius 1 centered at the origin.

If  $f: S \rightarrow S'$  is a mapping, and  $T$  is a subset of  $S$ , then the set of all values  $f(t)$  with  $t$  in  $T$  is called the **image** of  $T$  under  $f$ , and is denoted by  $f(T)$ .

**Example.** The image of the mapping

$$\theta \mapsto (\cos \theta, \sin \theta)$$

is the circle of radius 1 centered at the origin.

**Example.** The image of the mapping

$$x \mapsto (x, x^2), \quad x \text{ in } \mathbf{R},$$

is the graph of the function  $x \mapsto x^2$ , and is a parabola.

**Example.** Let  $a, b$  be positive numbers. Let

$$F_{a,b}: \mathbf{R}^2 \rightarrow \mathbf{R}^2$$

be the mapping such that

$$F_{a,b}(x, y) = (ax, by).$$

The image of the circle of radius 1 centered at the origin is an ellipse.

**Example.** Consider the mapping

$$f: \mathbf{R} \rightarrow \mathbf{R}^2$$

given by

$$f(t) = (t^2, t^3).$$

We view this as a curve, whose coordinates are given as a function of  $t$ . Thus we also write

$$x(t) = t^2 \quad \text{and} \quad y(t) = t^3$$

to denote the dependence of the coordinates on  $t$ . We view such a mapping as a parametrization of a curve. We can draw the image of this mapping in the plane. We make a small table of values as usual.

$t$	$x(t)$	$y(t)$	$t$	$x(t)$	$y(t)$
1	1	1	-1	1	-1
2	4	8	-2	4	-8
3	9	27	-3	9	-27

For any  $t$ , the value  $x(t)$  is  $\geq 0$ , and increases with  $t$ . When  $t$  is positive,  $y(t)$  is positive. When  $t$  is negative,  $y(t)$  is negative. Thus we see in Fig. 14-1 what the image of the mapping looks like.

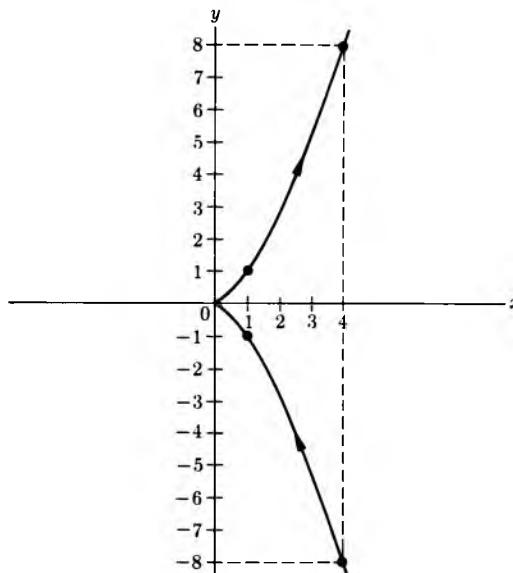


Fig. 14-1

**Example.** Suppose that a stone is thrown from a tall building, in the horizontal direction. Then gravity pulls the stone down. We want to give the coordinates  $(x(t), y(t))$  of the stone as a function of time  $t$ . Let the building be 50 ft tall. Then

$$x(t) = ct,$$

where  $c$  is some positive constant depending on the strength of the throw. The action of gravity determines the second coordinate to be

$$y(t) = 50 - \frac{1}{2}gt^2$$

where  $g$  is constant. Thus we have a mapping

$$t \mapsto \left( ct, 50 - \frac{1}{2}gt^2 \right),$$

defined for  $t \geq 0$  and for  $t \leq t_0$ , where  $t_0$  is the time at which the stone strikes the ground. See the picture in Fig. 14-2.

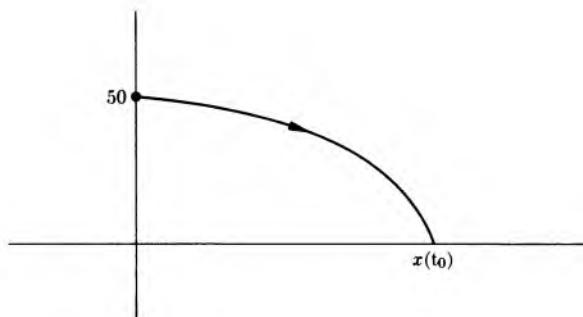


Fig. 14-2

Our mapping associates with each time  $t$  in the given range the position of the stone at time  $t$ . If we want to find the time at which the stone hits the ground, we have to find the time  $t$  such that

$$50 - \frac{1}{2}gt^2 = 0,$$

or in other words,

$$\frac{1}{2}gt^2 = 50.$$

We obtain

$$t = \sqrt{\frac{100}{g}}.$$

If the distance is measured in feet, and the time in seconds, then the value of  $g$  is 32. Hence in this case our value for  $t$  is

$$\sqrt{\frac{100}{32}} = \sqrt{\frac{25}{8}} \text{ sec.}$$

**Example.** Let  $S = \mathbf{R}^2$ . Then the association

$$X \mapsto d(X, O)$$

which to each point  $X$  associates its distance from the origin is a mapping, whose values are real numbers  $\geq 0$ .

In general, we shall follow the convention that a mapping from a set  $S$  into the real numbers is called a **function**. Thus

$$X \mapsto d(X, O)$$

is called the **distance function**. Of course, given a point  $P$ , we obtain another function

$$X \mapsto d(X, P),$$

whose value at  $X$  is the distance between  $X$  and  $P$ .

**Example.** The mapping

$$(x, y) \mapsto x^2 + \cos(xy)$$

is a function, defined on  $\mathbf{R}^2$ .

**Example.** At a given time  $t$ , we let  $f(t)$  be the temperature of a certain body. Then  $f$  is a function of time.

## EXERCISES

1. Draw the image of the mapping  $\mathbf{R} \rightarrow \mathbf{R}^2$  given by:

- a)  $\theta \mapsto (\cos 2\theta, \sin 2\theta)$ ,
- b)  $\theta \mapsto (5 \cos \theta, 5 \sin \theta)$ ,
- c)  $\theta \mapsto (2 \cos \theta, 3 \sin \theta)$ ,
- d)  $\theta \mapsto (4 \cos \theta, 5 \sin \theta)$ ,
- e)  $\theta \mapsto (a \cos \theta, b \sin \theta)$  if  $a, b$  are positive numbers.

2. Draw the image of the following mappings of  $\mathbf{R}$  into  $\mathbf{R}^2$ .

- a)  $t \mapsto (1 - t^2, t)$
- b)  $t \mapsto (t^2, t^4)$
- c)  $t \mapsto (t^4, t^8)$
- d)  $t \mapsto (t^2, -t^4)$

Show in each case that the image is part of a parabola.

3. A stone is thrown from a building 50 ft tall in such a way that its coordinates  $(x(t), y(t))$  at time  $t$ , until it hits the ground, are given by:

$$x(t) = 3t, \quad y(t) = 50 - 16t^2.$$

- a) Find the time at which it hits the ground.
- b) Find the distance of the point where it hits the ground from the origin.
- c) Show that the path of the stone is part of a parabola and give the equation of this parabola.
- d) Find the time at which the stone is 30 ft above the ground.
- e) Find the time at which the stone is 20 ft above the ground.

4. A particle starts from a point  $(0, 6)$  in the plane. It is attracted by a magnet along the  $x$ -axis, and repelled by a magnet along the  $y$ -axis in such a way that its coordinates are given by

$$x(t) = 2t \quad \text{and} \quad y(t) = 6 - 4t^2.$$

Sketch the path of the particle.

- a) Find the time at which it stands at a distance of 2 units above the  $x$ -axis.
  - b) Show that the path is part of a parabola and give the equation of this parabola.
  - c) Find the time  $t_0$  at which the particle hits the  $x$ -axis.
  - d) Find the distance of the point at which the particle hits the  $x$ -axis from the origin.
5. A particle starts from the point  $(0, 6)$  in the plane. It is attracted by a magnet below the  $x$ -axis, and repelled by a magnet along the  $y$ -axis in such a way that its coordinates are given as a function of  $t$  by

$$x(t) = 3t \quad \text{and} \quad y(t) = 6 - 15t^3.$$

- a) Find the time at which the particle hits the  $x$ -axis.
- b) Give a simple equation in terms of  $x$  and  $y$  such that the coordinates  $(x(t), y(t))$  of the particle satisfy this equation. Sketch the graph of this equation.
- c) Find the distance of the point at which the particle hits the  $x$ -axis from the origin.
- d) Find the time at which the particle is at a distance of 2 units from the  $x$ -axis, and below the  $x$ -axis.
- e) Find the time at which the particle is at a distance of 5 units below the  $x$ -axis.
- f) Find the time at which the particle is at a distance of 7 units below the  $x$ -axis.
6. Draw the image of the straight line  $y = 2$  under the mapping

$$(x, y) \mapsto (2^y \cos x, 2^y \sin x).$$

In general what will be the image of a line  $y = \text{constant}$  under this mapping? What is the image of a line  $x = \text{constant}$  under this mapping? Draw these images for

- |                   |                  |
|-------------------|------------------|
| a) $y = 3,$       | b) $y = -4,$     |
| c) $x = \pi/4,$   | d) $x = \pi/3,$  |
| e) $x = -3\pi/2,$ | f) $x = 2\pi/3.$ |

7. Let  $0 \leq r_1 < r_2$  and let  $0 \leq \theta_1 < \theta_2 \leq 2\pi$ . Let  $S$  be the set of points in  $\mathbf{R}^2$  with coordinates  $(r, \theta)$  such that

$$r_1 \leq r \leq r_2 \quad \text{and} \quad \theta_1 \leq \theta \leq \theta_2.$$

Thus  $S$  is a rectangle. Let  $F: \mathbf{R}^2 \rightarrow \mathbf{R}^2$  be the mapping given by

$$F(r, \theta) = (r \cos \theta, r \sin \theta).$$

- a) Draw the image of  $S$  under  $F$ .
- b) Draw the special case when  $r_1 = 0$ .
- c) Draw the special case when  $\theta_1 = 0$  and  $\theta_2 = \pi/2$ .

- d) Draw the special case when  $\theta_1 = 0$ ,  $\theta_2 = 2\pi$ , and  $0 < r_1 < r_2$ .  
e) Draw the special case when  $r_1 = 2$ ,  $r_2 = 3$ ,  $\theta_1 = \pi/4$ ,  $\theta_2 = 3\pi/4$ .

Fool around with other special cases so that you get the hang of this system. The map  $F$  in this exercise is called the **polar coordinate map**.

## §2. FORMALISM OF MAPPINGS

Mappings in general satisfy a formalism similar to that developed for mappings of the plane into itself. We shall now repeat this formalism briefly.

We have the **identity mapping**  $I_S$  for any set  $S$ , such that

$$I_S(x) = x \text{ for all } x \text{ in } S.$$

Sometimes we omit the subscript  $S$  on  $I$  if the reference to  $S$  is made clear by the context.

Let

$$f: S \rightarrow T \quad \text{and} \quad g: U \rightarrow V$$

be mappings. Assume that  $T$  is a subset of  $U$ . Then we can form the **composite mapping**  $g \circ f$ , whose value at an element  $x$  of  $S$  is

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

Note that  $f(x)$  is an element of  $T$ , and since  $T$  is a subset of  $U$ , we can take  $g(f(x))$ , whose value is now in  $V$ .

**Example.** Let  $f$  and  $g$  be given by

$$f(x) = x^2 \quad \text{and} \quad g(u) = \sin u.$$

Then

$$g(f(x)) = g(x^2) = \sin(x^2).$$

**Example.** Let  $f$  and  $g$  be given by

$$f(x) = 3x - 2 \quad \text{and} \quad g(u) = \frac{1}{\cos u}.$$

Then

$$g(f(x)) = \frac{1}{\cos(3x - 2)}.$$

Observe that  $g$  is defined only for those numbers  $u$  such that  $u \neq \frac{\pi}{2} + n\pi$ , and  $n$  is an integer. Thus the composite function is defined only for those numbers  $x$  such that

$$3x - 2 \neq \frac{\pi}{2} + n\pi,$$

and  $n$  is an integer.

**Example.** Let  $I: S \rightarrow S$  be the identity mapping, and let

$$f: S \rightarrow S$$

be any mapping. Then

$$I \circ f = f \circ I = f.$$

This is clear, because, for any  $x$  in  $S$ , we have

$$(I \circ f)(x) = I(f(x)) = f(x),$$

and similarly,

$$(f \circ I)(x) = f(I(x)) = f(x).$$

We note that the identity mapping behaves like the number 1 in a context where composition of mappings behaves like multiplication. This analogy is again evident in our next assertion.

**Composition of mappings is associative.** *This means: If*

$$f: S \rightarrow T, \quad g: T \rightarrow U, \quad h: U \rightarrow V$$

*are mappings, then*

$$(h \circ g) \circ f = h \circ (g \circ f).$$

The proof is exactly the same as the proof for the associativity of isometries of the plane into itself. Namely for any  $x$  in  $S$ , we have

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

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

Let

$$f: S \rightarrow T$$

be a mapping. By an **inverse mapping** for  $f$ , we mean a mapping

$$g: T \rightarrow S$$

such that the composites of  $f$  and  $g$  are the identities of  $S$  and  $T$ , respectively; that is,

$$g \circ f = I_S \quad \text{and} \quad f \circ g = I_T.$$

We have had many examples of inverse mappings with translations, rotations, and the like. We also have examples with functions.

**Example.** Let  $\mathbf{R}^+$  denote the positive real numbers. Let

$$f: \mathbf{R}^+ \rightarrow \mathbf{R}^+$$

be the square, i.e.  $f(x) = x^2$ . Then  $f$  has an inverse function, namely the square root. If  $x > 0$  and  $g$  is the square root, then  $g(x^2) = x$  and  $f(\sqrt{x}) = x$ . Thus  $f, g$  are inverse to each other.

**Example.** The exponential function and the logarithm are inverse to each other (when taken to the same base). In other words,

$$a^{\log_a x} = x \quad \text{and} \quad \log_a(a^x) = x.$$

This is merely the definition of the logarithm.

Thus we see that the notion of inverse mapping or inverse function includes many special cases already considered.

If  $f: S \rightarrow T$  is a mapping having an inverse mapping, then we denote this inverse by

$$f^{-1}: T \rightarrow S.$$

The elementary proof concerning the inverse of a number can be applied to give the uniqueness of the inverse mapping. Indeed, suppose that

$$g: T \rightarrow S \quad \text{and} \quad h: T \rightarrow S$$

are inverse mappings for  $f$ . Then we must have  $g = h$ . We leave the proof as an exercise.

The section on permutations will provide further examples for mappings, as well as an extension of the theory in an interesting direction. You may well want to read our next remarks in connection with permutations, and their iterations.

Let  $f: S \rightarrow S$  be a mapping of a set into itself. We may then iterate  $f$ . If  $x$  is an element of  $S$ , then we can form the iterates of the values,

$$f(x), \quad f(f(x)), \quad f(f(f(x))), \dots$$

We use the exponential notation, and write

$$\begin{aligned}f^2(x) &= f(f(x)) \\f^3(x) &= f(f(f(x))) \\&\vdots\end{aligned}$$

In general, we write

$$f^k(x) = f(f(f(\dots f(x) \dots)))$$

for the iteration of  $f$  taken  $k$  times, applied to  $x$ . Thus  $f^k$  is again a function defined on  $S$ , with values in  $S$ . The value of  $f^k$  at  $x$  is  $f^k(x)$ . Observe that  $f^{k+1} = f \circ f^k$ , or, in terms of  $x$ ,

$$f^{k+1}(x) = f(f^k(x)).$$

**Example.** Let  $f: \mathbf{R} \rightarrow \mathbf{R}$  be the function such that  $f(x) = x + 1$ . Then

$$f^2(x) = f(f(x)) = f(x + 1) = x + 2.$$

Similarly,

$$f^3(x) = f(f^2(x)) = f(x + 2) = x + 3.$$

In general, we see that

$$f^k(x) = x + k$$

for any positive integer  $k$ .

**Example.** Let  $F: \mathbf{R}^2 \rightarrow \mathbf{R}^2$  be the mapping given by

$$F(x, y) = (x + 3, y - 4).$$

If we denote  $(x, y)$  by  $X$  and if we let  $A = (3, -4)$ , then we can abbreviate the formula describing  $F$  as follows:

$$F(X) = X + A.$$

Thus we see that  $F$  is translation by  $A$ . Iterating  $F$  is analogous to the process used in the preceding example. For instance,

$$F^2(X) = F(X + A) = X + A + A = X + 2A,$$

and in general, for any positive integer  $k$ , we have

$$F^k(X) = X + kA.$$

**Example.** Let  $f: \mathbf{R} \rightarrow \mathbf{R}$  be the map given by  $f(x) = x^3$ . Then

$$f^2(x) = f(f(x)) = f(x^3) = (x^3)^3 = x^9.$$

Also,

$$f^3(x) = f(f^2(x)) = f(x^9) = (x^9)^3 = x^{27}.$$

In general, for any positive integer  $k$ , we have

$$f^k(x) = x^{3^k}.$$

Just as with mappings of the plane into itself, we have a general rule of exponent for mappings; namely, if  $m, n$  are positive integers and

$$f: S \rightarrow S$$

is a mapping, then

$$f^{m+n} = f^m \circ f^n.$$

This simply means that if we iterate the mapping  $m + n$  times, it amounts to iterating it first  $n$  times and then  $m$  times. Thus our formula is clear. We define

$$f^0 = I \quad (= I_S, \text{ identity mapping})$$

and we see that our formula also holds if  $m$  or  $n$  is 0. This is an immediate consequence of the definitions.

In fact, this formula also holds for negative integers  $m$  or  $n$ , if we make the appropriate definition. Namely, assume that  $f^{-1}$  exists. Then we define

$$f^{-k} = (f^{-1})^k$$

to be the composite of  $f^{-1}$  with itself  $k$  times. Note that

$$f^k \circ f^{-k} = f^0 = I.$$

This just means that if we apply  $f^{-1}$  to an element  $k$  times, and then apply  $f$  itself  $k$  times, we get that element. We can draw the picture as follows. Let  $x$  be an element of  $S$ . Then

$$f^k \circ f^{-k}(x) = \underbrace{f(f(\dots f}_{k \text{ times}} \underbrace{f^{-1}(f^{-1}(\dots f^{-1}(x))))}_{k \text{ times}}).$$

But the  $f \circ f^{-1}$  in the middle is the identity  $I$ . Similarly, we combine  $f \circ f^{-1}$  to get  $I$ , and we do this  $k$  times to see that

$$f^k \circ f^{-k}(x) = x.$$

**Note.** Later we shall discuss a more formal way of giving the above proof, namely induction. After you have read about induction, you will then understand the following argument which proves that

$$f^k \circ f^{-k} = I$$

by induction. It is true for  $k = 1$  by definition. Assume it for a positive integer  $k$ . Then

$$\begin{aligned} f^{k+1} \circ f^{-(k+1)} &= f \circ f^k \circ (f^{-1})^k \circ f^{-1} \\ &= f \circ I \circ f^{-1} \quad (\text{by induction}) \\ &= f \circ f^{-1} = I. \end{aligned}$$

**Example.** Let  $f: \mathbf{R} \rightarrow \mathbf{R}$  be the map such that  $f(x) = x + 3$ . Then

$$f^k(x) = x + 3k \quad \text{and} \quad f^{-k}(x) = x - 3k.$$

**Example.** Let  $A = (3, -4)$  and let  $F: \mathbf{R}^2 \rightarrow \mathbf{R}^2$  be the map such that

$$F(X) = X + A.$$

Then

$$F^k(X) = X + kA \quad \text{and} \quad F^{-k}(X) = X - kA.$$

**Example.** Let  $f: \mathbf{R} \rightarrow \mathbf{R}$  be the map such that  $f(x) = x^3$ . Then

$$f^k(x) = x^{3^k} \quad \text{and} \quad f^{-k}(x) = x^{3^{-k}}.$$

*In general, for any integers  $m, n$  (positive, negative, or zero) and any map  $f: S \rightarrow S$  which has an inverse, we have the formula*

$$f^{m+n} = f^m \circ f^n.$$

This can also be proved by induction, but we omit the proof.

**Example.** Inserting special numbers into the above formula, we have

$$f^5 \circ f^{-3} = f^2.$$

**Example.** Suppose that  $f^m = f^n$  for some integers  $m, n$ , and assume that  $f$  has an inverse mapping. Then, composing both sides with  $f^{-n}$ , we obtain

$$f^{m-n} = I.$$

Observe again how composition of mappings is analogous to multiplication.

**Example.** If  $f$  has an inverse mapping, and if  $f^4 = f^{-5}$ , then

$$f^9 = I.$$

**Example.** Let  $f_1, \dots, f_m$  be mappings of a set  $S$  such that

$$f_i^2 = I$$

for each  $i = 1, \dots, m$ . Let  $f$  be a map such that

$$f_1 \circ f_2 \circ \dots \circ f_m \circ f = I.$$

Compose on the left with  $f_1$ , which is nothing but  $f_1^{-1}$ . We obtain

$$f_2 \circ \dots \circ f_m \circ f = f_1.$$

Now compose with  $f_2$  on the left. We get

$$f_3 \circ \dots \circ f_m \circ f = f_2 \circ f_1.$$

Proceeding in this way, composing with  $f_3, \dots, f_m$  successively, we find that

$$f = f_m \circ f_{m-1} \circ \dots \circ f_2 \circ f_1.$$

This procedure will be used in the next section when we deal with permutations.

## EXERCISES

- Let  $f: S \rightarrow T$  and  $g: S \rightarrow T$  be mappings. Let

$$h: T \rightarrow U$$

be a mapping having an inverse mapping denoted by

$$h^{-1}: U \rightarrow T.$$

If

$$h \circ f = h \circ g,$$

prove that  $f = g$ . This is the **cancellation law** for mappings.

2. Let  $f: S \rightarrow T$  be a mapping having an inverse mapping. Prove the following statements.
- If  $x, y$  are elements of  $S$  and  $f(x) = f(y)$ , then  $x = y$ .
  - If  $z$  is an element of  $T$ , then there exists an element  $x$  of  $S$  such that  $f(x) = z$ .
3. Let  $a, b$  be non-zero numbers. Let  $F_{a,b}: \mathbf{R}^2 \rightarrow \mathbf{R}^2$  be the mapping such that

$$F_{a,b}(x, y) = (ax, by).$$

Show that  $F_{a,b}$  has an inverse mapping.

4. Let  $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$  be the mapping defined by

$$f(x, y) = (2x - y, y + x).$$

Show that  $f$  has an inverse mapping. [Hint: Let  $u = 2x - y$  and  $v = y + x$ . Solve for  $x, y$  in terms of  $u$  and  $v$ .]

5. Let  $f: \mathbf{R}^2 \rightarrow \mathbf{R}^2$  be the mapping defined by

$$f(x, y) = (3x + y, 2x - 4y).$$

Show that  $f$  has an inverse mapping.

6. Let  $f: S \rightarrow S$  be a mapping which has an inverse mapping.

- If  $f^3 = I$  and  $f^5 = I$ , show that  $f = I$ .
- If  $f^2 = I$  and  $f^7 = I$ , show that  $f = I$ .
- If  $f^4 = I$  and  $f^{11} = I$ , show that  $f = I$ .

7. Let  $f, g$  be mappings of a set  $S$  into itself, and assume that they have inverse mappings. Assume also that  $f \circ g = g \circ f$ . Express each one of the following in the form  $f^m \circ g^n$  where  $m, n$  are integers.

- $f^3 \circ g^2 \circ f^{-2} \circ f^5 \circ g^{-5}$
- $f^7 \circ g \circ g^4 \circ f^{-6} \circ g^3$
- $f^4 \circ g^5 \circ f^{-5} \circ g^{-7} \circ g^2 \circ f^2$
- $f^4 \circ f^{-8} \circ g^2 \circ f^3 \circ g^3 \circ f^{-2}$

8. a) Let  $f, g$  be mappings of a set  $S$  into itself, and assume that they have inverse mappings. Prove that  $f \circ g$  has an inverse mapping, and express it in terms of  $f^{-1}, g^{-1}$ .
- b) Let  $f_1, \dots, f_m$  be maps of  $S$  into itself, and assume that each  $f_i$  has an inverse mapping. Show that the composite

$$f_1 \circ f_2 \circ \dots \circ f_m$$

has an inverse mapping, and express this inverse mapping in terms of the maps  $f_i^{-1}$ .

9. Let  $f$  be a mapping of a set  $S$  into itself, and assume that  $f$  has an inverse mapping.
- If  $f^5 = I$ , express  $f^{-1}$  as a positive power of  $f$ .
  - In general, if  $f^n = I$ , for some positive power of  $f$ , express  $f^{-1}$  as a positive power of  $f$ .
10. Let  $f, g$  be mappings of a set  $S$  into itself. Assume that  $f^2 = g^2 = I$  and that  $f \circ g = g \circ f$ . Prove that  $(f \circ g)^2 = I$ . Prove that  $(f \circ g)^3 = I$ . What about  $(f \circ g)^n$  for any positive integer  $n$ ? What about  $(f \circ g)^n$  when  $n$  is a negative integer?

### §3. PERMUTATIONS

Let  $J_n$  be the set of integers  $1, 2, \dots, n$ , that is the set of integers  $k$  such that  $1 \leq k \leq n$ . By a **permutation** of  $J_n$ , we mean a mapping

$$\sigma: J_n \rightarrow J_n$$

having the following property. If  $i, j$  are in  $J_n$  and  $i \neq j$ , then  $\sigma(i) \neq \sigma(j)$ . Thus the image of a permutation  $\sigma$  consists of  $n$  distinct integers

$$\sigma(1), \quad \sigma(2), \quad \dots, \quad \sigma(n),$$

which must therefore be the integers  $1, 2, \dots, n$  in a different order. Thus we denote such a permutation  $\sigma$  by the symbols

$$\begin{bmatrix} 1 & 2 & 3 & \cdots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \cdots & \sigma(n) \end{bmatrix}.$$

**Example.** The permutation of  $J_4$  given by

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{bmatrix}$$

is such that

$$\sigma(1) = 3, \quad \sigma(2) = 1, \quad \sigma(3) = 4, \quad \sigma(4) = 2.$$

If  $\sigma$  and  $\sigma'$  are permutations, we denote their composite by  $\sigma\sigma'$ , omitting the little circle between them for simplicity and to emphasize the analogy with “multiplication”. In fact, we also call this composite the **product** of  $\sigma$  and  $\sigma'$ . Watch out! It is not always true that  $\sigma\sigma' = \sigma'\sigma$ .

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{bmatrix} \quad \text{and} \quad \sigma' = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{bmatrix}.$$

To determine  $\sigma'\sigma$ , we just look at the effect of this composite on each one of the numbers 1, 2, 3, 4. Thus with arrows:

$$\begin{aligned} 1 &\xrightarrow{\sigma} 2 \xrightarrow{\sigma'} 4, \\ 2 &\xrightarrow{\sigma} 3 \xrightarrow{\sigma'} 2, \\ 3 &\xrightarrow{\sigma} 4 \xrightarrow{\sigma'} 1, \\ 4 &\xrightarrow{\sigma} 1 \xrightarrow{\sigma'} 3. \end{aligned}$$

Thus we have

$$\sigma'\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{bmatrix}.$$

On the other hand, in a similar way, you find that

$$\sigma\sigma' = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 3 & 2 \end{bmatrix}.$$

Thus  $\sigma\sigma' \neq \sigma'\sigma$ .

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{bmatrix} \quad \text{and} \quad \sigma' = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}.$$

Then

$$\begin{aligned} \sigma\sigma'(1) &= \sigma(\sigma'(1)) = \sigma(3) = 3, \\ \sigma\sigma'(2) &= \sigma(\sigma'(2)) = \sigma(1) = 2, \\ \sigma\sigma'(3) &= \sigma(\sigma'(3)) = \sigma(2) = 1, \end{aligned}$$

so that we can write

$$\sigma\sigma' = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}.$$

If  $\sigma$  is a permutation of  $J_n$ , then we have already mentioned that

$$\sigma(1), \quad \sigma(2), \quad \dots, \quad \sigma(n)$$

are simply the elements of  $J_n$  in a different order. Thus to each element  $j$  of  $J_n$  there exists a unique element  $i$  such that

$$\sigma(i) = j.$$

We can therefore define the **inverse permutation** of  $\sigma$ , as with inverse mappings, to be that permutation  $\sigma^{-1}$  such that

$$\sigma\sigma^{-1} = \sigma^{-1}\sigma = I \text{ (the identity permutation).}$$

We have  $\sigma(i) = j$  if and only if  $\sigma^{-1}(j) = i$ .

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}.$$

Since  $\sigma(1) = 3$  we have  $\sigma^{-1}(3) = 1$ . Since  $\sigma(2) = 1$  we have  $\sigma^{-1}(1) = 2$ . Since  $\sigma(3) = 2$  we have  $\sigma^{-1}(2) = 3$ . Hence

$$\sigma^{-1} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix}.$$

**Example.** The identity permutation is of course given by

$$I = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}.$$

If  $\sigma_1, \dots, \sigma_r$  are permutations of  $J_n$ , then the inverse of the composite permutation

$$\sigma_1 \cdots \sigma_r$$

is the permutation

$$\sigma_r^{-1} \cdots \sigma_1^{-1}.$$

Just multiply one with the other, and you will find that all the factors cancel out to give the identity.

For instance, with three permutations, we have

$$(\sigma_1\sigma_2\sigma_3)^{-1} = \sigma_3^{-1}\sigma_2^{-1}\sigma_1^{-1}$$

because

$$\sigma_1\sigma_2\sigma_3\sigma_3^{-1}\sigma_2^{-1}\sigma_1^{-1} = I,$$

and similarly on the other side.

There is an important special case of a permutation, namely the **transposition** which interchanges two distinct numbers  $i \neq j$ , and leaves the others fixed.

**Example.** The permutation

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{bmatrix}$$

is a transposition, which interchanges 3 and 2, and leaves 1, 4 fixed.

If  $\tau$  is a transposition, then it is clear that

$$\tau^2 = I.$$

Thus in this case, we have

$$\tau^{-1} = \tau.$$

**Theorem 1.** Every permutation of  $J_n$  can be expressed as a product of transpositions.

*Proof.* Let  $\sigma$  be a permutation of  $J_n$ . Let  $\sigma(n) = k$ . If  $k = n$ , we let  $\tau_n$  be the identity. If  $k \neq n$ , we let  $\tau_n$  be the transposition which interchanges  $k$  and  $n$ , and leaves the other integers fixed. Then  $\tau_n \sigma$  leaves  $n$  fixed, and may therefore be viewed as a permutation of  $J_{n-1}$ . We now repeat our procedure. We let  $\tau_{n-1}$  be either the identity, or a transposition of  $J_{n-1}$  such that

$$\tau_{n-1} \tau_n \sigma$$

leaves  $n$  and  $n - 1$  fixed. We continue in this way, finding

$$\tau_n, \quad \tau_{n-1}, \quad \dots, \quad \tau_2, \quad \tau_1$$

which are either the identity, or transpositions. Finally

$$\tau_1 \tau_2 \cdots \tau_n \sigma$$

leaves every one of the numbers  $1, \dots, n$  fixed, and is therefore equal to the identity. Thus

$$\tau_1 \tau_2 \cdots \tau_n \sigma = I.$$

Now multiply on the left by  $\tau_1$ , then by  $\tau_2$ , and so on. Since  $\tau_i^2 = I$  for each  $i$ , we find that

$$\sigma = \tau_n \cdots \tau_1.$$

In this product, we may omit those  $\tau_i$  which are the identity, and we find that  $\sigma$  has been expressed as a product of transpositions.

**Note.** It may happen, of course, as when  $\sigma$  is the identity already, that every  $\tau_i$  is the identity, so that in our product, we have simply  $\sigma = I$ . Thus.

in a sense, there is no transposition in this product. It is a matter of convention about our use of language how we deal with this case. The best convention is to agree that  $I$  is the product of zero transpositions, and allow this possibility to be included in our expression “product of transpositions”. After all, we did not say how many transpositions, and it could have been zero. With this convention, the formulation we have given to Theorem 1 is correct.

The procedure used to prove Theorem 1 gives us an effective way in practice to express a permutation as a product of transpositions.

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}.$$

We want to express  $\sigma$  as a product of transpositions. Let  $\tau$  be the transposition which interchanges 3 and 1, and leaves 2 fixed. Then we find that

$$\tau\sigma = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{bmatrix}$$

so that  $\tau\sigma$  is a transposition, which we denote by  $\tau'$ . We can then write  $\tau\sigma = \tau'$ , so that composing on the left with  $\tau$  yields

$$\sigma = \tau\tau'$$

because  $\tau^2 = I$ . This is the desired product.

**Example.** We want to express the permutation

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{bmatrix}$$

as a product of transpositions. Let  $\tau_1$  be the transposition which interchanges 1 and 2, and leaves 3, 4 fixed. Then

$$\tau_1\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 2 \end{bmatrix}.$$

Now let  $\tau_2$  be the transposition which interchanges 2 and 3, and leaves 1, 4 fixed. Then

$$\tau_2\tau_1\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{bmatrix},$$

and we see that  $\tau_2\tau_1\sigma$  is a transposition, which we may denote by  $\tau_3$ . Then we get  $\tau_2\tau_1\sigma = \tau_3$  so that multiplying on the left with  $\tau_2$  and  $\tau_1$  successively we get

$$\sigma = \tau_1\tau_2\tau_3.$$

This is the desired expression.

In expressing a permutation as a product of transpositions, there is of course no uniqueness. This is possible in many ways.

**Example.** Take the permutation  $\sigma$  of the preceding example. Let  $\tau_4$  be the transposition which interchanges 1 and 3. Then

$$\tau_4\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix}.$$

If again  $\tau_1$  is the transposition which interchanges 1 and 2, and  $\tau_3$  is the transposition which interchanges 3 and 4, we see that

$$\tau_4\sigma = \tau_1\tau_3.$$

Hence composing on the left with  $\tau_4$  yields

$$\sigma = \tau_4\tau_1\tau_3.$$

Even though we don't have uniqueness of the expression of  $\sigma$  as a product of transpositions, still an interesting phenomenon occurs. Suppose that we have written  $\sigma$  as a product of transpositions in two ways:

$$\sigma = \tau_1 \cdots \tau_r = \tau'_1 \cdots \tau'_s.$$

The number of transpositions occurring in these expressions may not be the same, but it turns out that if  $r$  is even, then  $s$  must also be even, and if  $r$  is odd, then  $s$  must also be odd. In other words, we have:

**Theorem 2.** *Let  $\sigma$  be a permutation of  $J_n$ . In any expressions of  $\sigma$  as a product of transpositions, the number of transpositions occurring in such a product is either always even or always odd.*

The proof for Theorem 2 will involve a little more theory about permutations than we have right now, and we postpone it to the end. However, to use Theorem 2 is very easy, and we make some comments about that.

Suppose that a permutation  $\sigma$  can be expressed as a product of an even number of transpositions. Then we say that  $\sigma$  is an **even permutation**. If it can be expressed as a product of an odd number of transpositions, then we say that  $\sigma$  is an **odd permutation**. If

$$\sigma = \tau_1 \cdots \tau_m$$

is an expression of  $\sigma$  as a product of transposition, then we call

$$(-1)^m$$

the **sign** of  $\sigma$ . This sign is 1 or  $-1$  according as  $\sigma$  is even or odd.

**Example.** The sign of the permutation

$$\sigma = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}$$

is 1, and  $\sigma$  is an even permutation (cf. a previous example where we expressed this permutation as a product of 2 transpositions).

**Example.** The sign of the permutation

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{bmatrix}$$

is  $-1$ , and  $\sigma$  is an odd permutation (cf. a previous example where we expressed this permutation as a product of 3 transpositions).

*If you are not interested in the proof of Theorem 2, or in the further theory of permutations, you may omit the rest of this section.*

We now start on the discussion which will lead to the proof of Theorem 2. It involves looking more closely into the structure of permutations.

Let  $\sigma$  be a permutation of  $J_n$  and let  $i$  be one of the integers in  $J_n$ . Let us look at what happens to  $i$  when we apply successive powers of  $\sigma$ . We obtain

$$i, \quad \sigma(i), \quad \sigma^2(i), \quad \dots$$

Since  $J_n$  has only a finite number of elements, it follows that these numbers cannot be all distinct. Let us see this on an example.

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 1 & 2 \end{bmatrix}.$$

Let us start with  $i = 1$ . Then

$$\begin{aligned} \sigma(1) &= 3, \\ \sigma^2(1) &= \sigma(3) = 4, \\ \sigma^3(1) &= \sigma(4) = 1. \end{aligned}$$

Now we see that we start over again, namely

$$\begin{aligned} \sigma^4(1) &= \sigma(\sigma^3(1)) = \sigma(1) = 3, \\ \sigma^5(1) &= 4, \\ \sigma^6(1) &= 1. \end{aligned}$$

We see that we are going around in a circle, and we can even represent the numbers 1, 3, 4 on a circle as in Fig. 14-3.

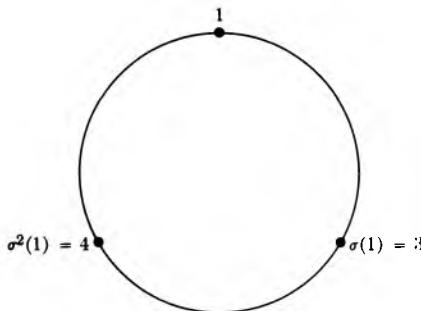


Fig. 14-3

In general, starting with any  $i$ , we have an analogous situation. Since the set  $J_n$  has only a finite number of elements, it follows that the numbers

$$i, \sigma(i), \sigma^2(i), \sigma^3(i), \dots$$

are not all distinct. Hence there exist two positive integers  $r, s$  with  $r < s$  such that  $\sigma^r(i) = \sigma^s(i)$ . Applying  $\sigma^{-r}$  to both sides, we get

$$i = \sigma^{s-r}(i).$$

Thus some power of  $\sigma$  leaves  $i$  fixed. Let  $k$  be the smallest positive integer such that  $\sigma^k(i) = i$ . We contend that the numbers

$$i, \sigma(i), \sigma^2(i), \dots, \sigma^{k-1}(i)$$

are distinct.

*Proof.* If  $\sigma^m(i) = \sigma^n(i)$  with two integers  $m, n$  such that

$$1 \leq m < n \leq k,$$

then  $\sigma^{n-m}(i) = i$ , and  $n - m < k$ . This contradicts our assumption that  $k$  is the smallest positive integer such that  $\sigma^k(i) = i$ , and therefore proves our contention.

We now see that the general situation is analogous to that of the example. and we can represent the numbers

$$i, \sigma(i), \sigma^2(i), \dots, \sigma^{k-1}(i)$$

as going around a circle as illustrated in Fig. 14-4. We shall call the set consisting of

$$i, \sigma(i), \sigma^2(i), \dots, \sigma^{k-1}(i)$$

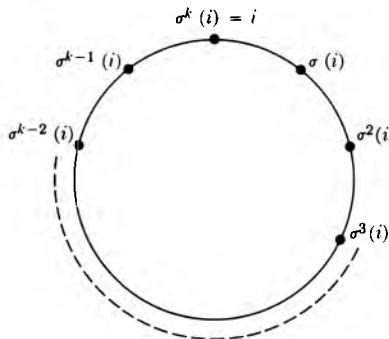


Fig. 14-4

an **orbit** of  $\sigma$ , or, specifically, the orbit to which  $i$  belongs. The effect of  $\sigma$  on this orbit will be denoted by square brackets,

$$\gamma = [i, \sigma(i), \dots, \sigma^{k-1}(i)],$$

and will be called the **cycle** of the orbit. We call  $k$  the **period** of  $i$  under  $\sigma$ , or also the **period** of the cycle, or the **length** of the cycle.

**Example.** Consider again the permutation

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 1 & 2 \end{bmatrix}.$$

Then the orbit of 1 consists of 1, 3, 4, and its cycle is denoted by

$$[1, 3, 4].$$

The cycle of 2 is given by

$$[2, 5]$$

because

$$\sigma(2) = 5 \quad \text{and} \quad \sigma(5) = 2.$$

According to our terminology, we see that 1 has period 3 under  $\sigma$ , whereas 2 has period 2. The cycle to which 1 belongs has length 3, and the cycle to which 5 belongs has length 2.

Note that in the orbit of 1, we could have selected 3 and started considering

$$3, \sigma(3), \sigma^2(3), \sigma^3(3) = 3,$$

so that the cycle  $[1, 3, 4]$  can also be written

$$[3, 4, 1].$$

Similarly, we have  $[2, 5] = [5, 2]$ .

**Example.** Let  $\sigma$  be the identity permutation of  $J_n$ . Then each orbit of  $\sigma$  consists of a single element, because  $\sigma(i) = i$  for each  $i$ .

**Example.** Let  $\tau$  be a transposition, which interchanges the two numbers  $a, b$  and leaves the others fixed. Then the orbit of  $a$  (or  $b$ ) consists of  $a, b$  and its cycle is denoted by  $[a, b]$ . All the other orbits consist of one element. Thus we may write for simplicity

$$\tau = [a, b].$$

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 1 \end{bmatrix}.$$

Then  $\sigma$  has just one orbit, namely  $J_5$  itself, and its cycle is

$$[1, 2, 3, 4, 5].$$

We have the usual picture representing  $\sigma$  in Fig. 14-5.

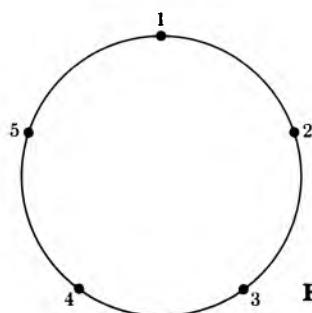


Fig. 14-5

As with transpositions, it will be useful to write

$$\sigma = [3, 4, 5, 1, 2].$$

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 1 & 2 \end{bmatrix}.$$

Then  $\sigma$  has two orbits, whose cycles are

$$[1, 3, 4] \quad \text{and} \quad [2, 5].$$

Observe that the two orbits have no elements in common. This is a special case of a general fact, which we shall state and prove below. The effect of

$\sigma$  is determined by its effect on the orbits, and hence we shall use the notation

$$\sigma = [1, 3, 4][2, 5].$$

We shall call this the **orbit decomposition** of  $\sigma$ .

**Theorem 3.** *If  $a, b$  are elements of  $J_n$  and  $\sigma$  is a permutation of  $J_n$ , then the orbits to which  $a$  and  $b$  belong either coincide, or have no element in common, i.e. are disjoint.*

*Proof.* Suppose that the orbits have some element in common. This means that we have

$$\sigma^r(a) = \sigma^s(b)$$

for some positive integers  $r, s$ . Then

$$a = \sigma^{s-r}(b),$$

and thus we see that  $a$  occurs as some  $\sigma^m(b)$ . But then all the numbers

$$\sigma(a) = \sigma^{m+1}(b), \quad \sigma^2(a) = \sigma^{m+2}(b), \dots$$

occur in the orbit of  $b$ , which means that the orbit of  $a$  is contained in the orbit of  $b$ . Conversely, we see in the same way that the orbit of  $b$  is contained in the orbit of  $a$ , and thus the two orbits must be equal. This proves our theorem.

Let  $S$  be a given orbit of  $\sigma$  and let  $i$  be any element of  $S$ . Let us denote the cycle of the orbit by

$$\gamma = [i, \sigma(i), \dots, \sigma^{k-1}(i)].$$

If  $S_1, \dots, S_m$  are the disjoint orbits of  $\sigma$ , then they have corresponding cycles  $\gamma_1, \dots, \gamma_m$  and we write symbolically

$$\sigma = \gamma_1 \gamma_2 \cdots \gamma_m.$$

We call this the **orbit decomposition** of  $\sigma$ , as in our example above.

**Example.** Let

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 6 & 2 & 1 & 4 & 3 \end{bmatrix}.$$

Then the two orbit cycles are

$$[1, 5, 4] \quad \text{and} \quad [2, 6, 3].$$

Thus we get the orbit decomposition of  $\sigma$ ; namely,

$$\sigma = [1, 5, 4][2, 6, 3].$$

In this case, both cycles have length 3, and every element of  $J_6$  has period 3 under  $\sigma$ .

**Example.** We can write the orbit decomposition of the identity of  $J_n$  as

$$I = [1][2] \cdots [n].$$

Let  $\sigma$  be a permutation of  $J_n$  and let  $S_1, \dots, S_m$  be its orbits. We then have the representation of  $\sigma$  as a product of cycles

$$\sigma = \gamma_1 \cdots \gamma_m$$

corresponding to these orbits. We ask: What are the orbits of  $\tau\sigma$ , if  $\tau$  is a transposition? Two cases arise. Let  $\tau = [a, b]$ .

*Case 1.* The two numbers  $a, b$  belong to the same orbit of  $\sigma$ , say  $S_1$ . Let

$$\gamma_1 = [i_1, \dots, i_k]$$

be the cycle of this orbit, and suppose that

$$a = i_r \quad \text{and} \quad b = i_s$$

with  $r < s$ . Thus we can express symbolically the effect of  $\tau\sigma$  on the orbit  $S_1$  by

$$\tau\gamma_1 = [i_r, i_s][i_1, \dots, i_{r-1}, i_r, \dots, i_{s-1}, i_s, \dots, i_k].$$

Now follow through the effect of  $\tau\gamma_1$  on each one of the numbers above. We see that under  $\tau\gamma_1$ , we have:

$$i_1 \mapsto i_2 \mapsto \cdots \mapsto i_{r-1} \mapsto i_s \mapsto i_{s+1} \mapsto \cdots \mapsto i_k \mapsto i_1,$$

thus forming a cycle

$$\gamma'_1 = [i_1, i_2, \dots, i_{r-1}, i_s, i_{s+1}, \dots, i_k].$$

On the other hand, we see that under  $\tau\gamma_1$ , we have:

$$i_r \mapsto i_{r+1} \mapsto \cdots \mapsto i_{s-1} \mapsto i_r,$$

thus forming another cycle

$$\gamma''_1 = [i_r, i_{r+1}, \dots, i_{s-1}]$$

which is disjoint from  $\gamma'_1$ . Furthermore,  $\tau$  has no effect on the cycles  $\gamma_2, \dots, \gamma_m$ . Hence

$$\tau\sigma = \tau\gamma_1\gamma_2 \cdots \gamma_m = \gamma'_1\gamma''_1\gamma_2 \cdots \gamma_m.$$

Thus  $\tau$  splits the orbits  $S_1$  into two distinct orbits  $S'_1$  and  $S''_1$  while leaving all other orbits fixed. In the present case, we see that  $\tau\sigma$  has one more orbit than  $\sigma$ .

*Case 2.* The two numbers  $a, b$  belong to distinct orbits of  $\sigma$ , say  $S_1$  and  $S_2$ .

If you work out the same type of argument as in Case 1, you will see that the effect of  $\tau$  on the orbits of  $\sigma$  is to join  $S_1$  and  $S_2$ , so that  $\tau\sigma$  has one less orbit than  $\sigma$ . You just have to follow through what happens to each one of the integers in the cycles

$$\gamma_1 = [i_1, \dots, i_r] \quad \text{and} \quad \gamma_2 = [j_1, \dots, j_s]$$

corresponding to the two orbits  $S_1$  and  $S_2$ . You will see that  $\tau$  has the effect of joining these orbits into a single orbit for  $\tau\sigma$ . Hence in this case,  $\tau\sigma$  has one less orbit than  $\sigma$ .

How does all this apply to the proof of Theorem 2? Very simply. Suppose that a given permutation  $\sigma$  of  $J_n$  can be written as a product of an odd number of transpositions, say

$$\sigma = \tau_1 \cdots \tau_p,$$

where  $p$  is odd. Let us start with the identity  $I$ , which has  $n$  orbits. As we multiply  $I$  successively by  $\tau_p, \tau_{p-1}, \dots, \tau_1$ , we either increase the number of orbits by 1, or decrease it by 1, each time. Consequently, we obtain the relation:

$$\text{Number of orbits of } \sigma = n + \text{an odd integer.}$$

If it were possible to express  $\sigma$  as a product

$$\sigma = \tau'_1 \cdots \tau'_q,$$

where  $q$  is even, then arguing similarly, we would obtain the relation:

$$\text{Number of orbits of } \sigma = n + \text{an even integer.}$$

However the integer

$$(\text{Number of orbits of } \sigma) - n$$

is either odd or even, and cannot be both. Consequently we cannot express  $\sigma$  both as a product of an even number of transpositions and an odd number of transpositions. This concludes the proof of Theorem 2.

The general rule following from this discussion can be stated as follows.

**Theorem 4.** *Let  $\sigma$  be a permutation of  $J_n$  and suppose that  $\sigma$  has  $k$  orbits. If  $n - k$  is even, the  $\sigma$  is an even permutation. If  $n - k$  is odd, then  $\sigma$  is an odd permutation.*

**Example.** Theorem 4 gives us another method for determining the sign of a permutation. For instance, take  $n = 6$ . In a previous example we have seen that the permutation

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 6 & 2 & 1 & 4 & 3 \end{bmatrix}$$

has two orbits, whose cycles are

$$[1, 5, 4] \quad \text{and} \quad [2, 6, 3].$$

Since  $6 - 2 = 4$  is even, it follows that  $\sigma$  is an even permutation.

**Example.** Let  $n = 5$ . The permutation

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 1 & 2 \end{bmatrix}$$

has two orbits, whose cycles are

$$[1, 3, 4] \quad \text{and} \quad [2, 5].$$

Since  $5 - 2 = 3$  is odd, it follows that  $\sigma$  is an odd permutation.

## EXERCISES

1. Express the following permutations as products of transpositions, and determine their signs.

a)  $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix}$

b)  $\begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}$

c)  $\begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$

d)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{bmatrix}$

e)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix}$

f)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{bmatrix}$

g)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{bmatrix}$

h)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{bmatrix}$

i)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{bmatrix}$

2. Express the following permutations as products of transpositions and determine their signs.

a)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 5 & 1 & 3 \end{bmatrix}$

b)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 1 & 3 & 2 \end{bmatrix}$

c) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 2 & 1 & 4 \end{bmatrix}$$

e) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 5 & 3 & 4 \end{bmatrix}$$

g) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1 \end{bmatrix}$$

i) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 3 & 1 & 2 & 4 \end{bmatrix}$$

d) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 3 & 1 & 2 \end{bmatrix}$$

f) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 1 & 2 & 5 \end{bmatrix}$$

h) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 1 & 5 & 2 & 3 \end{bmatrix}$$

3. Express the following permutations as products of transpositions and determine their signs.

a) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 6 & 4 & 5 \end{bmatrix}$$

c) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 4 & 1 & 6 & 2 \end{bmatrix}$$

e) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 3 & 1 & 2 & 4 & 6 \end{bmatrix}$$

g) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 4 & 5 & 6 & 3 \end{bmatrix}$$

i) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 6 & 5 & 3 & 1 \end{bmatrix}$$

b) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 6 & 1 & 5 & 3 & 4 \end{bmatrix}$$

d) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 3 & 6 & 1 & 2 & 5 \end{bmatrix}$$

f) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 5 & 4 & 3 & 6 & 1 \end{bmatrix}$$

h) 
$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 1 & 2 & 6 \end{bmatrix}$$

4. In each one of the cases of Exercise 1, write the inverse of the permutation.
5. In each one of the cases of Exercise 2, write the inverse of the permutation.
6. In each one of the cases of Exercise 3, write the inverse of the permutation.
7. In each one of the cases of Exercise 1, write the orbit decomposition of the permutation.
8. In each one of the cases of Exercise 2, write the orbit decomposition of the permutation.
9. In each one of the cases of Exercise 3, write the orbit decomposition of the permutation.
10. Prove that the number of odd permutations of  $J_n$  for  $n \geq 2$  is equal to the number of even permutations. [Hint: Let  $\sigma_1, \dots, \sigma_m$  be all the distinct even permutations. Let  $\tau$  be a transposition. Prove that

$$\tau\sigma_1, \dots, \tau\sigma_m$$

are all distinct, and constitute all the odd permutations.]

11. After you have read the section on induction, prove that the number of permutations of  $J_n$  is equal to  $n!$ . [Hint: By induction. It is true for  $n = 1$ . Assume it for  $n$ . Let  $\tau_k$  ( $k = 1, \dots, n$ ) be the transposition which interchanges  $n + 1$  and  $k$  for  $k = 1, \dots, n$ . Let  $S_n$  be the set of permutations of  $J_n$ . Show that the permutations

$$I\sigma, \tau_1\sigma, \dots, \tau_n\sigma$$

for  $\sigma$  in  $S_n$  give all distinct permutations of  $J_{n+1}$ . Hence the number of elements in  $S_{n+1}$  is equal to  $n + 1$  times the number of elements in  $S_n$ , namely  $(n + 1)n! = (n + 1)!$ .]

12. In Exercises 1, 2, and 3, find the sign of the permutation by the orbit method of the last two examples of this section, and Theorem 4. Which do you think is the faster way of determining the sign of the permutation? This way, or the old way, expressing  $\sigma$  as explicitly as a product of transpositions?

# 15 Complex Numbers

## §1. THE COMPLEX PLANE

The set of complex numbers is a set whose elements can be added and multiplied, so that the sum of complex numbers is a complex number, the product of complex numbers is a complex number, and so that addition and multiplication satisfy the following properties:

*Addition is commutative and associative.*

*Multiplication is commutative, associative, and distributive with respect to addition.*

*Every real number is a complex number, and if  $a, b$  are real numbers, then their sum and product as complex numbers are the same as their sum and product as real numbers, respectively.*

*If 1 is the real number one, then  $1z = z$  for every complex number  $z$ . Similarly,  $0z = 0$ .*

*Each complex number  $z$  has an additive inverse, namely  $(-1)z$ , so that*

$$z + (-1)z = 0.$$

*There exists a complex number  $i$  such that  $i^2 = -1$ .*

*Every complex number can be written in the form*

$$a + bi,$$

*where  $a, b$  are real numbers.*

Thus in the complex numbers we can take a square root of  $-1$ . Remember the arguments we gave when we discussed square roots in the real numbers. These arguments apply here, to show that

$$i, \quad -i$$

are the only two complex numbers whose square is  $-1$ .

Just as we represented real numbers on a line, we can represent complex numbers in the plane. Namely, if  $z = a + bi$  is a complex number, we view  $z$  as the point  $(a, b)$  in the plane. Thus  $i$  is represented by the point  $(0, 1)$  as shown in Fig. 15-1.

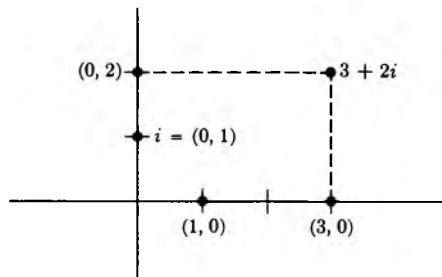


Fig. 15-1

We could, in fact, define a complex number to be a pair of real numbers  $(a, b)$ , and define  $i$  to be the pair  $(0, 1)$ . Then we could define addition of complex numbers just as we defined addition of points in the plane, componentwise. Thus if

$$z = (x, y) \quad \text{and} \quad w = (u, v)$$

with  $x, y, u, v$  real, then

$$z + w = (x + u, y + v).$$

With this definition, we identify a real number  $x$  with the point  $(x, 0)$  in the plane. If  $c$  is a real number, we keep our old definition of multiplication, and define

$$cz = (cx, cy).$$

Then we see that every complex number  $z$  as above can be written as

$$x + iy.$$

We would still have to define multiplication of complex numbers. If we assume that complex numbers exist satisfying our list of properties, then the product of two complex numbers

$$z = x + iy \quad \text{and} \quad w = u + iv$$

is given by

$$\begin{aligned} (x + iy)(u + iv) &= x(u + iv) + iy(u + iv) \\ &= xu + ixv + iyu + i^2yv \\ &= xu - yv + (xv + yu)i. \end{aligned}$$

If we want to define the multiplication of points

$$(x, y)(u, v)$$

to satisfy the properties listed above, then we must take the definition that this product is equal to the point

$$(xu - yv, xv + yu).$$

We could then verify by brute force that all the properties listed above are valid, having defined multiplication this way. This would be boring, so we omit it, and just assume all the properties.

From now on, we assume that complex numbers exist, and can be represented as points in the plane. We identify a real number  $a$  with the point  $(a, 0)$ . For this reason, the plane is sometimes called the **complex plane**.

In multiplying complex numbers, we use the rule  $i^2 = -1$  to simplify a product and to put it in the form  $a + bi$ .

**Example.** Let

$$z = 2 + 3i \quad \text{and} \quad w = 1 - i.$$

Then

$$\begin{aligned} zw &= (2 + 3i)(1 - i) = 2(1 - i) + 3i(1 - i) \\ &= 2 - 2i + 3i - 3i^2 \\ &= 2 + 3 + i \\ &= 5 + i. \end{aligned}$$

Let  $z = a + bi$  be a complex number. We define its **complex conjugate** (or simply its **conjugate**) to be the complex number

$$\bar{z} = a - bi.$$

Thus if  $z = 2 + 3i$ , then  $\bar{z} = 2 - 3i$ . We see at once that

$$z\bar{z} = a^2 + b^2.$$

Hence viewing  $z$  as a point in the complex plane, we see that  $z\bar{z}$  is the square of the distance of this point from the origin  $(0, 0)$ .

Let  $z = a + bi$  be a complex number  $\neq 0$ . Let

$$w = \frac{\bar{z}}{a^2 + b^2}.$$

Then, using the rules for multiplication, we obtain

$$zw = wz = 1,$$

because

$$z \frac{\bar{z}}{a^2 + b^2} = \frac{z\bar{z}}{a^2 + b^2} = 1.$$

This number  $w$  is called the (**multiplicative**) **inverse** of  $z$ , and is denoted by  $z^{-1}$ , or  $1/z$ . The proof given before that the multiplicative inverse is unique applies now. In other words, there exists one and only one complex number  $w$  such that  $wz = 1$ .

If  $z, w$  are complex numbers and  $z \neq 0$ , then we write  $w/z$  instead of  $z^{-1}w$ , just as we did with real numbers.

**Example.** To find the inverse of  $1 + i$ , we note that the conjugate of  $1 + i$  is  $1 - i$ , and that

$$(1 + i)(1 - i) = 2.$$

Hence

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

**Theorem 1.** Let  $z, w$  be complex numbers. Then

$$\overline{zw} = \bar{z} \overline{w}, \quad \overline{z + w} = \bar{z} + \bar{w}, \quad \overline{\bar{z}} = z.$$

*Proof.* The proofs follow immediately from the definitions of addition, multiplication, and the complex conjugate. We leave them to you as exercises.

Let  $z = x + yi$  be a complex number, with real  $x, y$ . We shall call  $x$  the **real part** of  $z$ , and  $y$  the **imaginary part** of  $z$ . These are denoted by  $\text{Re}(z)$  and  $\text{Im}(z)$ , respectively.

Thus if  $z = x + iy$ , then

$$z + \bar{z} = 2x = 2\text{Re}(z).$$

Similarly,

$$z - \bar{z} = 2y = 2\text{Im}(z).$$

We define the **absolute value** of a complex number  $z = x + iy$  to be

$$|z| = \sqrt{x^2 + y^2}.$$

Thus the absolute value of  $z$  is simply the distance from the origin to the point  $(x, y)$  in the plane. In terms of the absolute value, we can write

$$z^{-1} = \frac{\bar{z}}{|z|^2},$$

provided  $z \neq 0$ . Indeed, we observe that

$$z\bar{z} = |z|^2.$$

**Theorem 2.** *The absolute value of complex numbers satisfies the following properties. If  $z, w$  are complex numbers, then*

$$\begin{aligned}|zw| &= |z| |w| \\ |z + w| &\leq |z| + |w|.\end{aligned}$$

*Proof.* We have:

$$|zw|^2 = zw\bar{z}\bar{w} = z\bar{z}w\bar{w} = |z|^2|w|^2.$$

Taking the square root we conclude that  $|zw| = |z| |w|$ , thus proving the first assertion. As for the second, it is simply the triangle inequality. A simple algebraic proof can be given, but we omit it. Draw the picture of the points  $z, w, z + w$  in the plane to see the parallelogram and the triangle.

## EXERCISES

1. Express the following complex numbers in the form  $x + iy$  where  $x$  and  $y$  are real numbers.
  - a)  $(-1 + 3i)^{-1}$
  - b)  $(1 + i)(1 - i)$
  - c)  $(1 + i)(2 - i)$
  - d)  $(i - 1)(2 - i)$
  - e)  $(7 + \pi i)(\pi + i)$
  - f)  $(2i + 1)\pi i$
  - g)  $(\sqrt{2}i)(\pi + 3i)$
  - h)  $(i + 1)(i - 2)(i + 3)$
2. Express the following complex numbers in the form  $x + iy$ , where  $x, y$  are real numbers.
  - a)  $(1 + i)^{-1}$
  - b)  $\frac{1}{3+i}$
  - c)  $\frac{2+i}{2-i}$
  - d)  $\frac{1}{2-i}$
  - e)  $\frac{1+i}{i}$
  - f)  $\frac{i}{1+i}$
  - g)  $\frac{2i}{3-i}$
  - h)  $\frac{1}{-1+i}$
3. Let  $z$  be a complex number  $\neq 0$ . What is the absolute value of  $z/\bar{z}$ ?
4. Prove the statements of Theorem 1.
5. Show that for any complex number  $z = x + iy$ , with  $x, y$  real, we have

$$\operatorname{Im}(z) \leq |\operatorname{Im}(z)| \leq |z|.$$

## §2. POLAR FORM

Let  $z = x + iy$  be a complex number, which we view as a point  $(x, y)$  in the plane. Let  $r = |z|$ . Then

$$\frac{z}{|z|} = \frac{1}{r} z$$

has absolute value 1, and consequently can be viewed as a point on the circle of radius 1 centered at the origin. Hence there exists some real number  $\theta$  such that

$$\frac{z}{r} = (\cos \theta, \sin \theta).$$

This shows that we can write

$$z = r(\cos \theta + i \sin \theta).$$

Every complex number can be written as a product of a real number  $\geq 0$  and a complex number of absolute value 1.

**Example.** Let  $z = 1 + i\sqrt{3}$ . Then  $|z| = 2$ . Hence

$$z = 2(\cos \theta, \sin \theta).$$

Note that  $\cos \theta = \frac{1}{2}$  and  $\sin \theta = \sqrt{3}/2$ . Hence  $\theta = \pi/3$ . We may say that the polar coordinates of  $z$  are  $(2, \pi/3)$ .

We define the symbols  $e^{i\theta}$  by

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

Then we have

$$z = x + iy = re^{i\theta}.$$

This is called the **polar form** of  $z$ , and is illustrated in Fig. 15-2. The notation  $e^{i\theta}$  would be terrible if the next theorem were not true. But since the next theorem is true, the notation is quite good.

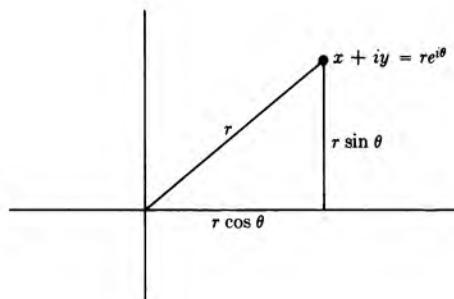


Fig. 15-2

**Theorem 3.** Let  $\theta, \varphi$  be real numbers. Then

$$e^{i\theta} e^{i\varphi} = e^{i(\theta+\varphi)}.$$

*Proof.* We shall see that this amounts to the addition formula for the sine and cosine. By definition, we have

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

Using the addition formula for the sine and cosine, we see that the preceding expression is equal to

$$\cos \theta \cos \varphi - \sin \theta \sin \varphi + i(\sin \theta \cos \varphi + \sin \varphi \cos \theta).$$

This is exactly the same complex number that we get by multiplying

$$(\cos \theta + i \sin \theta)(\cos \varphi + i \sin \varphi).$$

This proves Theorem 3.

Thus exponentiation with a complex number of the form  $i\theta$  obeys the same basic rule as ordinary exponentiation.

Let  $z$  and  $w$  be complex numbers. Let us write them in polar form; namely

$$z = r e^{i\theta} \quad \text{and} \quad w = s e^{i\varphi},$$

where  $r, s, \theta, \varphi$  are real numbers. Then their product is equal to

$$zw = r e^{i\theta} s e^{i\varphi} = rs e^{i(\theta+\varphi)}.$$

Thus, to multiply complex numbers, we may say roughly that we multiply their absolute values and add their angles.

**Example.** Find a complex number whose square is  $4e^{i\pi/2}$ .

Let  $z = 2e^{i\pi/4}$ . Then, using Theorem 3, we find that

$$z^2 = 4e^{i\pi/2}.$$

**Example.** Let  $n$  be a positive integer. Find a complex number  $w$  such that  $w^n = e^{i\pi/2}$ .

It is clear that the complex number  $w = e^{i\pi/2n}$  satisfies our requirements.

**Example.** We have

$$e^{i\pi} = -1 \quad \text{and} \quad e^{2\pi i} = 1.$$

This follows at once from the definition and the known values of the sine and cosine.

## EXERCISES

1. Put the following complex numbers in polar form.
 

a) $1 + i$	b) $\sqrt{3} + i$	c) $-3$	d) $4i$
e) $1 - i\sqrt{3}$	f) $-5i$	g) $-7$	h) $-1 - i$
2. Put the following complex numbers in the ordinary form  $x + iy$ . Also plot them as points in the plane.
 

a) $e^{3\pi i}$	b) $e^{2\pi i/3}$	c) $3e^{i\pi/4}$	d) $e^{-i\pi/3}$
e) $e^{2\pi i/6}$	f) $e^{-i\pi/2}$	g) $e^{-i\pi}$	h) $e^{-5i\pi/4}$
3. Let  $z$  be a complex number  $\neq 0$ . Show that there are precisely two distinct complex numbers whose square is  $z$ .
4. Let  $z$  be a complex number  $\neq 0$ . Let  $n$  be a positive integer. Show that there are  $n$  distinct complex numbers  $w$  such that  $w^n = z$ . Write these complex numbers in polar form. The proof given that a polynomial of degree  $\leq n$  has at most  $n$  roots applies to the complex case, and thus we see that there are no other complex numbers  $w$  such that  $w^n = z$  other than those you have presumably written down.
5. Write in polar form the  $n$  complex numbers  $w$  such that  $w^n = 1$ . Plot all of these as points in the plane for  $n = 2, 3, 4, 5$ .
6. If  $\theta$  is real, show that

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \quad \text{and} \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

# 16 Induction and Summations

## §1. INDUCTION

Induction is an axiom which allows us to give proofs that certain properties are true for all integers.

Suppose that we wish to prove a certain assertion concerning positive integers  $n$ . Let  $A(n)$  denote the assertion concerning the integer  $n$ . To prove it for all  $n$ , it suffices to prove the following.

**IND 1.** *The assertion  $A(1)$  is true (i.e. the assertion concerning the integer 1 is true).*

**IND 2.** *Assuming the assertion proved for all positive integers  $\leq n$ , prove it for  $n + 1$ , i.e. prove that  $A(n + 1)$  is true.*

The combination of these two steps is known as induction. For instance, **IND 1** gives us a starting point, and **IND 2** allows us to prove  $A(2)$  from  $A(1)$ , then  $A(3)$  from  $A(2)$ , and so forth, proceeding stepwise.

We shall now give examples.

**Example.** For all integers  $n \geq 1$ , we have

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

*Proof.* By induction. The assertion  $A(n)$  is the assertion of the theorem. When  $n = 1$ , it simply states that

$$1 = \frac{1(1 + 1)}{2},$$

which is clearly true.

Assume now that the assertion is true for the integer  $n$ . Then

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

Putting the expression on the right of the equality sign over a common denominator 2, we see that it is equal to

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

Hence assuming  $A(n)$ , we have proved that

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

which is none other than assertion  $A(n+1)$ . This proves our result.

**Notational remark.** We have just written the sum of the first  $n$  positive integers above using three dots to denote intermediate integers. There is a notation which avoids the use of such dots, using a capital Greek sigma sign ( $\sum$  for sum). Thus we write

$$\sum_{k=1}^n k = 1 + 2 + \cdots + n.$$

Similarly, we write

$$\sum_{k=1}^n k^2 = 1^2 + 2^2 + \cdots + n^2,$$

or also

$$\sum_{k=1}^n \sin k = \sin 1 + \sin 2 + \cdots + \sin n.$$

If  $f$  is any function, we would write

$$\sum_{k=1}^n f(k) = f(1) + f(2) + \cdots + f(n).$$

We can write the distributivity with respect to a sum of  $n$  terms with the sigma notation as follows. For any number  $c$ , we have

$$c \sum_{k=1}^n f(k) = \sum_{k=1}^n cf(k).$$

For instance,

$$5 \sum_{k=1}^n k^3 = \sum_{k=1}^n 5k^3.$$

Expanding with use of dots, we can also write this as

$$5(1^3 + 2^3 + \cdots + n^3) = 5 \cdot 1^3 + 5 \cdot 2^3 + \cdots + 5 \cdot n^3.$$

**Example.** Let  $f$  be a function defined for all real numbers such that

$$f(x + y) = f(x)f(y)$$

for all numbers  $x, y$ . Let  $f(1) = a$ . We want to prove by induction that  $f(n) = a^n$  for all positive integers  $n$ . This is true for  $n = 1$ . Assume the result for an integer  $n$ . Then

$$f(n + 1) = f(n)f(1) = a^n a = a^{n+1},$$

thereby proving our assertion.

**Example.** We want a simple formula for the number of ways of selecting  $k$  objects out of a set of  $n$  objects. This number is denoted by  $C_k^n$  and is called a **binomial coefficient**. The reason for this name is simple. Suppose that we wish to expand the product

$$(x + y)^n = (x + y)(x + y) \cdots (x + y)$$

as a sum of terms involving powers of  $x$  and  $y$ . In this expansion, we select  $x$  from  $k$  factors, and we select  $y$  from  $n - k$  factors. We then take the sum over all possible such selections. Thus we find that

$$(x + y)^n = \sum_{k=0}^n C_k^n x^k y^{n-k}.$$

It turns out that  $C_k^n$  has a simple expression. Let  $n!$  denote the product of the first  $n$  integers, so that

$$\begin{aligned} 1! &= 1, \\ 2! &= 1 \cdot 2 = 2, \\ 3! &= 1 \cdot 2 \cdot 3 = 6, \\ 4! &= 1 \cdot 2 \cdot 3 \cdot 4 = 24, \\ 5! &= 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120, \end{aligned}$$

and so on. By convention, we *define*  $0! = 1$ . Then the value for  $C_k^n$  is given by

$$C_k^n = \frac{n!}{k!(n - k)!}.$$

For instance,

$$C_2^4 = \frac{4!}{2!(4 - 2)!} = \frac{24}{2 \cdot 2} = 6.$$

It is very simple to prove by induction that  $C_k^n$  has the value described above, and we shall let you have fun with that, in Exercise 9. Instead of the notation

$C_k^n$ , one also uses the notation

$$\binom{n}{k}$$

for the binomial coefficient.

As a matter of convenience, if  $n$  is an integer  $\geq 0$  and  $k$  is an integer such that  $k < 0$  or  $k > n$ , then we define

$$\binom{n}{k} = 0.$$

## EXERCISES

1. Prove that, for all integers  $n \geq 1$ , we have

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

2. Prove that, for all integers  $n \geq 1$ , we have

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

b)  $1^3 + 2^3 + 3^3 + \cdots + n^3 = \left[\frac{n(n + 1)}{2}\right]^2$ .

3. Prove that

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

4. Prove that  $n(n^2 + 5)$  is divisible by 6 for all integers  $n \geq 1$ .

5. Prove that, for  $x \neq 1$ , we have

$$(1 + x)(1 + x^2)(1 + x^4) \cdots (1 + x^{2^n}) = \frac{1 - x^{2^{n+1}}}{1 - x}.$$

6. Let  $f$  be a function defined for all real numbers such that

$$f(xy) = f(x) + f(y)$$

for all real numbers  $x, y$ . Show that

$$f(x^n) = nf(x)$$

for all  $x$ .

7. Let  $f$  be a function defined for all numbers such that

$$f(xy) = f(x)f(y)$$

for all real numbers  $x, y$ . Show that  $f(x^n) = f(x)^n$  for all positive integers  $n$  and all real numbers  $x$ .

8. Using Exercises 1, 2, and 3, write out simple expressions giving the values for the following sums.

a)  $\sum_{k=1}^{2n} k^2$

b)  $\sum_{k=1}^{2n} k^3$

c)  $\sum_{k=1}^{2n} (2k - 1)$

d)  $\sum_{k=1}^{m-1} k^2$

e)  $\sum_{k=1}^{m-1} k^3$

f)  $\sum_{k=1}^m (2k - 1)$

9. **Binomial coefficients.** Let

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

where  $n, k$  are integers  $\geq 0$ ,  $0 \leq k \leq n$ , and  $0!$  is defined to be 1. Prove the following assertions.

a)  $\binom{n}{k} = \binom{n}{n-k}$

b)  $\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}$  (for  $k > 0$ )

- c) Prove by induction that for all numbers  $x, y$ , we have

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}.$$

10. Write down the numerical values for the following binomial coefficients:

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

$$\binom{4}{0}, \binom{4}{1}, \binom{4}{2}, \binom{4}{3}, \binom{4}{4},$$

$$\binom{5}{0}, \binom{5}{1}, \binom{5}{2}, \binom{5}{3}, \binom{5}{4}, \binom{5}{5}.$$

11. Write out in full the expansions of:

a)  $(x+y)^3$

b)  $(x+y)^4$

c)  $(x+y)^5$

Observe that you might already have done these in Chapter 1, §3.

**12. Theorem.** All billiard balls have the same color.

*Proof:* By induction, on the number  $n$  of billiard balls. Our theorem is certainly true for  $n = 1$ , i.e. for one billiard ball. Assume it for  $n$  billiard balls. We prove it for  $n + 1$ . Look at the first  $n$  billiard balls among those  $n + 1$ . By induction, they have the same color. Now look at the last  $n$  among those  $n + 1$ . They have the same color. Hence all  $n + 1$  have the same color (Fig. 16-1).

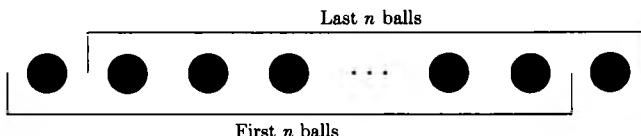


Fig. 16-1

What is wrong with this “proof”?

13. Let  $E$  be a set with  $n$  elements, and let  $F$  be a set with  $m$  elements. Show that the total number of mappings from  $E$  to  $F$  is  $m^n$ . [Hint: Use induction on  $n$ . First take  $n = 1$ . How many ways are there of mapping a single element into a set with  $m$  elements? Then assume the result for  $n$ , and prove it for  $n + 1$ , taking into account the number of possible ways of mapping the  $(n + 1)$ -th element of  $E$  into  $F$ .]

## §2. SUMMATIONS

We have already met simple cases of summations in the preceding section, when we computed the values of certain sums. We shall give here other applications of such summations.

We go back to the old problem of computing volumes, similar to the problem of computing areas and lengths discussed in Chapter 7. We shall treat a typical example, and let you work out others along similar lines. Our purpose is to compute the volume of a cone, of a ball, and more generally of a solid obtained by revolving a curve around an axis, which we take to be the  $x$ -axis.

We work out the volume of a cone in detail. In fact, we consider a special cone as shown on Fig. 16-2.

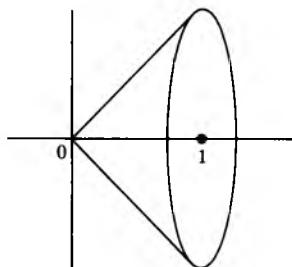


Fig. 16-2

We suppose that the height of the cone is equal to 1, and that the radius of the base is equal to 1 also. Then the cone is obtained by revolving the line  $y = x$  around the  $x$ -axis, between the values  $x = 0$  and  $x = 1$ .

We wish to prove that the volume of this cone is equal to  $\pi/3$ . We use the method discovered by Archimedes. We first observe that the volume of a cylinder whose base has radius  $r$  and of height  $h$  as shown in Fig. 16-3 is  $\pi r^2 h$ .

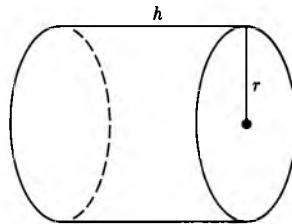


Fig. 16-3

Using this, we approximate the cone by cylinders as follows. Cut up the interval on the  $x$ -axis between 0 and 1 into  $n$  segments of equal length  $1/n$ . The end points of these segments are the rational numbers

$$0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, \frac{n}{n} = 1.$$

Using general notation, we can say that the small segments are the segments

$$\left[ \frac{k}{n}, \frac{k+1}{n} \right],$$

for  $k = 0, \dots, n$ . We use the notation  $[a, b]$  to denote the set of all numbers  $x$  such that  $a \leq x \leq b$ .

We draw the cylinder whose height is  $1/n$  and whose base is the disc of radius  $k/n$  centered at the point

$$\left(\frac{k}{n}, 0\right)$$

as shown in Fig. 16-4(a) and (b).

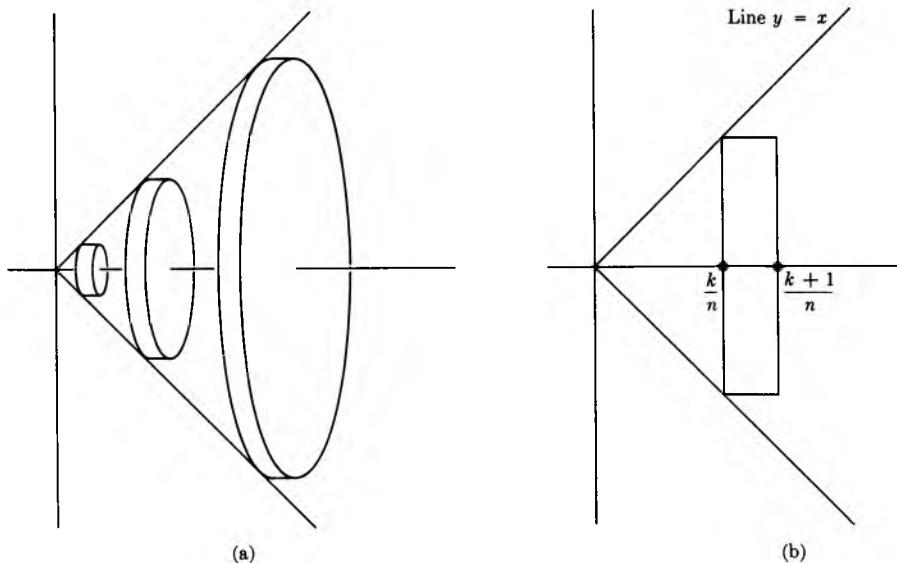


Fig. 16-4

Fig. 16-4(b) is a cross section of Fig. 16-4(a), viewed from the side. The volume of each small cylinder is equal to

$$(1) \quad \pi \frac{1}{n} \left(\frac{k}{n}\right)^2.$$

Hence the sum of the volumes of the small cylinders is equal to

$$(2) \quad \sum_{k=0}^n \pi \frac{1}{n} \left(\frac{k}{n}\right)^2 = \frac{\pi}{n^3} \sum_{k=0}^n k^2.$$

Of course, in the present case, the term in this sum corresponding to  $k = 0$  is equal to 0, so our sum is really equal to the sum

$$(3) \quad \frac{\pi}{n^3} \sum_{k=1}^n k^2.$$

If we use Exercise 2 of §1, then we get a simple expression for the value of

the sum, namely

$$(4) \quad \frac{\pi}{n^3} \frac{1}{6} n(n+1)(2n+1),$$

which is equal to

$$(5) \quad \frac{\pi}{6} 1 \left(1 + \frac{1}{n}\right) \left(2 + \frac{1}{n}\right).$$

You see this by dividing each expression  $n, n+1, 2n+1$  by  $n$ , thus using up the  $n^3$  in the denominator.

Now as we let  $n$  get arbitrarily large, we see that the expression of (5) approaches the volume of the cone, and also approaches

$$\frac{\pi}{6} \cdot 2 = \frac{\pi}{3}.$$

This achieves what we wanted.

To get the volume of an arbitrary cone, we can proceed in two ways. Suppose that the cone has a base whose radius is  $r$ , and let the height be  $h$ . Thus we draw the cone as in Fig. 16-5.

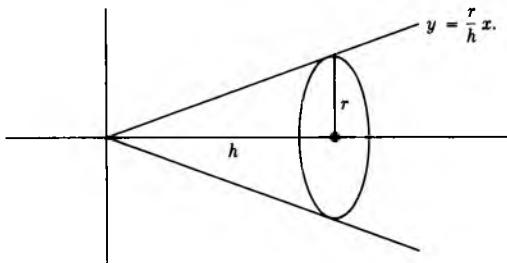


Fig. 16-5

We see that the cone is obtained by rotating the line

$$y = \frac{r}{h} x$$

around the  $x$ -axis, between the values  $x = 0$  and  $x = h$ . We could either follow the same method we used above, decomposing the segment  $[0, h]$  into  $n$  segments of length  $h/n$ , forming the cylinders as before, taking the sum, and then taking the limit. We leave this as Exercise 1. We would then get:

*The volume of a cone whose base has radius  $r$  and whose height is  $h$  is equal to  $\frac{1}{3}\pi r^2 h$ .*

But we can also proceed using mixed dilations, following the ideas of the exercises of Chapter 7, §1. Let us denote by  $C(r, h)$  the cone whose base has radius  $r$  and whose height is  $h$ . It should be intuitively clear (and we shall

give a coordinate argument below) that  $C(r, h)$  is obtained by a mixed dilation of the cone  $C(1, 1)$  whose volume was obtained above. Namely, a point in 3-space is described by three coordinates  $(x, y, z)$ . If  $a, b, c$  are three positive numbers, we let

$$F_{a,b,c} : \mathbf{R}^3 \rightarrow \mathbf{R}^3$$

be the map such that

$$F_{a,b,c}(x, y, z) = (ax, by, cz).$$

Then the cone  $C(r, h)$  is the image of the cone  $C(1, 1)$  under the mixed dilation

$$F_{h,r,r}.$$

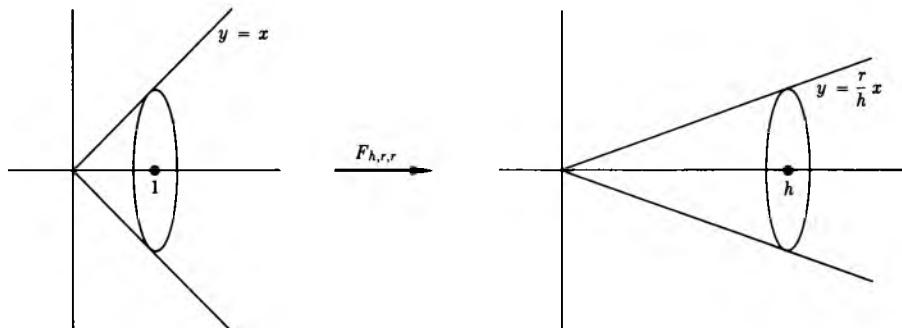


Fig. 16-6

Referring back to Chapter 7, suppose that we have a mixed dilation  $F_{a,b,c}$  as above. If  $R$  is a rectangular solid whose sides have lengths  $t, u, v$ , then the dilation of  $R$  by  $F_{a,b,c}$  is a rectangular solid whose sides have lengths  $at, bu, cv$ . The volume of  $R$  is  $tuv$ , and the volume of the dilated solid  $F_{a,b,c}(R)$  is  $atbucv = abc t u v$ . In other words, the volume gets multiplied by  $abc$  under the mixed dilation. If  $S$  is an arbitrary solid in 3-space, and if we approximate  $S$  by rectangular solids, then we conclude that a similar result holds for  $S$ . In other words, if  $V$  is the volume of  $S$ , then

$$\text{volume of } F_{a,b,c}(S) = abcV.$$

In the application to the cone, we have  $a = h$ ,  $b = r$ ,  $c = r$ , so that

$$abc = r^2h.$$

We can then conclude that the volume of  $C(r, h)$  is equal to  $hr^2$  times the volume of  $C(1, 1)$ . Thus we obtain the value

$$\text{Volume of } C(r, h) = \frac{1}{3} \pi r^2 h.$$

Let us now look more precisely at the reason why  $C(r, h)$  is the dilation of the cone  $C(1, 1)$  by the mixed dilation  $F_{h,r,r}$ . First we must express the cone in terms of coordinates. Let  $(x, y, z)$  be the coordinates of a point in 3-space. If this point lies in the cone  $C(r, h)$ , then

$$0 \leq x \leq h.$$

For each such value of  $x$ , a point  $(x, y, z)$  lies in the cone if and only if

$$y^2 + z^2 \leq \left(\frac{r}{h}x\right)^2.$$

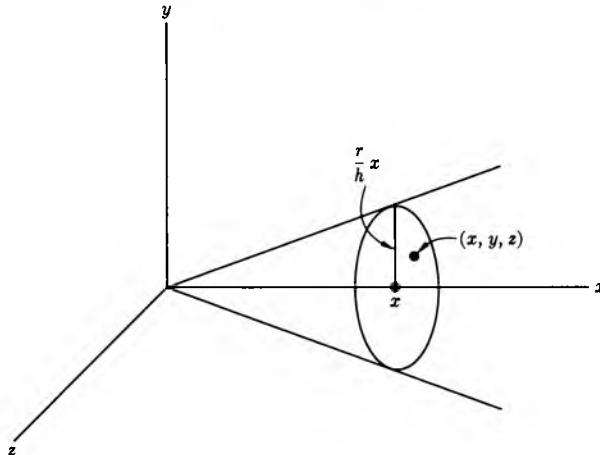


Fig. 16-7

Thus the points on the cone  $C(r, h)$  can be described as those points  $(x, y, z)$  such that

$$( * ) \quad \begin{aligned} 0 &\leq x \leq h, \\ \left(\frac{y}{r}\right)^2 + \left(\frac{z}{r}\right)^2 &\leq \left(\frac{x}{h}\right)^2. \end{aligned}$$

Let

$$x' = \frac{x}{h}, \quad y' = \frac{y}{r}, \quad z' = \frac{z}{r}.$$

If  $(x, y, z)$  satisfy (\*), i.e. lie in  $C(r, h)$ , then  $(x', y', z')$  satisfy

$$(**) \quad \begin{aligned} 0 &\leq x'^2 \leq 1, \\ y'^2 + z'^2 &\leq x'^2. \end{aligned}$$

Hence  $(x', y', z')$  is a point in the cone  $C(1, 1)$ . Conversely, it is also clear that if  $(x', y', z')$  is a point in  $C(1, 1)$ , then  $(x, y, z)$  is a point of  $C(r, h)$ . Hence  $C(r, h)$  is the image of  $C(1, 1)$  under the mixed dilation

$$F_{h,r,r}.$$

This concludes our argument.

## EXERCISES

1. Get the formula for the volume of a cone by approximating an arbitrary cone with cylinders.
2. Rotate the curve  $y = 3x$  about the  $x$ -axis. What is the volume of the solid obtained
  - a) when  $0 \leq x \leq 2$ ?
  - b) when  $0 \leq x \leq 5$ ?
  - c) when  $0 \leq x \leq c$  with an arbitrary positive number  $c$ ?
3. Rotate the curve  $y = \sqrt{x}$  about the  $x$ -axis. What is the volume of the solid obtained when  $0 \leq x \leq h$  and  $h$  has the value:
  - a)  $h = 1$ ?
  - b)  $h = 2$ ?
  - c)  $h = 3$ ?
  - d) arbitrary  $h$ ?
4. Rotate the curve  $y = \sqrt{r^2 - x^2}$  about the  $x$ -axis. What is the volume of the solid obtained when  $0 \leq x \leq r$  and  $r$  has the value:
  - a)  $r = 1$ ,
  - b)  $r = 2$ ,
  - c)  $r = 3$ ,
  - d)  $r$  is arbitrary?

5. Look again at Exercise 4. What is the solid obtained? You should now be able to see that the volume of a spherical ball of radius  $r$  is equal to

$$\frac{4}{3} \pi r^3.$$

6. Find the area between the curve  $y = x^2$  and the  $x$ -axis, from the origin to the following values of  $x$ :

- a)  $x = 1$ ,      b)  $x = 2$ ,      c)  $x = 3$ ,      d)  $x = 4$ .

Use a sum of areas of small rectangles as indicated in Fig. 16-8.

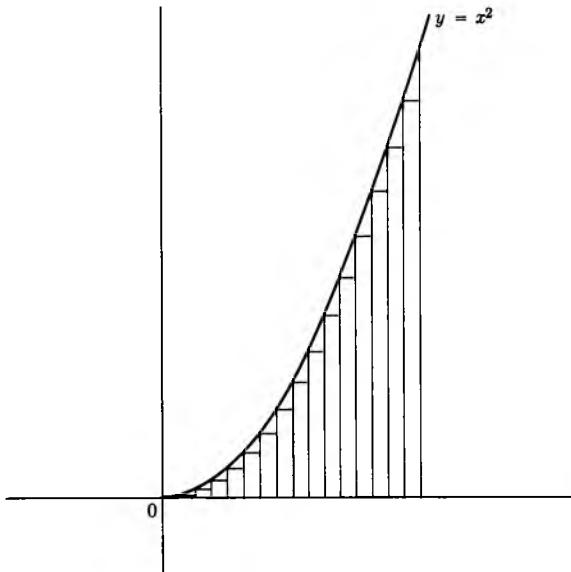


Fig. 16-8

7. Find the area between the curve  $y = x^3$  and the  $x$ -axis between the origin and the following values of  $x$ :

- a)  $x = 1$ ,      b)  $x = 2$ ,      c)  $x = 3$ ,      d)  $x = 4$ .

8. Let  $S$  be the region in the plane consisting of all points  $(x, y)$  such that

$$0 \leq x \leq 1 \quad \text{and} \quad 0 \leq y \leq x^2.$$

Let  $T$  be the region in the plane consisting of all points  $(x, y)$  such that

$$0 \leq x \leq 1 \quad \text{and} \quad 0 \leq y \leq 3x^2.$$

Express  $T$  as the image of  $S$  under a mixed dilation. What is the area of  $T$ ?

9. Let  $c$  be a number  $> 0$ . What is the area of the region lying between the curve  $y = cx^2$  and the  $x$ -axis, from the origin to the following values of  $x$ :

- a)  $x = 1$ ,      b)  $x = 2$ ,      c)  $x = 3$ .

### §3. GEOMETRIC SERIES

Let  $c$  be a number with  $c \neq 1$ . We consider the sum

$$\sum_{k=0}^n c^k = 1 + c + c^2 + \cdots + c^n.$$

Observe that

$$(1 + c + c^2 + \cdots + c^n)(1 - c) = 1 - c^{n+1}.$$

This is immediate, using distributivity. There is a cancellation of all terms in the product, except  $1 - c^{n+1}$ . Thus we have

$$\sum_{k=0}^n c^k = \frac{1}{1 - c} - \frac{c^{n+1}}{1 - c}.$$

What happens when  $n$  becomes very large? If  $c > 1$ , then the power  $c^{n+1}$  becomes large, and we don't regard this as interesting.

Suppose however that  $0 \leq c < 1$ . Then  $c^{n+1}$  approaches 0. For instance, suppose that  $c = \frac{1}{2}$ . Then

$$c^2 = \frac{1}{4}, \quad c^3 = \frac{1}{8}, \quad c^4 = \frac{1}{16}, \quad \dots$$

We see that the denominator is an increasing power of 2, and that the fraction for  $c^{n+1}$  thus approaches 0. Hence

$$\frac{c^{n+1}}{1 - c}$$

approaches 0 as  $n$  becomes arbitrarily large. Thus we may say that the sum

$$1 + c + c^2 + \cdots + c^n$$

approaches

$$\frac{1}{1 - c}$$

as  $n$  becomes arbitrarily large. It is convenient to abbreviate this by the symbols

$$\sum_{k=1}^{\infty} c^k = \frac{1}{1 - c}.$$

The symbol  $\infty$  is called “infinity”, and is explained only in terms of the context we have just introduced. Also, instead of saying “as  $n$  becomes arbitrarily large”, we also say “as  $n$  approaches infinity”.

**Warning.** There is no number called “infinity”. The use of the word infinity is meaningful only as an abbreviation in the context just described.

**Example.** Let  $c = \frac{1}{5}$ . Then

$$1 + \frac{1}{5} + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^3 + \cdots = \frac{1}{1 - \frac{1}{5}} = \frac{5}{4}.$$

Here, we write three dots  $\dots$  instead of using the summation sign.

The symbols

$$\sum_{k=1}^{\infty} c^k$$

are also called the **geometric series**. This symbol has a numerical value when  $0 < c < 1$ , and this value is then  $1/(1 - c)$  as above.

## EXERCISES

1. Find the value of the geometric series for the following values of  $c$ .

- |                  |                  |                  |                  |
|------------------|------------------|------------------|------------------|
| a) $\frac{1}{3}$ | b) $\frac{1}{4}$ | c) $\frac{1}{5}$ | d) $\frac{1}{6}$ |
| e) $\frac{3}{4}$ | f) $\frac{2}{3}$ | g) $\frac{2}{5}$ | h) $\frac{3}{5}$ |

2. Argue in a manner similar to that of the text to give a value to the geometric series when  $-1 < c \leq 0$ . Give the general formula, and also

give the specific numerical values of the geometric series for the following numbers  $c$ .

a)  $-\frac{1}{3}$

b)  $-\frac{1}{4}$

c)  $-\frac{1}{5}$

d)  $-\frac{1}{6}$

e)  $-\frac{3}{4}$

f)  $-\frac{2}{3}$

g)  $-\frac{2}{5}$

h)  $-\frac{3}{5}$

3. Let  $c$  be a complex number such that  $0 \leq |c| < 1$ . Again argue in a similar way to give a value to the geometric series.
4. What is the value of the sum

$$\sum_{k=0}^n c^k,$$

when  $c$  is equal to  $re^{i\theta}$  and  $0 \leq r < 1$ ? Express your answer in the same form as that used to discuss the geometric series. Similarly, give the value of the series

$$\sum_{k=0}^{\infty} c^k,$$

when  $c = re^{i\theta}$  and  $0 \leq r < 1$ .

5. Let  $c = e^{2\pi i/n}$  for some positive integer  $n$ . What is the value of

$$1 + c + c^2 + \cdots + c^{n-1}?$$

6. Let  $c$  be a complex number  $\neq 1$ , such that  $c^n = 1$  for some positive integer  $n$ . What is the value of

$$1 + c + c^2 + \cdots + c^{n-1}?$$

7. Consider the sums

$$\sum_{k=1}^n \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}.$$

Starting with  $\frac{1}{3}$ , group the terms of this sum by taking the first two, then the next four, then the next eight, and so on, by groups of  $2^d$  where  $d = 1, 2, \dots$ . In each such grouping, replace each term by the fraction furthest to the right, i.e. replace 1 by  $\frac{1}{2}$ , then  $\frac{1}{3}$  by  $\frac{1}{4}$ , then  $\frac{1}{5}, \frac{1}{6}, \frac{1}{7}$  by  $\frac{1}{8}$ , then the next group of fractions by  $\frac{1}{16}$ , and so on. In this way, show that these sums can be made to have arbitrarily large values, for sufficiently large  $n$ .

8. Consider the sums

$$\sum_{k=1}^n \frac{1}{k^2} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2}.$$

Group the terms differently, putting together  $\frac{1}{2^2}$  and  $\frac{1}{3^3}$ ; then  $\frac{1}{4^2}$  up to  $\frac{1}{7^2}$ , then  $\frac{1}{8^2}$  up to  $\frac{1}{15^2}$ , and so on. Prove that these sums are  $\leq 2$ , no matter what value  $n$  has.

9. A ball is thrown to a height of 6 ft and falls back down. Each time it rebounds to a height equal to four-fifths of the preceding height. What distance will the ball have traveled after it touches the ground for the
  - a) 3rd time?
  - b) 5th time?
  - c) 20th time?

d) Assuming that the ball goes on bouncing forever according to the above prescription, what distance does it travel?
10. A ball falls from a height of 10 ft and rebounds each time to a height equal to one-fifth of its previous height. What distance will the ball have traveled when it touches the ground for the
  - a) 5th time?
  - b) 20th time?

c) Assuming that the ball goes on bouncing forever according to the above prescription, what distance will it travel?



# 17 Determinants

## §1. MATRICES

An array of four numbers

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is called a  $2 \times 2$  matrix (read “two by two matrix”). For instance

$$\begin{pmatrix} 3 & -5 \\ 2 & 7 \end{pmatrix}$$

is a  $2 \times 2$  matrix. In a sense, a  $2 \times 2$  matrix can be viewed as a generalization of our “pairs”, and this can be done in two ways: by looking at the rows or at the columns of the matrix. In our example, we have two rows,

$$(3, -5) \quad \text{and} \quad (2, 7),$$

as well as two columns,

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -5 \\ 7 \end{pmatrix}.$$

These are called the first and second columns, respectively.

For this chapter, we call a pair of numbers, written either vertically or horizontally, a **vector**. If it is written vertically, we call it a **column vector**. If it is written horizontally, we call it a **row vector**.

**Example.** The row vectors of the matrix

$$\begin{pmatrix} -6 & 8 \\ 5 & -3 \end{pmatrix}$$

are  $(-6, 8)$  and  $(5, -3)$ . The column vectors of this same matrix are

$$\begin{pmatrix} -6 \\ 5 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 8 \\ -3 \end{pmatrix}.$$

We add column vectors componentwise, just as we added row vectors. For instance,

$$\begin{pmatrix} -1 \\ 6 \end{pmatrix} + \begin{pmatrix} 3 \\ -2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

Similarly, we multiply a column vector by a number, componentwise. For instance,

$$-7 \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} -14 \\ -21 \end{pmatrix}.$$

Thus writing pairs of numbers vertically is just a notational convenience to distinguish rows and columns; it does not affect the basic nature of the algebraic operations among them.

For a general  $2 \times 2$  matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

we call

$$\begin{pmatrix} a \\ c \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} b \\ d \end{pmatrix}$$

its **first** and **second column**, respectively. The numbers  $a, b, c, d$  are called the **components** of the matrix. We write the components  $a, b, c, d$  of a  $2 \times 2$  matrix also with subscripts:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

The first subscript indicates the row, and the second subscript indicates the column. For instance,  $a_{12}$  means the component of the matrix in the first row and second column. Similarly,  $a_{22}$  means the component of the second row and second column.

**Example.** In the matrix

$$A = \begin{pmatrix} 3 & -5 \\ 7 & 1 \end{pmatrix},$$

we have  $a_{11} = 3$ ,  $a_{21} = 7$ ,  $a_{12} = -5$ , and  $a_{22} = 1$ .

With this subscript notation, we denote the rows by

$$A_1 = (a_{11}, a_{12}) \quad \text{and} \quad A_2 = (a_{21}, a_{22}).$$

We denote the columns by

$$A^1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} \quad \text{and} \quad A^2 = \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}.$$

Thus we use a superscript like  $A^1$ ,  $A^2$  to denote columns instead of rows in a matrix  $A$ .

We can do similar things with a  $3 \times 3$  matrix, which is an array of numbers as follows.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

This  $3 \times 3$  matrix has three rows denoted by  $A_1$ ,  $A_2$ ,  $A_3$  and three columns, denoted by  $A^1$ ,  $A^2$ ,  $A^3$ . Its **components** are the numbers  $a_{ij}$ , where the indices  $i, j$  range over the integers 1, 2, 3.

**Example.** The three columns of the matrix

$$\begin{pmatrix} 3 & -1 & 4 \\ 2 & 5 & 7 \\ -8 & 2 & 3 \end{pmatrix}$$

are

$$A^1 = \begin{pmatrix} 3 \\ 2 \\ -8 \end{pmatrix}, \quad A^2 = \begin{pmatrix} -1 \\ 5 \\ 2 \end{pmatrix}, \quad A^3 = \begin{pmatrix} 4 \\ 7 \\ 3 \end{pmatrix}.$$

The three rows are

$$A_1 = (3, -1, 4), \quad A_2 = (2, 5, 7), \quad A_3 = (-8, 2, 3).$$

For this matrix, we have  $a_{21} = 2$  and  $a_{12} = -1$ .

Of course, we could now generalize further, and deal with  $4 \times 4$  matrices, etc., but in the rest of the chapter, we study only  $2 \times 2$  and  $3 \times 3$  matrices. For the higher theory, look up my book *Introduction to Linear Algebra*.

It is natural to flip a  $2 \times 2$  matrix or a  $3 \times 3$  matrix across the diagonal, and to change rows into columns, or columns into rows. This process is

called **transposition**. Thus the **transpose** of the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is the matrix

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

To give a numerical example, the transpose of the matrix

$$\begin{pmatrix} 3 & -8 \\ 7 & 15 \end{pmatrix}$$

is the matrix

$$\begin{pmatrix} 3 & 7 \\ -8 & 15 \end{pmatrix}.$$

Similarly, the transpose of a matrix

$$(1) \quad \begin{pmatrix} 4 & -3 & 2 \\ -1 & 5 & 7 \\ 9 & -8 & 14 \end{pmatrix}$$

is the matrix

$$(2) \quad \begin{pmatrix} 4 & -1 & 9 \\ -3 & 5 & -8 \\ 2 & 7 & 14 \end{pmatrix}.$$

The **transpose** of a matrix  $A$  is denoted by  $'A$ . The matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

is sometimes denoted by  $(a_{ij})$ , where the first index  $i$  denotes the row, and the second index  $j$  denotes the column. If  $A = (a_{ij})$ , then its transpose is denoted by

$$'A = (a_{ji}),$$

with a reversal of the indices  $i$  and  $j$ .

With this notation, we call the components  $a_{ii}$  of the matrix its **diagonal components**. Observe that the diagonal components remain unchanged when we transpose a matrix.

**Example.** The diagonal components of the matrix

$$\begin{pmatrix} 2 & 1 & 3 \\ -1 & 5 & 7 \\ -4 & 6 & -8 \end{pmatrix}$$

are

$$a_{11} = 2, \quad a_{22} = 5, \quad a_{33} = -8.$$

### EXERCISES

In each of the following cases, write down the second row and first column of the indicated matrix. Also write down its transpose.

1.  $\begin{pmatrix} 2 & -5 \\ -3 & -7 \end{pmatrix}$

2.  $\begin{pmatrix} 3 & -7 \\ 8 & 1 \end{pmatrix}$

3.  $\begin{pmatrix} -4 & 6 \\ 5 & -9 \end{pmatrix}$

4.  $\begin{pmatrix} 3 & 5 & 6 \\ -1 & 2 & 3 \\ 7 & 3 & -2 \end{pmatrix}$

5.  $\begin{pmatrix} -1 & 3 & -4 \\ 2 & 1 & 6 \\ 5 & -8 & -2 \end{pmatrix}$

6.  $\begin{pmatrix} -3 & 4 & 6 \\ -2 & -1 & -7 \\ \frac{1}{2} & 3 & \frac{1}{3} \end{pmatrix}$

7. Find the sum of the first two columns in the matrix of Exercise 4.
8. Find the sum of the second and third rows in the matrix of
  - a) Exercise 4,
  - b) Exercise 5,
  - c) Exercise 6.
9. Find the sum of the second and third columns in the matrix of
  - a) Exercise 4,
  - b) Exercise 5,
  - c) Exercise 6.
10. Find the sum of the first and third columns in the matrix of
  - a) Exercise 4,
  - b) Exercise 5,
  - c) Exercise 6.

## §2. DETERMINANTS OF ORDER 2

Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

be a  $2 \times 2$  matrix. We define its **determinant** to be the number

$$ad - bc.$$

**Example.** The determinant of the matrix

$$\begin{pmatrix} 3 & 6 \\ 1 & 7 \end{pmatrix}$$

is equal to  $3 \cdot 7 - 1 \cdot 6 = 15$ .

**Example.** The determinant of the matrix

$$\begin{pmatrix} -2 & -5 \\ 4 & 8 \end{pmatrix}$$

is equal to  $(-2) \cdot 8 - (-5) \cdot 4 = -16 + 20 = 4$ .

The determinant of the matrix  $A$  is denoted by the symbols

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

Thus we have

$$\begin{vmatrix} -2 & -5 \\ 4 & 8 \end{vmatrix} = 4.$$

This notation is useful when we want to exhibit the components of the matrix. In other circumstances, it is useful to denote the determinant of  $A$  by the symbols

$$D(A),$$

which are short.

The determinant is an important tool for solving linear equations. We see this in the next theorem.

**Theorem 1.** *Let  $a, b, c, d, u, v$  be numbers. Assume that the determinant  $ad - bc$  is not equal to 0. Then the system of linear equations*

$$\begin{aligned} ax + by &= u, \\ cx + dy &= v \end{aligned}$$

*has a unique solution.*

*Proof.* The proof follows the method which we have already met in Chapter 3, consisting of eliminating a variable. For instance, multiply the first equation by  $d$ , multiply the second by  $b$ , and subtract the second from the first. The term involving  $y$  then becomes equal to 0, and we find that

$$adx - bcx = du - bv.$$

But  $adx - bcx = x(ad - bc)$ . Hence we find that if  $(x, y)$  is a solution, then

$$(1) \quad x = \frac{du - bv}{ad - bc}.$$

Eliminating  $x$  instead of  $y$  from the equations, we would find similarly that

$$(2) \quad y = \frac{av - cu}{ad - bc}.$$

What these arguments show is that if the equations have a solution, then this solution is given by formulas (1) and (2). But, conversely, computing  $ax + by$  with these values for  $x$  and  $y$ , we find  $u$ . Similarly, computing  $cx + dy$ , we find  $v$ . For instance:

$$ax + by = a \frac{du - bv}{ad - bc} + b \frac{av - cu}{ad - bc} = \frac{adu - abv + bav - bcu}{ad - bc} = u,$$

because first  $abv$  cancels, and then  $ad - bc$  cancels from the numerator and denominator, to yield  $u$ . This proves our theorem.

Observe that the solutions for  $x$  and  $y$  are quotients of expressions which are themselves determinants. For instance, we can rewrite the expression for  $x$  and  $y$  in the form

$$x = \frac{\begin{vmatrix} u & b \\ v & d \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}} \quad y = \frac{\begin{vmatrix} a & u \\ c & v \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}.$$

Note how the column

$$\begin{pmatrix} u \\ v \end{pmatrix}$$

occurs as the first column in the numerator for  $x$ , and how it occurs as the second column in the numerator for  $y$ . In both cases, the denominator is the determinant  $D(A) = ad - bc$ . These facts are typical of the more general case to be studied later for  $3 \times 3$  determinants.

**Example.** The system of equations

$$\begin{aligned} 3x - 2y &= 5, \\ 4x + 7y &= -4 \end{aligned}$$

has a solution, given by

$$x = \frac{\begin{vmatrix} 5 & -2 \\ -4 & 7 \end{vmatrix}}{\begin{vmatrix} 3 & -2 \\ 4 & 7 \end{vmatrix}} \quad \text{and} \quad y = \frac{\begin{vmatrix} 3 & 5 \\ 4 & -4 \end{vmatrix}}{\begin{vmatrix} 3 & -2 \\ 4 & 7 \end{vmatrix}}.$$

In practice, for the  $2 \times 2$  case, it is easier to work out the elimination method without determinants each time. However, we have described the method of determinants because in cases involving more unknowns, this method does become useful sometimes.

**Theorem 2.** If  $A$  is a  $2 \times 2$  matrix, then the determinant of  $A$  is equal to the determinant of the transpose of  $A$ . In other words,

$$D(A) = D({}^t A).$$

*Proof.* This is immediate from the definition of the determinant. We have

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} \quad \text{and} \quad |{}^t A| = \begin{vmatrix} a & c \\ b & d \end{vmatrix},$$

and

$$ad - bc = ad - cb.$$

Of course, the property expressed in Theorem 2 is very simple. We give it here because it is satisfied by  $3 \times 3$  determinants which will be studied later.

## EXERCISES

1. Compute the following determinants.

$$\text{a) } \begin{vmatrix} 3 & -5 \\ 4 & 2 \end{vmatrix} \qquad \text{b) } \begin{vmatrix} 2 & -1 \\ -3 & 4 \end{vmatrix} \qquad \text{c) } \begin{vmatrix} -3 & 4 \\ 2 & -1 \end{vmatrix}$$

d) 
$$\begin{vmatrix} -5 & 3 \\ 4 & 6 \end{vmatrix}$$

e) 
$$\begin{vmatrix} 3 & 3 \\ -7 & -8 \end{vmatrix}$$

f) 
$$\begin{vmatrix} -5 & -4 \\ 6 & 3 \end{vmatrix}$$

2. Compute the determinant

$$\begin{vmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{vmatrix}$$

for any real number  $\theta$ .

3. Compute the determinant

$$\begin{vmatrix} \cos \theta & \sin \theta \\ \sin \theta & \cos \theta \end{vmatrix}$$

when

- a)  $\theta = \pi$ ,      b)  $\theta = \pi/2$ ,      c)  $\theta = \pi/3$ ,      d)  $\theta = \pi/4$ .

### §3. PROPERTIES OF $2 \times 2$ DETERMINANTS

Consider a  $2 \times 2$  matrix  $A$  with columns  $A^1, A^2$ . The determinant  $D(A)$  has interesting properties with respect to these columns, which we shall describe. Thus it is useful to use the notation

$$D(A) = D(A^1, A^2)$$

to emphasize the dependence of the determinant on its columns. If the two columns are denoted by

$$B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

then we would write

$$D(B, C) = \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix} = b_1c_2 - c_1b_2.$$

We may view the determinant as a certain type of “product” between the columns  $B$  and  $C$ . To what extent does this product satisfy the same rules as the product of numbers. Answer: To some extent, which we now determine precisely.

To begin with, this “product” satisfies distributivity. In the determinant notation, this means:

**D1.** If  $B = B' + B''$ , i.e.

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b'_1 \\ b'_2 \end{pmatrix} + \begin{pmatrix} b''_1 \\ b''_2 \end{pmatrix},$$

then

$$D(B' + B'', C) = D(B', C) + D(B'', C).$$

Similarly, if  $C = C' + C''$ , then

$$D(B, C' + C'') = D(B, C') + D(B, C'').$$

*Proof.* Of course, the proof is quite simple using the definition of the determinant. We have:

$$\begin{aligned} D(B' + B'', C) &= \begin{vmatrix} b'_1 + b''_1 & c_1 \\ b'_2 + b''_2 & c_2 \end{vmatrix} \\ &= (b'_1 + b''_1)c_2 - (b'_2 + b''_2)c_1 \\ &= b'_1c_2 + b''_1c_2 - b'_2c_1 - b''_2c_1 \\ &= D(B', C) + D(B'', C). \end{aligned}$$

Distributivity on the other side is proved similarly.

**D2.** If  $x$  is a number, then

$$D(xB, C) = x \cdot D(B, C) = D(B, xC).$$

*Proof.* We have:

$$\begin{aligned} D(xB, C) &= \begin{vmatrix} xb_1 & c_1 \\ xb_2 & c_2 \end{vmatrix} = xb_1c_2 - xb_2c_1 = x(b_1c_2 - b_2c_1) \\ &= xD(B, C). \end{aligned}$$

Again, the other equality is proved similarly.

The two vectors

$$E^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad E^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

will be called the **unit vectors**. The matrix formed by them, namely

$$E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

will be called the **unit matrix**. We have:

**D3.** *If  $E$  is the unit matrix, then  $D(E) = D(E^1, E^2) = 1$ .*

This is obvious.

**D4.** *If the two columns of the matrix are equal, then the determinant is equal to 0. In other words,*

$$D(B, B) = 0.$$

*Proof.* This is obvious, because

$$\begin{vmatrix} b_1 & b_1 \\ b_2 & b_2 \end{vmatrix} = b_1 b_2 - b_2 b_1 = 0.$$

These four basic properties are fundamental, and other properties can be deduced from them, without going back to the definition of the determinant in terms of the components of the matrix.

**D5.** *If we add a multiple of one column to the other, then the value of the determinant does not change. In other words, let  $x$  be a number. Then*

$$D(B + xC, C) = D(B, C) \quad \text{and} \quad D(B, C + xB) = D(B, C).$$

Written out in terms of components, the first relation reads

$$\begin{vmatrix} b_1 + xc_1 & c_1 \\ b_2 + xc_2 & c_2 \end{vmatrix} = \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}.$$

*Proof.* Using **D1**, **D2**, **D4** in succession, we find that

$$\begin{aligned} D(B + xC, C) &= D(B, C) + D(xC, C) \\ &= D(B, C) + xD(C, C) = D(B, C). \end{aligned}$$

A similar proof applies to  $D(B, C + xB)$ .

**D6.** *If the two columns are interchanged, then the value of the determinant changes by a sign. In other words, we have*

$$D(B, C) = -D(C, B).$$

*Proof.* Again, we use **D1**, **D2**, **D4** successively, and get:

$$\begin{aligned} 0 &= D(B + C, B + C) = D(B, B + C) + D(C, B + C) \\ &= D(B, B) + D(B, C) + D(C, B) + D(C, C) \\ &= D(B, C) + D(C, B). \end{aligned}$$

This proves that  $D(B, C) = -D(C, B)$ , as desired.

Of course, you can also give a proof using the components of the matrix. Do this as an exercise. However, there is some point in doing it as above, because in the study of determinants in the higher-dimensional case later, a proof with components becomes much messier, while the proof following the same pattern as the one we have given remains neat.

Observe that **D6** shows that our determinant, viewed as a “product,” is *not* commutative. Commutativity would mean that

$$D(B, C) = D(C, B),$$

and this is *not true*. Note that **D6** was deduced from distributivity and the special property **D4**. The property expressed by **D6** is often called **anti-commutativity**—“anti” because of the minus sign which appears.

We can find the values of  $x$  and  $y$  in Theorem 1, by using the properties we have just proved. This new argument is the one which generalizes later. It runs as follows. We wish to solve the system of linear equations

$$(*) \quad \begin{aligned} a_1x + b_1y &= c_1 \\ a_2x + b_2y &= c_2. \end{aligned}$$

Observe how we have numbered the coefficients in such a way that we can write the columns easily, in an abbreviated fashion. Namely,  $(*)$  can be written in the form

$$(**) \quad xA + yB = C,$$

where

$$A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad C = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Suppose that  $x, y$  are solutions of the system (\*\*). Using our properties of the determinant, we have:

$$\begin{aligned} D(C, B) &= D(xA + yB, B) = D(xA, B) + D(yB, B) \\ &= xD(A, B) + yD(B, B) \\ &= xD(A, B). \end{aligned}$$

Hence we find that

$$x = \frac{D(C, B)}{D(A, B)} = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}.$$

If you compare this with the solution found previously you will find, of course, that we have obtained the same value, namely a quotient of determinants whose denominator is the determinant of the coefficients of the equations. Similarly, you can find the value for  $y$  using the same method. Do it as an exercise.

## EXERCISES

1. Prove the other half of **D1**, i.e. distributivity on the side other than that given in the text.
2. Prove the other half of **D2**.
3. Prove the other half of **D5**.
4. Using the same method as at the end of the section, find the value for  $y$  as a quotient of two determinants.
5. Solve the linear equations of Chapter 2, §1 by determinants.
6. Let  $c$  be a number, and let  $A$  be a  $2 \times 2$  matrix. Define  $cA$  to be the matrix obtained by multiplying all components of  $A$  by  $c$ . How does  $D(cA)$  differ from  $D(A)$ ?
7. Let  $A = (a_1, a_2)$  and  $B = (b_1, b_2)$ . Define their **dot product**  $A \cdot B$  by the formula

$$A \cdot B = a_1b_1 + a_2b_2.$$

For instance,  $(3, 1) \cdot (-4, 5) = -3 \cdot 4 + 1 \cdot 5 = -7$ . Thus the dot product is a number. Prove that this product is commutative and distributive with respect to addition. If  $A \neq (0, 0)$ , prove that  $A \cdot A > 0$ . Give an example of  $A \neq O$  and  $B \neq O$  such that  $A \cdot B = 0$ . Compute  $A \cdot B$  for the following values of  $A$  and  $B$ .

- a)  $A = (-4, 3)$  and  $B = (5, -2)$
- b)  $A = (-2, -1)$  and  $B = (-3, 4)$

#### §4. DETERMINANTS OF ORDER 3

We shall define the determinant for  $3 \times 3$  matrices, and we shall see that it satisfies properties analogous to those of the  $2 \times 2$  case.

Let

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

be a  $3 \times 3$  matrix. We define its **determinant** according to the formula known as the **expansion by a row**, say the first row. That is, we define

$$(1) \quad D(A) = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

and we denote  $D(A)$  also with the two vertical bars

$$D(A) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

We may describe the sum in (1) as follows. Let  $A_{ij}$  be the matrix obtained from  $A$  by deleting the  $i$ -th row and  $j$ -th column. Then the sum for  $D(A)$  can be written as

$$a_{11}D(A_{11}) - a_{12}D(A_{12}) + a_{13}D(A_{13}).$$

In other words, each term consists of the product of an element of the first row and the determinant of the  $2 \times 2$  matrix obtained by deleting the first row and the  $j$ -th column, and putting the appropriate sign to this term as shown.

**Example.** Let

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 4 \\ -3 & 2 & 5 \end{pmatrix}.$$

Then

$$A_{11} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \end{pmatrix}, \quad A_{12} = \begin{pmatrix} 1 & 4 \\ -3 & 5 \end{pmatrix}, \quad A_{13} = \begin{pmatrix} 1 & 1 \\ -3 & 2 \end{pmatrix}$$

and our formula for the determinant of  $A$  yields

$$\begin{aligned} D(A) &= 2 \begin{vmatrix} 1 & 4 \\ 2 & 5 \end{vmatrix} - 1 \begin{vmatrix} 1 & 4 \\ -3 & 5 \end{vmatrix} + 0 \begin{vmatrix} 1 & 1 \\ -3 & 2 \end{vmatrix} \\ &= 2(5 - 8) - 1(5 + 12) + 0 \\ &= -23. \end{aligned}$$

Thus the determinant is a number. To compute this number in the above example, we computed the determinants of the  $2 \times 2$  matrices explicitly. We can also expand these in the general definition, and thus we find a six-term expression for the determinant of a general  $3 \times 3$  matrix  $A = (a_{ij})$ , namely:

$$(2) \quad \boxed{D(A) = a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.}$$

Do not memorize (2). Remember only (1), and write down (2) only when needed for specific purposes.

We could have used the other rows to expand the determinant, instead of the first row. For instance, the expansion according to the second row is given by

$$\begin{aligned} -a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{22} \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} - a_{23} \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} \\ = -a_{21}D(A_{21}) + a_{22}D(A_{22}) - a_{23}D(A_{23}). \end{aligned}$$

Again, each term is the product of  $a_{2j}$  with the determinant of the  $2 \times 2$  matrix obtained by deleting the second row and  $j$ -th column, together with the appropriate sign in front of each term. This sign is determined according to the pattern:

$$\begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix}.$$

If you write down the two terms for each one of the  $2 \times 2$  determinants in the expansion according to the second row, you will obtain six terms, and you will find immediately that they give you the same value which we wrote down in formula (2). Thus expanding according to the second row gives the same value for the determinant as expanding according to the first row.

Furthermore, we can also expand according to any one of the columns. For instance, expanding according to the first column, we find that

$$a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}$$

yields precisely the same six terms as in (2), if you write down each one of the two terms corresponding to each one of the  $2 \times 2$  determinants in the above expression.

**Example.** We compute the determinant

$$\begin{vmatrix} 3 & 0 & 1 \\ 1 & 2 & 5 \\ -1 & 4 & 2 \end{vmatrix}$$

by expanding according to the second column. The determinant is equal to

$$2 \begin{vmatrix} 3 & 1 \\ -1 & 2 \end{vmatrix} - 4 \begin{vmatrix} 3 & 1 \\ 1 & 5 \end{vmatrix} = 2(6 - (-1)) - 4(15 - 1) = -42.$$

Note that the presence of 0 in the first row and second column eliminates one term in the expansion, since this term is equal to 0.

If we expand the above determinant according to the third column, we find the same value, namely

$$+1 \begin{vmatrix} 1 & 2 \\ -1 & 4 \end{vmatrix} - 5 \begin{vmatrix} 3 & 0 \\ -1 & 4 \end{vmatrix} + 2 \begin{vmatrix} 3 & 0 \\ 1 & 2 \end{vmatrix} = -42.$$

**Theorem 3.** If  $A$  is a  $3 \times 3$  matrix, then  $D(A) = D({}^t A)$ . In other words, the determinant of  $A$  is equal to the determinant of the transpose of  $A$ .

*Proof.* This is true because expanding  $D(A)$  according to rows or columns gives the same value, namely the expression in (2).

## EXERCISES

1. Write down the expansion of a  $3 \times 3$  determinant according to the third row, the second column, and the third column, and verify in each case that you get the same six terms as in (2).

2. Compute the following determinants by expanding according to the second row, and also according to the third column, as a check for your computation. Of course, you should find the same value.

a) 
$$\begin{vmatrix} 2 & 1 & 2 \\ 0 & 3 & -1 \\ 4 & 1 & 1 \end{vmatrix}$$

b) 
$$\begin{vmatrix} 3 & -1 & 5 \\ -1 & 2 & 1 \\ -2 & 4 & 3 \end{vmatrix}$$

c) 
$$\begin{vmatrix} 2 & 4 & 3 \\ -1 & 3 & 0 \\ 0 & 2 & 1 \end{vmatrix}$$

d) 
$$\begin{vmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 2 & 7 \end{vmatrix}$$

e) 
$$\begin{vmatrix} -1 & 5 & 3 \\ 4 & 0 & 0 \\ 2 & 7 & 8 \end{vmatrix}$$

f) 
$$\begin{vmatrix} 3 & 1 & 2 \\ 4 & 5 & 1 \\ -1 & 2 & -3 \end{vmatrix}$$

3. Compute the following determinants.

a) 
$$\begin{vmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{vmatrix}$$

b) 
$$\begin{vmatrix} -3 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & -8 \end{vmatrix}$$

c) 
$$\begin{vmatrix} 6 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & -2 \end{vmatrix}$$

4. Let  $a, b, c$  be numbers. In terms of  $a, b, c$ , what is the value of the determinant

$$\begin{vmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{vmatrix} ?$$

5. Compute the following determinants.

a) 
$$\begin{vmatrix} 3 & 1 & -5 \\ 0 & 4 & 1 \\ 0 & 0 & -2 \end{vmatrix}$$

b) 
$$\begin{vmatrix} 4 & -6 & 7 \\ 0 & 2 & -8 \\ 0 & 0 & -9 \end{vmatrix}$$

c) 
$$\begin{vmatrix} 6 & 0 & 0 \\ -4 & -5 & 0 \\ 7 & 20 & -3 \end{vmatrix}$$

d) 
$$\begin{vmatrix} 5 & 0 & 0 \\ 4 & 2 & 0 \\ -17 & 19 & -3 \end{vmatrix}$$

6. In terms of the components of the matrix, what is the value of the determinant:

a) 
$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{vmatrix} ?$$

b) 
$$\begin{vmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{vmatrix} ?$$

## §5. PROPERTIES OF $3 \times 3$ DETERMINANTS

We shall now see that  $3 \times 3$  determinants satisfy the properties **D1** through **D6**, listed previously for  $2 \times 2$  determinants. These properties are concerned with the columns of the matrix, and hence it is useful to use the same notation which we used before. If  $A^1, A^2, A^3$  are the columns of the  $3 \times 3$  matrix  $A$ , then we write

$$D(A) = D(A^1, A^2, A^3).$$

For the rest of this section, we assume that our column and row vectors have dimension 3; that is, that they have three components. Thus any column vector  $B$  in this section can be written in the form

$$B = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

**D1.** Suppose that the first column can be written as a sum,

$$A^1 = B + C,$$

that is,

$$\begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} + \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}.$$

Then

$$D(B + C, A^2, A^3) = D(B, A^2, A^3) + D(C, A^2, A^3).$$

and the analogous rule holds with respect to the second and third columns.

*Proof.* We use the definition of the determinant, namely the expansion according to the first row. We see that each term splits into a sum of two terms corresponding to  $B$  and  $C$ . For instance,

$$a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} = b_1 \begin{vmatrix} a_{22} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + c_1 \begin{vmatrix} a_{22} & a_{23} \\ a_{31} & a_{33} \end{vmatrix},$$

$$a_{12} \begin{vmatrix} b_2 + c_2 & a_{23} \\ b_3 + c_3 & a_{33} \end{vmatrix} = a_{12} \begin{vmatrix} b_2 & a_{23} \\ b_3 & a_{33} \end{vmatrix} + a_{12} \begin{vmatrix} c_2 & a_{23} \\ c_3 & a_{33} \end{vmatrix},$$

$$a_{13} \begin{vmatrix} b_2 + c_2 & a_{22} \\ b_3 + c_3 & a_{32} \end{vmatrix} = a_{13} \begin{vmatrix} b_2 & a_{22} \\ b_3 & a_{32} \end{vmatrix} + a_{13} \begin{vmatrix} c_2 & a_{22} \\ c_3 & a_{32} \end{vmatrix}.$$

Summing with the appropriate sign yields the desired relation.

**D2.** If  $x$  is a number, then

$$D(xA^1, A^2, A^3) = x \cdot D(A^1, A^2, A^3),$$

and similarly for the other columns.

*Proof.* We have:

$$\begin{aligned} D(xA^1, A^2, A^3) &= xa_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} xa_{21} & a_{23} \\ xa_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} xa_{21} & a_{22} \\ xa_{31} & a_{32} \end{vmatrix} \\ &= x \cdot D(A^1, A^2, A^3). \end{aligned}$$

The proof is similar for the other columns.

In the  $3 \times 3$  case, we also have the **unit vectors**, namely

$$E^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad E^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad E^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

and the **unit  $3 \times 3$  matrix**, namely

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

**D3.** If  $E$  is the unit matrix, then  $D(E) = D(E^1, E^2, E^3) = 1$ .

*Proof.* This is obvious from the expansion according to the first row.

**D4.** If two columns of the matrix are equal, then the determinant is equal to 0.

*Proof.* Suppose that  $A^1 = A^2$ , and look at the expansion of the determinant according to the first row. Then  $a_{11} = a_{12}$ , and the first two terms cancel. The third term is equal to 0 because it involves a  $2 \times 2$  determinant whose two columns are equal. The proof for the other cases is similar. (Other cases:  $A^2 = A^3$  and  $A^1 = A^3$ .)

Observe that to prove our basic four properties, we needed to use the definition of the determinant, i.e. its expansion according to the first row. For the remaining properties, we can give a proof which is not based directly on this expansion, but only on the formalism of D1 through D4. This has the advantage of making the arguments easier, and in fact of making them completely analogous to those used in the  $2 \times 2$  case. We carry them out.

**D5.** *If we add a multiple of one column to another, then the value of the determinant does not change. In other words, let  $x$  be a number. Then for instance*

$$D(A^1, A^2 + xA^1, A^3) = D(A^1, A^2, A^3),$$

*and similarly in all other cases.*

*Proof.* We have

$$\begin{aligned} D(A^1, A^2, + xA^1, A^3) &= D(A^1, A^2, A^3) + D(A^1, xA^1, A^3) && \text{(by D1)} \\ &= D(A^1, A^2, A^3) + x \cdot D(A^1, A^1, A^3) && \text{(by D2)} \\ &= D(A^1, A^2, A^3) && \text{(by D4).} \end{aligned}$$

This proves what we wanted. The proofs of the other cases are similar.

**D6.** *If two adjacent columns are interchanged, then the determinant changes by a sign. In other words, we have*

$$D(A^1, A^3, A^2) = -D(A^1, A^2, A^3),$$

*and similarly in the other case.*

*Proof.* We use the same method as before. We find:

$$\begin{aligned} 0 &= D(A^1, A^2 + A^3, A^2 + A^3) \\ &= D(A^1, A^2, A^2 + A^3) + D(A^1, A^3, A^2 + A^3) \\ &= D(A^1, A^2, A^2) + D(A^1, A^2, A^3) + D(A^1, A^3, A^2) + D(A^1, A^3, A^3) \\ &= D(A^1, A^2, A^3) + D(A^1, A^3, A^2), \end{aligned}$$

using **D1** and **D4**. This proves **D6** in this case, and the other cases are proved similarly.

Using these rules, especially **D5**, we can compute determinants a little more efficiently. For instance, we have already noticed that when a 0 occurs in the given matrix, we can expand according to the row (or column) in which this 0 occurs, and it eliminates one term. Using **D5** repeatedly, we can change the matrix so as to get as many zeros as possible, and then reduce the computation to one term.

**Example.** Compute the determinant

$$\begin{vmatrix} 3 & 0 & 1 \\ 1 & 2 & 5 \\ -1 & 4 & 2 \end{vmatrix}.$$

We already have 0 in the first row. We subtract two times the second row from the third row. Our determinant is then equal to

$$\begin{vmatrix} 3 & 0 & 1 \\ 1 & 2 & 5 \\ -3 & 0 & -8 \end{vmatrix}.$$

We expand according to the second column. The expansion has only one term  $\neq 0$ , with a + sign, and that is:

$$2 \begin{vmatrix} 3 & 1 \\ -3 & -8 \end{vmatrix}.$$

The  $2 \times 2$  determinant can be evaluated by our definition of  $ad - bc$ , and we find the value

$$2(-24 - (-3)) = -42.$$

**Example.** We compute the determinant

$$\begin{vmatrix} 4 & 7 & 10 \\ 3 & 7 & 5 \\ 5 & -1 & 10 \end{vmatrix}.$$

We subtract two times the second row from the first row, and then from the third row, yielding

$$\begin{vmatrix} -2 & -7 & 0 \\ 3 & 7 & 5 \\ -1 & -15 & 0 \end{vmatrix},$$

which we expand according to the third column, and get

$$-5(30 - 7) = -5(23) = -115.$$

Note that the term has a minus sign, determined by our usual pattern of signs.

### EXERCISES

1. Write out in full and prove property **D1** with respect to the second column and the third column.

2. Same thing for property **D2**.
3. Prove the two cases not treated in the text for property **D4**.
4. Prove **D5** in the case
  - a) you add a multiple of the third column to the first;
  - b) you add a multiple of the second column to the first;
  - c) you add a multiple of the third column to the second.
5. Prove **D6** in the second case.
6. If you interchange the first and third columns of the given matrix, how does its determinant change? What about interchanging the first and third row?
7. State **D5** and **D6** for rows.
8. Compute the following determinants, making the computation as easy as you can.
 

a) $\begin{vmatrix} 4 & -9 & 2 \\ 4 & -9 & 2 \\ 3 & 1 & 5 \end{vmatrix}$	b) $\begin{vmatrix} 4 & -1 & 1 \\ 2 & 0 & 0 \\ 1 & 5 & 7 \end{vmatrix}$	c) $\begin{vmatrix} 2 & -1 & 4 \\ 1 & 1 & 5 \\ 1 & 2 & 3 \end{vmatrix}$
d) $\begin{vmatrix} 3 & 1 & 1 \\ 2 & 5 & 5 \\ 8 & 7 & 7 \end{vmatrix}$	e) $\begin{vmatrix} 2 & 1 & 1 \\ 3 & 1 & 5 \\ 4 & -2 & 3 \end{vmatrix}$	f) $\begin{vmatrix} -4 & 4 & 2 \\ 5 & 1 & 3 \\ 2 & 1 & 4 \end{vmatrix}$
g) $\begin{vmatrix} 7 & 3 & 2 \\ 1 & -1 & 1 \\ 2 & 1 & 3 \end{vmatrix}$	h) $\begin{vmatrix} 3 & 2 & 1 \\ 1 & 1 & 1 \\ -1 & 3 & 4 \end{vmatrix}$	i) $\begin{vmatrix} -2 & -1 & 1 \\ 3 & 1 & -1 \\ -1 & 2 & 3 \end{vmatrix}$
j) $\begin{vmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{vmatrix}$	k) $\begin{vmatrix} -4 & 1 & 2 \\ 3 & 2 & 1 \\ -1 & -1 & 1 \end{vmatrix}$	l) $\begin{vmatrix} -1 & 3 & 2 \\ 3 & -1 & 1 \\ 6 & -2 & 2 \end{vmatrix}$
9. Let  $c$  be a number and multiply each component  $a_{ij}$  of a  $3 \times 3$  matrix  $A$  by  $c$ , thus obtaining a new matrix which we denote by  $cA$ . How does  $D(A)$  differ from  $D(cA)$ ?
10. Let  $x_1, x_2, x_3$  be numbers. Show that

$$\begin{vmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{vmatrix} = (x_2 - x_1)(x_3 - x_2)(x_3 - x_1).$$

11. Suppose that  $A^1$  is a sum of three columns, say

$$A^1 = B^1 + B^2 + B^3.$$

Using **D1** twice, prove that

$$\begin{aligned} D(B^1 + B^2 + B^3, A^2, A^3) \\ = D(B^1, A^2, A^3) + D(B^2, A^2, A^3) + D(B^3, A^2, A^3). \end{aligned}$$

Using summation notation, we can write this in the form

$$D(B^1 + B^2 + B^3, A^2, A^3) = \sum_{j=1}^3 D(B^j, A^2, A^3),$$

which is shorter. In general, suppose that

$$A^1 = \sum_{j=1}^n B^j$$

is a sum of  $n$  columns. Using the summation notation, express similarly

$$D(A^1, A^2, A^3)$$

as a sum of (how many?) terms.

12. Let  $x_j$  ( $j = 1, 2, 3$ ) be numbers. Let

$$A^1 = x_1 C^1 + x_2 C^2 + x_3 C^3.$$

Prove that

$$D(A^1, A^2, A^3) = \sum_{j=1}^3 x_j D(C^j, A^2, A^3).$$

State and prove the analogous statement when

$$A^1 = \sum_{j=1}^n x_j C^j.$$

13. State the analogous property to that of Exercise 12 with respect to the second column. Then with respect to the third column.

## §6. CRAMER'S RULE

We now come to solving linear equations in three unknowns using determinants. For practical purposes, the simple-minded elimination method of Chapter 2, §2, is the easiest to use when the equations are explicitly given with numerical coefficients, and you don't need any harder theory, like that of determinants. For theoretical purposes, however, and for more advanced computations, the theory of determinant can often be used advantageously, and that is the reason why we now go into it.

We can write a system of linear equations

$$(*) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

using the vector notation, in the form

$$(**) \quad x_1 A^1 + x_2 A^2 + x_3 A^3 = B,$$

where  $A^1, A^2, A^3$  are the columns of the matrix of coefficients  $a_{ij}$ , and  $B$  is the column formed by  $b_1, b_2, b_3$ .

**Theorem 4.** Let  $x_1, x_2, x_3$  be numbers which are solutions of the system of linear equations above. Let  $A = (a_{ij})$  be its matrix of coefficients. If  $D(A) \neq 0$ , then

$$\begin{aligned} x_1 &= \frac{D(B, A^2, A^3)}{D(A^1, A^2, A^3)}, & x_2 &= \frac{D(A^1, B, A^3)}{D(A^1, A^2, A^3)}, \\ x_3 &= \frac{D(A^1, A^2, B)}{D(A^1, A^2, A^3)}. \end{aligned}$$

Or, written out in terms of the components, say for  $x_1$ , we have

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}}.$$

**Note.** The practicality of our notation now becomes really apparent. It is obviously much longer and more tiring to write out all the components than to write out the expressions with the abbreviation for the columns as in  $D(B, A^1, A^2)$ , etc.

*Proof of Theorem 4.* We use the same technique as for the proof of the  $2 \times 2$  case. We have:

$$\begin{aligned} D(B, A^2, A^3) &= D(x_1A^1 + x_2A^2 + x_3A^3, A^2, A^3) \\ &= x_1D(A^1, A^2, A^3) + x_2D(A^2, A^2, A^3) + x_3D(A^3, A^2, A^3) \\ &= x_1D(A^1, A^2, A^3). \end{aligned}$$

The first equality follows from **D1** and **D2**. As an exercise, write out the missing steps (Exercise 1). The second equality follows from **D4**. If  $D(A) \neq 0$ , then we get the desired expression for  $x_1$ . We treat  $x_2$  and  $x_3$  similarly (Exercise 2).

Theorem 4 is known as **Cramer's rule**.

Observe that Theorem 4 does not tell us that a solution exists. It tells us that if a solution exists, then it is given by the formulas as in the theorem. In fact, the following is true.

**Theorem 5.** Let  $A$  be a  $3 \times 3$  matrix whose determinant is not 0. Then for any column  $B$  there exist numbers  $x_1, x_2, x_3$  such that

$$x_1A^1 + x_2A^2 + x_3A^3 = B.$$

In other words, the system of linear equations (\*) has a solution.

The proof for Theorem 5 is slightly more involved than in the analogous  $2 \times 2$  case, and we shall omit it.

**Example.** We solve the following system of linear equations by Cramer's rule.

$$\begin{aligned} 3x + 2y + 4z &= 1 \\ 2x - y + z &= 0 \\ x + 2y + 3z &= 1 \end{aligned}$$

We have:

$$x = \frac{\begin{vmatrix} 1 & 2 & 4 \\ 0 & -1 & 1 \\ 1 & 2 & 3 \end{vmatrix}}{\begin{vmatrix} 3 & 2 & 4 \\ 2 & -1 & 1 \\ 1 & 2 & 3 \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} 3 & 1 & 4 \\ 2 & 0 & 1 \\ 1 & 1 & 3 \end{vmatrix}}{\begin{vmatrix} 3 & 2 & 4 \\ 2 & -1 & 1 \\ 1 & 2 & 3 \end{vmatrix}}, \quad z = \frac{\begin{vmatrix} 3 & 2 & 1 \\ 2 & -1 & 0 \\ 1 & 2 & 1 \end{vmatrix}}{\begin{vmatrix} 3 & 2 & 4 \\ 2 & -1 & 1 \\ 1 & 2 & 3 \end{vmatrix}}.$$

Observe how the column

$$B = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

shifts from the first column when solving for  $x$ , to the second column when solving for  $y$ , to the third column when solving for  $z$ . The denominator in all three expressions is the same, namely it is the determinant of the matrix of coefficients of the equations.

We know how to compute  $3 \times 3$  determinants, and we then find

$$x = -\frac{1}{5}, \quad y = 0, \quad z = \frac{2}{5}.$$

## EXERCISES

- Fill in the missing steps in the proof of Cramer's rule. Cf. Exercises 11 and 12 of the preceding section.
- Write out in full the proof of Cramer's rule for  $x_2$  and  $x_3$ . It is very similar to the proof for  $x_1$  in the text.
- Let  $A^1, A^2, A^3$  be columns of a  $3 \times 3$  matrix  $A$ , and assume that there exist numbers  $x_1, x_2, x_3$  not all 0 such that

$$x_1 A^1 + x_2 A^2 + x_3 A^3 = 0.$$

Prove that  $D(A) = 0$ .

- Solve the linear equations of Chapter 2, §2 by Cramer's rule.

# *Index*



# *Index*

- absolute value, 66
- addition, 61, 218
- addition formulas, 272
- additive inverse, 8, 61, 62, 219
- angle, 111
- area, 173
- associativity, 8, 61, 153, 218, 352
- ball, 227
- base, 334, 338
- beginning point, 230
- binomial coefficient, 39, 385, 387
- cancellation law, 14, 45, 357
- circle, 109, 203
- circumference, 180
- closed interval, 78
- coefficients, 322
- column, 401
- common divisor, 29
- commutativity, 8, 61, 218
- complex conjugate, 377
- complex numbers, 375
- component, 403
- composite, 351
- composition, 150, 351
- congruent, 26, 166
- constant function, 313
- constant mapping, 134
- coordinates, 191, 194
- cosine, 252
- cotangent, 270
- Cramer's rule, 425
- cross multiplication, 27, 43
- cross multiplication of inequalities, 80
- cycle, 367
- degree of polynomial, 322
- degrees, 115
- determinant, 406
- dilation, 136, 201, 215, 225
- directed segment, 230
- direction, 231
- distance, 107, 197, 348
- distance preserving, 143
- divide, 24
- divisible, 24
- element, 99
- ellipse, 178, 297
- empty set, 100
- end points, 78, 230
- Euclidean algorithm, 326
- even function, 317
- even integers, 23
- even permutation, 364
- expansion of determinant, 414
- exponential function, 333
- factorial, 39
- fixed point, 146
- function, 313

- geometric series, 397  
 graphs, 288, 305, 329  
 greater than, 75
- hyperbola, 300
- identity mapping, 135, 351  
 if and only if, 94  
 image, 134, 313, 346  
 implication, 75  
 induction, 383  
 integers, 6  
 interval, 78  
 inverse, 155, 353, 361  
 irrational, 37  
 isometry, 143, 222
- leading coefficient, 322  
 length of cycle, 367  
 line, 237, 281  
 line segment, 229  
 linear equations, 53  
 located vector, 230  
 logarithm, 338  
 logic, 94  
 lowest form, 29
- mapping, 134, 345  
 matrix, 279, 401  
 minus, 6  
 multiplication, 34, 62  
 multiplication table, 159  
 multiplicative inverse, 42, 62
- natural number, 5  
 negative, 33, 64  
 negative integers, 6  
 $n$ -th root, 71
- odd function, 317  
 odd integers, 22  
 odd permutation, 364  
 open interval, 78  
 orbit, 367  
 orbit decomposition, 369
- ordinary equation for the line, 246, 281  
 origin, 5, 191
- parabola, 291  
 parallel, 236  
 parallelogram law, 219  
 parametric representation, 239  
 period, 367  
 permutation, 359  
 perpendicular bisector, 127  
 plane, 201  
 polar coordinates, 260, 351  
 polar form, 380  
 polynomial, 318  
 positive, 64  
 positive integers, 5  
 power, 18, 71  
 product of functions, 315  
 Pythagoras theorem, 125
- quadrant, 194  
 quadratic equations, 83  
 quadratic formula, 86
- radian, 249  
 rational function, 328  
 rational numbers, 26  
 rational points, 206  
 rationalize, 68  
 ray, 110, 231  
 reading books, 93  
 rectangle, 123  
 reflection, 135, 216, 225  
 remainder, 326  
 right angle, 116  
 right triangle, 121  
 root, 71, 319  
 rotation, 137, 301, 305  
 row, 401
- same direction, 233  
 segment, 108  
 set, 99  
 similarity transformation, 137  
 sine, 252  
 slope, 309

- square root, 65  
straight angle, 115  
sum of functions, 315  
summation, 388  
  
tangent, 266  
translation, 141, 219  
transpose, 404  
transposition, 362  
triangle, 120  
  
unit matrix, 411  
unit vector, 411  
  
value, 134, 313  
vertex of ray, 110, 231  
volume, 392  
  
zero, 5, 219  
zero angle, 114  
zero function, 315



# *Answers to Selected Exercises*



# *Answers to Selected Exercises*

## *Chapter 1, §2*

1.  $(a + b) + (c + d) = a + b + c + d$  by associativity  
=  $a + b + d + c = a + d + b + c$   
by commutativity  
=  $(a + d) + (b + c)$  by associativity
3.  $(a - b) + (c - d) = a - b + c - d$  by associativity  
=  $a + c - b - d$  by commutativity  
=  $(a + c) + (-b - d)$  by associativity
6.  $(a - b) + (c - d) = a - b + c - d$  by associativity  
=  $-b + a + c - d = -b + a - d + c$   
=  $-b - d + a + c$  by commutativity  
=  $-(b + d) + (a + c)$  by associativity and N5
11. It suffices to show that  $(a + b + c) + (-a) + (-b) + (-c) = 0$ .  
But  $(a + b + c) + (-a) + (-b) + (-c)$   
=  $a + b + c - a - b - c$  by associativity  
=  $a + b - a + c - b - c$   
=  $a - a + b + c - b - c$  by commutativity  
=  $a - a + b - b + c - c$   
=  $(a - a) + (b - b) + (c - c)$  by associativity  
=  $0 + 0 + 0 = 0$

13. Same type of proof. We must show that  $(a - b) + b - a = 0$ . We have:

$$\begin{aligned} a - b + b - a &= a - b - a + b \\ &= a - a - b + b \text{ by commutativity} \\ &= (a - a) + (-b + b) \text{ by associativity} \\ &= 0 + 0 = 0 \end{aligned}$$

15.  $x = -3$     17.  $x = 5$     19.  $x = -3$

22. Let  $a + b = a + c$ . Adding  $-a$  to both sides, we get  $-a + a + b = -a + a + c$ , that is,  $(-a + a) + b = (-a + a) + c$  by associativity. Then  $0 + b = 0 + c$ , i.e.,  $b = c$ .

- 23.** Add  $-a$  to both sides of the equation  $a + b = a$ . We obtain  $-a + a + b = -a + a = 0$ , so  $0 + b = 0$  and  $b = 0$ .

*Chapter 1, §3*

1. a)  $2^8 3^3 a^7 b^4$       c)  $2^{10} 3^3 a^6 b^{10}$       e)  $2^4 3^5 a^6 b^8$
2.  $(a + b)^3 = (a + b)^2(a + b) = (a^2 + 2ab + b^2)(a + b)$   
 $= (a^2 + 2ab + b^2)a + (a^2 + 2ab + b^2)b$   
 $= a^3 + 2aba + b^2a + a^2b + 2ab^2 + b^3$   
 $= a^3 + 3a^2b + 3b^2a + b^3$
- $(a - b)^3 = (a - b)^2(a - b) = (a^2 - 2ab + b^2)(a - b)$   
 $= (a^2 - 2ab + b^2)a - (a^2 - 2ab + b^2)b$   
 $= a^3 - 2aba + b^2a - a^2b + 2ab^2 - b^3$   
 $= a^3 - 3a^2b + 3ab^2 - b^3$
3.  $(a + b)^4 = (a + b)^3(a + b) = (a^3 + 3a^2b + 3ab^2 + b^3)(a + b)$   
 $= (a^3 + 3a^2b + 3ab^2 + b^3)a + (a^3 + 3a^2b + 3ab^2 + b^3)b$   
 $= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$
- $(a - b)^4 = (a - b)^3(a - b) = (a^3 - 3a^2b + 3ab^2 - b^3)(a - b)$   
 $= a^4 - 4a^3b + 6a^2b^2 - 4ab^3 + b^4$
4. 16 $x^2 - 16x + 4$       6. 4 $x^2 + 20x + 25$       8.  $x^2 - 1$
12.  $x^4 + 2x^2 + 1$       14.  $x^4 + 4x^2 + 4$       20.  $2x^3 + 3x^2 - 9x - 10$
22.  $6x^3 + 25x^2 + 3x - 4$       24.  $4x^3 + 3x^2 - 25x + 6$       30. 3,200,000
32. a) 145,800,000      b) 48,600,000

*Chapter 1, §4*

1. a) If  $a = 2n$ ,  $b = 2m$ , then  $a + b = 2n + 2m = 2(n + m)$ .  
Setting  $k = n + m$  gives  $a + b = 2k$ , whence  $a + b$  is even.
- b) If  $a = 2n + 1$ ,  $b = 2m$ , then  $a + b = 2n + 1 + 2m = 2(n + m) + 1$ .  
Setting  $k = n + m$  gives  $a + b = 2k + 1$ , and  $a + b$  is odd.
2. Let  $a = 2n$ ,  $ab = 2nb$ .  
Setting  $k = nb$  gives  $ab = 2k$ , whence  $ab$  is even.
3. Let  $a = 2n$ ,  $a^3 = (2n)^3 = 2^3 n^3$ .  
Setting  $k = 2^2 n^3$  gives  $a^3 = 2k$ , whence  $a^3$  is even.
4. Let  $a = 2n + 1$ ,  $a^3 = (2n + 1)^3 = 8n^3 + 12n^2 + 6n + 1$ .  
Setting  $k = 4n^3 + 6n^2 + 3n$  gives  $a^3 = 2k + 1$ , whence  $a^3$  is odd.
5. Let  $n = 2k$ ,  $(-1)^n = (-1)^{2k} = ((-1)^2)^k = 1^k = 1$ , since  $(-1)^2 = +1$ .  
Then  $1^k = 1$  for all integers,  $k$ .
6. Let  $n = 2k + 1$ .  
Then  $(-1)^n = (-1)^{2k+1} = (-1)^{2k}(-1) = 1 \cdot (-1) = -1$ .
7. Let  $m = 2k + 1$ ,  $n = 2q + 1$ .  
Then  $mn = (2k + 1)(2q + 1) = 4kq + 2q + 2k + 1$ .  
Setting  $t = 2kq + q + k$  gives  $mn = 2t + 1$ . Hence  $mn$  is odd.
8. 4      10. 5      12. 1      14. 2

21. 50 is not divisible by 3. Thus  $50 = 3^0 \cdot 50; 0$ .
24. a) There exist integers  $k, q$  such that  $a - b = 5k, x - y = 5q$ . Then  $(a - b) + (x - y) = 5(k + q)$ , whence  $(a + x) - (b + y) = 5(k + q)$ , i.e.,  $a + x \equiv b + y \pmod{5}$ .
- b) There exist integers  $k, q$  such that  $a = b + 5k, x = y + 5q$ . Then  $ax = (b + 5k)(y + 5q) = by + 5(ky + bq + 5kq)$ . Setting  $t = ky + bq + 5kq$  gives  $ax = by + 5t$ , i.e.,  $ax \equiv by \pmod{5}$ .
26. For any positive integer  $a$ , either:
- a)  $a = 3k$ ,    b)  $a = 3k + 1$ , or    c)  $a = 3k + 2$ .
- a) Suppose  $a = 3k$ . Then  $a^2 = (3k)^2 = 9k^2 = 3(3k^2)$ .  $a^2$  is divisible by 3 and  $a$  is divisible by 3.
- b) Suppose  $a = 3k + 1$ . Then  $a^2 = (3k + 1)^2 = 9k^2 + 6k + 1 = 3(3k^2 + 2k) + 1$ .  $a^2$  is not divisible by 3 and  $a$  is not divisible by 3.
- c) Suppose  $a = 3k + 2$ . Then  $a^2 = (3k + 2)^2 = 9k^2 + 12k + 4 = 3(3k^2 + 4k + 1) + 1$ .  $a^2$  is not divisible by 3 and  $a$  is not divisible by 3.

### Chapter 1, §5

1. a)  $a = \frac{3}{8}$     b)  $a = -\frac{35}{3}$     c)  $a = -\frac{3}{20}$

2. a)  $x = \frac{5}{3}$     b)  $x = \frac{5}{2}$     c)  $x = -\frac{2}{7}$

3. a)  $\frac{2}{5}$     b)  $\frac{1}{3}$     c)  $\frac{6}{5}$     d)  $\frac{10}{3}$

4. Let  $b = n/m$  (this is a rational number). Then  $ab = m/n \cdot n/m = 1$ ;  $ba = n/m \cdot m/n = 1$ .

5. a) 14    b)  $\frac{11}{3}$     c)  $\frac{9}{32}$     d)  $\frac{1}{3}$     e)  $\frac{5}{4}$     f)  $\frac{1}{3}$     g)  $\frac{3}{5}$

6. a)  $\frac{50}{63}$     b)  $-\frac{20}{31}$

7. a)  $5! = 120$ ;  $6! = 720$ ;  $7! = 5,040$ ;  $8! = 40,320$

b)  $\binom{3}{0} = 1$ ;  $\binom{3}{1} = 3$ ;  $\binom{3}{2} = 3$ ;  $\binom{3}{3} = 1$ ;

$\binom{4}{0} = 1$ ;  $\binom{4}{1} = 4$ ;  $\binom{4}{2} = 6$ ;  $\binom{4}{3} = 4$ ;  $\binom{4}{4} = 1$ ;

$\binom{5}{0} = 1$ ;  $\binom{5}{1} = 5$ ;  $\binom{5}{2} = 10$ ;  $\binom{5}{3} = 10$ ;  $\binom{5}{4} = 5$ ;  $\binom{5}{5} = 1$

c)  $\binom{m}{m-n} = \frac{m!}{(m-n)!(m-(m-n))!} = \frac{m!}{(m-n)!(m-m+n)!}$

$$= \frac{m!}{(m-n)!n!} = \binom{m}{n}$$

d)  $\binom{m}{n} + \binom{m}{n-1} = \frac{m!}{n!(m-n)!} + \frac{m!}{(m-n+1)!(n-1)!}$

[common denominator  $n!(m-n+1)!$ ]

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

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

8. Suppose that there exists a rational number  $a = m/n$ , written in lowest form, such that  $a^3 = (m/n)^3 = 2$ . Then  $m^3 = 2n^3$ . Thus  $m^3$  is even, and hence  $m$  is even (proof similar to the one about  $m^2$  and  $m$ ). We can write  $m = 2p$  for some integer  $p$ . Thus  $m^3 = 2(4p^3)$ . Going back to  $m^3 = 2n^3$ , this yields  $2n^3 = 2(4p^3)$ , that is,  $n^3 = 4p^3 = 2(2p^3)$ . Consequently,  $n^3$  is even and  $n$  is even. Thus both  $m$  and  $n$  are even, which is impossible.
9. If  $a^4 = 2$ , then  $a^4 = (a^2)^2$ . But  $a^2$  is also a rational number, and Theorem 4 shows that  $(a^2)^2 = 2$  is impossible.
11. b)  $a = 1.4141$     12. b)  $a = 1.7321$     13. b)  $a = 2.236$   
 14. b)  $1.260$     16. a)  $3\text{ g}$     b)  $3/256\text{ g}$     c)  $3/2,048\text{ g}$   
 19. a)  $9 \cdot 10^5$     b)  $7.29 \cdot 10^5$     e) Between 110 and 120 min  
 20. b) 22,200    f) Between 4 and 5 mo  
 21. b) The population triples in 50 yr.

### Chapter 1, §6

1. a)  $-\frac{15}{19}$     c)  $\frac{25}{2}$     e)  $-\frac{13}{7}$
2. a)  $\frac{1}{x+y} - \frac{1}{x-y} = \frac{(x-y) - (x+y)}{(x+y)(x-y)} = \frac{x-y-x-y}{(x^2-y^2)} = \frac{-2y}{x^2-y^2}$   
 b)  $(x-1)(1+x+x^2) = x(1+x+x^2) - 1(1+x+x^2)$   
 $= x + x^2 + x^3 - 1 - x - x^2$   
 $= (x-x) + (x^2-x^2) + x^3 - 1$   
 Thus  $(x-1)(1+x+x^2) = x^3 - 1$ , or  

$$\frac{x^3 - 1}{x-1} = 1 + x + x^2.$$
- c)  $(x-1)(1+x+x^2+x^3)$   
 $= x(1+x+x^2+x^3) - 1(1+x+x^2+x^3)$   
 $= x + x^2 + x^3 + x^4 - 1 - x - x^2 - x^3 = x^4 - 1$   
 Thus  $(x-1)(1+x+x^2+x^3) = x^4 - 1$  or  

$$\frac{x^4 - 1}{x-1} = 1 + x + x^2 + x^3.$$
- d)  $(x-1)(x^{n-1} + x^{n-2} + \cdots + x + 1)$   
 $= x(x^{n-1} + x^{n-2} + \cdots + x + 1) - (x^{n-1} + x^{n-2} + \cdots + x + 1)$   
 $= x^n + x^{n-1} + \cdots + x - x^{n-1} - x^{n-2} - \cdots - x - 1 = x^n - 1$   
 Thus  $(x-1)(x^{n-1} + x^{n-2} + \cdots + x + 1) = x^n - 1$  or  

$$\frac{x^n - 1}{x-1} = x^{n-1} + \cdots + x + 1.$$
3. b)  $\frac{2x}{x+5} - \frac{3x+1}{2x+1} = \frac{2x(2x+1) - (x+5)(3x+1)}{(x+5)(2x+1)}$   
 Numerator gives  $(4x^2 + 2x) - (3x^2 + x + 15x + 5)$   
 $= x^2 - 14x - 5.$

Denominator gives  $2x^2 + x + 10x + 5 = 2x^2 + 11x + 5$ .

The quotient is equal to  $\frac{x^2 - 14x - 5}{2x^2 + 11x + 5}$ .

4. a) Use cross multiplication.

$$\begin{aligned}(x^2 + xy + y^2)(x - y) &= (x^2 + xy + y^2) - (x^2 + xy + y^2)y \\&= x^3 + x^2y + xy^2 - x^2y - xy^2 - y^3 \\&= x^3 - y^3\end{aligned}$$

$$\begin{aligned}\text{b)} \quad (x^3 + x^2y + xy^2 + y^3)(x - y) &= (x^3 + x^2y + xy^2 + y^3)x - (x^3 + x^2y + xy^2 + y^3)y \\&= x^4 + x^3y + x^2y^2 + xy^3 - x^3y - x^2y^2 - xy^3 - y^4 \\&= x^4 - y^4\end{aligned}$$

$$\begin{aligned}\text{c)} \quad x^2 &= \frac{(1 - t^2)^2}{(1 + t^2)^2} = \frac{1 - 2t^2 + t^4}{(1 + t^2)^2}, \quad y^2 = \frac{4t^2}{(1 + t^2)^2} \\&\text{Hence } x^2 + y^2 = \frac{1 - 2t^2 + t^4 + 4t^2}{(1 + t^2)^2} = \frac{1 + 2t^2 + t^4}{(1 + t^2)^2} \\&\quad = \frac{(1 + t^2)^2}{(1 + t^2)^2} = 1.\end{aligned}$$

5. b)  $(x + 1)(x^4 - x^3 + x^2 - x + 1)$

$$\begin{aligned}&= x(x^4 - x^3 + x^2 - x + 1) + (x^4 - x^3 + x^2 - x + 1) \\&= x^5 - x^4 + x^3 - x^2 + x + x^4 - x^3 + x^2 - x + 1 \\&= x^5 + 1\end{aligned}$$

Thus  $(x^5 + 1)/(x + 1) = x^4 - x^3 + x^2 - x + 1$ .

$$\begin{aligned}\text{c)} \quad (x + 1)(x^{n-1} - x^{n-2} + x^{n-3} - \dots - x + 1) &= x(x^{n-1} - x^{n-2} + \dots - x + 1) + (x^{n-1} - x^{n-2} + \dots - x + 1) \\&= x^n - x^{n-1} + x^{n-2} + \dots - x^2 + x + x^{n-1} - x^{n-2} \\&\quad + \dots + x^2 - x + 1 \\&= x^n + 1\end{aligned}$$

Hence  $(x^n + 1)/(x + 1) = x^{n-1} - x^{n-2} + x^{n-3} + \dots - x + 1$ .

6. 25/8 sec    7. a) 9/20 lb/in<sup>3</sup>    b) 9/2 lb/in<sup>3</sup>

8. a) 0°C    c) (335/9)°C    e) -40°C

9. a) 32°F    c) -40°F    d) 98.6°F    e) 104°F    f) 212°F

11. a) 9 cm<sup>3</sup>    12.  $\frac{5}{2}$  hr    13. 500 tickets at \$5.00 and 800 tickets at \$2.00

14. a) 120 g    15. a) 18 qt    18. a)  $\frac{5}{2}$  gal

## Chapter 2, §1

1.  $x = \frac{5}{3}$  and  $y = \frac{1}{3}$     3.  $x = -2$  and  $y = 1$

5.  $x = -\frac{1}{2}$  and  $y = -\frac{3}{2}$

9. a) Multiply the first equation by  $d$  and the second by  $b$ . Subtract each side of the second equation from the corresponding side of the first. The terms with  $y$  cancel, and you get  $adx - bcx = d - 2b$ , whence  $x = (d - 2b)/(ad - bc)$ . Multiply the first equation by  $c$  and the

second by  $a$ . Subtract the first from the second. The terms with  $x$  cancel, and you get  $ady - bcy = 2a - c$ , whence

$$y = \frac{2a - c}{ad - bc}.$$

### Chapter 2, §2

1.  $x = 1, y = \frac{1}{2}, z = -\frac{1}{2}$
3.  $x = 51/67, y = 29/67, z = 25/67$
5.  $x = \frac{1}{2}, y = -9/16, z = 13/16$
9.  $x = 205/43, y = -130/43, z = -94/43$
11.  $x = 32/11, y = -38/11, z = -51/11$

### Chapter 3, §2

1. a)  $a$  positive and  $a$  positive implies by POS 1 that  $aa = a^2$  is positive.  
b) If  $b$  is negative, then  $-b$  is positive. By POS 1,  $a(-b)$  is positive.  
But  $a(-b) = -ab$ . Hence  $-ab$  is positive and  $ab$  is negative.  
c)  $a$  negative implies  $-a$  positive;  $b$  negative implies  $-b$  positive. By POS 1,  $(-a)(-b)$  is positive. But  $(-a)(-b) = ab$ .
2. a) If  $a^{-1}$  were negative then  $aa^{-1} = 1$  would be negative, which is impossible. Hence  $a^{-1}$  is positive.
3. If  $a^{-1}$  were positive, then  $aa^{-1}$  would be negative, which is impossible because  $aa^{-1} = 1$ . Hence  $a^{-1}$  is negative.
4.  $\left(\sqrt{\frac{a}{b}}\right)^2 = \frac{a}{b}$  on one hand;  $\left(\frac{\sqrt{a}}{\sqrt{b}}\right)^2 = \frac{(\sqrt{a})^2}{(\sqrt{b})^2} = \frac{a}{b}$  on the other. Since  $\frac{\sqrt{a}}{\sqrt{b}}$  is positive, it must be equal to  $\sqrt{\frac{a}{b}}$ .
5.  $\frac{1}{1 - \sqrt{2}} = \frac{1 + \sqrt{2}}{(1 - \sqrt{2})(1 + \sqrt{2})} = \frac{1 + \sqrt{2}}{1 - (\sqrt{2})^2} = \frac{1 + \sqrt{2}}{-1}$   
 $= -(1 + \sqrt{2}).$
7.  $\frac{1}{3 + \sqrt{5}} = \frac{3 - \sqrt{5}}{3^2 - (\sqrt{5})^2} = \frac{3 - \sqrt{5}}{4}$ . Hence  $c = \frac{3}{4}$  and  $d = -\frac{1}{4}$ .
10.  $(x + y\sqrt{5})(z + w\sqrt{5}) = x(z + w\sqrt{5}) + y\sqrt{5}(z + w\sqrt{5})$   
 $= xz + xw\sqrt{5} + yz\sqrt{5} + 5yw$   
 $= (xz + 5yw) + (xw + yz)\sqrt{5}$

Let  $c = xz + 5yw$ ,  $d = xw + yz$ . These are rational numbers since  $x, y, z, w$  are rational.

12. a)  $\frac{x + 1}{2(\sqrt{2x + 3} - 2)}$    c)  $\frac{-1}{\sqrt{x - h} + \sqrt{x}}$    f)  $\frac{2}{\sqrt{x + 2h} + \sqrt{x}}$
13. a)  $x = 3$  and  $x = -1$    c)  $x = 7$  and  $x = -1$    e)  $x = -1$  and  $x = -7$
14. b)  $x = \frac{1}{3}$  and  $x = -1$    d)  $x = \frac{2}{3}$  and  $x = 0$

15. a)  $\frac{x - y}{x - 2\sqrt{x}\sqrt{y} + y}$  c)  $\frac{1}{x - \sqrt{x+1}\sqrt{x-1}}$

e)  $\frac{x + y - 1}{x + y + 3 + 4\sqrt{x+y}}$

16. a)  $\frac{x + y + 2\sqrt{x}\sqrt{y}}{x - y}$  c)  $x + \sqrt{x-1}\sqrt{x+1}$

17.  $\sqrt{x-1} = 3 + \sqrt{x}$  implies that  $x-1 = (3 + \sqrt{x})^2 = 9 + 6\sqrt{x} + x$ .  
Thus  $\sqrt{x} = -10/6$ . This is impossible since  $\sqrt{x}$  cannot be negative.

18.  $\sqrt{x-1} = 3 - \sqrt{x}$  implies that  $x-1 = 9 - 6\sqrt{x} + x$ . Thus,  
 $\sqrt{x} = 10/6 = 5/3$ , i.e.,  $x = 25/9$ .

19. a) No  $x$  b) No  $x$  c)  $x = 1$

20. For any number  $x$  we know that  $|x| = |-x|$ , because  $|x| = \sqrt{x^2} = \sqrt{(-x)^2} = |-x|$ . But  $b-a = -(a-b)$ , so put  $x = a-b$ .

### Chapter 3, §3

1. a)  $2^2 a^1 b^{-4}$  or  $2^2 3^0 a^1 b^{-4}$  c)  $2^{-1} 3^1 a^{-2} b^{-2}$  3. 8 4. No

5. No; for instance,  $(\sqrt{2})^5 = 4\sqrt{2}$ . If this were a rational number  $r$ , then  $\sqrt{2} = r/4$  would be rational, and this is not true.

6. b) 2 c) 27 g) 125 7. a) .3 c) .25 8. a)  $\frac{4}{9}$  b)  $\frac{2}{3}$

9. a)  $x = 2 + 5^{1/3}$  or  $x = 2 + \sqrt[3]{5}$  c)  $x = 14/3$  or  $x = 16/3$

f)  $x = \frac{-5\sqrt{8} \pm 1}{3\sqrt{8}}$  or  $x = \frac{-40 \pm \sqrt{8}}{24}$

### Chapter 3, §4

1. If  $a-b > 0$  and  $c < 0$ , then  $(a-b)c < 0$ . But  $(a-b) = ac-bc$ , so that  $ac < bc$ .

2.  $ac < bc < bd$ , using IN 2 twice.

3.  $ac > bc > bd$ , using IN 3 twice.

4. a)  $x$  positive implies  $y$  positive. Then  $1/x$  and  $1/y$  are also positive. Multiply each side of the inequality  $x < y$  by  $1/x$ . We get  $1 = x/x < y/x$ . Multiply each side of this last inequality by  $1/y$ . We get  $1/y < 1/x$ , as desired.

b) Multiply each side of the inequality  $a/b < c/d$  by  $d$ , and then by  $b$ . You get  $ad < bc$ . Conversely, multiply each side of the inequality  $ad < bc$  by  $1/d$  and  $1/b$ .

5. We must verify that  $(b+c)-(a+c)$  is positive. But  $(b+c)-(a+c) = b+c-a-c = b-a$ , which is positive by assumption. Also,  $(b-c)-(a-c) = b-c-a+c = b-a$ , which is positive. Hence  $(b-c) > (a-c)$ .

6. We have  $a^2 = aa < ab < bb = b^2$ , using IN 2 twice. Next, multiplying each side of the inequality  $a^2 < b^2$  by  $a$ , we get  $a^3 < ab^2$ , and multiply-

ing each side of the inequality  $a < b$  by  $b^2$  we get  $ab^2 < b^3$ , whence  $a^3 < b^3$ . Continuing in this way yields the general result.

7. Let  $r = a^{1/n}$  and  $s = b^{1/n}$ . If  $r \geq s$ , then we know from Exercise 6 that  $r^n \geq s^n$ , or in other words,  $a \geq b$ . But this contradicts our assumption. Hence  $r < s$ .
8. a) Use Exercise 4 (b). The inequality  $a/b < (a+c)/(b+d)$  is equivalent with  $a(b+d) < b(a+c)$  by cross multiplication, and this is equivalent with  $ab+ad < ba+bc$ , which is equivalent with  $ad < bc$ , which is true by multiplying each side of the inequality  $a/b < c/d$  by  $b$  and  $d$ .

As for the inequality on the right-hand side, it is equivalent by cross multiplication with  $d(a+c) < c(b+d)$ , which is equivalent with  $da+dc < cb+cd$ , which is equivalent with  $da < cb$ , which is equivalent by cross multiplication with  $a/b < c/d$ , which is true by assumption.

- b) The left inequality is equivalent by cross multiplication with  $a(b+rd) < b(a+rc)$ , which is equivalent with  $ab+ard < ba+brc$ , which is equivalent with  $ard < brc$ . Multiplying both sides by  $r^{-1}$ , which is positive, this last inequality is equivalent with  $ad < bc$ , which is true by cross multiplication of the assumed inequality  $a/b < c/d$ .

The argument for the right inequality works again by cross multiplication.

- c) By cross multiplication, it suffices to prove that  $(a+rc)(b+sd) < (a+sc)(b+rd)$ , and this is equivalent with  $ab+rcb+asd+rscd < ab+scb+ard+rscd$ , which is equivalent with  $rcb+asd < scb+ard$ , which is equivalent with  $r(bc-ad) < s(bc-ad)$ . Since  $r < s$  and  $bc-ad > 0$ , this last inequality is true, and our assertion is proved.

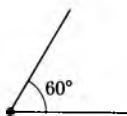
9.  $3x - 1 > 0$ ,  $3x > 1$ ,  $x > \frac{1}{3}$     11.  $x > -1$     13.  $x < -\frac{1}{3}$   
 19.  $-\frac{1}{2} < x < 2$     21.  $-\frac{1}{2} < x < 0$     23.  $\frac{1}{3} < x < \frac{1}{2}$   
 25.  $-1 < x < 1$     27.  $-\sqrt{3} < x < \sqrt{3}$     29.  $-1 > x > 1$

#### Chapter 4

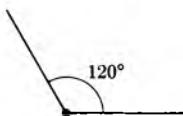
1.  $\frac{-3 \pm \sqrt{17}}{2}$     2.  $\frac{3 \pm \sqrt{17}}{2}$   
 3. Impossible with real numbers,  $(4 \pm \sqrt{-4})/2$     4.  $x = 5$  and  $x = -1$   
 7.  $\frac{-3 \pm \sqrt{57}}{6}$   
 8. Impossible with real numbers,  $(-5 \pm \sqrt{-31})/4$   
 13.  $\frac{-3 \pm \sqrt{9 + 4\sqrt{2}}}{2}$     14.  $\frac{3 \pm \sqrt{9 + 4\sqrt{5}}}{2}$

*Chapter 5, §2*

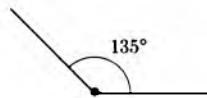
1. a)



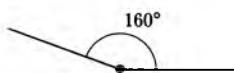
b)



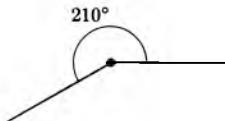
c)



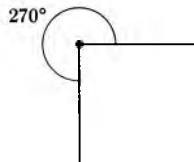
d)



e)



h)

Areas: a)  $1 \text{ in}^2$     b)  $2 \text{ in}^2$     c)  $135/60 \text{ in}^2$ 

2. a)  $\pi r^2 \left( \frac{\theta_2 - \theta_1}{360} \right)$     b)  $\pi(r_2^2 - r_1^2)$     c)  $\pi(r_2^2 - r_1^2) \left( \frac{\theta_2 - \theta_1}{360} \right)$

3. a)  $4\pi/9$     c)  $\pi/3$     4. a)  $21\pi$     c)  $32\pi$     5. a)  $16\pi/9$     c)  $4\pi/3$

6. a)  $7\pi/36$     c)  $77\pi/72$

*Chapter 5, §3*

1. a)  $2\sqrt{2}$     c)  $4\sqrt{2}$     2. a)  $\sqrt{5}$     c)  $\sqrt{65}$     e)  $r\sqrt{34}$

3. a)  $\sqrt{3}$  b)  $2\sqrt{3}$  e)  $r\sqrt{3}$  4. a)  $\sqrt{50}$  c)  $\sqrt{38}$

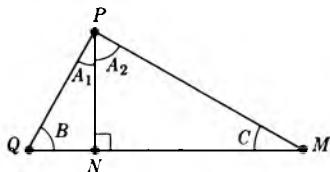
5. a)  $\sqrt{a^2 + b^2 + c^2}$  b)  $r\sqrt{a^2 + b^2 + c^2}$

6.  $100\sqrt{26}$  ft 7. a)  $\sqrt{51}$  ft 8. a)  $10\sqrt{40}$  ft

9.  $d(O, P)^2 + d(O, M)^2 = d(P, M)^2$   
 $d(O, Q)^2 + d(O, M)^2 = d(M, Q)^2$  (I)

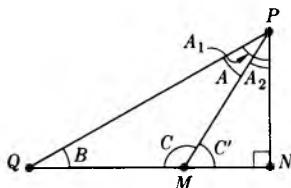
If  $d(O, P) = d(O, Q)$ , that is,  $d(O, P)^2 = d(O, Q)^2$ , then (I) gives  $d(P, M)^2 = d(M, Q)^2$ , that is,  $d(P, M) = d(M, Q)$ .

10. In the first case, the picture is as follows.



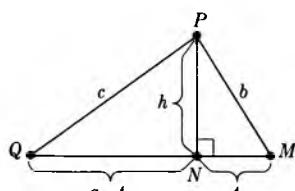
The triangles  $\triangle PQN$  and  $\triangle PMN$  are right triangles, and hence  $m(A_1) + m(B) = 90^\circ$ ,  $m(A_2) + m(C) = 90^\circ$ . Adding these, we find  $m(A_1) + m(A_2) + m(B) + m(C) = 180^\circ$ . But  $m(A) = m(A_1) + m(A_2)$ , so that we proved what we wanted.

In the second case, the picture is as follows.



Then  $\triangle PQN$  is a right triangle, and so is  $\triangle PMN$ , with right angle at N. Hence  $(*) m(A_1) + m(B) = 90^\circ$ ,  $m(A_2) + m(C') = 90^\circ$ , where  $C'$  is the supplementary angle to  $C$ , i.e.,  $m(C) + m(C') = 180^\circ$ , so that  $m(C) = 180^\circ - m(C')$ . Subtracting the expressions in (\*), we find  $m(A_1) - m(A_2) + m(B) - m(C') = 0^\circ$ . But  $m(A) = m(A_1) - m(A_2)$ . Substituting the value for  $m(C')$ , we get  $m(A) + m(B) + m(C) = 180^\circ$ .

11. a)

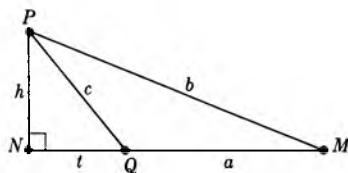


Area of  $\triangle PQN = \frac{1}{2}h(a - t)$

Area of  $\triangle PMN = \frac{1}{2}ht$

Area of  $\triangle PQM = \text{area of } \triangle PQN + \text{area of } \triangle PMN$   
 $= \frac{1}{2}h(a - t + t) = \frac{1}{2}ha$

b)

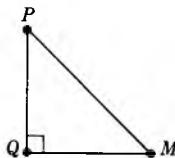


$$\text{Area of } \triangle PNM = \frac{1}{2}h(a + t)$$

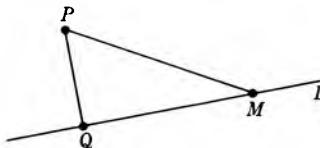
$$\text{Area of } \triangle PNQ = \frac{1}{2}ht$$

$$\begin{aligned}\text{Area of } \triangle PQM &= \text{area of } \triangle PNM - \text{area of } \triangle PNQ \\ &= \frac{1}{2}h(a + t - t) = \frac{1}{2}ha\end{aligned}$$

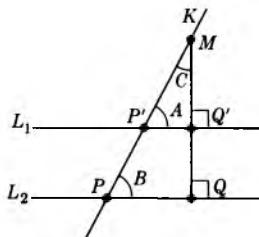
12. a) By Pythagoras  $d(P, M)^2 = d(P, Q)^2 + d(Q, M)^2$ . Then  $d(P, M)^2 \geq d(P, Q)^2$ . Thus  $d(P, M) \geq d(P, Q)$ , and similarly  $d(P, M) \geq d(Q, M)$ .



- b)  $d(P, M)^2 = d(P, Q)^2 + d(Q, M)^2$ . The distance is minimum when  $d(Q, M) = 0$ , i.e.,  $d(Q, M)^2 = 0$ . This occurs exactly when  $M = Q$ .



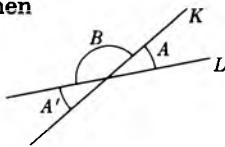
13. a) We can draw a perpendicular, from a point  $M$  on  $K$ , to  $L_2$  and this line is also perpendicular to  $L_1$ . Then  $\triangle MQ'P'$  is a right triangle, so  $m(C) + m(A) + 90^\circ = 180^\circ$ . Also  $\triangle MQP$  is a right triangle, so  $m(C) + m(B) + 90^\circ = 180^\circ$ . Combining these equations gives  $m(A) = m(B)$ .



- b) See preceding Theorem 1, §3.

- c) Let  $B$  be the angle as shown on the figure. Then

$m(A) + m(B) = 180^\circ$  and  
 $m(A') + m(B) = 180^\circ$ . Hence  
 $m(A) = m(A')$ . You could also  
argue by using (a) and (b).



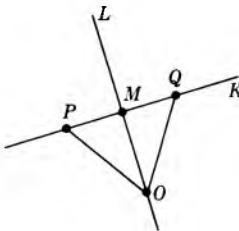
*Chapter 6, §1*

1. a) All points of the plane are fixed.
- b) The center of the reflection is the only fixed point.
- c) All the points on the reflection line are fixed.
- d) The center of the rotation is the only fixed point.
- e) No fixed point.
- f) The center of the dilation is the only fixed point if  $r \neq 1$ . If  $r = 1$ , the dilation is the identity.
2. a)  $(-1) 360 + 330$    c)  $(-1) 360 + 180$    g)  $(0) 360 + 120$   
i)  $(-2) 360 + 320$

*Chapter 6, §2*

4. The lines  $F(L)$  and  $F(K)$  cannot have a point in common, otherwise this point would be of the form  $F(P) = F(Q)$  for some point  $P$  on  $L$  and  $Q$  on  $K$ . But the distance between  $P$  and  $Q$  is the same as the distance between  $F(P)$  and  $F(Q)$ , and it would then follow that  $d(P, Q) = 0$ , so  $P = Q$ , which is impossible. Hence  $F(L)$  and  $F(K)$  have no point in common, and are therefore parallel.

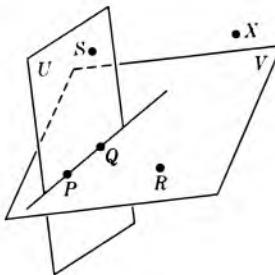
5.



Let  $M$  be the point of intersection of  $L$  and  $K$ . Let  $O$  be a point of  $L$ , not on  $K$ . Let  $P, Q$  be points on  $K$  lying on opposite sides of  $M$ , at the same distance from  $M$ . Thus  $d(O, P) = d(O, Q)$ . Since  $F$  is an isometry, we have  $d(F(O), F(P)) = d(F(O), F(Q))$ . Furthermore,  $F(P)$  and  $F(Q)$  are distinct, lie on  $F(K)$ , at the same distance from  $F(M)$ , again because  $F$  is an isometry. From the corollary of Pythagoras, it follows that  $F(O)$  lies on the perpendicular bisector of the segment between  $F(P)$  and  $F(Q)$ . The line  $F(L)$  is the unique line which passes through  $F(O)$  and  $F(M)$ , and is therefore perpendicular to the line  $F(K)$ , which is the unique line passing through  $F(P)$  and  $F(Q)$ .

6. Theorems 1 and 2 are valid.

**Theorem 4.** *If  $P, Q, R, S$  are four points which do not lie in a plane and are fixed by an isometry  $f$  of 3-space, then  $f$  is the identity.*



*Proof.* Observe first that  $P, Q, R$  do not lie on a line. Otherwise, this line and  $S$  lie in a plane, which would contradict the assumption of the theorem. Let  $V$  be the plane passing through  $P, Q, R$ . Arguing as above, the points  $S, P, Q$  do not lie on a line. Let  $U$  be the plane passing through  $S, P, Q$ . (We assume that if three points are not on a line, then there exists one and only one plane passing through them.) By Theorem 3, the isometry  $f$  leaves  $U$  and  $V$  fixed. Let  $X$  be a point in 3-space. We want to show that  $f(X) = X$ . If  $X$  lies in  $U$  or  $V$ , we are done. Assume that  $X$  does not lie in  $U$  or  $V$ . If  $X$  lies on the line between  $S$  and  $R$ , then  $X$  is fixed by Theorem 2. Assume that  $X$  does not lie on the line between  $S$  and  $R$ . Let  $W$  be the plane passing through  $X, S, R$ . We assume that the intersection of two planes is a line. Let  $L$  be the intersection of  $W$  and  $U$ . Since  $L$  lies in  $U$ , it is fixed by  $f$ . Since  $R$  is not in the plane  $U$ , it follows that  $R$  is not in  $L$ . We assume that given a line and a point not on the line, there exists a unique plane passing through the line and the point. Thus  $W$  is the unique plane passing through  $R$  and  $L$ . Any two points on  $L$  are fixed by  $f$ , and hence by Theorem 3, the plane  $W$  is fixed by  $f$ . Since  $X$  lies in  $W$ , we conclude that  $X$  is fixed by  $f$ , thereby proving Theorem 4.

### Chapter 6, §3

1.  $n = 2$
5. a) 4   c) 8   e) 24

### Chapter 6, §4

1. a)  $F(S)$  is contained in  $S'$ . *Proof.* Let  $Q$  be on  $S$ ,  $d(P, Q) = r$ . Since  $F$  is an isometry, we have  $d(F(P), F(Q)) = r$  and  $F(Q)$  is on  $S'$ .  $S'$  is contained in  $F(S)$ . *Proof.* Let  $Q'$  be on  $S'$ . Using  $F^{-1}$ , we have similarly  $F^{-1}(Q')$  on  $S$ , and thus  $F(F^{-1}(Q')) = Q'$  is in  $F(S)$ .

2. There exists a translation  $T$  such that  $T(P') = P$ . This reduces our proof to the case when  $P = P'$ , which we now assume. Since  $d(P, Q) = d(P, Q')$ , there exists a rotation  $R$  with respect to  $P$  such that  $R(Q) = Q'$ . This concludes the proof.

3. Since  $F \circ H = F \circ G$ , we get  $F^{-1} \circ (F \circ H) = F^{-1} \circ (F \circ G)$ .

Using the associativity of mappings we get  $(F^{-1} \circ F) \circ H = (F^{-1} \circ F) \circ G$ , whence  $H = G$ .

4. a)  $F^2 = I$  and  $F^3 = I$ . But  $F^3 = F \circ F^2$ , thus  $F^3 = I = F \circ I$ , i.e.  $F = I$ .

$F^8 = F^3 \circ F^5$ , thus  $I = F^3 \circ I$ , i.e.  $F^3 = I$ .  $F^5 = F^3 \circ F^2$ , thus  $I = I \circ F^2$ , i.e.  $F^2 = I$ .  $F^3 = F \circ F^2$ , thus  $I = F \circ I$ , i.e.  $F = I$ .

5.  $F(P) = G(P)$  implies  $P = (F^{-1} \circ G)(P)$ .  $F(Q) = G(Q)$  implies  $Q = (F^{-1} \circ G)(Q)$ .  $F(M) = G(M)$  implies  $M = (F^{-1} \circ G)(M)$ .  $F^{-1} \circ G$  is an isometry with 3 fixed points  $P, Q, M$ . Hence  $F^{-1} \circ G = I$ , whence  $F \circ F^{-1} \circ G = F$ , i.e.  $F = G$ .

8.  $F^{-1} = F^{n-1}$

$$9. V = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix}, \quad H \circ V = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{bmatrix}$$

$$10. H \circ G \circ H = G^3, \quad n = 3, \quad G = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{bmatrix},$$

$$G^3 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{bmatrix}, \quad H \circ G = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{bmatrix},$$

$$H \circ G^3 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{bmatrix}, \quad G^2 \circ H = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix}$$

11.	$I$	$G$	$G^2$	$G^3$	$H$	$HG$	$HG^2$	$HG^3$
$I$	$I$	$G$	$G^2$	$G^3$	$H$	$HG$	$HG^2$	$HG^3$
$G$	$G$	$G^2$	$G^3$	$I$	$HG^3$	$H$	$HG$	$HG^2$
$G^2$	$G^2$	$G^3$	$I$	$G$	$HG^2$	$HG^3$	$H$	$HG$
$G^3$	$G^3$	$I$	$G$	$G^2$	$HG$	$HG^2$	$HG^3$	$H$
$H$	$H$	$HG$	$HG^2$	$HG^3$	$I$	$G$	$G^2$	$G^3$
$HG$	$HG$	$HG^2$	$HG^3$	$H$	$G^3$	$I$	$G$	$G^2$
$HG^2$	$HG^2$	$HG^3$	$H$	$HG$	$G^2$	$G^3$	$I$	$G$
$HG^3$	$HG^3$	$H$	$HG$	$HG^2$	$G$	$G^2$	$G^3$	$I$

$$13. I = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}, \quad VG = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}, \quad VG^2 = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{bmatrix}$$

14.  $k = 5$ ,  $G^6 = I$

15. b)  $G^2 = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 1 & 2 \end{bmatrix}$ ,  $H = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 3 & 2 & 1 & 6 & 5 \end{bmatrix}$ ,

$$HG = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 6 & 5 & 4 \end{bmatrix}$$

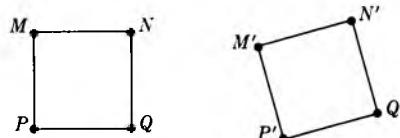
c)  $VG = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix}$ ,  $VG^4 = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 6 & 5 & 4 \end{bmatrix}$

*Chapter 6, §5*

1. By Theorems 4, 5, and 6 an isometry  $F$  can be written as a composite of isometries  $F_1$ , or  $F_1 \circ F_2$ , or  $F_1 \circ F_2 \circ F_3$ , such that each one of  $F_1, F_2, F_3$  has an inverse (because they are the identity, or a translation, or a reflection, or a rotation). Then  $F$  itself has an inverse, given in these cases by  $F_1^{-1}$ ,  $F_2^{-1} \circ F_1^{-1}$ , and  $F_3^{-1} \circ F_2^{-1} \circ F_1^{-1}$ .

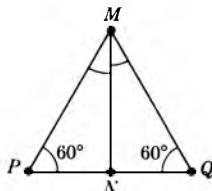
*Chapter 6, §6*

2.  $S$  congruent to  $S'$  means that there exists an isometry  $F_1$  such that  $F_1(S) = S'$ . Furthermore,  $S'$  congruent to  $S''$  means that there exists an isometry  $F_2$  such that  $F_2(S') = S''$ . But then  $(F_2 \circ F_1)(S) = F_2(F_1(S)) = S''$ . Since  $F_2 \circ F_1$  is an isometry, it follows that  $S$  is congruent to  $S''$ .
3. Let the two squares have corners at  $P, Q, M, N$  and  $P', Q', M', N'$  respectively, as shown. Since  $\overline{PQ}$  and  $\overline{P'Q'}$  have the same length, there exists an isometry which maps  $P$  on  $P'$  and  $Q$  on  $Q'$ , say by Theorem 8.



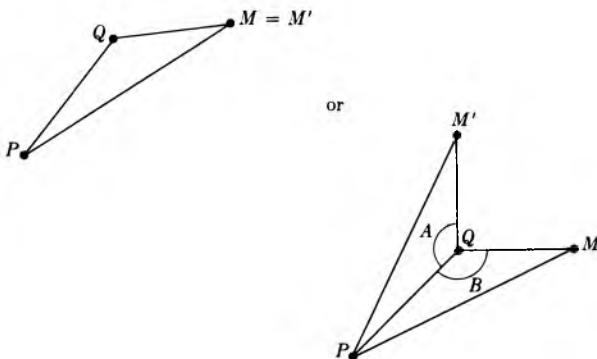
Thus we may assume that  $P = P'$  and  $Q = Q'$ . In this case, either  $M = M'$  or  $M \neq M'$ . If  $M = M'$ , then  $N$  is the point of intersection of the line through  $M$  parallel to  $\overline{PQ}$ , and the line through  $Q$ , perpendicular to  $\overline{PQ}$ . Similarly,  $N'$  is this same point of intersection, so that  $N = N'$ . It follows that in the present case, the squares coincide. If  $M \neq M'$ , then the reflection through the line  $L_{PQ}$  maps  $M$  on  $M'$ , and reduces our problem to the preceding case.

5.



We draw the triangle  $\triangle PQM$ . Let the line perpendicular to  $\overline{PQ}$ , passing through  $M$ , intersect  $\overline{PQ}$  at the point  $N$ , as shown. Then  $\triangle MNP$  and  $\triangle MNQ$  are right triangles. By the known theorem on the sum of the angles of a right triangle, it follows that the angles  $\angle PMN$  and  $\angle NMQ$  both have  $30^\circ$ . Hence if we reflect the line  $L_{PM}$  through the line  $L_{MN}$ , we obtain the line  $L_{MQ}$ . This reflection maps the line  $L_{PQ}$  on itself. Since  $P$  is the intersection of  $L_{PM}$  and  $L_{PQ}$ , its image by the reflection is the intersection of  $L_{MQ}$  and  $L_{PQ}$ , which is  $Q$ . Hence our reflection maps  $\overline{MP}$  on  $\overline{MQ}$ , and these two sides therefore have the same length. The same argument applies to another pair of sides.

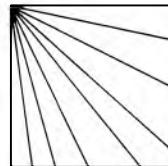
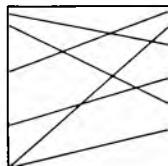
7. By an isometry we can map  $P$  on  $P'$  and  $Q$  on  $Q'$ . This reduces us to the case when  $P = P'$  and  $Q = Q'$ , which we now assume. The picture is then as follows.



The angles of the two triangles with vertex at  $Q$  have the same measure. If  $M$  and  $M'$  lie on the same side of the line  $L_{PQ}$ , then they lie on the same ray with vertex at  $Q$ . Since  $d(Q, M) = d(Q, M')$ , it follows that  $M = M'$ , and we are done. If  $M$  and  $M'$  lie on opposite sides of the line  $L_{PQ}$ , then we reflect through this line to get them on the same side, and this reduces our problem to the case already taken care of.

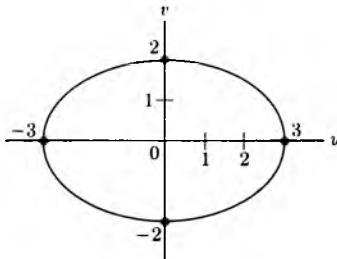
9. It is the set of all line segments  $\overline{PQ}$ , where  $P$  lies on one side of the square and  $Q$  lies on the other. There are other ways also. For instance,

the set of all line segments from one corner to the points on the opposite two sides. These are shown on the next figures.



### Chapter 7, §1

1. a)  $1 \text{ in}^2$  b)  $4 \text{ in}^2$
2. a)  $4/3 \text{ in}^2$  c)  $36/3 \text{ in}^2$  d)  $\frac{1}{3} \text{ in}^2$
3. Multiplied by  $ab$ .
4. Area is  $\pi ab$ ; each square of area 1 is dilated into a square of area  $ab$ .  
Hence the area  $\pi$  of the disc of radius 1 is dilated into  $\pi ab$ .



5. a)  $28\pi$  c)  $\pi\sqrt{18}$
6. a)  $\pi/\sqrt{12}$  c)  $\pi/6$
7.  $F_r(x, y, z) = (rx, ry, rz)$ .
8. a) Each side dilates by  $r$ , thus the volume is multiplied by  $r^3$ .  
b) Using the approximation by small cubes, the volume will be multiplied by  $r^3$ .  
c)  $\frac{4}{3}\pi r^3$
9.  $x^2 + y^2 + z^2 = 1$     10.  $V = abc$     11. c)  $abcV$     12.  $56\pi$
13.  $\frac{20}{3}\sqrt{6}\pi$     14.  $\frac{4}{3}\pi abc$

### Chapter 8, §1

3.  $x < 0$  and  $y > 0$
4.  $x < 0$  and  $y < 0$

### Chapter 8, §2

1.  $\sqrt{17}$     3.  $\sqrt{61}$     5.  $\sqrt{106}$

11.  $d(P, Q) = 0$ , so that  $d(P, Q)^2 = 0 = (x_1 - x_2)^2 + (y_1 - y_2)^2$ . Thus  $(x_1 - x_2)^2 = -(y_1 - y_2)^2$  is impossible with real numbers unless  $x_1 - x_2 = 0$  and  $y_1 - y_2 = 0$ , i.e.  $x_1 = x_2$  and  $y_1 = y_2$ , which means  $P = Q$ .
12.  $d(A, B) = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2}$   
 $d(rA, rB) = \sqrt{(ra_1 - ra_2)^2 + (rb_1 - rb_2)^2}$   
 $= \sqrt{r^2[(a_1 - a_2)^2 + (b_1 - b_2)^2]}$   
 Thus  $d(rA, rB) = r d(A, B)$ .
13. a)  $\sqrt{41}$  c)  $\sqrt{50}$

*Chapter 8, §3*

1.  $(x + 3)^2 + (y - 1)^2 = 4$     3.  $(x + 1)^2 + (y + 2)^2 = \frac{1}{5}$   
 7.  $C = (1, 2)$  and  $r = 5$     9.  $C = (-1, 9)$  and  $r = \sqrt{8}$   
 11.  $C = (5, 0)$  and  $r = \sqrt{10}$   
 13.  $(x + 1)^2 + y^2 = 6$ ;  $C = (-1, 0)$  and  $r = \sqrt{6}$   
 15.  $(x + 2)^2 + (y - 2)^2 = 28$ ;  $C = (-2, 2)$  and  $r = \sqrt{28}$   
 17.  $(x - 1)^2 + (y + 5/2)^2 = \frac{133}{4}$ ;  $C = (1, -5/2)$  and  $r = \frac{\sqrt{133}}{2}$   
 20. a)  $(x - 1)^2 + (y + 3)^2 + (z - 2)^2 = 1$   
 c)  $(x + 1)^2 + (y - 1)^2 + (z - 4)^2 = 9$   
 e)  $(x + 2)^2 + (y + 1)^2 + (z + 3)^2 = 4$

*Chapter 8, §4*

2. By cross multiplication for inequalities, it suffices to prove that  $(1 - s^2)(1 + t^2) > (1 + s^2)(1 - t^2)$ , which is equivalent to  $1 - s^2 + t^2 - s^2t^2 > 1 + s^2 - t^2 - s^2t^2$ . This is equivalent to  $2t^2 > 2s^2$ , which is true because  $0 \leq s < t$ .
3. c)  $x = 15/17$  and  $y = 8/17$
4. a)  $\frac{1 - t^2}{1 + t^2}$  approaches  $-1$     b)  $\frac{1 - t^2}{1 + t^2}$  approaches  $-1$
5. In both cases  $\frac{2t}{1 + t^2}$  approaches 0.

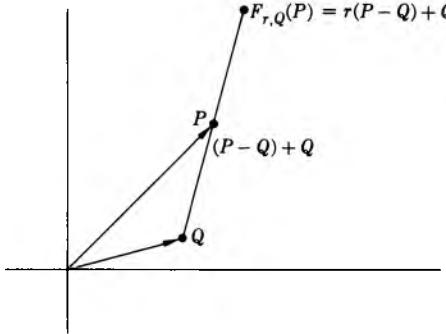
*Chapter 9, §1*

1. a)  $cA = (-12, 20)$     c)  $cA = (-8, -10)$     e)  $cA = (4, 5)$   
 2.  $bA = (ba_1, ba_2)$  and  $cA = (ca_1, ca_2)$ .  $bA = cA$  gives  $ba_1 = ca_1$  and  $ba_2 = ca_2$ , whence  $b = c$ .

3.  $d^2(-A, -B) = (-a_1 - (-b_1))^2 + (-a_2 - (-b_2))^2$   
 $= (b_1 - a_1)^2 + (b_2 - a_2)^2 = d^2(A, B)$
4. a)  $cA = (ca_1, ca_2, ca_3)$ , stretching all the coordinates by  $c$   
b)  $-A = (-a_1, -a_2, -a_3)$

*Chapter 9, §2*

1.  $A + B = (4, 6)$  and  $A - B = (-2, 2)$
3.  $A + B = (2, 3)$  and  $A - B = (-4, 1)$
5.  $A + B = (-2, 3)$  and  $A - B = (0, -1)$
11. We have  $T_A(P) = P + A$  and  $T_A(Q) = Q + A$ . Hence  
 $d(T_A(P), T_A(Q)) = |T_A(P) - T_A(Q)| = |P + A - Q - A|$   
 $= |P - Q| = d(P, Q).$   
Under translation by  $A$ , we have  $d(M, O) = d(M + A, A)$ , that is,  
 $|M| = |M + A - A|$ , valid for all points of the disc.
14.  $R_Q(P) = Q - A = Q - (P - Q) = Q - P + Q = 2Q - P$
15. a) Reflection of  $M$  through  $O$  is  $-M$ . The translation of  $-M$  by  $2Q$  is  $2Q - M$ .  
b) By Exercise 14, we must have  $-P + A = 2Q - P$ . Hence  $Q = \frac{1}{2}A$ .
16. a)  $F_{r,Q}(P) = r(P - Q) + Q$



- b) Translation by  $(1 - r)Q$
17. Reflection of  $M$  through  $Q$  is  $2Q - M$ . If  $d(M, A) = r$ , that is  $|M - A| = r$ , then  $|(2Q - M) - (2Q - A)| = |M - A| = r$ . The reflection of  $S(r, A)$  is the circle of center  $2Q - A$  and of radius  $r$ .
18. By  $-A$ , because  $T_{-A} \circ T_A = T_A \circ T_{-A} = I$ .
19. a)  $F_r^{-1} = F_{1/r}$ , because

$$(F_r \circ F_{1/r})(P) = r \circ \left( \frac{1}{r} P \right) = P,$$

$$(F_{1/r} \circ F_r)(P) = \frac{1}{r} \circ (rP) = P$$

20.  $(T_A \circ T_B)(P) = T_A(P + B) = (P + B) + A = P + (B + A)$   
 $= T_{A+B}(P);$  translation by  $A + B$
21. a)  $R(P) = -P,$  so that  $(R \circ R)(P) = -(-P) = P$  and  $R^{-1} = R$   
 b)  $(R \circ T_A \circ R^{-1})(P) = (R \circ T_A)(-P) = R(A - P) = P - A$   
 $= T_A(P);$  translation by  $-A$
22. a) Image is  $(1, -2)$  c) Image is  $(-2, 4)$
23.  $|R_x(P) - R_x(Q)|^2 = (p_1 - q_1)^2 + (-p_2 - (-q_2))^2$   
 $= (p_1 - q_1)^2 + (-p_2 + q_2)^2 = |P - Q|^2$
25. a)  $T_A(P) = P$  means  $P + A = P,$  in which case  $A = 0,$  i.e.  $T_A = I.$   
 Then all points are fixed.  
 b)  $R_o(P) = -P$  means  $-P = P,$  i.e.  $P = O.$   
 c)  $R_P(M) = M$  means  $2P - M = M,$  i.e.  $P = M.$   
 d)  $R_x(P) = P$  means  $(x, -y) = (x, y),$  i.e.  $y = 0.$   
 e)  $R_y(P) = P$  means  $(-x, y) = (x, y),$  i.e.  $x = 0.$
33.  $r = |A|;$   $\left| \frac{1}{r} A \right| = \left( \frac{1}{r} \right) |A|$  (since  $r > 0) = \frac{|A|}{|A|} = 1$

### Chapter 10, §2

1. a)  $(2, 2)$  b)  $(5/3, 3)$  c)  $(7/3, 1)$
5. The segment  $\overline{PQ}$  consists of all points of the form  $tP + (1 - t)Q,$   $0 \leq t \leq 1.$  Apply  $T_A$  to such points. We get  $tP + (1 - t)Q + A = t(P + A) + (1 - t)(Q + A),$   $0 \leq t \leq 1,$  because  $tA + (1 - t)A = A.$  Therefore the image of  $\overline{PQ}$  by  $T_A$  consists of all points on the segment whose end points are  $P + A$  and  $Q + A.$
12.  $\overrightarrow{PQ}$  and  $\overrightarrow{NM}$  have opposite directions means that  $(Q - P) = c(M - N),$  for some number  $c < 0.$   $A$  and  $B$  have opposite directions means  $A = cB,$  for some number  $c < 0.$
13. a)  $(\frac{5}{3}, 2, \frac{13}{3})$  b)  $(\frac{1}{3}, 3, \frac{11}{3})$

### Chapter 10, §3

1. a)  $(2t + 1, 6t - 1)$  b)  $(\frac{4}{3}, 0)$  c)  $(0, -4)$
3. a)  $(3t - 4, 3t - 2)$  b)  $(-2, 0)$  c)  $(0, 2)$
7.  $(43, -15)$  8.  $(-\frac{1}{7}, -\frac{5}{7})$  11. a)  $(\frac{7}{4}, \frac{5}{4})$
12.  $A$  and  $B$  are parallel if and only if  $A = cB,$  with some number  $c \neq 0;$  that is,  $(a_1, a_2) = (cb_1, cb_2).$  This means  $a_1 = cb_1$  and  $a_2 = cb_2,$  or in other words,  $a_1a_2 = a_2b_1 = a_1b_2,$  whence  $c(a_2b_1 - a_1b_2) = 0.$  This implies  $a_1b_2 - a_2b_1 = 0.$
13. The first line consists of all points  $(p_1 + ta_1, p_2 + ta_2),$   $t$  in  $\mathbf{R}.$   
 The second line consists of all points  $(q_1 + sb_1, q_2 + sb_2),$   $t$  in  $\mathbf{R}.$

The common point is determined by the values of  $t, s$  such that

$$\begin{cases} p_1 + ta_1 = q_1 + sb_1 \\ p_2 + ta_2 = q_2 + sb_2 \end{cases} \quad \text{or} \quad \begin{cases} ta_1 - sb_1 = q_1 - p_1 \\ ta_2 - sb_2 = q_2 - p_2. \end{cases}$$

Since  $a_1 b_2 \neq a_2 b_1$  [Exercise 12, lines not parallel] we get the values

$$s = \frac{a_2(p_1 - q_1) - a_1(p_2 - q_2)}{a_2 b_1 - a_1 b_2} \quad \text{and} \quad t = \frac{b_1(p_2 - q_2) - b_2(p_1 - q_1)}{a_2 b_1 - a_1 b_2}.$$

14. a)  $\left( \frac{-1 - 3\sqrt{159}}{5}, \frac{-3 + \sqrt{159}}{5} \right)$  and  $\left( \frac{-1 + 3\sqrt{159}}{5}, \frac{-3 - \sqrt{159}}{5} \right)$

b)  $\left( \frac{5 + \sqrt{103}}{2}, \frac{-5 + \sqrt{103}}{2} \right)$  and  $\left( \frac{5 - \sqrt{103}}{2}, \frac{-5 - \sqrt{103}}{2} \right)$

15. a)  $\left( \frac{-1 - 6\sqrt{6}}{5}, \frac{-3 + 2\sqrt{6}}{5} \right)$  and  $\left( \frac{-1 + 6\sqrt{6}}{5}, \frac{-3 - 2\sqrt{6}}{5} \right)$

b) No intersection

16. a), b), c), d) No intersection

18. The line is described by  $(p + at, q + bt)$ ,  $t$  in  $\mathbf{R}$ . The intersections are given by those values of  $t$  such that  $(p + at)^2 + (q + bt)^2 = r^2$ , that is  $(a^2 + b^2)t^2 + 2(ap + bq)t + (p^2 + q^2 - r^2) = 0$ . Then

$$t = \frac{-(ap + bq) \pm \sqrt{(ap + bq)^2 + (a^2 + b^2)(r^2 - p^2 - q^2)}}{a^2 + b^2}.$$

The square root is defined since  $(ap + bq)^2 \geq 0$ ,  $(a^2 + b^2) \geq 0$ ,  $r^2 \geq p^2 + q^2$ . Finally  $P + tA$  are the two points, for the two values of  $t$  above.

### Chapter 10, §4

1.  $3x + 4y = 13$     2.  $2x + 7y = 31$     3.  $3x - 11y = 10$

### Chapter 11, §1

1. a)  $\pi/12$     c)  $7\pi/12$     2. a)  $\pi/9$     b)  $2\pi/9$   
 3. a)  $315^\circ$     c)  $100^\circ$     e)  $120^\circ$     g)  $300^\circ$

### Chapter 11, §2

1. a)

$n$	1	2	3	4	5	6	7	8	9	10	11	12
$\sin \frac{n\pi}{6}$	$\frac{1}{2}$	$\sqrt{3}/2$	1	$\sqrt{3}/2$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\sqrt{3}/2$	-1	$-\sqrt{3}/2$	$-\frac{1}{2}$	0
$\cos \frac{n\pi}{6}$	$\sqrt{3}/2$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\sqrt{3}/2$	-1	$-\sqrt{3}/2$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\sqrt{3}/2$	1
$\frac{n\pi}{6}$	$30^\circ$	$60^\circ$	$90^\circ$	$120^\circ$	$150^\circ$	$180^\circ$	$210^\circ$	$240^\circ$	$270^\circ$	$300^\circ$	$330^\circ$	$360^\circ$

b)

$n$	1	2	3	4	5	6	7	8
$\sin \frac{n\pi}{4}$	$\sqrt{2}/2$	1	$\sqrt{2}/2$	0	$-\sqrt{2}/2$	-1	$-\sqrt{2}/2$	0
$\cos \frac{n\pi}{4}$	$\sqrt{2}/2$	0	$-\sqrt{2}/2$	-1	$-\sqrt{2}/2$	0	$\sqrt{2}/2$	1
$\frac{n\pi}{4}$	$45^\circ$	$90^\circ$	$135^\circ$	$180^\circ$	$225^\circ$	$270^\circ$	$315^\circ$	$360^\circ$

2. a)

$n$	1	2	3	4	5	6	7	8	9	10	11	12
$\sin \frac{-n\pi}{6}$	$-\frac{1}{2}$	$-\sqrt{3}/2$	-1	$-\sqrt{3}/2$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\sqrt{3}/2$	1	$\sqrt{3}/2$	$\frac{1}{2}$	0
$\cos \frac{-n\pi}{6}$	$\sqrt{3}/2$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\sqrt{3}/2$	-1	$-\sqrt{3}/2$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\sqrt{3}/2$	1
$\frac{-n\pi}{6}$	$330^\circ$	$300^\circ$	$270^\circ$	$240^\circ$	$210^\circ$	$180^\circ$	$150^\circ$	$120^\circ$	$90^\circ$	$60^\circ$	$30^\circ$	$0^\circ$

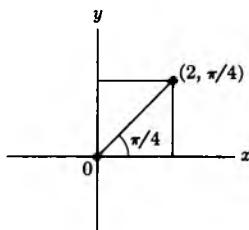
b)

$n$	1	2	3	4	5	6	7	8
$\sin \frac{-n\pi}{4}$	$-\sqrt{2}/2$	-1	$-\sqrt{2}/2$	0	$\sqrt{2}/2$	1	$\sqrt{2}/2$	0
$\cos \frac{-n\pi}{4}$	$\sqrt{2}/2$	0	$-\sqrt{2}/2$	-1	$-\sqrt{2}/2$	0	$\sqrt{2}/2$	1
$\frac{-n\pi}{4}$	$315^\circ$	$270^\circ$	$225^\circ$	$180^\circ$	$135^\circ$	$90^\circ$	$45^\circ$	$0^\circ$

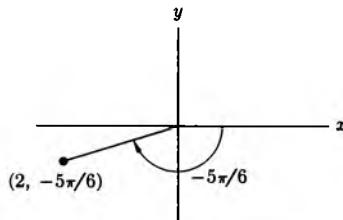
3. Sine is negative. Cosine is positive.

4. a)  $400\sqrt{3}$  ft5. a)  $5/3$  mi

6. a) 1 mi

7. a)  $x = \sqrt{2}$  and  $y = \sqrt{2}$ 

d)  $x = -\sqrt{3}$  and  $y = -1$



8. a)  $r = \sqrt{2}$ ,  $\theta = \pi/4$  c)  $r = 6$ ,  $\theta = \pi/3$

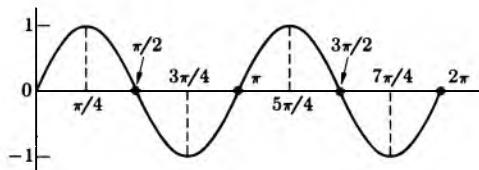
9. a)  $40/3 = 13.3$  b) 10 c)  $5/2 = 2.5$  10. a) .8 b) 1.8

11. a) .3 b) 1.2 12. a)  $\pi/6$  c)  $\pi/3$

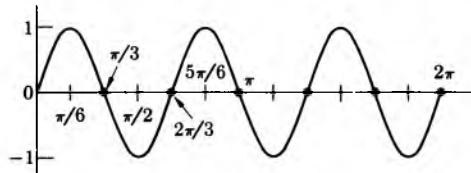
13. a)  $-\pi/6$  c)  $-\pi/3$  e)  $-\pi/4$

*Chapter 11, §3*

2.



3. b)



*Chapter 11, §4*

1.

$n$	1	2	3	4	5	6	7	8	9	10	11	12
$\tan \frac{n\pi}{6}$	$\sqrt{3}/3$	$\sqrt{3}$		$-\sqrt{3}$	$-\sqrt{3}/3$	0	$\sqrt{3}/3$	$\sqrt{3}$		$-\sqrt{3}$	$-\sqrt{3}/3$	0
$\frac{n\pi}{6}$	$30^\circ$	$60^\circ$	$90^\circ$	$120^\circ$	$150^\circ$	$180^\circ$	$210^\circ$	$240^\circ$	$270^\circ$	$300^\circ$	$330^\circ$	$360^\circ$

$n$	1	2	3	4	5	6	7	8
$\tan \frac{n\pi}{4}$	1		-1	0	1		-1	0
$\frac{n\pi}{4}$	$45^\circ$	$90^\circ$	$135^\circ$	$180^\circ$	$225^\circ$	$270^\circ$	$315^\circ$	$360^\circ$

The tangent is not defined at those points where the space is left blank.

2. Note that  $1/\cos x$  decreases from arbitrarily large values to 1, in the interval  $-\pi/2 \leq x \leq 0$ . Furthermore,  $-\sin x$  decreases from 1 to 0 in this interval. Since  $\cos x$  and  $-\sin x$  are both positive in this interval, we conclude that  $-\tan x = -\sin x/\cos x$  decreases from arbitrarily large values to 0. Hence  $\tan x$  itself increases from arbitrarily large negative values to 0.

5.  $1 + \tan^2 x = 1 + \frac{\sin^2 x}{\cos^2 x} = \frac{\sin^2 x + \cos^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$

7. a)  $\sqrt{2}/2$  and  $-\sqrt{2}/2$  c)  $\frac{1}{2}$  and  $-\frac{1}{2}$

8. a)  $\sqrt{2}/2$  and  $-\sqrt{2}/2$  c)  $\sqrt{3}/2$  and  $-\sqrt{3}/2$

9. a) 500 ft c)  $\frac{500\sqrt{3}}{3}$  ft 11. a)  $\frac{1}{2\sqrt{100 + \frac{1}{4}}}$

12. a)  $2\sqrt{3}$  c)  $7 - 2\sqrt{3}$  f)  $16 - 7\sqrt{3}$

14. a)  $6 \tan \theta$  b)  $6\sqrt{1 + \tan^2 \theta}$  c)  $(7 - 6 \tan \theta)\sqrt{1 + \frac{1}{\tan^2 \theta}}$

15. a)  $(6 \sin \theta)/\sqrt{1 - \sin^2 \theta}$  b)  $6/\sqrt{1 - \sin^2 \theta}$

### Chapter 11, §5

1.  $\frac{\sqrt{2}}{4}[1 + \sqrt{3}]$

3. a)  $(\sqrt{3} - 1)(\sqrt{2}/4)$  c)  $(\sqrt{3} + 1)(\sqrt{2}/4)$  e)  $(\sqrt{3} - 1)(\sqrt{2}/4)$

4. a)  $\sin(x + x) = \sin x \cos x + \sin x \cos x = 2 \sin x \cos x$

b)  $\cos(x + x) = \cos x \cos x - \sin x \sin x = \cos^2 x - \sin^2 x$

c)  $\cos 2x = \cos^2 x + (\cos^2 x - 1) = 2 \cos^2 x - 1$ ,

i.e.  $\cos^2 x = \frac{1 + \cos 2x}{2}$

d)  $\cos 2x = (1 - \sin^2 x) - \sin^2 x = 1 - 2 \sin^2 x$ ,

i.e.  $\sin^2 x = \frac{1 - \cos 2x}{2}$

5. a)  $1.4\sqrt{0.51}$  b) 0.96 6. a) 0.02 b) 0.88 c) 0.68

7. a)  $\sqrt{0.85}$  b)  $\sqrt{0.8}$

10.  $\sin 3x = 3 \sin x \cos^2 x - \sin^3 x$

$\sin 4x = 4 \sin x \cos^3 x - 4 \sin^3 x \cos x$

$\sin 5x = \sin x [5 \cos^4 x - 10 \sin^2 x \cos^2 x - \sin^4 x]$

12.  $\pi/4$ , because  $2 \sin \theta \cos \theta = \sin 2\theta$  has a maximum value 1 for  $\theta = \pi/4$ .

13. a)  $\sin(m + n)x = \sin mx \cos nx + \sin nx \cos mx$ ,

$\sin(m - n)x = \sin mx \cos nx - \sin nx \cos mx$ .

Adding gives  $\sin mx \cos nx = \frac{1}{2}[\sin(m + n)x + \sin(m - n)x]$ .

*Chapter 11, §6*

1.  $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$

3.  $\begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$

5.  $\begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$

7.  $\begin{pmatrix} 0 & +1 \\ -1 & 0 \end{pmatrix}$

11. 1.  $x' = -3, y' = -1$

3.  $x' = (3 - \sqrt{3})/2, y' = (3\sqrt{3} + 1)/2$

5.  $x' = -2\sqrt{2}/2, y' = \sqrt{2}$

12. 1.  $x' = -5, y' = 2$

3.  $x' = (5 + 2\sqrt{3})/2, y' = (5\sqrt{3} - 2)/2$

5.  $x' = -3\sqrt{2}/2, y' = 7\sqrt{2}/2$

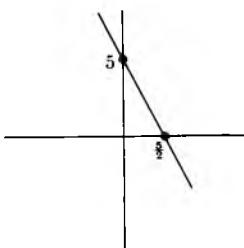
21.  $D_r = \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}$  and  $F_{a,b} = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$

22.  $G_\varphi = \begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}$   $G_\psi = \begin{pmatrix} \cos \psi & -\sin \psi \\ \sin \psi & \cos \psi \end{pmatrix}$

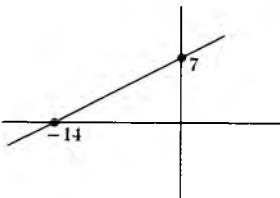
$G_\varphi G_\psi = \begin{pmatrix} \cos(\varphi + \psi) & -\sin(\varphi + \psi) \\ \sin(\varphi + \psi) & \cos(\varphi + \psi) \end{pmatrix}$

*Chapter 12, §1*

1. a)



2. a)



3. a)  $(\frac{7}{8}, \frac{11}{8})$

4. a)  $y = -\frac{8}{3}x - \frac{5}{3}$  b)  $y = -\frac{3}{2}x - 4$

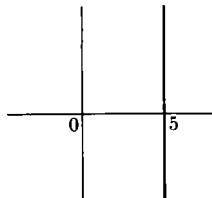
5. a)  $x = \sqrt{2}$  b)  $y = \frac{9}{3 + \sqrt{3}}x + \frac{27}{3 + \sqrt{3}} + 5$

8. a)  $(\frac{3}{5}, \frac{2}{5})$

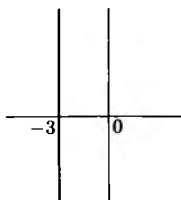
9.  $y = 4x - 3$

11.  $y = -\frac{1}{2}x + 3 + \frac{\sqrt{2}}{2}$

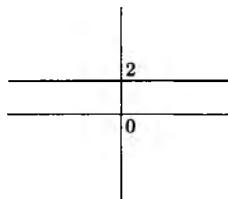
13.



15.



17.

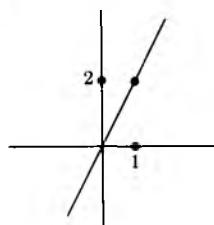


19.  $-\frac{1}{4}$     20.  $-8$     21.  $\frac{-2}{\sqrt{2} - 2}$

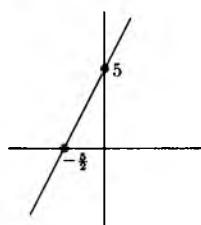
23.  $y = \frac{2}{\sqrt{2} - \pi}x - \frac{2\pi}{\sqrt{2} - \pi} + 1$

26.  $y = (3 + \sqrt{2})x + 3 + 2\sqrt{2}$

27. a)



c)



**28.** If they are parallel, then  $a = c$ . Any point  $(x, y)$  in common would be such that  $ax + b = cx + d$ , so  $b = d$  which is impossible. If they are not parallel, then  $a \neq c$ . Let  $(x, y)$  be a point in common. Then  $ax + b = cx + d$ , whence  $x(a - c) = d - b$ , so  $x = (d - b)/(a - c)$ . Then  $y = ax + b = (ad - ab)/(a - c) + b = (ad - bc)/(a - c)$ . Substituting in the original equations shows that these values of  $x, y$  satisfy the equations.

**29.** a)  $(-4, -7)$  c)  $(-\frac{1}{3}, \frac{7}{3})$       **30.**  $a_2/a_1$  if  $a_1 \neq 0$ . Independent of  $P$ .

**31.** a)  $(\frac{4}{5}, \frac{3}{5})$  and  $(0, -1)$  c)  $(1, 1)$  and  $(\frac{1}{5}, -\frac{7}{5})$

### Chapter 12, §2

**5.**  $x' = x - 2$ ,  $y' = y + 1$ , and  $x'^2 + y'^2 = 25$

**6.**  $x' = x$ ,  $y' = y - 1$ , and  $x'^2 + y'^2 = 9$

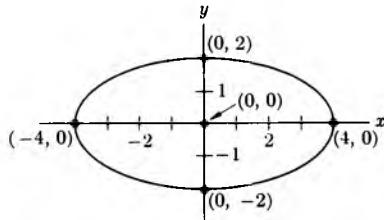
**8.**  $y' = y + \frac{25}{8}$ ,  $x' = x + \frac{1}{4}$ , and  $y' = 2x'^2$

**10.**  $y' = y + 4$ ,  $x' = x - 1$ , and  $y' = x'^2$

### Chapter 12, §3

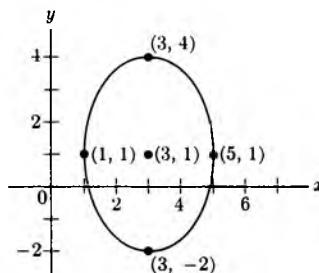
**1.** Center  $(0, 0)$

Extremities  $\begin{cases} (0, 2) & \text{and } (0, -2) \\ (-4, 0) & \text{and } (4, 0) \end{cases}$



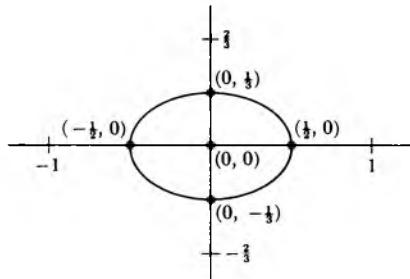
**2.** Center  $(3, 1)$

Extremities  $\begin{cases} (1, 1) & \text{and } (5, 1) \\ (3, 4) & \text{and } (3, -2) \end{cases}$



**7.** Center  $(0, 0)$

$$\text{Extremities } \begin{cases} (0, \frac{1}{3}) & \text{and } (0, -\frac{1}{3}) \\ (-\frac{1}{2}, 0) & \text{and } (\frac{1}{2}, 0) \end{cases}$$



**8.** Center  $(0, 0)$

$$\text{Extremities } \begin{cases} (0, \frac{1}{4}) & \text{and } (0, -\frac{1}{4}) \\ (-\frac{1}{5}, 0) & \text{and } (\frac{1}{5}, 0) \end{cases}$$

**11.** Center  $(1, -2)$

$$\text{Extremities } \begin{cases} (\frac{1}{2}, -2) & \text{and } (\frac{3}{2}, -2) \\ (1, -\frac{7}{4}) & \text{and } (1, -\frac{9}{4}) \end{cases}$$

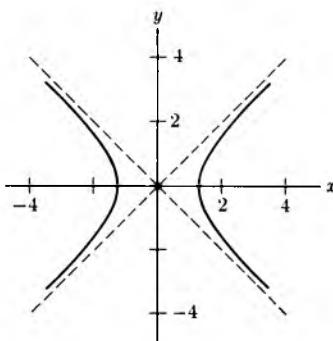
**12.** Center  $(-3, -1)$

$$\text{Extremities } \begin{cases} (-3, -\frac{2}{3}) & \text{and } (-3, -\frac{4}{3}) \\ (-\frac{13}{4}, -1) & \text{and } (-\frac{11}{4}, -1) \end{cases}$$

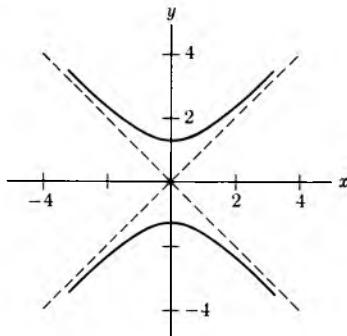
*Chapter 12, §5*

**2.**  $u^2 - v^2 = 2$

**3. a)**  $u^2 - v^2 = 2$



b)  $v^2 - u^2 = 2$



5. a)  $u^2 - v^2 = 2$  b)  $u^2 - v^2 = 4$

6. a)  $v^2 - u^2 = 2$  b)  $v^2 - u^2 = 4$

8. For any point  $(x, y)$  we have  $F_r((x, y)) = (rx, ry)$  and

$$G_\theta(F_r((x, y))) = (rx \cos \theta - ry \sin \theta, rx \sin \theta + ry \cos \theta).$$

If you compute the coordinates for  $F_r(G_\theta((x, y)))$  in a similar way, you will find the same expression.

### Chapter 13, §1

1.  $f\left(\frac{3}{4}\right) = \frac{4}{3}$ ,  $f\left(-\frac{2}{3}\right) = -\frac{3}{2}$     2.  $x \neq \pm\sqrt{2}$ ,  $f(5) = \frac{1}{23}$

3. All real numbers,  $f(27) = 3$     4. a)  $f(1) = 1$

5. a)  $f\left(\frac{1}{2}\right) = 1$  b)  $f(2) = 4$     7.  $x \geq 0$ ,  $f(16) = 2$

8. a) Odd b) Even

9. Let  $E(x) = (f(x) + f(-x))/2$  and  $O(x) = (f(x) - f(-x))/2$ . Then  $E(x) + O(x) = f(x)$ . Furthermore,  $E(-x) = (f(-x) + f(x))/2 = E(x)$ , so  $E$  is even;  $O(-x) = (f(-x) - f(x))/2 = -O(x)$ , so  $O$  is odd.

10. a) Odd b) Even c) Odd

12. a) Even. *Proof:* Let  $f, g$  be odd functions. Then

$$(fg)(-x) = f(-x)g(-x) = -f(x)(-g(x)) = f(x)g(x) = (fg)(x).$$

### Chapter 13, §2

1. a) 2 b) 5 d) 7 e) 6

h)  $f(x)g(x) = (a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0)(b_m x^m + b_{m-1} x^{m-1} + \cdots + b_0)$  with  $a_n \neq 0$  and  $b_m \neq 0$ . Thus  $\deg f = n$  and  $\deg g = m$ . Then  $f(x)g(x) = a_n b_m x^{m+n} + \text{terms of lower degree}$ . Hence  $\deg (fg) = m + n = \deg f + \deg g$ , because the leading coefficient  $a_n b_m$  is  $\neq 0$ .

2. a)  $\left(x - \frac{-3 - \sqrt{17}}{2}\right)\left(x - \frac{-3 + \sqrt{17}}{2}\right)$  e)  $3(x + 1)(x - \frac{1}{3})$

g)  $3\left(x + \frac{3 + \sqrt{57}}{6}\right)\left(x + \frac{3 - \sqrt{57}}{6}\right)$

3. We must have  $\deg g + \deg h = \deg f = 3$ . If  $\deg g = 1$ , then  $g$  has a root (why?), say  $c$ , that is  $g(c) = 0$ . But then  $f(c) = g(c)h(c) = 0$ , so that  $c$  is a root of  $f$ . If  $g$  has degree 2, then  $h$  has degree 1, and we can argue in the same way with  $h$ . These are the only possibilities, because  $g$  cannot have degree 3, otherwise  $\deg h = 0 < 1$ , contrary to assumption.

5.  $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ ;

$f(c) = c^n + a_{n-1}c^{n-1} + \cdots + a_0 = 0$ .

Thus  $c[c^{n-1} + a_{n-1}c^{n-2} + \cdots + a_1] = -a_0$ , and  $c$  divides  $a_0$ .

6. a) 1 b) 1 and  $-1$  c) 1

d) 1 if  $n$  is odd. 1 and  $-1$  if  $n$  is even. f) No root

7. a)  $q(x) = 4x^2 + 8x + 15$  and  $r(x) = 32$

b)  $q(x) = 4x$  and  $r(x) = 3x + 2$

c)  $q(x) = 4x$  and  $r(x) = -5x + 2$

8. a)  $q(x) = 6x - 1$  and  $r(x) = x^2 + 4x + 4$

b)  $q(x) = 6x^2 - x + 31$  and  $r(x) = 160 - 7x$

c)  $q(x) = 6x^3 - 13x^2 + 25x - 52$  and  $r(x) = 109$

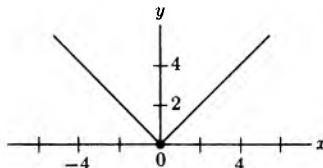
d)  $q(x) = 2x^3 - x^2 + \frac{2}{3}x - \frac{8}{9}$  and  $r(x) = \frac{53}{9}$

9. Rational functions.

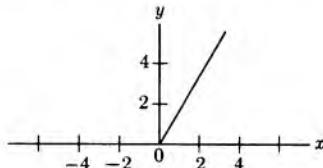
a)  $\frac{3x^2 - 2x + 23}{x^2 - 2x - 15}$  b)  $\frac{11x^2 - 3x - 10}{6x^2 + 10x + 4}$  c)  $\frac{4x^3 + 16x^2 + x + 9}{x^2 + 16x + 5}$

Chapter 13, §3

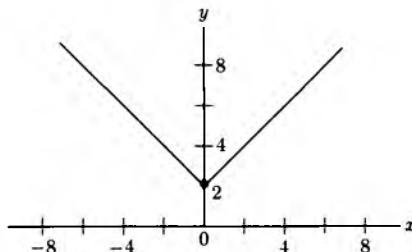
9.



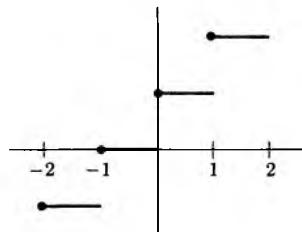
10.



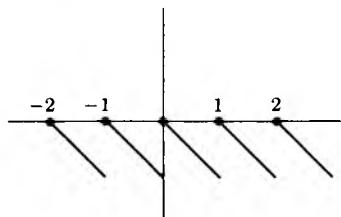
12.



25.



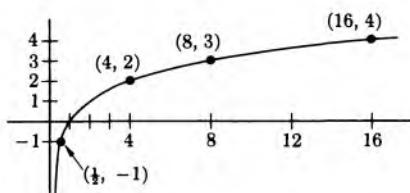
28.

*Chapter 13, §4*

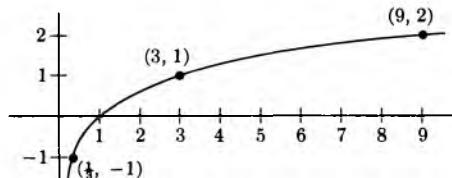
3.  $b^x$  is steeper.
4. a)  $5/a^3$
5.  $P(t) = 10^5(3^{t/50})$ , and  $t$  is in years.  
a)  $9 \cdot 10^5$  b)  $27 \cdot 10^5$  c)  $81 \cdot 10^5$

*Chapter 13, §5*

1. a)



b)



2. a) 6 b) -3 c) 2

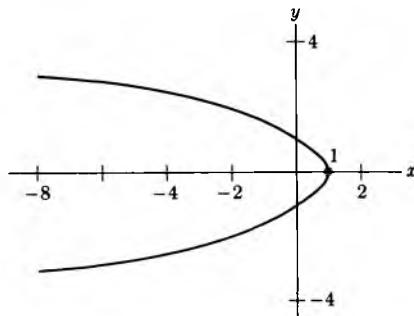
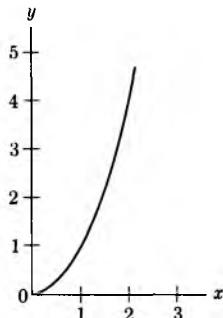
3.  $e^{(\log a)x} = (e^{\log a})^x = a^x$ . Hence  $\log a^x = (\log a)x = x \log a$ .

a) .36 b) 1.5 c) 0.1 d) 0.4 e) 1.8

4.  $c = 5e^{-4}$  5.  $t = \frac{1}{5} \log 2$

6.  $12 \frac{\log 10}{\log 2}$  min 8.  $\frac{3 \log 2}{\log 10 - \log 9}$  min

10. 20.7 11. 11.05 hr

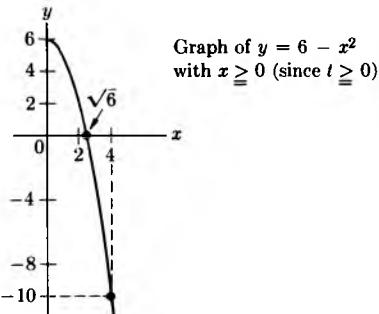
*Chapter 14, §1*2. a) If  $t = y$ , then  $x = 1 - y^2$  (parabola).b) If  $x = t^2$ , then  $y = x^2$  with  $y \geq 0$  and  $x \geq 0$ , half a parabola.

3. a)  $\frac{5}{4}\sqrt{2}$  sec d)  $\sqrt{\frac{5}{4}}$

4. a) 1

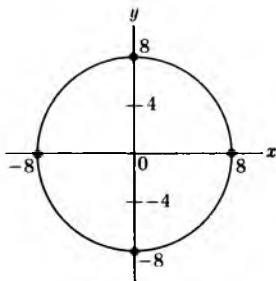
b) The points  $(x(t), y(t))$  satisfy the equation  $y = 6 - x^2$ , which is a parabola.

c)  $\frac{\sqrt{6}}{2}$

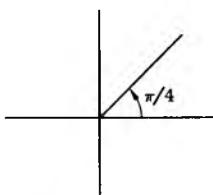


6. Image of a line  $y = c$  is a circle centered at the origin of radius  $2^c$ . Image of a line  $x = c$  is a ray of slope  $\tan c$ . The ray is open, i.e. does not contain the origin, because for all numbers  $y$ , the number  $2^y$  is  $> 0$ !!

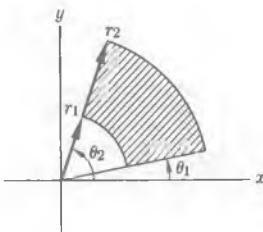
a)



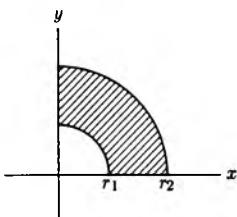
c)



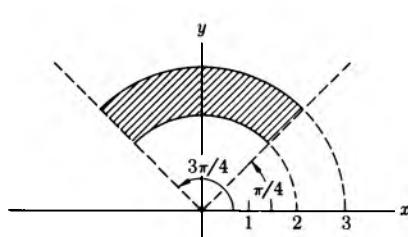
7. a)



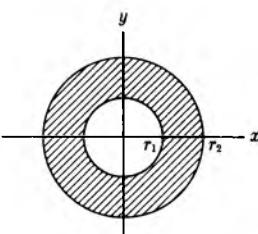
c)



d)



e)



## Chapter 14, §2

1.  $h \circ f = h \circ g$ . Multiplying on the left by  $h^{-1}$  gives  $h^{-1} \circ h \circ f = h^{-1} \circ h \circ g$ , so  $I_T \circ f = I_T \circ g$ , i.e.  $f = g$ .
2. a)  $f(x) = f(y)$ ;  $f^{-1}(f(x)) = f^{-1}(f(y))$ , that is  $x = y$   
b) Let  $x = f^{-1}(z)$ . Then  $f(x) = f(f^{-1}(z)) = z$ .
3. Let  $G: (x, y) \mapsto \left(\frac{x}{a}, \frac{y}{b}\right)$ . Then  $(G \circ F)(x, y) = G(ax, by) = (x, y)$   
 $(F \circ G)(x, y) = F\left(\frac{x}{a}, \frac{y}{b}\right) = (x, y)$ .
4.  $\begin{cases} u = 2x - y \\ v = y + x \end{cases}$  i.e.  $x = \frac{u+v}{3}$  and  $y = \frac{2v-u}{3}$   
 Let  $g: (u, v) \mapsto \left(\frac{u+v}{3}, \frac{2v-u}{3}\right)$ . Then  $(g \circ f)(x, y) = (x, y)$   
 $(f \circ g)(x, y) = f\left(\frac{x+y}{3}, \frac{2y-x}{3}\right) = (x, y)$ .

6. a)  $f^5 = f^3 \circ f^2 = I = I \circ f^2$ , thus  $f^2 = I$

$f^3 = f \circ f^2 = I = f \circ I$ , thus  $f = I$

7. a)  $f^6 \circ g^{-3}$  c)  $f \circ g^0 = f$

8. a)  $g^{-1} \circ f^{-1}$  is the inverse;  $(g^{-1} \circ f^{-1}) \circ (f \circ g) = g^{-1} \circ g = I$

$(f \circ g) \circ (g^{-1} \circ f^{-1}) = f \circ f^{-1} = I$

b)  $f_m^{-1} \circ f_{m-1}^{-1} \circ \dots \circ f_1^{-1}$  is the inverse;  $(f_1 \circ f_2 \dots \circ f_m) \underbrace{(f_m^{-1} \circ \dots \circ f_1^{-1})}_{I} = I$

$(f_m^{-1} \circ \dots \circ f_1^{-1}) \underbrace{(f_1 \circ f_2 \dots \circ f_m)}_I = I$

*Chapter 14, §3*

1. a)  $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{bmatrix}$ ; sign +1

c) It is already a transposition; sign -1

e)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{bmatrix}$ ; sign +1

h)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{bmatrix}$ ; sign -1

2. a)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 5 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 3 & 1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 3 & 4 & 5 \end{bmatrix}$ ; sign -1

c)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 3 & 1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 5 & 3 & 4 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 4 & 5 \end{bmatrix}$ ; sign +1

3. a)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 5 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 6 & 5 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 4 & 5 & 6 \end{bmatrix}$ ; sign -1

c)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 6 & 3 & 4 & 5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 2 & 3 & 1 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 5 & 3 & 4 & 2 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 4 & 5 & 6 \end{bmatrix}$ ; sign +1

4. a)  $\begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}$  c)  $\begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$  e)  $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix}$

5. a)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 1 & 5 & 2 & 3 \end{bmatrix}$  c)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 1 & 5 & 2 \end{bmatrix}$  e)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 5 & 3 \end{bmatrix}$

6. a)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 5 & 6 & 4 \end{bmatrix}$  c)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 6 & 1 & 3 & 2 & 5 \end{bmatrix}$  e)  $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 2 & 5 & 1 & 6 \end{bmatrix}$

7. a) [1 2 3] c) [1 3] [2] e) [1 2] [3 4] g) [1 4 3] [2]

8. a) [1 2 4] [3 5] c) [1 3 2 5 4] d) [1 4] [2 5] [3]

9. a) [1 3] [2] [4 6 5] b) [1 2 6 4 5 3] e) [1 3 4] [2 5 6]

10. Let  $S_e$  be the set of all *distinct* even permutations.

Let  $S_e = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$  and let  $S' = \{\tau\sigma_1, \tau\sigma_2, \dots, \tau\sigma_m\}$ , with  $\tau$  being a transposition. Each  $\sigma_i$  is even. Hence  $\tau\sigma_i$  is odd. We now prove that all the elements of the set  $S'$  are distinct. If  $\tau\sigma_k = \tau\sigma_j$ , then multiplying on the left by  $\tau$  shows that  $\sigma_k = \sigma_j$ , so that  $k = j$  and  $\sigma_k = \sigma_j$ . Finally, we must prove that  $S'$  contains all odd permutations. Let  $\sigma$  be an odd permutation. Then  $\tau\sigma$  is even and  $\tau\sigma = \sigma_k$  for some  $k$ . Hence  $\sigma = \tau^{-1}\sigma_k = \tau\sigma_k$ , and  $\sigma$  is in  $S'$ .

11. Property is true for  $n = 1$ . Assume it for  $n$ . Suppose that  $\tau_i\sigma = \tau_j\sigma'$  for some indices  $i, j$  and some  $\sigma, \sigma'$  in  $S_n$ . Since  $\tau_i\sigma(n+1) = i$  and  $\tau_j\sigma'(n+1) = j$ , it follows that  $\tau_i = \tau_j$ . Multiplying on the left by  $\tau_i (= \tau_j)$ , we conclude that  $\sigma = \sigma'$ . If  $\sigma' = \tau_i\sigma$ , then again looking at the effect on the number  $n+1$ , we conclude that this cannot be so. Hence the permutations  $\sigma, \tau_1\sigma, \dots, \tau_n\sigma$  with  $\sigma$  in  $S_n$  are all distinct.

Furthermore, if  $\gamma$  is a permutation of  $J_{n+1}$ , then either  $\gamma$  leaves  $n+1$  fixed, in which case  $\gamma$  is already in  $S_n$ , or  $\gamma(n+1) = i$  for some  $i$  with  $1 \leq i \leq n$ . In this case  $\tau_i\gamma$  leaves  $n+1$  fixed, so  $\tau_i\gamma = \sigma$  is an element of  $S_n$ , and  $\gamma = \tau_i\sigma$ . Thus we have found all the permutations of  $J_{n+1}$ . We assume by induction that there are  $n!$  permutations in  $S_n$ . To each  $\sigma$  in  $S_n$  we have associated the  $n+1$  permutations  $\sigma, \tau_1\sigma, \dots, \tau_n\sigma$  in  $S_{n+1}$ . Hence the total number of permutations in  $S_{n+1}$  is

$$(n+1)n! = (n+1)!.$$

### Chapter 15, §1

1. a)  $(-1 - 3i)/10$  c)  $3 + i$  e)  $6\pi + i(7 + \pi^2)$
2. a)  $(1 - i)/2$  c)  $(3 + 4i)/5$  e)  $1 - i$       3. 1
4.  $z = x + iy$  and  $w = u + iv$ ; thus  $\bar{z} = x - iy$  and  $\bar{w} = u - iv$   
 $zw = (xu - yv) + i(yu + xv)$  and  $\bar{z}\bar{w}(xu - yv) - i(yu + xv)$   
 $\bar{z}\bar{w} = (x - iy)(u - iv) = (xu - yv) - i(yu + xv)$   
 $\bar{z} + \bar{w} = (x - iy) + (u - iv) = (x + u) - i(y + v)$   
 $\bar{z} + \bar{w} = [(x + u) + i(y + v)] = (x + u) - i(y + v)$   
 $\bar{z} = (x - iy) = x + iy = z$
5.  $\operatorname{Im}(z) = y$ ; we always have  $y \leq |y|$ , thus  $\operatorname{Im}(z) \leq |\operatorname{Im}(z)|$ . Next,  $y^2 \leq x^2 + y^2$  because  $x^2 \geq 0$ . Hence  $|y| = \sqrt{y^2} \leq \sqrt{x^2 + y^2} = |z|$ . These two inequalities together constitute the desired inequalities, namely  $\operatorname{Im}(z) \leq |\operatorname{Im}(z)| \leq |z|$ .

### Chapter 15, §2

1. a)  $\sqrt{2}e^{i\pi/4}$  c)  $3e^{i\pi}$  e)  $2e^{-i\pi/3}$
2. a)  $-1$  c)  $\frac{3\sqrt{2} + i3\sqrt{2}}{2}$  e)  $\frac{1 + i\sqrt{3}}{2}$

3. Suppose  $\alpha, \beta$  are complex numbers such that  $\alpha^2 = \beta^2 = z$ . Then  $\alpha^2 - \beta^2 = 0$  so that  $(\alpha + \beta)(\alpha - \beta) = 0$ . Hence  $\alpha = \beta$  or  $\alpha = -\beta$ . So there are at most two such complex numbers. But if  $z = re^{i\theta}$ , then  $\alpha = r^{1/2}e^{i\theta/2}$  and  $-\alpha$  have squares equal to  $z$ .
4. Let  $z = re^{i\theta}$ . Let  $w_k = r^{1/n}e^{i\theta/n + 2k\pi i/n}$  where  $k = 1, \dots, n$ . Then  $w_1, \dots, w_n$  have  $n$ -th powers equal to  $z$ .
5.  $w_n = e^{2k\pi i/n}$  with  $k$  from 0 to  $(n - 1)$
6.  $e^{i\theta} = \cos \theta + i \sin \theta$   
 $e^{-i\theta} = \cos \theta - i \sin \theta$  } . Adding the two equations yields  
 $\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2}$ . Subtracting them yields  $\sin \theta = \frac{1}{2i}[e^{i\theta} - e^{-i\theta}]$ .

### Chapter 16, §1

1. Let  $A(n) = \sum_{k=0}^{n-1} (2k + 1)$ . We want to show  $A(n) = n^2$ . For  $A(1)$  we have  $1 = 1$ . By induction  $A(n+1) = A(n) + 2n + 1 = n^2 + 2n + 1 = (n + 1)^2$ .
2. b) Let  $A(n) = \sum_{k=1}^n k^3$ . We want to prove that  $A(n) = \left[ \frac{n(n+1)}{2} \right]^2$ .  
For  $A(1)$  we have  $1 = \left[ \frac{1(2)}{2} \right]^2 = 1$ . By induction,  $A(n+1) = A(n) + (n+1)^3 = \frac{1}{4}(n+1)^2[n^2 + 4n + 4]$ . Thus  $A(n+1) = \frac{1}{4}(n+1)^2(n+2)^2 = \left[ \frac{(n+1)(n+2)}{2} \right]^2$ .
5. Let  $A(n)$  be the product  $\prod_{k=0}^n (1 + x^{2^k})$ . We want to show  

$$A(n) = \frac{1 - x^{2^{n+1}}}{1 - x}$$
. We have  $A(0) = 1 + x = \frac{1 - x^2}{1 - x}$ .  
By induction  $A(n+1) = A(n)(1 + x^{2^{n+1}})$   
 $= \frac{(1 - x^{2^{n+1}})(1 + x^{2^{n+1}})}{1 - x}$   
 $= \frac{1 - x^{2^{n+1}+1}}{1 - x} = \frac{1 - x^{2^{n+2}}}{1 - x}$ .

6. Let  $A(n)$  be the property:  $f(x^n) = nf(x)$ .  $A(1)$  is true since  $f(x) = f(x)$ . By induction let us prove  $A(n+1)$ . We have:

$$f(x^{n+1}) = f(x^n x) = f(x^n)f(x) = nf(x) + f(x) = (n + 1)f(x).$$

8. a)  $\frac{n(2n+1)(4n+1)}{3}$

c)  $4n^2$  e)  $\frac{[m(n-1)]^2}{4}$

9. a)  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$   
 $= \frac{n!}{(n-k)!k!} = \binom{n}{n-k}$

b)  $\binom{n}{k-1} + \binom{n}{k} = \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!}$   
 $= \frac{n!}{(k-1)!(n-k)!} \left[ \frac{1}{n-k+1} + \frac{1}{k} \right]$   
 $= \frac{n!}{(k-1)!(n-k)!} \cdot \frac{n+1}{k(n-k+1)}$   
 $= \frac{(n+1)!}{k!(n-k+1)!}$   
 $= \binom{n+1}{k}$

c) Let  $A(n)$  be the property:  $(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$ .  $A(1)$  is true, because  $(x+y) = \binom{1}{0}y + \binom{1}{1}x = x+y$ .  $A(n+1)$  is proved using induction:

$$\begin{aligned}(x+y)^{n+1} &= (x+y)^n(x+y) = \left[ \sum_{k=0}^n \binom{n}{k} x^k y^{n-k} \right] (x+y) \\&= \sum_{k=0}^n \binom{n}{k} x^{k+1} y^{n-k} + \sum_{k=0}^n \binom{n}{k} x^k y^{n-k+1} \\&= \sum_{j=1}^{n+1} \binom{n}{j-1} x^j y^{n+1-j} + \sum_{j=0}^n \binom{n}{j} x^j y^{n+1-j} \\&= \sum_{j=0}^{n+1} \binom{n+1}{j} x^j y^{n+1-j}\end{aligned}$$

10. a) 3 c) 1 e) 4

11. a)  $x^3 + 3x^2y + 3xy^2 + y^3$  b)  $x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$   
c)  $x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$

12. The inductive argument does not apply going from  $n = 1$  to  $n = 2$ . There is no "middle" ball!!

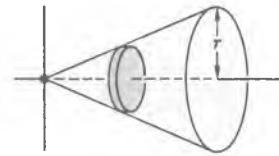
13.  $A(1)$  is obviously true. By induction, there are  $m^n$  ways of mapping  $n$  elements into  $F$ , and for each of these ways there are  $m$  ways of mapping the remaining element. Thus there are  $m^n \cdot m$  ways or  $m^{n+1}$  of mapping  $E$  into  $F$ .

## Chapter 16, §2

1. Divide the height into  $n$  segments of length  $\frac{h}{n}$ .

The radius of a small cylinder is  $\frac{r(kh)/n}{h} = \frac{rk}{n}$ .

The volume of each cylinder is  $\pi \left(\frac{rk}{n}\right)^2 \cdot \frac{h}{n}$ .



Using Exercise 2(a) of §1, the sum of these volumes is

$$\sum_{k=1}^n \pi \left(\frac{rk}{n}\right)^2 \cdot \frac{h}{n} = \frac{\pi hr^2}{n^3} \sum_{k=1}^n k^2 = \frac{\pi hr^2}{6} \left[ \left(1 + \frac{1}{n}\right) \left(2 + \frac{1}{n}\right) \right].$$

As  $n$  becomes arbitrarily large the sum approaches  $\frac{\pi hr^2}{3}$ .

2. a)  $24\pi$  c)  $3\pi c^3$  3. a)  $\pi/2$  c)  $9\pi/2$  d)  $\pi h^2/2$

4. a)  $2\pi/3$  d)  $2\pi r^3/3$  5. It is half a ball. 6. a)  $\frac{1}{3}$  c) 9

7. a)  $\frac{1}{4}$  c)  $81/4$

8.  $S$  is mapped into  $T$  with  $F_{1,3}$ :  $(x, y) \mapsto (x, 3y)$ . The area is 1.

9. a)  $c/3$

## Chapter 16, §3

1. a)  $3/2$  c)  $5/4$  e) 4

2.  $(1 - c + c^2 - c^3 + \cdots + (-1)^n c^n)(1 + c) = 1 + (-1)^n c^{n+1}$ . Hence

$$\sum_{k=0}^n (-1)^k c^k = \frac{1}{1+c} + \frac{(-1)^n c^{n+1}}{1+c}.$$

But  $c^{n+1}$  approaches zero as  $n$  becomes large. Hence

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n (-1)^k c^k = \frac{1}{1+c}.$$

a)  $3/4$  c)  $5/6$  e)  $4/7$

4.  $\sum_{k=0}^n r e^{ik\theta} = \frac{1 - r^{n+1} e^{i\theta(n+1)}}{1 - r e^{i\theta}}$ ;  $\sum_{k=0}^{\infty} r e^{ik\theta} = \frac{1}{1 - r e^{i\theta}}$

5. 0

7.  $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots \geq 2 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + 8 \cdot \frac{1}{16} + \cdots$   
 $\geq 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots$

By additional groups of terms we add at least  $\frac{1}{2}$  each time. The sum can be made larger than any given value  $A$  by adding  $2A$  groups of terms.

$$8. S = 1 + \underbrace{\frac{1}{2^2} + \frac{1}{3^2}}_{\leq \frac{1}{2}} + \underbrace{\frac{1}{4^2} + \frac{1}{5^2}}_{\leq \frac{1}{4}} + \underbrace{\frac{1}{6^2} + \frac{1}{7^2}}_{\leq \frac{1}{4}} + \underbrace{\frac{1}{8^2} + \cdots + \frac{1}{15^2}}_{\leq \frac{1}{8}} + \cdots$$

Hence  $S \leq \sum_{k=0}^{\infty} \frac{1}{2^k} \leq 2.$

*Chapter 17, §1*

1. a) 26 c) -5 2. 1 3. b) -1 d) 0

*Chapter 17, §2*

$$1. D(B, C' + C'') = \begin{vmatrix} b_1 & c'_1 + c''_1 \\ b_2 & c'_2 + c''_2 \end{vmatrix} = b_1(c'_2 + c''_2) - b_2(c'_1 + c''_1) \\ = (b_1c'_2 - b_2c'_1) + (b_1c''_2 - b_2c''_1) \\ = D(B, C') + D(B, C'')$$

$$3. D(B, C + xB) = D(B, C) + D(B, xB) \text{ using D1} \\ = D(B, C) + xD(B, B) \text{ using D2} \\ = D(B, C) + 0 \text{ using D4}$$

$$4. D(C, A) = D(xA + yB, A) = D(xA, A) + D(yB, A) \\ = xD(A, A) + yD(B, A) = yD(B, A)$$

$$\text{Hence } y = \frac{D(C, A)}{D(B, A)}.$$

$$5. 1. x = \frac{5}{3} \text{ and } y = \frac{1}{3} \quad 3. x = -2 \text{ and } y = 1 \quad 5. x = -\frac{1}{2} \text{ and } y = -\frac{3}{2}$$

*Chapter 17, §4*

$$1. \text{ a) } D(A) = a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} - a_{32} \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} + a_{33} \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \\ = a_{31}a_{12}a_{23} - a_{31}a_{13}a_{22} - a_{32}a_{11}a_{23} + a_{32}a_{21}a_{13} \\ + a_{33}a_{11}a_{22} - a_{33}a_{12}a_{21}$$

$$\text{c) } D(A) = a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} - a_{23} \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} + a_{33} \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \\ = a_{13}a_{21}a_{32} - a_{13}a_{31}a_{22} - a_{23}a_{11}a_{32} + a_{23}a_{31}a_{12} \\ + a_{33}a_{11}a_{22} - a_{33}a_{21}a_{12}$$

$$2. \text{ a) } -20 \quad \text{c) } 4 \quad \text{e) } -76 \quad 3. \text{ a) } 140 \quad 4. abc \quad 5. \text{ a) } -24 \quad \text{c) } 90$$

$$6. \text{ a) } a_{11}a_{22}a_{33} \quad \text{b) } a_{11}a_{22}a_{33}$$

*Chapter 17, §5*

$$1. D(A^1, B + C, A^3) = -(b_1 + c_1) \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + (b_2 + c_2) \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} \\ - (b_3 + c_3) \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix}$$

$$D(A^1, B, A^3) = -b_1 \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + b_2 \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} - b_3 \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix}$$

$$D(A^1, C, A^3) = -c_1 \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + c_2 \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + c_3 \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix}$$

You could also get a proof by expanding according to any row or column.

4. a)  $D(A^1 + xA^3, A^2, A^3) = D(A^1, A^2, A^3) + D(xA^3, A^2, A^3)$   
 $= D(A^1, A^2, A^3) + xD(A^3, A^2, A^3)$   
 $= D(A^1, A^2, A^3) + 0$
5.  $D(A^1 + A^2, A^1 + A^2, A^3) = D(A^1, A^1, A^3) + D(A^1, A^2, A^3)$   
 $+ D(A^2, A^1, A^3) + D(A^2, A^3, A^3)$   
 $= 0 + D(A^1, A^2, A^3)$   
 $+ D(A^2, A^1, A^3) + 0$
6.  $D(A^1, A^2, A^3) = -D(A^1, A^3, A^2) = D(A^3, A^1, A^2)$   
 $= -D(A^3, A^2, A^1)$

Since  $D({}^t A) = D(A)$ , the same holds for rows.

7. If you add a multiple of one row to another, you do not change the value of the determinant. If you interchange two adjacent rows, then the determinant changes by a sign.

8. a) 0 b) 24 c) -12 d) 0 9.  $D(cA) = c^3 D(A)$
10.  $\begin{vmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{vmatrix} = \begin{vmatrix} 1 & x_1 & x_1^2 \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 \\ 1 & x_3 & x_3^2 \end{vmatrix} = \begin{vmatrix} 1 & x_1 & x_1^2 \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 \\ 0 & x_3 - x_1 & x_3^2 - x_1^2 \end{vmatrix}$   
 $= (x_2 - x_1)(x_3 - x_1)(x_3 + x_1 - x_2 - x_1)$   
 $= (x_2 - x_1)(x_3 - x_1)(x_3 - x_2)$
11.  $D(B^1 + B^2 + B^3, A^2, A^3) = D(B^1, A^2, A^3) + D(B^2 + B^3, A^2, A^3)$   
 $= D(B^1, A^2, A^3) + D(B^2, A^2, A^3)$   
 $+ D(B^3, A^2, A^3)$

$$D\left(\sum_{j=1}^n B_j, A^2, A^3\right) = \sum_{j=1}^n D(B_j, A^2, A^3)$$

### Chapter 17, §6

1.  $D(x_1 A^1 + x_2 A^2 + x_3 A^3, A^2, A^3)$   
 $= x_1 D(A^1, A^2, A^3) + x_2 D(A^2, A^2, A^3) + x_3 D(A^3, A^2, A^3)$   
 $= x_1 D(A^1, A^2, A^3)$
2. a)  $D(A^1, B, A^3) = x_1 D(A^1, A^1, A^3) + x_2 D(A^1, A^2, A^3)$   
 $+ x_3 D(A^1, A^3, A^3)$   
 $= 0 + x_2 D(A^1, A^2, A^3)$

Hence  $x_2 = \frac{D(A^1, B, A^3)}{D(A^1, A^2, A^3)}$ .

$$\begin{aligned}3. \quad & D(x_1 A^1 + x_2 A^2 + x_3 A^3, A^2, A^3) \\&= D(0, A^2, A^3) = 0 \\&= x_1 D(A^1, A^2, A^3) + x_2 D(A^2, A^2, A^3) + x_3 D(A^3, A^2, A^3) \\&= x_1 D(A^1, A^2, A^3) + 0 + 0\end{aligned}$$

Since at least one of the number  $x_1, x_2, x_3$  is not zero, we may assume that  $x_1 \neq 0$ . Then  $x_1 D(A^1, A^2, A^3) = 0$  implies  $D(A^1, A^2, A^3) = 0$ .





## THE AUTHOR

Serge Lang received the Ph.D. degree from Princeton University and is currently Professor of Mathematics at Columbia University. Professor Lang is the author of *A First Course in Calculus, Second Edition*; *A Second Course in Calculus, Second Edition*; *Introduction to Linear Algebra*; *Linear Algebra, Second Edition*; *Algebraic Structures*; *Analysis I*; *Algebra*; *Algebraic Number Theory*; *Analysis II*; *Introduction to Diophantine Approximations*; *Introduction to Transcendental Numbers*—and co-editor of *Collected Papers of Emil Artin*—all published by Addison-Wesley. His other books include *Abelian Varieties*, *Diophantine Geometry*, *Introduction to Algebraic Geometry*, and *Introduction to Differentiable Manifolds*.

## OTHER BOOKS OF INTEREST

### A FIRST COURSE IN CALCULUS, Second Edition

By Serge Lang

316 pp, 28 illus (1968)

This text covers the basic notions of derivative and integral and the basic techniques and applications which accompany them. This edition includes a discussion of complex numbers and greatly expanded exercise sets.

### A SECOND COURSE IN CALCULUS, Second Edition

By Serge Lang

305 pp, 116 illus (1968)

This book is a continuation of the author's *A First Course in Calculus, Second Edition* and deals principally with functions of several variables. The two books together are designed for a two-year course in analytic geometry, calculus, and linear algebra. They are also available as one volume: *A Complete Course in Calculus* (Addison-Wesley, 1968).

### INTRODUCTION TO LINEAR ALGEBRA

By Serge Lang

180 pp, 49 illus, paperbound (1970)

This text can be used for an independent short course in the elements of linear algebra or as a supplement to the calculus sequence. The author concentrates on the natural bases and has included excellent applications to geometry. This book will probably find its greatest use at the end of the freshman year after the student has completed the basics of one-variable calculus, although it can also serve as an introduction to some algebra previous to the calculus.

### LINEAR ALGEBRA, Second Edition

By Serge Lang

In press (1971)

For linear algebra courses at the undergraduate level, this book contains enough material for a one-year course. It covers vector spaces, matrices, linear maps, and determinants in greater detail than the first edition. Many new examples and exercises have been added to this edition, and the author has extensively revised the first half of the book. New material on unitary maps over the reals, the Jordan canonical form, and spectral theorem has also been introduced.

# **Book of Proof**

Third Edition

Richard Hammack

Richard Hammack (publisher)  
Department of Mathematics & Applied Mathematics  
P.O. Box 842014  
Virginia Commonwealth University  
Richmond, Virginia, 23284

## **Book of Proof**

Edition 3

© 2018 by Richard Hammack

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivative  
4.0 International License



Typeset in 11pt T<sub>E</sub>X Gyre Schola using PDFL<sup>A</sup>T<sub>E</sub>X

Cover by R. Hammack. The cover diagrams are based on a geometric construction that renders a correct perspective view of an object (here an octagonal column) from its floor plan. The method was invented by Piero della Francesca 1415–1492, a Renaissance painter and mathematician. Piero was a great expositor. His writings explained *why* his methods worked, not just how to apply them.

*To my students*

---

# Contents

---

<b>Preface</b>	<b>vii</b>
<b>Introduction</b>	<b>viii</b>

## *I Fundamentals*

<b>1. Sets</b>	<b>3</b>
1.1. Introduction to Sets	3
1.2. The Cartesian Product	8
1.3. Subsets	12
1.4. Power Sets	15
1.5. Union, Intersection, Difference	18
1.6. Complement	20
1.7. Venn Diagrams	22
1.8. Indexed Sets	25
1.9. Sets That Are Number Systems	30
1.10. Russell's Paradox	32
<b>2. Logic</b>	<b>34</b>
2.1. Statements	35
2.2. And, Or, Not	39
2.3. Conditional Statements	42
2.4. Biconditional Statements	46
2.5. Truth Tables for Statements	48
2.6. Logical Equivalence	50
2.7. Quantifiers	53
2.8. More on Conditional Statements	56
2.9. Translating English to Symbolic Logic	57
2.10. Negating Statements	59
2.11. Logical Inference	63
2.12. An Important Note	64
<b>3. Counting</b>	<b>65</b>
3.1. Lists	65
3.2. The Multiplication Principle	67
3.3. The Addition and Subtraction Principles	74
3.4. Factorials and Permutations	78
3.5. Counting Subsets	85
3.6. Pascal's Triangle and the Binomial Theorem	90
3.7. The Inclusion-Exclusion Principle	93
3.8. Counting Multisets	96
3.9. The Division and Pigeonhole Principles	104
3.10. Combinatorial Proof	108

## *II How to Prove Conditional Statements*

<b>4. Direct Proof</b>	<b>113</b>
4.1. Theorems	113
4.2. Definitions	115
4.3. Direct Proof	118
4.4. Using Cases	124
4.5. Treating Similar Cases	125
<b>5. Contrapositive Proof</b>	<b>128</b>
5.1. Contrapositive Proof	128
5.2. Congruence of Integers	131
5.3. Mathematical Writing	133
<b>6. Proof by Contradiction</b>	<b>137</b>
6.1. Proving Statements with Contradiction	138
6.2. Proving Conditional Statements by Contradiction	141
6.3. Combining Techniques	142
6.4. Some Words of Advice	143

## *III More on Proof*

<b>7. Proving Non-Conditional Statements</b>	<b>147</b>
7.1. If-and-Only-If Proof	147
7.2. Equivalent Statements	149
7.3. Existence Proofs; Existence and Uniqueness Proofs	150
7.4. Constructive Versus Non-Constructive Proofs	154
<b>8. Proofs Involving Sets</b>	<b>157</b>
8.1. How to Prove $a \in A$	157
8.2. How to Prove $A \subseteq B$	159
8.3. How to Prove $A = B$	162
8.4. Examples: Perfect Numbers	165
<b>9. Disproof</b>	<b>172</b>
9.1. Counterexamples	174
9.2. Disproving Existence Statements	176
9.3. Disproof by Contradiction	178
<b>10. Mathematical Induction</b>	<b>180</b>
10.1. Proof by Induction	182
10.2. Proof by Strong Induction	187
10.3. Proof by Smallest Counterexample	191
10.4. The Fundamental Theorem of Arithmetic	192
10.5. Fibonacci Numbers	193

*IV Relations, Functions and Cardinality*

<b>11. Relations</b>	<b>201</b>
11.1. Relations	201
11.2. Properties of Relations	205
11.3. Equivalence Relations	210
11.4. Equivalence Classes and Partitions	215
11.5. The Integers Modulo $n$	218
11.6. Relations Between Sets	221
<b>12. Functions</b>	<b>223</b>
12.1. Functions	223
12.2. Injective and Surjective Functions	228
12.3. The Pigeonhole Principle Revisited	233
12.4. Composition	235
12.5. Inverse Functions	238
12.6. Image and Preimage	242
<b>13. Proofs in Calculus</b>	<b>244</b>
13.1. The Triangle Inequality	245
13.2. Definition of a Limit	246
13.3. Limits That Do Not Exist	249
13.4. Limits Laws	251
13.5. Continuity and Derivatives	256
13.6. Limits at Infinity	258
13.7. Sequences	261
13.8. Series	265
<b>14. Cardinality of Sets</b>	<b>269</b>
14.1. Sets with Equal Cardinalities	269
14.2. Countable and Uncountable Sets	275
14.3. Comparing Cardinalities	280
14.4. The Cantor-Bernstein-Schröder Theorem	284
<b>Conclusion</b>	<b>291</b>
<b>Solutions</b>	<b>292</b>

---

## Preface to the Third Edition

---

My goal in writing this book has been to create a very inexpensive high-quality textbook. The book can be downloaded from my web page in PDF format for free, and the print version costs considerably less than comparable traditional textbooks.

In this third edition, Chapter 3 (on counting) has been expanded, and a new chapter on calculus proofs has been added. New examples and exercises have been added throughout. My decisions regarding revisions have been guided by both the Amazon reviews and emails from readers, and I am grateful for all comments.

I have taken pains to ensure that the third edition is compatible with the second. Exercises have not been reordered, although some have been edited for clarity and some new ones have been appended. (The one exception is that Chapter 3's reorganization shifted some exercises.) The chapter sequencing is identical between editions, with one exception: The final chapter on cardinality has become Chapter 14 in order to make way for the new Chapter 13 on calculus proofs. There has been a slight renumbering of the sections within chapters 10 and 11, but the numbering of the exercises within the sections is unchanged.

This core of this book is an expansion and refinement of lecture notes I developed while teaching proofs courses over the past 18 years at Virginia Commonwealth University (a large state university) and Randolph-Macon College (a small liberal arts college). I found the needs of these two audiences to be nearly identical, and I wrote this book for them. But I am mindful of a larger audience. I believe this book is suitable for almost any undergraduate mathematics program.

RICHARD HAMMACK

*Lawrenceville, Virginia*

*February 14, 2018*

---

## Introduction

---

This is a book about how to prove theorems.

Until this point in your education, mathematics has probably been presented as a primarily computational discipline. You have learned to solve equations, compute derivatives and integrals, multiply matrices and find determinants; and you have seen how these things can answer practical questions about the real world. In this setting your primary goal in using mathematics has been to compute answers.

But there is another side of mathematics that is more theoretical than computational. Here the primary goal is to understand mathematical structures, to prove mathematical statements, and even to invent or discover new mathematical theorems and theories. The mathematical techniques and procedures that you have learned and used up until now are founded on this theoretical side of mathematics. For example, in computing the area under a curve, you use the fundamental theorem of calculus. It is because this theorem is true that your answer is correct. However, in learning calculus you were probably far more concerned with how that theorem could be applied than in understanding why it is true. But how do we *know* it is true? How can we convince ourselves or others of its validity? Questions of this nature belong to the theoretical realm of mathematics. This book is an introduction to that realm.

This book will initiate you into an esoteric world. You will learn and apply the methods of thought that mathematicians use to verify theorems, explore mathematical truth and create new mathematical theories. This will prepare you for advanced mathematics courses, for you will be better able to understand proofs, write your own proofs and think critically and inquisitively about mathematics.

The book is organized into four parts, as outlined below.

## PART I Fundamentals

- Chapter 1: Sets
- Chapter 2: Logic
- Chapter 3: Counting

Chapters 1 and 2 lay out the language and conventions used in all advanced mathematics. Sets are fundamental because every mathematical structure, object, or entity can be described as a set. Logic is fundamental because it allows us to understand the meanings of statements, to deduce facts about mathematical structures and to uncover further structures. All subsequent chapters build on these first two chapters. Chapter 3 is included partly because its topics are central to many branches of mathematics, but also because it is a source of many examples and exercises that occur throughout the book. (However, the course instructor may choose to omit Chapter 3.)

## PART II Proving Conditional Statements

- Chapter 4: Direct Proof
- Chapter 5: Contrapositive Proof
- Chapter 6: Proof by Contradiction

Chapters 4 through 6 are concerned with three main techniques used for proving theorems that have the “conditional” form “*If P, then Q.*”

## PART III More on Proof

- Chapter 7: Proving Non-Conditional Statements
- Chapter 8: Proofs Involving Sets
- Chapter 9: Disproof
- Chapter 10: Mathematical Induction

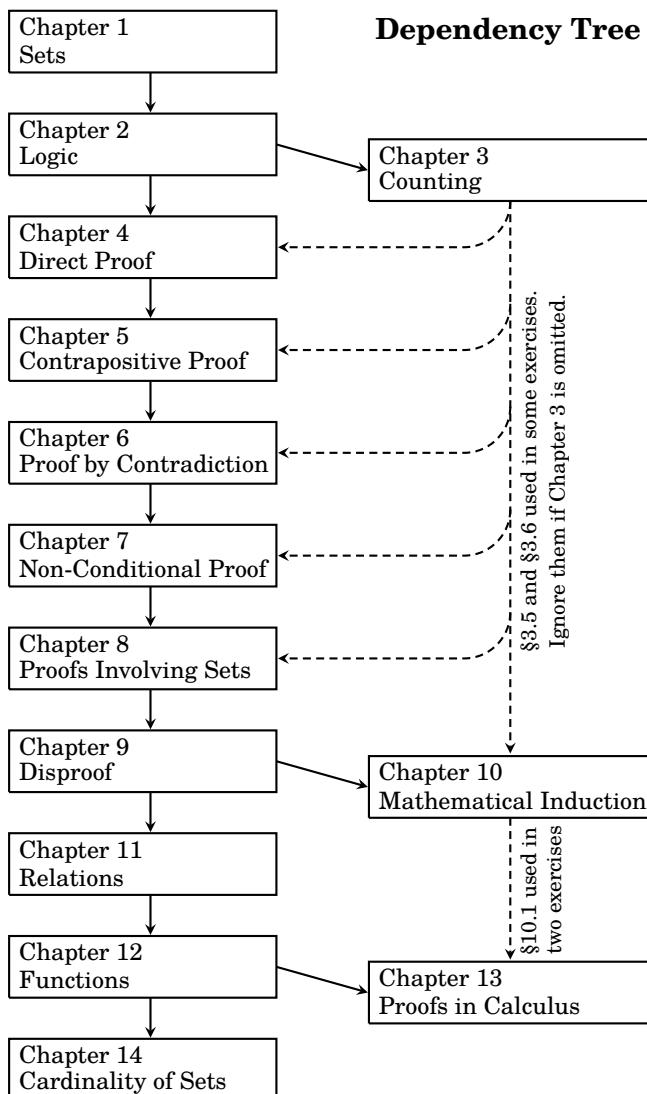
These chapters deal with useful variations, embellishments and consequences of the proof techniques introduced in Chapters 4 through 6.

## PART IV Relations, Functions and Cardinality

- Chapter 11: Relations
- Chapter 12: Functions
- Chapter 13: Proofs in Calculus
- Chapter 14: Cardinality of Sets

These final chapters are mainly concerned with the idea of *functions*, which are central to all of mathematics. Upon mastering this material you will be ready for advanced mathematics courses such as abstract algebra, analysis, topology, combinatorics and theory of computation.

The chapters are organized as in the following dependency tree. The left-hand column forms the core of the book; each chapter in this column uses material from all chapters above it. Chapters 3 and 13 may be omitted without loss of continuity. But the material in Chapter 3 is a great source of exercises, and the reader who omits it should ignore the later exercises that draw from it. Chapter 10, on induction, can also be omitted with no break in continuity. However, induction is a topic that most proof courses will include.



**To the instructor.** The book is designed for a three or four credit course. A course emphasizing discrete mathematics could cover chapters 1–12. A course that is more of a preparation for analysis could cover all but Chapter 3. The following timetable (for a fourteen-week semester) is a hybrid of these two options. Sections marked with \* may require only the briefest mention in class, or may be best left for the students to digest on their own.

Week	Monday	Wednesday	Friday
1	Section 1.1	Section 1.2	Sections 1.3, 1.4
2	Sections 1.5, 1.6, 1.7	Section 1.8	Sections 1.9*, 2.1
3	Section 2.2	Sections 2.3, 2.4	Sections 2.5, 2.6
4	Section 2.7	Sections 2.8*, 2.9	Sections 2.10, 2.11*, 2.12*
5	Sections 3.1, 3.2, 3.3	Section 3.4, 3.5	Sections 3.5, 3.6
6	EXAM	Sections 4.1, 4.2, 4.3	Sections 4.3, 4.4, 4.5*
7	Sections 5.1, 5.2, 5.3*	Section 6.1	Sections 6.2 6.3*
8	Sections 7.1, 7.2*, 7.3	Sections 8.1, 8.2	Section 8.3
9	Section 8.4	Sections 9.1, 9.2, 9.3*	Section 10.1
10	Sections 10.1, 10.4*	Sections 10.2, 10.3	EXAM
11	Sections 11.1, 11.2	Sections 11.3, 11.4	Sections 11.5, 11.6
12	Section 12.1	Section 12.2	Section 12.2
13	Sections 12.3, 12.4*	Section 12.5	Sections 12.5, 12.6*
14	Section 14.1	Section 14.2	Sections 14.3, 14.4*

The *entire* book could be covered in a 4-credit course, or in a 3-credit course pitched to a more mature audience.

**Acknowledgments.** I thank my students in VCU’s MATH 300 courses for offering feedback as they read the first edition of this book. Thanks especially to Cory Colbert and Lauren Pace for rooting out typographical mistakes and inconsistencies. I am especially indebted to Cory for reading early drafts of each chapter and catching numerous mistakes before I posted the final draft on my web page. Cory also created the index, suggested some interesting exercises, and wrote some solutions. Thanks to Moa Apagodu, Sean Cox, Brent Cody and Andy Lewis for suggesting many improvements while teaching from the book. I am indebted to Lon Mitchell, whose expertise with typesetting and on-demand publishing made the print version of this book a reality.

And thanks to countless readers all over the world who contacted me concerning errors and omissions. Because of you, this is a better book.



## *Part I*

---

### *Fundamentals*

---



# CHAPTER 1

---

## Sets

---

All of mathematics can be described with sets. This becomes more and more apparent the deeper into mathematics you go. It will be apparent in most of your upper level courses, and certainly in this course. The theory of sets is a language that is perfectly suited to describing and explaining all types of mathematical structures.

### 1.1 Introduction to Sets

A **set** is a collection of things. The things are called **elements** of the set. We are mainly concerned with sets whose elements are mathematical entities, such as numbers, points, functions, etc.

A set is often expressed by listing its elements between commas, enclosed by braces. For example, the collection  $\{2, 4, 6, 8\}$  is a set which has four elements, the numbers 2, 4, 6 and 8. Some sets have infinitely many elements. For example, consider the collection of all integers,

$$\{\dots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \dots\}.$$

Here the dots indicate a pattern of numbers that continues forever in both the positive and negative directions. A set is called an **infinite** set if it has infinitely many elements; otherwise it is called a **finite** set.

Two sets are **equal** if they contain exactly the same elements. Thus  $\{2, 4, 6, 8\} = \{4, 2, 8, 6\}$  because even though they are listed in a different order, the elements are identical; but  $\{2, 4, 6, 8\} \neq \{2, 4, 6, 7\}$ . Also

$$\{\dots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \dots\} = \{0, -1, 1, -2, 2, -3, 3, -4, 4, \dots\}.$$

We often let uppercase letters stand for sets. In discussing the set  $\{2, 4, 6, 8\}$  we might declare  $A = \{2, 4, 6, 8\}$  and then use  $A$  to stand for  $\{2, 4, 6, 8\}$ . To express that 2 is an element of the set  $A$ , we write  $2 \in A$ , and read this as “2 is an element of  $A$ ,” or “2 is in  $A$ ,” or just “2 in  $A$ .” We also have  $4 \in A$ ,  $6 \in A$  and  $8 \in A$ , but  $5 \notin A$ . We read this last expression as “5 is not an element of  $A$ ,” or “5 not in  $A$ .” Expressions like  $6, 2 \in A$  or  $2, 4, 8 \in A$  are used to indicate that several things are in a set.

Some sets are so significant that we reserve special symbols for them. The set of **natural numbers** (i.e., the positive whole numbers) is denoted by  $\mathbb{N}$ , that is,

$$\mathbb{N} = \{1, 2, 3, 4, 5, 6, 7, \dots\}.$$

### The set of integers

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, 4, \dots\}$$

is another fundamental set. The symbol  $\mathbb{R}$  stands for the set of all **real numbers**, a set that is undoubtedly familiar to you from calculus. Other special sets will be listed later in this section.

Sets need not have just numbers as elements. The set  $B = \{T, F\}$  consists of two letters, perhaps representing the values “true” and “false.” The set  $C = \{a, e, i, o, u\}$  consists of the lowercase vowels in the English alphabet. The set  $D = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$  has as elements the four corner points of a square on the  $x$ - $y$  coordinate plane. Thus  $(0, 0) \in D$ ,  $(1, 0) \in D$ , etc., but  $(1, 2) \notin D$  (for instance). It is even possible for a set to have other sets as elements. Consider  $E = \{1, \{2, 3\}, \{2, 4\}\}$ , which has three elements: the number 1, the set  $\{2, 3\}$  and the set  $\{2, 4\}$ . Thus  $1 \in E$  and  $\{2, 3\} \in E$  and  $\{2, 4\} \in E$ . But note that  $2 \notin E$ ,  $3 \notin E$  and  $4 \notin E$ .

Consider the set  $M = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\}$  of three two-by-two matrices. We have  $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \in M$ , but  $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \notin M$ . Letters can serve as symbols denoting a set’s elements: If  $a = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ ,  $b = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$  and  $c = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ , then  $M = \{a, b, c\}$ .

If  $X$  is a finite set, its **cardinality** or **size** is the number of elements it has, and this number is denoted as  $|X|$ . Thus for the sets above,  $|A| = 4$ ,  $|B| = 2$ ,  $|C| = 5$ ,  $|D| = 4$ ,  $|E| = 3$  and  $|M| = 3$ .

There is a special set that, although small, plays a big role. The **empty set** is the set  $\{\}$  that has no elements. We denote it as  $\emptyset$ , so  $\emptyset = \{\}$ . Whenever you see the symbol  $\emptyset$ , it stands for  $\{\}$ . Observe that  $|\emptyset| = 0$ . The empty set is the only set whose cardinality is zero.

Be careful in writing the empty set. Don’t write  $\{\emptyset\}$  when you mean  $\emptyset$ . These sets can’t be equal because  $\emptyset$  contains nothing while  $\{\emptyset\}$  contains one thing, namely the empty set. If this is confusing, think of a set as a box with things in it, so, for example,  $\{2, 4, 6, 8\}$  is a “box” containing four numbers. The empty set  $\emptyset = \{\}$  is an empty box. By contrast,  $\{\emptyset\}$  is a box with an empty box inside it. Obviously, there’s a difference: An empty box is not the same as a box with an empty box inside it. Thus  $\emptyset \neq \{\emptyset\}$ . (You might also note  $|\emptyset| = 0$  and  $|\{\emptyset\}| = 1$  as additional evidence that  $\emptyset \neq \{\emptyset\}$ .)

This box analogy can help us think about sets. The set  $F = \{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}\}$  may look strange but it is really very simple. Think of it as a box containing three things: an empty box, a box containing an empty box, and a box containing a box containing an empty box. Thus  $|F| = 3$ . The set  $G = \{\mathbb{N}, \mathbb{Z}\}$  is a box containing two boxes, the box of natural numbers and the box of integers. Thus  $|G| = 2$ .

A special notation called **set-builder notation** is used to describe sets that are too big or complex to list between braces. Consider the infinite set of even integers  $E = \{\dots, -6, -4, -2, 0, 2, 4, 6, \dots\}$ . In set-builder notation this set is written as

$$E = \{2n : n \in \mathbb{Z}\}.$$

We read the first brace as “*the set of all things of form*,” and the colon as “*such that*.” So the expression  $E = \{2n : n \in \mathbb{Z}\}$  reads as “*E equals the set of all things of form*  $2n$ , *such that*  $n$  is an element of  $\mathbb{Z}$ .” The idea is that  $E$  consists of all possible values of  $2n$ , where  $n$  takes on all values in  $\mathbb{Z}$ .

In general, a set  $X$  written with set-builder notation has the syntax

$$X = \{\text{expression} : \text{rule}\},$$

where the elements of  $X$  are understood to be all values of “expression” that are specified by “rule.” For example, above  $E$  is the set of all values of the expression  $2n$  that satisfy the rule  $n \in \mathbb{Z}$ . There can be many ways to express the same set. For example,  $E = \{2n : n \in \mathbb{Z}\} = \{n : n \text{ is an even integer}\} = \{n : n = 2k, k \in \mathbb{Z}\}$ . Another common way of writing it is

$$E = \{n \in \mathbb{Z} : n \text{ is even}\},$$

read “*E is the set of all n in  $\mathbb{Z}$  such that n is even*.” Some writers use a bar instead of a colon; for example,  $E = \{n \in \mathbb{Z} | n \text{ is even}\}$ . We use the colon.

**Example 1.1** Here are some further illustrations of set-builder notation.

1.  $\{n : n \text{ is a prime number}\} = \{2, 3, 5, 7, 11, 13, 17, \dots\}$
2.  $\{n \in \mathbb{N} : n \text{ is prime}\} = \{2, 3, 5, 7, 11, 13, 17, \dots\}$
3.  $\{n^2 : n \in \mathbb{Z}\} = \{0, 1, 4, 9, 16, 25, \dots\}$
4.  $\{x \in \mathbb{R} : x^2 - 2 = 0\} = \{\sqrt{2}, -\sqrt{2}\}$
5.  $\{x \in \mathbb{Z} : x^2 - 2 = 0\} = \emptyset$
6.  $\{x \in \mathbb{Z} : |x| < 4\} = \{-3, -2, -1, 0, 1, 2, 3\}$
7.  $\{2x : x \in \mathbb{Z}, |x| < 4\} = \{-6, -4, -2, 0, 2, 4, 6\}$
8.  $\{x \in \mathbb{Z} : |2x| < 4\} = \{-1, 0, 1\}$

Items 6–8 above highlight a conflict of notation that we must always be alert to. The expression  $|X|$  means *absolute value* if  $X$  is a number and *cardinality* if  $X$  is a set. The distinction should always be clear from context. Consider  $\{x \in \mathbb{Z} : |x| < 4\}$  in Example 1.1 (6) above. Here  $x \in \mathbb{Z}$ , so  $x$  is a number (not a set), and thus the bars in  $|x|$  must mean absolute value, not cardinality. On the other hand, suppose  $A = \{\{1, 2\}, \{3, 4, 5, 6\}, \{7\}\}$  and  $B = \{X \in A : |X| < 3\}$ . The elements of  $A$  are sets (not numbers), so the  $|X|$  in the expression for  $B$  must mean cardinality. Therefore  $B = \{\{1, 2\}, \{7\}\}$ .

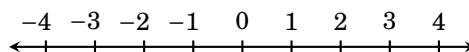
**Example 1.2** Describe the set  $A = \{7a + 3b : a, b \in \mathbb{Z}\}$ .

**Solution:** This set contains all numbers of form  $7a + 3b$ , where  $a$  and  $b$  are integers. Each such number  $7a + 3b$  is an integer, so  $A$  contains only integers. But *which* integers? If  $n$  is *any* integer, then  $n = 7n + 3(-2n)$ , so  $n = 7a + 3b$  where  $a = 7n$  and  $b = -2n$ . Therefore  $n \in A$ . We've now shown that  $A$  contains only integers, and also that every integer is an element of  $A$ . Consequently  $A = \mathbb{Z}$ .

We close this section with a summary of special sets. These are sets that are so common that they are given special names and symbols.

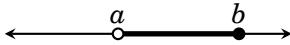
- The empty set:  $\emptyset = \{\}$
- The natural numbers:  $\mathbb{N} = \{1, 2, 3, 4, 5, \dots\}$
- The integers:  $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, 4, 5, \dots\}$
- The rational numbers:  $\mathbb{Q} = \left\{x : x = \frac{m}{n}, \text{ where } m, n \in \mathbb{Z} \text{ and } n \neq 0\right\}$
- The real numbers:  $\mathbb{R}$

We visualize the set  $\mathbb{R}$  of real numbers as an infinitely long number line.



Notice that  $\mathbb{Q}$  is the set of all numbers in  $\mathbb{R}$  that can be expressed as a fraction of two integers. You may be aware that  $\mathbb{Q} \neq \mathbb{R}$ , as  $\sqrt{2} \notin \mathbb{Q}$  but  $\sqrt{2} \in \mathbb{R}$ . (If not, this point will be addressed in Chapter 6.)

In calculus you encountered intervals on the number line. Like  $\mathbb{R}$ , these too are infinite sets of numbers. Any two numbers  $a, b \in \mathbb{R}$  with  $a < b$  give rise to various intervals. Graphically, they are represented by a darkened segment on the number line between  $a$  and  $b$ . A solid circle at an endpoint indicates that that number is included in the interval. A hollow circle indicates a point that is not included in the interval.

- Closed interval:  $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$  
- Open interval:  $(a, b) = \{x \in \mathbb{R} : a < x < b\}$  
- Half-open interval:  $(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$  
- Half-open interval:  $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$  
- Infinite interval:  $(a, \infty) = \{x \in \mathbb{R} : a < x\}$  
- Infinite interval:  $[a, \infty) = \{x \in \mathbb{R} : a \leq x\}$  
- Infinite interval:  $(-\infty, b) = \{x \in \mathbb{R} : x < b\}$  
- Infinite interval:  $(-\infty, b] = \{x \in \mathbb{R} : x \leq b\}$  

Each of these intervals is an infinite set containing infinitely many numbers as elements. For example, though its length is short, the interval  $(0.1, 0.2)$  contains infinitely many numbers, that is, all numbers between 0.1 and 0.2. It is an unfortunate notational accident that  $(a, b)$  can denote both an open interval on the line and a point on the plane. The difference is usually clear from context. In the next section we will see yet another meaning of  $(a, b)$ .

### Exercises for Section 1.1

**A.** Write each of the following sets by listing their elements between braces.

1.  $\{5x - 1 : x \in \mathbb{Z}\}$
2.  $\{3x + 2 : x \in \mathbb{Z}\}$
3.  $\{x \in \mathbb{Z} : -2 \leq x < 7\}$
4.  $\{x \in \mathbb{N} : -2 < x \leq 7\}$
5.  $\{x \in \mathbb{R} : x^2 = 3\}$
6.  $\{x \in \mathbb{R} : x^2 = 9\}$
7.  $\{x \in \mathbb{R} : x^2 + 5x = -6\}$
8.  $\{x \in \mathbb{R} : x^3 + 5x^2 = -6x\}$
9.  $\{x \in \mathbb{R} : \sin \pi x = 0\}$
10.  $\{x \in \mathbb{R} : \cos x = 1\}$
11.  $\{x \in \mathbb{Z} : |x| < 5\}$
12.  $\{x \in \mathbb{Z} : |2x| < 5\}$
13.  $\{x \in \mathbb{Z} : |6x| < 5\}$
14.  $\{5x : x \in \mathbb{Z}, |2x| \leq 8\}$
15.  $\{5a + 2b : a, b \in \mathbb{Z}\}$
16.  $\{6a + 2b : a, b \in \mathbb{Z}\}$

**B.** Write each of the following sets in set-builder notation.

17.  $\{2, 4, 8, 16, 32, 64, \dots\}$
18.  $\{0, 4, 16, 36, 64, 100, \dots\}$
19.  $\{\dots, -6, -3, 0, 3, 6, 9, 12, 15, \dots\}$
20.  $\{\dots, -8, -3, 2, 7, 12, 17, \dots\}$
21.  $\{0, 1, 4, 9, 16, 25, 36, \dots\}$
22.  $\{3, 6, 11, 18, 27, 38, \dots\}$
23.  $\{3, 4, 5, 6, 7, 8\}$
24.  $\{-4, -3, -2, -1, 0, 1, 2\}$
25.  $\{\dots, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, \dots\}$
26.  $\{\dots, \frac{1}{27}, \frac{1}{9}, \frac{1}{3}, 1, 3, 9, 27, \dots\}$
27.  $\{\dots, -\pi, -\frac{\pi}{2}, 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi, \frac{5\pi}{2}, \dots\}$
28.  $\{\dots, -\frac{3}{2}, -\frac{3}{4}, 0, \frac{3}{4}, \frac{3}{2}, 3, \frac{15}{4}, \frac{9}{2}, \dots\}$

C. Find the following cardinalities.

- |   |  |
|---|--|
| 29. $ \{\{1\}, \{2, \{3, 4\}\}, \emptyset\} $<br>30. $ \{\{1, 4\}, a, b, \{\{3, 4\}\}, \{\emptyset\}\} $<br>31. $ \{\{\{1\}, \{2, \{3, 4\}\}\}, \emptyset\} $<br>32. $ \{\{\{1, 4\}, a, b, \{\{3, 4\}\}, \{\emptyset\}\}\} $<br>33. $ \{x \in \mathbb{Z} :  x  < 10\} $ | 34. $ \{x \in \mathbb{N} :  x  < 10\} $<br>35. $ \{x \in \mathbb{Z} : x^2 < 10\} $<br>36. $ \{x \in \mathbb{N} : x^2 < 10\} $<br>37. $ \{x \in \mathbb{N} : x^2 < 0\} $<br>38. $ \{x \in \mathbb{N} : 5x \leq 20\} $ |
|---|--|

D. Sketch the following sets of points in the  $x$ - $y$  plane.

- |   |   |
|---|---|
| 39. $\{(x, y) : x \in [1, 2], y \in [1, 2]\}$<br>40. $\{(x, y) : x \in [0, 1], y \in [1, 2]\}$<br>41. $\{(x, y) : x \in [-1, 1], y = 1\}$<br>42. $\{(x, y) : x = 2, y \in [0, 1]\}$<br>43. $\{(x, y) :  x  = 2, y \in [0, 1]\}$<br>44. $\{(x, x^2) : x \in \mathbb{R}\}$<br>45. $\{(x, y) : x, y \in \mathbb{R}, x^2 + y^2 = 1\}$ | 46. $\{(x, y) : x, y \in \mathbb{R}, x^2 + y^2 \leq 1\}$<br>47. $\{(x, y) : x, y \in \mathbb{R}, y \geq x^2 - 1\}$<br>48. $\{(x, y) : x, y \in \mathbb{R}, x > 1\}$<br>49. $\{(x, x+y) : x \in \mathbb{R}, y \in \mathbb{Z}\}$<br>50. $\{(x, \frac{x^2}{y}) : x \in \mathbb{R}, y \in \mathbb{N}\}$<br>51. $\{(x, y) \in \mathbb{R}^2 : (y-x)(y+x) = 0\}$<br>52. $\{(x, y) \in \mathbb{R}^2 : (y-x^2)(y+x^2) = 0\}$ |
|---|---|

## 1.2 The Cartesian Product

Given two sets  $A$  and  $B$ , it is possible to “multiply” them to produce a new set denoted as  $A \times B$ . This operation is called the *Cartesian product*. To understand it, we must first understand the idea of an ordered pair.

**Definition 1.1** An **ordered pair** is a list  $(x, y)$  of two things  $x$  and  $y$ , enclosed in parentheses and separated by a comma.

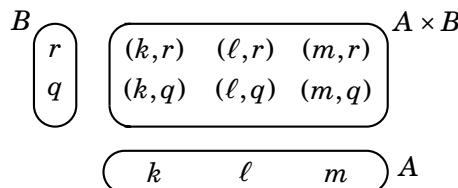
For example,  $(2, 4)$  is an ordered pair, as is  $(4, 2)$ . These ordered pairs are different because even though they have the same things in them, the order is different. We write  $(2, 4) \neq (4, 2)$ . Right away you can see that ordered pairs can be used to describe points on the plane, as was done in calculus, but they are not limited to just that. The things in an ordered pair don’t have to be numbers. You can have ordered pairs of letters, such as  $(\ell, m)$ , ordered pairs of sets such as  $(\{2, 5\}, \{3, 2\})$ , even ordered pairs of ordered pairs like  $((2, 4), (4, 2))$ . The following are also ordered pairs:  $(2, \{1, 2, 3\})$ ,  $(\mathbb{R}, (0, 0))$ . Any list of two things enclosed by parentheses is an ordered pair. Now we are ready to define the Cartesian product.

**Definition 1.2** The **Cartesian product** of two sets  $A$  and  $B$  is another set, denoted as  $A \times B$  and defined as  $A \times B = \{(a, b) : a \in A, b \in B\}$ .

Thus  $A \times B$  is a set of ordered pairs of elements from  $A$  and  $B$ . For example, if  $A = \{k, \ell, m\}$  and  $B = \{q, r\}$ , then

$$A \times B = \{(k, q), (k, r), (\ell, q), (\ell, r), (m, q), (m, r)\}.$$

Figure 1.1 shows how to make a schematic diagram of  $A \times B$ . Line up the elements of  $A$  horizontally and line up the elements of  $B$  vertically, as if  $A$  and  $B$  form an  $x$ - and  $y$ -axis. Then fill in the ordered pairs so that each element  $(x, y)$  is in the column headed by  $x$  and the row headed by  $y$ .

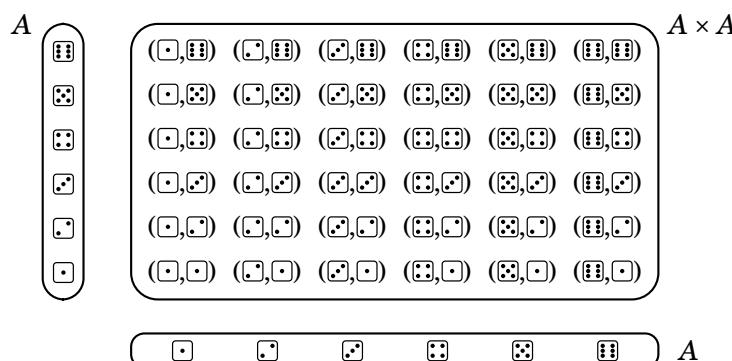


**Figure 1.1.** A diagram of a Cartesian product

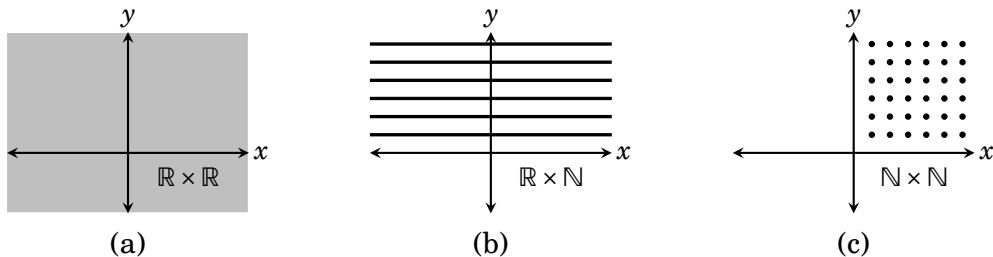
For another example,  $\{0, 1\} \times \{2, 1\} = \{(0, 2), (0, 1), (1, 2), (1, 1)\}$ . If you are a visual thinker, you may wish to draw a diagram similar to Figure 1.1. The rectangular array of such diagrams give us the following general fact.

**Fact 1.1** If  $A$  and  $B$  are finite sets, then  $|A \times B| = |A| \cdot |B|$ .

**Example 1.3** Let  $A = \{\square, \blacksquare, \blacksquare\square, \square\square, \blacksquare\blacksquare, \blacksquare\square\square\}$  be the set consisting of the six faces of a dice. The Cartesian product  $A \times A$  is diagrammed below. By Fact 1.1 (or by simple counting),  $|A \times A| = 6 \cdot 6 = 36$ . We might think of  $A \times A$  as the set of possible outcomes in rolling a dice two times in a row. Each element of the product is an ordered pair of form (result of 1st roll, result of 2nd roll). Such constructions are useful in the study of probability.



The set  $\mathbb{R} \times \mathbb{R} = \{(x, y) : x, y \in \mathbb{R}\}$  should be very familiar. It can be viewed as the set of points on the Cartesian plane, as drawn in Figure 1.2(a). The set  $\mathbb{R} \times \mathbb{N} = \{(x, y) : x \in \mathbb{R}, y \in \mathbb{N}\}$  can be regarded as all of the points on the plane whose second coordinate is a natural number. This is illustrated in Figure 1.2(b), which shows that  $\mathbb{R} \times \mathbb{N}$  looks like infinitely many horizontal lines at integer heights above the  $x$ -axis. The set  $\mathbb{N} \times \mathbb{N}$  is the set of all points on the plane whose coordinates are both natural numbers. It looks like a grid of dots in the first quadrant, as illustrated in Figure 1.2(c).



**Figure 1.2.** Drawings of some Cartesian products

It is even possible for one factor of a Cartesian product to be a Cartesian product itself, as in  $\mathbb{R} \times (\mathbb{N} \times \mathbb{Z}) = \{(x, (y, z)) : x \in \mathbb{R}, (y, z) \in \mathbb{N} \times \mathbb{Z}\}$ .

We can also define Cartesian products of three or more sets by moving beyond ordered pairs. An **ordered triple** is a list  $(x, y, z)$ . The Cartesian product of the three sets  $\mathbb{R}$ ,  $\mathbb{N}$  and  $\mathbb{Z}$  is  $\mathbb{R} \times \mathbb{N} \times \mathbb{Z} = \{(x, y, z) : x \in \mathbb{R}, y \in \mathbb{N}, z \in \mathbb{Z}\}$ . Of course there is no reason to stop with ordered triples. In general,

$$A_1 \times A_2 \times \cdots \times A_n = \{(x_1, x_2, \dots, x_n) : x_i \in A_i \text{ for each } i = 1, 2, \dots, n\}.$$

Be mindful of parentheses. There is a slight difference between  $\mathbb{R} \times (\mathbb{N} \times \mathbb{Z})$  and  $\mathbb{R} \times \mathbb{N} \times \mathbb{Z}$ . The first is a Cartesian product of two sets; its elements are ordered pairs  $(x, (y, z))$ . The second is a Cartesian product of three sets; its elements are ordered triples  $(x, y, z)$ . To be sure, in many situations there is no harm in blurring the distinction between expressions like  $(x, (y, z))$  and  $(x, y, z)$ , but for now we regard them as different.

For any set  $A$  and positive integer  $n$ , the **Cartesian power**  $A^n$  is

$$A^n = A \times A \times \cdots \times A = \{(x_1, x_2, \dots, x_n) : x_1, x_2, \dots, x_n \in A\}.$$

In this way,  $\mathbb{R}^2$  is the familiar Cartesian plane and  $\mathbb{R}^3$  is three-dimensional space. You can visualize how, if  $\mathbb{R}^2$  is the plane, then  $\mathbb{Z}^2 = \{(m, n) : m, n \in \mathbb{Z}\}$  is a grid of points on the plane. Likewise, as  $\mathbb{R}^3$  is 3-dimensional space,  $\mathbb{Z}^3 = \{(m, n, p) : m, n, p \in \mathbb{Z}\}$  is a grid of points in space.

In other courses you may encounter sets that are very similar to  $\mathbb{R}^n$ , but yet have slightly different shades of meaning. Consider, for example, the set of all two-by-three matrices with entries from  $\mathbb{R}$ :

$$M = \left\{ \begin{bmatrix} u & v & w \\ x & y & z \end{bmatrix} : u, v, w, x, y, z \in \mathbb{R} \right\}.$$

This is not really all that different from the set

$$\mathbb{R}^6 = \{(u, v, w, x, y, z) : u, v, w, x, y, z \in \mathbb{R}\}.$$

The elements of these sets are merely certain arrangements of six real numbers. Despite their similarity, we maintain that  $M \neq \mathbb{R}^6$ , for two-by-three matrices are not the same things as sequences of six numbers.

**Example 1.4** Represent the two sides of a coin by the set  $S = \{\text{H}, \text{T}\}$ . The possible outcomes of tossing the coin seven times in a row can be described with the Cartesian power  $S^7$ . A typical element of  $S^7$  looks like

$$(\text{H}, \text{H}, \text{T}, \text{H}, \text{T}, \text{T}, \text{T}),$$

meaning a head was tossed first, then another head, then a tail, then a head followed by three tails. Note that  $|S^7| = 2^7 = 128$ , so there are 128 possible outcomes. If this is not clear, then it will be explained fully in Chapter 3.

## Exercises for Section 1.2

A. Write out the indicated sets by listing their elements between braces.

1. Suppose  $A = \{1, 2, 3, 4\}$  and  $B = \{a, c\}$ .
 

(a) $A \times B$	(c) $A \times A$	(e) $\emptyset \times B$	(g) $A \times (B \times B)$
(b) $B \times A$	(d) $B \times B$	(f) $(A \times B) \times B$	(h) $B^3$
2. Suppose  $A = \{\pi, e, 0\}$  and  $B = \{0, 1\}$ .
 

(a) $A \times B$	(c) $A \times A$	(e) $A \times \emptyset$	(g) $A \times (B \times B)$
(b) $B \times A$	(d) $B \times B$	(f) $(A \times B) \times B$	(h) $A \times B \times B$
3.  $\{x \in \mathbb{R} : x^2 = 2\} \times \{a, c, e\}$
4.  $\{n \in \mathbb{Z} : 2 < n < 5\} \times \{n \in \mathbb{Z} : |n| = 5\}$
5.  $\{x \in \mathbb{R} : x^2 = 2\} \times \{x \in \mathbb{R} : |x| = 2\}$
6.  $\{x \in \mathbb{R} : x^2 = x\} \times \{x \in \mathbb{N} : x^2 = x\}$
7.  $\{\emptyset\} \times \{0, \emptyset\} \times \{0, 1\}$
8.  $\{0, 1\}^4$

B. Sketch these Cartesian products on the  $x$ - $y$  plane  $\mathbb{R}^2$  (or  $\mathbb{R}^3$  for the last two).

9.  $\{1, 2, 3\} \times \{-1, 0, 1\}$
10.  $\{-1, 0, 1\} \times \{1, 2, 3\}$
11.  $[0, 1] \times [0, 1]$
12.  $[-1, 1] \times [1, 2]$
13.  $\{1, 1.5, 2\} \times [1, 2]$
14.  $[1, 2] \times \{1, 1.5, 2\}$
15.  $\{1\} \times [0, 1]$
16.  $[0, 1] \times \{1\}$
17.  $\mathbb{N} \times \mathbb{Z}$
18.  $\mathbb{Z} \times \mathbb{Z}$
19.  $[0, 1] \times [0, 1] \times [0, 1]$
20.  $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\} \times [0, 1]$

### 1.3 Subsets

It can happen that every element of a set  $A$  is an element of another set  $B$ . For example, each element of  $A = \{0, 2, 4\}$  is also an element of  $B = \{0, 1, 2, 3, 4\}$ . When  $A$  and  $B$  are related this way we say that  $A$  is a *subset* of  $B$ .

**Definition 1.3** Suppose  $A$  and  $B$  are sets. If every element of  $A$  is also an element of  $B$ , then we say  $A$  is a **subset** of  $B$ , and we denote this as  $A \subseteq B$ . We write  $A \not\subseteq B$  if  $A$  is *not* a subset of  $B$ , that is, if it is *not* true that every element of  $A$  is also an element of  $B$ . Thus  $A \not\subseteq B$  means that there is at least one element of  $A$  that is *not* an element of  $B$ .

**Example 1.5** Be sure you understand why each of the following is true.

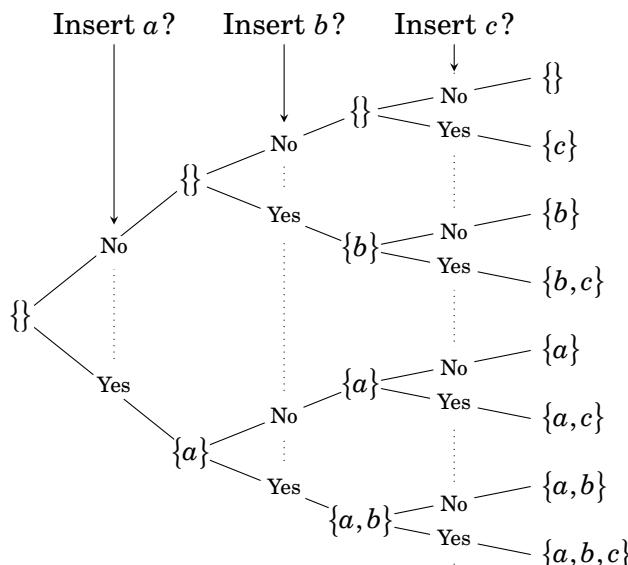
1.  $\{2, 3, 7\} \subseteq \{2, 3, 4, 5, 6, 7\}$
2.  $\{2, 3, 7\} \not\subseteq \{2, 4, 5, 6, 7\}$
3.  $\{2, 3, 7\} \subseteq \{2, 3, 7\}$
4.  $\{(x, \sin(x)) : x \in \mathbb{R}\} \subseteq \mathbb{R}^2$
5.  $\{1, 3, 5, 7, 11, 13, 17, \dots\} \subseteq \mathbb{N}$
6.  $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$
7.  $\mathbb{R} \times \mathbb{N} \subseteq \mathbb{R} \times \mathbb{R}$
8.  $A \subseteq A$  for any set  $A$
9.  $\emptyset \subseteq \emptyset$ .

This brings us to a significant fact: If  $B$  is any set whatsoever, then  $\emptyset \subseteq B$ . To see why this is true, look at the last sentence of Definition 1.3. It says that  $\emptyset \not\subseteq B$  would mean that there is at least one element of  $\emptyset$  that is not an element of  $B$ . But this cannot be so because  $\emptyset$  contains no elements! Thus it is not the case that  $\emptyset \not\subseteq B$ , so it must be that  $\emptyset \subseteq B$ .

**Fact 1.2** The empty set is a subset of all sets, that is,  $\emptyset \subseteq B$  for any set  $B$ .

Here is another way to look at it. Imagine a subset of  $B$  as a thing you make by starting with braces  $\{\}$ , then filling them with selections from  $B$ . For example, to make one particular subset of  $B = \{a, b, c\}$ , start with  $\{\}$ , select  $b$  and  $c$  from  $B$  and insert them into  $\{\}$  to form the subset  $\{b, c\}$ . Alternatively, you could have chosen just  $a$  to make  $\{a\}$ , and so on. But one option is to simply select nothing from  $B$ . This leaves you with the subset  $\{\}$ . Thus  $\{\} \subseteq B$ . More often we write it as  $\emptyset \subseteq B$ .

This idea of “making” a subset can help us list out all the subsets of a given set  $B$ . As an example, let  $B = \{a, b, c\}$ . Let’s list all of its subsets. One way of approaching this is to make a tree-like structure. Begin with the subset  $\{\}$ , which is shown on the left of Figure 1.3. Considering the element  $a$  of  $B$ , we have a choice: insert it into  $\{\}$ , or not. The lines from  $\{\}$  point to what we get depending whether or not we insert  $a$ , either  $\{\}$  or  $\{a\}$ . Now move on to the element  $b$  of  $B$ . For each of the sets just formed we can either insert or not insert  $b$ , and the lines on the diagram point to the resulting sets  $\{\}$ ,  $\{b\}$ ,  $\{a\}$ , or  $\{a, b\}$ . Finally, to each of these sets, we can either insert  $c$  or not insert it, and this gives us, on the far right-hand column, the sets  $\{\}$ ,  $\{c\}$ ,  $\{b\}$ ,  $\{b, c\}$ ,  $\{a\}$ ,  $\{a, c\}$ ,  $\{a, b\}$  and  $\{a, b, c\}$ . These are the eight subsets of  $B = \{a, b, c\}$ .



**Figure 1.3.** A “tree” for listing subsets

We can see from the way this tree branches that if it happened that  $B = \{a\}$ , then  $B$  would have just two subsets, those in the second column of the diagram. If it happened that  $B = \{a, b\}$ , then  $B$  would have four subsets, those in the third column, and so on. At each branching of the tree, the number of subsets doubles. So in general, if  $|B| = n$ , then  $B$  has  $2^n$  subsets.

**Fact 1.3** If a finite set has  $n$  elements, then it has  $2^n$  subsets.

For a slightly more complex example, consider listing the subsets of  $B = \{1, 2, \{1, 3\}\}$ . This  $B$  has just three elements: 1, 2 and  $\{1, 3\}$ . At this point you probably don't even have to draw a tree to list out  $B$ 's subsets. You just make all the possible selections from  $B$  and put them between braces to get

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

These are the eight subsets of  $B$ . Exercises like this help you identify what is and isn't a subset. You know immediately that a set such as  $\{1, 3\}$  is *not* a subset of  $B$  because it can't be made by inserting elements from  $B$  into  $\{\}$ , as the 3 is not an element of  $B$  and thus is not a valid selection. Notice that although  $\{1, 3\} \notin B$ , it *is* true that  $\{1, 3\} \in B$ . Also,  $\{\{1, 3\}\} \subseteq B$ .

**Example 1.6** Be sure you understand why the following statements are true. Each illustrates an aspect of set theory that you've learned so far.

1.  $1 \in \{1, \{1\}\}$  ..... 1 is the first element listed in  $\{1, \{1\}\}$
2.  $1 \notin \{1, \{1\}\}$  ..... because 1 is not a set
3.  $\{1\} \in \{1, \{1\}\}$  .....  $\{1\}$  is the second element listed in  $\{1, \{1\}\}$
4.  $\{1\} \subseteq \{1, \{1\}\}$  ..... make subset  $\{1\}$  by selecting 1 from  $\{1, \{1\}\}$
5.  $\{\{1\}\} \notin \{1, \{1\}\}$  ..... because  $\{1, \{1\}\}$  contains only 1 and  $\{1\}$ , and not  $\{\{1\}\}$
6.  $\{\{1\}\} \subseteq \{1, \{1\}\}$  ..... make subset  $\{\{1\}\}$  by selecting  $\{1\}$  from  $\{1, \{1\}\}$
7.  $\mathbb{N} \notin \mathbb{N}$  .....  $\mathbb{N}$  is a set (not a number) and  $\mathbb{N}$  contains only numbers
8.  $\mathbb{N} \subseteq \mathbb{N}$  ..... because  $X \subseteq X$  for every set  $X$
9.  $\emptyset \notin \mathbb{N}$  ..... because the set  $\mathbb{N}$  contains only numbers and no sets
10.  $\emptyset \subseteq \mathbb{N}$  ..... because  $\emptyset$  is a subset of every set
11.  $\mathbb{N} \in \{\mathbb{N}\}$  ..... because  $\{\mathbb{N}\}$  has just one element, the set  $\mathbb{N}$
12.  $\mathbb{N} \notin \{\mathbb{N}\}$  ..... because, for instance,  $1 \in \mathbb{N}$  but  $1 \notin \{\mathbb{N}\}$
13.  $\emptyset \notin \{\mathbb{N}\}$  ..... note that the only element of  $\{\mathbb{N}\}$  is  $\mathbb{N}$ , and  $\mathbb{N} \neq \emptyset$
14.  $\emptyset \subseteq \{\mathbb{N}\}$  ..... because  $\emptyset$  is a subset of every set
15.  $\emptyset \in \{\emptyset, \mathbb{N}\}$  .....  $\emptyset$  is the first element listed in  $\{\emptyset, \mathbb{N}\}$
16.  $\emptyset \subseteq \{\emptyset, \mathbb{N}\}$  ..... because  $\emptyset$  is a subset of every set
17.  $\{\mathbb{N}\} \subseteq \{\emptyset, \mathbb{N}\}$  ..... make subset  $\{\mathbb{N}\}$  by selecting  $\mathbb{N}$  from  $\{\emptyset, \mathbb{N}\}$
18.  $\{\mathbb{N}\} \notin \{\emptyset, \{\mathbb{N}\}\}$  ..... because  $\mathbb{N} \notin \{\emptyset, \{\mathbb{N}\}\}$
19.  $\{\mathbb{N}\} \in \{\emptyset, \{\mathbb{N}\}\}$  .....  $\{\mathbb{N}\}$  is the second element listed in  $\{\emptyset, \{\mathbb{N}\}\}$
20.  $\{(1, 2), (2, 2), (7, 1)\} \subseteq \mathbb{N} \times \mathbb{N}$  ..... each of  $(1, 2)$ ,  $(2, 2)$ ,  $(7, 1)$  is in  $\mathbb{N} \times \mathbb{N}$

Though they should help you understand the concept of subset, the above examples are somewhat artificial. But in general, subsets arise very naturally. For instance, consider the unit circle  $C = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ .

This is a subset  $C \subseteq \mathbb{R}^2$ . Likewise the graph of a function  $y = f(x)$  is a set of points  $G = \{(x, f(x)) : x \in \mathbb{R}\}$ , and  $G \subseteq \mathbb{R}^2$ . Surely sets such as  $C$  and  $G$  are more easily understood or visualized when regarded as subsets of  $\mathbb{R}^2$ . Mathematics is filled with such instances where it is important to regard one set as a subset of another.

### Exercises for Section 1.3

**A.** List all the subsets of the following sets.

- |                          |   |
|--------------------------|---|
| 1. $\{1, 2, 3, 4\}$      | 5. $\{\emptyset\}$                              |
| 2. $\{1, 2, \emptyset\}$ | 6. $\{\mathbb{R}, \mathbb{Q}, \mathbb{N}\}$     |
| 3. $\{\{\mathbb{R}\}\}$  | 7. $\{\mathbb{R}, \{\mathbb{Q}, \mathbb{N}\}\}$ |
| 4. $\emptyset$           | 8. $\{\{0, 1\}, \{0, 1, \{2\}\}, \{0\}\}$       |

**B.** Write out the following sets by listing their elements between braces.

- |   |  |
|---|--|
| 9. $\{X : X \subseteq \{3, 2, a\} \text{ and }  X  = 2\}$ | 11. $\{X : X \subseteq \{3, 2, a\} \text{ and }  X  = 4\}$ |
| 10. $\{X \subseteq \mathbb{N} :  X  \leq 1\}$             | 12. $\{X : X \subseteq \{3, 2, a\} \text{ and }  X  = 1\}$ |

**C.** Decide if the following statements are true or false. Explain.

- |   |   |
|---|---|
| 13. $\mathbb{R}^3 \subseteq \mathbb{R}^3$ | 15. $\{(x, y) : x - 1 = 0\} \subseteq \{(x, y) : x^2 - x = 0\}$ |
| 14. $\mathbb{R}^2 \subseteq \mathbb{R}^3$ | 16. $\{(x, y) : x^2 - x = 0\} \subseteq \{(x, y) : x - 1 = 0\}$ |

### 1.4 Power Sets

Given a set, you can form a new set with the *power set* operation.

**Definition 1.4** If  $A$  is a set, the **power set** of  $A$  is another set, denoted as  $\mathcal{P}(A)$  and defined to be the set of all subsets of  $A$ . In symbols,  $\mathcal{P}(A) = \{X : X \subseteq A\}$ .

For example, suppose  $A = \{1, 2, 3\}$ . The power set of  $A$  is the set of all subsets of  $A$ . We learned how to find these subsets in the previous section, and they are  $\{\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}$  and  $\{1, 2, 3\}$ . Therefore the power set of  $A$  is

$$\mathcal{P}(A) = \left\{ \emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\} \right\}.$$

As we saw in the previous section, if a finite set  $A$  has  $n$  elements, then it has  $2^n$  subsets, and thus its power set has  $2^n$  elements.

**Fact 1.4** If  $A$  is a finite set, then  $|\mathcal{P}(A)| = 2^{|A|}$ .

**Example 1.7** You should examine the following statements and make sure you understand how the answers were obtained. In particular, notice that in each instance the equation  $|\mathcal{P}(A)| = 2^{|A|}$  is true.

1.  $\mathcal{P}(\{0,1,3\}) = \{\emptyset, \{0\}, \{1\}, \{3\}, \{0,1\}, \{0,3\}, \{1,3\}, \{0,1,3\}\}$
2.  $\mathcal{P}(\{1,2\}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$
3.  $\mathcal{P}(\{1\}) = \{\emptyset, \{1\}\}$
4.  $\mathcal{P}(\emptyset) = \{\emptyset\}$
5.  $\mathcal{P}(\{a\}) = \{\emptyset, \{a\}\}$
6.  $\mathcal{P}(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\}$
7.  $\mathcal{P}(\{a\}) \times \mathcal{P}(\{\emptyset\}) = \{(\emptyset, \emptyset), (\emptyset, \{\emptyset\}), (\{a\}, \emptyset), (\{a\}, \{\emptyset\})\}$
8.  $\mathcal{P}(\mathcal{P}(\{\emptyset\})) = \{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \{\emptyset, \{\emptyset\}\}\}$
9.  $\mathcal{P}(\{1, \{1,2\}\}) = \{\emptyset, \{1\}, \{\{1,2\}\}, \{1, \{1,2\}\}\}$
10.  $\mathcal{P}(\{\mathbb{Z}, \mathbb{N}\}) = \{\emptyset, \{\mathbb{Z}\}, \{\mathbb{N}\}, \{\mathbb{Z}, \mathbb{N}\}\}$

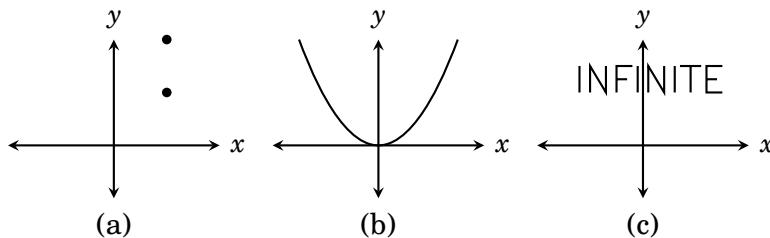
Next are some that are **wrong**. See if you can determine why they are wrong and make sure you understand the explanation on the right.

11.  $\mathcal{P}(1) = \{\emptyset, \{1\}\}$  ..... meaningless because 1 is not a set
12.  $\mathcal{P}(\{1, \{1,2\}\}) = \{\emptyset, \{1\}, \{1,2\}, \{1, \{1,2\}\}\}$  ..... wrong because  $\{1,2\} \not\subseteq \{1, \{1,2\}\}$
13.  $\mathcal{P}(\{1, \{1,2\}\}) = \{\emptyset, \{\{1\}\}, \{\{1,2\}\}, \{\emptyset, \{1,2\}\}\}$  ... wrong because  $\{\{1\}\} \not\subseteq \{1, \{1,2\}\}$

If  $A$  is finite, then it is possible (though maybe not practical) to list out  $\mathcal{P}(A)$  between braces as was done in the above example. That is not possible if  $A$  is infinite. For example, consider  $\mathcal{P}(\mathbb{N})$ . If you start listing its elements you quickly discover that  $\mathbb{N}$  has infinitely many subsets, and it's not clear how (or if) they could be arranged in a list with a definite pattern:

$$\begin{aligned}\mathcal{P}(\mathbb{N}) = & \{\emptyset, \{1\}, \{2\}, \dots, \{1,2\}, \{1,3\}, \dots, \{39,47\}, \\ & \dots, \{3, 87, 131\}, \dots, \{2, 4, 6, 8, \dots\}, \dots ? \dots\}.\end{aligned}$$

The set  $\mathcal{P}(\mathbb{R}^2)$  is mind boggling. Think of  $\mathbb{R}^2 = \{(x,y) : x, y \in \mathbb{R}\}$  as the set of all points on the Cartesian plane. A subset of  $\mathbb{R}^2$  (that is, an *element* of  $\mathcal{P}(\mathbb{R}^2)$ ) is a set of points in the plane. Let's look at some of these sets. Since  $\{(1,2), (1,1)\} \subseteq \mathbb{R}^2$ , we know that  $\{(1,2), (1,1)\} \in \mathcal{P}(\mathbb{R}^2)$ . We can even draw a picture of this subset, as in Figure 1.4(a). For another example, the graph of the equation  $y = x^2$  is the set of points  $G = \{(x, x^2) : x \in \mathbb{R}\}$  and this is a subset of  $\mathbb{R}^2$ , so  $G \in \mathcal{P}(\mathbb{R}^2)$ . Figure 1.4(b) is a picture of  $G$ . Because this can be done for any function, the graph of any imaginable function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is an element of  $\mathcal{P}(\mathbb{R}^2)$ .



**Figure 1.4.** Three of the many, many sets in  $\mathcal{P}(\mathbb{R}^2)$

In fact, any black-and-white image on the plane can be thought of as a subset of  $\mathbb{R}^2$ , where the black points belong to the subset and the white points do not. So the text “INFINITE” in Figure 1.4(c) is a subset of  $\mathbb{R}^2$  and therefore an element of  $\mathcal{P}(\mathbb{R}^2)$ . By that token,  $\mathcal{P}(\mathbb{R}^2)$  contains a copy of the page you are reading now.

Thus, in addition to containing every imaginable function and every imaginable black-and-white image,  $\mathcal{P}(\mathbb{R}^2)$  also contains the full text of every book that was ever written, those that are yet to be written and those that will never be written. Inside of  $\mathcal{P}(\mathbb{R}^2)$  is a detailed biography of your life, from beginning to end, as well as the biographies of all of your unborn descendants. It is startling that the five symbols used to write  $\mathcal{P}(\mathbb{R}^2)$  can express such an incomprehensibly large set.

Homework: Think about  $\mathcal{P}(\mathcal{P}(\mathbb{R}^2))$ .

### Exercises for Section 1.4

**A.** Write the following sets by listing their elements between braces.

- |   |   |
|---|---|
| 1. $\mathcal{P}(\{\{a,b\}, \{c\}\})$                | 7. $\mathcal{P}(\{a,b\}) \times \mathcal{P}(\{0,1\})$     |
| 2. $\mathcal{P}(\{1,2,3,4\})$                       | 8. $\mathcal{P}(\{1,2\} \times \{3\})$                    |
| 3. $\mathcal{P}(\{\{\emptyset\}, 5\})$              | 9. $\mathcal{P}(\{a,b\} \times \{0\})$                    |
| 4. $\mathcal{P}(\{\mathbb{R}, \mathbb{Q}\})$        | 10. $\{X \in \mathcal{P}(\{1,2,3\}) :  X  \leq 1\}$       |
| 5. $\mathcal{P}(\mathcal{P}(\{2\}))$                | 11. $\{X \subseteq \mathcal{P}(\{1,2,3\}) :  X  \leq 1\}$ |
| 6. $\mathcal{P}(\{1,2\}) \times \mathcal{P}(\{3\})$ | 12. $\{X \in \mathcal{P}(\{1,2,3\}) : 2 \in X\}$          |

**B.** Suppose that  $|A| = m$  and  $|B| = n$ . Find the following cardinalities.

- |  |   |
|--|---|
| 13. $ \mathcal{P}(\mathcal{P}(\mathcal{P}(A))) $ | 17. $ \{X \in \mathcal{P}(A) :  X  \leq 1\} $                     |
| 14. $ \mathcal{P}(\mathcal{P}(A)) $              | 18. $ \mathcal{P}(A \times \mathcal{P}(B)) $                      |
| 15. $ \mathcal{P}(A \times B) $                  | 19. $ \mathcal{P}(\mathcal{P}(\mathcal{P}(A \times \emptyset))) $ |
| 16. $ \mathcal{P}(A) \times \mathcal{P}(B) $     | 20. $ \{X \subseteq \mathcal{P}(A) :  X  \leq 1\} $               |

## 1.5 Union, Intersection, Difference

Just as numbers are combined with operations such as addition, subtraction and multiplication, there are various operations that can be applied to sets. The Cartesian product (defined in Section 1.2) is one such operation; given sets  $A$  and  $B$ , we can combine them with  $\times$  to get a new set  $A \times B$ . Here are three new operations called union, intersection and difference.

**Definition 1.5** Suppose  $A$  and  $B$  are sets.

The **union** of  $A$  and  $B$  is the set  $A \cup B = \{x : x \in A \text{ or } x \in B\}$ .

The **intersection** of  $A$  and  $B$  is the set  $A \cap B = \{x : x \in A \text{ and } x \in B\}$ .

The **difference** of  $A$  and  $B$  is the set  $A - B = \{x : x \in A \text{ and } x \notin B\}$ .

In words, the union  $A \cup B$  is the set of all things that are in  $A$  or in  $B$  (or in both). The intersection  $A \cap B$  is the set of all things in both  $A$  and  $B$ . The difference  $A - B$  is the set of all things that are in  $A$  but not in  $B$ .

**Example 1.8** Suppose  $A = \{a, b, c, d, e\}$ ,  $B = \{d, e, f\}$  and  $C = \{1, 2, 3\}$ .

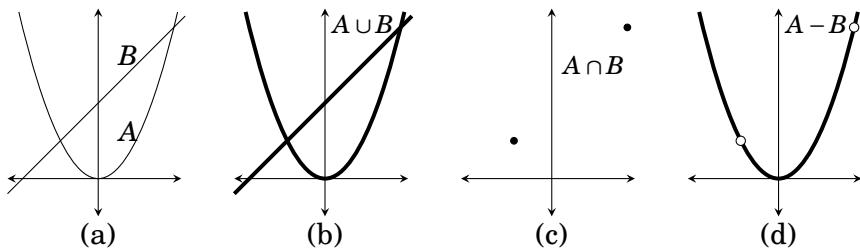
1.  $A \cup B = \{a, b, c, d, e, f\}$
2.  $A \cap B = \{d, e\}$
3.  $A - B = \{a, b, c\}$
4.  $B - A = \{f\}$
5.  $(A - B) \cup (B - A) = \{a, b, c, f\}$
6.  $A \cup C = \{a, b, c, d, e, 1, 2, 3\}$
7.  $A \cap C = \emptyset$
8.  $A - C = \{a, b, c, d, e\}$
9.  $(A \cap C) \cup (A - C) = \{a, b, c, d, e\}$
10.  $(A \cap B) \times B = \{(d, d), (d, e), (d, f), (e, d), (e, e), (e, f)\}$
11.  $(A \times C) \cap (B \times C) = \{(d, 1), (d, 2), (d, 3), (e, 1), (e, 2), (e, 3)\}$

Parts 12–15 use interval notation (Section 1.1), so  $[2, 5] = \{x \in \mathbb{R} : 2 \leq x \leq 5\}$ , etc. Sketching these on the number line may aid your understanding.

12.  $[2, 5] \cup [3, 6] = [2, 6]$
13.  $[2, 5] \cap [3, 6] = [3, 5]$
14.  $[2, 5] - [3, 6] = [2, 3)$
15.  $[0, 3] - [1, 2] = [0, 1) \cup (2, 3]$

Observe that for any sets  $X$  and  $Y$  it is always true that  $X \cup Y = Y \cup X$  and  $X \cap Y = Y \cap X$ , but in general  $X - Y \neq Y - X$ .

**Example 1.9** Let  $A = \{(x, x^2) : x \in \mathbb{R}\}$  be the graph of the equation  $y = x^2$  and let  $B = \{(x, x+2) : x \in \mathbb{R}\}$  be the graph of the equation  $y = x + 2$ . These sets are subsets of  $\mathbb{R}^2$ . They are sketched together in Figure 1.5(a). Figure 1.5(b) shows  $A \cup B$ , the set of all points  $(x, y)$  that are on one (or both) of the two graphs. Observe that  $A \cap B = \{(-1, 1), (2, 4)\}$  consists of just two elements, the two points where the graphs intersect, as illustrated in Figure 1.5(c). Figure 1.5(d) shows  $A - B$ , which is the set  $A$  with “holes” where  $B$  crossed it. In set builder notation, we could write  $A \cup B = \{(x, y) : x \in \mathbb{R}, y = x^2 \text{ or } y = x + 2\}$  and  $A - B = \{(x, x^2) : x \in \mathbb{R} - \{-1, 2\}\}$ .



**Figure 1.5.** The union, intersection and difference of sets  $A$  and  $B$

### Exercises for Section 1.5

1. Suppose  $A = \{4, 3, 6, 7, 1, 9\}$ ,  $B = \{5, 6, 8, 4\}$  and  $C = \{5, 8, 4\}$ . Find:
 

<b>(a)</b> $A \cup B$	<b>(d)</b> $A - C$	<b>(g)</b> $B \cap C$
<b>(b)</b> $A \cap B$	<b>(e)</b> $B - A$	<b>(h)</b> $B \cup C$
<b>(c)</b> $A - B$	<b>(f)</b> $A \cap C$	<b>(i)</b> $C - B$
2. Suppose  $A = \{0, 2, 4, 6, 8\}$ ,  $B = \{1, 3, 5, 7\}$  and  $C = \{2, 8, 4\}$ . Find:
 

<b>(a)</b> $A \cup B$	<b>(d)</b> $A - C$	<b>(g)</b> $B \cap C$
<b>(b)</b> $A \cap B$	<b>(e)</b> $B - A$	<b>(h)</b> $C - A$
<b>(c)</b> $A - B$	<b>(f)</b> $A \cap C$	<b>(i)</b> $C - B$
3. Suppose  $A = \{0, 1\}$  and  $B = \{1, 2\}$ . Find:
 

<b>(a)</b> $(A \times B) \cap (B \times B)$	<b>(d)</b> $(A \cap B) \times A$	<b>(g)</b> $\mathcal{P}(A) - \mathcal{P}(B)$
<b>(b)</b> $(A \times B) \cup (B \times B)$	<b>(e)</b> $(A \times B) \cap B$	<b>(h)</b> $\mathcal{P}(A \cap B)$
<b>(c)</b> $(A \times B) - (B \times B)$	<b>(f)</b> $\mathcal{P}(A) \cap \mathcal{P}(B)$	<b>(i)</b> $\mathcal{P}(A \times B)$
4. Suppose  $A = \{b, c, d\}$  and  $B = \{a, b\}$ . Find:
 

<b>(a)</b> $(A \times B) \cap (B \times B)$	<b>(d)</b> $(A \cap B) \times A$	<b>(g)</b> $\mathcal{P}(A) - \mathcal{P}(B)$
<b>(b)</b> $(A \times B) \cup (B \times B)$	<b>(e)</b> $(A \times B) \cap B$	<b>(h)</b> $\mathcal{P}(A \cap B)$
<b>(c)</b> $(A \times B) - (B \times B)$	<b>(f)</b> $\mathcal{P}(A) \cap \mathcal{P}(B)$	<b>(i)</b> $\mathcal{P}(A) \times \mathcal{P}(B)$

5. Sketch the sets  $X = [1, 3] \times [1, 3]$  and  $Y = [2, 4] \times [2, 4]$  on the plane  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $X \cup Y$ ,  $X \cap Y$ ,  $X - Y$  and  $Y - X$ . (Hint:  $X$  and  $Y$  are Cartesian products of intervals. You may wish to review how you drew sets like  $[1, 3] \times [1, 3]$  in the exercises for Section 1.2.)
  6. Sketch the sets  $X = [-1, 3] \times [0, 2]$  and  $Y = [0, 3] \times [1, 4]$  on the plane  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $X \cup Y$ ,  $X \cap Y$ ,  $X - Y$  and  $Y - X$ .
  7. Sketch the sets  $X = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$  and  $Y = \{(x, y) \in \mathbb{R}^2 : x \geq 0\}$  on  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $X \cup Y$ ,  $X \cap Y$ ,  $X - Y$  and  $Y - X$ .
  8. Sketch the sets  $X = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$  and  $Y = \{(x, y) \in \mathbb{R}^2 : -1 \leq y \leq 0\}$  on  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $X \cup Y$ ,  $X \cap Y$ ,  $X - Y$  and  $Y - X$ .
  9. Is the statement  $(\mathbb{R} \times \mathbb{Z}) \cap (\mathbb{Z} \times \mathbb{R}) = \mathbb{Z} \times \mathbb{Z}$  true or false? What about the statement  $(\mathbb{R} \times \mathbb{Z}) \cup (\mathbb{Z} \times \mathbb{R}) = \mathbb{R} \times \mathbb{R}$ ?
  10. Do you think the statement  $(\mathbb{R} - \mathbb{Z}) \times \mathbb{N} = (\mathbb{R} \times \mathbb{N}) - (\mathbb{Z} \times \mathbb{N})$  is true, or false? Justify.
- 

## 1.6 Complement

This section introduces yet another set operation, called the *set complement*. The definition requires the idea of a *universal set*, which we now discuss.

When dealing with a set, we almost always regard it as a subset of some larger set. For example, consider the set of prime numbers  $P = \{2, 3, 5, 7, 11, 13, \dots\}$ . If asked to name some things that are *not* in  $P$ , we might mention some composite numbers like 4 or 6 or 423. It probably would not occur to us to say that Vladimir Putin is not in  $P$ . True, Vladimir Putin is not in  $P$ , but he lies entirely outside of the discussion of what is a prime number and what is not. We have an unstated assumption that

$$P \subseteq \mathbb{N}$$

because  $\mathbb{N}$  is the most natural setting in which to discuss prime numbers. In this context, anything not in  $P$  should still be in  $\mathbb{N}$ . This larger set  $\mathbb{N}$  is called the **universal set** or **universe** for  $P$ .

Almost every useful set in mathematics can be regarded as having some natural universal set. For instance, the unit circle is the set  $C = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ , and since all these points are in the plane  $\mathbb{R}^2$  it is natural to regard  $\mathbb{R}^2$  as the universal set for  $C$ . In the absence of specifics, if  $A$  is a set, then its universal set is often denoted as  $U$ . We are now ready to define the complement operation.

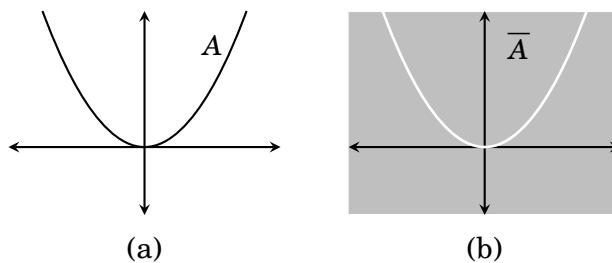
**Definition 1.6** Let  $A$  be a set with a universal set  $U$ . The **complement** of  $A$ , denoted  $\overline{A}$ , is the set  $\overline{A} = U - A$ .

**Example 1.10** If  $P$  is the set of prime numbers, then

$$\overline{P} = \mathbb{N} - P = \{1, 4, 6, 8, 9, 10, 12, \dots\}.$$

Thus  $\overline{P}$  is the set of composite numbers and 1.

**Example 1.11** Let  $A = \{(x, x^2) : x \in \mathbb{R}\}$  be the graph of the equation  $y = x^2$ . Figure 1.6(a) shows  $A$  in its universal set  $\mathbb{R}^2$ . The complement of  $A$  is  $\overline{A} = \mathbb{R}^2 - A = \{(x, y) \in \mathbb{R}^2 : y \neq x^2\}$ , illustrated by the shaded area in Figure 1.6(b).



**Figure 1.6.** A set and its complement

### Exercises for Section 1.6

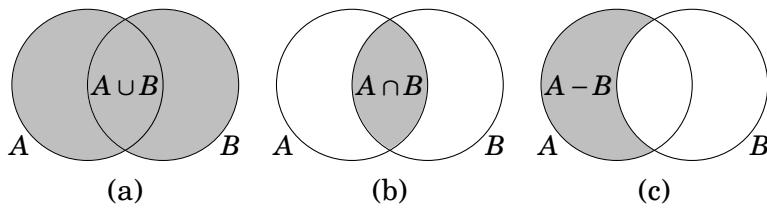
1. Let  $A = \{4, 3, 6, 7, 1, 9\}$  and  $B = \{5, 6, 8, 4\}$  have universal set  $U = \{0, 1, 2, \dots, 10\}$ . Find:
 

<b>(a)</b> $\overline{A}$	<b>(d)</b> $A \cup \overline{A}$	<b>(g)</b> $\overline{A} - \overline{B}$
<b>(b)</b> $\overline{B}$	<b>(e)</b> $A - \overline{A}$	<b>(h)</b> $\overline{A} \cap B$
<b>(c)</b> $A \cap \overline{A}$	<b>(f)</b> $A - \overline{B}$	<b>(i)</b> $\overline{A} \cap B$
2. Let  $A = \{0, 2, 4, 6, 8\}$  and  $B = \{1, 3, 5, 7\}$  have universal set  $U = \{0, 1, 2, \dots, 8\}$ . Find:
 

<b>(a)</b> $\overline{A}$	<b>(d)</b> $A \cup \overline{A}$	<b>(g)</b> $\overline{A} \cap \overline{B}$
<b>(b)</b> $\overline{B}$	<b>(e)</b> $A - \overline{A}$	<b>(h)</b> $\overline{A} \cap B$
<b>(c)</b> $A \cap \overline{A}$	<b>(f)</b> $\overline{A} \cup B$	<b>(i)</b> $\overline{A} \times B$
3. Sketch the set  $X = [1, 3] \times [1, 2]$  on the plane  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $\overline{X}$  and  $\overline{X} \cap ([0, 2] \times [0, 3])$ .
4. Sketch the set  $X = [-1, 3] \times [0, 2]$  on the plane  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $\overline{X}$  and  $\overline{X} \cap ([0, 2] \times [-1, 3])$ .
5. Sketch the set  $X = \{(x, y) \in \mathbb{R}^2 : 1 \leq x^2 + y^2 \leq 4\}$  on the plane  $\mathbb{R}^2$ . On a separate drawing, shade in the set  $\overline{X}$ .
6. Sketch the set  $X = \{(x, y) \in \mathbb{R}^2 : y < x^2\}$  on  $\mathbb{R}^2$ . Shade in the set  $\overline{X}$ .

## 1.7 Venn Diagrams

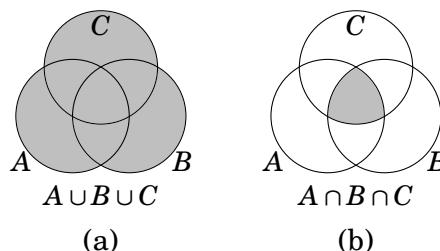
In thinking about sets, it is sometimes helpful to draw informal, schematic diagrams of them. In doing this we often represent a set with a circle (or oval), which we regard as enclosing all the elements of the set. Such diagrams can illustrate how sets combine using various operations. For example, Figures 1.7(a–c) show two sets  $A$  and  $B$  that overlap in a middle region. The sets  $A \cup B$ ,  $A \cap B$  and  $A - B$  are shaded. Such graphical representations of sets are called **Venn diagrams**, after their inventor, British logician John Venn, 1834–1923.



**Figure 1.7.** Venn diagrams for two sets

Though you are unlikely to draw Venn diagrams as a part of a proof of any theorem, you will probably find them to be useful “scratch work” devices that help you to understand how sets combine, and to develop strategies for proving certain theorems or solving certain problems. The remainder of this section uses Venn diagrams to explore how three sets can be combined using  $\cup$  and  $\cap$ .

Let’s begin with the set  $A \cup B \cup C$ . Our definitions suggest this should consist of all elements which are in one or more of the sets  $A$ ,  $B$  and  $C$ . Figure 1.8(a) shows a Venn diagram for this. Similarly, we think of  $A \cap B \cap C$  as all elements common to each of  $A$ ,  $B$  and  $C$ , so in Figure 1.8(b) the region belonging to all three sets is shaded.

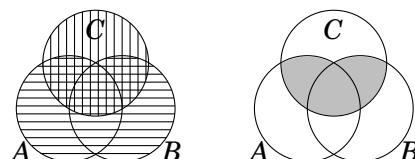


**Figure 1.8.** Venn diagrams for three sets

We can also think of  $A \cap B \cap C$  as the two-step operation  $(A \cap B) \cap C$ . In this expression the set  $A \cap B$  is represented by the region common to both  $A$  and  $B$ , and when we intersect *this* with  $C$  we get Figure 1.8(b). This is a visual representation of the fact that  $A \cap B \cap C = (A \cap B) \cap C$ . Similarly, we have  $A \cap B \cap C = A \cap (B \cap C)$ . Likewise,  $A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$ .

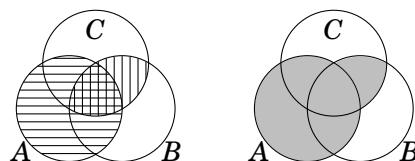
Notice that in these examples, where the expression either contains only the symbol  $\cup$  or only the symbol  $\cap$ , the placement of the parentheses is irrelevant, so we are free to drop them. It is analogous to the situations in algebra involving expressions  $(a+b)+c = a+(b+c)$  or  $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ . We tend to drop the parentheses and write simply  $a + b + c$  or  $a \cdot b \cdot c$ . By contrast, in an expression like  $(a+b) \cdot c$  the parentheses are absolutely essential because  $(a+b) \cdot c$  and  $a+(b \cdot c)$  are generally not equal.

Now let's use Venn diagrams to help us understand the expressions  $(A \cup B) \cap C$  and  $A \cup (B \cap C)$ , which use a mix of  $\cup$  and  $\cap$ . Figure 1.9 shows how to draw a Venn diagram for  $(A \cup B) \cap C$ . In the drawing on the left, the set  $A \cup B$  is shaded with horizontal lines, while  $C$  is shaded with vertical lines. Thus the set  $(A \cup B) \cap C$  is represented by the cross-hatched region where  $A \cup B$  and  $C$  overlap. The superfluous shadings are omitted in the drawing on the right showing the set  $(A \cup B) \cap C$ .



**Figure 1.9.** How to make a Venn diagram for  $(A \cup B) \cap C$

Now think about  $A \cup (B \cap C)$ . In Figure 1.10 the set  $A$  is shaded with horizontal lines, and  $B \cap C$  is shaded with vertical lines. The union  $A \cup (B \cap C)$  is represented by the totality of all shaded regions, as shown on the right.



**Figure 1.10.** How to make a Venn diagram for  $A \cup (B \cap C)$

Compare the diagrams for  $(A \cup B) \cap C$  and  $A \cup (B \cap C)$  in Figures 1.9 and 1.10. The fact that the diagrams are different indicates that  $(A \cup B) \cap C \neq A \cup (B \cap C)$  in general. Thus an expression such as  $A \cup B \cap C$  is absolutely meaningless because we can't tell whether it means  $(A \cup B) \cap C$  or  $A \cup (B \cap C)$ . In summary, Venn diagrams have helped us understand the following.

### Important Points:

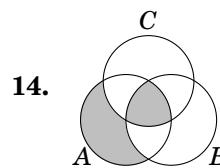
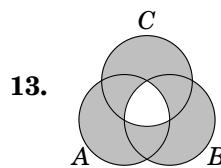
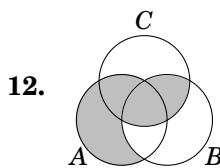
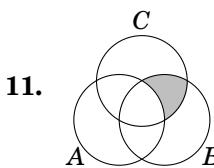
- If an expression involving sets uses only  $\cup$ , then parentheses are optional.
- If an expression involving sets uses only  $\cap$ , then parentheses are optional.
- If an expression uses both  $\cup$  and  $\cap$ , then parentheses are **essential**.

In the next section we will study types of expressions that use only  $\cup$  or only  $\cap$ . These expressions will not require the use of parentheses.

### Exercises for Section 1.7

1. Draw a Venn diagram for  $\overline{A}$ , where  $A$  is a subset of a universal set  $U$ .
2. Draw a Venn diagram for  $B - A$ .
3. Draw a Venn diagram for  $(A - B) \cap C$ .
4. Draw a Venn diagram for  $(A \cup B) - C$ .
5. Draw Venn diagrams for  $A \cup (B \cap C)$  and  $(A \cup B) \cap (A \cup C)$ . Based on your drawings, do you think  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ ?
6. Draw Venn diagrams for  $A \cap (B \cup C)$  and  $(A \cap B) \cup (A \cap C)$ . Based on your drawings, do you think  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ ?
7. Suppose sets  $A$  and  $B$  are in a universal set  $U$ . Draw Venn diagrams for  $\overline{A \cap B}$  and  $\overline{A} \cap \overline{B}$ . Based on your drawings, do you think it's true that  $\overline{A \cap B} = \overline{A} \cup \overline{B}$ ?
8. Suppose sets  $A$  and  $B$  are in a universal set  $U$ . Draw Venn diagrams for  $\overline{A \cup B}$  and  $\overline{A} \cap \overline{B}$ . Based on your drawings, do you think it's true that  $\overline{A \cup B} = \overline{A} \cap \overline{B}$ ?
9. Draw a Venn diagram for  $(A \cap B) - C$ .
10. Draw a Venn diagram for  $(A - B) \cup C$ .

Following are Venn diagrams for expressions involving sets  $A$ ,  $B$  and  $C$ . Write a corresponding expression.



## 1.8 Indexed Sets

When a mathematical problem involves lots of sets, it is often convenient to keep track of them by using subscripts (also called indices). Thus instead of denoting three sets as  $A$ ,  $B$  and  $C$ , we might instead write them as  $A_1$ ,  $A_2$  and  $A_3$ . These are called **indexed sets**.

Although we defined union and intersection to be operations that combine two sets, you by now have no difficulty forming unions and intersections of three or more sets. (For instance, in the previous section we drew Venn diagrams for the intersection and union of three sets.) But let's take a moment to write down careful definitions. Given sets  $A_1, A_2, \dots, A_n$ , the set  $A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n$  consists of everything that is in *at least one* of the sets  $A_i$ . Likewise  $A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n$  consists of everything that is common to *all* of the sets  $A_i$ . Here is a careful definition.

**Definition 1.7** Suppose  $A_1, A_2, \dots, A_n$  are sets. Then

$$\begin{aligned} A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n &= \{x : x \in A_i \text{ for } \textit{at least one} \text{ set } A_i, \text{ for } 1 \leq i \leq n\}, \\ A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n &= \{x : x \in A_i \text{ for } \textit{every} \text{ set } A_i, \text{ for } 1 \leq i \leq n\}. \end{aligned}$$

But if the number  $n$  of sets is large, these expressions can get messy. To overcome this, we now develop some notation akin to sigma notation. You already know that sigma notation is a convenient symbolism for expressing sums of many numbers. Given numbers  $a_1, a_2, a_3, \dots, a_n$ , then

$$\sum_{i=1}^n a_i = a_1 + a_2 + a_3 + \dots + a_n.$$

Even if the list of numbers is infinite, the sum

$$\sum_{i=1}^{\infty} a_i = a_1 + a_2 + a_3 + \dots + a_i + \dots$$

is often still meaningful. The notation we are about to introduce is very similar to this. Given sets  $A_1, A_2, A_3, \dots, A_n$ , we define

$$\bigcup_{i=1}^n A_i = A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n \quad \text{and} \quad \bigcap_{i=1}^n A_i = A_1 \cap A_2 \cap A_3 \cap \dots \cap A_n.$$

**Example 1.12** Suppose  $A_1 = \{0, 2, 5\}$ ,  $A_2 = \{1, 2, 5\}$  and  $A_3 = \{2, 5, 7\}$ . Then

$$\bigcup_{i=1}^3 A_i = A_1 \cup A_2 \cup A_3 = \{0, 1, 2, 5, 7\} \quad \text{and} \quad \bigcap_{i=1}^3 A_i = A_1 \cap A_2 \cap A_3 = \{2, 5\}.$$

This notation is also used when the list of sets  $A_1, A_2, A_3, A_4, \dots$  is infinite:

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup A_3 \cup \dots = \{x : x \in A_i \text{ for at least one set } A_i \text{ with } 1 \leq i\}.$$

$$\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap A_3 \cap \dots = \{x : x \in A_i \text{ for every set } A_i \text{ with } 1 \leq i\}.$$

**Example 1.13** This example involves the following infinite list of sets.

$$A_1 = \{-1, 0, 1\}, \quad A_2 = \{-2, 0, 2\}, \quad A_3 = \{-3, 0, 3\}, \quad \dots, \quad A_i = \{-i, 0, i\}, \quad \dots$$

Observe that  $\bigcup_{i=1}^{\infty} A_i = \mathbb{Z}$ , and  $\bigcap_{i=1}^{\infty} A_i = \{0\}$ .

Here is a useful twist on our new notation. We can write

$$\bigcup_{i=1}^3 A_i = \bigcup_{i \in \{1, 2, 3\}} A_i,$$

which is understood to be the union of the sets  $A_i$  for  $i = 1, 2, 3$ . Likewise:

$$\bigcap_{i=1}^3 A_i = \bigcap_{i \in \{1, 2, 3\}} A_i$$

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i \in \mathbb{N}} A_i$$

$$\bigcap_{i=1}^{\infty} A_i = \bigcap_{i \in \mathbb{N}} A_i$$

Here we are taking the union or intersection of a collection of sets  $A_i$  where  $i$  is an element of some set, be it  $\{1, 2, 3\}$  or  $\mathbb{N}$ . In general, the way this works is that we will have a collection of sets  $A_i$  for  $i \in I$ , where  $I$  is the set of possible subscripts. The set  $I$  is called an **index set**.

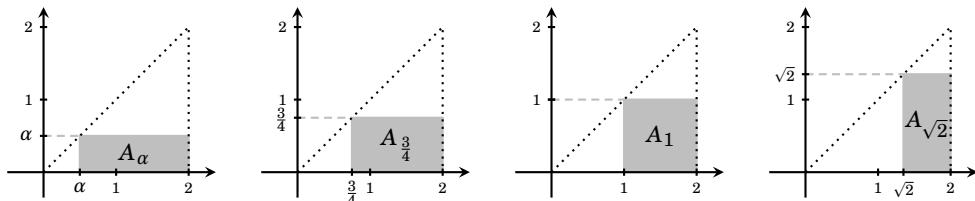
It is important to realize that the set  $I$  need not even consist of integers. (We could subscript with letters or real numbers, etc.) Since we are programmed to think of  $i$  as an integer, let's make a slight notational change: Use  $\alpha$ , not  $i$ , to stand for an element of  $I$ . Thus we are dealing with a collection of sets  $A_\alpha$  for  $\alpha \in I$ . This leads to the following definition.

**Definition 1.8** If  $A_\alpha$  is a set for every  $\alpha$  in some index set  $I \neq \emptyset$ , then

$$\bigcup_{\alpha \in I} A_\alpha = \{x : x \in A_\alpha \text{ for at least one set } A_\alpha \text{ with } \alpha \in I\}$$

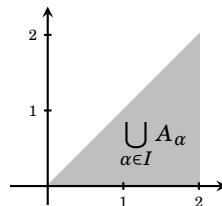
$$\bigcap_{\alpha \in I} A_\alpha = \{x : x \in A_\alpha \text{ for every set } A_\alpha \text{ with } \alpha \in I\}.$$

**Example 1.14** In this example, all sets  $A_\alpha$  are all subsets of the plane  $\mathbb{R}^2$ . Each  $\alpha$  belongs to the index set  $I = [0, 2] = \{x \in \mathbb{R} : 0 \leq x \leq 2\}$ , which is the set of all real numbers between 0 and 2. For each number  $\alpha \in I$ , define  $A_\alpha$  to be the set  $A_\alpha = [\alpha, 2] \times [0, \alpha]$ , which is the rectangle on the  $xy$ -plane whose base runs from  $\alpha$  to 2 on the  $x$ -axis, and whose height is  $\alpha$ . Some of these are shown shaded below. (The dotted diagonal line  $y = x$  is not a part of any of the sets, but is shown for clarity, as the upper left corner of each  $A_\alpha$  touches it.) Note that these sets are not indexed with just integers. For example, as  $\sqrt{2} \in I$ , there is a set  $A_{\sqrt{2}}$ , which is shown below on the right.



Note that  $A_0 = [0, 2] \times [0, 0] = [0, 2] \times \{0\}$  is the interval  $[0, 2]$  on the  $x$ -axis (a “flat” rectangle). Also,  $A_2 = [2, 2] \times [0, 2] = \{2\} \times [0, 2]$  is the vertical side of the dotted triangle in the above pictures.

Now consider the infinite union  $\bigcup_{\alpha \in I} A_\alpha$ . It is the shaded triangle shown below, because any point  $(x, y)$  on this triangle belongs to the set  $A_x$ , and is therefore in the union. (And any point not on the triangle is not in any  $A_x$ .)



Now let's work out the intersection  $\bigcap_{\alpha \in I} A_\alpha$ . Notice that the point  $(2, 0)$  on the  $x$ -axis is the lower right corner of any set  $A_\alpha$ , so  $(2, 0) \in A_\alpha$  for any  $\alpha \in I$ . Therefore the point  $(2, 0)$  is in the intersection of all the  $A_\alpha$ . But any other point  $(x, y) \neq (2, 0)$  on the triangle does not belong to all of the sets  $A_\alpha$ . The reason is that if  $x < 2$ , then  $(x, y) \notin A_\alpha$  for any  $x < \alpha \leq 2$ . (Check this.) And if  $x = 2$ , then  $(x, y) \notin A_\alpha$  for any  $0 < \alpha \leq y$ . Consequently

$$\bigcap_{\alpha \in I} A_\alpha = \{(2, 0)\}.$$

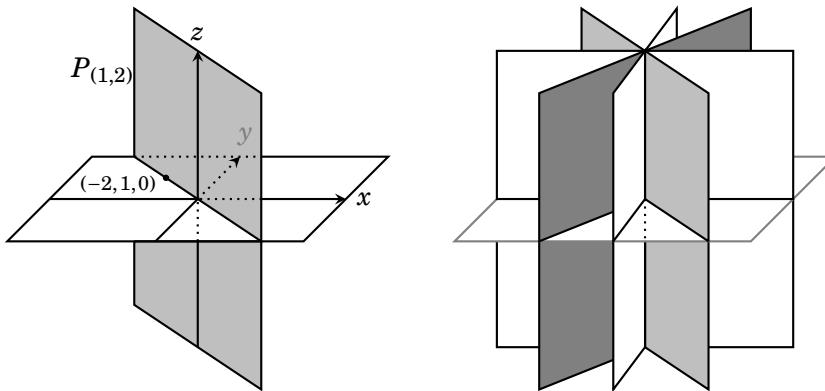
This intersection consists of only one element, the point  $(2, 0)$ .

**Example 1.15** Here our sets are indexed by  $\mathbb{R}^2$ . For any  $(a, b) \in \mathbb{R}^2$ , let  $P_{(a,b)}$  be the following subset of  $\mathbb{R}^3$ :

$$P_{(a,b)} = \{(x, y, z) \in \mathbb{R}^3 : ax + by = 0\}.$$

In words, given a point  $(a, b) \in \mathbb{R}^2$ , the corresponding set  $P_{(a,b)}$  consists of all points  $(x, y, z)$  in  $\mathbb{R}^3$  that satisfy the equation  $ax + by = 0$ . From previous math courses you will recognize this as a plane in  $\mathbb{R}^3$ , that is,  $P_{(a,b)}$  is a plane in  $\mathbb{R}^3$ . Moreover, since any point  $(0, 0, z)$  on the  $z$ -axis automatically satisfies  $ax + by = 0$ , each  $P_{(a,b)}$  contains the  $z$ -axis.

Figure 1.11 (left) shows the set  $P_{(1,2)} = \{(x, y, z) \in \mathbb{R}^3 : x + 2y = 0\}$ . It is the vertical plane that intersects the  $xy$ -plane at the line  $x + 2y = 0$ .



**Figure 1.11.** The sets  $P_{(a,b)}$  are vertical planes containing the  $z$ -axis.

For any point  $(a, b) \in \mathbb{R}^2$  with  $(a, b) \neq (0, 0)$ , we can visualize  $P_{(a,b)}$  as the vertical plane that cuts the  $xy$ -plane at the line  $ax + by = 0$ . Figure 1.11 (right) shows a few of the  $P_{(a,b)}$ . Since any two such planes intersect along the  $z$ -axis, and because the  $z$ -axis is a subset of every  $P_{(a,b)}$ , it is immediately clear that

$$\bigcap_{(a,b) \in \mathbb{R}^2} P_{(a,b)} = \{(0, 0, z) : z \in \mathbb{R}\} = \text{"the } z\text{-axis".}$$

For the union, note that any given point  $(a, b, c) \in \mathbb{R}^3$  belongs to the set  $P_{(-b,a)}$  because  $(x, y, z) = (a, b, c)$  satisfies the equation  $-bx + ay = 0$ . (In fact, any  $(a, b, c)$  belongs to the special set  $P_{(0,0)} = \mathbb{R}^3$ , which is the only  $P_{(a,b)}$  that is not a plane.) Since any point in  $\mathbb{R}^3$  belongs to some  $P_{(a,b)}$  we have

$$\bigcup_{(a,b) \in \mathbb{R}^2} P_{(a,b)} = \mathbb{R}^3.$$

## Exercises for Section 1.8

- Suppose  $A_1 = \{a, b, d, e, g, f\}$ ,  $A_2 = \{a, b, c, d\}$ ,  $A_3 = \{b, d, a\}$  and  $A_4 = \{a, b, h\}$ .
    - $\bigcup_{i=1}^4 A_i =$
    - $\bigcap_{i=1}^4 A_i =$
  - Suppose  $\begin{cases} A_1 &= \{0, 2, 4, 8, 10, 12, 14, 16, 18, 20, 22, 24\}, \\ A_2 &= \{0, 3, 6, 9, 12, 15, 18, 21, 24\}, \\ A_3 &= \{0, 4, 8, 12, 16, 20, 24\}. \end{cases}$ 
    - $\bigcup_{i=1}^3 A_i =$
    - $\bigcap_{i=1}^3 A_i =$
  - For each  $n \in \mathbb{N}$ , let  $A_n = \{0, 1, 2, 3, \dots, n\}$ .
    - $\bigcup_{i \in \mathbb{N}} A_i =$
    - $\bigcap_{i \in \mathbb{N}} A_i =$
  - For each  $n \in \mathbb{N}$ , let  $A_n = \{-2n, 0, 2n\}$ .
    - $\bigcup_{i \in \mathbb{N}} A_i =$
    - $\bigcap_{i \in \mathbb{N}} A_i =$
  - (a)  $\bigcup_{i \in \mathbb{N}} [i, i+1] =$  (b)  $\bigcap_{i \in \mathbb{N}} [i, i+1] =$
  - (a)  $\bigcup_{i \in \mathbb{N}} [0, i+1] =$  (b)  $\bigcap_{i \in \mathbb{N}} [0, i+1] =$
  - (a)  $\bigcup_{i \in \mathbb{N}} \mathbb{R} \times [i, i+1] =$  (b)  $\bigcap_{i \in \mathbb{N}} \mathbb{R} \times [i, i+1] =$
  - (a)  $\bigcup_{\alpha \in \mathbb{R}} \{\alpha\} \times [0, 1] =$  (b)  $\bigcap_{\alpha \in \mathbb{R}} \{\alpha\} \times [0, 1] =$
  - (a)  $\bigcup_{X \in \mathcal{P}(\mathbb{N})} X =$  (b)  $\bigcap_{X \in \mathcal{P}(\mathbb{N})} X =$
  - (a)  $\bigcup_{x \in [0, 1]} [x, 1] \times [0, x^2] =$  (b)  $\bigcap_{x \in [0, 1]} [x, 1] \times [0, x^2] =$
  - Is  $\bigcap_{\alpha \in I} A_\alpha \subseteq \bigcup_{\alpha \in I} A_\alpha$  always true for any collection of sets  $A_\alpha$  with index set  $I$ ?
  - If  $\bigcap_{\alpha \in I} A_\alpha = \bigcup_{\alpha \in I} A_\alpha$ , what do you think can be said about the relationships between the sets  $A_\alpha$ ?
  - If  $J \neq \emptyset$  and  $J \subseteq I$ , does it follow that  $\bigcup_{\alpha \in J} A_\alpha \subseteq \bigcup_{\alpha \in I} A_\alpha$ ? What about  $\bigcap_{\alpha \in J} A_\alpha \subseteq \bigcap_{\alpha \in I} A_\alpha$ ?
  - If  $J \neq \emptyset$  and  $J \subseteq I$ , does it follow that  $\bigcap_{\alpha \in I} A_\alpha \subseteq \bigcap_{\alpha \in J} A_\alpha$ ? Explain.

## 1.9 Sets That Are Number Systems

In practice, the sets we tend to be most interested in often have special properties and structures. For example, the sets  $\mathbb{Z}$ ,  $\mathbb{Q}$  and  $\mathbb{R}$  are familiar number systems: Given such a set, any two of its elements can be added (or multiplied, etc.) together to produce another element in the set. These operations obey the familiar commutative, associative and distributive properties that we all have dealt with for years. Such properties lead to the standard algebraic techniques for solving equations. Even though we are concerned with the idea of proof, we will not find it necessary to define, prove or verify such properties and techniques; we will accept them as the ground rules upon which our further deductions are based.

We also accept as fact the natural ordering of the elements of  $\mathbb{N}$ ,  $\mathbb{Z}$ ,  $\mathbb{Q}$  and  $\mathbb{R}$ , so that (for example) the meaning of “ $5 < 7$ ” is understood and does not need to be justified or explained. Similarly, if  $x \leq y$  and  $a \neq 0$ , we know that  $ax \leq ay$  or  $ax \geq ay$ , depending on whether  $a$  is positive or negative.

Another thing that our ingrained understanding of number order tells us is that any non-empty subset of  $\mathbb{N}$  has a smallest element. In other words, if  $A \subseteq \mathbb{N}$  and  $A \neq \emptyset$ , then there is an element  $x_0 \in A$  that is smaller than every other element of  $A$ . (To find it, start at 1, then move in increments to 2, 3, 4, etc., until you hit a number  $x_0 \in A$ ; this is the smallest element of  $A$ .) Similarly, given  $b \in \mathbb{Z}$ , any non-empty subset  $A \subseteq \{b, b+1, b+2, b+3, \dots\}$  has a smallest element. This fact is sometimes called the **well-ordering principle**. There is no need to remember this term, but do be aware that we will use this simple, intuitive idea often in proofs, usually without a second thought.

The well-ordering principle seems innocent enough, but it actually says something very fundamental and special about the positive integers  $\mathbb{N}$ . In fact, the corresponding statement for the positive real numbers is false: The subset  $A = \{\frac{1}{n} : n \in \mathbb{N}\}$  of the positive reals has no smallest element because for any  $x_0 = \frac{1}{n} \in A$  we might pick, there is a smaller element  $\frac{1}{n+1} \in A$ .

One consequence of the well-ordering principle (as we will see below) is the familiar fact that any integer  $a$  can be divided by a non-zero integer  $b$ , resulting in a quotient  $q$  and remainder  $r$ . For example,  $b = 3$  goes into  $a = 17$   $q = 5$  times with remainder  $r = 2$ . In symbols,  $17 = 5 \cdot 3 + 2$ , or  $a = qb + r$ . This significant fact is called the **division algorithm**.

**Fact 1.5 (Division Algorithm)** Given integers  $a$  and  $b$  with  $b > 0$ , there exist unique integers  $q$  and  $r$  for which  $a = qb + r$  and  $0 \leq r < b$ .

Although there is no harm in accepting the division algorithm without proof, note that it does follow from the well-ordering principle. Here's how: Given integers  $a, b$  with  $b > 0$ , form the set

$$A = \{a - xb : x \in \mathbb{Z}, 0 \leq a - xb \leq a\} \subseteq \{0, 1, 2, 3, \dots\}.$$

This is the set of non-negative integers got by subtracting multiples of  $b$  from  $a$ . (For example, if  $a = 17$  and  $b = 3$ , then we get  $A = \{2, 5, 8, 11, 14, 17\}$  by subtracting multiples of 3 from 17. Note that the remainder  $r = 2$  of  $17 \div 3$  is the smallest element of  $A$ .) In general, the well-ordering principle says the set  $A = \{a - xb : x \in \mathbb{Z}, 0 \leq a - xb \leq a\}$  has a smallest element  $r$ . Then  $r = a - qb$  for some  $x = q \in \mathbb{Z}$ , so  $a = qb + r$ . Because  $r \in A \subseteq \{0, 1, 2, 3, \dots\}$ , we know  $0 \leq r$ . In addition, it cannot happen that  $r \geq b$ : If this were the case, then the non-negative number  $r - b = (a - qb) - b = a - (q+1)b$  having form  $a - xb$  would be a smaller element of  $A$  than  $r$ , and  $r$  was explicitly chosen as the smallest element of  $A$ . Since it is not true that  $r \geq b$ , it must be that  $r < b$ . Therefore  $0 \leq r < b$ . We've now produced a  $q$  and an  $r$  for which  $a = qb + r$  and  $0 \leq r < b$ . (Exercise 28 of Chapter 7 asks you to prove  $q$  and  $r$  are *unique* in the sense that no other values of  $q$  and  $r$  have these properties.)

Moving on, it is time to clarify a small issue. This chapter asserted that all of mathematics can be described with sets. But at the same time we maintained that some mathematical entities are not sets. (For instance, our approach was to say that an individual number, such as 5, is not itself a set, though it may be an *element* of a set.) We have made this distinction because we need a place to stand as we explore sets: After all, it would appear suspiciously circular to declare that every mathematical entity is a set, and then go on to define a set as a collection whose members are sets!

But to most mathematicians, saying "The number 5 is not a set," is like saying "The number 5 is not a number."

The truth is that any number *can* itself be understood as a set. One way to do this is to begin with the identification  $0 = \emptyset$ . Then  $1 = \{\emptyset\} = \{0\}$ , and  $2 = \{\emptyset, \{\emptyset\}\} = \{0, 1\}$ , and  $3 = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = \{0, 1, 2\}$ . In general the natural number  $n$  is the set  $n = \{0, 1, 2, \dots, n-1\}$  of the  $n$  numbers (which are themselves sets) that come before it.

We will not undertake such a study here, but the elements of the number systems  $\mathbb{Z}$ ,  $\mathbb{Q}$  and  $\mathbb{R}$  can all be defined in terms of sets. (Even the operations of addition, multiplication, etc., can be defined in set-theoretic terms.) In fact, mathematics itself can be regarded as the study of things that can be described as sets. Any mathematical entity is a set, whether or not we choose to think of it that way.

### 1.10 Russell's Paradox

This section contains some background information that may be interesting, but is not used in the remainder of the book.

The philosopher and mathematician Bertrand Russell (1872–1970) did groundbreaking work on the theory of sets and the foundations of mathematics. He was probably among the first to understand how the misuse of sets can lead to bizarre and paradoxical situations. He is famous for an idea that has come to be known as **Russell's paradox**.

Russell's paradox involves the following set of sets:

$$A = \{X : X \text{ is a set and } X \notin X\}. \quad (1.1)$$

In words,  $A$  is the set of all sets that do not include themselves as elements. Most sets we can think of are in  $A$ . The set  $\mathbb{Z}$  of integers is not an integer (i.e.,  $\mathbb{Z} \notin \mathbb{Z}$ ) and therefore  $\mathbb{Z} \in A$ . Also  $\emptyset \in A$  because  $\emptyset$  is a set and  $\emptyset \notin \emptyset$ .

Is there a set that is not in  $A$ ? Consider  $B = \{\underbrace{\{\{\dots\}\}}_B\}$ . Think of  $B$  as a box containing a box, containing a box, containing a box, and so on, forever. Or a set of identical Russian dolls, nested one inside the other, endlessly. The curious thing about  $B$  is that it has just one element, namely  $B$  itself:

$$B = \{\underbrace{\{\{\dots\}\}}_B\}.$$

Thus  $B \in B$ . As  $B$  does not satisfy  $B \notin B$ , Equation (1.1) says  $B \notin A$ .

Russell's paradox arises from the question “*Is  $A$  an element of  $A$ ?*”

For a set  $X$ , Equation (1.1) says  $X \in A$  means the same thing as  $X \notin X$ . So for  $X = A$ , the previous line says  $A \in A$  means the same thing as  $A \notin A$ . Conclusions: If  $A \in A$  is true, then it is false. If  $A \in A$  is false, then it is true. This is Russell's paradox.

Initially Russell's paradox sparked a crisis among mathematicians. How could a mathematical statement be both true and false? This seemed to be in opposition to the very essence of mathematics.

The paradox instigated a very careful examination of set theory and an evaluation of what can and cannot be regarded as a set. Eventually mathematicians settled upon a collection of axioms for set theory—the so-called **Zermelo-Fraenkel axioms**. One of these axioms is the well-ordering principle of the previous section. Another, the axiom of foundation, states that no non-empty set  $X$  is allowed to have the property  $X \cap x \neq \emptyset$  for all its elements  $x$ . This rules out such circularly defined “sets” as  $B = \{B\}$  mentioned above. If we adhere to these axioms, then situations like Russell's

paradox disappear. Most mathematicians accept all this on faith and happily ignore the Zermelo-Fraenkel axioms. Paradoxes like Russell's do not tend to come up in everyday mathematics—you have to go out of your way to construct them.

Still, Russell's paradox reminds us that precision of thought and language is an important part of doing mathematics. The next chapter deals with the topic of logic, a codification of thought and language.

**Additional Reading on Sets.** For a lively account of Bertrand Russell's life and work (including his paradox), see the graphic novel *Logicomix: An Epic Search For Truth*, by Apostolos Doxiadis and Christos Papadimitriou. Also see cartoonist Jessica Hagy's online strip *Indexed*—it is based largely on Venn diagrams.

# CHAPTER 2

---

## Logic

---

**L**ogic is a systematic way of thinking that allows us to parse the meanings of sentences and to deduce new information from old information. You use logic informally in everyday life and certainly also in doing mathematics. For example, say you are working with a certain circle (call it “Circle X”), and suppose you have available the following two pieces of information.

1. Circle X has a radius of 3 units.
2. If any circle has radius  $r$ , then its area is  $\pi r^2$  square units.

You have no trouble putting these two facts together to get:

3. Circle X has area  $9\pi$  square units.

In doing this you are using logic to combine existing information to produce new information. Because deducing new information is central to mathematics, logic plays a fundamental role. This chapter is intended to give you a sufficient mastery of it.

It is important to realize that logic is a process of deducing information correctly, *not* just deducing correct information. For example, suppose we were mistaken and Circle X actually had a radius of 4, not 3. Let’s look at our exact same argument again.

1. Circle X has a radius of 3 units.
2. If any circle has radius  $r$ , then its area is  $\pi r^2$  square units.

---

3. Circle X has area  $9\pi$  square units.

The sentence “*Circle X has a radius of 3 units.*” is now untrue, and so is our conclusion “*Circle X has area  $9\pi$  square units.*” But the logic is perfectly correct; the information was combined correctly, even if some of it was false. This distinction between correct logic and correct information is significant because it is often important to follow the consequences of an incorrect assumption. Ideally, we want both our logic *and* our information to be correct, but the point is that they are different things.

In proving theorems, we apply logic to information that is considered obviously true (such as “*Any two points determine exactly one line.*”) or is already known to be true (e.g., the Pythagorean theorem). If our logic is correct, then anything we deduce from such information will also be true (or at least as true as the “obviously true” information we began with).

## 2.1 Statements

The study of logic begins with statements. A **statement** is a sentence or a mathematical expression that is either definitely true or definitely false. You can think of statements as pieces of information that are either correct or incorrect. Thus statements are pieces of information that we might apply logic to in order to produce other pieces of information (which are also statements).

**Example 2.1** Here are some examples of statements. They are all true.

If a circle has radius  $r$ , then its area is  $\pi r^2$  square units.

Every even number is divisible by 2.

$$2 \in \mathbb{Z}$$

$$\sqrt{2} \notin \mathbb{Z}$$

$$\mathbb{N} \subseteq \mathbb{Z}$$

The set  $\{0, 1, 2\}$  has three elements.

Some right triangles are isosceles.

**Example 2.2** Here are some additional statements. They are all false.

All right triangles are isosceles.

$$5 = 2$$

$$\sqrt{2} \notin \mathbb{R}$$

$$\mathbb{Z} \subseteq \mathbb{N}$$

$$\{0, 1, 2\} \cap \mathbb{N} = \emptyset$$

**Example 2.3** Here non-statements are paired with similar statements.

NOT a statement:	Statement:
Add 5 to both sides.	Adding 5 to both sides of $x - 5 = 37$ gives $x = 42$ .
$\mathbb{Z}$	$42 \in \mathbb{Z}$
42	42 is not a number.
What is the solution of $2x = 84$ ?	The solution of $2x = 84$ is 42.

**Example 2.4** We will often use the letters  $P$ ,  $Q$ ,  $R$  and  $S$  to stand for specific statements. When more letters are needed we can use subscripts. Here are more statements, designated with letters. You decide which of them are true and which are false.

$P$  : For every integer  $n > 1$ , the number  $2^n - 1$  is prime.

$Q$  : Every polynomial of degree  $n$  has at most  $n$  roots.

$R$  : The function  $f(x) = x^2$  is continuous.

$S_1$  :  $\mathbb{Z} \subseteq \emptyset$

$S_2$  :  $\{0, -1, -2\} \cap \mathbb{N} = \emptyset$

Designating statements with letters (as was done above) is a very useful shorthand. In discussing a particular statement, such as “*The function  $f(x) = x^2$  is continuous*,” it is convenient to just refer to it as  $R$  to avoid having to write or say it many times.

Statements can contain variables. Here is an example.

$P$  : If an integer  $x$  is a multiple of 6, then  $x$  is even.

This is a sentence that is true. (All multiples of 6 are even, so no matter which multiple of 6 the integer  $x$  happens to be, it is even.) Since the sentence  $P$  is definitely true, it is a statement. When a sentence or statement  $P$  contains a variable such as  $x$ , we sometimes denote it as  $P(x)$  to indicate that it is saying something about  $x$ . Thus the above statement can be denoted as

$P(x)$  : If an integer  $x$  is a multiple of 6, then  $x$  is even.

A statement or sentence involving two variables might be denoted  $P(x, y)$ , and so on.

It is quite possible for a sentence containing variables to not be a statement. Consider the following example.

$Q(x)$  : The integer  $x$  is even.

Is this a statement? Whether it is true or false depends on just which integer  $x$  is. It is true if  $x = 4$  and false if  $x = 7$ , etc. But without any stipulations on the value of  $x$  it is impossible to say whether  $Q(x)$  is true or false. Since it is neither definitely true nor definitely false,  $Q(x)$  cannot be a statement. A sentence such as this, whose truth depends on the value of one or more variables, is called an **open sentence**. The variables in an open sentence (or statement) can represent any type of entity, not just numbers. Here is an open sentence where the variables are functions:

$R(f,g)$ : The function  $f$  is the derivative of the function  $g$ .

This open sentence is true if  $f(x) = 2x$  and  $g(x) = x^2$ . It is false if  $f(x) = x^3$  and  $g(x) = x^2$ , etc. A sentence such as  $R(f,g)$  (that involves variables) can be denoted either as  $R(f,g)$  or just  $R$ . We use the expression  $R(f,g)$  when we want to emphasize that the sentence involves variables.

We will have more to say about open sentences later, but for now let's return to statements.

Statements are everywhere in mathematics. Any result or theorem that has been proved true is a statement. The quadratic formula and the Pythagorean theorem are both statements:

$$P : \text{The solutions of the equation } ax^2 + bx + c = 0 \text{ are } x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

$$Q : \text{If a right triangle has legs of lengths } a \text{ and } b \text{ and hypotenuse of length } c, \text{ then } a^2 + b^2 = c^2.$$

Here is a very famous statement, so famous, in fact, that it has a name. It is called **Fermat's last theorem** after Pierre Fermat, a seventeenth-century French mathematician who scribbled it in the margin of a book.

$R$  : For all numbers  $a, b, c, n \in \mathbb{N}$  with  $n > 2$ , it is the case that  $a^n + b^n \neq c^n$ .

Fermat believed this statement to be true. He noted that he could prove it was true, except the book's margin was too narrow to contain his proof. It is doubtful that he really had a correct proof in mind, for after his death many generations of brilliant mathematicians tried unsuccessfully to prove that his statement was true (or false). Finally, in 1993, Andrew Wiles of Princeton University announced that he had devised a proof. Wiles had worked on the problem for over seven years, and his proof runs through hundreds of pages. The moral of this story is that some true statements are not obviously true.

Here is another statement famous enough to be named. It was first posed in the eighteenth century by the German mathematician Christian Goldbach, and thus is called the **Goldbach conjecture**:

$S$  : Every even integer greater than 2 is a sum of two prime numbers.

You must agree that  $S$  is either true or false. It appears to be true, because when you examine even numbers that are bigger than 2, they seem to be sums of two primes:  $4 = 2 + 2$ ,  $6 = 3 + 3$ ,  $8 = 3 + 5$ ,  $10 = 5 + 5$ ,  $12 = 5 + 7$ ,  $100 = 17 + 83$  and so on. But that's not to say there isn't some large even

number that's not the sum of two primes. If such a number exists, then  $S$  is false. The thing is, in the over 260 years since Goldbach first posed this problem, no one has been able to determine whether it's true or false. But since it is clearly either true or false,  $S$  is a statement.

This book is about the methods that can be used to prove that  $S$  (or any other statement) is true or false. To prove that  $S$  is true, start with obvious statements (or other statements that have been proven true) and use logic to deduce more and more complex statements until finally we obtain the statement  $S$ . Of course some statements are more difficult to prove than others, and  $S$  appears to be notoriously difficult; we will concentrate on statements that are easier to prove.

But the point is this: In proving that statements are true, we use logic to help us understand statements and to combine pieces of information to produce new pieces of information. In the next several sections we explore some standard ways that statements can be combined to form new statements, or broken down into simpler statements.

### **Exercises for Section 2.1**

Decide whether or not the following are statements. In the case of a statement, say if it is true or false, if possible.

1. Every real number is an even integer.
2. Every even integer is a real number.
3. If  $x$  and  $y$  are real numbers and  $5x = 5y$ , then  $x = y$ .
4. Sets  $\mathbb{Z}$  and  $\mathbb{N}$ .
5. Sets  $\mathbb{Z}$  and  $\mathbb{N}$  are infinite.
6. Some sets are finite.
7. The derivative of any polynomial of degree 5 is a polynomial of degree 6.
8.  $\mathbb{N} \notin \mathcal{P}(\mathbb{N})$ .
9.  $\cos(x) = -1$
10.  $(\mathbb{R} \times \mathbb{N}) \cap (\mathbb{N} \times \mathbb{R}) = \mathbb{N} \times \mathbb{N}$
11. The integer  $x$  is a multiple of 7.
12. If the integer  $x$  is a multiple of 7, then it is divisible by 7.
13. Either  $x$  is a multiple of 7, or it is not.
14. Call me Ishmael.
15. In the beginning, God created the heaven and the earth.

## 2.2 And, Or, Not

The word “and” can be used to combine two statements to form a new statement. Consider for example the following sentence.

$R_1$  : The number 2 is even **and** the number 3 is odd.

We recognize this as a true statement, based on our ingrained understanding of the meaning of the word “and.” Notice that  $R_1$  is made up of two simpler statements:

$P$  : The number 2 is even.

$Q$  : The number 3 is odd.

These are joined together by the word “and” to form the more complex statement  $R_1$ . The statement  $R_1$  asserts that  $P$  and  $Q$  are both true. Since both  $P$  and  $Q$  are in fact true, the statement  $R_1$  is also true.

Had one or both of  $P$  and  $Q$  been false, then  $R_1$  would be false. For instance, each of the following statements is false.

$R_2$  : The number 1 is even **and** the number 3 is odd.

$R_3$  : The number 2 is even **and** the number 4 is odd.

$R_4$  : The number 3 is even **and** the number 2 is odd.

From these examples we see that any two statements  $P$  and  $Q$  can be combined to form a new statement “ $P$  **and**  $Q$ .” In the spirit of using letters to denote statements, we now introduce the special symbol  $\wedge$  to stand for the word “and.” Thus if  $P$  and  $Q$  are statements,  $P \wedge Q$  stands for the statement “ $P$  **and**  $Q$ .” The statement  $P \wedge Q$  is true if both  $P$  and  $Q$  are true; otherwise it is false. This is summarized in the following table, called a **truth table**.

$P$	$Q$	$P \wedge Q$
$T$	$T$	$T$
$T$	$F$	$F$
$F$	$T$	$F$
$F$	$F$	$F$

In this table,  $T$  stands for “True,” and  $F$  stands for “False.” ( $T$  and  $F$  are called **truth values**.) Each line lists one of the four possible combinations of truth values for  $P$  and  $Q$ , and the column headed by  $P \wedge Q$  tells whether the statement  $P \wedge Q$  is true or false in each case.

Statements can also be combined using the word “or.” Consider the following four statements.

$S_1$  : The number 2 is even **or** the number 3 is odd.

$S_2$  : The number 1 is even **or** the number 3 is odd.

$S_3$  : The number 2 is even **or** the number 4 is odd.

$S_4$  : The number 3 is even **or** the number 2 is odd.

In mathematics, the assertion “ $P$  **or**  $Q$ ” is always understood to mean that one *or both* of  $P$  and  $Q$  is true. Thus statements  $S_1$ ,  $S_2$ ,  $S_3$  are all true, while  $S_4$  is false. The symbol  $\vee$  is used to stand for the word “or.” So if  $P$  and  $Q$  are statements,  $P \vee Q$  represents the statement “ $P$  **or**  $Q$ .” Here is the truth table.

$P$	$Q$	$P \vee Q$
$T$	$T$	$T$
$T$	$F$	$T$
$F$	$T$	$T$
$F$	$F$	$F$

It is important to be aware that the meaning of “or” expressed in the above table differs from the way it is often used in everyday conversation. For example, suppose a university official makes the following threat:

You pay your tuition **or** you will be withdrawn from school.

You understand that this means that either you pay your tuition *or* you will be withdrawn from school, *but not both*. In mathematics we never use the word “or” in such a sense. For us “or” means exactly what is stated in the table for  $\vee$ . Thus  $P \vee Q$  being true means *one or both* of  $P$  and  $Q$  is true. If we ever need to express the fact that exactly one of  $P$  and  $Q$  is true, we use one of the following constructions:

**$P$  or  $Q$ , but not both.**

**Either  $P$  or  $Q$ .**

**Exactly one of  $P$  or  $Q$ .**

If the university official were a mathematician, he might have qualified his statement in one of the following ways.

Pay your tuition **or** you will be withdrawn from school, **but not both**.

**Either** you pay your tuition **or** you will be withdrawn from school.

To conclude this section, we mention another way of obtaining new statements from old ones. Given any statement  $P$ , we can form the new statement “**It is not true that  $P$ .**” For example, consider the following statement.

The number 2 is even.

This statement is true. Now change it by inserting the words “It is not true that” at the beginning:

**It is not true that** the number 2 is even.

This new statement is false.

For another example, starting with the false statement “ $2 \in \emptyset$ ,” we get the true statement “It is not true that  $2 \in \emptyset$ .”

We use the symbol  $\sim$  to stand for the words “It’s not true that,” so  $\sim P$  means “**It’s not true that  $P$ .**” We can read  $\sim P$  simply as “not  $P$ .” Unlike  $\wedge$  and  $\vee$ , which combine two statements, the symbol  $\sim$  just alters a single statement. Thus its truth table has just two lines, one for each possible value of  $P$ .

$P$	$\sim P$
T	F
F	T

The statement  $\sim P$  is called the **negation** of  $P$ . The negation of a specific statement can be expressed in numerous ways. Consider

$P$  : The number 2 is even.

Here are several ways of expressing its negation.

$\sim P$  : It’s not true that the number 2 is even.

$\sim P$  : It is false that the number 2 is even.

$\sim P$  : The number 2 is not even.

In this section we’ve learned how to combine or modify statements with the operations  $\wedge$ ,  $\vee$  and  $\sim$ . Of course we can also apply these operations to open sentences or a mixture of open sentences and statements. For example,  $(x \text{ is an even integer}) \wedge (3 \text{ is an odd integer})$  is an open sentence that is a combination of an open sentence and a statement.

### Exercises for Section 2.2

Express each statement or open sentence in one of the forms  $P \wedge Q$ ,  $P \vee Q$ , or  $\sim P$ . Be sure to also state exactly what statements  $P$  and  $Q$  stand for.

1. The number 8 is both even and a power of 2.
2. The matrix  $A$  is not invertible.
3.  $x \neq y$
4.  $x < y$
5.  $y \geq x$
6. There is a quiz scheduled for Wednesday or Friday.
7. The number  $x$  equals zero, but the number  $y$  does not.
8. At least one of the numbers  $x$  and  $y$  equals 0.
9.  $x \in A - B$
10.  $x \in A \cup B$
11.  $A \in \{X \in \mathcal{P}(\mathbb{N}) : |\overline{X}| < \infty\}$
12. Happy families are all alike, but each unhappy family is unhappy in its own way.  
(Leo Tolstoy, *Anna Karenina*)
13. Human beings want to be good, but not too good, and not all the time.  
(George Orwell)
14. A man should look for what is, and not for what he thinks should be.  
(Albert Einstein)

### 2.3 Conditional Statements

There is yet another way to combine two statements. Suppose we have in mind a specific integer  $a$ . Consider the following statement about  $a$ .

$R$  : If the integer  $a$  is a multiple of 6, then  $a$  is divisible by 2.

We immediately spot this as a true statement based on our knowledge of integers and the meanings of the words “if” and “then.” If integer  $a$  is a multiple of 6, then  $a$  is even, so therefore  $a$  is divisible by 2. Notice that  $R$  is built up from two simpler statements:

$P$  : The integer  $a$  is a multiple of 6.

$Q$  : The integer  $a$  is divisible by 2.

$R$  : If  $P$ , then  $Q$ .

In general, given any two statements  $P$  and  $Q$  whatsoever, we can form the new statement “If  $P$ , then  $Q$ .” This is written symbolically as  $P \Rightarrow Q$  which we read as “If  $P$ , then  $Q$ ,” or “ $P$  implies  $Q$ .” Like  $\wedge$  and  $\vee$ , the symbol  $\Rightarrow$  has a very specific meaning. When we assert that the statement  $P \Rightarrow Q$  is true, we mean that if  $P$  is true then  $Q$  must also be true. (In other words we mean that the condition  $P$  being true forces  $Q$  to be true.) A statement of form  $P \Rightarrow Q$  is called a **conditional** statement because it means  $Q$  will be true *under the condition* that  $P$  is true.

Think of  $P \Rightarrow Q$  as a promise that whenever  $P$  is true,  $Q$  will be true also. There is only one way this promise can be broken (i.e., be false), namely if  $P$  is true but  $Q$  is false. So the truth table for the promise  $P \Rightarrow Q$  is as follows:

$P$	$Q$	$P \Rightarrow Q$
$T$	$T$	$T$
$T$	$F$	$F$
$F$	$T$	$T$
$F$	$F$	$T$

Perhaps you are bothered by how  $P \Rightarrow Q$  is true in the last two lines. Here is an example to explain it. Suppose your professor makes this promise:

If you pass the final exam, then you will pass the course.

Your professor is making the promise

(You pass the exam)  $\Rightarrow$  (You pass the course).

Under what circumstances did she lie? There are four possible scenarios, depending on whether or not you passed the exam and whether or not you passed the course. These scenarios are tallied in the following table.

You pass exam	You pass course	(You pass exam) $\Rightarrow$ (You pass course)
$T$	$T$	$T$
$T$	$F$	$F$
$F$	$T$	$T$
$F$	$F$	$T$

The first row is the scenario in which you pass the exam and you pass the course. Clearly the professor kept her promise, so the  $T$  in the third column indicates that she told the truth. In the second row, you passed the exam but failed the course. In this case your professor broke her promise, and the  $F$  in the third column indicates that what she said was untrue.

The third row describes the scenario in which you failed the exam but still passed the course. How could that happen? Maybe your professor felt sorry for you. But that doesn't make her a liar. Her only promise was that if you passed the exam then you would pass the course. She did not say passing the exam was the *only way* to pass the course. Since she didn't lie, then she told the truth, so there is a  $T$  in the third column.

Finally look at the fourth row: you failed the exam and failed the course. Your professor certainly **did not** lie to you. Hence the  $T$  in the third column.

For another example, consider this statement:

If this month is September, **then** there is an equinox this month.

An *equinox* is a day for which there are equal hours of darkness and light. There are two equinoxes per year, one in September and the other in March. The above statement is thus unquestionably true, for it asserts correctly that if the current month is September, then an equinox will occur this month. In symbolic form, our statement is

(This month is September)  $\Rightarrow$  (There is an equinox this month).

This statement is true, but the open sentences  $P$ : “*This month is September*,” and  $Q$ : “*There is an equinox this month*,” are either true or false, depending on what month it is. But  $P \Rightarrow Q$  is always true. This is shown below for three (out of 12) months. Notice how  $P \Rightarrow Q$  is true, even when  $P$  is false.

	This month is September	There is an equinox this month	$\left( \begin{array}{l} \text{This month} \\ \text{is September} \end{array} \right) \Rightarrow \left( \begin{array}{l} \text{There is an} \\ \text{equinox} \\ \text{this month} \end{array} \right)$
Sept.	$T$	$T$	$T$
March	$F$	$T$	$T$
May	$F$	$F$	$T$

As  $P \Rightarrow Q$  is a true statement in this particular example, there is no month with  $P$  true and  $Q$  false. (Unless we imagine that Earth is destroyed by an asteroid before September 21, a possibility that we shall not entertain.)

In mathematics, whenever we encounter the construction “*If  $P$ , then  $Q$* ,” it means exactly what the truth table for  $\Rightarrow$  expresses. Of course there are other grammatical constructions that also mean  $P \Rightarrow Q$ . Here is a summary of the main ones. The meaning of each is encapsulated by the table for  $\Rightarrow$ .

If  $P$ , then  $Q$ .

$Q$  if  $P$ .

$Q$  whenever  $P$ .

$Q$ , provided that  $P$ .

Whenever  $P$ , then also  $Q$ .

$P$  is a sufficient condition for  $Q$ .

For  $Q$ , it is sufficient that  $P$ .

$Q$  is a necessary condition for  $P$ .

For  $P$ , it is necessary that  $Q$ .

$P$  only if  $Q$ .

$$P \Rightarrow Q$$

$P$	$Q$	$P \Rightarrow Q$
$T$	$T$	$T$
$T$	$F$	$F$
$F$	$T$	$T$
$F$	$F$	$T$

These can all be used in the place of (and mean exactly the same thing as) “*If P, then Q.*” You should analyze the meaning of each one and convince yourself that it captures the meaning of  $P \Rightarrow Q$ . For example,  $P \Rightarrow Q$  means the condition of  $P$  being true is enough (i.e., sufficient) to make  $Q$  true; hence “*P is a sufficient condition for Q.*”

The wording can be tricky. An everyday situation may help clarify it. For example, consider your professor’s promise:

(You pass the exam)  $\Rightarrow$  (You pass the course).

This means that your passing the exam is a sufficient (though perhaps not necessary) condition for your passing the course. Thus your professor might just as well have phrased her promise in one of the following ways.

Passing the exam is a sufficient condition for passing the course.

For you to pass the course, it is sufficient that you pass the exam.

However, when we want to say “*If P, then Q*” in everyday conversation, we do not normally express this as “*Q is a necessary condition for P*” or “*P only if Q*.” But such constructions are not uncommon in mathematics. To understand why they make sense, notice that  $P \Rightarrow Q$  being true means that it’s impossible that  $P$  is true but  $Q$  is false, so in order for  $P$  to be true it is necessary that  $Q$  is true; hence “*Q is a necessary condition for P*.” And this means that  $P$  can only be true if  $Q$  is true, i.e., “*P only if Q*.”

---

### Exercises for Section 2.3

Without changing their meanings, convert each of the following sentences into a sentence having the form “*If P, then Q.*”

1. A matrix is invertible provided that its determinant is not zero.
2. For a function to be continuous, it is sufficient that it is differentiable.
3. For a function to be continuous, it is necessary that it is integrable.
4. A function is rational if it is a polynomial.
5. An integer is divisible by 8 only if it is divisible by 4.
6. Whenever a surface has only one side, it is non-orientable.
7. A series converges whenever it converges absolutely.
8. A geometric series with ratio  $r$  converges if  $|r| < 1$ .
9. A function is integrable provided the function is continuous.
10. The discriminant is negative only if the quadratic equation has no real solutions.
11. You fail only if you stop writing. (Ray Bradbury)

12. People will generally accept facts as truth only if the facts agree with what they already believe. (Andy Rooney)
  13. Whenever people agree with me I feel I must be wrong. (Oscar Wilde)
- 

## 2.4 Biconditional Statements

It is important to understand that  $P \Rightarrow Q$  is not the same as  $Q \Rightarrow P$ . To see why, suppose that  $a$  is some integer and consider the statements

$$\begin{aligned} (a \text{ is a multiple of } 6) &\Rightarrow (a \text{ is divisible by } 2), \\ (a \text{ is divisible by } 2) &\Rightarrow (a \text{ is a multiple of } 6). \end{aligned}$$

The first statement asserts that if  $a$  is a multiple of 6 then  $a$  is divisible by 2. This is clearly true, for any multiple of 6 is even and therefore divisible by 2. The second statement asserts that if  $a$  is divisible by 2 then it is a multiple of 6. This is not necessarily true, for  $a = 4$  (for instance) is divisible by 2, yet not a multiple of 6. Therefore the meanings of  $P \Rightarrow Q$  and  $Q \Rightarrow P$  are in general quite different. The conditional statement  $Q \Rightarrow P$  is called the **converse** of  $P \Rightarrow Q$ , so a conditional statement and its converse express entirely different things.

But sometimes, if  $P$  and  $Q$  are just the right statements, it can happen that  $P \Rightarrow Q$  and  $Q \Rightarrow P$  are both necessarily true. For example, consider the statements

$$\begin{aligned} (a \text{ is even}) &\Rightarrow (a \text{ is divisible by } 2), \\ (a \text{ is divisible by } 2) &\Rightarrow (a \text{ is even}). \end{aligned}$$

No matter what value  $a$  has, both of these statements are true. Since both  $P \Rightarrow Q$  and  $Q \Rightarrow P$  are true, it follows that  $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$  is true.

Let's introduce a new symbol  $\Leftrightarrow$  to express the meaning of the statement  $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$ . The expression  $P \Leftrightarrow Q$  is understood to have exactly the same meaning as  $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$ . According to the previous section,  $Q \Rightarrow P$  is read as " $P$  if  $Q$ ," and  $P \Rightarrow Q$  can be read as " $P$  only if  $Q$ ." Therefore we pronounce  $P \Leftrightarrow Q$  as " $P$  if and only if  $Q$ ." For example, given an integer  $a$ , we have the true statement

$$(a \text{ is even}) \Leftrightarrow (a \text{ is divisible by } 2),$$

which we can read as "*The integer  $a$  is even if and only if  $a$  is divisible by 2.*"

The truth table for  $\Leftrightarrow$  is shown below. Notice that in the first and last rows, both  $P \Rightarrow Q$  and  $Q \Rightarrow P$  are true (according to the truth table for  $\Rightarrow$ ), so  $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$  is true, and hence  $P \Leftrightarrow Q$  is true. However, in the middle two rows one of  $P \Rightarrow Q$  or  $Q \Rightarrow P$  is false, so  $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$  is false, making  $P \Leftrightarrow Q$  false.

$P$	$Q$	$P \Leftrightarrow Q$
$T$	$T$	$T$
$T$	$F$	$F$
$F$	$T$	$F$
$F$	$F$	$T$

Compare the statement  $R : (a \text{ is even}) \Leftrightarrow (a \text{ is divisible by 2})$  with this truth table. If  $a$  is even then the two statements on either side of  $\Leftrightarrow$  are true, so according to the table  $R$  is true. If  $a$  is odd then the two statements on either side of  $\Leftrightarrow$  are false, and again according to the table  $R$  is true. Thus  $R$  is true no matter what value  $a$  has. In general,  $P \Leftrightarrow Q$  being true means  $P$  and  $Q$  are both true or both false.

Not surprisingly, there are many ways of saying  $P \Leftrightarrow Q$  in English. The following constructions all mean  $P \Leftrightarrow Q$ :

- $P$  if and only if  $Q$ .  
 $P$  is a necessary and sufficient condition for  $Q$ .  
For  $P$  it is necessary and sufficient that  $Q$ .  
 $P$  is equivalent to  $Q$ .  
If  $P$ , then  $Q$ , and conversely.
- $P \Leftrightarrow Q$

The first three of these just combine constructions from the previous section to express that  $P \Rightarrow Q$  and  $Q \Rightarrow P$ . In the last one, the words “...and conversely” mean that in addition to “If  $P$ , then  $Q$ ” being true, the converse statement “If  $Q$ , then  $P$ ” is also true.

### Exercises for Section 2.4

Without changing their meanings, convert each of the following sentences into a sentence having the form “ $P$  if and only if  $Q$ .”

1. For matrix  $A$  to be invertible, it is necessary and sufficient that  $\det(A) \neq 0$ .
2. If a function has a constant derivative then it is linear, and conversely.
3. If  $xy = 0$  then  $x = 0$  or  $y = 0$ , and conversely.
4. If  $a \in \mathbb{Q}$  then  $5a \in \mathbb{Q}$ , and if  $5a \in \mathbb{Q}$  then  $a \in \mathbb{Q}$ .
5. For an occurrence to become an adventure, it is necessary and sufficient for one to recount it. (Jean-Paul Sartre)

## 2.5 Truth Tables for Statements

You should now know the truth tables for  $\wedge$ ,  $\vee$ ,  $\sim$ ,  $\Rightarrow$  and  $\Leftrightarrow$ . They should be *internalized* as well as memorized. You must understand the symbols thoroughly, for we now combine them to form more complex statements.

For example, suppose we want to convey that one or the other of  $P$  and  $Q$  is true but they are not both true. No single symbol expresses this, but we could combine them as

$$(P \vee Q) \wedge \sim(P \wedge Q),$$

which literally means:

*P or Q is true, and it is not the case that both P and Q are true.*

This statement will be true or false depending on the truth values of  $P$  and  $Q$ . In fact we can make a truth table for the entire statement. Begin as usual by listing the possible true/false combinations of  $P$  and  $Q$  on four lines. The statement  $(P \vee Q) \wedge \sim(P \wedge Q)$  contains the individual statements  $(P \vee Q)$  and  $(P \wedge Q)$ , so we next tally their truth values in the third and fourth columns. The fifth column lists values for  $\sim(P \wedge Q)$ , and these are just the opposites of the corresponding entries in the fourth column. Finally, combining the third and fifth columns with  $\wedge$ , we get the values for  $(P \vee Q) \wedge \sim(P \wedge Q)$  in the sixth column.

$P$	$Q$	$(P \vee Q)$	$(P \wedge Q)$	$\sim(P \wedge Q)$	$(P \vee Q) \wedge \sim(P \wedge Q)$
$T$	$T$	$T$	$T$	$F$	$F$
$T$	$F$	$T$	$F$	$T$	$T$
$F$	$T$	$T$	$F$	$T$	$T$
$F$	$F$	$F$	$F$	$T$	$F$

This truth table tells us that  $(P \vee Q) \wedge \sim(P \wedge Q)$  is true precisely when one but not both of  $P$  and  $Q$  are true, so it has the meaning we intended. (Notice that the middle three columns of our truth table are just “helper columns” and are not necessary parts of the table. In writing truth tables, you may choose to omit such columns if you are confident about your work.)

For another example, consider the following familiar statement about real numbers  $x$  and  $y$ :

The product  $xy$  equals zero if and only if  $x = 0$  or  $y = 0$ .

This can be modeled as  $(xy = 0) \Leftrightarrow (x = 0 \vee y = 0)$ . If we introduce letters  $P, Q$  and  $R$  for the statements  $xy = 0$ ,  $x = 0$  and  $y = 0$ , it becomes  $P \Leftrightarrow (Q \vee R)$ . Notice that the parentheses are necessary here, for without them we wouldn’t know whether to read the statement as  $P \Leftrightarrow (Q \vee R)$  or  $(P \Leftrightarrow Q) \vee R$ .

Making a truth table for  $P \Leftrightarrow (Q \vee R)$  entails a line for each T/F combination for the three statements  $P$ ,  $Q$  and  $R$ . The eight possible combinations are tallied in the first three columns of the following table.

$P$	$Q$	$R$	$Q \vee R$	$P \Leftrightarrow (Q \vee R)$
$T$	$T$	$T$	$T$	$T$
$T$	$T$	$F$	$T$	$T$
$T$	$F$	$T$	$T$	$T$
$T$	$F$	$F$	$F$	$F$
$F$	$T$	$T$	$T$	$F$
$F$	$T$	$F$	$T$	$F$
$F$	$F$	$T$	$T$	$F$
$F$	$F$	$F$	$F$	$T$

We fill in the fourth column using our knowledge of the truth table for  $\vee$ . Finally the fifth column is filled in by combining the first and fourth columns with our understanding of the truth table for  $\Leftrightarrow$ . The resulting table gives the true/false values of  $P \Leftrightarrow (Q \vee R)$  for all values of  $P$ ,  $Q$  and  $R$ .

Notice that when we plug in various values for  $x$  and  $y$ , the statements  $P : xy = 0$ ,  $Q : x = 0$  and  $R : y = 0$  have various truth values, but the statement  $P \Leftrightarrow (Q \vee R)$  is always true. For example, if  $x = 2$  and  $y = 3$ , then  $P$ ,  $Q$  and  $R$  are all false. This scenario is described in the last row of the table, and there we see that  $P \Leftrightarrow (Q \vee R)$  is true. Likewise if  $x = 0$  and  $y = 7$ , then  $P$  and  $Q$  are true and  $R$  is false, a scenario described in the second line of the table, where again  $P \Leftrightarrow (Q \vee R)$  is true. There is a simple reason why  $P \Leftrightarrow (Q \vee R)$  is true for any values of  $x$  and  $y$ : It is that  $P \Leftrightarrow (Q \vee R)$  represents  $(xy = 0) \Leftrightarrow (x = 0 \vee y = 0)$ , which is a *true mathematical statement*. It is absolutely impossible for it to be false.

This may make you wonder about the lines in the table where  $P \Leftrightarrow (Q \vee R)$  is false. Why are they there? The reason is that  $P \Leftrightarrow (Q \vee R)$  can also represent a false statement. To see how, imagine that at the end of the semester your professor makes the following promise.

You pass the class if and only if you get an “A” on the final or you get a “B” on the final.

This promise has the form  $P \Leftrightarrow (Q \vee R)$ , so its truth values are tabulated in the above table. Imagine it turned out that you got an “A” on the exam but failed the course. Then surely your professor lied to you. In fact,  $P$  is false,  $Q$  is true and  $R$  is false. This scenario is reflected in the sixth line of the table, and indeed  $P \Leftrightarrow (Q \vee R)$  is false (i.e., it is a lie).

The moral of this example is that people can lie, but true mathematical statements *never* lie.

We close this section with a word about the use of parentheses. The symbol  $\sim$  is analogous to the minus sign in algebra. It negates the expression it precedes. Thus  $\sim P \vee Q$  means  $(\sim P) \vee Q$ , not  $\sim(P \vee Q)$ . In  $\sim(P \vee Q)$ , the value of the entire expression  $P \vee Q$  is negated.

### Exercises for Section 2.5

Write a truth table for the logical statements in problems 1–9:

1.  $P \vee (Q \Rightarrow R)$
4.  $\sim(P \vee Q) \vee (\sim P)$
7.  $(P \wedge \sim P) \Rightarrow Q$
2.  $(Q \vee R) \Leftrightarrow (R \wedge Q)$
5.  $(P \wedge \sim P) \vee Q$
8.  $P \vee (Q \wedge \sim R)$
3.  $\sim(P \Rightarrow Q)$
6.  $(P \wedge \sim P) \wedge Q$
9.  $\sim(\sim P \vee \sim Q)$
10. Suppose the statement  $((P \wedge Q) \vee R) \Rightarrow (R \vee S)$  is false. Find the truth values of  $P, Q, R$  and  $S$ . (This can be done without a truth table.)
11. Suppose  $P$  is false and that the statement  $(R \Rightarrow S) \Leftrightarrow (P \wedge Q)$  is true. Find the truth values of  $R$  and  $S$ . (This can be done without a truth table.)

### 2.6 Logical Equivalence

In contemplating the truth table for  $P \Leftrightarrow Q$ , you probably noticed that  $P \Leftrightarrow Q$  is true exactly when  $P$  and  $Q$  are both true or both false. In other words,  $P \Leftrightarrow Q$  is true precisely when at least one of the statements  $P \wedge Q$  or  $\sim P \wedge \sim Q$  is true. This may tempt us to say that  $P \Leftrightarrow Q$  means the same thing as  $(P \wedge Q) \vee (\sim P \wedge \sim Q)$ .

To see if this is really so, we can write truth tables for  $P \Leftrightarrow Q$  and  $(P \wedge Q) \vee (\sim P \wedge \sim Q)$ . In doing this, it is more efficient to put these two statements into the same table, as follows. (This table has helper columns for the intermediate expressions  $\sim P$ ,  $\sim Q$ ,  $(P \wedge Q)$  and  $(\sim P \wedge \sim Q)$ .)

$P$	$Q$	$\sim P$	$\sim Q$	$(P \wedge Q)$	$(\sim P \wedge \sim Q)$	$(P \wedge Q) \vee (\sim P \wedge \sim Q)$	$P \Leftrightarrow Q$
T	T	F	F	T	F	T	T
T	F	F	T	F	F	F	F
F	T	T	F	F	F	F	F
F	F	T	T	F	T	T	T

The table shows that  $P \Leftrightarrow Q$  and  $(P \wedge Q) \vee (\sim P \wedge \sim Q)$  have the same truth value, no matter the values  $P$  and  $Q$ . It is as if  $P \Leftrightarrow Q$  and  $(P \wedge Q) \vee (\sim P \wedge \sim Q)$  are algebraic expressions that are equal no matter what is “plugged into”

variables  $P$  and  $Q$ . We express this state of affairs by writing

$$P \Leftrightarrow Q = (P \wedge Q) \vee (\sim P \wedge \sim Q)$$

and saying that  $P \Leftrightarrow Q$  and  $(P \wedge Q) \vee (\sim P \wedge \sim Q)$  are **logically equivalent**.

In general, two statements are **logically equivalent** if their truth values match up line-for-line in a truth table.

Logical equivalence is important because it can give us different (and potentially useful) ways of looking at the same thing. As an example, the following table shows that  $P \Rightarrow Q$  is logically equivalent to  $(\sim Q) \Rightarrow (\sim P)$ .

$P$	$Q$	$\sim P$	$\sim Q$	$(\sim Q) \Rightarrow (\sim P)$	$P \Rightarrow Q$
T	T	F	F	T	T
T	F	F	T	F	F
F	T	T	F	T	T
F	F	T	T	T	T

The fact  $P \Rightarrow Q = (\sim Q) \Rightarrow (\sim P)$  is useful because so many theorems have the form  $P \Rightarrow Q$ . As we will see in Chapter 5, proving such a theorem may be easier if we express it in the logically equivalent form  $(\sim Q) \Rightarrow (\sim P)$ .

Two pairs of logically equivalent statements in particular are significant enough to have a special name: **DeMorgan's laws**.

### Fact 2.1 (DeMorgan's Laws)

1.  $\sim(P \wedge Q) = (\sim P) \vee (\sim Q)$
2.  $\sim(P \vee Q) = (\sim P) \wedge (\sim Q)$

The first of DeMorgan's laws is verified by the following table. You are asked to verify the second in one of the exercises.

$P$	$Q$	$\sim P$	$\sim Q$	$P \wedge Q$	$\sim(P \wedge Q)$	$(\sim P) \vee (\sim Q)$
T	T	F	F	T	F	F
T	F	F	T	F	T	T
F	T	T	F	F	T	T
F	F	T	T	F	T	T

DeMorgan's laws are actually very natural and intuitive. Consider the statement  $\sim(P \wedge Q)$ , which we can interpret as meaning that *it is not the case that both  $P$  and  $Q$  are true*. If it is not the case that both  $P$  and  $Q$  are true, then at least one of  $P$  or  $Q$  is false, in which case  $(\sim P) \vee (\sim Q)$  is true. Thus  $\sim(P \wedge Q)$  means the same thing as  $(\sim P) \vee (\sim Q)$ .

DeMorgan's laws can be very useful. Suppose we happen to know that some statement of form  $\sim(P \vee Q)$  is true. The second of DeMorgan's laws tells us that  $(\sim Q) \wedge (\sim P)$  is also true, hence  $\sim P$  and  $\sim Q$  are both true as well. Quickly obtaining this additional information can be extremely useful.

Here is a summary of some significant logical equivalences. Those that are not immediately obvious can be verified with truth tables.

$$P \Rightarrow Q = (\sim Q) \Rightarrow (\sim P) \quad \text{Contrapositive law} \quad (2.1)$$

$$\begin{aligned} \sim(P \wedge Q) &= \sim P \vee \sim Q \\ \sim(P \vee Q) &= \sim P \wedge \sim Q \end{aligned} \quad \left. \right\} \quad \text{DeMorgan's laws} \quad (2.2)$$

$$\begin{aligned} P \wedge Q &= Q \wedge P \\ P \vee Q &= Q \vee P \end{aligned} \quad \left. \right\} \quad \text{Commutative laws} \quad (2.3)$$

$$\begin{aligned} P \wedge(Q \vee R) &= (P \wedge Q) \vee (P \wedge R) \\ P \vee(Q \wedge R) &= (P \vee Q) \wedge (P \vee R) \end{aligned} \quad \left. \right\} \quad \text{Distributive laws} \quad (2.4)$$

$$\begin{aligned} P \wedge(Q \wedge R) &= (P \wedge Q) \wedge R \\ P \vee(Q \vee R) &= (P \vee Q) \vee R \end{aligned} \quad \left. \right\} \quad \text{Associative laws} \quad (2.5)$$

Notice how the distributive law  $P \wedge(Q \vee R) = (P \wedge Q) \vee (P \wedge R)$  has the same structure as the distributive law  $p \cdot (q + r) = p \cdot q + p \cdot r$  from algebra. Concerning the associative laws, the fact that  $P \wedge(Q \wedge R) = (P \wedge Q) \wedge R$  means that the position of the parentheses is irrelevant, and we can write this as  $P \wedge Q \wedge R$  without ambiguity. Similarly, we may drop the parentheses in an expression such as  $P \vee(Q \vee R)$ .

But parentheses are essential when there is a mix of  $\wedge$  and  $\vee$ , as in  $P \vee(Q \wedge R)$ . Indeed,  $P \vee(Q \wedge R)$  and  $(P \vee Q) \wedge R$  are **not** logically equivalent. (See Exercise 13 for Section 2.6, below.)

## Exercises for Section 2.6

**A.** Use truth tables to show that the following statements are logically equivalent.

- |  |   |
|--|---|
| 1. $P \wedge(Q \vee R) = (P \wedge Q) \vee (P \wedge R)$ | 5. $\sim(P \vee Q \vee R) = (\sim P) \wedge(\sim Q) \wedge(\sim R)$                 |
| 2. $P \vee(Q \wedge R) = (P \vee Q) \wedge (P \vee R)$   | 6. $\sim(P \wedge Q \wedge R) = (\sim P) \vee(\sim Q) \vee(\sim R)$                 |
| 3. $P \Rightarrow Q = (\sim P) \vee Q$                   | 7. $P \Rightarrow Q = (P \wedge \sim Q) \Rightarrow (Q \wedge \sim Q)$              |
| 4. $\sim(P \vee Q) = (\sim P) \wedge(\sim Q)$            | 8. $\sim P \Leftrightarrow Q = (P \Rightarrow \sim Q) \wedge(\sim Q \Rightarrow P)$ |

**B.** Decide whether or not the following pairs of statements are logically equivalent.

- |  |  |
|--|--|
| 9. $P \wedge Q$ and $\sim(\sim P \vee \sim Q)$                             | 12. $\sim(P \Rightarrow Q)$ and $P \wedge \sim Q$                          |
| 10. $(P \Rightarrow Q) \vee R$ and $\sim((P \wedge \sim Q) \wedge \sim R)$ | 13. $P \vee(Q \wedge R)$ and $(P \vee Q) \wedge R$                         |
| 11. $(\sim P) \wedge(P \Rightarrow Q)$ and $\sim(Q \Rightarrow P)$         | 14. $P \wedge(Q \vee \sim Q)$ and $(\sim P) \Rightarrow (Q \wedge \sim Q)$ |

## 2.7 Quantifiers

Using symbols  $\wedge$ ,  $\vee$ ,  $\sim$ ,  $\Rightarrow$  and  $\Leftrightarrow$ , we can deconstruct many English sentences into a symbolic form. As we have seen, this symbolic form can help us understand the logical structure of sentences and how different sentences may actually have the same meaning (as in logical equivalence).

But these symbols alone are not powerful enough to capture the full meaning of every statement. To see why, imagine that we are dealing with an infinite set  $S = \{x_1, x_2, x_3, \dots\}$  of integers. Suppose we want to express the statement “*Every element of S is odd.*” We would have to write

$$P(x_1) \wedge P(x_2) \wedge P(x_3) \wedge P(x_4) \wedge \dots,$$

where  $P(x)$  is the open sentence “*x is odd.*” And if we wanted the express “*There is at least one element of S that is odd,*” we’d have to write

$$P(x_1) \vee P(x_2) \vee P(x_3) \vee P(x_4) \vee \dots.$$

The problem is that these expressions trail on forever.

To overcome this defect, we will introduce two new symbols  $\forall$  and  $\exists$ . The symbol  $\forall$  stands for the phrase “*for all*” and  $\exists$  stands for “*there exists*.” Thus the statement “*Every element of S is odd.*” is written symbolically as

$$\forall x \in S, P(x),$$

and “*There is at least one element of S that is odd,*” is written succinctly as

$$\exists x \in S, P(x),$$

These new symbols are called *quantifiers*.

**Definition 2.1** The symbols  $\forall$  and  $\exists$  are called **quantifiers**.

$\forall$  stands for the phrase “*For all*” or “*For every*,” or “*For each*,”

$\exists$  stands for the phrase “*There exists a*” or “*There is a*.”

Thus the statement

For every  $n \in \mathbb{Z}$ ,  $2n$  is even,

can be expressed in either of the following ways:

$\forall n \in \mathbb{Z}$ ,  $2n$  is even,

$\forall n \in \mathbb{Z}$ ,  $E(2n)$ .

Likewise, a statement such as

There exists a subset  $X$  of  $\mathbb{N}$  for which  $|X| = 5$ .

can be translated as

$$\exists X, (X \subseteq \mathbb{N}) \wedge (|X| = 5) \quad \text{or} \quad \exists X \subseteq \mathbb{N}, |X| = 5 \quad \text{or} \quad \exists X \in \mathcal{P}(\mathbb{N}), |X| = 5.$$

The symbols  $\forall$  and  $\exists$  are called quantifiers because they refer in some sense to the quantity (i.e., all or some) of the variable that follows them. The symbol  $\forall$  is called the **universal quantifier** and  $\exists$  is called the **existential quantifier**. Statements containing them are called **quantified statements**. A statement beginning with  $\forall$  is called a **universally quantified statement**, and one beginning with  $\exists$  is called an **existentially quantified statement**.

**Example 2.5** The following English statements are paired with their translations into symbolic form.

Every integer that is not odd is even.

$$\forall n \in \mathbb{Z}, \sim(n \text{ is odd}) \Rightarrow (n \text{ is even}), \quad \text{or} \quad \forall n \in \mathbb{Z}, \sim O(n) \Rightarrow E(n).$$

There is an integer that is not even.

$$\exists n \in \mathbb{Z}, \sim E(n).$$

For every real number  $x$ , there is a real number  $y$  for which  $y^3 = x$ .

$$\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, y^3 = x.$$

Given any two rational numbers  $a$  and  $b$ , the product  $ab$  is rational.

$$\forall a, b \in \mathbb{Q}, ab \in \mathbb{Q}.$$

Given a set  $S$  (such as, but not limited to,  $\mathbb{N}$ ,  $\mathbb{Z}$ ,  $\mathbb{Q}$ , etc.), a quantified statement of form  $\forall x \in S, P(x)$  is understood to be true if  $P(x)$  is true for every  $x \in S$ . If there is at least one  $x \in S$  for which  $P(x)$  is false, then  $\forall x \in S, P(x)$  is a false statement. Similarly,  $\exists x \in S, P(x)$  is true provided that  $P(x)$  is true for at least one element  $x \in S$ ; otherwise it is false. Thus each statement in Example 2.5 is true. Here are some examples of quantified statements that are false:

**Example 2.6** The following false quantified statements are paired with their translations.

Every integer is even.

$$\forall n \in \mathbb{Z}, E(n).$$

There is an integer  $n$  for which  $n^2 = 2$ .

$$\exists n \in \mathbb{Z}, n^2 = 2.$$

For every real number  $x$ , there is a real number  $y$  for which  $y^2 = x$ .

$$\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, y^2 = x.$$

Given any two rational numbers  $a$  and  $b$ , the number  $\sqrt{ab}$  is rational.

$$\forall a, b \in \mathbb{Q}, \sqrt{ab} \in \mathbb{Q}.$$

**Example 2.7** When a statement contains two quantifiers you must be very alert to their order, for reversing the order can change the meaning. Consider the following statement from Example 2.5.

$$\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, y^3 = x.$$

This statement is true, for no matter what number  $x$  is there exists a number  $y = \sqrt[3]{x}$  for which  $y^3 = x$ . Now reverse the order of the quantifiers to get the new statement

$$\exists y \in \mathbb{R}, \forall x \in \mathbb{R}, y^3 = x.$$

This new statement says that there exists a particular number  $y$  with the property that  $y^3 = x$  for *every* real number  $x$ . Since no number  $y$  can have this property, the statement is false. The two statements above have entirely different meanings.

Quantified statements are often misused in casual conversation. Maybe you've heard someone say "*All students do not pay full tuition.*" when they mean "*Not all students pay full tuition.*" While the mistake is perhaps marginally forgivable in casual conversation, it must never be made in a mathematical context. Do not say "*All integers are not even.*" because that means there are no even integers. Instead, say "*Not all integers are even.*"

## Exercises for Section 2.7

Write the following as English sentences. Say whether they are true or false.

- |   |   |
|---|---|
| 1. $\forall x \in \mathbb{R}, x^2 > 0$  | 6. $\exists n \in \mathbb{N}, \forall X \in \mathcal{P}(\mathbb{N}),  X  < n$ |
| 2. $\forall x \in \mathbb{R}, \exists n \in \mathbb{N}, x^n \geq 0$           | 7. $\forall X \subseteq \mathbb{N}, \exists n \in \mathbb{Z},  X  = n$        |
| 3. $\exists a \in \mathbb{R}, \forall x \in \mathbb{R}, ax = x$               | 8. $\forall n \in \mathbb{Z}, \exists X \subseteq \mathbb{N},  X  = n$        |
| 4. $\forall X \in \mathcal{P}(\mathbb{N}), X \subseteq \mathbb{R}$            | 9. $\forall n \in \mathbb{Z}, \exists m \in \mathbb{Z}, m = n + 5$            |
| 5. $\forall n \in \mathbb{N}, \exists X \in \mathcal{P}(\mathbb{N}),  X  < n$ | 10. $\exists m \in \mathbb{Z}, \forall n \in \mathbb{Z}, m = n + 5$           |

## 2.8 More on Conditional Statements

It is time to address a very important point about conditional statements that contain variables. To motivate this, let's return to the following example concerning integers  $x$ :

$$(x \text{ is a multiple of } 6) \Rightarrow (x \text{ is even}).$$

As noted earlier, since every multiple of 6 is even, this is a true statement no matter what integer  $x$  is. We could even underscore this fact by writing this true statement as

$$\forall x \in \mathbb{Z}, (x \text{ is a multiple of } 6) \Rightarrow (x \text{ is even}).$$

But now switch things around to get the different statement

$$(x \text{ is even}) \Rightarrow (x \text{ is a multiple of } 6).$$

This is true for some values of  $x$  such as  $-6, 12, 18$ , etc., but false for others (such as  $2, 4$ , etc.). Thus we do not have a statement, but rather an open sentence. (Recall from Section 2.1 that an *open sentence* is a sentence whose truth value depends on the value of a certain variable or variables.) However, by putting a universal quantifier in front we get

$$\forall x \in \mathbb{Z}, (x \text{ is even}) \Rightarrow (x \text{ is a multiple of } 6),$$

which is definitely false, so this new expression is a statement, *not an open sentence*. In general, given any two open sentences  $P(x)$  and  $Q(x)$  about integers  $x$ , the expression  $\forall x \in \mathbb{Z}, P(x) \Rightarrow Q(x)$  is either true or false, so it is a statement, not an open sentence.

Now we come to the very important point. In mathematics, whenever  $P(x)$  and  $Q(x)$  are open sentences concerning elements  $x$  in some set  $S$  (depending on context), an expression of form  $P(x) \Rightarrow Q(x)$  is understood to be the *statement*  $\forall x \in S, P(x) \Rightarrow Q(x)$ . In other words, if a conditional statement is not explicitly quantified then there is an implied universal quantifier in front of it. This is done because statements of the form  $\forall x \in S, P(x) \Rightarrow Q(x)$  are so common in mathematics that we would get tired of putting the  $\forall x \in S$  in front of them.

Thus the following sentence is a true statement (as it is true for all  $x$ ).

If  $x$  is a multiple of 6, then  $x$  is even.

Likewise, the next sentence is a false statement (as it is not true for all  $x$ ).

If  $x$  is even, then  $x$  is a multiple of 6.

This leads to the following significant interpretation of a conditional statement, which is more general than (but consistent with) its definition in Section 2.3.

**Definition 2.2** If  $P$  and  $Q$  are statements or open sentences, then

“If  $P$ , then  $Q$ ,”

is a statement. This statement is true if it’s impossible for  $P$  to be true while  $Q$  is false. It is false if there is at least one instance in which  $P$  is true but  $Q$  is false.

Thus the following are **true** statements:

If  $x \in \mathbb{R}$ , then  $x^2 + 1 > 0$ .

If a function  $f$  is differentiable on  $\mathbb{R}$ , then  $f$  is continuous on  $\mathbb{R}$ .

Likewise, the following are **false** statements:

If  $p$  is a prime number, then  $p$  is odd. (2 is prime.)

If  $f$  is a rational function, then  $f$  has an asymptote. ( $x^2$  is rational.)

## 2.9 Translating English to Symbolic Logic

In writing (and reading) proofs of theorems, we must always be alert to the logical structure and meanings of the sentences. Sometimes it is necessary or helpful to parse them into expressions involving logic symbols. This may be done mentally or on scratch paper, or occasionally even explicitly within the body of a proof. The purpose of this section is to give you sufficient practice in translating English sentences into symbolic form so that you can better understand their logical structure. Here are some examples:

**Example 2.8** Consider the Mean Value Theorem from Calculus:

If  $f$  is continuous on the interval  $[a, b]$  and differentiable on  $(a, b)$ , then there is a number  $c \in (a, b)$  for which  $f'(c) = \frac{f(b)-f(a)}{b-a}$ .

Here is a translation to symbolic form:

$$\left( (f \text{ cont. on } [a, b]) \wedge (f \text{ is diff. on } (a, b)) \right) \Rightarrow \left( \exists c \in (a, b), f'(c) = \frac{f(b)-f(a)}{b-a} \right).$$

**Example 2.9** Consider Goldbach's conjecture, from Section 2.1:

Every even integer greater than 2 is the sum of two primes.

This can be translated in the following ways, where  $P$  is the set of prime numbers and  $S = \{4, 6, 8, 10, \dots\}$  is the set of even integers greater than 2.

$$(n \in S) \Rightarrow (\exists p, q \in P, n = p + q)$$

$$\forall n \in S, \exists p, q \in P, n = p + q$$

These translations of Goldbach's conjecture illustrate an important point. The first has the basic structure  $(n \in S) \Rightarrow Q(n)$  and the second has structure  $\forall n \in S, Q(n)$ , yet they have exactly the same meaning. This is significant. Every universally quantified statement can be expressed as a conditional statement.

**Fact 2.2** Suppose  $S$  is a set and  $Q(x)$  is a statement about  $x$  for each  $x \in S$ . The following statements mean the same thing:

$$\forall x \in S, Q(x)$$

$$(x \in S) \Rightarrow Q(x).$$

This fact is significant because so many theorems have the form of a conditional statement. (The Mean Value Theorem is an example.) In proving a theorem we have to think carefully about what it says. Sometimes a theorem will be expressed as a universally quantified statement, but it will be more convenient to think of it as a conditional statement. Understanding the above fact allows us to switch between the two forms.

The section closes with some final points. In translating a statement, be attentive to its intended meaning. Don't jump into, for example, automatically replacing every "and" with  $\wedge$  and "or" with  $\vee$ . An example:

At least one of the integers  $x$  and  $y$  is even.

Don't be led astray by the presence of the word "and." The meaning of the statement is that one or both of the numbers is even, so it should be translated with "or," not "and":

$$(x \text{ is even}) \vee (y \text{ is even}).$$

Finally, the logical meaning of "but" can be captured by "and." The sentence "*The integer  $x$  is even, but the integer  $y$  is odd,*" is translated as

$$(x \text{ is even}) \wedge (y \text{ is odd}).$$

---

### Exercises for Section 2.9

Translate each of the following sentences into symbolic logic.

1. If  $f$  is a polynomial and its degree is greater than 2, then  $f'$  is not constant.
  2. The number  $x$  is positive but the number  $y$  is not positive.
  3. If  $x$  is prime, then  $\sqrt{x}$  is not a rational number.
  4. For every prime number  $p$  there is another prime number  $q$  with  $q > p$ .
  5. For every positive number  $\varepsilon$ , there is a positive number  $\delta$  for which  $|x - a| < \delta$  implies  $|f(x) - f(a)| < \varepsilon$ .
  6. For every positive number  $\varepsilon$  there is a positive number  $M$  for which  $|f(x) - b| < \varepsilon$ , whenever  $x > M$ .
  7. There exists a real number  $a$  for which  $a + x = x$  for every real number  $x$ .
  8. I don't eat anything that has a face.
  9. If  $x$  is a rational number and  $x \neq 0$ , then  $\tan(x)$  is not a rational number.
  10. If  $\sin(x) < 0$ , then it is not the case that  $0 \leq x \leq \pi$ .
  11. There is a Providence that protects idiots, drunkards, children and the United States of America. (Otto von Bismarck)
  12. You can fool some of the people all of the time, and you can fool all of the people some of the time, but you can't fool all of the people all of the time. (Abraham Lincoln)
  13. Everything is funny as long as it is happening to somebody else. (Will Rogers)
- 

### 2.10 Negating Statements

Given a statement  $R$ , the statement  $\sim R$  is called the **negation** of  $R$ . If  $R$  is a complex statement, then it is often the case that its negation  $\sim R$  can be written in a simpler or more useful form. The process of finding this form is called **negating  $R$** . In proving theorems it is often necessary to negate certain statements. We now investigate how to do this.

We have already examined part of this topic. **DeMorgan's laws**

$$\sim(P \wedge Q) = (\sim P) \vee (\sim Q) \quad (2.6)$$

$$\sim(P \vee Q) = (\sim P) \wedge (\sim Q) \quad (2.7)$$

(from Section 2.6) can be viewed as rules that tell us how to negate the statements  $P \wedge Q$  and  $P \vee Q$ . Here are some examples that illustrate how DeMorgan's laws are used to negate statements involving "and" or "or."

**Example 2.10** Consider negating the following statement.

$R$  : You can solve it by factoring or with the quadratic formula.

Now,  $R$  means (You can solve it by factoring)  $\vee$  (You can solve it with Q.F.), which we will denote as  $P \vee Q$ . The negation of this is

$$\sim(P \vee Q) = (\sim P) \wedge (\sim Q).$$

Therefore, in words, the negation of  $R$  is

$\sim R$  : You can't solve it by factoring and you can't solve it with the quadratic formula.

Maybe you can find  $\sim R$  without invoking DeMorgan's laws. That is good; you have internalized DeMorgan's laws and are using them unconsciously.

**Example 2.11** We will negate the following sentence.

$R$  : The numbers  $x$  and  $y$  are both odd.

This statement means ( $x$  is odd)  $\wedge$  ( $y$  is odd), so its negation is

$$\begin{aligned}\sim((x \text{ is odd}) \wedge (y \text{ is odd})) &= \sim(x \text{ is odd}) \vee \sim(y \text{ is odd}) \\ &= (x \text{ is even}) \vee (y \text{ is even}).\end{aligned}$$

Therefore the negation of  $R$  can be expressed in the following ways:

$\sim R$  : The number  $x$  is even or the number  $y$  is even.

$\sim R$  : At least one of  $x$  and  $y$  is even.

Now let's move on to a slightly different kind of problem. It's often necessary to find the negations of quantified statements. For example, consider  $\sim(\forall x \in \mathbb{N}, P(x))$ . Reading this in words, we have the following:

It is not the case that  $P(x)$  is true for all natural numbers  $x$ .

This means  $P(x)$  is false for at least one  $x$ . In symbols, this is  $\exists x \in \mathbb{N}, \sim P(x)$ . Thus  $\sim(\forall x \in \mathbb{N}, P(x)) = \exists x \in \mathbb{N}, \sim P(x)$ . Similarly, you can reason out that  $\sim(\exists x \in \mathbb{N}, P(x)) = \forall x \in \mathbb{N}, \sim P(x)$ . In general:

$$\sim(\forall x \in S, P(x)) = \exists x \in S, \sim P(x), \tag{2.8}$$

$$\sim(\exists x \in S, P(x)) = \forall x \in S, \sim P(x). \tag{2.9}$$

Be sure that you *understand* these two logical equivalences. They conform to our everyday use of language, but they pin down the meaning in a mathematically precise way.

**Example 2.12** Consider negating the following statement.

$R$  : The square of every real number is non-negative.

Symbolically,  $R$  can be expressed as  $\forall x \in \mathbb{R}, x^2 \geq 0$ , and thus its negation is  $\sim(\forall x \in \mathbb{R}, x^2 \geq 0) = \exists x \in \mathbb{R}, \sim(x^2 \geq 0) = \exists x \in \mathbb{R}, x^2 < 0$ . In words, this is

$\sim R$  : There exists a real number whose square is negative.

Observe that  $R$  is true and  $\sim R$  is false. Maybe you can get  $\sim R$  immediately, without using Equation (2.8) as we did above. If so, that is good; if not, you should be there soon.

**Example 2.13** If a statement has multiple quantifiers, negating it involves several iterations of Equations (2.8) and (2.9). Consider the following:

$S$  : For every real number  $x$  there is a real number  $y$  for which  $y^3 = x$ .

This statement asserts any real number  $x$  has a cube root  $y$ , so it's true. Symbolically  $S$  can be expressed as

$$\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, y^3 = x.$$

Let's work out the negation of this statement.

$$\begin{aligned} \sim(\forall x \in \mathbb{R}, \exists y \in \mathbb{R}, y^3 = x) &= \exists x \in \mathbb{R}, \sim(\exists y \in \mathbb{R}, y^3 = x) \\ &= \exists x \in \mathbb{R}, \forall y \in \mathbb{R}, \sim(y^3 = x) \\ &= \exists x \in \mathbb{R}, \forall y \in \mathbb{R}, y^3 \neq x. \end{aligned}$$

Thus the negation is a (false) statement that can be written in either of the following ways.

$\sim S$  : There is a real number  $x$  such that for all real numbers  $y$ ,  $y^3 \neq x$ .

$\sim S$  : There is a real number  $x$  for which  $y^3 \neq x$  for all real numbers  $y$ .

In writing proofs you will occasionally have to negate a conditional statement  $P \Rightarrow Q$ . The remainder of this section describes how to do this. To begin, look at the expression  $\sim(P \Rightarrow Q)$ , which literally says " $P \Rightarrow Q$  is false." You know from the truth table for  $\Rightarrow$  that the only way that  $P \Rightarrow Q$  can be false is if  $P$  is true and  $Q$  is false. Therefore

$$\sim(P \Rightarrow Q) = P \wedge \sim Q. \quad (2.10)$$

(In fact, in Exercise 12 of Section 2.6, you used a truth table to verify that these two statements are indeed logically equivalent.)

**Example 2.14** Negate the following statement about a particular (i.e., constant) number  $a$ .

$R$  : If  $a$  is odd then  $a^2$  is odd.

Using Equation (2.10), we get the following negation.

$\sim R$  :  $a$  is odd and  $a^2$  is not odd.

**Example 2.15** This example is like the previous one, but the constant  $a$  is replaced by a variable  $x$ . We will negate the following statement.

$R$  : If  $x$  is odd then  $x^2$  is odd.

As in Section 2.8, we interpret this as the universally quantified statement

$R$  :  $\forall x \in \mathbb{Z}, (x \text{ odd}) \Rightarrow (x^2 \text{ odd})$ .

By Equations (2.8) and (2.10), we get the following negation for  $R$ .

$$\begin{aligned}\sim (\forall x \in \mathbb{Z}, (x \text{ odd}) \Rightarrow (x^2 \text{ odd})) &= \exists x \in \mathbb{Z}, \sim ((x \text{ odd}) \Rightarrow (x^2 \text{ odd})) \\ &= \exists x \in \mathbb{Z}, (x \text{ odd}) \wedge \sim (x^2 \text{ odd}).\end{aligned}$$

Translating back into words, we have

$\sim R$  : There is an odd integer  $x$  whose square is not odd.

Notice that  $R$  is true and  $\sim R$  is false.

The above Example 2.15 showed how to negate a conditional statement  $P(x) \Rightarrow Q(x)$ . This type of problem can sometimes be embedded in more complex negation. See Exercise 5 below (and its solution).

## Exercises for Section 2.10

Negate the following sentences.

1. The number  $x$  is positive, but the number  $y$  is not positive.
2. If  $x$  is prime, then  $\sqrt{x}$  is not a rational number.
3. For every prime number  $p$ , there is another prime number  $q$  with  $q > p$ .
4. For every positive number  $\varepsilon$ , there is a positive number  $\delta$  such that  $|x - a| < \delta$  implies  $|f(x) - f(a)| < \varepsilon$ .
5. For every positive number  $\varepsilon$ , there is a positive number  $M$  for which  $|f(x) - b| < \varepsilon$  whenever  $x > M$ .
6. There exists a real number  $a$  for which  $a + x = x$  for every real number  $x$ .
7. I don't eat anything that has a face.

8. If  $x$  is a rational number and  $x \neq 0$ , then  $\tan(x)$  is not a rational number.
  9. If  $\sin(x) < 0$ , then it is not the case that  $0 \leq x \leq \pi$ .
  10. If  $f$  is a polynomial and its degree is greater than 2, then  $f'$  is not constant.
  11. You can fool all of the people all of the time.
  12. Whenever I have to choose between two evils, I choose the one I haven't tried yet. (Mae West)
- 

## 2.11 Logical Inference

Suppose we know that a conditional statement  $P \Rightarrow Q$  is true. This tells us that whenever  $P$  is true,  $Q$  will also be true. By itself,  $P \Rightarrow Q$  being true does not tell us that either  $P$  or  $Q$  is true (they could both be false, or  $P$  could be false and  $Q$  true). But if in addition we happen to know that  $P$  is true, then  $Q$  must be true. This is called a **logical inference**: From two true statements we infer that a third statement is true. In essence, statements  $P \Rightarrow Q$  and  $P$  are “added together” to get  $Q$ . We can indicate this by stacking  $P \Rightarrow Q$  and  $P$  one atop the other with a line separating them from  $Q$ . The intended meaning is that  $P \Rightarrow Q$  combined with  $P$  produces  $Q$ .

$$\begin{array}{c} P \Rightarrow Q \\ P \\ \hline Q \end{array}$$

This is a very frequently-used pattern of thought. (In fact, it is exactly the pattern we used in the example on page 34.) This rule even has a name. It is called the **modus ponens** rule.

Two other logical inferences, called **modus tollens** and **elimination** are listed below. In each case you should convince yourself (based on your knowledge of the relevant truth tables) that the truth of the statements above the line forces the statement below the line to be true.

MODUS PONENS

$$\begin{array}{c} P \Rightarrow Q \\ P \\ \hline \end{array}$$

MODUS TOLLENS

$$\begin{array}{c} P \Rightarrow Q \\ \sim Q \\ \hline \end{array}$$

ELIMINATION

$$\begin{array}{c} P \vee Q \\ \sim P \\ \hline \end{array}$$

$$Q$$

$$\sim P$$

$$Q$$

It is important to internalize these rules. (You surely already use at least modus ponens and elimination in daily life anyway.) But you need not

remember their names; few working mathematicians can recall the names, though they use the rules constantly. The names are not important, but the rules are.

Three additional logical inferences are listed below. The first states the obvious fact that if  $P$  and  $Q$  are both true, then so is the statement  $P \wedge Q$ . On the other hand,  $P \wedge Q$  being true forces  $P$  (also  $Q$ ) to be true. Finally, if  $P$  is true, then  $P \vee Q$  must be true, no matter what statement  $Q$  is.

$$\begin{array}{c} P \\ Q \\ \hline P \wedge Q \end{array} \qquad \begin{array}{c} P \wedge Q \\ \hline P \end{array} \qquad \begin{array}{c} P \\ \hline P \vee Q \end{array}$$

These inferences are so intuitively obvious that they scarcely need to be mentioned. However, they represent certain patterns of reasoning that we will frequently apply to sentences in proofs, so we should be cognizant of the fact that we are using them.

## 2.12 An Important Note

It is important to be aware of the reasons that we study logic. There are three very significant reasons. First, the truth tables we studied tell us the exact meanings of the words such as “and,” “or,” “not” and so on. For instance, whenever we use or read the “*If..., then*” construction in a mathematical context, logic tells us exactly what is meant. Second, the rules of inference provide a system in which we can produce new information (statements) from known information. Finally, logical rules such as DeMorgan’s laws help us correctly change certain statements into (potentially more useful) statements with the same meaning. Thus, logic helps us understand the meanings of statements, and it also produces new meaningful statements.

Logic is the glue that holds strings of statements together and pins down the exact meaning of certain key phrases such as the “*If..., then*” or “*For all*” constructions. Logic is the common language that all mathematicians use, so we must have a firm grip on it in order to write and understand mathematics.

But despite its fundamental role, logic’s place is in the background of what we do, not the forefront. From here on, the beautiful symbols  $\wedge$ ,  $\vee$ ,  $\Rightarrow$ ,  $\Leftrightarrow$ ,  $\sim$ ,  $\forall$  and  $\exists$  are rarely written. But we are aware of their meanings constantly. When reading or writing a sentence involving mathematics we parse it with these symbols, either mentally or on scratch paper, so as to understand the true and unambiguous meaning.

# CHAPTER 3

---

## Counting

---

**I**t may seem peculiar that a college-level text has a chapter on counting. At its most basic level, counting is a process of pointing to each object in a collection and calling off “*one, two, three,...*” until the quantity of objects is determined. How complex could that be? Actually, counting can become quite subtle, and in this chapter we explore some of its more sophisticated aspects. Our goal is still to answer the question “*How many?*” but we introduce mathematical techniques that bypass the actual process of counting individual objects. Sets play a big role in our discussions because the things we need to count are often naturally grouped together into a set. The concept of a *list* is also extremely useful.

### 3.1 Lists

A **list** is an ordered sequence of objects. A list is denoted by an opening parenthesis, followed by the objects, separated by commas, followed by a closing parenthesis. For example  $(a, b, c, d, e)$  is a list consisting of the first five letters of the English alphabet, in order. The objects  $a, b, c, d, e$  are called the **entries** of the list; the first entry is  $a$ , the second is  $b$ , and so on. If the entries are rearranged we get a different list, so, for instance,

$$(a, b, c, d, e) \neq (b, a, c, d, e).$$

A list is somewhat like a set, but instead of being a mere collection of objects, the entries of a list have a definite *order*. For sets we have

$$\{a, b, c, d, e\} = \{b, a, c, d, e\},$$

but—as noted above—the analogous equality for lists does not hold.

Unlike sets, lists can have repeated entries. Thus  $(5, 3, 5, 4, 3, 3)$  is a perfectly acceptable list, as is  $(S, O, S)$ . The **length** of a list is its number of entries. So  $(5, 3, 5, 4, 3, 3)$  has length six, and  $(S, O, S)$  has length three.

For more examples,  $(a, 15)$  is a list of length two. And  $(0, (0, 1, 1))$  is a list of length two whose second entry is a list of length three. Two lists are **equal** if they have exactly the same entries in exactly the same positions. Thus equal lists have the same number of entries. If two lists have different lengths, then they can not be equal. Thus  $(0, 0, 0, 0, 0, 0) \neq (0, 0, 0, 0, 0)$ . Also

$$(g, r, o, c, e, r, y, l, i, s, t) \quad \neq \quad \left( \begin{array}{c} \text{bread} \\ \text{milk} \\ \text{eggs} \\ \text{bananas} \\ \text{coffee} \end{array} \right)$$

because the list on the left has length eleven but the list on the right has just one entry (a piece of paper with some words on it).

There is one very special list which has no entries at all. It is called the **empty list** and is denoted  $()$ . It is the only list whose length is zero.

For brevity we often write lists without parentheses, or even commas. For instance, we may write  $(S, O, S)$  as *SOS* if there is no risk of confusion. But be alert that doing this can lead to ambiguity: writing  $(9, 10, 11)$  as  $9\ 10\ 11$  may cause us to confuse it with  $(9, 1, 0, 1, 1)$ . Here it's best to retain the parenthesis/comma notation or at least write the list as  $9, 10, 11$ . A list of symbols written without parentheses and commas is called a **string**.

The process of tossing a coin ten times may be described by a string such as *HHTHTTTHHT*. Tossing it twice could lead to any of the outcomes *HH*, *HT*, *TH* or *TT*. Tossing it zero times is described by the empty list  $()$ .

Imagine rolling a dice five times and recording the outcomes. This might be described by the list  $(\square, \blacksquare, \blacksquare, \square, \blacksquare)$ , meaning that you rolled  $\square$  first, then  $\blacksquare$ , then  $\blacksquare$ , etc. We might abbreviate this list as  $\square\blacksquare\blacksquare\square\blacksquare$ , or  $3, 5, 3, 1, 6$ .

Now imagine rolling a pair of dice, one white and one black. A typical outcome might be modeled as a set like  $\{\square, \blacksquare\}$ . Rolling the pair six times might be described with a list of six such outcomes:

$$(\{\square, \blacksquare\}, \{\square, \blacksquare\}, \{\square, \blacksquare\}, \{\square, \blacksquare\}, \{\square, \blacksquare\}, \{\square, \blacksquare\}).$$

We might abbreviate this list as  $\square\blacksquare\blacksquare\blacksquare\blacksquare\blacksquare$ .

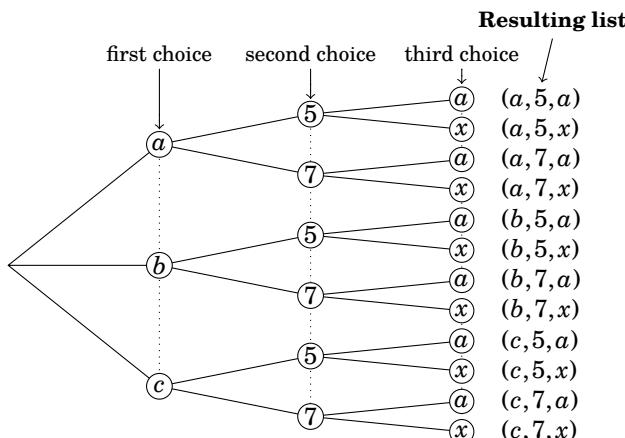
We study lists because many real-world phenomena can be described and understood in terms of them. Your phone number can be identified as a list of ten digits. (Order is essential, for rearranging the digits can produce a different phone number.) A *byte* is another important example of a list. A byte is simply a length-eight list of 0's and 1's. The world of information technology revolves around bytes. And the examples above show that multi-step processes (such as rolling a dice twice) can be modeled as lists.

We now explore methods of counting or enumerating lists and processes.

### 3.2 The Multiplication Principle

Many practical problems involve counting the number of possible lists that satisfy some condition or property.

For example, suppose we make a list of length three having the property that the first entry must be an element of the set  $\{a, b, c\}$ , the second entry must be in  $\{5, 7\}$  and the third entry must be in  $\{a, x\}$ . Thus  $(a, 5, a)$  and  $(b, 5, a)$  are two such lists. How many such lists are there all together? To answer this question, imagine making the list by selecting the first entry, then the second and finally the third. This is described in Figure 3.1. The choices for the first list entry are  $a, b$  or  $c$ , and the left of the diagram branches out in three directions, one for each choice. Once this choice is made there are two choices (5 or 7) for the second entry, and this is described graphically by two branches from each of the three choices for the first entry. This pattern continues for the choice for the third entry, which is either  $a$  or  $x$ . Thus, in the diagram there are  $3 \cdot 2 \cdot 2 = 12$  paths from left to right, each corresponding to a particular choice for each entry in the list. The corresponding lists are tallied at the far-right end of each path. So, to answer our original question, there are 12 possible lists with the stated properties.



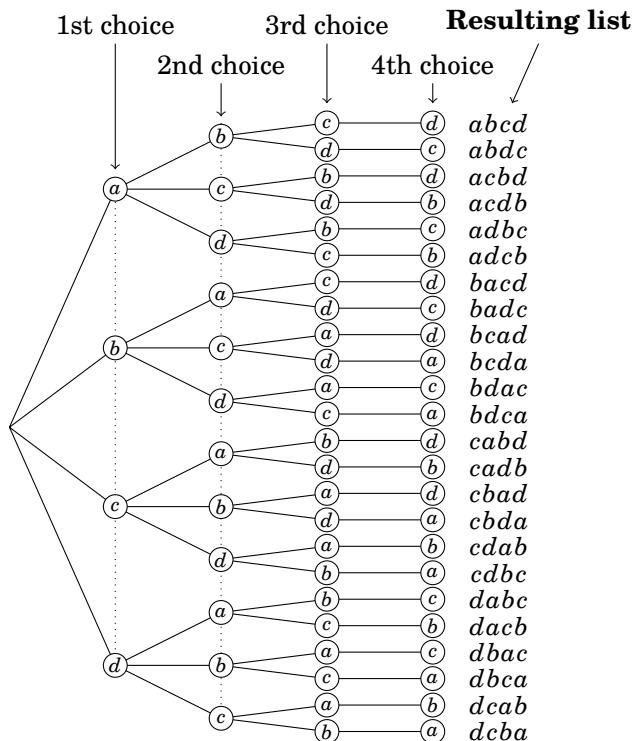
**Figure 3.1.** Constructing lists of length 3

In the above example there are 3 choices for the first entry, 2 choices for the second entry, and 2 for the third, and the total number of possible lists is the product of choices  $3 \cdot 2 \cdot 2 = 12$ . This kind of reasoning is an instance of what we will call the *multiplication principle*. We will do one more example before stating this important idea.

Consider making a list of length 4 from the four letters  $\{a, b, c, d\}$ , where the list is not allowed to have a repeated letter. For example,  $abcd$  and  $cadb$  are allowed, but  $aabc$  and  $cacb$  are not allowed. How many such lists are there?

Let's analyze this question with a tree representing the choices we have for each list entry. In making such a list we could start with the first entry: we have 4 choices for it, namely  $a, b, c$  or  $d$ , and the left side of the tree branches out to each of these choices. But once we've chosen a letter for the first entry, we can't use that letter in the list again, so there are only 3 choices for the second entry. And once we've chosen letters for the first and second entries we can't use these letters in the third entry, so there are just 2 choices for it. By the time we get to the fourth entry we are forced to use whatever letter we have left; there is only 1 choice.

The situation is described fully in the below tree showing how to make all allowable lists by choosing 4 letters for the first entry, 3 for the second entry, 2 for the third entry and 1 for the fourth entry. We see that the total number of lists is the product  $4 \cdot 3 \cdot 2 \cdot 1 = 24$ .



**Figure 3.2.** Constructing lists from letters in  $\{a, b, c, d\}$ , without repetition.

These trees show that the number of lists constructible by some specified process equals the product of the numbers of choices for each list entry. We summarize this kind of reasoning as an important fact.

**Fact 3.1 (Multiplication Principle)** Suppose in making a list of length  $n$  there are  $a_1$  possible choices for the first entry,  $a_2$  possible choices for the second entry,  $a_3$  possible choices for the third entry, and so on. Then the total number of different lists that can be made this way is the product  $a_1 \cdot a_2 \cdot a_3 \cdots a_n$ .

In using the multiplication principle you **do not** need to draw a tree with  $a_1 \cdot a_2 \cdots a_n$  branches. Just multiply the numbers!

**Example 3.1** A standard license plate consists of three letters followed by four numbers. For example, *JRB-4412* and *MMX-8901* are two standard license plates. How many different standard license plates are possible?

**Solution:** A license plate such as *JRB-4412* corresponds to a length-7 list (*J,R,B,4,4,1,2*), so we just need to count how many such lists are possible. We use the multiplication principle. There are  $a_1 = 26$  possibilities (one for each letter of the alphabet) for the first entry of the list. Similarly, there are  $a_2 = 26$  possibilities for the second entry and  $a_3 = 26$  possibilities for the third. There are  $a_4 = 10$  possibilities for the fourth entry. Likewise  $a_5 = a_6 = a_7 = 10$ . So there is a total of  $a_1 \cdot a_2 \cdot a_3 \cdot a_4 \cdot a_5 \cdot a_6 \cdot a_7 = 26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 175,760,000$  possible standard license plates.

**Example 3.2** In ordering a café latte, you have a choice of whole, skim or soy milk; small, medium or large; and either one or two shots of espresso. How many choices do you have in ordering one drink?

**Solution:** Your choice is modeled by a list of form (milk, size, shots). There are 3 choices for the first entry, 3 for the second and 2 for the third. By the multiplication principle, the number of choices is  $3 \cdot 3 \cdot 2 = 18$ .

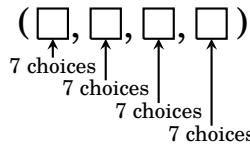
There are two types of list-counting problems. On one hand, there are situations in which list entries can be repeated, as in license plates or telephone numbers. The sequence *CCX-4144* is a perfectly valid license plate in which the symbols *C* and *4* appear more than once. On the other hand, for some lists repeated symbols do not make sense or are not allowed, as in the (milk, size, shots) list from Example 3.2. We say *repetition is allowed* in the first type of list and *repetition is not allowed* in the second kind of list. (We will call a list in which repetition is not allowed a **non-repetitive list**.) The next example illustrates the difference.

**Example 3.3** Consider lists of length 4 made with symbols  $A, B, C, D, E, F, G$ .

- How many such lists are possible if repetition is allowed?
- How many such lists are possible if repetition is **not** allowed?
- How many are there if repetition is **not** allowed and the list has an  $E$ ?
- How many are there if repetition is allowed and the list has an  $E$ ?

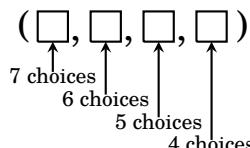
**Solutions:**

- Imagine the list as containing four boxes that we fill with selections from the letters  $A, B, C, D, E, F$  and  $G$ , as illustrated below.



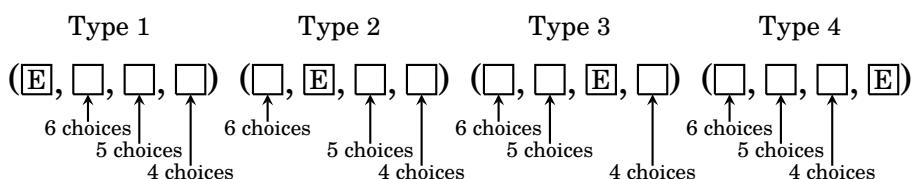
We have 7 choices in filling each box. The multiplication principle says the total number of lists that can be made this way is  $7 \cdot 7 \cdot 7 \cdot 7 = 2401$ .

- This problem is the same as the previous one except that repetition is not allowed. We have seven choices for the first box, but once it is filled we can no longer use the symbol that was placed in it. Hence there are only six possibilities for the second box. Once the second box has been filled we have used up two of our letters, and there are only five left to choose from in filling the third box. Finally, when the third box is filled we have only four possible letters for the last box.



Thus there are  $7 \cdot 6 \cdot 5 \cdot 4 = 840$  lists in which repetition does not occur.

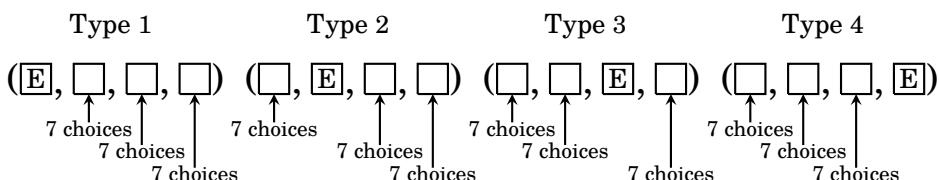
- We are asked to count the length-4 lists in which repetition is not allowed and the symbol  $E$  must appear somewhere in the list. Thus  $E$  occurs once and only once in each list. Let us divide these lists into four categories depending on whether the  $E$  occurs as the first, second, third or fourth entry. These four types of lists are illustrated below.



Consider lists of the first type, in which the  $E$  appears in the first entry. We have six remaining choices ( $A, B, C, D, F$  or  $G$ ) for the second entry, five choices for the third entry and four choices for the fourth entry. Hence there are  $6 \cdot 5 \cdot 4 = 120$  lists having an  $E$  in the first entry. As shown above, there are also  $6 \cdot 5 \cdot 4 = 120$  lists having an  $E$  in the second, third or fourth entry. So there are  $120 + 120 + 120 + 120 = 480$  lists with exactly one  $E$ .

- (d) Now we seek the number of length-4 lists where repetition is allowed and the list must contain an  $E$ . Here is our strategy: By Part (a) of this exercise there are  $7 \cdot 7 \cdot 7 \cdot 7 = 7^4 = 2401$  lists with repetition allowed. Obviously this is not the answer to our current question, for many of these lists contain no  $E$ . We will subtract from 2401 the number of lists that **do not** contain an  $E$ . In making a list that does not contain an  $E$ , we have six choices for each list entry (because we can choose any one of the six letters  $A, B, C, D, F$  or  $G$ ). Thus there are  $6 \cdot 6 \cdot 6 \cdot 6 = 6^4 = 1296$  lists without an  $E$ . So the answer to our question is that there are  $2401 - 1296 = 1105$  lists with repetition allowed that contain at least one  $E$ .

Before moving on from Example 3.3, let's address an important point. Perhaps you wondered if Part (d) could be solved in the same way as Part (c). Let's try doing it that way. We want to count the length-4 lists (repetition allowed) that contain at least one  $E$ . The following diagram is adapted from Part (c). The only difference is that there are now seven choices in each slot because we are allowed to repeat any of the seven letters.



We get a total of  $7^3 + 7^3 + 7^3 + 7^3 = 1372$  lists, an answer that is larger than the (correct) value of 1105 from our solution to Part (d) above. It is easy to see what went wrong. The list  $(E, E, A, B)$  is of type 1 *and* type 2, so it got counted *twice*. Similarly  $(E, E, C, E)$  is of type 1, 2 and 4, so it got counted three times. In fact, you can find many similar lists that were counted multiple times. In solving counting problems, we must always be careful to avoid this kind of double-counting or triple-counting, or worse.

The next section presents two new counting principles that codify the kind of thinking we used in parts (c) and (d) above. Combined with the multiplication principle, they solve complex counting problems in ways that avoid the pitfalls of double counting. But first, one more example of the multiplication principle highlights another pitfall to be alert to.

**Example 3.4** A non-repetitive list of length 5 is to be made from the symbols  $A, B, C, D, E, F, G$ . The first entry must be either a  $B, C$  or  $D$ , and the last entry must be a vowel. How many such lists are possible?

**Solution:** Start by making a list of five boxes. The first box must contain either  $B, C$  or  $D$ , so there are three choices for it.

$$(\square, \square, \square, \square, \square)$$

↑  
3 choices

Now there are 6 letters left for the remaining 4 boxes. The knee-jerk action is to fill them in, one at a time, using up an additional letter each time.

$$(\square, \square, \square, \square, \square)$$

↑  
3 choices  
↑  
6 choices  
↑  
5 choices  
↑  
4 choices

But when we get to the last box, there is a problem. It is supposed to contain a vowel, but for all we know we have already used up one or both vowels in the previous boxes. The multiplication principle breaks down because there is no way to tell how many choices there are for the last box.

The correct way to solve this problem is to fill in the first and last boxes (the ones that have restrictions) first.

$$(\square, \square, \square, \square, \square)$$

↑  
3 choices                      ↑  
                                  2 choices

Then fill the remaining middle boxes with the 5 remaining letters.

$$(\square, \square, \square, \square, \square)$$

↑  
3 choices  
↑  
5 choices  
↑  
4 choices  
↑  
3 choices                      ↑  
                                  2 choices

By the multiplication principle, there are  $3 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 360$  lists.

The new principles to be introduced in the next section are usually used in conjunction with the multiplication principle. So work a few exercises now to test your understanding of it.

---

### Exercises for Section 3.2

1. Consider lists made from the letters  $T, H, E, O, R, Y$ , with repetition allowed.
    - (a) How many length-4 lists are there?
    - (b) How many length-4 lists are there that begin with  $T$ ?
    - (c) How many length-4 lists are there that do not begin with  $T$ ?
  2. Airports are identified with 3-letter codes. For example, Richmond, Virginia has the code  $RIC$ , and Memphis, Tennessee has  $MEM$ . How many different 3-letter codes are possible?
  3. How many lists of length 3 can be made from the symbols  $A, B, C, D, E, F$  if...
    - (a) ... repetition is allowed.
    - (b) ... repetition is not allowed.
    - (c) ... repetition is not allowed and the list must contain the letter  $A$ .
    - (d) ... repetition is allowed and the list must contain the letter  $A$ .
  4. In ordering coffee you have a choice of regular or decaf; small, medium or large; here or to go. How many different ways are there to order a coffee?
  5. This problem involves 8-digit binary strings such as  $10011011$  or  $00001010$  (i.e., 8-digit numbers composed of 0's and 1's).
    - (a) How many such strings are there?
    - (b) How many such strings end in 0?
    - (c) How many such strings have 1's for their second and fourth digits?
    - (d) How many such strings have 1's for their second **or** fourth digits?
  6. You toss a coin, then roll a dice, and then draw a card from a 52-card deck. How many different outcomes are there? How many outcomes are there in which the dice lands on ? How many outcomes are there in which the dice lands on an odd number? How many outcomes are there in which the dice lands on an odd number and the card is a King?
  7. This problem concerns 4-letter codes made from the letters  $A, B, C, D, \dots, Z$ .
    - (a) How many such codes can be made?
    - (b) How many such codes have no two consecutive letters the same?
  8. A coin is tossed 10 times in a row. How many possible sequences of heads and tails are there?
  9. A new car comes in a choice of five colors, three engine sizes and two transmissions. How many different combinations are there?
  10. A dice is tossed four times in a row. There are many possible outcomes, such as or . How many different outcomes are possible?
-

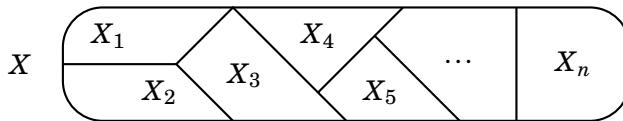
### 3.3 The Addition and Subtraction Principles

We now discuss two new counting principles, the addition and subtraction principles. Actually, they are not *entirely* new—you've used them intuitively for years. Here we give names to these two fundamental thought patterns, and phrase them in the language of sets. Doing this helps us recognize when we are using them, and, more importantly, it helps us see new situations in which they can be used.

The *addition principle* simply asserts that if a set can be broken into pieces, then the size of the set is the sum of the sizes of the pieces.

**Fact 3.2 (Addition Principle)**

Suppose a finite set  $X$  can be decomposed as a union  $X = X_1 \cup X_2 \cup \dots \cup X_n$ , where  $X_i \cap X_j = \emptyset$  whenever  $i \neq j$ . Then  $|X| = |X_1| + |X_2| + \dots + |X_n|$ .



In our first example we will rework an instance where we used the addition principle naturally, without comment: in Part (c) of Example 3.3.

**Example 3.5** How many length-4 non-repetitive lists can be made from the symbols  $A, B, C, D, E, F, G$ , if the list must contain an  $E$ ?

In Example 3.3 (c) our approach was to divide these lists into four types, depending on whether the  $E$  is in the first, second, third or fourth position.

Type 1	Type 2	Type 3	Type 4
$\boxed{E} \quad \square \quad \square \quad \square$ 6    5    4	$\square \quad \boxed{E} \quad \square \quad \square$ 6    5    4	$\square \quad \square \quad \boxed{E} \quad \square$ 6    5    4	$\square \quad \square \quad \square \quad \boxed{E}$ 6    5    4

Then we used the multiplication principle to count the lists of type 1. There are 6 choices for the second entry, 5 for the third, and 4 for the fourth. This is indicated above, where the number below a box is the number of choices we have for that position. The multiplication principle implies that there are  $6 \cdot 5 \cdot 4 = 120$  lists of type 1. Similarly there are  $6 \cdot 5 \cdot 4 = 120$  lists of types 2, 3, and 4.

$X_1$	$EABC$ $EACB$ $EBAC$ $\vdots$	$X_2$	$AEBC$ $AECB$ $BEAC$ $\vdots$	$X_3$	$ABEC$ $ACEB$ $BAEC$ $\vdots$	$X_4$	$ABCE$ $ACBE$ $BACE$ $\vdots$
-------	--	-------	--	-------	--	-------	--

We then used the addition principle intuitively, conceiving of the lists to be counted as the elements of a set  $X$ , broken up into parts  $X_1, X_2, X_3$  and  $X_4$ , which are the lists of types 1, 2, 3 and 4, respectively.

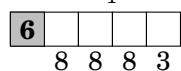
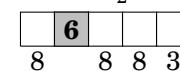
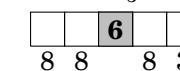
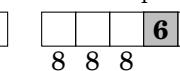
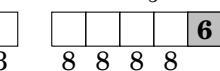
The addition principle says that the number of lists that contain an  $E$  is  $|X| = |X_1| + |X_2| + |X_3| + |X_4| = 120 + 120 + 120 + 120 = \mathbf{480}$ .

We use the addition principle when we need to count the things in some set  $X$ . If we can find a way to break  $X$  up as  $X = X_1 \cup X_2 \cup \dots \cup X_n$ , where each  $X_i$  is easier to count than  $X$ , then the addition principle gives an answer of  $|X| = |X_1| + |X_2| + |X_3| + \dots + |X_n|$ .

But for this to work the intersection of any two pieces  $X_i$  must be  $\emptyset$ , as stated in Fact 3.2. For instance, if  $X_1$  and  $X_2$  shared an element, then that element would be counted once in  $|X_1|$  and again in  $|X_2|$ , and we'd get  $|X| < |X_1| + |X_2| + \dots + |X_n|$ . (This is precisely the double counting issue mentioned after Example 3.3.)

**Example 3.6** How many **even** 5-digit numbers are there for which no digit is 0, and the digit 6 appears exactly once? For instance, 55634 and 16118 are such numbers, but not 63304 (has a 0), nor 63364 (too many 6's), nor 55637 (not even).

**Solution:** Let  $X$  be the set of all such numbers. The answer will be  $|X|$ , so our task is to find  $|X|$ . Put  $X = X_1 \cup X_2 \cup X_3 \cup X_4 \cup X_5$ , where  $X_i$  is the set of those numbers in  $X$  whose  $i$ th digit is 6, as diagramed below. Note  $X_i \cap X_j = \emptyset$  whenever  $i \neq j$  because the numbers in  $X_i$  have their 6 in a different position than the numbers in  $X_j$ . Our plan is to use the multiplication principle to compute each  $|X_i|$ , and follow this with the addition principle.

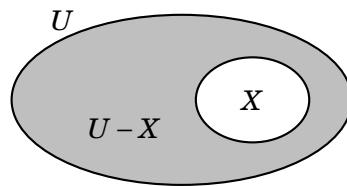
$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
				

The first digit of any number in  $X_1$  is 6, and the three digits following it can be any of the ten digits except 0 (not allowed) or 6 (already appears). Thus there are eight choices for each of three digits following the first 6. But because any number in  $X_1$  is even, its final digit must be one of 2, 4 or 8, so there are just three choices for this final digit. By the multiplication principle,  $|X_1| = 8 \cdot 8 \cdot 8 \cdot 3 = 1536$ . Likewise  $|X_2| = |X_3| = |X_4| = 8 \cdot 8 \cdot 8 \cdot 3 = 1536$ .

But  $X_5$  is slightly different because we do not choose the final digit, which is already 6. The multiplication principle gives  $|X_5| = 8 \cdot 8 \cdot 8 \cdot 8 = 4096$ .

The addition principle gives our final answer. The number of even 5-digit numbers with no 0's and one 6 is  $|X| = |X_1| + |X_2| + |X_3| + |X_4| + |X_5| = 1536 + 1536 + 1536 + 1536 + 4096 = \mathbf{10,240}$ .

Now we introduce our next counting method, the *subtraction principle*. To set it up, imagine that a set  $X$  is a subset of a universal set  $U$ , as shown on the right. The complement  $\bar{X} = U - X$  is shaded. Suppose we wanted to count the things in this shaded region. Surely this is the number of things in  $U$  minus the number of things in  $X$ , which is to say  $|\bar{X}| = |U| - |X|$ . That is the subtraction principle.



**Fact 3.3 (Subtraction Principle)**

If  $X$  is a subset of a finite set  $U$ , then  $|\bar{X}| = |U| - |X|$ .

In other words, if  $X \subseteq U$  then  $|U - X| = |U| - |X|$ .

The subtraction principle is used in situations where it is easier to count the things in some set  $U$  that we wish to *exclude* from consideration than it is to count those things that *are* included. We have seen this kind of thinking before. We quietly and naturally used it in part (d) of Example 3.3. For convenience we repeat that example now, casting it into the language of the subtraction principle.

**Example 3.7** How many length-4 lists can be made from the symbols  $A, B, C, D, E, F, G$  if the list has at least one  $E$ , and repetition is allowed?

**Solution:** Such a list might contain one, two, three or four  $E$ 's, which could occur in various positions. This is a fairly complex situation.

But it is very easy to count the set  $U$  of all lists of length 4 made from  $A, B, C, D, E, F, G$  if we don't care whether or not the lists have any  $E$ 's. The multiplication principle says  $|U| = 7 \cdot 7 \cdot 7 \cdot 7 = 2401$ .

It is equally easy to count the set  $X$  of those lists that *contain no*  $E$ 's. The multiplication principle says  $|X| = 6 \cdot 6 \cdot 6 \cdot 6 = 1296$ .

We are interested in those lists that have at least one  $E$ , and this is the set  $U - X$ . By the subtraction principle, the answer to our question is  $|U - X| = |U| - |X| = 2401 - 1296 = 1105$ .

As we continue with counting we will have many opportunities to use the multiplication, addition and subtraction principles. Usually these will arise in the context of other counting principles that we have yet to explore. It is thus important that you solidify the current ideas now, by working some exercises before moving on.

---

### Exercises for Section 3.3

1. Five cards are dealt off of a standard 52-card deck and lined up in a row. How many such lineups are there that have at least one red card? How many such lineups are there in which the cards are either all black or all hearts?
  2. Five cards are dealt off of a standard 52-card deck and lined up in a row. How many such lineups are there in which all 5 cards are of the same suit?
  3. Five cards are dealt off of a standard 52-card deck and lined up in a row. How many such lineups are there in which all 5 cards are of the same color (i.e., all black or all red)?
  4. Five cards are dealt off of a standard 52-card deck and lined up in a row. How many such lineups are there in which exactly one of the 5 cards is a queen?
  5. How many integers between 1 and 9999 have no repeated digits? How many have at least one repeated digit?
  6. Consider lists made from the symbols  $A, B, C, D, E$ , with repetition allowed.
    - (a) How many such length-5 lists have at least one letter repeated?
    - (b) How many such length-6 lists have at least one letter repeated?
  7. A password on a certain site must be five characters long, made from letters of the alphabet, and have at least one upper case letter. How many different passwords are there? What if there must be a mix of upper and lower case?
  8. This problem concerns lists made from the letters  $A, B, C, D, E, F, G, H, I, J$ .
    - (a) How many length-5 lists can be made from these letters if repetition is not allowed and the list must begin with a vowel?
    - (b) How many length-5 lists can be made from these letters if repetition is not allowed and the list must begin and end with a vowel?
    - (c) How many length-5 lists can be made from these letters if repetition is not allowed and the list must contain exactly one  $A$ ?
  9. Consider lists of length 6 made from the letters  $A, B, C, D, E, F, G, H$ . How many such lists are possible if repetition is not allowed and the list contains two consecutive vowels?
  10. Consider the lists of length six made with the symbols  $P, R, O, F, S$ , where repetition is allowed. (For example, the following is such a list:  $(P,R,O,O,F,S)$ .) How many such lists can be made if the list must end in an  $S$  and the symbol  $O$  is used more than once?
  11. How many integers between 1 and 1000 are divisible by 5? How many are not divisible by 5?
  12. Six math books, four physics books and three chemistry books are arranged on a shelf. How many arrangements are possible if all books of the same subject are grouped together?
-

### 3.4 Factorials and Permutations

In working examples from the previous two sections you may have noticed that we often need to count the number of non-repetitive lists of length  $n$  that are made from  $n$  symbols. This kind of problem occurs so often that a special idea, called a *factorial*, is used to handle it.

The table below motivates this. The first column lists successive integer values  $n$ , from 0 onward. The second contains a set  $\{a, b, \dots\}$  of  $n$  symbols. The third column shows all the possible non-repetitive lists of length  $n$  that can be made from these symbols. Finally, the last column tallies up how many lists there are of that type. When  $n = 0$  there is only one list of length 0 that can be made from 0 symbols, namely the empty list  $()$ . Thus the value 1 is entered in the last column of that row.

$n$	Symbols	Non-repetitive lists of length $n$ made from the symbols	$n!$
0	$\{\}$	$()$	1
1	$\{a\}$	$a$	1
2	$\{a, b\}$	$ab, ba$	2
3	$\{a, b, c\}$	$abc, acb, bac, bca, cab, cba$	6
4	$\{a, b, c, d\}$	$abcd, acbd, bacd, bcad, cabd, cbad, abdc, acdb, badc, beda, cadb, cbda, adbc, adcb, bdac, bdca, cdab, cdba, dabc, dacb, dbac, dbca, dcab, dcba,$	24
:	:	:	:

For  $n > 0$ , the number that appears in the last column can be computed using the multiplication principle. The number of non-repetitive lists of length  $n$  that can be made from  $n$  symbols is  $n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1$ . Thus, for instance, the number in the last column of the row for  $n = 4$  is  $4 \cdot 3 \cdot 2 \cdot 1 = 24$ .

The number that appears in the last column of Row  $n$  is called the **factorial** of  $n$ . It is denoted with the special symbol  $n!$ , which we pronounce as “ $n$  factorial.” Here is the definition:

**Definition 3.1** If  $n$  is a non-negative integer, then  $n!$  is the number of lists of length  $n$  that can be made from  $n$  symbols, without repetition. Thus  $0! = 1$  and  $1! = 1$ . If  $n > 1$ , then  $n! = n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1$ .

It follows that

$0!$	=	1
$1!$	=	1
$2!$	=	$2 \cdot 1 = 2$
$3!$	=	$3 \cdot 2 \cdot 1 = 6$
$4!$	=	$4 \cdot 3 \cdot 2 \cdot 1 = 24$
$5!$	=	$5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$
$6!$	=	$6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$ , and so on.

Students are often tempted to say  $0! = 0$ , but this is wrong. The correct value is  $0! = 1$ , as the above definition and table show. Here is another way to see that  $0!$  must equal 1: Notice that  $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5 \cdot (4 \cdot 3 \cdot 2 \cdot 1) = 5 \cdot 4!$ . Also  $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 4 \cdot (3 \cdot 2 \cdot 1) = 4 \cdot 3!$ . Generalizing this, we get a formula.

$$n! = n \cdot (n - 1)! \quad (3.1)$$

Plugging in  $n = 1$  gives  $1! = 1 \cdot (1 - 1)! = 1 \cdot 0!$ , that is,  $1! = 1 \cdot 0!$ . If we mistakenly thought  $0!$  were 0, this would give the incorrect result  $1! = 0$ .

**Example 3.8** This problem involves making lists of length seven from the letters  $a, b, c, d, e, f$  and  $g$

- (a) How many such lists are there if repetition is not allowed?
- (b) How many such lists are there if repetition is not allowed and the first two entries must be vowels?
- (c) How many such lists are there in which repetition is allowed, and the list must contain at least one repeated letter?

To answer the first question, note that there are seven letters, so the number of lists is  $7! = 5040$ . To answer the second question, notice that the set  $\{a, b, c, d, e, f, g\}$  contains two vowels and five consonants. Thus in making the list the first two entries must be filled by vowels and the final five must be filled with consonants. By the multiplication principle, the number of such lists is  $2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 2!5! = 440$ .

To answer part (c) we use the subtraction principle. Let  $U$  be the set of all lists made from  $a, b, c, d, e, f, g$ , with repetition allowed. The multiplication principle gives  $|U| = 7 \cdot 7 \cdot 7 \cdot 7 \cdot 7 \cdot 7 \cdot 7 = 7^7 = 823,543$ . Notice that  $U$  includes lists that are non-repetitive, like  $(a,g,f,b,d,c,e)$ , as well as lists that have some repetition, like  $(f,g,b,g,a,a,a)$ . We want to find the number of lists that have at least one repeated letter, so we will subtract away from  $U$  all those lists that have no repetition. Let  $X \subseteq U$  be those lists that have no repetition, so  $|X| = 7!$ . Thus the answer to our question is  $|U - X| = |U| - |X| = 7^7 - 7! = 823,543 - 5040 = 818,503$ .

In part (a) of Example 3.8 we counted the number of non-repetitive lists made from all seven of the symbols in the set  $X = \{a, b, c, d, e, f, g\}$ , and there were  $7! = 5040$  such lists. Any such list, such as *cedagf*, *gfedcba* or *abcdefg* is simply an arrangement of the elements of  $X$  in a row. There is a name for such an arrangement. It is called a *permutation* of  $X$ .

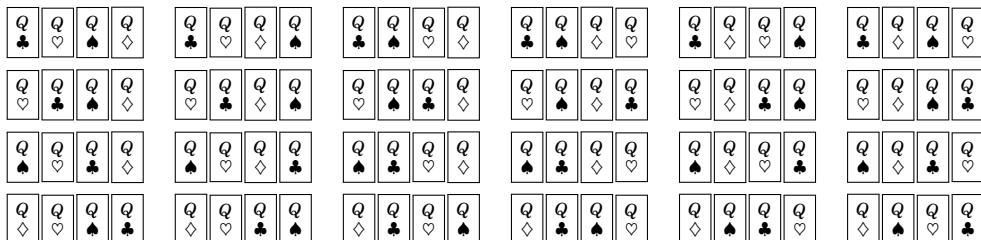
A **permutation** of a set is an arrangement of all of the set's elements in a row, that is, a list without repetition that uses every element of the set. For example, the permutations of the set  $X = \{1, 2, 3\}$  are the six lists

$$123, 132, 213, 231, 312, 321.$$

That we get six different permutations of  $X$  is predicted by Definition 3.1, which says there are  $3! = 3 \cdot 2 \cdot 1 = 6$  non-repetitive lists that can be made from the three symbols in  $X$ .

Think of the numbers 1, 2 and 3 as representing three books. The above shows that there are six ways to arrange them on a shelf.

From a deck of cards you take the four queens and lay them in a row. By the multiplication principle there are  $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$  ways to do this, that is, there are 24 permutations of the set of four Queen cards.



In general, a set with  $n$  elements will have  $n!$  different permutations. Above, the set  $\{1, 2, 3\}$  has  $3! = 6$  permutations, while  $\{\diamondsuit, \spadesuit, \heartsuit, \clubsuit\}$  has  $4! = 24$  permutations. The set  $\{a, b, c, d, e, f, g\}$  has  $7! = 5040$  permutations, though there's not much point in listing them all out. The important thing is that the factorial counts the number of permutations.

In saying a permutation of a set is an arrangement of its elements in a *row*, we are speaking informally because sometimes the elements are not literally in a row. Imagine a classroom of 20 desks, in four rows of five desks each. Let  $X$  be a class (set) of 20 students. If the students walk in and seat themselves, one per desk, we can regard this as a permutation of the 20 students because we can number the desks  $1, 2, 3, \dots, 20$  and in this sense the students have arranged themselves in a list of length 20. There are  $20! = 2,432,902,008,176,640,000$  permutations of the students.

Now we discuss a variation of the idea of a permutation of a set  $X$ . Imagine taking some number  $k \leq |X|$  of elements from the set  $X$  and then arranging *them* in a row. The result is what we call a  $k$ -permutation of  $X$ . A **permutation** of  $X$  is a non-repetitive list made from all elements of  $X$ . A  **$k$ -permutation** of  $X$  is a non-repetitive list made from  $k$  elements of  $X$ .

For example, take  $X = \{a, b, c, d\}$ . The 1-permutations of  $X$  are the lists we could make with just one element from  $X$ . There are only 4 such lists:

$$a \quad b \quad c \quad d.$$

The 2-permutations of  $X$  are the non-repetitive lists that we could make from two elements of  $X$ . There are 12 of them:

$$ab \quad ac \quad ad \quad ba \quad bc \quad bd \quad ca \quad cb \quad cd \quad da \quad db \quad dc.$$

Even before writing them all down, we'd know there are 12 of them because in making a non-repetitive length-2 list from  $X$  we have 4 choices for the first element, then 3 choices for the second, so by the multiplication principle the total number of 2-permutations of  $X$  is  $4 \cdot 3 = 12$ .

Now let's count the number of 3-permutations of  $X$ . They are the length-3 non-repetitive lists made from elements of  $X$ . The multiplication principle says there will be  $4 \cdot 3 \cdot 2 = 24$  of them. Here they are:

$$\begin{array}{ccccccc} abc & acb & bac & bca & cab & cba \\ abd & adb & bad & bda & dab & dba \\ acd & adc & cad & cda & dac & dca \\ bcd & bdc & cbd & cdb & dbc & dc b \end{array}$$

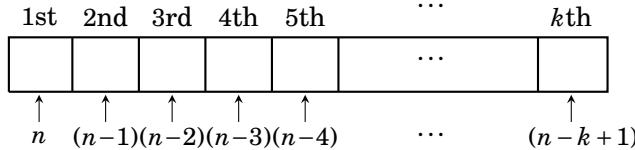
The 4-permutations of  $X$  are the non-repetitive lists made from all 4 elements of  $X$ . These are simply the  $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$  permutations of  $X$ .

Let's go back and think about the *0-permutations* of  $X$ . They are the non-repetitive lists of length 0 made from the elements of  $X$ . Of course there is only one such list, namely the empty list () .

Now we are going to introduce some notation. The expression  $P(n, k)$  denotes the number of  $k$ -permutations of an  $n$ -element set. By the examples on this page we have  $P(4, 0) = 1$ ,  $P(4, 1) = 4$ ,  $P(4, 2) = 12$ ,  $P(4, 3) = 24$ , and  $P(4, 4) = 24$ .

What about, say,  $P(4, 5)$ ? This is the number of 5-permutations of a 4-element set, that is, the number of non-repetitive length-5 lists that can be made from 4 symbols. There is no such list, so  $P(4, 5) = 0$ .

If  $n > 0$ , then  $P(n, k)$  can be computed with the multiplication principle. In making a non-repetitive length- $k$  list from  $n$  symbols we have  $n$  choices for the 1st entry,  $n - 1$  for the 2nd,  $n - 2$  for the 3rd, and  $n - 3$  for the 4th.



Notice that the number of choices for the  $i$ th position is  $n - i + 1$ . For example, the 5th position has  $n - 5 + 1 = n - 4$  choices. Continuing in this pattern, the last ( $k$ th) entry has  $n - k + 1$  choices. Therefore

$$P(n, k) = n(n - 1)(n - 2) \cdots (n - k + 1). \quad (3.2)$$

All together there are  $k$  factors in this product, so to compute  $P(n, k)$  just perform  $n(n - 1)(n - 2)(n - 3) \cdots$  until you've multiplied  $k$  numbers. Examples:

$$\begin{aligned} P(10, 1) &= 10 = 10 \\ P(10, 2) &= 10 \cdot 9 = 90 \\ P(10, 3) &= 10 \cdot 9 \cdot 8 = 720 \\ P(10, 4) &= 10 \cdot 9 \cdot 8 \cdot 7 = 5040 \\ &\vdots \quad \vdots \quad \vdots \\ P(10, 10) &= 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 3,628,800 \\ P(10, 11) &= 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 0 = 0. \end{aligned}$$

Note  $P(10, 11) = 0$ , as the 11th factor in the product is 0. This makes sense because  $P(10, 11)$  is the number of non-repetitive length-11 lists made from just 10 symbols. There are no such lists, so  $P(10, 11) = 0$  is right. In fact you can check that Equation (3.2) gives  $P(n, k) = 0$  whenever  $k > n$ .

Also notice above that  $P(10, 10) = 10!$ . In general  $P(n, n) = n!$ .

We now derive another formula for  $P(n, k)$ , one that works for  $0 \leq k \leq n$ . Using Equation (3.2) with cancellation and the definition of a factorial,

$$\begin{aligned} P(n, k) &= n(n - 1)(n - 2) \cdots (n - k + 1) \\ &= \frac{n(n - 1)(n - 2) \cdots (n - k + 1)(n - k)(n - k - 1) \cdots 3 \cdot 2 \cdot 1}{(n - k)(n - k - 1) \cdots 3 \cdot 2 \cdot 1} = \frac{n!}{(n - k)!}. \end{aligned}$$

To illustrate, let's find  $P(8, 5)$  in two ways. Equation (3.2) says  $P(8, 5) = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 = 6720$ . By the above formula,  $P(8, 5) = \frac{8!}{(8 - 5)!} = \frac{8!}{3!} = \frac{40,320}{6} = 6720$ .

We summarize these ideas in the following definition and fact.

**Fact 3.4** A **k-permutation** of an  $n$ -element set is a non-repetitive length- $k$  list made from elements of the set. Informally we think of a  $k$ -permutation as an arrangement of  $k$  of the set's elements in a row.

The number of  $k$ -permutations of an  $n$ -element set is denoted  $P(n, k)$ , and

$$P(n, k) = n(n - 1)(n - 2) \cdots (n - k + 1).$$

$$\text{If } 0 \leq k \leq n, \text{ then } P(n, k) = n(n - 1)(n - 2) \cdots (n - k + 1) = \frac{n!}{(n - k)!}.$$

Notice that  $P(n, 0) = \frac{n!}{(n - 0)!} = \frac{n!}{n!} = 1$ , which makes sense because only one list of length 0 can be made from  $n$  symbols, namely the empty list. Also  $P(0, 0) = \frac{0!}{(0 - 0)!} = \frac{0!}{0!} = \frac{1}{1} = 1$ , which is to be expected because there is only one list of length 0 that can be made with 0 symbols, again the empty list.

**Example 3.9** Ten contestants run a marathon. All finish, and there are no ties. How many different possible rankings are there for first-, second- and third-place?

**Solution:** Call the contestants  $A, B, C, D, E, F, G, H, I$  and  $J$ . A ranking of winners can be regarded as a 3-permutation of the set of 10 contestants. For example,  $ECH$  means  $E$  in first-place,  $C$  in second-place and  $H$  in third. Thus there are  $P(10, 3) = 10 \cdot 9 \cdot 8 = 720$  possible rankings.

**Example 3.10** You deal five cards off of a standard 52-card deck, and line them up in a row. How many such lineups are there that either consist of all red cards, or all clubs?

**Solution:** There are 26 red cards. The number of ways to line up five of them is  $P(26, 5) = 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 343,200$ .

There are 13 club cards (which are black). The number of ways to line up five of them is  $P(13, 5) = 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 = 154,440$ .

By the addition principle, the answer to our question is that there are  $P(26, 5) + P(13, 5) = 497,640$  lineups that are either all red cards, or all club cards.

Notice that we do not need to use the notation  $P(n, k)$  to solve the problems on this page. Straightforward applications of the multiplication and addition principles would suffice. However, the  $P(n, k)$  notation often proves to be a convenient shorthand.

### Exercises for Section 3.4

1. What is the smallest  $n$  for which  $n!$  has more than 10 digits?
2. For which values of  $n$  does  $n!$  have  $n$  or fewer digits?
3. How many 5-digit positive integers are there in which there are no repeated digits and all digits are odd?
4. Using only pencil and paper, find the value of  $\frac{100!}{95!}$ .
5. Using only pencil and paper, find the value of  $\frac{120!}{118!}$ .
6. There are two 0's at the end of  $10! = 3,628,800$ . Using only pencil and paper, determine how many 0's are at the end of the number  $100!$ .
7. Find how many 9-digit numbers can be made from the digits 1, 2, 3, 4, 5, 6, 7, 8, 9 if repetition is not allowed and all the odd digits occur first (on the left) followed by all the even digits (i.e., as in 137598264, but not 123456789).
8. Compute how many 7-digit numbers can be made from the digits 1, 2, 3, 4, 5, 6, 7 if there is no repetition and the odd digits must appear in an unbroken sequence. (Examples: 3571264 or 2413576 or 2467531, etc., but **not** 7234615.)
9. How many permutations of the letters  $A, B, C, D, E, F, G$  are there in which the three letters ABC appear consecutively, in alphabetical order?
10. How many permutations of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 are there in which the digits alternate even and odd? (For example, 2183470965.)
11. You deal 7 cards off of a 52-card deck and line them up in a row. How many possible lineups are there in which not all cards are red?
12. You deal 7 cards off of a 52-card deck and line them up in a row. How many possible lineups are there in which no card is a club?
13. How many lists of length six (with no repetition) can be made from the 26 letters of the English alphabet?
14. Five of ten books are arranged on a shelf. In how many ways can this be done?
15. In a club of 15 people, we need to choose a president, vice-president, secretary, and treasurer. In how many ways can this be done?
16. How many 4-permutations are there of the set  $\{A,B,C,D,E,F\}$  if whenever  $A$  appears in the permutation, it is followed by  $E$ ?
17. Three people in a group of ten line up at a ticket counter to buy tickets. How many lineups are possible?
18. There is a very interesting function  $\Gamma : [0, \infty) \rightarrow \mathbb{R}$  called the **gamma function**. It is defined as  $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ . It has the remarkable property that if  $x \in \mathbb{N}$ , then  $\Gamma(x) = (x - 1)!$ . Check that this is true for  $x = 1, 2, 3, 4$ . Notice that this function provides a way of extending factorials to numbers other than integers. Since  $\Gamma(n) = (n - 1)!$  for all  $n \in \mathbb{N}$ , we have the formula  $n! = \Gamma(n + 1)$ . But  $\Gamma$  can be evaluated at any number in  $[0, \infty)$ , not just at integers, so we have a formula for  $n!$  for any real number  $n \in [0, \infty)$ . Extra credit: Compute  $\pi!$ .

### 3.5 Counting Subsets

The previous section dealt with counting lists made by selecting  $k$  entries from a set of  $n$  elements. We turn now to a related question: How many subsets can be made by selecting  $k$  elements from a set with  $n$  elements?

To see the difference between these two problems, take  $A = \{a, b, c, d, e\}$ . Consider the non-repetitive lists made from selecting two elements from  $A$ . Fact 3.4 says there are  $P(5, 2) = 5 \cdot 4 = 20$  such lists, namely

$$(a, b), (a, c), (a, d), (a, e), (b, c), (b, d), (b, e), (c, d), (c, e), (d, e), \\ (b, a), (c, a), (d, a), (e, a), (c, b), (d, b), (e, b), (d, c), (e, c), (e, d).$$

But there are only ten 2-element subsets of  $A$ . They are

$$\{a, b\}, \{a, c\}, \{a, d\}, \{a, e\}, \{b, c\}, \{b, d\}, \{b, e\}, \{c, d\}, \{c, e\}, \{d, e\}.$$

The reason that there are more lists than subsets is that changing the order of the entries of a list produces a different list, but changing the order of the elements of a set does not change the set. Using elements  $a, b \in A$ , we can make two lists  $(a, b)$  and  $(b, a)$ , but only one subset  $\{a, b\}$ .

This section is concerned with counting subsets, not lists. As noted above, the basic question is this: How many subsets can be made by choosing  $k$  elements from an  $n$ -element set? We begin with some notation that gives a name to the answer to this question.

**Definition 3.2** If  $n$  and  $k$  are integers, then  $\binom{n}{k}$  denotes the number of subsets that can be made by choosing  $k$  elements from an  $n$ -element set. We read  $\binom{n}{k}$  as “ $n$  choose  $k$ .” (Some textbooks write  $C(n, k)$  instead of  $\binom{n}{k}$ .)

This is illustrated in the following table that tallies the  $k$ -element subsets of the 4-element set  $A = \{a, b, c, d\}$ , for various values of  $k$ .

$k$	$k$ -element subsets of $A = \{a, b, c, d\}$	$\binom{4}{k}$
-1		$\binom{4}{-1} = 0$
0	$\emptyset$	$\binom{4}{0} = 1$
1	$\{a\}, \{b\}, \{c\}, \{d\}$	$\binom{4}{1} = 4$
2	$\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}$	$\binom{4}{2} = 6$
3	$\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}$	$\binom{4}{3} = 4$
4	$\{a, b, c, d\}$	$\binom{4}{4} = 1$
5		$\binom{4}{5} = 0$

The values of  $k$  appear in the far-left column of the table. To the right of each  $k$  are all of the subsets (if any) of  $A$  of size  $k$ . For example, when  $k = 1$ , set  $A$  has four subsets of size  $k$ , namely  $\{a\}$ ,  $\{b\}$ ,  $\{c\}$  and  $\{d\}$ . Therefore  $\binom{4}{1} = 4$ . When  $k = 2$  there are six subsets of size  $k$  so  $\binom{4}{2} = 6$ .

When  $k = 0$ , there is only one subset of  $A$  that has cardinality  $k$ , namely the empty set,  $\emptyset$ . Therefore  $\binom{4}{0} = 1$ .

Notice that if  $k$  is negative or greater than  $|A|$ , then  $A$  has no subsets of cardinality  $k$ , so  $\binom{4}{k} = 0$  in these cases. In general  $\binom{n}{k} = 0$  whenever  $k < 0$  or  $k > n$ . In particular this means  $\binom{n}{k} = 0$  if  $n$  is negative.

Although it was not hard to work out the values of  $\binom{4}{k}$  by writing out subsets in the above table, this method of actually listing sets would not be practical for computing  $\binom{n}{k}$  when  $n$  and  $k$  are large. We need a formula. To find one, we will now carefully work out the value of  $\binom{5}{3}$  in a way that highlights a pattern that points the way to a formula for any  $\binom{n}{k}$ .

To begin, note that  $\binom{5}{3}$  is the number of 3-element subsets of  $\{a, b, c, d, e\}$ . These are listed in the top row of the table below, where we see  $\binom{5}{3} = 10$ . The column under each subset tallies the  $3! = 6$  permutations of that subset. The first subset  $\{a, b, c\}$  has  $3! = 6$  permutations; these are listed below it. The second column tallies the permutations of  $\{a, b, d\}$ , and so on.

	$\binom{5}{3}$									
	$\leftarrow \qquad \qquad \qquad \rightarrow\right)$									
	$\{a, b, c\} \{a, b, d\} \{a, b, e\} \{a, c, d\} \{a, c, e\} \{a, d, e\} \{b, c, d\} \{b, c, e\} \{b, d, e\} \{c, d, e\}$									
3!	abc	abd	abe	acd	ace	ade	bcd	bce	bde	cde
	acb	adb	aeb	adc	aec	aed	bdc	bec	bed	ced
	bac	bad	bae	cad	cae	dae	cbd	cbe	dbe	dce
	bca	bda	bea	cda	cea	dea	cdb	ceb	deb	dec
	cba	dba	eba	dca	eca	eda	dcg	ecb	edb	edc
	cab	dab	eab	dac	eac	ead	dbc	ebc	ebd	ecd

The body of this table has  $\binom{5}{3}$  columns and  $3!$  rows, so it has a total of  $3!\binom{5}{3}$  lists. But notice also that the table consists of every 3-permutation of  $\{a, b, c, d, e\}$ . Fact 3.4 says that there are  $P(5, 3) = \frac{5!}{(5-3)!}$  such 3-permutations. Thus the total number of lists in the table can be written as either  $3!\binom{5}{3}$  or  $\frac{5!}{(5-3)!}$ , which is to say  $3!\binom{5}{3} = \frac{5!}{(5-3)!}$ . Dividing both sides by  $3!$  yields

$$\binom{5}{3} = \frac{5!}{3!(5-3)!}.$$

Working this out, you will find that it does give the correct value of 10.

But there was nothing special about the values 5 and 3. We could do the above analysis for any  $\binom{n}{k}$  instead of  $\binom{5}{3}$ . The table would have  $\binom{n}{k}$  columns and  $k!$  rows. We would get

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

We have established the following fact, which holds for all  $k, n \in \mathbb{Z}$ .

**Fact 3.5** If  $0 \leq k \leq n$ , then  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ . Otherwise  $\binom{n}{k} = 0$ .

Let's now use our new knowledge to work some exercises.

**Example 3.11** How many size-4 subsets does  $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$  have?

The answer is  $\binom{9}{4} = \frac{9!}{4!(9-4)!} = \frac{9!}{4!5!} = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5!}{4!5!} = \frac{9 \cdot 8 \cdot 7 \cdot 6}{4!} = \frac{9 \cdot 8 \cdot 7 \cdot 6}{24} = 126$ .

**Example 3.12** How many 5-element subsets of  $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$  have exactly two even elements?

**Solution:** Making a 5-element subset of  $A$  with exactly two even elements is a 2-step process. First select two of the four even elements from  $A$ . There are  $\binom{4}{2} = 6$  ways to do this. Next, there are  $\binom{5}{3} = 10$  ways select three of the five odd elements of  $A$ . By the multiplication principle, there are  $\binom{4}{2}\binom{5}{3} = 6 \cdot 10 = 60$  ways to select two even and three odd elements from  $A$ . So there are **60** 5-element subsets of  $A$  with exactly two even elements.

**Example 3.13** A single 5-card hand is dealt off of a standard 52-card deck. How many different 5-card hands are possible?

**Solution:** Think of the deck as a set  $D$  of 52 cards. Then a 5-card hand is just a 5-element subset of  $D$ . There are many such subsets, such as

$$\left\{ \begin{array}{|c|c|c|c|c|} \hline & 7 & 2 & 3 & A \\ \hline & \clubsuit & \clubsuit & \heartsuit & \spadesuit \\ \hline \end{array}, \begin{array}{|c|c|c|c|c|} \hline & 7 & 2 & 3 & 5 \\ \hline & \clubsuit & \clubsuit & \heartsuit & \diamondsuit \\ \hline \end{array}, \begin{array}{|c|c|c|c|c|} \hline & 7 & 2 & 3 & K \\ \hline & \clubsuit & \clubsuit & \heartsuit & \spadesuit \\ \hline \end{array}, \begin{array}{|c|c|c|c|c|} \hline & 7 & 2 & 3 & Q \\ \hline & \clubsuit & \clubsuit & \heartsuit & \spadesuit \\ \hline \end{array}, \begin{array}{|c|c|c|c|c|} \hline & 7 & 2 & 3 & J \\ \hline & \clubsuit & \clubsuit & \heartsuit & \spadesuit \\ \hline \end{array} \right\}.$$

Thus the number of 5-card hands is the number of 5-element subsets of  $D$ , which is

$$\binom{52}{5} = \frac{52!}{5! \cdot 47!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47!}{5! \cdot 47!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5!} = 2,598,960.$$

**Answer:** There are **2,598,960** different five-card hands that can be dealt from a deck of 52 cards.

**Example 3.14** This problem concerns 5-card hands that can be dealt off of a 52-card deck. How many such hands are there in which two of the cards are clubs and three are hearts?

**Solution:** Such a hand is described by a list of length two of the form

$$\left( \left\{ \begin{array}{|c|} \hline * \\ \hline \clubsuit \\ \hline \end{array} \right|, \begin{array}{|c|} \hline * \\ \hline \heartsuit \\ \hline \end{array} \right\}, \left\{ \begin{array}{|c|} \hline * \\ \hline \heartsuit \\ \hline \end{array} \right|, \begin{array}{|c|} \hline * \\ \hline \heartsuit \\ \hline \end{array} \right\} \right),$$

where the first entry is a 2-element subset of the set of 13 club cards, and the second entry is a 3-element subset of the set of 13 heart cards. There are  $\binom{13}{2}$  choices for the first entry and  $\binom{13}{3}$  choices for the second, so by the multiplication principle there are  $\binom{13}{2}\binom{13}{3} = \frac{13!}{2!11!} \frac{13!}{3!10!} = 22,308$  such lists. Thus there are **22,308** such 5-card hands.

**Example 3.15** A lottery features a bucket of 36 balls numbered 1 through 36. Six balls will be drawn randomly. For \$1 you buy a ticket with six blanks:  $\boxed{\phantom{0}}\boxed{\phantom{0}}\boxed{\phantom{0}}\boxed{\phantom{0}}\boxed{\phantom{0}}\boxed{\phantom{0}}$ . You fill in the blanks with six different numbers between 1 and 36. You win \$1,000,000 if you chose the same numbers that are drawn, regardless of order. What are your chances of winning?

**Solution:** In filling out the ticket you are choosing six numbers from a set of 36 numbers. Thus there are  $\binom{36}{6} = \frac{36!}{6!(36-6)!} = 1,947,792$  different combinations of numbers you might write. Only one of these will be a winner. **Your chances of winning are one in 1,947,792.**

**Example 3.16** How many 7-digit binary strings (0010100, 1101011, etc.) have an odd number of 1's?

**Solution:** Let  $A$  be the set of all 7-digit binary strings with an odd number of 1's, so the answer will be  $|A|$ . To find  $|A|$ , we break  $A$  into smaller parts. Notice any string in  $A$  will have either one, three, five or seven 1's. Let  $A_1$  be the set of 7-digit binary strings with only one 1. Let  $A_3$  be the set of 7-digit binary strings with three 1's. Let  $A_5$  be the set of 7-digit binary strings with five 1's, and let  $A_7$  be the set of 7-digit binary strings with seven 1's. Then  $A = A_1 \cup A_3 \cup A_5 \cup A_7$ . Any two of the sets  $A_i$  have empty intersection, so the addition principle gives  $|A| = |A_1| + |A_3| + |A_5| + |A_7|$ .

Now we must compute the individual terms of this sum. Take  $A_3$ , the set of 7-digit binary strings with three 1's. Such a string can be formed by selecting three out of seven positions for the 1's and putting 0's in the other spaces. Thus  $|A_3| = \binom{7}{3}$ . Similarly  $|A_1| = \binom{7}{1}$ ,  $|A_5| = \binom{7}{5}$ , and  $|A_7| = \binom{7}{7}$ .

**Answer:**  $|A| = |A_1| + |A_3| + |A_5| + |A_7| = \binom{7}{1} + \binom{7}{3} + \binom{7}{5} + \binom{7}{7} = 7 + 35 + 21 + 1 = 64$ . There are **64** 7-digit binary strings with an odd number of 1's.

---

### Exercises for Section 3.5

1. Suppose a set  $A$  has 37 elements. How many subsets of  $A$  have 10 elements? How many subsets have 30 elements? How many have 0 elements?
  2. Suppose  $A$  is a set for which  $|A| = 100$ . How many subsets of  $A$  have 5 elements? How many subsets have 10 elements? How many have 99 elements?
  3. A set  $X$  has exactly 56 subsets with 3 elements. What is the cardinality of  $X$ ?
  4. Suppose a set  $B$  has the property that  $|\{X : X \in \mathcal{P}(B), |X| = 6\}| = 28$ . Find  $|B|$ .
  5. How many 16-digit binary strings contain exactly seven 1's? (Examples of such strings include 0111000011110000 and 0011001100110010, etc.)
  6.  $|\{X \in \mathcal{P}(\{0,1,2,3,4,5,6,7,8,9\}) : |X| = 4\}| =$
  7.  $|\{X \in \mathcal{P}(\{0,1,2,3,4,5,6,7,8,9\}) : |X| < 4\}| =$
  8. This problem concerns lists made from the symbols  $A, B, C, D, E, F, G, H, I$ .
    - (a) How many length-5 lists can be made if there is no repetition and the list is in alphabetical order? (Example:  $BDEFI$  or  $ABCGH$ , but not  $BACGH$ .)
    - (b) How many length-5 lists can be made if repetition is not allowed and the list is **not** in alphabetical order?
  9. This problem concerns lists of length 6 made from the letters  $A, B, C, D, E, F$ , without repetition. How many such lists have the property that the  $D$  occurs before the  $A$ ?
  10. A department consists of 5 men and 7 women. From this department you select a committee with 3 men and 2 women. In how many ways can you do this?
  11. How many positive 10-digit integers contain no 0's and exactly three 6's?
  12. Twenty-one people are to be divided into two teams, the Red Team and the Blue Team. There will be 10 people on Red Team and 11 people on Blue Team. In how many ways can this be done?
  13. Suppose  $n, k \in \mathbb{Z}$ , and  $0 \leq k \leq n$ . Use Fact 3.5, the formula  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ , to show that  $\binom{n}{k} = \binom{n}{n-k}$ .
  14. Suppose  $n, k \in \mathbb{Z}$ , and  $0 \leq k \leq n$ . Use Definition 3.2 alone (without using Fact 3.5) to show that  $\binom{n}{k} = \binom{n}{n-k}$ .
  15. How many 10-digit binary strings are there that do not have exactly four 1's?
  16. How many 6-element subsets of  $A = \{0,1,2,3,4,5,6,7,8,9\}$  have exactly three even elements? How many do not have exactly three even elements?
  17. How many 10-digit binary strings are there that have exactly four 1's or exactly five 1's? How many do not have exactly four 1's or exactly five 1's?
  18. How many 10-digit binary strings have an even number of 1's?
  19. A 5-card poker hand is called a *flush* if all cards are the same suit. How many different flushes are there?
-

### 3.6 Pascal's Triangle and the Binomial Theorem

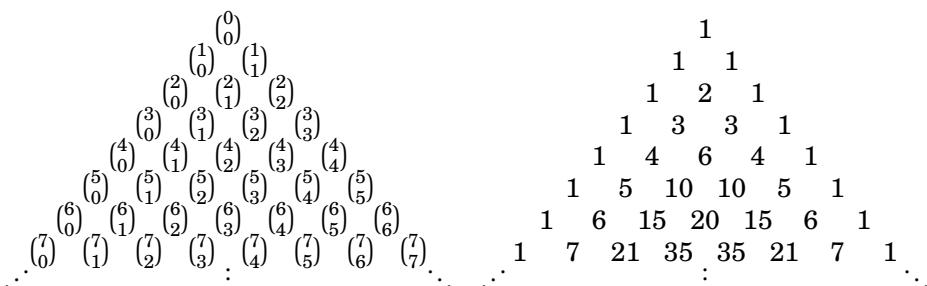
There are some beautiful and significant patterns among the numbers  $\binom{n}{k}$ . We now investigate a pattern based on one equation in particular. It happens that

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \quad (3.3)$$

for any integers  $n$  and  $k$  with  $1 \leq k \leq n$ .

To see why this is true, notice that the left-hand side  $\binom{n+1}{k}$  is the number of  $k$ -element subsets of the set  $A = \{0, 1, 2, 3, \dots, n\}$ , which has  $n+1$  elements. Such a subset either contains 0 or it does not. The  $\binom{n}{k-1}$  on the right is the number of subsets of  $A$  that contain 0, because to make such a subset we can start with  $\{0\}$  and append it an additional  $k-1$  numbers selected from  $\{1, 2, 3, \dots, n\}$ , and there are  $\binom{n}{k-1}$  ways to do this. Also, the  $\binom{n}{k}$  on the right is the number of subsets of  $A$  that **do not** contain 0, for it is the number of ways to select  $k$  elements from  $\{1, 2, 3, \dots, n\}$ . In light of all this, Equation (3.3) just states the obvious fact that the number of  $k$ -element subsets of  $A$  equals the number of  $k$ -element subsets that contain 0 plus the number of  $k$ -element subsets that do not contain 0.

Having seen why Equation (3.3) is true, we now highlight it by arranging the numbers  $\binom{n}{k}$  in a triangular pattern. The left-hand side of Figure 3.3 shows the numbers  $\binom{n}{k}$  arranged in a pyramid with  $\binom{0}{0}$  at the apex, just above a row containing  $\binom{1}{k}$  with  $k=0$  and  $k=1$ . Below this is a row listing the values of  $\binom{2}{k}$  for  $k=0, 1, 2$ , and so on.



**Figure 3.3.** Pascal's triangle

Any number  $\binom{n+1}{k}$  for  $0 < k < n$  in this pyramid is just below and between the two numbers  $\binom{n}{k-1}$  and  $\binom{n}{k}$  in the previous row. But Equation (3.3) says  $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$ . Therefore any number (other than 1) in the pyramid is the sum of the two numbers immediately above it.

This pattern is especially evident on the right of Figure 3.3, where each  $\binom{n}{k}$  is worked out. Notice how 21 is the sum of the numbers 6 and 15 above it. Similarly, 5 is the sum of the 1 and 4 above it and so on.

This arrangement is called **Pascal's triangle**, after Blaise Pascal, 1623–1662, a French philosopher and mathematician who discovered many of its properties. We've shown only the first eight rows, but the triangle extends downward forever. We can always add a new row at the bottom by placing a 1 at each end and obtaining each remaining number by adding the two numbers above its position. Doing this in Figure 3.3 (right) gives a new bottom row

$$1 \quad 8 \quad 28 \quad 56 \quad 70 \quad 56 \quad 28 \quad 8 \quad 1.$$

This row consists of the numbers  $\binom{8}{k}$  for  $0 \leq k \leq 8$ , and we have computed them without the formula  $\binom{8}{k} = \frac{8!}{k!(8-k)!}$ . Any  $\binom{n}{k}$  can be computed this way.

The very top row (containing only 1) of Pascal's triangle is called *Row 0*. Row 1 is the next down, followed by Row 2, then Row 3, etc. Thus Row  $n$  lists the numbers  $\binom{n}{k}$  for  $0 \leq k \leq n$ . Exercises 3.5.13 and 3.5.14 established

$$\binom{n}{k} = \binom{n}{n-k}, \quad (3.4)$$

for each  $0 \leq k \leq n$ . In words, the  $k$ th entry of Row  $n$  of Pascal's triangle equals the  $(n-k)$ th entry. This means that Pascal's triangle is symmetric with respect to the vertical line through its apex, as is evident in Figure 3.3.

		1			1		
		1	1			$1x + 1y$	
		1	2	1		$1x^2 + 2xy + 1y^2$	
		1	3	3	1	$1x^3 + 3x^2y + 3xy^2 + 1y^3$	
		1	4	6	4	1	$1x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + 1y^4$
		1	5	10	10	5	1
⋮		⋮		⋮	⋮	⋮	⋮

**Figure 3.4.** The  $n^{th}$  row of Pascal's triangle lists the coefficients of  $(x+y)^n$

Notice that Row  $n$  appears to be a list of the coefficients of  $(x+y)^n$ . For example  $(x+y)^2 = 1x^2 + 2xy + 1y^2$ , and Row 2 lists the coefficients 1 2 1. Also  $(x+y)^3 = 1x^3 + 3x^2y + 3xy^2 + 1y^3$ , and Row 3 is 1 3 3 1. See Figure 3.4, which suggestss that the numbers in Row  $n$  are the coefficients of  $(x+y)^n$ .

In fact this turns out to be true for every  $n$ . This fact is known as the **binomial theorem**, and it is worth mentioning here. It tells how to raise a binomial  $x + y$  to a non-negative integer power  $n$ .

**Theorem 3.1 (Binomial Theorem)** If  $n$  is a non-negative integer, then  $(x + y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \binom{n}{3}x^{n-3}y^3 + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n$ .

For now we will be content to accept the binomial theorem without proof. (You will be asked to prove it in an exercise in Chapter 10.) You may find it useful from time to time. For instance, you can use it if you ever need to expand an expression such as  $(x + y)^7$ . To do this, look at Row 7 of Pascal's triangle in Figure 3.3 and apply the binomial theorem to get

$$(x + y)^7 = x^7 + 7x^6y + 21x^5y^2 + 35x^4y^3 + 35x^3y^4 + 21x^2y^5 + 7xy^6 + y^7.$$

For another example,

$$\begin{aligned}(2a - b)^4 &= ((2a) + (-b))^4 \\ &= (2a)^4 + 4(2a)^3(-b) + 6(2a)^2(-b)^2 + 4(2a)(-b)^3 + (-b)^4 \\ &= 16a^4 - 32a^3b + 24a^2b^2 - 8ab^3 + b^4.\end{aligned}$$

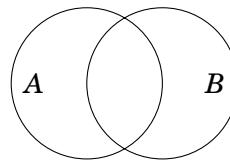
### Exercises for Section 3.6

1. Write out Row 11 of Pascal's triangle.
2. Use the binomial theorem to find the coefficient of  $x^8y^5$  in  $(x + y)^{13}$ .
3. Use the binomial theorem to find the coefficient of  $x^8$  in  $(x + 2)^{13}$ .
4. Use the binomial theorem to find the coefficient of  $x^6y^3$  in  $(3x - 2y)^9$ .
5. Use the binomial theorem to show  $\sum_{k=0}^n \binom{n}{k} = 2^n$ .
6. Use Definition 3.2 (page 85) and Fact 1.3 (page 13) to show  $\sum_{k=0}^n \binom{n}{k} = 2^n$ .
7. Use the binomial theorem to show  $\sum_{k=0}^n 3^k \binom{n}{k} = 4^n$ .
8. Use Fact 3.5 (page 87) to derive Equation 3.3 (page 90).
9. Use the binomial theorem to show  $\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \binom{n}{4} - \cdots + (-1)^n \binom{n}{n} = 0$ , for  $n > 0$ .
10. Show that the formula  $k \binom{n}{k} = n \binom{n-1}{k-1}$  is true for all integers  $n, k$  with  $0 \leq k \leq n$ .
11. Use the binomial theorem to show  $9^n = \sum_{k=0}^n (-1)^k \binom{n}{k} 10^{n-k}$ .
12. Show that  $\binom{n}{k} \binom{k}{m} = \binom{n}{m} \binom{n-m}{k-m}$ .
13. Show that  $\binom{n}{3} = \binom{2}{2} + \binom{3}{2} + \binom{4}{2} + \binom{5}{2} + \cdots + \binom{n-1}{2}$ .
14. The first five rows of Pascal's triangle appear in the digits of powers of 11:  $11^0 = 1$ ,  $11^1 = 11$ ,  $11^2 = 121$ ,  $11^3 = 1331$  and  $11^4 = 14641$ . Why is this so? Why does the pattern not continue with  $11^5$ ?

### 3.7 The Inclusion-Exclusion Principle

Many counting problems involve computing the cardinality of a union  $A \cup B$  of two finite sets. We examine this kind of problem now.

First we develop a formula for  $|A \cup B|$ . It is tempting to say that  $|A \cup B|$  must equal  $|A| + |B|$ , but that is not quite right. If we count the elements of  $A$  and then count the elements of  $B$  and add the two figures together, we get  $|A| + |B|$ . But if  $A$  and  $B$  have some elements in common, then we have counted each element in  $A \cap B$  twice.



Therefore  $|A| + |B|$  exceeds  $|A \cup B|$  by  $|A \cap B|$ , and consequently  $|A \cup B| = |A| + |B| - |A \cap B|$ . This can be a useful equation.

#### Fact 3.6 Inclusion-Exclusion Formula

If  $A$  and  $B$  are finite sets, then  $|A \cup B| = |A| + |B| - |A \cap B|$ .

Notice that the sets  $A$ ,  $B$  and  $A \cap B$  are all generally smaller than  $A \cup B$ , so Fact 3.6 has the potential of reducing the problem of determining  $|A \cup B|$  to three simpler counting problems. It is called the *inclusion-exclusion* formula because elements in  $A \cap B$  are included (twice) in  $|A| + |B|$ , then excluded when  $|A \cap B|$  is subtracted. Notice that if  $A \cap B = \emptyset$ , then we do in fact get  $|A \cup B| = |A| + |B|$ . (This is an instance of the addition principle!) Conversely, if  $|A \cup B| = |A| + |B|$ , then it must be that  $A \cap B = \emptyset$ .

**Example 3.17** A 3-card hand is dealt off of a standard 52-card deck. How many different such hands are there for which all three cards are red or all three cards are face cards?

**Solution:** Let  $A$  be the set of 3-card hands where all three cards are red (i.e., either  $\heartsuit$  or  $\diamondsuit$ ). Let  $B$  be the set of 3-card hands in which all three cards are face cards (i.e.,  $J, K$  or  $Q$  of any suit). These sets are illustrated below.

$$A = \left\{ \left\{ \begin{matrix} 5 \\ \heartsuit \end{matrix}, \begin{matrix} K \\ \diamondsuit \end{matrix}, \begin{matrix} 2 \\ \heartsuit \end{matrix} \right\}, \left\{ \begin{matrix} K \\ \heartsuit \end{matrix}, \begin{matrix} J \\ \heartsuit \end{matrix}, \begin{matrix} Q \\ \heartsuit \end{matrix} \right\}, \left\{ \begin{matrix} A \\ \diamondsuit \end{matrix}, \begin{matrix} 6 \\ \diamondsuit \end{matrix}, \begin{matrix} 6 \\ \heartsuit \end{matrix} \right\}, \dots \right\} \quad (\text{Red cards})$$

$$B = \left\{ \left\{ \begin{matrix} K \\ \spadesuit \end{matrix}, \begin{matrix} K \\ \diamondsuit \end{matrix}, \begin{matrix} J \\ \clubsuit \end{matrix} \right\}, \left\{ \begin{matrix} K \\ \heartsuit \end{matrix}, \begin{matrix} J \\ \heartsuit \end{matrix}, \begin{matrix} Q \\ \heartsuit \end{matrix} \right\}, \left\{ \begin{matrix} Q \\ \diamondsuit \end{matrix}, \begin{matrix} Q \\ \clubsuit \end{matrix}, \begin{matrix} Q \\ \heartsuit \end{matrix} \right\}, \dots \right\} \quad (\text{Face cards})$$

We seek the number of 3-card hands that are all red or all face cards, and this number is  $|A \cup B|$ . By Fact 3.6,  $|A \cup B| = |A| + |B| - |A \cap B|$ . Let's examine  $|A|$ ,  $|B|$  and  $|A \cap B|$  separately. Any hand in  $A$  is formed by selecting three cards from the 26 red cards in the deck, so  $|A| = \binom{26}{3}$ . Similarly, any hand in  $B$  is formed by selecting three cards from the 12 face cards in the deck, so  $|B| = \binom{12}{3}$ . Now think about  $A \cap B$ . It contains all the 3-card hands made up of cards that are red face cards.

$$A \cap B = \left\{ \left\{ \begin{array}{c} K \\ \heartsuit \\ \diamondsuit \end{array}, \begin{array}{c} K \\ \diamondsuit \\ \heartsuit \end{array}, \begin{array}{c} J \\ \heartsuit \\ \diamondsuit \end{array} \right\}, \left\{ \begin{array}{c} K \\ \heartsuit \\ \heartsuit \end{array}, \begin{array}{c} J \\ \heartsuit \\ \heartsuit \end{array}, \begin{array}{c} Q \\ \heartsuit \\ \heartsuit \end{array} \right\}, \left\{ \begin{array}{c} Q \\ \diamondsuit \\ \diamondsuit \end{array}, \begin{array}{c} J \\ \diamondsuit \\ \diamondsuit \end{array}, \begin{array}{c} Q \\ \diamondsuit \\ \heartsuit \end{array} \right\}, \dots \right\} \quad (\text{Red face cards})$$

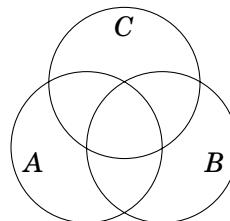
The deck has only 6 red face cards, so  $|A \cap B| = \binom{6}{3}$ .

Now we can answer our question. The number of 3-card hands that are all red or all face cards is  $|A \cup B| = |A| + |B| - |A \cap B| = \binom{26}{3} + \binom{12}{3} - \binom{6}{3} = 2600 + 220 - 20 = \mathbf{2800}$ .

**Example 3.18** A 3-card hand is dealt off of a standard 52-card deck. How many different such hands are there for which it is **not** the case that all 3 cards are red or all three cards are face cards?

**Solution:** We will use the subtraction principle combined with our answer to Example 3.17, above. The total number of 3-card hands is  $\binom{52}{3} = \frac{52!}{3!(52-3)!} = \frac{52!}{3!49!} = \frac{52 \cdot 51 \cdot 50}{3!} = 26 \cdot 51 \cdot 17 = 22,542$ . To get our answer, we must subtract from this the number of 3-card hands that are all red or all face cards, that is, we must subtract the answer from Example 3.17. Thus the answer to our question is  $22,542 - 2800 = \mathbf{19,742}$ .

There is an analogue of Fact 3.6 that involves three sets. Consider three sets  $A$ ,  $B$  and  $C$ , as represented in the following Venn Diagram.



Using the same kind of reasoning that resulted in Fact 3.6, you can convince yourself that

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|. \quad (3.5)$$

There's probably not much harm in ignoring this one for now, but if you find this kind of thing intriguing you should definitely take a course in combinatorics. (Ask your instructor!)

---

### Exercises for Section 3.7

1. At a certain university 523 of the seniors are history majors or math majors (or both). There are 100 senior math majors, and 33 seniors are majoring in both history and math. How many seniors are majoring in history?
  2. How many 4-digit positive integers are there for which there are no repeated digits, or for which there may be repeated digits, but all digits are odd?
  3. How many 4-digit positive integers are there that are even or contain no 0's?
  4. This problem involves lists made from the letters  $T, H, E, O, R, Y$ , with repetition allowed.
    - (a) How many 4-letter lists are there that don't begin with  $T$ , or don't end in  $Y$ ?
    - (b) How many 4-letter lists are there in which the sequence of letters  $T, H, E$  appears consecutively (in that order)?
    - (c) How many 6-letter lists are there in which the sequence of letters  $T, H, E$  appears consecutively (in that order)?
  5. How many 7-digit binary strings begin in 1 or end in 1 or have exactly four 1's?
  6. Is the following statement true or false? Explain. If  $A_1 \cap A_2 \cap A_3 = \emptyset$ , then  $|A_1 \cup A_2 \cup A_3| = |A_1| + |A_2| + |A_3|$ .
  7. Consider 4-card hands dealt off of a standard 52-card deck. How many hands are there for which all 4 cards are of the same suit or all 4 cards are red?
  8. Consider 4-card hands dealt off of a standard 52-card deck. How many hands are there for which all 4 cards are of different suits or all 4 cards are red?
  9. A 4-letter list is made from the letters  $L, I, S, T, E, D$  according to the following rule: Repetition is allowed, and the first two letters on the list are vowels or the list ends in  $D$ . How many such lists are possible?
  10. How many 6-digit numbers are even or are divisible by 5?
  11. How many 7-digit numbers are even or have exactly three digits equal to 0?
  12. How many 5-digit numbers are there in which three of the digits are 7, or two of the digits are 2?
  13. How many 8-digit binary strings end in 1 or have exactly four 1's?
  14. How many 3-card hands (from a standard 52-card deck) have the property that it is **not** the case that all cards are black or all cards are of the same suit?
  15. How many 10-digit binary strings begin in 1 or end in 1?
-

### 3.8 Counting Multisets

You have in your pocket four pennies, two nickels, a dime and three quarters. You might be tempted to regard this collection as a set

$$\{1, 1, 1, 1, 5, 5, 10, 25, 25\}.$$

But this is not a valid model of your collection of change, because a set cannot have repeated elements. To overcome this difficulty, we make a new construction called a **multiset**. A multiset is like a set, except that elements can be repeated. We will use square brackets  $[ ]$  instead of braces  $\{ \}$  to denote multisets. For example, your multiset of change is

$$[1, 1, 1, 1, 5, 5, 10, 25, 25].$$

A multiset is a hybrid of a set and a list; in a multiset, elements can be repeated, but order does not matter. For instance

$$\begin{aligned}[1, 1, 1, 1, 5, 5, 10, 25, 25] &= [25, 5, 1, 1, 10, 1, 1, 5, 25] \\ &= [25, 10, 25, 1, 5, 1, 5, 1, 1].\end{aligned}$$

Given a multiset  $A$ , its **cardinality**  $|A|$  is the number of elements it has, including repetition. So if  $A = [1, 1, 1, 1, 5, 5, 10, 25, 25]$ , then  $|A| = 9$ . The **multiplicity** of an element  $x \in A$  is the number of times that  $x$  appears, so  $1 \in A$  has multiplicity 4, while 5 and 25 each have multiplicity 2, and 10 has multiplicity 1. Notice that every set can be regarded as a multiset for which each element has multiplicity 1. In this sense we can think of  $\emptyset = \{\} = []$  as the multiset that has no elements.

To illustrate the idea of multisets, consider the multisets of cardinality 2 that can be made from the symbols  $\{a, b, c, d\}$ . They are

$$[a, a] \quad [a, b] \quad [a, c] \quad [a, d] \quad [b, b] \quad [b, c] \quad [b, d] \quad [c, c] \quad [c, d] \quad [d, d].$$

We have listed them so that the letters in each multiset are in alphabetical order (remember, we can order the elements of a multiset in any way we choose), and the 10 multisets are arranged in dictionary order.

For multisets of cardinality 3 made from  $\{a, b, c, d\}$ , we have

$$\begin{array}{cccccc}[a, a, a] & [a, a, b] & [a, a, c] & [a, a, d] & [a, b, b] \\ [a, b, c] & [a, b, d] & [a, c, c] & [a, c, d] & [a, d, d] \\ [b, b, b] & [b, b, c] & [b, b, d] & [b, c, c] & [b, c, d] \\ [b, d, d] & [c, c, c] & [c, c, d] & [c, d, d] & [d, d, d].\end{array}$$

Though  $X = \{a, b, c, d\}$  has no subsets of cardinality 5, there are many multisets of cardinality 5 made from these elements, including  $[a, a, a, a, a]$ ,  $[a, a, b, c, d]$  and  $[b, c, c, d, d]$ , and so on. Exactly how many are there?

This is the first question about multisets that we shall tackle: Given a finite set  $X$ , how many cardinality- $k$  multisets can be made from  $X$ ?

Let's start by counting the cardinality-5 multisets made from symbols  $X = \{a, b, c, d\}$ . (Our approach will lead to a general formula.) We know we can write any such multiset with its letters in alphabetical order. Tweaking the notation slightly, we could write any such multiset with bars separating the groupings of  $a, b, c, d$ , as shown in the table below. Notice that if a symbol does not appear in the multiset, we still write the bar that would have separated it from the others.

Multiset	with separating bars	encoding
$[a, a, b, c, d]$	$aa b c d$	$* *   *   *   *$
$[a, b, b, c, d]$	$a bb c d$	$*  * *   *   *$
$[a, b, c, c, d]$	$a b cc d$	$*  *   * *   *$
$[a, a, c, c, d]$	$aa  cc d$	$* *    * *   *$
$[b, b, d, d, d]$	$ bb  ddd$	$  * *    * * *$
$[a, a, a, a, a]$	$aaaaa   $	$* * * * *    $

This suggests that we can encode the multisets as lists made from the two symbols  $*$  and  $|$ , with an  $*$  for each element of the multiset, as follows.

$$\begin{array}{cccc} * \text{ for each } a & * \text{ for each } b & * \text{ for each } c & * \text{ for each } d \\ \overbrace{* \cdots \cdots *} & \overbrace{| * \cdots \cdots *} & \overbrace{| * \cdots \cdots *} & \overbrace{| * \cdots \cdots *} \end{array}$$

For examples see the right-hand column of the table. Any such encoding is a list made from 5 stars and 3 bars, so the list has a total of 8 entries. How many such lists are there? We can form such a list by choosing 3 of the 8 positions for the bars, and filling the remaining three positions with stars. Therefore the number of such lists is  $\binom{8}{3} = \frac{8!}{3!5!} = 56$ .

That is our answer. **There are 56 cardinality-5 multisets** that can be made from the symbols in  $X = \{a, b, c, d\}$ .

If we wanted to count the cardinality-3 multisets made from  $X$ , then the exact same reasoning would apply, but with 3 stars instead of 5. We'd be counting the length-6 lists with 3 stars and 3 bars. There are  $\binom{6}{3} = \frac{6!}{3!3!} = 20$  such lists. So there are 20 cardinality-3 multisets made from  $X = \{a, b, c, d\}$ . This agrees with our accounting on the previous page.

In general, given a set  $X = \{x_1, x_2, \dots, x_n\}$  of  $n$  elements, any cardinality- $k$  multiset made from its elements can be encoded in a star-and-bar list

$$\underbrace{* * * *}_{\text{* for each } x_1} | \underbrace{* * * *}_{\text{* for each } x_2} | \underbrace{* * * *}_{\text{* for each } x_3} | \dots \dots | \underbrace{* * * *}_{\text{* for each } x_n}.$$

Such a list has  $k$  stars (one for each element of the multiset) and  $n - 1$  separating bars (a bar between each of the  $n$  groupings of stars). Therefore its length is  $k + n - 1$ . We can make such a list by selecting  $n - 1$  list positions out of  $k + n - 1$  positions for the bars and inserting stars in the left-over positions. Thus there are  $\binom{k+n-1}{n-1}$  such lists. Alternatively we could choose  $k$  positions for the stars and fill in the remaining  $n - k$  with bars, so there are  $\binom{k+n-1}{k}$  such lists. Note that  $\binom{k+n-1}{k} = \binom{k+n-1}{n-1}$  by Equation (3.4) on page 91.

Let's summarize our reckoning.

**Fact 3.7** The number of  $k$ -element multisets that can be made from the elements of an  $n$ -element set  $X = \{x_1, x_2, \dots, x_n\}$  is

$$\binom{k+n-1}{k} = \binom{k+n-1}{n-1}.$$

This works because any cardinality- $k$  multiset made from the  $n$  elements of  $X$  can be encoded in a star-and-bar list of length  $k + n - 1$ , having form

$$\underbrace{* * * *}_{\text{* for each } x_1} | \underbrace{* * * *}_{\text{* for each } x_2} | \underbrace{* * * *}_{\text{* for each } x_3} | \dots \dots | \underbrace{* * * *}_{\text{* for each } x_n}$$

with  $k$  stars and  $n - 1$  bars separating the  $n$  groupings of stars. Such a list can be made by selecting  $n - 1$  positions for the bars, and filling the remaining positions with stars, and there are  $\binom{n+k-1}{n-1}$  ways to do this.

For example, the number of 2-element multisets that can be made from the 4-element set  $X = \{a, b, c, d\}$  is  $\binom{2+4-1}{2} = \binom{5}{2} = 10$ . This agrees with our accounting of them on page 96. The number of 3-element multisets that can be made from the elements of  $X$  is  $\binom{3+4-1}{3} = \binom{6}{3} = 20$ . Again this agrees with our list of them on page 96.

The number of 1-element multisets made from  $X$  is  $\binom{1+4-1}{1} = \binom{4}{1} = 4$ . Indeed, the four multisets are  $[a], [b], [c]$  and  $[d]$ . The number of 0-element multisets made from  $X$  is  $\binom{0+4-1}{0} = \binom{3}{0} = 1$ . This is right, because there is only one such multiset, namely  $\emptyset$ .

**Example 3.19** A bag contains 20 identical red marbles, 20 identical green marbles, and 20 identical blue marbles. You reach in and grab 20 marbles. There are many possible outcomes. You could have 11 reds, 4 greens and 5 blues. Or you could have 20 reds, 0 greens and 0 blues, etc. All together, how many outcomes are possible?

**Solution:** Each outcome can be thought of as a 20-element multiset made from the elements of the 3-element set  $X = \{R, G, B\}$ . For example, 11 reds, 4 greens and 5 blues would correspond to the multiset

$$[ R, G, G, G, G, B, B, B, B, B ].$$

The outcome consisting of 10 reds and 10 blues corresponds to the multiset

$$[ R, R, R, R, R, R, R, R, R, B, B, B, B, B, B, B, B, B ].$$

Thus the total number of outcomes is the number of 20-element multisets made from the elements of the 3-element set  $X = \{R, G, B\}$ . By Fact 3.7, the answer is  $\binom{20+3-1}{20} = \binom{22}{20} = 231$  possible outcomes.

Rather than remembering the formula in Fact 3.7, it is probably best to work out a new stars-and-bars model as needed. This is because it is often easy to see how a particular problem can be modeled with stars and bars, and once they have been set up, the formula in Fact 3.7 falls out automatically.

For instance, we could solve Example 3.19 by noting that each outcome has a star-and-bar encoding using 20 stars and 2 bars. (The outcome  $[R, R, R, R, R, R, R, R, R, R, G, G, G, G, B, B, B, B, B]$  can be encoded in stars and bars as  $* * * * * * * * * | * * * * | * * * * *,$  etc.) We can form such a list by choosing 2 out of 22 slots for bars and filling the remaining 20 slots with stars. There are  $\binom{22}{2} = 231$  ways of doing this.

Our next example involves counting the number of *non-negative* integer solutions of the equation  $w + x + y + z = 20$ . By a *non-negative integer solution* to the equation, we mean an assignment of non-negative integers to the variables that makes the equation true. For example, one solution is  $w = 7, x = 3, y = 5, z = 5$ . We can write this solution compactly as  $(w, x, y, z) = (7, 3, 5, 5)$ . Two other solutions are  $(w, x, y, z) = (1, 3, 1, 15)$  and  $(w, x, y, z) = (0, 20, 0, 0)$ . We would not include  $(w, x, y, z) = (1, -1, 10, 10)$  as a solution because even though it satisfies the equation, the value of  $x$  is negative. How many solutions are there all together? The next example presents a way of solving this type of question.

**Example 3.20** How many non-negative integer solutions does the equation  $w + x + y + z = 20$  have?

**Solution:** We can model a solution with stars and bars. For example, encode the solution  $(w, x, y, z) = (3, 4, 5, 8)$  as

$$\overbrace{***}^3 | \overbrace{****}^4 | \overbrace{*****}^5 | \overbrace{*****}^8.$$

In general, any solution  $(w, x, y, z) = (a, b, c, d)$  gets encoded as

$$\overbrace{***\dots*}^a | \overbrace{***\dots*}^b | \overbrace{***\dots*}^c | \overbrace{***\dots*}^d,$$

where all together there are 20 stars and 3 bars. So, for instance the solution  $(w, x, y, z) = (0, 0, 10, 10)$  gets encoded as  $||*****|*****|*****|*****$ , and the solution  $(w, x, y, z) = (7, 3, 5, 5)$  is encoded as  $*****|***|***|***$ . Thus we can describe any non-negative integer solution to the equation as a list of length  $20 + 3 = 23$  that has 20 stars and 3 bars. We can make any such list by choosing 3 out of 23 spots for the bars, and filling the remaining 20 spots with stars. The number of ways to do this is  $\binom{23}{3} = \frac{23!}{3!20!} = \frac{23 \cdot 22 \cdot 21}{3 \cdot 2} = 23 \cdot 11 \cdot 7 = 1771$ . Thus there are **1771** non-negative integer solutions of  $w + x + y + z = 20$ .

For another approach to this example, model solutions of  $w + x + y + z = 20$  as 20-element multisets made from the elements of  $\{w, x, y, z\}$ . For example, solution  $(5, 5, 4, 6)$  corresponds to  $[w, w, w, w, w, x, x, x, x, x, y, y, y, y, z, z, z, z, z]$ . By Fact 3.7, there are  $\binom{20+4-1}{20} = \binom{23}{20} = 1771$  such multisets, so this is the number of solutions to  $w + x + y + z = 20$ .

**Example 3.21** This problem concerns the lists  $(w, x, y, z)$  of integers with the property that  $0 \leq w \leq x \leq y \leq z \leq 10$ . That is, each entry is an integer between 0 and 10, and the entries are ordered from smallest to largest. For example,  $(0, 3, 3, 7)$ ,  $(1, 1, 1, 1)$  and  $(2, 3, 6, 9)$  have this property, but  $(2, 3, 6, 4)$  does not. How many such lists are there?

**Solution:** We can encode such a list with 10 stars and 4 bars, where  $w$  is the number of stars to the left of the first bar,  $x$  is the number of stars to the left of the second bar,  $y$  is the number of stars to the left of the third bar, and  $z$  is the number of stars to the left of the fourth bar.

For example,  $(2, 3, 6, 9)$  is encoded as  $* * | * | * * * | * * * | *$ , and  $(1, 2, 3, 4)$  is encoded as  $* | * | * | * | * * * * *$ .

Here are some other examples of lists paired with their encodings.

(0,3,3,7)	* * *     * * * *   * * *
(1,1,1,1)	*       * * * * * * * *
(9,9,9,10)	* * * * * * * *       *

Such encodings are lists of length 14, with 10 stars and 4 bars. We can make such a list by choosing 4 of the 14 slots for the bars and filling the remaining slots with stars. The number of ways to do this is  $\binom{14}{4} = 1001$ . Answer: There are **1001** such lists.

We will examine one more type of multiset problem. To motivate it, consider the permutations of the letters of the word “BOOK.” At first glance there are 4 letters, so we should get  $4! = 24$  permutations. But this is not quite right because two of the letters are identical. We could interchange the two O’s but still have the same permutation. To get a grip on the problem, let’s make one of the letters lower case: BOoK. Now our 24 permutations are listed below in the oval.

BOoK	KOoB	OoKB	OoBK	OBoK	OKoB	OKBo	OBKo	BKOo	KBOo	KOBo	BOKo
BoOK	KoOB	oOKB	oOBK	oBOK	oKOB	oKBO	oBKO	BKoO	KBoO	KBoO	KoBO
BOOK	KOOB	OOKB	OOBK	OBOK	OKOB	OKBO	OBKO	BKOO	KBOO	KOBO	BOKO

The columns in the oval correspond to the same permutation of the letters of BOOK, as indicated in the row below the oval. Thus there are actually  $\frac{4!}{2} = \frac{24}{2} = 12$  permutations of the letters of BOOK.

This is actually a problem about multisets. The letters in “BOOK” form a multiset [B,O,O,K], and we have determined that there are 12 permutations of this multiset.

For another motivational example, consider the permutations of the letters of the word BANANA. Here there are two N’s and three A’s. Though some of the letters look identical, think of them as distinct physical objects that we can permute into different orderings. It helps to subscript the letters to emphasize that they are actually six distinct objects:

$$B A_1 N_1 A_2 N_2 A_3.$$

Now, there are  $6! = 720$  permutations of these six letters. It’s not practical to write out all of them, but we can get a sense of the problem by making a partial listing in the box below.

BA <sub>1</sub> N <sub>1</sub> A <sub>2</sub> N <sub>2</sub> A <sub>3</sub>	A <sub>1</sub> B N <sub>1</sub> A <sub>2</sub> N <sub>2</sub> A <sub>3</sub>	...
BA <sub>1</sub> N <sub>1</sub> A <sub>3</sub> N <sub>2</sub> A <sub>2</sub>	A <sub>1</sub> B N <sub>1</sub> A <sub>3</sub> N <sub>2</sub> A <sub>2</sub>	...
BA <sub>2</sub> N <sub>1</sub> A <sub>1</sub> N <sub>2</sub> A <sub>3</sub>	A <sub>2</sub> B N <sub>1</sub> A <sub>1</sub> N <sub>2</sub> A <sub>3</sub>	...
BA <sub>2</sub> N <sub>1</sub> A <sub>3</sub> N <sub>2</sub> A <sub>1</sub>	A <sub>2</sub> B N <sub>1</sub> A <sub>3</sub> N <sub>2</sub> A <sub>1</sub>	...
BA <sub>3</sub> N <sub>1</sub> A <sub>2</sub> N <sub>2</sub> A <sub>1</sub>	A <sub>3</sub> B N <sub>1</sub> A <sub>2</sub> N <sub>2</sub> A <sub>1</sub>	...
BA <sub>3</sub> N <sub>1</sub> A <sub>1</sub> N <sub>2</sub> A <sub>2</sub>	A <sub>3</sub> B N <sub>1</sub> A <sub>1</sub> N <sub>2</sub> A <sub>2</sub>	720 permutations of BA <sub>1</sub> N <sub>1</sub> A <sub>2</sub> N <sub>2</sub> A <sub>3</sub>
BA <sub>1</sub> N <sub>2</sub> A <sub>2</sub> N <sub>1</sub> A <sub>3</sub>	A <sub>1</sub> B N <sub>2</sub> A <sub>2</sub> N <sub>1</sub> A <sub>3</sub>	...
BA <sub>1</sub> N <sub>2</sub> A <sub>3</sub> N <sub>1</sub> A <sub>2</sub>	A <sub>1</sub> B N <sub>2</sub> A <sub>3</sub> N <sub>1</sub> A <sub>2</sub>	...
BA <sub>2</sub> N <sub>2</sub> A <sub>1</sub> N <sub>1</sub> A <sub>3</sub>	A <sub>2</sub> B N <sub>2</sub> A <sub>1</sub> N <sub>1</sub> A <sub>3</sub>	...
BA <sub>2</sub> N <sub>2</sub> A <sub>3</sub> N <sub>1</sub> A <sub>1</sub>	A <sub>2</sub> B N <sub>2</sub> A <sub>3</sub> N <sub>1</sub> A <sub>1</sub>	...
BA <sub>3</sub> N <sub>2</sub> A <sub>2</sub> N <sub>1</sub> A <sub>1</sub>	A <sub>3</sub> B N <sub>2</sub> A <sub>2</sub> N <sub>1</sub> A <sub>1</sub>	...
BA <sub>3</sub> N <sub>2</sub> A <sub>1</sub> N <sub>1</sub> A <sub>2</sub>	A <sub>3</sub> B N <sub>2</sub> A <sub>1</sub> N <sub>1</sub> A <sub>2</sub>	...

BANANA

ABNANA

The first column lists the permutations of BA<sub>1</sub>N<sub>1</sub>A<sub>2</sub>N<sub>2</sub>A<sub>3</sub> corresponding to the word BANANA. By the multiplication principle, the column has  $3!2! = 12$  permutations because the three  $A_i$ 's can be permuted in 3! ways within their positions, and the two  $N_i$ 's can be permuted in 2! ways. Similarly, the second column lists the  $3!2! = 12$  permutations corresponding to the “word” ABNANA.

All together there are  $6! = 720$  permutations of BA<sub>1</sub>N<sub>1</sub>A<sub>2</sub>N<sub>2</sub>A<sub>3</sub>, and groupings of 12 of them correspond to particular permutations of BANANA. Therefore the total number of permutations of BANANA is  $\frac{6!}{3!2!} = \frac{720}{12} = 60$ .

The kind of reasoning used here generalizes to the following fact.

**Fact 3.8** Suppose a multiset  $A$  has  $n$  elements, with multiplicities  $p_1, p_2, \dots, p_k$ . Then the total number of permutations of  $A$  is

$$\frac{n!}{p_1! p_2! \cdots p_k!}.$$

**Example 3.22** Count the permutations of the letters in MISSISSIPPI.

**Solution:** Think of this word as an 11-element multiset with one M, four I's, four S's and two P's. By Fact 3.8, it has  $\frac{11!}{1!4!4!2!} = 34,650$  permutations.

**Example 3.23** Determine the number of permutations of the multiset [1, 1, 1, 1, 5, 5, 10, 25, 25].

**Solution:** By Fact 3.8 the answer is  $\frac{9!}{4!2!1!2!} = 3780$ .

---

### Exercises for Section 3.8

1. How many 10-element multisets can be made from the symbols  $\{1, 2, 3, 4\}$ ?
  2. How many 2-element multisets can be made from the 26 letters of the alphabet?
  3. You have a dollar in pennies, a dollar in nickels, a dollar in dimes, and a dollar in quarters. You give a friend four of your coins. In how many ways can this be done?
  4. A bag contains 20 identical red balls, 20 identical blue balls, 20 identical green balls, and 20 identical white balls. You reach in and grab 15 balls. How many different outcomes are possible?
  5. A bag contains 20 identical red balls, 20 identical blue balls, 20 identical green balls, and **one** white ball. You reach in and grab 15 balls. How many different outcomes are possible?
  6. A bag contains 20 identical red balls, 20 identical blue balls, 20 identical green balls, one white ball, and one black ball. You reach in and grab 20 balls. How many different outcomes are possible?
  7. In how many ways can you place 20 identical balls into five different boxes?
  8. How many lists  $(x, y, z)$  of three integers are there with  $0 \leq x \leq y \leq z \leq 100$ ?
  9. A bag contains 50 pennies, 50 nickels, 50 dimes and 50 quarters. You reach in and grab 30 coins. How many different outcomes are possible?
  10. How many non-negative integer solutions does  $u + v + w + x + y + z = 90$  have?
  11. How many integer solutions does the equation  $w + x + y + z = 100$  have if  $w \geq 4$ ,  $x \geq 2$ ,  $y \geq 0$  and  $z \geq 0$ ?
  12. How many integer solutions does the equation  $w + x + y + z = 100$  have if  $w \geq 7$ ,  $x \geq 0$ ,  $y \geq 5$  and  $z \geq 4$ ?
  13. How many length-6 lists can be made from the symbols  $\{A, B, C, D, E, F, G\}$ , if repetition is allowed and the list is in alphabetical order? (Examples: BBCEGG, but not BBBAGG.)
  14. How many permutations are there of the letters in the word “PEPPERMINT”?
  15. How many permutations are there of the letters in the word “TENNESSEE”?
  16. A community in Canada’s Northwest Territories is known in the local language as “TUKTUYAAQTUUQ.” How many permutations does this name have?
  17. You roll a dice six times in a row. How many possible outcomes are there that have two 1’s three 5’s and one 6?
  18. Flip a coin ten times in a row. How many outcomes have 3 heads and 7 tails?
  19. In how many ways can you place 15 identical balls into 20 different boxes if each box can hold at most one ball?
  20. You distribute 25 identical pieces of candy among five children. In how many ways can this be done?
  21. How many numbers between 10,000 and 99,999 contain one or more of the digits 3, 4 and 8, but no others?
-

### 3.9 The Division and Pigeonhole Principles

Our final fundamental counting principle is called the **division principle**. Before discussing it, we need some notation. Given a number  $x$ , its **floor**  $\lfloor x \rfloor$  is  $x$  rounded down to the nearest integer. Thus  $\lfloor \frac{10}{4} \rfloor = 2$ , and  $\lfloor 9.31 \rfloor = 9$ , and  $\lfloor 7 \rfloor = 7$ , etc. The **ceiling** of  $x$ , denoted  $\lceil x \rceil$ , is  $x$  rounded up to the nearest integer. Thus  $\lceil \frac{10}{4} \rceil = 3$ , and  $\lceil 9.31 \rceil = 10$ , and  $\lceil 7 \rceil = 7$ .

The division principle is often illustrated by a simple situation involving pigeons. Imagine  $n$  pigeons that live in  $k$  pigeonholes, or boxes. (Possibly  $n \neq k$ .) At night all the pigeons fly into the boxes. When this happens, some of the  $k$  boxes may contain more than one pigeon, and some may be empty. But no matter what, the average number of pigeons per box is  $\frac{n}{k}$ . Obviously, at least one of the boxes contains  $\frac{n}{k}$  or more pigeons. (Because not all the boxes can contain fewer than the average number of pigeons per box.) And because a box must contain a whole number of pigeons, we round up to conclude that at least one box contains  $\lceil \frac{n}{k} \rceil$  or more pigeons.

Similarly, at least one box must contain  $\frac{n}{k}$  or fewer pigeons, because not all boxes can contain more than the average number of pigeons per box. Rounding down, at least one box contains  $\lfloor \frac{n}{k} \rfloor$  or fewer pigeons.

We call this line of reasoning the *division principle*. (Some texts call it the *strong form of the pigeonhole principle*.)

#### Fact 3.9 (Division Principle)

Suppose  $n$  objects are placed into  $k$  boxes.

Then at least one box contains  $\lceil \frac{n}{k} \rceil$  or more objects,  
and at least one box contains  $\lfloor \frac{n}{k} \rfloor$  or fewer objects.

This has a useful variant. If  $n > k$ , then  $\frac{n}{k} > 1$ , so  $\lceil \frac{n}{k} \rceil > 1$ , and this means some box contains more than one object. On the other hand, if  $n < k$  then  $\frac{n}{k} < 1$ , so  $\lfloor \frac{n}{k} \rfloor < 1$ , meaning at least one box is empty. Thus the division principle yields the following consequence, called the *pigeonhole principle*.

#### Fact 3.10 (Pigeonhole Principle)

Suppose  $n$  objects are placed into  $k$  boxes.

If  $n > k$ , then at least one box contains more than one object.

If  $n < k$ , then at least one box is empty.

The pigeonhole principle is named for the scenario in which  $n$  pigeons fly into  $k$  pigeonholes (or boxes). If there are more pigeons than boxes ( $n > k$ )

then some box gets more than one pigeon. And if there are fewer pigeons than boxes ( $n < k$ ) then there must be at least one empty box.

Like the multiplication, addition and subtraction principles, the division and pigeonhole principles are intuitive and obvious, but they can prove things that are not obvious. The challenge is seeing where and how to apply them. Our examples will start simple and get progressively more complex.

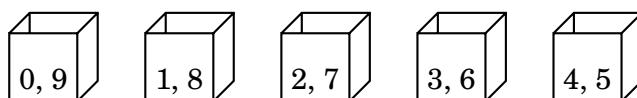
For an extremely simple application, notice that in any group of 13 people, at least two of them were born on the same month. Although this is obvious, it really does follow from the pigeonhole principle. Think of the 13 people as objects, and put each person in the “box” that is his birth month. As there are more people than boxes (months), at least one box (month) has two or more people in it, meaning at least two of the 13 people were born in the same month.

Further, for any group of 100 people, the division principle says that there is a month in which  $\lceil \frac{100}{12} \rceil = 9$  or more of the people were born. It also guarantees a month in which  $\lfloor \frac{100}{12} \rfloor = 8$  or fewer of the people were born.

**Example 3.24** Pick six numbers between 0 and 9 (inclusive). Show that two of them must add up to 9.

For example, suppose you picked 0, 1, 3, 5, 7 and 8. Then  $1 + 8 = 9$ . If you picked 4, 5, 6, 7, 8, 9, then  $4 + 5 = 9$ . The problem asks us to show that this happens no matter how we pick the numbers.

**Solution:** Pick six numbers between 0 and 9. Here’s why two of them sum to 9: Imagine five boxes, each marked with two numbers, as shown below. Each box is labeled so that the two numbers written on it sum to 9.

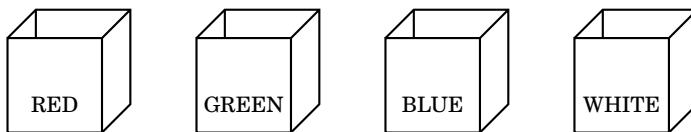


For each number that was picked, put it in the box having that number written on it. For example, if we picked 7, it goes in the box labeled “2, 7.” (The number 2, if picked, would go in that box too.) In this way we place the six chosen numbers in five boxes. As there are more numbers than boxes, the pigeonhole principle says that some box has more than one (hence two) of the picked numbers in it. Those two numbers sum to 9.

Notice that if we picked only five numbers from 0 to 9, then it’s possible that no two sum to 9: we could be unlucky and pick 0, 1, 2, 3, 4. But the pigeonhole principle ensures that if six are picked then two do sum to 9.

**Example 3.25** A store has a gumball machine containing a large number of red, green, blue and white gumballs. You get one gumball for each nickel you put into the machine. The store offers the following deal: You agree to buy some number of gumballs, and if 13 or more of them have the same color you get \$5. What is the fewest number of gumballs you need to buy to be 100% certain that you will make money on the deal?

**Solution:** Let  $n$  be the number of gumballs that you buy. Imagine sorting your  $n$  gumballs into four boxes labeled RED, GREEN, BLUE, and WHITE. (That is, red balls go in the red box, green balls go in the green box, etc.)



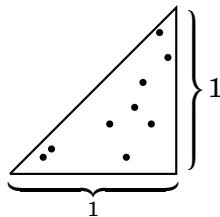
The division principle says that one box contains  $\lceil \frac{n}{4} \rceil$  or more gumballs. Provided  $\lceil \frac{n}{4} \rceil \geq 13$ , you will know you have 13 gumballs of the same color. This happens if  $\frac{n}{4} > 12$  (so the ceiling of  $\frac{n}{4}$  rounds to a value larger than 12). Therefore you need  $n > 4 \cdot 12 = 48$ , so if  $n = 49$  you know you have at least  $\lceil \frac{49}{4} \rceil = \lceil 12.25 \rceil = 13$  gumballs of the same color.

**Answer:** Buy 49 gumballs for 49 nickels, which is \$2.45. You get \$5, and therefore have made \$2.55.

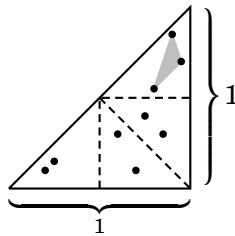
Note that if you bought just 48 gumballs, you might win, but there is a chance that you'd get 12 gumballs of each color and miss out on the \$5. And if you bought more than 49, you'd still get the \$5, but you would have spent more nickels.

Explicitly mentioning the boxes in the above solution is not necessary. Some people prefer to draw a conclusion based averaging alone. They might solve the problem by letting  $n$  be the number of gumballs bought, so  $n = r + g + b + w$ , where  $r$  is the number of them that are red,  $g$  is the number that are green,  $b$  is the number of blues and  $w$  is the number of whites. Then the average number of gumballs of a particular color is  $\frac{r+g+b+w}{4} = \frac{n}{4}$ . We need this to be greater than 12 to ensure 13 of the same color, and the smallest number that does the job is  $n = 49$ . This is still the division principle, in a pure form.

**Example 3.26** Nine points are randomly placed on the right triangle shown below. Show that three of these points form a triangle whose area is  $\frac{1}{8}$  square unit or less. (We allow triangles with zero area, in which case the three points lie on a line.)



**Solution:** Divide the triangle into four smaller triangles, as indicated by the dashed lines below. Each of these four triangles has an area of



$\frac{1}{2}bh = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$  square units. Think of these smaller triangles as “boxes.” So we have placed 9 points in 4 boxes. (If one of the 9 points happens to be on a dashed line, say it belongs to the box below or to its left.) The division principle says one of the boxes has at least  $\lceil \frac{9}{4} \rceil = 3$  of the points in it. Those three points form a triangle whose area is no larger than the area of the “box” that it is in. Thus these three points form a triangle whose area is  $\frac{1}{8}$  or less.

### Exercises for Section 3.9

1. Show that if six numbers are chosen at random, then at least two of them will have the same remainder when divided by 5.
2. You deal a pile of cards, face down, from a standard 52-card deck. What is the least number of cards the pile must have before you can be assured that it contains at least five cards of the same suit?
3. What is the fewest number of times you must roll a six-sided dice before you can be assured that 10 or more of the rolls resulted in the same number?
4. Select any five points on a square whose side-length is one unit. Show that at least two of these points are within  $\frac{\sqrt{2}}{2}$  units of each other.
5. Prove that any set of seven distinct natural numbers contains a pair of numbers whose sum or difference is divisible by 10.
6. Given a sphere  $S$ , a *great circle* of  $S$  is the intersection of  $S$  with a plane through its center. Every great circle divides  $S$  into two parts. A *hemisphere* is the union of the great circle and one of these two parts. Show that if five points are placed arbitrarily on  $S$ , then there is a hemisphere that contains four of them.

### 3.10 Combinatorial Proof

**Combinatorial proof** is a method of proving two different expressions are equal by showing that they are both answers to the same counting question. We have already used combinatorial proof (without *calling* it combinatorial proof) in proving Pascal's formula  $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$  on page 90.

There we argued that the left-hand side  $\binom{n+1}{k}$  is, by definition, the number of  $k$ -element subsets of the set  $S = \{0, 1, 2, \dots, n\}$  with  $|S| = n+1$ . But the right-hand side also gives the number of  $k$ -element subsets of  $S$ , because such a subset either contains 0 or it does not. We can make any  $k$ -element subset of  $S$  that contains 0 by starting with 0 and selecting  $k-1$  other elements from  $\{1, 2, \dots, n\}$ , in  $\binom{n}{k-1}$  ways. We can make any  $k$ -element subset that does not contain 0 by selecting  $k$  elements from  $\{1, 2, \dots, n\}$ , and there are  $\binom{n}{k}$  ways to do this. Thus,

$$\underbrace{\binom{n+1}{k}}_{\substack{\text{number of} \\ k\text{-element} \\ \text{subsets of} \\ S = \{0, 1, \dots, n\}}} = \underbrace{\binom{n}{k-1}}_{\substack{\text{number of} \\ k\text{-element} \\ \text{subsets of} \\ S \text{ with } 0}} + \underbrace{\binom{n}{k}}_{\substack{\text{number of} \\ k\text{-element} \\ \text{subsets of} \\ S \text{ without } 0}}$$

Both sides count the number of  $k$ -element subsets of  $S$ , so they are equal. This is combinatorial proof.

**Example 3.27** Use combinatorial proof to show  $\binom{n}{k} = \binom{n}{n-k}$ .

**Solution.** First, by definition, if  $k < 0$  or  $k > n$ , then both sides are 0, and thus equal. Therefore for the rest of the proof we can assume  $0 \leq k \leq n$ .

The left-hand side  $\binom{n}{k}$  is the number of  $k$ -element subsets of  $S = \{1, 2, \dots, n\}$ . Every  $k$ -element subset  $X \subseteq S$  pairs with a unique  $(n-k)$ -element subset  $\bar{X} = S - X \subseteq S$ . Thus the number of  $k$ -element subsets of  $S$  equals the number of  $(n-k)$ -element subsets of  $S$ , which is to say  $\binom{n}{k} = \binom{n}{n-k}$ .

We could also derive  $\binom{n}{k} = \binom{n}{n-k}$  by using the formula for  $\binom{n}{k}$  and quickly get

$$\binom{n}{n-k} = \frac{n!}{(n-k)! (n-(n-k))!} = \frac{n!}{(n-k)! k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

But you may feel that the combinatorial proof is “slicker” because it uses the *meanings* of the terms. Often it is flat-out easier than using formulas, as in the next example.

Our next example will prove that  $\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}$ , for any positive integer  $n$ , which is to say that  $\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}$ . For example, if  $n = 5$ , this statement asserts  $\binom{5}{0}^2 + \binom{5}{1}^2 + \binom{5}{2}^2 + \binom{5}{3}^2 + \binom{5}{4}^2 + \binom{5}{5}^2 = \binom{10}{5}$ , which is

$$1^2 + 5^2 + 10^2 + 10^2 + 5^2 + 1^2 = \binom{10}{5},$$

and this is true, as both sides equal 252. In general, the statement says the squares of the entries in the  $n$ th row of Pascal's triangle add up to  $\binom{2n}{n}$ .

**Example 3.28** Use a combinatorial proof to show that  $\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}$ .

First, the right-hand side  $\binom{2n}{n}$  is the number of ways to select  $n$  things from a set  $S$  that has  $2n$  elements.

Now let's count this a different way. Divide  $S$  into two equal-sized parts,  $S = A \cup B$ , where  $|A| = n$  and  $|B| = n$ , and  $A \cap B = \emptyset$ .

For any fixed  $k$  with  $0 \leq k \leq n$ , we can select  $n$  things from  $S$  by taking  $k$  things from  $A$  and  $n - k$  things from  $B$  for a total of  $k + (n - k) = n$  things. By the multiplication principle, we get  $\binom{n}{k} \binom{n}{n-k}$   $n$ -element subsets of  $S$  this way.

As  $k$  could be any number from 0 to  $n$ , the number of ways to select  $n$  things from  $S$  is thus

$$\underbrace{\binom{n}{0} \binom{n}{n-0}}_{\substack{0 \text{ from } A \\ n \text{ from } B}} + \underbrace{\binom{n}{1} \binom{n}{n-1}}_{\substack{1 \text{ from } A \\ n-1 \text{ from } B}} + \underbrace{\binom{n}{2} \binom{n}{n-2}}_{\substack{2 \text{ from } A \\ n-2 \text{ from } B}} + \underbrace{\binom{n}{3} \binom{n}{n-3}}_{\substack{3 \text{ from } A \\ n-3 \text{ from } B}} + \cdots + \underbrace{\binom{n}{n} \binom{n}{0}}_{\substack{n \text{ from } A \\ 0 \text{ from } B}}.$$

But because  $\binom{n}{n-k} = \binom{n}{k}$ , this expression equals  $\binom{n}{0} \binom{n}{0} + \binom{n}{1} \binom{n}{1} + \binom{n}{2} \binom{n}{2} + \cdots + \binom{n}{n} \binom{n}{n}$ , which is  $\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2 = \sum_{k=0}^n \binom{n}{k}^2$ .

In summary, we've counted the ways to choose  $n$  elements from the set  $S$  with two methods. One method gives  $\binom{2n}{n}$ , and the other gives  $\sum_{k=0}^n \binom{n}{k}^2$ .

Therefore  $\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}$ .

Be on the lookout for opportunities to use combinatorial proof, and watch for it in your readings outside of this course. Also, try some of the exercises below. Sometimes it takes some creative thinking and false starts before you hit on an idea that works, but once you find it the solution is usually remarkably simple.

### Exercises for Section 3.10

Use combinatorial proof to solve the following problems. You may assume that any variables  $m, n, k$  and  $p$  are non-negative integers.

1. Show that  $1(n - 0) + 2(n - 1) + 3(n - 2) + 4(n - 3) + \cdots + (n - 1)2 + (n - 0)1 = \binom{n+2}{3}$ .
2. Show that  $1 + 2 + 3 + \cdots + n = \binom{n+1}{2}$ .
3. Show that  $\binom{n}{2}\binom{n-2}{k-2} = \binom{n}{k}\binom{k}{2}$ .
4. Show that  $P(n, k) = P(n - 1, k) + k \cdot P(n - 1, k - 1)$ .
5. Show that  $\binom{2n}{2} = 2\binom{n}{2} + n^2$ .
6. Show that  $\binom{3n}{3} = 3\binom{n}{3} + 6n\binom{n}{2} + n^3$ .
7. Show that  $\sum_{k=0}^p \binom{m}{k} \binom{n}{p-k} = \binom{m+n}{p}$ .
8. Show that  $\sum_{k=0}^m \binom{m}{k} \binom{n}{p+k} = \binom{m+n}{m+p}$ .
9. Show that  $\sum_{k=m}^n \binom{k}{m} = \binom{n+1}{m+1}$ .
10. Show that  $\sum_{k=1}^n k \binom{n}{k} = n2^{n-1}$ .
11. Show that  $\sum_{k=0}^n 2^k \binom{n}{k} = 3^n$ .
12. Show that  $\sum_{k=0}^n \binom{n}{k} \binom{k}{m} = \binom{n}{m} 2^{n-m}$ .

## *Part II*

---

### *How to Prove Conditional Statements*

---



# CHAPTER 4

---

## Direct Proof

---

**I**t is time to prove some theorems. There are various strategies for doing this; we now examine the most straightforward approach, a technique called *direct proof*. As we begin, it is important to keep in mind the meanings of three key terms: *Theorem*, *proof* and *definition*.

A **theorem** is a mathematical statement that is true and can be (and has been) verified as true. A **proof** of a theorem is a written verification that shows that the theorem is definitely and unequivocally true. A proof should be understandable and convincing to anyone who has the requisite background and knowledge. This knowledge includes an understanding of the meanings of the mathematical words, phrases and symbols that occur in the theorem and its proof. It is crucial that both the writer of the proof and the readers of the proof agree on the exact meanings of all the words, for otherwise there is an intolerable level of ambiguity. A **definition** is an exact, unambiguous explanation of the meaning of a mathematical word or phrase. We will elaborate on the terms *theorem* and *definition* in the next two sections, and then finally we will be ready to begin writing proofs.

### 4.1 Theorems

A **theorem** is a statement that is true and has been proved to be true. You have encountered many theorems in your mathematical education. Here are some theorems taken from an undergraduate calculus text. They will be familiar to you, though you may not have read all the proofs.

**Theorem:** Let  $f$  be differentiable on an open interval  $I$  and let  $c \in I$ . If  $f(c)$  is the maximum or minimum value of  $f$  on  $I$ , then  $f'(c) = 0$ .

**Theorem:** If  $\sum_{k=1}^{\infty} a_k$  converges, then  $\lim_{k \rightarrow \infty} a_k = 0$ .

**Theorem:** Suppose  $f$  is continuous on the interval  $[a, b]$ . Then  $f$  is integrable on  $[a, b]$ .

**Theorem:** Every absolutely convergent series converges.

Observe that each of these theorems either has the conditional form “*If  $P$ , then  $Q$* ,” or can be put into that form. The first theorem has an initial sentence “*Let  $f$  be differentiable on an open interval  $I$ , and let  $c \in I$* ,” which sets up some notation, but a conditional statement follows it. The third theorem has form “*Suppose  $P$ . Then  $Q$* ,” but this means the same thing as “*If  $P$ , then  $Q$* .” The last theorem can be re-expressed as “*If a series is absolutely convergent, then it is convergent*.”

For another example, the equation  $\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}$  from Example 3.28 (page 109) is best phrased as a conditional statement, to make clear the assumption that  $n$  is an *integer*.

**Theorem:** If  $n$  is a non-negative integer, then  $\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}$ .

A theorem of form “*If  $P$ , then  $Q$* ,” can be regarded as a device that produces new information from  $P$ . Whenever we are dealing with a situation in which  $P$  is true, then the theorem guarantees that, in addition,  $Q$  is true. Since this kind of expansion of information is useful, theorems of form “*If  $P$ , then  $Q$* ,” are very common.

But not *every* theorem is a conditional statement. Some have the form of the biconditional  $P \Leftrightarrow Q$ , but, as we know, that can be expressed as *two* conditional statements. Other theorems simply state facts about specific things. For example, here is another theorem from your study of calculus.

**Theorem:** The harmonic series  $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots$  diverges.

It would be difficult (or at least awkward) to restate this as a conditional statement. Still, it is true that most theorems are conditional statements, so much of this book will concentrate on that type of theorem.

It is important to be aware that there are a number of words that mean essentially the same thing as the word “theorem,” but are used in slightly different ways. In general the word “theorem” is reserved for a statement that is considered important or significant (the Pythagorean theorem, for example). A statement that is true but not as significant is sometimes called a **proposition**. A **lemma** is a theorem whose main purpose is to help prove another theorem. A **corollary** is a result that is an immediate consequence of a theorem or proposition. It is not important that you remember all these words now, for their meanings will become clear with usage.

Our main task is to learn how to prove theorems. As the above examples suggest, proving theorems requires a clear understanding of the conditional statement, and that is the main reason we studied it so extensively in Chapter 2. It is also crucial to understand the role of definitions.

## 4.2 Definitions

A proof of a theorem should be absolutely convincing. Ambiguity must be avoided. Everyone must agree on the exact meaning of each mathematical term. In Chapter 1 we defined the meanings of the sets  $\mathbb{N}$ ,  $\mathbb{Z}$ ,  $\mathbb{R}$ ,  $\mathbb{Q}$  and  $\emptyset$ , as well as the meanings of the symbols  $\in$  and  $\subseteq$ , and we shall make frequent use of these things. Here is another definition that we use often.

**Definition 4.1** An integer  $n$  is **even** if  $n = 2a$  for some integer  $a \in \mathbb{Z}$ .

Thus, for example, 10 is even because  $10 = 2 \cdot 5$ . Also, according to the definition, 7 is not even because there is no integer  $a$  for which  $7 = 2a$ . While there would be nothing wrong with defining an integer to be odd if it's not even, the following definition is more concrete.

**Definition 4.2** An integer  $n$  is **odd** if  $n = 2a + 1$  for some integer  $a \in \mathbb{Z}$ .

Thus 7 is odd because  $7 = 2 \cdot 3 + 1$ . We will use these definitions whenever the concept of even or odd numbers arises. If in a proof a certain number is even, the definition allows us to write it as  $2a$  for an appropriate integer  $a$ . If some quantity has form  $2b + 1$  where  $b$  is an integer, then the definition tells us the quantity is odd.

**Definition 4.3** Two integers have the **same parity** if they are both even or they are both odd. Otherwise they have **opposite parity**.

Thus 5 and  $-17$  have the same parity, as do 8 and 0; but 3 and 4 have opposite parity.

Two points about definitions are in order. First, in this book the word or term being defined appears in boldface type. Second, it is common to express definitions as conditional statements even though the biconditional would more appropriately convey the meaning. Consider the definition of an even integer. You understand full well that if  $n$  is even then  $n = 2a$  (for  $a \in \mathbb{Z}$ ), and if  $n = 2a$ , then  $n$  is even. Thus, technically the definition should read "*An integer  $n$  is even if and only if  $n = 2a$  for some  $a \in \mathbb{Z}$ .*" However, it is an almost-universal convention that definitions are phrased in the conditional form, even though they are interpreted as being in the biconditional form. There is really no good reason for this, other than economy of words. It is the standard way of writing definitions, and we have to get used to it.

Here is another definition that we will use often.

**Definition 4.4** Suppose  $a$  and  $b$  are integers. We say that  $a$  **divides**  $b$ , written  $a \mid b$ , if  $b = ac$  for some  $c \in \mathbb{Z}$ . In this case we also say that  $a$  is a **divisor** of  $b$ , and that  $b$  is a **multiple** of  $a$ .

For example, 5 divides 15 because  $15 = 5 \cdot 3$ . We write this as  $5 \mid 15$ . Similarly  $8 \mid 32$  because  $32 = 8 \cdot 4$ , and  $-6 \mid 6$  because  $6 = -6 \cdot -1$ . However, 6 does not divide 9 because there is no integer  $c$  for which  $9 = 6 \cdot c$ . We express this as  $6 \nmid 9$ , which we read as “6 *does not divide* 9.”

Be careful of your interpretation of the symbols. There is a big difference between the expressions  $a \mid b$  and  $a/b$ . The expression  $a \mid b$  is a *statement*, while  $a/b$  is a fraction. For example,  $8 \mid 16$  is true and  $8 \mid 20$  is false. By contrast,  $8/16 = 0.5$  and  $8/20 = 0.4$  are numbers, not statements. Be careful not to write one when you mean the other.

Every integer has a set of integers that divide it. For example, the set of divisors of 6 is  $\{a \in \mathbb{Z} : a \mid 6\} = \{-6, -3, -2, -1, 1, 2, 3, 6\}$ . The set of divisors of 5 is  $\{-5, -1, 1, 5\}$ . The set of divisors of 0 is  $\mathbb{Z}$ . This brings us to the following definition, with which you are already familiar.

**Definition 4.5** A number  $n \in \mathbb{N}$  is **prime** if it has exactly two positive divisors, 1 and  $n$ . If  $n$  has more than two positive divisors, it is called **composite**. (Thus  $n$  is composite if and only if  $n = ab$  for  $1 < a, b < n$ .)

For example, 2 is prime, as are 5 and 17. The definition implies that 1 is neither prime nor composite, as it only has one positive divisor, namely 1.

**Definition 4.6** The **greatest common divisor** of integers  $a$  and  $b$ , denoted  $\gcd(a, b)$ , is the largest integer that divides both  $a$  and  $b$ .

The **least common multiple** of non-zero integers  $a$  and  $b$ , denoted  $\text{lcm}(a, b)$ , is the smallest integer in  $\mathbb{N}$  that is a multiple of both  $a$  and  $b$ .

So  $\gcd(18, 24) = 6$ ,  $\gcd(5, 5) = 5$  and  $\gcd(32, -8) = 8$ . Also  $\gcd(50, 18) = 2$ , but  $\gcd(50, 9) = 1$ . Note that  $\gcd(0, 6) = 6$ , because, although every integer divides 0, the largest divisor of 6 is 6.

The expression  $\gcd(0, 0)$  is problematic. Every integer divides 0, so the only conclusion is that  $\gcd(0, 0) = \infty$ . We circumvent this irregularity by simply agreeing to consider  $\gcd(a, b)$  only when  $a$  and  $b$  are not both zero.

Continuing our examples,  $\text{lcm}(4, 6) = 12$ , and  $\text{lcm}(7, 7) = 7$ .

Of course not all terms can be defined. If every word in a definition were defined, there would be separate definitions for the words that appeared

in those definitions, and so on, until the chain of defined terms became circular. Thus we accept some ideas as being so intuitively clear that they require no definitions or verifications. For example, we will not find it necessary to define what an integer (or a real number) is. Nor will we define addition, multiplication, subtraction and division, though we will use these operations freely. We accept and use such things as the distributive and commutative properties of addition and multiplication, as well as other standard properties of arithmetic and algebra.

As mentioned in Section 1.9, we accept as fact the natural ordering of the elements of  $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$  and  $\mathbb{R}$ , so that (for example) statements such as “ $5 < 7$ ,” and “ $x < y$  implies  $-x > -y$ ,” do not need to be justified.

In addition, we accept the following fact without justification or proof.

**Fact 4.1** If  $a$  and  $b$  are integers, then so are their sum, product and difference. That is, if  $a, b \in \mathbb{Z}$ , then  $a + b \in \mathbb{Z}$ ,  $a - b \in \mathbb{Z}$  and  $ab \in \mathbb{Z}$

It follows that any combination of integers using the operations  $+$ ,  $-$  and  $\cdot$  is an integer. For example, if  $a, b$  and  $c$  are integers, then  $a^2b - ca + b \in \mathbb{Z}$ .

We will also accept as obvious the fact that any integer  $a$  can be divided by a positive integer  $b$ , resulting in a unique quotient  $q$  and remainder  $r$ . For example,  $b = 3$  goes into  $a = 17$   $q = 5$  times with remainder  $r = 2$ . In symbols,  $17 = 5 \cdot 3 + 2$ , or  $a = qb + r$ . This fact, called the *division algorithm*, was mentioned on page 30.

**(The Division Algorithm)** Given integers  $a$  and  $b$  with  $b > 0$ , there exist unique integers  $q$  and  $r$  for which  $a = qb + r$  and  $0 \leq r < b$ .

Another fact that we will accept without proof (at least for now) is that every natural number greater than 1 has a unique factorization into primes. For example, the number 1176 can be factored into primes as  $1176 = 2 \cdot 2 \cdot 2 \cdot 3 \cdot 7 \cdot 7 = 2^3 \cdot 3 \cdot 7^2$ . By *unique* we mean that *any* factorization of 1176 into primes will have exactly the same factors (i.e., three 2's, one 3 and two 7's). Thus, for example, there is no valid factorization of 1176 that has a factor of 5. You may be so used to factoring numbers into primes that it seems obvious that there cannot be different prime factorizations of the same number, but in fact this is a fundamental result whose proof is not transparent. Nonetheless, we will be content to assume that every natural number greater than 1 has a unique factorization into primes. (We will revisit the issue of a proof in Section 10.3.)

We will introduce other accepted facts, as well as definitions, as needed.

### 4.3 Direct Proof

This section explains a simple way to prove theorems or propositions that have the form of conditional statements. The technique is called **direct proof**. To simplify the discussion, our first examples will involve proving statements that are almost obviously true. Thus we will call the statements *propositions* rather than theorems. (Remember, a proposition is a statement that, although true, is not as significant as a theorem.)

To understand how the technique of direct proof works, suppose we have some proposition of the following form.

**Proposition** If  $P$ , then  $Q$ .

This proposition is a conditional statement of form  $P \Rightarrow Q$ . Our goal is to show that this conditional statement is true. To see how to proceed, look at the truth table for  $P \Rightarrow Q$ .

$P$	$Q$	$P \Rightarrow Q$
$T$	$T$	$T$
$T$	$F$	$F$
$F$	$T$	$T$
$F$	$F$	$T$

The table shows that if  $P$  is false, the statement  $P \Rightarrow Q$  is automatically true. This means that if we are concerned with showing  $P \Rightarrow Q$  is true, we don't have to worry about the situations where  $P$  is false (as in the last two lines of the table) because the statement  $P \Rightarrow Q$  will be automatically true in those cases. But we must be very careful about the situations where  $P$  is true (as in the first two lines of the table). We must show that the condition of  $P$  being true forces  $Q$  to be true also, for that means the second line of the table cannot happen.

This gives a fundamental outline for proving statements of the form  $P \Rightarrow Q$ . Begin by assuming that  $P$  is true (remember, we don't need to worry about  $P$  being false) and show this forces  $Q$  to be true.

#### Outline for Direct Proof

**Proposition** If  $P$ , then  $Q$ .

*Proof.* Suppose  $P$ .

⋮

Therefore  $Q$ . ■

So the setup for direct proof is remarkably simple. The first line of the proof is the sentence “*Suppose P.*” The last line is the sentence “*Therefore Q.*” Between the first and last line we use logic, definitions and standard math facts to transform the statement  $P$  to the statement  $Q$ . It is common to use the word “*Proof*” to indicate the beginning of a proof, and the symbol ■ to indicate the end.

As our first example, let’s prove that if  $x$  is odd then  $x^2$  is also odd. (Granted, this is not a terribly impressive result, but we will move on to more significant things in due time.) The first step in the proof is to fill in the outline for direct proof. This is a lot like painting a picture, where the basic structure is sketched in first. We leave some space between the first and last line of the proof. The following series of frames indicates the steps you might take to fill in this space with a logical chain of reasoning.

**Proposition** If  $x$  is odd, then  $x^2$  is odd.

*Proof.* Suppose  $x$  is odd.

Therefore  $x^2$  is odd. ■

Now that we have written the first and last lines, we need to fill in the space with a chain of reasoning that shows that  $x$  being odd forces  $x^2$  to be odd.

In doing this it’s always advisable to use any definitions that apply. The first line says  $x$  is odd, and by Definition 4.2 it must be that  $x = 2a + 1$  for some  $a \in \mathbb{Z}$ , so we write this in as our second line.

**Proposition** If  $x$  is odd, then  $x^2$  is odd.

*Proof.* Suppose  $x$  is odd.

*Then  $x = 2a + 1$  for some  $a \in \mathbb{Z}$ , by definition of an odd number.*

Therefore  $x^2$  is odd. ■

Now jump down to the last line, which says  $x^2$  is odd. Think about what the line immediately above it would have to be in order for us to conclude that  $x^2$  is odd. By the definition of an odd number, we would have to have  $x^2 = 2a + 1$  for some  $a \in \mathbb{Z}$ . However, the symbol  $a$  now appears earlier in the proof in a different context, so we should use a different symbol, say  $b$ .

**Proposition** If  $x$  is odd, then  $x^2$  is odd.

*Proof.* Suppose  $x$  is odd.

Then  $x = 2a + 1$  for some  $a \in \mathbb{Z}$ , by definition of an odd number.

Thus  $x^2 = 2b + 1$  for an integer  $b$ .

Therefore  $x^2$  is odd, by definition of an odd number. ■

We are almost there. We can bridge the gap as follows.

**Proposition** If  $x$  is odd, then  $x^2$  is odd.

*Proof.* Suppose  $x$  is odd.

Then  $x = 2a + 1$  for some  $a \in \mathbb{Z}$ , by definition of an odd number.

Thus  $x^2 = (2a + 1)^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + 1$ .

So  $x^2 = 2b + 1$  where  $b$  is the integer  $b = 2a^2 + 2a$ .

Thus  $x^2 = 2b + 1$  for an integer  $b$ .

Therefore  $x^2$  is odd, by definition of an odd number. ■

Finally, we may wish to clean up our work and write the proof in paragraph form. Here is our final version.

**Proposition** If  $x$  is odd, then  $x^2$  is odd.

*Proof.* Suppose  $x$  is odd. Then  $x = 2a + 1$  for some  $a \in \mathbb{Z}$ , by definition of an odd number. Thus  $x^2 = (2a + 1)^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + 1$ , so  $x^2 = 2b + 1$  where  $b = 2a^2 + 2a \in \mathbb{Z}$ . Therefore  $x^2$  is odd, by definition of an odd number. ■

At least initially, it's generally a good idea to write the first and last line of your proof first, and then fill in the gap, sometimes jumping alternately between top and bottom until you meet in the middle, as we did above. This way you are constantly reminded that you are aiming for the statement at the bottom. Sometimes you will leave too much space, sometimes not enough. Sometimes you will get stuck before figuring out what to do. This is normal. Mathematicians do scratch work just as artists do sketches for their paintings.

Here is another example. Consider proving the following proposition.

**Proposition** Let  $a, b$  and  $c$  be integers. If  $a | b$  and  $b | c$ , then  $a | c$ .

Let's apply the basic outline for direct proof. To clarify the procedure we will write the proof in stages again.

**Proposition** Let  $a, b$  and  $c$  be integers. If  $a | b$  and  $b | c$ , then  $a | c$ .

*Proof.* Suppose  $a | b$  and  $b | c$ .

Therefore  $a | c$ . ■

Our first step is to apply Definition 4.4 to the first line. The definition says  $a | b$  means  $b = ac$  for some integer  $c$ , but since  $c$  already appears in a different context on the first line, we must use a different letter, say  $d$ . Similarly let's use a new letter  $e$  in the definition of  $b | c$ .

**Proposition** Let  $a, b$  and  $c$  be integers. If  $a | b$  and  $b | c$ , then  $a | c$ .

*Proof.* Suppose  $a | b$  and  $b | c$ .

*By Definition 4.4, we know  $a | b$  means  $b = ad$  for some  $d \in \mathbb{Z}$ .*

*Likewise,  $b | c$  means  $c = be$  for some  $e \in \mathbb{Z}$ .*

Therefore  $a | c$ . ■

We have almost bridged the gap. The line immediately above the last line should show that  $a | c$ . According to Definition 4.4, this line should say that  $c = ax$  for some integer  $x$ . We can get this equation from the lines at the top, as follows.

**Proposition** Let  $a, b$  and  $c$  be integers. If  $a | b$  and  $b | c$ , then  $a | c$ .

*Proof.* Suppose  $a | b$  and  $b | c$ .

*By Definition 4.4, we know  $a | b$  means  $b = ad$  for some  $d \in \mathbb{Z}$ .*

*Likewise,  $b | c$  means  $c = be$  for some  $e \in \mathbb{Z}$ .*

*Thus  $c = be = (ad)e = a(de)$ , so  $c = ax$  for the integer  $x = de$ .*

Therefore  $a | c$ . ■

The next example is presented all at once rather than in stages.

**Proposition** If  $x$  is an even integer, then  $x^2 - 6x + 5$  is odd.

*Proof.* Suppose  $x$  is an even integer.

Then  $x = 2a$  for some  $a \in \mathbb{Z}$ , by definition of an even integer.

$$\text{So } x^2 - 6x + 5 = (2a)^2 - 6(2a) + 5 = 4a^2 - 12a + 5 = 4a^2 - 12a + 4 + 1 = 2(2a^2 - 6a + 2) + 1.$$

Therefore we have  $x^2 - 6x + 5 = 2b + 1$ , where  $b = 2a^2 - 6a + 2 \in \mathbb{Z}$ .

Consequently  $x^2 - 6x + 5$  is odd, by definition of an odd number. ■

One doesn't normally use a separate line for each sentence in a proof, but for clarity we will often do this in the first few chapters of this book.

Our next example illustrates a standard technique for showing two quantities are equal. If we can show  $m \leq n$  and  $n \leq m$  then it follows that  $m = n$ . In general, the reasoning involved in showing  $m \leq n$  can be quite different from that of showing  $n \leq m$ .

Recall Definition 4.6 of a least common multiple on page 116.

**Proposition** If  $a, b, c \in \mathbb{N}$ , then  $\text{lcm}(ca, cb) = c \cdot \text{lcm}(a, b)$ .

*Proof.* Assume  $a, b, c \in \mathbb{N}$ . Let  $m = \text{lcm}(ca, cb)$  and  $n = c \cdot \text{lcm}(a, b)$ . We will show  $m = n$ . By definition,  $\text{lcm}(a, b)$  is a positive multiple of both  $a$  and  $b$ , so  $\text{lcm}(a, b) = ax = by$  for some  $x, y \in \mathbb{N}$ . From this we see that  $n = c \cdot \text{lcm}(a, b) = cax = cby$  is a positive multiple of both  $ca$  and  $cb$ . But  $m = \text{lcm}(ca, cb)$  is the *smallest* positive multiple of both  $ca$  and  $cb$ . Thus  $m \leq n$ .

On the other hand, as  $m = \text{lcm}(ca, cb)$  is a multiple of both  $ca$  and  $cb$ , we have  $m = cax = cby$  for some  $x, y \in \mathbb{Z}$ . Then  $\frac{1}{c}m = ax = by$  is a multiple of both  $a$  and  $b$ . Therefore  $\text{lcm}(a, b) \leq \frac{1}{c}m$ , so  $c \cdot \text{lcm}(a, b) \leq m$ , that is,  $n \leq m$ .

We've shown  $m \leq n$  and  $n \leq m$ , so  $m = n$ . The proof is complete. ■

The examples we've looked at so far have all been proofs of statements about integers. In our next example, we are going to prove that if  $x$  and  $y$  are positive real numbers for which  $x \leq y$ , then  $\sqrt{x} \leq \sqrt{y}$ . You may feel that the proof is not as "automatic" as the proofs we have done so far. Finding the right steps in a proof can be challenging, and that is part of the fun.

**Proposition** Let  $x$  and  $y$  be positive numbers. If  $x \leq y$ , then  $\sqrt{x} \leq \sqrt{y}$ .

*Proof.* Suppose  $x \leq y$ . Subtracting  $y$  from both sides gives  $x - y \leq 0$ .

This can be written as  $\sqrt{x^2} - \sqrt{y^2} \leq 0$ .

Factor this as a difference of two squares to get  $(\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) \leq 0$ .

Dividing both sides by the positive number  $\sqrt{x} + \sqrt{y}$  produces  $\sqrt{x} - \sqrt{y} \leq 0$ .

Adding  $\sqrt{y}$  to both sides gives  $\sqrt{x} \leq \sqrt{y}$ . ■

This proposition tells us that whenever  $x \leq y$ , we can take the square root of both sides and be assured that  $\sqrt{x} \leq \sqrt{y}$ . This can be useful, as we will see in our next proposition.

That proposition will concern the expression  $2\sqrt{xy} \leq x + y$ . Notice when you substitute random positive values for the variables, the expression is true. For example, for  $x = 6$  and  $y = 4$ , the left side is  $2\sqrt{6 \cdot 4} = 4\sqrt{6} \approx 9.79$ , which is less than the right side  $6 + 4 = 10$ . Is it true that  $2\sqrt{xy} \leq x + y$  for any positive  $x$  and  $y$ ? How could we prove it?

To see how, let's first cast this into the form of a conditional statement: If  $x$  and  $y$  are positive real numbers, then  $2\sqrt{xy} \leq x + y$ . The proof begins with the assumption that  $x$  and  $y$  are positive, and ends with  $2\sqrt{xy} \leq x + y$ . In mapping out a strategy, it can be helpful to work backwards, working from  $2\sqrt{xy} \leq x + y$  to something that is obviously true. Then the steps can be reversed in the proof. In this case, squaring both sides of  $2\sqrt{xy} \leq x + y$  gives us

$$4xy \leq x^2 + 2xy + y^2.$$

Now subtract  $4xy$  from both sides and factor:

$$\begin{aligned} 0 &\leq x^2 - 2xy + y^2 \\ 0 &\leq (x - y)^2. \end{aligned}$$

But this last line is clearly true, since the square of  $x - y$  cannot be negative! This gives us a strategy for the proof, which follows.

**Proposition** If  $x$  and  $y$  are positive real numbers, then  $2\sqrt{xy} \leq x + y$ .

*Proof.* Suppose  $x$  and  $y$  are positive real numbers.

Then  $0 \leq (x - y)^2$ , that is,  $0 \leq x^2 - 2xy + y^2$ .

Adding  $4xy$  to both sides gives  $4xy \leq x^2 + 2xy + y^2$ .

Factoring yields  $4xy \leq (x + y)^2$ .

Previously we proved that such an inequality still holds after taking the square root of both sides; doing so produces  $2\sqrt{xy} \leq x + y$ . ■

Notice that in the last step of the proof we took the square root of both sides of  $4xy \leq (x + y)^2$  and got  $\sqrt{4xy} \leq \sqrt{(x + y)^2}$ , and the fact that this did not reverse the symbol  $\leq$  followed from our previous proposition. This is an important point. Often the proof of a proposition or theorem uses another proposition or theorem (that has already been proved).

## 4.4 Using Cases

In proving a statement is true, we sometimes have to examine multiple cases before showing the statement is true in all possible scenarios. This section illustrates a few examples.

Our examples will concern the expression  $1 + (-1)^n(2n - 1)$ . Here is a table showing its value for various integers for  $n$ . Notice that  $1 + (-1)^n(2n - 1)$  is a multiple of 4 in every line.

$n$	$1 + (-1)^n(2n - 1)$
1	0
2	4
3	-4
4	8
5	-8
6	12
7	-12

Is  $1 + (-1)^n(2n - 1)$  always a multiple of 4? We prove the answer is “yes” in our next example. Notice, however, that the expression  $1 + (-1)^n(2n - 1)$  behaves differently depending on whether  $n$  is even or odd, for in the first case  $(-1)^n = 1$ , and in the second  $(-1)^n = -1$ . Thus the proof must examine these two possibilities separately.

**Proposition** If  $n \in \mathbb{N}$ , then  $1 + (-1)^n(2n - 1)$  is a multiple of 4.

*Proof.* Suppose  $n \in \mathbb{N}$ .

Then  $n$  is either even or odd. Let’s consider these two cases separately.

**Case 1.** Suppose  $n$  is even. Then  $n = 2k$  for some  $k \in \mathbb{Z}$ , and  $(-1)^n = 1$ .

Thus  $1 + (-1)^n(2n - 1) = 1 + (1)(2 \cdot 2k - 1) = 4k$ , which is a multiple of 4.

**Case 2.** Suppose  $n$  is odd. Then  $n = 2k + 1$  for some  $k \in \mathbb{Z}$ , and  $(-1)^n = -1$ .

Thus  $1 + (-1)^n(2n - 1) = 1 - (2(2k + 1) - 1) = -4k$ , which is a multiple of 4.

These cases show that  $1 + (-1)^n(2n - 1)$  is always a multiple of 4. ■

Now let’s examine the flip side of the question. We just proved that  $1 + (-1)^n(2n - 1)$  is always a multiple of 4, but can we get *every* multiple of 4 this way? The following proposition and proof give an affirmative answer.

**Proposition** Every multiple of 4 equals  $1 + (-1)^n(2n - 1)$  for some  $n \in \mathbb{N}$ .

*Proof.* In conditional form, the proposition is as follows:

If  $k$  is a multiple of 4, then there is an  $n \in \mathbb{N}$  for which  $1 + (-1)^n(2n - 1) = k$ .

What follows is a proof of this conditional statement.

Suppose  $k$  is a multiple of 4.

This means  $k = 4a$  for some integer  $a$ .

We must produce an  $n \in \mathbb{N}$  for which  $1 + (-1)^n(2n - 1) = k$ .

This is done by cases, depending on whether  $a$  is zero, positive or negative.

**Case 1.** Suppose  $a = 0$ . Let  $n = 1$ . Then  $1 + (-1)^n(2n - 1) = 1 + (-1)^1(2 - 1) = 0 = 4 \cdot 0 = 4a = k$ .

**Case 2.** Suppose  $a > 0$ . Let  $n = 2a$ , which is in  $\mathbb{N}$  because  $a$  is positive. Also  $n$  is even, so  $(-1)^n = 1$ . Thus  $1 + (-1)^n(2n - 1) = 1 + (2n - 1) = 2n = 2(2a) = 4a = k$ .

**Case 3.** Suppose  $a < 0$ . Let  $n = 1 - 2a$ , which is an element of  $\mathbb{N}$  because  $a$  is negative, making  $1 - 2a$  positive. Also  $n$  is odd, so  $(-1)^n = -1$ . Thus  $1 + (-1)^n(2n - 1) = 1 - (2n - 1) = 1 - (2(1 - 2a) - 1) = 4a = k$ .

The above cases show that no matter whether a multiple  $k = 4a$  of 4 is zero, positive or negative,  $k = 1 + (-1)^n(2n - 1)$  for some  $n \in \mathbb{N}$ . ■

## 4.5 Treating Similar Cases

Occasionally two or more cases in a proof will be so similar that writing them separately seems tedious or unnecessary. Here is an example:

**Proposition** If two integers have opposite parity, then their sum is odd.

*Proof.* Suppose  $m$  and  $n$  are two integers with opposite parity.

We need to show that  $m + n$  is odd. This is done in two cases, as follows.

**Case 1.** Suppose  $m$  is even and  $n$  is odd. Thus  $m = 2a$  and  $n = 2b + 1$  for some integers  $a$  and  $b$ . Therefore  $m + n = 2a + 2b + 1 = 2(a + b) + 1$ , which is odd (by Definition 4.2).

**Case 2.** Suppose  $m$  is odd and  $n$  is even. Thus  $m = 2a + 1$  and  $n = 2b$  for some integers  $a$  and  $b$ . Therefore  $m + n = 2a + 1 + 2b = 2(a + b) + 1$ , which is odd (by Definition 4.2).

In either case,  $m + n$  is odd. ■

The two cases in this proof are entirely alike except for the order in which the even and odd terms occur. It is entirely appropriate to just do one case and indicate that the other case is nearly identical. The phrase “*Without loss of generality...*” is a common way of signaling that the proof is treating just one of several nearly identical cases. Here is a second version of the above example.

**Proposition** If two integers have opposite parity, then their sum is odd.

*Proof.* Suppose  $m$  and  $n$  are two integers with opposite parity.

We need to show that  $m + n$  is odd.

Without loss of generality, suppose  $m$  is even and  $n$  is odd.

Thus  $m = 2a$  and  $n = 2b + 1$  for some integers  $a$  and  $b$ .

Therefore  $m + n = 2a + 2b + 1 = 2(a + b) + 1$ , which is odd (by Definition 4.2). ■

In reading proofs in other texts, you may sometimes see the phrase “Without loss of generality” abbreviated as “WLOG.” However, in the interest of transparency we will avoid writing it this way. In a similar spirit, it is advisable—at least until you become more experienced in proof writing—that you write out all cases, no matter how similar they appear to be.

Please check your understanding by doing the following exercises. The odd numbered problems have complete proofs in the Solutions section in the back of the text.

## Exercises for Chapter 4

Use the method of direct proof to prove the following statements.

1. If  $x$  is an even integer, then  $x^2$  is even.
2. If  $x$  is an odd integer, then  $x^3$  is odd.
3. If  $a$  is an odd integer, then  $a^2 + 3a + 5$  is odd.
4. Suppose  $x, y \in \mathbb{Z}$ . If  $x$  and  $y$  are odd, then  $xy$  is odd.
5. Suppose  $x, y \in \mathbb{Z}$ . If  $x$  is even, then  $xy$  is even.
6. Suppose  $a, b, c \in \mathbb{Z}$ . If  $a \mid b$  and  $a \mid c$ , then  $a \mid (b + c)$ .
7. Suppose  $a, b \in \mathbb{Z}$ . If  $a \mid b$ , then  $a^2 \mid b^2$ .
8. Suppose  $a$  is an integer. If  $5 \mid 2a$ , then  $5 \mid a$ .
9. Suppose  $a$  is an integer. If  $7 \mid 4a$ , then  $7 \mid a$ .
10. Suppose  $a$  and  $b$  are integers. If  $a \mid b$ , then  $a \mid (3b^3 - b^2 + 5b)$ .
11. Suppose  $a, b, c, d \in \mathbb{Z}$ . If  $a \mid b$  and  $c \mid d$ , then  $ac \mid bd$ .
12. If  $x \in \mathbb{R}$  and  $0 < x < 4$ , then  $\frac{4}{x(4-x)} \geq 1$ .
13. Suppose  $x, y \in \mathbb{R}$ . If  $x^2 + 5y = y^2 + 5x$ , then  $x = y$  or  $x + y = 5$ .
14. If  $n \in \mathbb{Z}$ , then  $5n^2 + 3n + 7$  is odd. (Try cases.)
15. If  $n \in \mathbb{Z}$ , then  $n^2 + 3n + 4$  is even. (Try cases.)
16. If two integers have the same parity, then their sum is even. (Try cases.)
17. If two integers have opposite parity, then their product is even.
18. Suppose  $x$  and  $y$  are positive real numbers. If  $x < y$ , then  $x^2 < y^2$ .
19. Suppose  $a, b$  and  $c$  are integers. If  $a^2 \mid b$  and  $b^3 \mid c$ , then  $a^6 \mid c$ .

20. If  $a$  is an integer and  $a^2 \mid a$ , then  $a \in \{-1, 0, 1\}$ .
21. If  $p$  is prime and  $k$  is an integer for which  $0 < k < p$ , then  $p$  divides  $\binom{p}{k}$ .
22. If  $n \in \mathbb{N}$ , then  $n^2 = 2\binom{n}{2} + \binom{n}{1}$ . (You may need a separate case for  $n = 1$ .)
23. If  $n \in \mathbb{N}$ , then  $\binom{2n}{n}$  is even.
24. If  $n \in \mathbb{N}$  and  $n \geq 2$ , then the numbers  $n! + 2, n! + 3, n! + 4, n! + 5, \dots, n! + n$  are all composite. (Thus for any  $n \geq 2$ , one can find  $n$  consecutive composite numbers. This means there are arbitrarily large “gaps” between prime numbers.)
25. If  $a, b, c \in \mathbb{N}$  and  $c \leq b \leq a$ , then  $\binom{a}{b}\binom{b}{c} = \binom{a}{b-c}\binom{a-b+c}{c}$ .
26. Every odd integer is a difference of two squares. (Example  $7 = 4^2 - 3^2$ , etc.)
27. Suppose  $a, b \in \mathbb{N}$ . If  $\gcd(a, b) > 1$ , then  $b \mid a$  or  $b$  is not prime.
28. Let  $a, b, c \in \mathbb{Z}$ . Suppose  $a$  and  $b$  are not both zero, and  $c \neq 0$ . Prove that  $c \cdot \gcd(a, b) \leq \gcd(ca, cb)$ .

# CHAPTER 5

---

## Contrapositive Proof

---

We now examine an alternative to direct proof called **contrapositive proof**. Like direct proof, the technique of contrapositive proof is used to prove conditional statements of the form “*If P, then Q.*” Although it is possible to use direct proof exclusively, there are occasions where contrapositive proof is much easier.

### 5.1 Contrapositive Proof

To understand how contrapositive proof works, imagine that you need to prove a proposition of the following form.

**Proposition** If  $P$ , then  $Q$ .

This is a conditional statement of form  $P \Rightarrow Q$ . Our goal is to show that this conditional statement is true. Recall that in Section 2.6 we observed that  $P \Rightarrow Q$  is logically equivalent to  $\sim Q \Rightarrow \sim P$ . For convenience, we duplicate the truth table that verifies this fact.

$P$	$Q$	$\sim Q$	$\sim P$	$P \Rightarrow Q$	$\sim Q \Rightarrow \sim P$
T	T	F	F	T	T
T	F	T	F	F	F
F	T	F	T	T	T
F	F	T	T	T	T

According to the table, statements  $P \Rightarrow Q$  and  $\sim Q \Rightarrow \sim P$  are different ways of expressing exactly the same thing. The expression  $\sim Q \Rightarrow \sim P$  is called the **contrapositive form** of  $P \Rightarrow Q$ . (Do not confuse the two words *contrapositive* and *converse*. Recall from Section 2.4 that the *converse* of  $P \Rightarrow Q$  is the statement  $Q \Rightarrow P$ , which is *not* logically equivalent to  $P \Rightarrow Q$ .)

Since  $P \Rightarrow Q$  is logically equivalent to  $\sim Q \Rightarrow \sim P$ , it follows that to prove  $P \Rightarrow Q$  is true, it suffices to instead prove that  $\sim Q \Rightarrow \sim P$  is true. If we were to use direct proof to show  $\sim Q \Rightarrow \sim P$  is true, we would assume  $\sim Q$  is true use this to deduce that  $\sim P$  is true. This in fact is the basic approach of contrapositive proof, summarized as follows.

### Outline for Contrapositive Proof

**Proposition** If  $P$ , then  $Q$ .

*Proof.* Suppose  $\sim Q$ .

⋮

Therefore  $\sim P$ . ■

So the setup for contrapositive proof is very simple. The first line of the proof is the sentence “*Suppose  $Q$  is not true.*” (Or something to that effect.) The last line is the sentence “*Therefore  $P$  is not true.*” Between the first and last line we use logic and definitions to transform the statement  $\sim Q$  to the statement  $\sim P$ .

To illustrate this new technique, and to contrast it with direct proof, we now prove a proposition in two ways: first with direct proof and then with contrapositive proof.

**Proposition** Suppose  $x \in \mathbb{Z}$ . If  $7x + 9$  is even, then  $x$  is odd.

*Proof.* (Direct) Suppose  $7x + 9$  is even.

Thus  $7x + 9 = 2a$  for some integer  $a$ .

Subtracting  $6x + 9$  from both sides, we get  $x = 2a - 6x - 9$ .

Thus  $x = 2a - 6x - 9 = 2a - 6x - 10 + 1 = 2(a - 3x - 5) + 1$ .

Consequently  $x = 2b + 1$ , where  $b = a - 3x - 5 \in \mathbb{Z}$ .

Therefore  $x$  is odd. ■

Here is a contrapositive proof of the same statement:

**Proposition** Suppose  $x \in \mathbb{Z}$ . If  $7x + 9$  is even, then  $x$  is odd.

*Proof.* (Contrapositive) Suppose  $x$  is not odd.

Thus  $x$  is even, so  $x = 2a$  for some integer  $a$ .

Then  $7x + 9 = 7(2a) + 9 = 14a + 8 + 1 = 2(7a + 4) + 1$ .

Therefore  $7x + 9 = 2b + 1$ , where  $b$  is the integer  $7a + 4$ .

Consequently  $7x + 9$  is odd.

Therefore  $7x + 9$  is not even. ■

Though the proofs have equal length, you may feel that the contrapositive proof was smoother. This is because it is easier to transform information about  $x$  into information about  $7x+9$  than the other way around. For our next example, consider the following proposition concerning an integer  $x$ :

**Proposition** Suppose  $x \in \mathbb{Z}$ . If  $x^2 - 6x + 5$  is even, then  $x$  is odd.

A direct proof would be problematic. We would begin by assuming that  $x^2 - 6x + 5$  is even, so  $x^2 - 6x + 5 = 2a$ . Then we would need to transform this into  $x = 2b + 1$  for  $b \in \mathbb{Z}$ . But it is not quite clear how that could be done, for it would involve isolating an  $x$  from the quadratic expression. However the proof becomes very simple if we use contrapositive proof.

**Proposition** Suppose  $x \in \mathbb{Z}$ . If  $x^2 - 6x + 5$  is even, then  $x$  is odd.

*Proof.* (Contrapositive) Suppose  $x$  is not odd.

Thus  $x$  is even, so  $x = 2a$  for some integer  $a$ .

So  $x^2 - 6x + 5 = (2a)^2 - 6(2a) + 5 = 4a^2 - 12a + 5 = 4a^2 - 12a + 4 + 1 = 2(2a^2 - 6a + 2) + 1$ .

Therefore  $x^2 - 6x + 5 = 2b + 1$ , where  $b$  is the integer  $2a^2 - 6a + 2$ .

Consequently  $x^2 - 6x + 5$  is odd.

Therefore  $x^2 - 6x + 5$  is not even. ■

In summary, since  $x$  being not odd ( $\sim Q$ ) resulted in  $x^2 - 6x + 5$  being not even ( $\sim P$ ), then  $x^2 - 6x + 5$  being even ( $P$ ) means that  $x$  is odd ( $Q$ ). Thus we have proved  $P \Rightarrow Q$  by proving  $\sim Q \Rightarrow \sim P$ . Here is another example:

**Proposition** Suppose  $x, y \in \mathbb{R}$ . If  $y^3 + yx^2 \leq x^3 + xy^2$ , then  $y \leq x$ .

*Proof.* (Contrapositive) Suppose it is not true that  $y \leq x$ , so  $y > x$ .

Then  $y - x > 0$ . Multiply both sides of  $y - x > 0$  by the positive value  $x^2 + y^2$ .

$$\begin{aligned} (y - x)(x^2 + y^2) &> 0(x^2 + y^2) \\ yx^2 + y^3 - x^3 - xy^2 &> 0 \\ y^3 + yx^2 &> x^3 + xy^2 \end{aligned}$$

Therefore  $y^3 + yx^2 > x^3 + xy^2$ , so it is not true that  $y^3 + yx^2 \leq x^3 + xy^2$ . ■

Proving “*If  $P$ , then  $Q$* ,” with the contrapositive approach necessarily involves the negated statements  $\sim P$  and  $\sim Q$ . In working with these we may have to use the techniques for negating statements (e.g., DeMorgan’s laws) discussed in Section 2.10. We consider such an example next.

**Proposition** Suppose  $x, y \in \mathbb{Z}$ . If  $5 \nmid xy$ , then  $5 \nmid x$  and  $5 \nmid y$ .

*Proof.* (Contrapositive) Suppose it is not true that  $5 \nmid x$  **and**  $5 \nmid y$ .

By DeMorgan's law, it is not true that  $5 \nmid x$  **or** it is not true that  $5 \nmid y$ .

Therefore  $5 \mid x$  or  $5 \mid y$ . We consider these possibilities separately.

**Case 1.** Suppose  $5 \mid x$ . Then  $x = 5a$  for some  $a \in \mathbb{Z}$ .

From this we get  $xy = (5a)y = 5(ay)$ , and that means  $5 \mid xy$ .

**Case 2.** Suppose  $5 \mid y$ . Then  $y = 5a$  for some  $a \in \mathbb{Z}$ .

From this we get  $xy = x(5a) = 5(xa)$ , and that means  $5 \mid xy$ .

The above cases show that  $5 \mid xy$ , so it is not true that  $5 \nmid xy$ . ■

## 5.2 Congruence of Integers

Now is a good time to introduce a new definition that occurs in many branches of mathematics and will surely play a role in some of your later courses. But our main reason for introducing it is that it provides more opportunities to practice writing proofs, both direct and contrapositive.

**Definition 5.1** Given integers  $a$  and  $b$  and  $n \in \mathbb{N}$ , we say that  $a$  and  $b$  are **congruent modulo n** if  $n \mid (a - b)$ . We express this as  $a \equiv b \pmod{n}$ . If  $a$  and  $b$  are not congruent modulo  $n$ , we write this as  $a \not\equiv b \pmod{n}$ .

**Example 5.1** Here are some examples:

1.  $9 \equiv 1 \pmod{4}$  because  $4 \mid (9 - 1)$ .
2.  $6 \equiv 10 \pmod{4}$  because  $4 \mid (6 - 10)$ .
3.  $14 \not\equiv 8 \pmod{4}$  because  $4 \nmid (14 - 8)$ .
4.  $20 \equiv 4 \pmod{8}$  because  $8 \mid (20 - 4)$ .
5.  $17 \equiv -4 \pmod{3}$  because  $3 \mid (17 - (-4))$ .

In practical terms,  $a \equiv b \pmod{n}$  means that  $a$  and  $b$  have the same remainder when divided by  $n$ . For example, we saw above that  $6 \equiv 10 \pmod{4}$  and indeed 6 and 10 both have remainder 2 when divided by 4. Also we saw  $14 \not\equiv 8 \pmod{4}$ , and sure enough 14 has remainder 2 when divided by 4, while 8 has remainder 0.

To see that this is true in general, note that if  $a$  and  $b$  both have the same remainder  $r$  when divided by  $n$ , then it follows that  $a = kn + r$  and  $b = \ell n + r$  for some  $k, \ell \in \mathbb{Z}$ . Then  $a - b = (kn + r) - (\ell n + r) = n(k - \ell)$ . But  $a - b = n(k - \ell)$  means  $n \mid (a - b)$ , so  $a \equiv b \pmod{n}$ . Conversely, this chapter's Exercise 32 asks you to show that if  $a \equiv b \pmod{n}$ , then  $a$  and  $b$  have the same remainder when divided by  $n$ .

We conclude this section with several proofs involving congruence of integers, but you will also test your skills with other proofs in the exercises.

**Proposition** Let  $a, b \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$ , then  $a^2 \equiv b^2 \pmod{n}$ .

*Proof.* We will use direct proof. Suppose  $a \equiv b \pmod{n}$ .

By definition of congruence of integers, this means  $n \mid (a - b)$ .

Then by definition of divisibility, there is an integer  $c$  for which  $a - b = nc$ .

Now multiply both sides of this equation by  $a + b$ .

$$\begin{aligned} a - b &= nc \\ (a - b)(a + b) &= nc(a + b) \\ a^2 - b^2 &= nc(a + b) \end{aligned}$$

Since  $c(a + b) \in \mathbb{Z}$ , the above equation tells us  $n \mid (a^2 - b^2)$ .

According to Definition 5.1, this gives  $a^2 \equiv b^2 \pmod{n}$ . ■

Let's pause to consider this proposition's meaning. It says  $a \equiv b \pmod{n}$  implies  $a^2 \equiv b^2 \pmod{n}$ . In other words, it says that if integers  $a$  and  $b$  have the same remainder when divided by  $n$ , then  $a^2$  and  $b^2$  also have the same remainder when divided by  $n$ . As an example of this, 6 and 10 have the same remainder (2) when divided by  $n = 4$ , and their squares 36 and 100 also have the same remainder (0) when divided by  $n = 4$ . The proposition promises this will happen for all  $a, b$  and  $n$ . In our examples we tend to concentrate more on how to prove propositions than on what the propositions mean. This is reasonable since our main goal is to learn how to prove statements. But it is helpful to sometimes also think about the meaning of what we prove.

**Proposition** Let  $a, b, c \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$ , then  $ac \equiv bc \pmod{n}$ .

*Proof.* We employ direct proof. Suppose  $a \equiv b \pmod{n}$ . By Definition 5.1, it follows that  $n \mid (a - b)$ . Therefore, by definition of divisibility, there exists an integer  $k$  for which  $a - b = nk$ . Multiply both sides of this equation by  $c$  to get  $ac - bc = nkc$ . Thus  $ac - bc = n(kc)$  where  $kc \in \mathbb{Z}$ , which means  $n \mid (ac - bc)$ . By Definition 5.1, we have  $ac \equiv bc \pmod{n}$ . ■

Contrapositive proof seems to be the best approach in the next example, since it will eliminate the symbols  $\mid$  and  $\not\equiv$ .

**Proposition** Suppose  $a, b \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $12a \not\equiv 12b \pmod{n}$ , then  $n \nmid 12$ .

*Proof.* (Contrapositive) Suppose  $n \mid 12$ . Then  $12 = nc$  for some  $c \in \mathbb{Z}$ . Thus

$$12(a - b) = nc(a - b)$$

From this,  $12a - 12b = n(ca - cb)$ . Because  $ca - cb \in \mathbb{Z}$ , we get  $n \mid (12a - 12b)$ . This in turn means  $12a \equiv 12b \pmod{n}$ . ■

### 5.3 Mathematical Writing

Now that we have begun writing proofs, it is a good time to contemplate the craft of writing. Unlike logic and mathematics, where there is a clear-cut distinction between what is right or wrong, the difference between good and bad writing is sometimes a matter of opinion. But there are some standard guidelines that will make your writing clearer. Some are listed below.

#### 1. Begin each sentence with a word, not a mathematical symbol.

The reason is that sentences begin with capital letters, but mathematical symbols are case sensitive. Because  $x$  and  $X$  can have entirely different meanings, putting such symbols at the beginning of a sentence can lead to ambiguity. Here are some examples of bad usage (marked with  $\times$ ) and good usage (marked with  $\checkmark$ ):

$A$  is a subset of  $B$ .  $\times$

The set  $A$  is a subset of  $B$ .  $\checkmark$

$x$  is an integer, so  $2x + 5$  is an integer.  $\times$

Because  $x$  is an integer,  $2x + 5$  is an integer.  $\checkmark$

$x^2 - x + 2 = 0$  has two solutions.  $\times$

$X^2 - x + 2 = 0$  has two solutions.  $\times$  (and silly too)

The equation  $x^2 - x + 2 = 0$  has two solutions.  $\checkmark$

#### 2. End each sentence with a period, even when the sentence ends with a mathematical symbol or expression.

Euler proved that  $\sum_{k=1}^{\infty} \frac{1}{k^s} = \prod_{p \in P} \frac{1}{1 - \frac{1}{p^s}}$ .  $\times$

Euler proved that  $\sum_{k=1}^{\infty} \frac{1}{k^s} = \prod_{p \in P} \frac{1}{1 - \frac{1}{p^s}}$ .  $\checkmark$

Mathematical statements (equations, etc.) are like English phrases that happen to contain special symbols, so use normal punctuation.

**3. Separate mathematical symbols and expressions with words.**

Not doing this can cause confusion by making distinct expressions appear to merge. Compare the clarity of the following examples.

Because  $x^2 - 1 = 0$ ,  $x = 1$  or  $x = -1$ . ✗

Because  $x^2 - 1 = 0$ , it follows that  $x = 1$  or  $x = -1$ . ✓

Unlike  $A \cup B$ ,  $A \cap B$  equals  $\emptyset$ . ✗

Unlike  $A \cup B$ , the set  $A \cap B$  equals  $\emptyset$ . ✓

**4. Avoid misuse of symbols.** Symbols such as  $=$ ,  $\leq$ ,  $\subseteq$ ,  $\in$ , etc., are not words. While it is appropriate to use them in mathematical expressions, they are out of place in other contexts.

Since the two sets are  $=$ , one is a subset of the other. ✗

Since the two sets are equal, one is a subset of the other. ✓

The empty set is  $\subseteq$  of every set. ✗

The empty set is a subset of every set. ✓

Since  $a$  is odd and  $x$  odd  $\Rightarrow x^2$  odd,  $a^2$  is odd. ✗

Since  $a$  is odd and any odd number squared is odd,  $a^2$  is odd. ✓

**5. Avoid using unnecessary symbols.** Mathematics is confusing enough without them. Don't muddy the water even more.

No set  $X$  has negative cardinality. ✗

No set has negative cardinality. ✓

**6. Use the first person plural.** In mathematical writing, it is common to use the words “we” and “us” rather than “I,” “you” or “me.” It is as if the reader and writer are having a conversation, with the writer guiding the reader through the details of the proof.**7. Use the active voice.** This is just a suggestion, but the active voice makes your writing more lively. (And briefer too.)

The value  $x = 3$  is obtained through division of both sides by 5. ✗

Dividing both sides by 5, we get  $x = 3$ . ✓

**8. Explain each new symbol.** In writing a proof, you must explain the meaning of every new symbol you introduce. Failure to do this can lead to ambiguity, misunderstanding and mistakes. For example, consider the following two possibilities for a sentence in a proof, where  $a$  and  $b$  have been introduced on a previous line.

Since  $a | b$ , it follows that  $b = ac$ . ✗

Since  $a | b$ , it follows that  $b = ac$  for some integer  $c$ . ✓

If you use the first form, then the reader may momentarily scan backwards looking for where the  $c$  entered into the picture, not realizing at first that it came from the definition of divides.

- 9. Watch out for “it.”** The pronoun “it” causes confusion when it is unclear what it refers to. If there is any possibility of confusion, you should avoid “it.” Here is an example:

Since  $X \subseteq Y$ , and  $0 < |X|$ , we see that it is not empty. ✗

Is “it”  $X$  or  $Y$ ? Either one would make sense, but which do we mean?

Since  $X \subseteq Y$ , and  $0 < |X|$ , we see that  $Y$  is not empty. ✓

- 10. Since, because, as, for, so.** In proofs, it is common to use these words as conjunctions joining two statements, and meaning that one statement is true and as a consequence the other true. The following statements all mean that  $P$  is true (or assumed to be true) and as a consequence  $Q$  is true also.

$Q$ since $P$	$Q$ because $P$	$Q$ , as $P$	$Q$ , for $P$	$P$ , so $Q$
Since $P, Q$	Because $P, Q$	as $P, Q$		

Notice that the meaning of these constructions is different from that of “*If  $P$ , then  $Q$* ,” for they are asserting not only that  $P$  implies  $Q$ , but **also** that  $P$  is true. Exercise care in using them. It must be the case that  $P$  and  $Q$  are both statements **and** that  $Q$  really does follow from  $P$ .

$x \in \mathbb{N}$ , so  $\mathbb{Z}$  ✗

$x \in \mathbb{N}$ , so  $x \in \mathbb{Z}$  ✓

- 11. Thus, hence, therefore, consequently.** These adverbs precede a statement that follows logically from previous sentences or clauses. Be sure that a statement follows them.

Therefore  $2k + 1$ . ✗

Therefore  $a = 2k + 1$ . ✓

- 12. Clarity is the gold standard of mathematical writing.** If you think breaking a rule makes your writing clearer, then break the rule.

Your mathematical writing will evolve with practice. One of the best ways to develop a good mathematical writing style is to read other people’s proofs. Adopt what works and avoid what doesn’t.

## Exercises for Chapter 5

- A.** Prove the following statements with contrapositive proof. (In each case, think about how a direct proof would work. In most cases contrapositive is easier.)
1. Suppose  $n \in \mathbb{Z}$ . If  $n^2$  is even, then  $n$  is even.
  2. Suppose  $n \in \mathbb{Z}$ . If  $n^2$  is odd, then  $n$  is odd.
  3. Suppose  $a, b \in \mathbb{Z}$ . If  $a^2(b^2 - 2b)$  is odd, then  $a$  and  $b$  are odd.
  4. Suppose  $a, b, c \in \mathbb{Z}$ . If  $a$  does not divide  $bc$ , then  $a$  does not divide  $b$ .
  5. Suppose  $x \in \mathbb{R}$ . If  $x^2 + 5x < 0$  then  $x < 0$ .
  6. Suppose  $x \in \mathbb{R}$ . If  $x^3 - x > 0$  then  $x > -1$ .
  7. Suppose  $a, b \in \mathbb{Z}$ . If both  $ab$  and  $a + b$  are even, then both  $a$  and  $b$  are even.
  8. Suppose  $x \in \mathbb{R}$ . If  $x^5 - 4x^4 + 3x^3 - x^2 + 3x - 4 \geq 0$ , then  $x \geq 0$ .
  9. Suppose  $n \in \mathbb{Z}$ . If  $3 \nmid n^2$ , then  $3 \nmid n$ .
  10. Suppose  $x, y, z \in \mathbb{Z}$  and  $x \neq 0$ . If  $x \nmid yz$ , then  $x \nmid y$  and  $x \nmid z$ .
  11. Suppose  $x, y \in \mathbb{Z}$ . If  $x^2(y+3)$  is even, then  $x$  is even or  $y$  is odd.
  12. Suppose  $a \in \mathbb{Z}$ . If  $a^2$  is not divisible by 4, then  $a$  is odd.
  13. Suppose  $x \in \mathbb{R}$ . If  $x^5 + 7x^3 + 5x \geq x^4 + x^2 + 8$ , then  $x \geq 0$ .
- B.** Prove the following statements using either direct or contrapositive proof.
14. If  $a, b \in \mathbb{Z}$  and  $a$  and  $b$  have the same parity, then  $3a + 7$  and  $7b - 4$  do not.
  15. Suppose  $x \in \mathbb{Z}$ . If  $x^3 - 1$  is even, then  $x$  is odd.
  16. Suppose  $x, y \in \mathbb{Z}$ . If  $x + y$  is even, then  $x$  and  $y$  have the same parity.
  17. If  $n$  is odd, then  $8 \mid (n^2 - 1)$ .
  18. If  $a, b \in \mathbb{Z}$ , then  $(a + b)^3 \equiv a^3 + b^3 \pmod{3}$ .
  19. Let  $a, b, c \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$  and  $a \equiv c \pmod{n}$ , then  $c \equiv b \pmod{n}$ .
  20. If  $a \in \mathbb{Z}$  and  $a \equiv 1 \pmod{5}$ , then  $a^2 \equiv 1 \pmod{5}$ .
  21. Let  $a, b \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$ , then  $a^3 \equiv b^3 \pmod{n}$ .
  22. Let  $a \in \mathbb{Z}$ ,  $n \in \mathbb{N}$ . If  $a$  has remainder  $r$  when divided by  $n$ , then  $a \equiv r \pmod{n}$ .
  23. Let  $a, b \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$ , then  $a^2 \equiv ab \pmod{n}$ .
  24. If  $a \equiv b \pmod{n}$  and  $c \equiv d \pmod{n}$ , then  $ac \equiv bd \pmod{n}$ .
  25. If  $n \in \mathbb{N}$  and  $2^n - 1$  is prime, then  $n$  is prime.
  26. If  $n = 2^k - 1$  for  $k \in \mathbb{N}$ , then every entry in Row  $n$  of Pascal's Triangle is odd.
  27. If  $a \equiv 0 \pmod{4}$  or  $a \equiv 1 \pmod{4}$ , then  $\binom{a}{2}$  is even.
  28. If  $n \in \mathbb{Z}$ , then  $4 \nmid (n^2 - 3)$ .
  29. If integers  $a$  and  $b$  are not both zero, then  $\gcd(a, b) = \gcd(a - b, b)$ .
  30. If  $a \equiv b \pmod{n}$ , then  $\gcd(a, n) = \gcd(b, n)$ .
  31. Suppose the division algorithm applied to  $a$  and  $b$  yields  $a = qb + r$ . Prove  $\gcd(a, b) = \gcd(r, b)$ .
  32. If  $a \equiv b \pmod{n}$ , then  $a$  and  $b$  have the same remainder when divided by  $n$ .

# CHAPTER 6

---

## Proof by Contradiction

---

We now explore a third method of proof: **proof by contradiction**. This method is not limited to proving just conditional statements—it can be used to prove any kind of statement whatsoever. The basic idea is to assume that the statement we want to prove is *false*, and then show that this assumption leads to nonsense. We are then led to conclude that we were wrong to assume the statement was false, so the statement must be true. As an example, consider the following proposition and its proof.

**Proposition** If  $a, b \in \mathbb{Z}$ , then  $a^2 - 4b \neq 2$ .

*Proof.* Suppose this proposition is *false*.

This conditional statement being *false* means there exist numbers  $a$  and  $b$  for which  $a, b \in \mathbb{Z}$  is true, but  $a^2 - 4b \neq 2$  is *false*.

In other words, there exist integers  $a, b \in \mathbb{Z}$  for which  $a^2 - 4b = 2$ .

From this equation we get  $a^2 = 4b + 2 = 2(2b + 1)$ , so  $a^2$  is even.

Because  $a^2$  is even, it follows that  $a$  is even, so  $a = 2c$  for some integer  $c$ .

Now plug  $a = 2c$  back into the boxed equation to get  $(2c)^2 - 4b = 2$ , so  $4c^2 - 4b = 2$ . Dividing by 2, we get  $2c^2 - 2b = 1$ .

Therefore  $1 = 2(c^2 - b)$ , and because  $c^2 - b \in \mathbb{Z}$ , it follows that 1 is even.

We know 1 is **not** even, so something went wrong.

But all the logic after the first line of the proof is correct, so it must be that the first line was incorrect. In other words, we were wrong to assume the proposition was *false*. Thus the proposition is *true*. ■

You may be a bit suspicious of this line of reasoning, but in the next section we will see that it is logically sound. For now, notice that at the end of the proof we deduced that 1 is even, which conflicts with our knowledge that 1 is odd. In essence, we have obtained the statement  $(1 \text{ is odd}) \wedge \sim(1 \text{ is odd})$ , which has the form  $C \wedge \sim C$ . Notice that no matter what statement  $C$  is, and whether or not it is true, the statement  $C \wedge \sim C$  is *false*. A statement—like this one—that cannot be *true* is called a **contradiction**. Contradictions play a key role in our new technique.

## 6.1 Proving Statements with Contradiction

Let's now see why the proof on the previous page is logically valid. In that proof we needed to show that a statement  $P : (a, b \in \mathbb{Z}) \Rightarrow (a^2 - 4b \neq 2)$  was true. The proof began with the assumption that  $P$  was false, that is that  $\sim P$  was true, and from this we deduced  $C \wedge \sim C$ . In other words we proved that  $\sim P$  being true forces  $C \wedge \sim C$  to be true, and this means that we proved that the *conditional statement*  $(\sim P) \Rightarrow (C \wedge \sim C)$  is true. To see that this is the same as proving  $P$  is true, look at the following truth table for  $(\sim P) \Rightarrow (C \wedge \sim C)$ . Notice that the columns for  $P$  and  $(\sim P) \Rightarrow (C \wedge \sim C)$  are exactly the same, so  $P$  is logically equivalent to  $(\sim P) \Rightarrow (C \wedge \sim C)$ .

$P$	$C$	$\sim P$	$C \wedge \sim C$	$(\sim P) \Rightarrow (C \wedge \sim C)$
T	T	F	F	T
T	F	F	F	T
F	T	T	F	F
F	F	T	F	F

Therefore to prove a statement  $P$ , it suffices to instead prove the conditional statement  $(\sim P) \Rightarrow (C \wedge \sim C)$ . This can be done with direct proof: Assume  $\sim P$  and deduce  $C \wedge \sim C$ . Here is the outline:

### Outline for Proof by Contradiction

**Proposition**  $P$ .

*Proof.* Suppose  $\sim P$ .

⋮

Therefore  $C \wedge \sim C$ . ■

A slightly unsettling feature of this method is that we may not know at the beginning of the proof what the statement  $C$  is going to be. In doing the scratch work for the proof, you assume that  $\sim P$  is true, then deduce new statements until you have deduced some statement  $C$  and its negation  $\sim C$ .

If this method seems confusing, look at it this way. In the first line of the proof we suppose  $\sim P$  is true, that is, we assume  $P$  is *false*. But if  $P$  is really true then this contradicts our assumption that  $P$  is false. But we haven't yet *proved*  $P$  to be true, so the contradiction is not obvious. We use logic and reasoning to transform the non-obvious contradiction  $\sim P$  to an obvious contradiction  $C \wedge \sim C$ .

The idea of proof by contradiction is ancient, going back at least to the Pythagoreans, who used it to prove that certain numbers are irrational. Our next example follows their logic to prove that  $\sqrt{2}$  is irrational. Recall that a number is rational if it is a fraction of integers, and it is irrational if it cannot be expressed as a fraction of integers. Here is the exact definition:

**Definition 6.1** A real number  $x$  is **rational** if  $x = \frac{a}{b}$  for some  $a, b \in \mathbb{Z}$ . Also,  $x$  is **irrational** if it is not rational, that is if  $x \neq \frac{a}{b}$  for every  $a, b \in \mathbb{Z}$ .

We are now ready to use contradiction to prove that  $\sqrt{2}$  is irrational. According to the outline, the first line of the proof should be “*Suppose that it is not true that  $\sqrt{2}$  is irrational.*” But it is helpful (though not mandatory) to tip our reader off to the fact that we are using proof by contradiction. One standard way of doing this is to make the first line “*Suppose for the sake of contradiction that it is not true that  $\sqrt{2}$  is irrational.*”

**Proposition** The number  $\sqrt{2}$  is irrational.

*Proof.* Suppose for the sake of contradiction that it is not true that  $\sqrt{2}$  is irrational. Then  $\sqrt{2}$  is rational, so there are integers  $a$  and  $b$  for which

$$\sqrt{2} = \frac{a}{b}. \quad (6.1)$$

Let this fraction be fully reduced; in particular, this means that  $a$  and  $b$  are not both even. (If they were both even, then the fraction could be further reduced by factoring 2's from the numerator and denominator and canceling.) Squaring both sides of Equation 6.1 gives  $2 = \frac{a^2}{b^2}$ , and therefore

$$a^2 = 2b^2. \quad (6.2)$$

From this it follows that  $a^2$  is even. But we proved earlier (Exercise 1 on page 136) that  $a^2$  being even implies  $a$  is even. Thus, as we know that  $a$  and  $b$  are not both even, it follows that  $b$  is odd. Now, since  $a$  is even there is an integer  $c$  for which  $a = 2c$ . Plugging this value for  $a$  into Equation (6.2), we get  $(2c)^2 = 2b^2$ , so  $4c^2 = 2b^2$ , and hence  $b^2 = 2c^2$ . This means  $b^2$  is even, so  $b$  is even also. But previously we deduced that  $b$  is odd. Thus we have the contradiction  $b$  is even and  $b$  is odd. ■

To appreciate the power of proof by contradiction, imagine trying to prove that  $\sqrt{2}$  is irrational without it. Where would we begin? What would be our initial assumption? There are no clear answers to these questions.

Proof by contradiction gives us a starting point: Assume  $\sqrt{2}$  is rational, and work from there.

In the above proof we got the contradiction ( $b$  is even)  $\wedge \sim(b$  is even) which has the form  $C \wedge \sim C$ . In general, your contradiction need not necessarily be of this form. Any statement that is clearly false is sufficient. For example  $2 \neq 2$  would be a fine contradiction, as would be  $4 \mid 2$ , provided that you could deduce them.

Here is another ancient example, dating back at least as far as Euclid:

**Proposition** There are infinitely many prime numbers.

*Proof.* For the sake of contradiction, suppose there are only finitely many prime numbers. Then we can list all the prime numbers as  $p_1, p_2, p_3, \dots, p_n$ , where  $p_1 = 2, p_2 = 3, p_3 = 5, p_4 = 7$  and so on. Thus  $p_n$  is the  $n$ th and largest prime number. Now consider the number  $a = (p_1 p_2 p_3 \cdots p_n) + 1$ , that is,  $a$  is the product of all prime numbers, plus 1. Now  $a$ , like any natural number greater than 1, has at least one prime divisor, and that means  $p_k \mid a$  for at least one of our  $n$  prime numbers  $p_k$ . Thus there is an integer  $c$  for which  $a = cp_k$ , which is to say

$$(p_1 p_2 p_3 \cdots p_{k-1} p_k p_{k+1} \cdots p_n) + 1 = cp_k.$$

Dividing both sides of this by  $p_k$  gives us

$$(p_1 p_2 p_3 \cdots p_{k-1} p_{k+1} \cdots p_n) + \frac{1}{p_k} = c,$$

so

$$\frac{1}{p_k} = c - (p_1 p_2 p_3 \cdots p_{k-1} p_{k+1} \cdots p_n).$$

The expression on the right is an integer, while the expression on the left is not an integer. This is a contradiction. ■

Proof by contradiction often works well in proving statements of the form  $\forall x, P(x)$ . The reason is that the proof set-up involves assuming  $\sim \forall x, P(x)$ , which as we know from Section 2.10 is equivalent to  $\exists x, \sim P(x)$ . This gives us a specific  $x$  for which  $\sim P(x)$  is true, and often that is enough to produce a contradiction. Here is an example:

**Proposition** For every real number  $x \in [0, \pi/2]$ , we have  $\sin x + \cos x \geq 1$ .

*Proof.* Suppose for the sake of contradiction that this is not true.

Then there exists an  $x \in [0, \pi/2]$  for which  $\sin x + \cos x < 1$ .

Since  $x \in [0, \pi/2]$ , neither  $\sin x$  nor  $\cos x$  is negative, so  $0 \leq \sin x + \cos x < 1$ . Thus  $0^2 \leq (\sin x + \cos x)^2 < 1^2$ , which gives  $0^2 \leq \sin^2 x + 2\sin x \cos x + \cos^2 x < 1^2$ . As  $\sin^2 x + \cos^2 x = 1$ , this becomes  $0 \leq 1 + 2\sin x \cos x < 1$ , so  $1 + 2\sin x \cos x < 1$ . Subtracting 1 from both sides gives  $2\sin x \cos x < 0$ . But this contradicts the fact that neither  $\sin x$  nor  $\cos x$  is negative. ■

## 6.2 Proving Conditional Statements by Contradiction

Since the previous two chapters dealt exclusively with proving conditional statements, we now formalize the procedure in which contradiction is used to prove a conditional statement. Suppose we want to prove a proposition of the following form.

**Proposition** If  $P$ , then  $Q$ .

Thus we need to prove that  $P \Rightarrow Q$  is true. Proof by contradiction begins with the assumption that  $\sim(P \Rightarrow Q)$  is true, that is, that  $P \Rightarrow Q$  is false. But we know that  $P \Rightarrow Q$  being false means that it is possible that  $P$  can be true while  $Q$  is false. Thus the first step in the proof is to assume  $P$  and  $\sim Q$ . Here is an outline:

### Outline for Proving a Conditional Statement with Contradiction

**Proposition** If  $P$ , then  $Q$ .

*Proof.* Suppose  $P$  and  $\sim Q$ .

⋮

Therefore  $C \wedge \sim C$ . ■

To illustrate this new technique, we revisit a familiar result: If  $a^2$  is even, then  $a$  is even. According to the outline, the first line of the proof should be “*For the sake of contradiction, suppose  $a^2$  is even and  $a$  is not even.*”

**Proposition** Suppose  $a \in \mathbb{Z}$ . If  $a^2$  is even, then  $a$  is even.

*Proof.* For the sake of contradiction, suppose  $a^2$  is even and  $a$  is not even. Then  $a^2$  is even, and  $a$  is odd.

Since  $a$  is odd, there is an integer  $c$  for which  $a = 2c + 1$ .

Then  $a^2 = (2c + 1)^2 = 4c^2 + 4c + 1 = 2(2c^2 + 2c) + 1$ , so  $a^2$  is odd.

Thus  $a^2$  is even and  $a^2$  is not even, a contradiction. ■

Here is another example.

**Proposition** If  $a, b \in \mathbb{Z}$  and  $a \geq 2$ , then  $a \nmid b$  or  $a \nmid (b+1)$ .

*Proof.* Suppose for the sake of contradiction there exist  $a, b \in \mathbb{Z}$  with  $a \geq 2$ , and for which it is not true that  $a \nmid b$  or  $a \nmid (b+1)$ .

By DeMorgan's law, we have  $a \mid b$  and  $a \mid (b+1)$ .

The definition of divisibility says there are  $c, d \in \mathbb{Z}$  with  $b = ac$  and  $b+1 = ad$ .

Subtracting one equation from the other gives  $ad - ac = 1$ , so  $a(d - c) = 1$ .

Since  $a$  is positive,  $d - c$  is also positive (otherwise  $a(d - c)$  would be negative).

Then  $d - c$  is a positive integer and  $a(d - c) = 1$ , so  $a = 1/(d - c) < 2$ .

Thus we have  $a \geq 2$  and  $a < 2$ , a contradiction. ■

### 6.3 Combining Techniques

Often in more complex proofs several proof techniques are combined within a single proof. For example, in proving a conditional statement  $P \Rightarrow Q$ , we might begin with direct proof and thus assume  $P$  to be true with the aim of ultimately showing  $Q$  is true. But the truth of  $Q$  might hinge on the truth of some other statement  $R$  which—together with  $P$ —would imply  $Q$ . We would then need to prove  $R$ , and we would use whichever proof technique seems most appropriate. This can lead to “proofs inside of proofs.” Consider the following example. The overall approach is direct, but inside the direct proof is a separate proof by contradiction.

**Proposition** Every non-zero rational number can be expressed as a product of two irrational numbers.

*Proof.* This proposition can be reworded as follows: If  $r$  is a non-zero rational number, then  $r$  is a product of two irrational numbers. In what follows, we prove this with direct proof.

Suppose  $r$  is a non-zero rational number. Then  $r = \frac{a}{b}$  for integers  $a$  and  $b$ . Also,  $r$  can be written as a product of two numbers as follows:

$$r = \sqrt{2} \cdot \frac{r}{\sqrt{2}}.$$

We know  $\sqrt{2}$  is irrational, so to complete the proof we must show  $\frac{r}{\sqrt{2}}$  is also irrational.

To show this, assume for the sake of contradiction that  $\frac{r}{\sqrt{2}}$  is rational. This means

$$\frac{r}{\sqrt{2}} = \frac{c}{d}$$

for integers  $c$  and  $d$ , so

$$\sqrt{2} = r \frac{d}{c}.$$

But we know  $r = \frac{a}{b}$ , which combines with the above equation to give

$$\sqrt{2} = r \frac{d}{c} = \frac{a}{b} \frac{d}{c} = \frac{ad}{bc}.$$

This means  $\sqrt{2}$  is rational, which is a contradiction because we know it is irrational. Therefore  $\frac{r}{\sqrt{2}}$  is irrational.

Hence  $r = \sqrt{2} \cdot \frac{r}{\sqrt{2}}$  is a product of two irrational numbers. ■

For another example of a proof-within-a-proof, try Exercise 5 at the end of this chapter (or see its solution). Exercise 5 asks you to prove that  $\sqrt{3}$  is irrational. This turns out to be slightly trickier than proving that  $\sqrt{2}$  is irrational.

## 6.4 Some Words of Advice

Despite the power of proof by contradiction, it's best to use it only when the direct and contrapositive approaches do not seem to work. The reason for this is that a proof by contradiction can often have hidden in it a simpler contrapositive proof, and if this is the case it's better to go with the simpler approach. Consider the following example.

**Proposition** Suppose  $a \in \mathbb{Z}$ . If  $a^2 - 2a + 7$  is even, then  $a$  is odd.

*Proof.* (Contradiction) Suppose  $a^2 - 2a + 7$  is even and  $a$  is not odd.

That is, suppose  $a^2 - 2a + 7$  is even and  $a$  is even.

Since  $a$  is even, there is an integer  $c$  for which  $a = 2c$ .

Then  $a^2 - 2a + 7 = (2c)^2 - 2(2c) + 7 = 2(2c^2 - 2c + 3) + 1$ , so  $a^2 - 2a + 7$  is odd.

Thus  $a^2 - 2a + 7$  is both even and odd, a contradiction. ■

Though there is nothing really wrong with this proof, notice that part of it assumes  $a$  is not odd and deduces that  $a^2 - 2a + 7$  is not even. That is the contrapositive approach! Thus it would be more efficient to proceed as follows, using contrapositive proof.

**Proposition** Suppose  $a \in \mathbb{Z}$ . If  $a^2 - 2a + 7$  is even, then  $a$  is odd.

*Proof.* (Contrapositive) Suppose  $a$  is not odd.

Then  $a$  is even, so there is an integer  $c$  for which  $a = 2c$ .

Then  $a^2 - 2a + 7 = (2c)^2 - 2(2c) + 7 = 2(2c^2 - 2c + 3) + 1$ , so  $a^2 - 2a + 7$  is odd.

Thus  $a^2 - 2a + 7$  is not even. ■

## Exercises for Chapter 6

**A.** Use the method of proof by contradiction to prove the following statements. (In each case, you should also think about how a direct or contrapositive proof would work. You will find in most cases that proof by contradiction is easier.)

1. Suppose  $n \in \mathbb{Z}$ . If  $n$  is odd, then  $n^2$  is odd.
2. Suppose  $n \in \mathbb{Z}$ . If  $n^2$  is odd, then  $n$  is odd.
3. Prove that  $\sqrt[3]{2}$  is irrational.
4. Prove that  $\sqrt{6}$  is irrational.
5. Prove that  $\sqrt{3}$  is irrational.
6. If  $a, b \in \mathbb{Z}$ , then  $a^2 - 4b - 2 \neq 0$ .
7. If  $a, b \in \mathbb{Z}$ , then  $a^2 - 4b - 3 \neq 0$ .
8. Suppose  $a, b, c \in \mathbb{Z}$ . If  $a^2 + b^2 = c^2$ , then  $a$  or  $b$  is even.
9. Suppose  $a, b \in \mathbb{R}$ . If  $a$  is rational and  $ab$  is irrational, then  $b$  is irrational.
10. There exist no integers  $a$  and  $b$  for which  $21a + 30b = 1$ .
11. There exist no integers  $a$  and  $b$  for which  $18a + 6b = 1$ .
12. For every positive  $x \in \mathbb{Q}$ , there is a positive  $y \in \mathbb{Q}$  for which  $y < x$ .
13. For every  $x \in [\pi/2, \pi]$ ,  $\sin x - \cos x \geq 1$ .
14. If  $A$  and  $B$  are sets, then  $A \cap (B - A) = \emptyset$ .
15. If  $b \in \mathbb{Z}$  and  $b \nmid k$  for every  $k \in \mathbb{N}$ , then  $b = 0$ .
16. If  $a$  and  $b$  are positive real numbers, then  $a + b \geq 2\sqrt{ab}$ .
17. For every  $n \in \mathbb{Z}$ ,  $4 \nmid (n^2 + 2)$ .
18. Suppose  $a, b \in \mathbb{Z}$ . If  $4 \mid (a^2 + b^2)$ , then  $a$  and  $b$  are not both odd.

**B.** Prove the following statements using any method from Chapters 4, 5 or 6.

19. The product of any five consecutive integers is divisible by 120. (For example, the product of 3, 4, 5, 6 and 7 is 2520, and  $2520 = 120 \cdot 21$ .)
20. We say that a point  $P = (x, y)$  in  $\mathbb{R}^2$  is **rational** if both  $x$  and  $y$  are rational. More precisely,  $P$  is rational if  $P = (x, y) \in \mathbb{Q}^2$ . An equation  $F(x, y) = 0$  is said to have a **rational point** if there exists  $x_0, y_0 \in \mathbb{Q}$  such that  $F(x_0, y_0) = 0$ . For example, the curve  $x^2 + y^2 - 1 = 0$  has rational point  $(x_0, y_0) = (1, 0)$ . Show that the curve  $x^2 + y^2 - 3 = 0$  has no rational points.
21. Exercise 20 (above) involved showing that there are no rational points on the curve  $x^2 + y^2 - 3 = 0$ . Use this fact to show that  $\sqrt{3}$  is irrational.
22. Explain why  $x^2 + y^2 - 3 = 0$  not having any rational solutions (Exercise 20) implies  $x^2 + y^2 - 3^k = 0$  has no rational solutions for  $k$  an odd, positive integer.
23. Use the above result to prove that  $\sqrt{3^k}$  is irrational for all odd, positive  $k$ .
24. The number  $\log_2 3$  is irrational.

## *Part III*

---

### *More on Proof*

---



# CHAPTER 7

---

## Proving Non-Conditional Statements

---

The last three chapters introduced three major proof techniques: direct, contrapositive and contradiction. These three techniques are used to prove statements of the form “*If P, then Q.*” As we know, most theorems and propositions have this conditional form, or they can be reworded to have this form. Thus the three main techniques are quite important. But some theorems and propositions cannot be put into conditional form. For example, some theorems have form “*P if and only if Q.*” Such theorems are biconditional statements, not conditional statements. In this chapter we examine ways to prove them. In addition to learning how to prove if-and-only-if theorems, we will also look at two other types of theorems.

### 7.1 If-and-Only-If Proof

Some propositions have the form

*P if and only if Q.*

We know from Section 2.4 that this statement asserts that **both** of the following conditional statements are true:

*If P, then Q.*

*If Q, then P.*

So to prove “*P if and only if Q.*” we must prove **two** conditional statements. Recall from Section 2.4 that  $Q \Rightarrow P$  is called the *converse* of  $P \Rightarrow Q$ . Thus we need to prove both  $P \Rightarrow Q$  and its converse. These are both conditional statements, so we may prove them with either direct, contrapositive or contradiction proof. Here is an outline:

#### Outline for If-and-Only-If Proof

**Proposition**  $P$  if and only if  $Q$ .

*Proof.*

[Prove  $P \Rightarrow Q$  using direct, contrapositive or contradiction proof.]

[Prove  $Q \Rightarrow P$  using direct, contrapositive or contradiction proof.] ■

Let's start with a very simple example. You already know that an integer  $n$  is odd if and only if  $n^2$  is odd, but let's prove it anyway, just to illustrate the outline. In this example we prove  $(n \text{ is odd}) \Rightarrow (n^2 \text{ is odd})$  using direct proof and  $(n^2 \text{ is odd}) \Rightarrow (n \text{ is odd})$  using contrapositive proof.

**Proposition** The integer  $n$  is odd if and only if  $n^2$  is odd.

*Proof.* First we show that  $n$  being odd implies that  $n^2$  is odd. Suppose  $n$  is odd. Then, by definition of an odd number,  $n = 2a + 1$  for some integer  $a$ . Thus  $n^2 = (2a + 1)^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + 1$ . This expresses  $n^2$  as twice an integer, plus 1, so  $n^2$  is odd.

Conversely, we need to prove that  $n^2$  being odd implies that  $n$  is odd. We use contrapositive proof. Suppose  $n$  is not odd. Then  $n$  is even, so  $n = 2a$  for some integer  $a$  (by definition of an even number). Thus  $n^2 = (2a)^2 = 2(2a^2)$ , so  $n^2$  is even because it's twice an integer. Thus  $n^2$  is not odd. We've now proved that if  $n$  is not odd, then  $n^2$  is not odd, and this is a contrapositive proof that if  $n^2$  is odd then  $n$  is odd. ■

In proving " $P$  if and only if  $Q$ ," you should begin a new paragraph when starting the proof of  $Q \Rightarrow P$ . Since this is the converse of  $P \Rightarrow Q$ , it's a good idea to begin the paragraph with the word "*Conversely*" (as we did above) to remind the reader that you've finished the first part of the proof and are moving on to the second. Likewise, it's a good idea to remind the reader of exactly what statement that paragraph is proving.

The next example uses direct proof in both parts of the proof.

**Proposition** Suppose  $a$  and  $b$  are integers. Then  $a \equiv b \pmod{6}$  if and only if  $a \equiv b \pmod{2}$  and  $a \equiv b \pmod{3}$ .

*Proof.* First we prove that if  $a \equiv b \pmod{6}$ , then  $a \equiv b \pmod{2}$  and  $a \equiv b \pmod{3}$ . Suppose  $a \equiv b \pmod{6}$ . This means  $6 | (a - b)$ , so there is an integer  $n$  for which

$$a - b = 6n.$$

From this we get  $a - b = 2(3n)$ , which implies  $2 | (a - b)$ , so  $a \equiv b \pmod{2}$ . But we also get  $a - b = 3(2n)$ , which implies  $3 | (a - b)$ , so  $a \equiv b \pmod{3}$ . Therefore  $a \equiv b \pmod{2}$  and  $a \equiv b \pmod{3}$ .

Conversely, suppose  $a \equiv b \pmod{2}$  and  $a \equiv b \pmod{3}$ . Since  $a \equiv b \pmod{2}$  we get  $2 | (a - b)$ , so there is an integer  $k$  for which  $a - b = 2k$ . Therefore  $a - b$  is even. Also, from  $a \equiv b \pmod{3}$  we get  $3 | (a - b)$ , so there is an integer  $\ell$  for which

$$a - b = 3\ell.$$

But since we know  $a - b$  is even, it follows that  $\ell$  must be even also, for if it were odd then  $a - b = 3\ell$  would be odd (because  $a - b$  would be the product of two odd integers). Hence  $\ell = 2m$  for some integer  $m$ . Thus  $a - b = 3\ell = 3 \cdot 2m = 6m$ . This means  $6 | (a - b)$ , so  $a \equiv b \pmod{6}$ . ■

Since if-and-only-if proofs simply combine methods with which we are already familiar, we will not do any further examples in this section. But it is of utmost importance that you practice your skill on some of this chapter's exercises.

## 7.2 Equivalent Statements

In other courses you will sometimes encounter a certain kind of theorem that is neither a conditional nor a biconditional statement. Instead, it asserts that a list of statements is “*equivalent*.” You saw this (or will see it) in your linear algebra textbook, which featured the following theorem:

**Theorem** Suppose  $A$  is an  $n \times n$  matrix. The following statements are equivalent:

- (a) The matrix  $A$  is invertible.
- (b) The equation  $Ax = \mathbf{b}$  has a unique solution for every  $\mathbf{b} \in \mathbb{R}^n$ .
- (c) The equation  $Ax = \mathbf{0}$  has only the trivial solution.
- (d) The reduced row echelon form of  $A$  is  $I_n$ .
- (e)  $\det(A) \neq 0$ .
- (f) The matrix  $A$  does not have 0 as an eigenvalue.

When a theorem asserts that a list of statements is “*equivalent*,” it is asserting that either the statements are all true, or they are all false. Thus the above theorem tells us that whenever we are dealing with a particular  $n \times n$  matrix  $A$ , then either the statements (a) through (f) are all true for  $A$ , or statements (a) through (f) are all false for  $A$ . For example, if we happen to know that  $\det(A) \neq 0$ , the theorem assures us that in addition to statement (e) being true, **all** the statements (a) through (f) are true. On the other hand, if it happens that  $\det(A) = 0$ , the theorem tells us that all statements (a) through (f) are false. In this way, the theorem multiplies our knowledge of  $A$  by a factor of six. Obviously that can be very useful.

What method would we use to prove such a theorem? In a certain sense, the above theorem is like an if-and-only-if theorem. An if-and-only-if theorem of form  $P \Leftrightarrow Q$  asserts that  $P$  and  $Q$  are either both true or both false, that is, that  $P$  and  $Q$  are equivalent. To prove  $P \Leftrightarrow Q$  we prove  $P \Rightarrow Q$  followed by  $Q \Rightarrow P$ , essentially making a “cycle” of implications from  $P$  to  $Q$ .

and back to  $P$ . Similarly, one approach to proving the theorem about the  $n \times n$  matrix would be to prove the conditional statement  $(a) \Rightarrow (b)$ , then  $(b) \Rightarrow (c)$ , then  $(c) \Rightarrow (d)$ , then  $(d) \Rightarrow (e)$ , then  $(e) \Rightarrow (f)$  and finally  $(f) \Rightarrow (a)$ . This pattern is illustrated below.

$$\begin{array}{ccccc} (a) & \Rightarrow & (b) & \Rightarrow & (c) \\ \uparrow & & & & \downarrow \\ (f) & \Leftarrow & (e) & \Leftarrow & (d) \end{array}$$

Notice that if these six implications have been proved, then it really does follow that the statements (a) through (f) are either all true or all false. If one of them is true, then the circular chain of implications forces them all to be true. On the other hand, if one of them (say (c)) is false, the fact that  $(b) \Rightarrow (c)$  is true forces (b) to be false. This combined with the truth of  $(a) \Rightarrow (b)$  makes (a) false, and so on counterclockwise around the circle.

Thus to prove that  $n$  statements are equivalent, it suffices to prove  $n$  conditional statements showing each statement implies another, in circular pattern. But it is not necessary that the pattern be circular. The following schemes would also do the job:

$$\begin{array}{ccccc} (a) & \Rightarrow & (b) & \Leftrightarrow & (c) \\ \uparrow & & \downarrow & & \\ (f) & \Leftarrow & (e) & \Leftrightarrow & (d) \end{array}$$

$$\begin{array}{ccccc} (a) & \Leftrightarrow & (b) & \Leftrightarrow & (c) \\ & & \Downarrow & & \\ (f) & \Leftrightarrow & (e) & \Leftrightarrow & (d) \end{array}$$

But a circular pattern yields the fewest conditional statements that must be proved. Whatever the pattern, each conditional statement can be proved with either direct, contrapositive or contradiction proof.

Though we shall not do any of these proofs in this text, you are sure to encounter them in subsequent courses.

### 7.3 Existence Proofs; Existence and Uniqueness Proofs

Up until this point, we have dealt with proving conditional statements or with statements that can be expressed with two or more conditional statements. Generally, these conditional statements have form  $P(x) \Rightarrow Q(x)$ . (Possibly with more than one variable.) We saw in Section 2.8 that this can be interpreted as a universally quantified statement  $\forall x, P(x) \Rightarrow Q(x)$ .

Thus, conditional statements are universally quantified statements, so in proving a conditional statement—whether we use direct, contrapositive or contradiction proof—we are really proving a universally quantified statement.

But how would we prove an *existentially* quantified statement? What technique would we employ to prove a theorem of the following form?

$$\exists x, R(x)$$

This statement asserts that there exists some specific object  $x$  for which  $R(x)$  is true. To prove  $\exists x, R(x)$  is true, all we would have to do is find and display an *example* of a specific  $x$  that makes  $R(x)$  true.

Though most theorems and propositions are conditional (or if-and-only-if) statements, a few have the form  $\exists x, R(x)$ . Such statements are called **existence statements**, and theorems that have this form are called **existence theorems**. To prove an existence theorem, all you have to do is provide a particular example that shows it is true. This is often quite simple. (But not always!) Here are some examples:

**Proposition** There exists an even prime number.

*Proof.* Observe that 2 is an even prime number. ■

Admittedly, this last proposition was a bit of an oversimplification. The next one is slightly more challenging.

**Proposition** There exists an integer that can be expressed as the sum of two perfect cubes in two different ways.

*Proof.* Consider the number 1729. Note that  $1^3 + 12^3 = 1729$  and  $9^3 + 10^3 = 1729$ . Thus the number 1729 can be expressed as the sum of two perfect cubes in two different ways. ■

Sometimes in the proof of an existence statement, a little verification is needed to show that the example really does work. For example, the above proof would be incomplete if we just asserted that 1729 can be written as a sum of two cubes in two ways without showing *how* this is possible.

**WARNING:** Although an example suffices to prove an existence statement, a single example does not prove a conditional statement.

Often an existence statement will be embedded inside of a conditional statement. Consider the following. (Recall the definition of gcd on page 116.)

If  $a, b \in \mathbb{N}$ , then there exist integers  $k$  and  $\ell$  for which  $\gcd(a, b) = ak + b\ell$ .

This is a conditional statement that has the form

$$a, b \in \mathbb{N} \implies \exists k, \ell \in \mathbb{Z}, \gcd(a, b) = ak + b\ell.$$

To prove it with direct proof, we would first assume that  $a, b \in \mathbb{N}$ , then prove the existence statement  $\exists k, \ell \in \mathbb{Z}, \gcd(a, b) = ak + b\ell$ . That is, we would produce two integers  $k$  and  $\ell$  (which depend on  $a$  and  $b$ ) for which  $\gcd(a, b) = ak + b\ell$ . Let's carry out this plan. (We will use this fundamental proposition several times later, so it is given a number.)

**Proposition 7.1** If  $a, b \in \mathbb{N}$ , then there exist integers  $k$  and  $\ell$  for which  $\gcd(a, b) = ak + b\ell$ .

*Proof.* (Direct) Suppose  $a, b \in \mathbb{N}$ . Consider the set  $A = \{ax + by : x, y \in \mathbb{Z}\}$ . This set contains both positive and negative integers, as well as 0. (Reason: Let  $y = 0$  and let  $x$  range over all integers. Then  $ax + by = ax$  ranges over all multiples of  $a$ , both positive, negative and zero.) Let  $d$  be the smallest *positive* element of  $A$ . Then, because  $d$  is in  $A$ , it must have the form  $d = ak + b\ell$  for some specific  $k, \ell \in \mathbb{Z}$ .

To finish, we will show  $d = \gcd(a, b)$ . We will first argue that  $d$  is a common divisor of  $a$  and  $b$ , and then that it is the *greatest* common divisor.

To see that  $d \mid a$ , use the division algorithm (page 30) to write  $a = qd + r$  for integers  $q$  and  $r$  with  $0 \leq r < d$ . The equation  $a = qd + r$  yields

$$\begin{aligned} r &= a - qd \\ &= a - q(ak + b\ell) \\ &= a(1 - qk) + b(-q\ell). \end{aligned}$$

Therefore  $r$  has form  $r = ax + by$ , so it belongs to  $A$ . But  $0 \leq r < d$  and  $d$  is the smallest positive number in  $A$ , so  $r$  can't be positive; hence  $r = 0$ . Updating our equation  $a = qd + r$ , we get  $a = qd$ , so  $d \mid a$ . Repeating this argument with  $b = qd + r$  shows  $d \mid b$ . Thus  $d$  is indeed a common divisor of  $a$  and  $b$ . It remains to show that it is the *greatest* common divisor.

As  $\gcd(a, b)$  divides  $a$  and  $b$ , we have  $a = \gcd(a, b) \cdot m$  and  $b = \gcd(a, b) \cdot n$  for some  $m, n \in \mathbb{Z}$ . So  $d = ak + b\ell = \gcd(a, b) \cdot mk + \gcd(a, b) \cdot n\ell = \gcd(a, b)(mk + n\ell)$ , and thus  $d$  is a multiple of  $\gcd(a, b)$ . Therefore  $d \geq \gcd(a, b)$ . But  $d$  can't be a larger common divisor of  $a$  and  $b$  than  $\gcd(a, b)$ , so  $d = \gcd(a, b)$ . ■

We conclude this section with a discussion of so-called *uniqueness proofs*. Some existence statements have form “*There is a unique  $x$  for which  $P(x)$ .*” Such a statement asserts that there is *exactly one* example  $x$  for which  $P(x)$  is true. To prove it, you must produce an example  $x = d$  for which  $P(d)$  is true, **and** you must show that  $d$  is the only such example. The next proposition illustrates this. In essence, it asserts that the set  $\{ax + by : x, y \in \mathbb{Z}\}$  consists precisely of all the multiples of  $\gcd(a, b)$ .

**Proposition** Suppose  $a, b \in \mathbb{N}$ . Then there exists a unique  $d \in \mathbb{N}$  for which: An integer  $m$  is a multiple of  $d$  if and only if  $m = ax + by$  for some  $x, y \in \mathbb{Z}$ .

*Proof.* Suppose  $a, b \in \mathbb{N}$ . Let  $d = \gcd(a, b)$ . We first show that an integer  $m$  is a multiple of  $d$  if and only if  $m = ax + by$  for some  $x, y \in \mathbb{Z}$ . Let  $m = dn$  be a multiple of  $d$ . By Proposition 7.1 (on the previous page), there are integers  $k$  and  $\ell$  for which  $d = ak + b\ell$ . Then  $m = dn = (ak + b\ell)n = a(kn) + b(\ell n)$ , so  $m = ax + by$  for integers  $x = kn$  and  $y = \ell n$ .

Conversely, suppose  $m = ax + by$  for some  $x, y \in \mathbb{Z}$ . Since  $d = \gcd(a, b)$  is a divisor of both  $a$  and  $b$ , we have  $a = dc$  and  $b = de$  for some  $c, e \in \mathbb{Z}$ . Then  $m = ax + by = dcx + dey = d(cx + ey)$ , and this is a multiple of  $d$ .

We have now shown that there is a natural number  $d$  with the property that  $m$  is a multiple of  $d$  if and only if  $m = ax + by$  for some  $x, y \in \mathbb{Z}$ . It remains to show that  $d$  is the *unique* such natural number. To do this, suppose  $d'$  is *any* natural number with the property that  $d$  has:

$$m \text{ is a multiple of } d' \iff m = ax + by \text{ for some } x, y \in \mathbb{Z}. \quad (7.1)$$

We next argue that  $d' = d$ ; that is,  $d$  is the *unique* natural number with the stated property. Because of (7.1),  $m = a \cdot 1 + b \cdot 0 = a$  is a multiple of  $d'$ . Likewise  $m = a \cdot 0 + b \cdot 1 = b$  is a multiple of  $d'$ . Hence  $a$  and  $b$  are both multiples of  $d'$ , so  $d'$  is a common divisor of  $a$  and  $b$ , and therefore

$$d' \leq \gcd(a, b) = d.$$

But also, by (7.1), the multiple  $m = d' \cdot 1 = d'$  of  $d'$  can be expressed as  $d' = ax + by$  for some  $x, y \in \mathbb{Z}$ . As noted in the second paragraph of the proof,  $a = dc$  and  $b = de$  for some  $c, e \in \mathbb{Z}$ . Thus  $d' = ax + by = dcx + dey = d(cx + ey)$ , so  $d'$  is a multiple of  $d$ . As  $d'$  and  $d$  are both positive, it follows that

$$d \leq d'.$$

We've now shown that  $d' \leq d$  and  $d \leq d'$ , so  $d = d'$ . The proof is complete. ■

## 7.4 Constructive Versus Non-Constructive Proofs

Existence proofs fall into two categories: constructive and non-constructive. Constructive proofs display an explicit example that proves the theorem; non-constructive proofs prove an example exists without actually giving it. We illustrate the difference with two proofs of the same fact: There exist *irrational* numbers  $x$  and  $y$  (possibly equal) for which  $x^y$  is *rational*.

**Proposition** There exist irrational numbers  $x, y$  for which  $x^y$  is rational.

*Proof.* Let  $x = \sqrt{2}^{\sqrt{2}}$  and  $y = \sqrt{2}$ . We know  $y$  is irrational, but it is not clear whether  $x$  is rational or irrational. On one hand, if  $x$  is irrational, then we have an irrational number to an irrational power that is rational:

$$x^y = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = \sqrt{2}^{\sqrt{2}\sqrt{2}} = \sqrt{2}^2 = 2.$$

On the other hand, if  $x$  is rational, then  $y^x = \sqrt{2}^{\sqrt{2}} = x$  is rational. Either way, we have an irrational number to an irrational power that is rational. ■

The above is a classic example of a **non-constructive** proof. It shows that there exist irrational numbers  $x$  and  $y$  for which  $x^y$  is rational without actually producing (or constructing) an example. It convinces us that one of  $(\sqrt{2}^{\sqrt{2}})^{\sqrt{2}}$  or  $\sqrt{2}^{\sqrt{2}}$  is an irrational number to an irrational power that is rational, but it does not say which one is the correct example. It thus proves that an example exists without explicitly stating one.

Next comes a **constructive proof** of this statement, one that produces (or constructs) two explicit irrational numbers  $x, y$  for which  $x^y$  is rational.

**Proposition** There exist irrational numbers  $x, y$  for which  $x^y$  is rational.

*Proof.* Let  $x = \sqrt{2}$  and  $y = \log_2 9$ . Then

$$x^y = \sqrt{2}^{\log_2 9} = \sqrt{2}^{\log_2 3^2} = \sqrt{2}^{2\log_2 3} = \left(\sqrt{2}^2\right)^{\log_2 3} = 2^{\log_2 3} = 3.$$

As 3 is rational, we have shown that  $x^y = 3$  is rational.

We know that  $x = \sqrt{2}$  is irrational. The proof will be complete if we can show that  $y = \log_2 9$  is irrational. Suppose for the sake of contradiction that  $\log_2 9$  is rational, so there are integers  $a$  and  $b$  for which  $\frac{a}{b} = \log_2 9$ . This means  $2^{a/b} = 9$ , so  $(2^{a/b})^b = 9^b$ , which reduces to  $2^a = 9^b$ . But  $2^a$  is even, while  $9^b$  is odd (because it is the product of the odd number 9 with itself  $b$  times). This is a contradiction; the proof is complete. ■

This existence proof has inside of it a separate proof (by contradiction) that  $\log_2 9$  is irrational. Such combinations of proof techniques are, of course, typical.

Be alert to constructive and non-constructive proofs as you read proofs in other books and articles, as well as to the possibility of crafting such proofs of your own.

---

## Exercises for Chapter 7

Prove the following statements. These exercises are cumulative, covering all techniques addressed in Chapters 4–7.

1. Suppose  $x \in \mathbb{Z}$ . Then  $x$  is even if and only if  $3x + 5$  is odd.
2. Suppose  $x \in \mathbb{Z}$ . Then  $x$  is odd if and only if  $3x + 6$  is odd.
3. Given an integer  $a$ , then  $a^3 + a^2 + a$  is even if and only if  $a$  is even.
4. Given an integer  $a$ , then  $a^2 + 4a + 5$  is odd if and only if  $a$  is even.
5. An integer  $a$  is odd if and only if  $a^3$  is odd.
6. Suppose  $x, y \in \mathbb{R}$ . Then  $x^3 + x^2y = y^2 + xy$  if and only if  $y = x^2$  or  $y = -x$ .
7. Suppose  $x, y \in \mathbb{R}$ . Then  $(x + y)^2 = x^2 + y^2$  if and only if  $x = 0$  or  $y = 0$ .
8. Suppose  $a, b \in \mathbb{Z}$ . Prove that  $a \equiv b \pmod{10}$  if and only if  $a \equiv b \pmod{2}$  and  $a \equiv b \pmod{5}$ .
9. Suppose  $a \in \mathbb{Z}$ . Prove that  $14 \mid a$  if and only if  $7 \mid a$  and  $2 \mid a$ .
10. If  $a \in \mathbb{Z}$ , then  $a^3 \equiv a \pmod{3}$ .
11. Suppose  $a, b \in \mathbb{Z}$ . Prove that  $(a - 3)b^2$  is even if and only if  $a$  is odd or  $b$  is even.
12. There exist a positive real number  $x$  for which  $x^2 < \sqrt{x}$ .
13. Suppose  $a, b \in \mathbb{Z}$ . If  $a + b$  is odd, then  $a^2 + b^2$  is odd.
14. Suppose  $a \in \mathbb{Z}$ . Then  $a^2 \mid a$  if and only if  $a \in \{-1, 0, 1\}$ .
15. Suppose  $a, b \in \mathbb{Z}$ . Prove that  $a + b$  is even if and only if  $a$  and  $b$  have the same parity.
16. Suppose  $a, b \in \mathbb{Z}$ . If  $ab$  is odd, then  $a^2 + b^2$  is even.
17. There is a prime number between 90 and 100.
18. There is a set  $X$  for which  $\mathbb{N} \in X$  and  $\mathbb{N} \subseteq X$ .
19. If  $n \in \mathbb{N}$ , then  $2^0 + 2^1 + 2^2 + 2^3 + 2^4 + \dots + 2^n = 2^{n+1} - 1$ .
20. There exists an  $n \in \mathbb{N}$  for which  $11 \mid (2^n - 1)$ .
21. Every real solution of  $x^3 + x + 3 = 0$  is irrational.
22. If  $n \in \mathbb{Z}$ , then  $4 \mid n^2$  or  $4 \mid (n^2 - 1)$ .
23. Suppose  $a, b$  and  $c$  are integers. If  $a \mid b$  and  $a \mid (b^2 - c)$ , then  $a \mid c$ .
24. If  $a \in \mathbb{Z}$ , then  $4 \nmid (a^2 - 3)$ .

- 25.** If  $p > 1$  is an integer and  $n \nmid p$  for each integer  $n$  for which  $2 \leq n \leq \sqrt{p}$ , then  $p$  is prime.
- 26.** The product of any  $n$  consecutive positive integers is divisible by  $n!$ .
- 27.** Suppose  $a, b \in \mathbb{Z}$ . If  $a^2 + b^2$  is a perfect square, then  $a$  and  $b$  are not both odd.
- 28.** Prove the division algorithm: If  $a, b \in \mathbb{N}$ , there exist *unique* integers  $q, r$  for which  $a = bq + r$ , and  $0 \leq r < b$ . (A proof of existence is given in Section 1.9, but uniqueness needs to be established too.)
- 29.** If  $a \mid bc$  and  $\gcd(a, b) = 1$ , then  $a \mid c$ .  
(Suggestion: Use the proposition on page 152.)
- 30.** Suppose  $a, b, p \in \mathbb{Z}$  and  $p$  is prime. Prove that if  $p \mid ab$  then  $p \mid a$  or  $p \mid b$ . (Suggestion: Use the proposition on page 152.)
- 31.** If  $n \in \mathbb{Z}$ , then  $\gcd(n, n+1) = 1$ .
- 32.** If  $n \in \mathbb{Z}$ , then  $\gcd(n, n+2) \in \{1, 2\}$ .
- 33.** If  $n \in \mathbb{Z}$ , then  $\gcd(2n+1, 4n^2+1) = 1$ .
- 34.** If  $\gcd(a, c) = \gcd(b, c) = 1$ , then  $\gcd(ab, c) = 1$ .  
(Suggestion: Use the proposition on page 152.)
- 35.** Suppose  $a, b \in \mathbb{N}$ . Then  $a = \gcd(a, b)$  if and only if  $a \mid b$ .
- 36.** Suppose  $a, b \in \mathbb{N}$ . Then  $a = \text{lcm}(a, b)$  if and only if  $b \mid a$ .

# CHAPTER 8

---

## Proofs Involving Sets

---

Students in their first advanced mathematics classes are often surprised by the extensive role that sets play and by the fact that most of the proofs they encounter are proofs about sets. Perhaps you've already seen such proofs in your linear algebra course, where a **vector space** was defined to be a *set* of objects (called vectors) that obey certain properties. Your text proved many things about vector spaces, such as the fact that the intersection of two vector spaces is also a vector space, and the proofs used ideas from set theory. As you go deeper into mathematics, you will encounter more and more ideas, theorems and proofs that involve sets. The purpose of this chapter is to give you a foundation that will prepare you for this new outlook.

We will discuss how to show that an object is an element of a set, how to prove one set is a subset of another and how to prove two sets are equal. As you read this chapter you may need to occasionally refer back to Chapter 1 to refresh your memory. For your convenience, the main definitions from Chapter 1 are summarized below. If  $A$  and  $B$  are sets, then

$$\begin{aligned} A \times B &= \{(x, y) : x \in A, y \in B\}, \\ A \cup B &= \{x : (x \in A) \vee (x \in B)\}, \\ A \cap B &= \{x : (x \in A) \wedge (x \in B)\}, \\ A - B &= \{x : (x \in A) \wedge (x \notin B)\}, \\ \overline{A} &= U - A. \end{aligned}$$

Recall that  $A \subseteq B$  means that every element of  $A$  is also an element of  $B$ . Also, the *power set* of  $A$  is the set of all subsets of  $A$ :

$$\mathcal{P}(A) = \{X : X \subseteq A\}.$$

### 8.1 How to Prove $a \in A$

We will begin with a review of set-builder notation, and then review how to show that a given object  $a$  is an element of some set  $A$ .

Generally, a set  $A$  will be expressed in set-builder notation  $A = \{x : P(x)\}$ , where  $P(x)$  is some open sentence about  $x$ . The set  $A$  is understood to have as elements all those things  $x$  for which  $P(x)$  is true. For example,

$$\{x : x \text{ is an odd integer}\} = \{\dots, -5, -3, -1, 1, 3, 5, \dots\}.$$

A common variation of this notation is to express a set as  $A = \{x \in S : P(x)\}$ . Here it is understood that  $A$  consists of all elements  $x$  of the (predetermined) set  $S$  for which  $P(x)$  is true. Keep in mind that, depending on context,  $x$  could be any kind of object (integer, ordered pair, set, function, etc.). There is also nothing special about the particular variable  $x$ ; any reasonable symbol  $x, y, k$ , etc., would do. Some examples follow.

$$\begin{aligned} \{n \in \mathbb{Z} : n \text{ is odd}\} &= \{\dots, -5, -3, -1, 1, 3, 5, \dots\} \\ \{x \in \mathbb{N} : 6|x\} &= \{6, 12, 18, 24, 30, \dots\} \\ \{(a, b) \in \mathbb{Z} \times \mathbb{Z} : b = a + 5\} &= \{\dots, (-2, 3), (-1, 4), (0, 5), (1, 6), \dots\} \\ \{X \in \mathcal{P}(\mathbb{Z}) : |X| = 1\} &= \{\dots, \{-1\}, \{0\}, \{1\}, \{2\}, \{3\}, \{4\}, \dots\} \end{aligned}$$

Now it should be clear how to prove that an object  $a$  belongs to a set  $\{x : P(x)\}$ . Since  $\{x : P(x)\}$  consists of all things  $x$  for which  $P(x)$  is true, to show that  $a \in \{x : P(x)\}$  we just need to show that  $P(a)$  is true. Likewise, to show  $a \in \{x \in S : P(x)\}$ , we need to confirm that  $a \in S$  and that  $P(a)$  is true. These ideas are summarized below. However, you should **not** memorize these methods, you should **understand** them. With contemplation and practice, using them becomes natural and intuitive.

### How to show $a \in \{x : P(x)\}$

Show that  $P(a)$  is true.

### How to show $a \in \{x \in S : P(x)\}$

1. Verify that  $a \in S$ .
2. Show that  $P(a)$  is true.

**Example 8.1** Let's investigate elements of  $A = \{x : x \in \mathbb{N} \text{ and } 7|x\}$ . This set has form  $A = \{x : P(x)\}$  where  $P(x)$  is the open sentence  $(x \in \mathbb{N}) \wedge (7|x)$ . Thus  $21 \in A$  because  $P(21)$  is true. Similarly,  $7, 14, 28, 35$ , etc., are all elements of  $A$ . But  $8 \notin A$  (for example) because  $P(8)$  is false. Likewise  $-14 \notin A$  because  $P(-14)$  is false.

**Example 8.2** Consider the set  $A = \{X \in \mathcal{P}(\mathbb{N}) : |X| = 3\}$ . We know that  $\{4, 13, 45\} \in A$  because  $\{4, 13, 45\} \in \mathcal{P}(\mathbb{N})$  and  $|\{4, 13, 45\}| = 3$ . Also  $\{1, 2, 3\} \in A$ ,  $\{10, 854, 3\} \in A$ , etc. However  $\{1, 2, 3, 4\} \notin A$  because  $|\{1, 2, 3, 4\}| \neq 3$ . Further,  $\{-1, 2, 3\} \notin A$  because  $\{-1, 2, 3\} \notin \mathcal{P}(\mathbb{N})$ .

**Example 8.3** Consider the set  $B = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{5}\}$ . Notice  $(8, 23) \in B$  because  $(8, 23) \in \mathbb{Z} \times \mathbb{Z}$  and  $8 \equiv 23 \pmod{5}$ . Likewise,  $(100, 75) \in B$ ,  $(102, 77) \in B$ , etc., but  $(6, 10) \notin B$ .

Now suppose  $n \in \mathbb{Z}$  and consider the ordered pair  $(4n+3, 9n-2)$ . Does this ordered pair belong to  $B$ ? To answer this, we first observe that  $(4n+3, 9n-2) \in \mathbb{Z} \times \mathbb{Z}$ . Next, we observe that  $(4n+3)-(9n-2) = -5n+5 = 5(1-n)$ , so  $5|(4n+3)-(9n-2)$ , which means  $(4n+3) \equiv (9n-2) \pmod{5}$ . Therefore we have established that  $(4n+3, 9n-2)$  meets the requirements for belonging to  $B$ , so  $(4n+3, 9n-2) \in B$  for every  $n \in \mathbb{Z}$ .

**Example 8.4** This illustrates another common way of defining a set. Consider the set  $C = \{3x^3 + 2 : x \in \mathbb{Z}\}$ . Elements of this set consist of all the values  $3x^3 + 2$  where  $x$  is an integer. Thus  $-22 \in C$  because  $-22 = 3(-2)^3 + 2$ . You can confirm  $-1 \in C$  and  $5 \in C$ , etc. Also  $0 \notin C$  and  $\frac{1}{2} \notin C$ , etc.

## 8.2 How to Prove $A \subseteq B$

In this course (and more importantly, beyond it) you will encounter many circumstances where it is necessary to prove that one set is a subset of another. This section explains how to do this. The methods we discuss should improve your skills in both writing your own proofs and in comprehending the proofs that you read.

Recall (Definition 1.3) that if  $A$  and  $B$  are sets, then  $A \subseteq B$  means that every element of  $A$  is also an element of  $B$ . In other words, it means *if  $a \in A$ , then  $a \in B$* . Therefore to prove that  $A \subseteq B$ , we just need to prove that the conditional statement

*If  $a \in A$ , then  $a \in B$*

is true. This can be proved directly, by assuming  $a \in A$  and deducing  $a \in B$ . The contrapositive approach is another option: Assume  $a \notin B$  and deduce  $a \notin A$ . Each of these two approaches is outlined below.

### How to Prove $A \subseteq B$ (**Direct approach**)

*Proof.* Suppose  $a \in A$ .

⋮

Therefore  $a \in B$ .

Thus  $a \in A$  implies  $a \in B$ ,  
so it follows that  $A \subseteq B$ . ■

### How to Prove $A \subseteq B$ (**Contrapositive approach**)

*Proof.* Suppose  $a \notin B$ .

⋮

Therefore  $a \notin A$ .

Thus  $a \notin B$  implies  $a \notin A$ ,  
so it follows that  $A \subseteq B$ . ■

In practice, the direct approach usually yields the most straightforward and easy proof, though occasionally the contrapositive is the most expedient. (You can even prove  $A \subseteq B$  by contradiction: Assume  $(a \in A) \wedge (a \notin B)$ , and deduce a contradiction.) The remainder of this section consists of examples with occasional commentary. Unless stated otherwise, we will use the direct approach in all proofs; pay special attention to how the above outline for the direct approach is used.

**Example 8.5** Prove that  $\{x \in \mathbb{Z} : 18|x\} \subseteq \{x \in \mathbb{Z} : 6|x\}$ .

*Proof.* Suppose  $a \in \{x \in \mathbb{Z} : 18|x\}$ .

This means that  $a \in \mathbb{Z}$  and  $18|a$ .

By definition of divisibility, there is an integer  $c$  for which  $a = 18c$ .

Consequently  $a = 6(3c)$ , and from this we deduce that  $6|a$ .

Therefore  $a$  is one of the integers that 6 divides, so  $a \in \{x \in \mathbb{Z} : 6|x\}$ .

We've shown  $a \in \{x \in \mathbb{Z} : 18|x\}$  implies  $a \in \{x \in \mathbb{Z} : 6|x\}$ , so it follows that  $\{x \in \mathbb{Z} : 18|x\} \subseteq \{x \in \mathbb{Z} : 6|x\}$ . ■

**Example 8.6** Prove that  $\{x \in \mathbb{Z} : 2|x\} \cap \{x \in \mathbb{Z} : 9|x\} \subseteq \{x \in \mathbb{Z} : 6|x\}$ .

*Proof.* Suppose  $a \in \{x \in \mathbb{Z} : 2|x\} \cap \{x \in \mathbb{Z} : 9|x\}$ .

By definition of intersection, this means  $a \in \{x \in \mathbb{Z} : 2|x\}$  and  $a \in \{x \in \mathbb{Z} : 9|x\}$ .

Since  $a \in \{x \in \mathbb{Z} : 2|x\}$  we know  $2|a$ , so  $a = 2c$  for some  $c \in \mathbb{Z}$ . Thus  $a$  is even.

Since  $a \in \{x \in \mathbb{Z} : 9|x\}$  we know  $9|a$ , so  $a = 9d$  for some  $d \in \mathbb{Z}$ .

As  $a$  is even,  $a = 9d$  implies  $d$  is even. (Otherwise  $a = 9d$  would be odd.)

Then  $d = 2e$  for some integer  $e$ , and we have  $a = 9d = 9(2e) = 6(3e)$ .

From  $a = 6(3e)$ , we conclude  $6|a$ , and this means  $a \in \{x \in \mathbb{Z} : 6|x\}$ .

We have shown that  $a \in \{x \in \mathbb{Z} : 2|x\} \cap \{x \in \mathbb{Z} : 9|x\}$  implies  $a \in \{x \in \mathbb{Z} : 6|x\}$ , so it follows that  $\{x \in \mathbb{Z} : 2|x\} \cap \{x \in \mathbb{Z} : 9|x\} \subseteq \{x \in \mathbb{Z} : 6|x\}$ . ■

**Example 8.7** Show  $\{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{6}\} \subseteq \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{3}\}$ .

*Proof.* Suppose  $(a, b) \in \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{6}\}$ .

This means  $(a, b) \in \mathbb{Z} \times \mathbb{Z}$  and  $a \equiv b \pmod{6}$ .

Consequently  $6|(a - b)$ , so  $a - b = 6c$  for some integer  $c$ .

It follows that  $a - b = 3(2c)$ , and this means  $3|(a - b)$ , so  $a \equiv b \pmod{3}$ .

Thus  $(a, b) \in \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{3}\}$ .

We've now seen that  $(a, b) \in \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{6}\}$  implies  $(a, b) \in \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{3}\}$ , so it follows that  $\{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{6}\} \subseteq \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x \equiv y \pmod{3}\}$ . ■

Some statements involving subsets are transparent enough that we often accept (and use) them without proof. For example, if  $A$  and  $B$  are any sets, then it's very easy to confirm  $A \cap B \subseteq A$ . (Reason: Suppose  $x \in A \cap B$ . Then  $x \in A$  and  $x \in B$  by definition of intersection, so in particular  $x \in A$ . Thus  $x \in A \cap B$  implies  $x \in A$ , so  $A \cap B \subseteq A$ .) Other statements of this nature include  $A \subseteq A \cup B$  and  $A - B \subseteq A$ , as well as conditional statements such as  $((A \subseteq B) \wedge (B \subseteq C)) \Rightarrow (A \subseteq C)$  and  $(X \subseteq A) \Rightarrow (X \subseteq A \cup B)$ . Our point of view in this text is that we do not need to prove such obvious statements unless we are explicitly asked to do so in an exercise. (Still, you should do some quick mental proofs to convince yourself that the above statements are true. If you don't see that  $A \cap B \subseteq A$  is true but that  $A \subseteq A \cap B$  is not necessarily true, then you need to spend more time on this topic.)

The next example will show that if  $A$  and  $B$  are sets, then  $\mathcal{P}(A) \cup \mathcal{P}(B) \subseteq \mathcal{P}(A \cup B)$ . Before beginning our proof, let's look at an example to see if this statement really makes sense. Suppose  $A = \{1, 2\}$  and  $B = \{2, 3\}$ . Then

$$\begin{aligned}\mathcal{P}(A) \cup \mathcal{P}(B) &= \{\emptyset, \{1\}, \{2\}, \{1, 2\}\} \cup \{\emptyset, \{2\}, \{3\}, \{2, 3\}\} \\ &= \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}\}.\end{aligned}$$

Also  $\mathcal{P}(A \cup B) = \mathcal{P}(\{1, 2, 3\}) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}$ . Thus, even though  $\mathcal{P}(A) \cup \mathcal{P}(B) \neq \mathcal{P}(A \cup B)$ , it is true that  $\mathcal{P}(A) \cup \mathcal{P}(B) \subseteq \mathcal{P}(A \cup B)$  for this particular  $A$  and  $B$ . Now let's prove  $\mathcal{P}(A) \cup \mathcal{P}(B) \subseteq \mathcal{P}(A \cup B)$  no matter what sets  $A$  and  $B$  are.

**Example 8.8** Prove that if  $A$  and  $B$  are sets, then  $\mathcal{P}(A) \cup \mathcal{P}(B) \subseteq \mathcal{P}(A \cup B)$ .

*Proof.* Suppose  $X \in \mathcal{P}(A) \cup \mathcal{P}(B)$ .

By definition of union, this means  $X \in \mathcal{P}(A)$  or  $X \in \mathcal{P}(B)$ .

Therefore  $X \subseteq A$  or  $X \subseteq B$  (by definition of power sets). We consider cases.

**Case 1.** Suppose  $X \subseteq A$ . Then  $X \subseteq A \cup B$ , and this means  $X \in \mathcal{P}(A \cup B)$ .

**Case 2.** Suppose  $X \subseteq B$ . Then  $X \subseteq A \cup B$ , and this means  $X \in \mathcal{P}(A \cup B)$ .

(We do not need to consider the case where  $X \subseteq A$  and  $X \subseteq B$  because that is taken care of by either of cases 1 or 2.) The above cases show that  $X \in \mathcal{P}(A \cup B)$ .

Thus we've shown that  $X \in \mathcal{P}(A) \cup \mathcal{P}(B)$  implies  $X \in \mathcal{P}(A \cup B)$ , and this completes the proof that  $\mathcal{P}(A) \cup \mathcal{P}(B) \subseteq \mathcal{P}(A \cup B)$ . ■

In our next example, we prove a conditional statement. Direct proof is used, and in the process we use our new technique for showing  $A \subseteq B$ .

**Example 8.9** Suppose  $A$  and  $B$  are sets. If  $\mathcal{P}(A) \subseteq \mathcal{P}(B)$ , then  $A \subseteq B$ .

*Proof.* We use direct proof. Assume  $\mathcal{P}(A) \subseteq \mathcal{P}(B)$ .

Based on this assumption, we must now show that  $A \subseteq B$ .

To show  $A \subseteq B$ , suppose that  $a \in A$ .

Then the one-element set  $\{a\}$  is a subset of  $A$ , so  $\{a\} \in \mathcal{P}(A)$ .

But then, since  $\mathcal{P}(A) \subseteq \mathcal{P}(B)$ , it follows that  $\{a\} \in \mathcal{P}(B)$ .

This means that  $\{a\} \subseteq B$ , hence  $a \in B$ .

We've shown that  $a \in A$  implies  $a \in B$ , so therefore  $A \subseteq B$ . ■

### 8.3 How to Prove $A = B$

In proofs it is often necessary to show that two sets are equal. There is a standard way of doing this. Suppose we want to show  $A = B$ . If we show  $A \subseteq B$ , then every element of  $A$  is also in  $B$ , but there is still a possibility that  $B$  could have some elements that are not in  $A$ , so we can't conclude  $A = B$ . But if *in addition* we also show  $B \subseteq A$ , then  $B$  can't contain anything that is not in  $A$ , so  $A = B$ . This is the standard procedure for proving  $A = B$ : Prove both  $A \subseteq B$  and  $B \subseteq A$ .

#### How to Prove $A = B$

*Proof.*

[Prove that  $A \subseteq B$ .]

[Prove that  $B \subseteq A$ .]

Therefore, since  $A \subseteq B$  and  $B \subseteq A$ ,  
it follows that  $A = B$ . ■

**Example 8.10** Prove that  $\{n \in \mathbb{Z} : 35|n\} = \{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\}$ .

*Proof.* First we show  $\{n \in \mathbb{Z} : 35|n\} \subseteq \{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\}$ . Suppose  $a \in \{n \in \mathbb{Z} : 35|n\}$ . This means  $35|a$ , so  $a = 35c$  for some  $c \in \mathbb{Z}$ . Thus  $a = 5(7c)$  and  $a = 7(5c)$ . From  $a = 5(7c)$  it follows that  $5|a$ , so  $a \in \{n \in \mathbb{Z} : 5|n\}$ . From  $a = 7(5c)$  it follows that  $7|a$ , which means  $a \in \{n \in \mathbb{Z} : 7|n\}$ . As  $a$  belongs to both  $\{n \in \mathbb{Z} : 5|n\}$  and  $\{n \in \mathbb{Z} : 7|n\}$ , we get  $a \in \{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\}$ . Thus we've shown that  $\{n \in \mathbb{Z} : 35|n\} \subseteq \{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\}$ .

Next we show  $\{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\} \subseteq \{n \in \mathbb{Z} : 35|n\}$ . Suppose that  $a \in \{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\}$ . By definition of intersection, this means that  $a \in \{n \in \mathbb{Z} : 5|n\}$  and  $a \in \{n \in \mathbb{Z} : 7|n\}$ . Therefore it follows that  $5|a$  and  $7|a$ . By definition of divisibility, there are integers  $c$  and  $d$  with  $a = 5c$  and  $a = 7d$ . Then  $a$  has both 5 and 7 as prime factors, so the prime factorization of  $a$

must include factors of 5 and 7. Hence  $5 \cdot 7 = 35$  divides  $a$ , so  $a \in \{n \in \mathbb{Z} : 35|n\}$ . We've now shown that  $\{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\} \subseteq \{n \in \mathbb{Z} : 35|n\}$ .

At this point we've shown that  $\{n \in \mathbb{Z} : 35|n\} \subseteq \{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\}$  and  $\{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\} \subseteq \{n \in \mathbb{Z} : 35|n\}$ , so we've proved  $\{n \in \mathbb{Z} : 35|n\} = \{n \in \mathbb{Z} : 5|n\} \cap \{n \in \mathbb{Z} : 7|n\}$ . ■

You know from algebra that if  $c \neq 0$  and  $ac = bc$ , then  $a = b$ . The next example shows that an analogous statement holds for sets  $A, B$  and  $C$ . The example asks us to prove a conditional statement. We will prove it with direct proof. In carrying out the process of direct proof, we will have to use the new techniques from this section.

**Example 8.11** Suppose  $A, B$ , and  $C$  are sets, and  $C \neq \emptyset$ . Prove that if  $A \times C = B \times C$ , then  $A = B$ .

*Proof.* Suppose  $A \times C = B \times C$ . We must now show  $A = B$ .

First we will show  $A \subseteq B$ . Suppose  $a \in A$ . Since  $C \neq \emptyset$ , there exists an element  $c \in C$ . Thus, since  $a \in A$  and  $c \in C$ , we have  $(a, c) \in A \times C$ , by definition of the Cartesian product. But then, since  $A \times C = B \times C$ , it follows that  $(a, c) \in B \times C$ . But  $(a, c) \in B \times C$  means  $a \in B$ , by definition of the Cartesian product. We have shown  $a \in A$  implies  $a \in B$ , so  $A \subseteq B$ .

Next we show  $B \subseteq A$ . We use the same argument as above, with the roles of  $A$  and  $B$  reversed. Suppose  $a \in B$ . Since  $C \neq \emptyset$ , there exists an element  $c \in C$ . Thus, since  $a \in B$  and  $c \in C$ , we have  $(a, c) \in B \times C$ . But then, since  $B \times C = A \times C$ , we have  $(a, c) \in A \times C$ . It follows that  $a \in A$ . We have shown  $a \in B$  implies  $a \in A$ , so  $B \subseteq A$ .

The previous two paragraphs have shown  $A \subseteq B$  and  $B \subseteq A$ , so  $A = B$ . In summary, we have shown that if  $A \times C = B \times C$ , then  $A = B$ . This completes the proof. ■

Now we'll look at another way that set operations are similar to operations on numbers. From algebra you are familiar with the distributive property  $a \cdot (b + c) = a \cdot b + a \cdot c$ . Replace the numbers  $a, b, c$  with sets  $A, B, C$ , and replace  $\cdot$  with  $\times$  and  $+$  with  $\cap$ . We get  $A \times (B \cap C) = (A \times B) \cap (A \times C)$ . This statement turns out to be true, as we now prove.

**Example 8.12** Given sets  $A, B$  and  $C$ , prove  $A \times (B \cap C) = (A \times B) \cap (A \times C)$ .

*Proof.* First we will show that  $A \times (B \cap C) \subseteq (A \times B) \cap (A \times C)$ .

Suppose  $(a, b) \in A \times (B \cap C)$ .

By definition of the Cartesian product, this means  $a \in A$  and  $b \in B \cap C$ .

By definition of intersection, it follows that  $b \in B$  and  $b \in C$ .

Thus, since  $a \in A$  and  $b \in B$ , it follows that  $(a, b) \in A \times B$  (by definition of  $\times$ ). Also, since  $a \in A$  and  $b \in C$ , it follows that  $(a, b) \in A \times C$  (by definition of  $\times$ ). Now we have  $(a, b) \in A \times B$  and  $(a, b) \in A \times C$ , so  $(a, b) \in (A \times B) \cap (A \times C)$ . We've shown that  $(a, b) \in A \times (B \cap C)$  implies  $(a, b) \in (A \times B) \cap (A \times C)$  so we have  $A \times (B \cap C) \subseteq (A \times B) \cap (A \times C)$ .

Next we will show that  $(A \times B) \cap (A \times C) \subseteq A \times (B \cap C)$ . Suppose  $(a, b) \in (A \times B) \cap (A \times C)$ . By definition of intersection, this means  $(a, b) \in A \times B$  and  $(a, b) \in A \times C$ . By definition of the Cartesian product,  $(a, b) \in A \times B$  means  $a \in A$  and  $b \in B$ . By definition of the Cartesian product,  $(a, b) \in A \times C$  means  $a \in A$  and  $b \in C$ . We now have  $b \in B$  and  $b \in C$ , so  $b \in B \cap C$ , by definition of intersection. Thus we've deduced that  $a \in A$  and  $b \in B \cap C$ , so  $(a, b) \in A \times (B \cap C)$ . In summary, we've shown that  $(a, b) \in (A \times B) \cap (A \times C)$  implies  $(a, b) \in A \times (B \cap C)$  so we have  $(A \times B) \cap (A \times C) \subseteq A \times (B \cap C)$ .

The previous two paragraphs show that  $A \times (B \cap C) \subseteq (A \times B) \cap (A \times C)$  and  $(A \times B) \cap (A \times C) \subseteq A \times (B \cap C)$ , so it follows that  $(A \times B) \cap (A \times C) = A \times (B \cap C)$ . ■

Occasionally you can prove two sets are equal by working out a series of equalities leading from one set to the other. This like showing two algebraic expressions are equal by manipulating one until you obtain the other. We illustrate this in the following example, which gives an alternate solution to the previous example. This approach is sometimes not applicable (or awkward), but when it works it can shorten a proof dramatically.

A quick note before beginning the example. Notice that any statement  $P$  is logically equivalent to  $P \wedge P$ . (Write out a truth table if you are in doubt.) At one point in the following example we will replace the expression  $x \in A$  with the logically equivalent statement  $(x \in A) \wedge (x \in A)$ .

**Example 8.13** Given sets  $A$ ,  $B$ , and  $C$ , prove  $A \times (B \cap C) = (A \times B) \cap (A \times C)$ .

*Proof.* Just observe the following sequence of equalities.

$$\begin{aligned}
 A \times (B \cap C) &= \{(x, y) : (x \in A) \wedge (y \in B \cap C)\} && (\text{def. of } \times) \\
 &= \{(x, y) : (x \in A) \wedge (y \in B) \wedge (y \in C)\} && (\text{def. of } \cap) \\
 &= \{(x, y) : (x \in A) \wedge (x \in A) \wedge (y \in B) \wedge (y \in C)\} && (P = P \wedge P) \\
 &= \{(x, y) : ((x \in A) \wedge (y \in B)) \wedge ((x \in A) \wedge (y \in C))\} && (\text{rearrange}) \\
 &= \{(x, y) : (x \in A) \wedge (y \in B)\} \cap \{(x, y) : (x \in A) \wedge (y \in C)\} && (\text{def. of } \cap) \\
 &= (A \times B) \cap (A \times C) && (\text{def. of } \times)
 \end{aligned}$$

This completes the proof. ■

The equation  $A \times (B \cap C) = (A \times B) \cap (A \times C)$  just obtained is a fundamental law that you may actually use fairly often as you continue with mathematics. Some similar equations are listed below. Each of these can be proved with this section's techniques, and the exercises will ask that you do so.

$$\left. \begin{array}{l} \overline{A \cap B} = \overline{A} \cup \overline{B} \\ A \cup B = \overline{\overline{A} \cap \overline{B}} \\ A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \\ A \times (B \cup C) = (A \times B) \cup (A \times C) \\ A \times (B \cap C) = (A \times B) \cap (A \times C) \end{array} \right\}$$

DeMorgan's laws for sets  
Distributive laws for sets

It is very good practice to prove these equations. Depending on your learning style, it is probably not necessary to commit them to memory. But don't forget them entirely. They may be useful later in your mathematical work. If so, you can look them up or re-derive them on the spot. If you go on to study mathematics deeply, you will at some point realize that you've internalized them without even being cognizant of it.

## 8.4 Examples: Perfect Numbers

Sometimes it takes a good bit of work and creativity to show that one set is a subset of another or that they are equal. We illustrate this now with examples from number theory involving what are called perfect numbers. Even though this topic is quite old, dating back more than 2000 years, it leads to some questions that are unanswered even today.

The problem involves adding up the positive divisors of a natural number. To begin the discussion, consider the number 12. If we add up the positive divisors of 12 that are less than 12, we obtain  $1 + 2 + 3 + 4 + 6 = 16$ , which is greater than 12. Doing the same thing for 15, we get  $1 + 3 + 5 = 9$ , which is less than 15. For the most part, given a natural number  $p$ , the sum of its positive divisors less than itself will either be greater than  $p$  or less than  $p$ . But occasionally the divisors add up to exactly  $p$ . If this happens, then  $p$  is said to be a *perfect number*.

**Definition 8.1** A number  $p \in \mathbb{N}$  is **perfect** if it equals the sum of its positive divisors less than itself. Some examples follow.

- The number 6 is perfect since  $6 = 1 + 2 + 3$ .
- The number 28 is perfect since  $28 = 1 + 2 + 4 + 7 + 14$ .
- The number 496 is perfect since  $496 = 1 + 2 + 4 + 8 + 16 + 31 + 62 + 124 + 248$ .

Though it would take a while to find it by trial-and-error, the next perfect number after 496 is 8128. You can check that 8128 is perfect. Its divisors are 1, 2, 4, 8, 16, 32, 64, 127, 254, 508, 1016, 2032, 4064 and indeed

$$8128 = 1 + 2 + 4 + 8 + 16 + 32 + 64 + 127 + 254 + 508 + 1016 + 2032 + 4064.$$

Are there other perfect numbers? How can they be found? Do they obey any patterns? These questions fascinated the ancient Greek mathematicians. In what follows we will develop an idea—recorded by Euclid—that partially answers these questions. Euclid lived millennia before set theory was even invented, so he certainly did not use sets. Nonetheless we will phrase his idea in the language of sets.

Since our goal is to understand what numbers are perfect, let's define the following set:

$$P = \{p \in \mathbb{N} : p \text{ is perfect}\}.$$

Therefore  $P = \{6, 28, 496, 8128, \dots\}$ , but it is unclear what numbers are in  $P$  other than the ones listed. Our goal is to gain a better understanding of just which numbers the set  $P$  includes. To do this, we will examine the following set  $A$ . It looks more complicated than  $P$ , but it will be very helpful for understanding  $P$ , as we will soon see.

$$A = \{2^{n-1}(2^n - 1) : n \in \mathbb{N}, \text{ and } 2^n - 1 \text{ is prime}\}$$

In words,  $A$  consists of every natural number of form  $2^{n-1}(2^n - 1)$ , where  $2^n - 1$  is prime. To get a feel for what numbers belong to  $A$ , look at the following table. For each natural number  $n$ , it tallies the corresponding numbers  $2^{n-1}$  and  $2^n - 1$ . If  $2^n - 1$  happens to be prime, then the product  $2^{n-1}(2^n - 1)$  is given; otherwise that entry is labeled with an  $*$ .

$n$	$2^{n-1}$	$2^n - 1$	$2^{n-1}(2^n - 1)$
1	1	1	*
2	2	3	6
3	4	7	28
4	8	15	*
5	16	31	496
6	32	63	*
7	64	127	8128
8	128	255	*
9	256	511	*
10	512	1023	*
11	1024	2047	*
12	2048	4095	*
13	4096	8191	33,550,336

Notice that the first four entries of  $A$  are the perfect numbers 6, 28, 496 and 8128. At this point you may want to jump to the conclusion that  $A = P$ . But it is a shocking fact that in over 2000 years no one has ever been able to determine whether or not  $A = P$ . But it is known that  $A \subseteq P$ , and we will now prove it. In other words, we are going to show that every element of  $A$  is perfect. (But by itself, that leaves open the possibility that there may be some perfect numbers in  $P$  that are not in  $A$ .)

The main ingredient for the proof will be the formula for the sum of a geometric series with common ratio  $r$ . You probably saw this most recently in Calculus II. The formula is

$$\sum_{k=0}^n r^k = \frac{r^{n+1} - 1}{r - 1}.$$

We will need this for the case  $r = 2$ , which is

$$\sum_{k=0}^n 2^k = 2^{n+1} - 1. \quad (8.1)$$

(See the solution for Exercise 19 in Section 7.4 for a proof of this formula.) Now we are ready to prove our result. Let's draw attention to its significance by calling it a theorem rather than a proposition.

**Theorem 8.1** If  $A = \{2^{n-1}(2^n - 1) : n \in \mathbb{N}, \text{ and } 2^n - 1 \text{ is prime}\}$  and  $P = \{p \in \mathbb{N} : p \text{ is perfect}\}$ , then  $A \subseteq P$ .

*Proof.* Assume  $A$  and  $P$  are as stated. To show  $A \subseteq P$ , we must show that  $p \in A$  implies  $p \in P$ . Thus suppose  $p \in A$ . By definition of  $A$ , this means

$$p = 2^{n-1}(2^n - 1) \quad (8.2)$$

for some  $n \in \mathbb{N}$  for which  $2^n - 1$  is prime. We want to show that  $p \in P$ , that is, we want to show  $p$  is perfect. Thus, we need to show that the sum of the positive divisors of  $p$  that are less than  $p$  add up to  $p$ . Notice that since  $2^n - 1$  is prime, any divisor of  $p = 2^{n-1}(2^n - 1)$  must have the form  $2^k$  or  $2^k(2^n - 1)$  for  $0 \leq k \leq n - 1$ . Thus the positive divisors of  $p$  are as follows:

$$\begin{array}{cccccc} 2^0, & 2^1, & 2^2, & \dots & 2^{n-2}, & 2^{n-1}, \\ 2^0(2^n - 1), & 2^1(2^n - 1), & 2^2(2^n - 1), & \dots & 2^{n-2}(2^n - 1), & 2^{n-1}(2^n - 1). \end{array}$$

Notice that this list starts with  $2^0 = 1$  and ends with  $2^{n-1}(2^n - 1) = p$ .

If we add up all these divisors except for the last one (which equals  $p$ ) we get the following:

$$\begin{aligned}
 \sum_{k=0}^{n-1} 2^k + \sum_{k=0}^{n-2} 2^k (2^n - 1) &= \sum_{k=0}^{n-1} 2^k + (2^n - 1) \sum_{k=0}^{n-2} 2^k \\
 &= (2^n - 1) + (2^n - 1)(2^{n-1} - 1) \quad (\text{by Equation (8.1)}) \\
 &= [1 + (2^{n-1} - 1)](2^n - 1) \\
 &= 2^{n-1}(2^n - 1) \\
 &= p \quad (\text{by Equation (8.2)}).
 \end{aligned}$$

This shows that the positive divisors of  $p$  that are less than  $p$  add up to  $p$ . Therefore  $p$  is perfect, by definition of a perfect number. Thus  $p \in P$ , by definition of  $P$ .

We have shown that  $p \in A$  implies  $p \in P$ , which means  $A \subseteq P$ . ■

Combined with the chart on the previous page, this theorem gives us a new perfect number! The element  $p = 2^{13-1}(2^{13} - 1) = 33,550,336$  in  $A$  is perfect.

Observe also that every element of  $A$  is a multiple of a power of 2, and therefore even. But this does not necessarily mean every perfect number is even, because we've only shown  $A \subseteq P$ , not  $A = P$ . For all we know there may be odd perfect numbers in  $P - A$  that are not in  $A$ .

Are there any odd perfect numbers? No one knows.

In over 2000 years, no one has ever found an odd perfect number, nor has anyone been able to prove that there are none. But it is known that the set  $A$  does contain every *even* perfect number. This fact was first proved by Euler, and we duplicate his reasoning in the next theorem, which proves that  $A = E$ , where  $E$  is the set of all *even* perfect numbers. It is a good example of how to prove two sets are equal.

For convenience, we are going to use a slightly different definition of a perfect number. A number  $p \in \mathbb{N}$  is **perfect** if its positive divisors add up to  $2p$ . For example, the number 6 is perfect since the sum of its divisors is  $1 + 2 + 3 + 6 = 2 \cdot 6$ . This definition is simpler than the first one because we do not have to stipulate that we are adding up the divisors that are *less than*  $p$ . Instead we add in the last divisor  $p$ , and that has the effect of adding an additional  $p$ , thereby doubling the answer.

**Theorem 8.2** If  $A = \{2^{n-1}(2^n - 1) : n \in \mathbb{N}, \text{ and } 2^n - 1 \text{ is prime}\}$  and  $E = \{p \in \mathbb{N} : p \text{ is perfect and even}\}$ , then  $A = E$ .

*Proof.* To show that  $A = E$ , we need to show  $A \subseteq E$  and  $E \subseteq A$ .

First we will show that  $A \subseteq E$ . Suppose  $p \in A$ . This means  $p$  is even, because the definition of  $A$  shows that every element of  $A$  is a multiple of a power of 2. Also,  $p$  is a perfect number because Theorem 8.1 states that every element of  $A$  is also an element of  $P$ , hence perfect. Thus  $p$  is an even perfect number, so  $p \in E$ . Therefore  $A \subseteq E$ .

Next we show that  $E \subseteq A$ . Suppose  $p \in E$ . This means  $p$  is an even perfect number. Write the prime factorization of  $p$  as  $p = 2^k 3^{n_1} 5^{n_2} 7^{n_3} \dots$ , where some of the powers  $n_1, n_2, n_3 \dots$  may be zero. But, as  $p$  is even, the power  $k$  must be greater than zero. It follows  $p = 2^k q$  for some positive integer  $k$  and an odd integer  $q$ . Now, our aim is to show that  $p \in A$ , which means we must show  $p$  has form  $p = 2^{n-1}(2^n - 1)$ . To get our current  $p = 2^k q$  closer to this form, let  $n = k + 1$ , so we now have

$$p = 2^{n-1}q. \quad (8.3)$$

List the positive divisors of  $q$  as  $d_1, d_2, d_3, \dots, d_m$ . (Where  $d_1 = 1$  and  $d_m = q$ .) Then the divisors of  $p$  are:

$$\begin{array}{ccccc} 2^0 d_1 & 2^0 d_2 & 2^0 d_3 & \dots & 2^0 d_m \\ 2^1 d_1 & 2^1 d_2 & 2^1 d_3 & \dots & 2^1 d_m \\ 2^2 d_1 & 2^2 d_2 & 2^2 d_3 & \dots & 2^2 d_m \\ 2^3 d_1 & 2^3 d_2 & 2^3 d_3 & \dots & 2^3 d_m \\ \vdots & \vdots & \vdots & & \vdots \\ 2^{n-1} d_1 & 2^{n-1} d_2 & 2^{n-1} d_3 & \dots & 2^{n-1} d_m. \end{array}$$

Since  $p$  is perfect, these divisors add up to  $2p$ . By Equation (8.3), their sum is  $2p = 2(2^{n-1}q) = 2^n q$ . Adding the divisors column-by-column, we get

$$\sum_{k=0}^{n-1} 2^k d_1 + \sum_{k=0}^{n-1} 2^k d_2 + \sum_{k=0}^{n-1} 2^k d_3 + \dots + \sum_{k=0}^{n-1} 2^k d_m = 2^n q.$$

Applying Equation (8.1), this becomes

$$\begin{aligned} (2^n - 1)d_1 + (2^n - 1)d_2 + (2^n - 1)d_3 + \dots + (2^n - 1)d_m &= 2^n q \\ (2^n - 1)(d_1 + d_2 + d_3 + \dots + d_m) &= 2^n q \\ d_1 + d_2 + d_3 + \dots + d_m &= \frac{2^n q}{2^n - 1}, \end{aligned}$$

so that

$$d_1 + d_2 + d_3 + \cdots + d_m = \frac{(2^n - 1 + 1)q}{2^n - 1} = \frac{(2^n - 1)q + q}{2^n - 1} = q + \frac{q}{2^n - 1}.$$

From this we see that  $\frac{q}{2^n - 1}$  is an integer. It follows that both  $q$  and  $\frac{q}{2^n - 1}$  are positive divisors of  $q$ . Since their sum equals the sum of *all* positive divisors of  $q$ , it follows that  $q$  has only two positive divisors,  $q$  and  $\frac{q}{2^n - 1}$ . Since one of its divisors must be 1, it must be that  $\frac{q}{2^n - 1} = 1$ , which means  $q = 2^n - 1$ . Now a number with just two positive divisors is prime, so  $q = 2^n - 1$  is prime. Plugging this into Equation (8.3) gives  $p = 2^{n-1}(2^n - 1)$ , where  $2^n - 1$  is prime. This means  $p \in A$ , by definition of  $A$ . We have now shown that  $p \in E$  implies  $p \in A$ , so  $E \subseteq A$ .

Since  $A \subseteq E$  and  $E \subseteq A$ , it follows that  $A = E$ . ■

Do not be alarmed if you feel that you wouldn't have thought of this proof. It took the genius of Euler to discover this approach.

We'll conclude this chapter with some facts about perfect numbers.

- The sixth perfect number is  $p = 2^{17-1}(2^{17} - 1) = 8589869056$ .
- The seventh perfect number is  $p = 2^{19-1}(2^{19} - 1) = 137438691328$ .
- The eighth perfect number is  $p = 2^{31-1}(2^{31} - 1) = 2305843008139952128$ .
- The twentieth perfect number is  $p = 2^{4423-1}(2^{4423} - 1)$ . It has 2663 digits.
- The twenty-third perfect number  $p = 2^{11,213-1}(2^{11,213} - 1)$  has 6957 digits.
- The fiftieth perfect number is  $p = 2^{77,232,917-1}(2^{77,232,917} - 1)$ .

As mentioned earlier, no one knows whether or not there are any odd perfect numbers. It is not even known whether there are finitely many or infinitely many perfect numbers. It is known that the last digit of every even perfect number is either a 6 or an 8. Perhaps this is something you'd enjoy proving.

We've seen that perfect numbers are closely related to prime numbers having the form  $2^n - 1$ . Such prime numbers are called **Mersenne primes**, after the French scholar Marin Mersenne (1588–1648), who popularized them. The first several Mersenne primes are  $2^2 - 1 = 3$ ,  $2^3 - 1 = 7$ ,  $2^5 - 1 = 31$ ,  $2^7 - 1 = 127$  and  $2^{13} - 1 = 8191$ . To date, only 50 Mersenne primes are known, the largest of which is  $2^{77,232,917} - 1$ . There is a substantial cash prize for anyone who finds a 51st. (See <http://www.mersenne.org/>.) You may have better luck with the exercises.

## Exercises for Chapter 8

Use the methods introduced in this chapter to prove the following statements.

1. Prove that  $\{12n : n \in \mathbb{Z}\} \subseteq \{2n : n \in \mathbb{Z}\} \cap \{3n : n \in \mathbb{Z}\}$ .
2. Prove that  $\{6n : n \in \mathbb{Z}\} = \{2n : n \in \mathbb{Z}\} \cap \{3n : n \in \mathbb{Z}\}$ .
3. If  $k \in \mathbb{Z}$ , then  $\{n \in \mathbb{Z} : n | k\} \subseteq \{n \in \mathbb{Z} : n | k^2\}$ .
4. If  $m, n \in \mathbb{Z}$ , then  $\{x \in \mathbb{Z} : mn | x\} \subseteq \{x \in \mathbb{Z} : m | x\} \cap \{x \in \mathbb{Z} : n | x\}$ .
5. If  $p$  and  $q$  are positive integers, then  $\{pn : n \in \mathbb{N}\} \cap \{qn : n \in \mathbb{N}\} \neq \emptyset$ .
6. Suppose  $A, B$  and  $C$  are sets. Prove that if  $A \subseteq B$ , then  $A - C \subseteq B - C$ .
7. Suppose  $A, B$  and  $C$  are sets. If  $B \subseteq C$ , then  $A \times B \subseteq A \times C$ .
8. If  $A, B$  and  $C$  are sets, then  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ .
9. If  $A, B$  and  $C$  are sets, then  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ .
10. If  $A$  and  $B$  are sets in a universal set  $U$ , then  $\overline{A \cap B} = \overline{A} \cup \overline{B}$ .
11. If  $A$  and  $B$  are sets in a universal set  $U$ , then  $\overline{A \cup B} = \overline{A} \cap \overline{B}$ .
12. If  $A, B$  and  $C$  are sets, then  $A - (B \cap C) = (A - B) \cup (A - C)$ .
13. If  $A, B$  and  $C$  are sets, then  $A - (B \cup C) = (A - B) \cap (A - C)$ .
14. If  $A, B$  and  $C$  are sets, then  $(A \cup B) - C = (A - C) \cup (B - C)$ .
15. If  $A, B$  and  $C$  are sets, then  $(A \cap B) - C = (A - C) \cap (B - C)$ .
16. If  $A, B$  and  $C$  are sets, then  $A \times (B \cup C) = (A \times B) \cup (A \times C)$ .
17. If  $A, B$  and  $C$  are sets, then  $A \times (B \cap C) = (A \times B) \cap (A \times C)$ .
18. If  $A, B$  and  $C$  are sets, then  $A \times (B - C) = (A \times B) - (A \times C)$ .
19. Prove that  $\{9^n : n \in \mathbb{Z}\} \subseteq \{3^n : n \in \mathbb{Z}\}$ , but  $\{9^n : n \in \mathbb{Z}\} \neq \{3^n : n \in \mathbb{Z}\}$ .
20. Prove that  $\{9^n : n \in \mathbb{Q}\} = \{3^n : n \in \mathbb{Q}\}$ .
21. Suppose  $A$  and  $B$  are sets. Prove  $A \subseteq B$  if and only if  $A - B = \emptyset$ .
22. Let  $A$  and  $B$  be sets. Prove that  $A \subseteq B$  if and only if  $A \cap B = A$ .
23. For each  $a \in \mathbb{R}$ , let  $A_a = \{(x, a(x^2 - 1)) \in \mathbb{R}^2 : x \in \mathbb{R}\}$ . Prove that  $\bigcap_{a \in \mathbb{R}} A_a = \{(-1, 0), (1, 0)\}$ .
24. Prove that  $\bigcap_{x \in \mathbb{R}} [3 - x^2, 5 + x^2] = [3, 5]$ .
25. Suppose  $A, B, C$  and  $D$  are sets. Prove that  $(A \times B) \cup (C \times D) \subseteq (A \cup C) \times (B \cup D)$ .
26. Prove  $\{4k + 5 : k \in \mathbb{Z}\} = \{4k + 1 : k \in \mathbb{Z}\}$ .
27. Prove  $\{12a + 4b : a, b \in \mathbb{Z}\} = \{4c : c \in \mathbb{Z}\}$ .
28. Prove  $\{12a + 25b : a, b \in \mathbb{Z}\} = \mathbb{Z}$ .
29. Suppose  $A \neq \emptyset$ . Prove that  $A \times B \subseteq A \times C$ , if and only if  $B \subseteq C$ .
30. Prove that  $(\mathbb{Z} \times \mathbb{N}) \cap (\mathbb{N} \times \mathbb{Z}) = \mathbb{N} \times \mathbb{N}$ .
31. Suppose  $B \neq \emptyset$  and  $A \times B \subseteq B \times C$ . Prove  $A \subseteq C$ .

# CHAPTER 9

---

## Disproof

---

Ever since Chapter 4 we have dealt with one major theme: Given a statement, prove that it is true. In every example and exercise we were handed a true statement and charged with the task of proving it. Have you ever wondered what would happen if you were given a *false* statement to prove? The answer is that no (correct) proof would be possible, for if it were, the statement would be true, not false.

But how would you convince someone that a statement is false? The mere fact that you could not produce a proof does not automatically mean the statement is false, for you know (perhaps all too well) that proofs can be difficult to construct. It turns out that there is a very simple and utterly convincing procedure that proves a statement is false. The process of carrying out this procedure is called **disproof**. Thus, this chapter is concerned with **disproving** statements.

Before describing the new method, we will set the stage with some relevant background information. First, we point out that mathematical statements can be divided into three categories, described below.

One category consists of all those statements that have been proved to be true. For the most part we regard these statements as significant enough to be designated with special names such as “theorem,” “proposition,” “lemma” and “corollary.” Some examples of statements in this category are listed in the left-hand box in the diagram on the following page. There are also some wholly uninteresting statements (such as  $2 = 2$ ) in this category, and although we acknowledge their existence we certainly do not dignify them with terms such as “theorem” or “proposition.”

At the other extreme is a category consisting of statements that are known to be false. Examples are listed in the box on the right. Since mathematicians are not very interested in them, these types of statements do not get any special names, other than the blanket term “false statement.”

But there is a third (and quite interesting) category between these two extremes. It consists of statements whose truth or falsity has not been determined. Examples include things like “*Every perfect number*

*is even,” or “Every even integer greater than 2 is the sum of two primes.”* (The latter statement is called the *Goldbach conjecture*. See Section 2.1.) Mathematicians have a special name for the statements in this category that they suspect (but haven’t yet proved) are true. Such statements are called **conjectures**.

### THREE TYPES OF STATEMENTS:

Known to be true (Theorems & propositions)	Truth unknown (Conjectures)	Known to be false
<p>Examples:</p> <ul style="list-style-type: none"> <li>• Pythagorean theorem</li> <li>• Fermat’s last theorem (Section 2.1)</li> <li>• The square of an odd number is odd.</li> <li>• The series <math>\sum_{k=1}^{\infty} \frac{1}{k}</math> diverges.</li> </ul>	<p>Examples:</p> <ul style="list-style-type: none"> <li>• All perfect numbers are even.</li> <li>• Any even number greater than 2 is the sum of two primes. (Goldbach’s conjecture, Section 2.1)</li> <li>• There are infinitely many prime numbers of form <math>2^n - 1</math>, with <math>n \in \mathbb{N}</math>.</li> </ul>	<p>Examples:</p> <ul style="list-style-type: none"> <li>• All prime numbers are odd.</li> <li>• Some quadratic equations have three solutions.</li> <li>• <math>0 = 1</math></li> <li>• There exist natural numbers <math>a, b</math> and <math>c</math> for which <math>a^3 + b^3 = c^3</math>.</li> </ul>

Mathematicians spend much of their time and energy attempting to prove or disprove conjectures. (They also expend considerable mental energy in creating new conjectures based on collected evidence or intuition.) When a conjecture is proved (or disproved) the proof or disproof will typically appear in a published paper, provided the conjecture is of sufficient interest. If it is proved, the conjecture attains the status of a theorem or proposition. If it is disproved, then no one is really very interested in it anymore—mathematicians do not care much for false statements. (Though some disproved conjectures are viewed as instructive examples or curiosities, especially if the conjecture had been considered significant.)

Most conjectures that mathematicians are interested in are quite difficult to prove or disprove. We are not at that level yet. In this text, the “conjectures” that you will encounter are the kinds of statements that an experienced mathematician would immediately spot as true or false, but you may have to do some work before figuring out a proof or disproof. But in keeping with the cloud of uncertainty that surrounds conjectures at the advanced levels of mathematics, most exercises in this chapter (and many beyond it) will ask you to prove or disprove statements without giving any hint as to whether they are true or false. Your job will be to decide whether or not they are true and to either prove or disprove them. The examples

in this chapter will illustrate the processes one typically goes through in deciding whether a statement is true or false, and then verifying that it's true or false.

You know the three major methods of proving a statement: direct proof, contrapositive proof and proof by contradiction. Now we are ready to understand the method of disproving a statement. Suppose you want to disprove a statement  $P$ . In other words you want to prove that  $P$  is *false*. The way to do this is to prove that  $\sim P$  is *true*, for if  $\sim P$  is true, it follows immediately that  $P$  has to be false.

**How to disprove  $P$ :** Prove  $\sim P$ .

Our approach is incredibly simple. To disprove  $P$ , prove  $\sim P$ . In theory, this proof can be carried out by the direct, contrapositive or contradiction approaches. However, in practice things can be even easier than that if we are disproving a universally quantified statement or a conditional statement. That is our next topic.

### 9.1 Disproving Universal Statements: Counterexamples

A conjecture may be described as a statement that we hope is a theorem. As we know, many theorems (hence many conjectures) are universally quantified statements. Thus it seems reasonable to begin our discussion by investigating how to disprove a universally quantified statement such as

$$\forall x \in S, P(x).$$

To disprove this statement, we must prove its negation. Its negation is

$$\sim (\forall x \in S, P(x)) = \exists x \in S, \sim P(x).$$

The negation is an existence statement. To prove the negation is true, we just need to produce an *example* of an  $x \in S$  that makes  $\sim P(x)$  true, that is, an  $x$  that makes  $P(x)$  false. This leads to the following outline for disproving a universally quantified statement.

**How to disprove  $\forall x \in S, P(x)$ .**

Produce an example of an  $x \in S$  that makes  $P(x)$  false.

Things are just as simple if we want to disprove a conditional statement  $P(x) \Rightarrow Q(x)$ . This statement asserts that for every  $x$  that makes  $P(x)$  true,  $Q(x)$  will also be true. The statement can only be false if there is an  $x$  that makes  $P(x)$  true and  $Q(x)$  false. This leads to our next outline for disproof.

**How to disprove  $P(x) \Rightarrow Q(x)$ .**

Produce an example of an  $x$  that makes  $P(x)$  true and  $Q(x)$  false.

In both of the above outlines, the statement is disproved simply by exhibiting an example that shows the statement is not always true. (Think of it as an example that exposes the statement as a promise that can be broken.) There is a special name for an example that disproves a statement: It is called a **counterexample**.

**Example 9.1** As our first example, we will work through the process of deciding whether or not the following conjecture is true.

**Conjecture** For every  $n \in \mathbb{Z}$ , the integer  $f(n) = n^2 - n + 11$  is prime.

In resolving the truth or falsity of a conjecture, it's a good idea to gather as much information about the conjecture as possible. In this case let's start by making a table that tallies the values of  $f(n)$  for some integers  $n$ .

$n$	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
$f(n)$	23	17	13	11	11	13	17	23	31	41	53	67	83	101

In every case,  $f(n)$  is prime, so you may begin to suspect that the conjecture is true. Before attempting a proof, let's try one more  $n$ . Unfortunately,  $f(11) = 11^2 - 11 + 11 = 11^2$  is not prime. The conjecture is false because  $n = 11$  is a counterexample. We summarize our disproof as follows:

*Disproof.* The statement “For every  $n \in \mathbb{Z}$ , the integer  $f(n) = n^2 - n + 11$  is prime,” is **false**. For a counterexample, note that for  $n = 11$ , the integer  $f(11) = 121 = 11 \cdot 11$  is not prime. ■

In disproving a statement with a counterexample, it is important to explain exactly how the counterexample makes the statement false. Our work would not have been complete if we had just said “for a counterexample, consider  $n = 11$ ,” and left it at that. We need to show that the answer  $f(11)$  is not prime. Showing the factorization  $f(11) = 11 \cdot 11$  suffices for this.

**Example 9.2** Either prove or disprove the following conjecture.

**Conjecture** If  $A$ ,  $B$  and  $C$  are sets, then  $A - (B \cap C) = (A - B) \cap (A - C)$ .

*Disproof.* This conjecture is false because of the following counterexample. Let  $A = \{1, 2, 3\}$ ,  $B = \{1, 2\}$  and  $C = \{2, 3\}$ . Notice that  $A - (B \cap C) = \{1, 3\}$  and  $(A - B) \cap (A - C) = \emptyset$ , so  $A - (B \cap C) \neq (A - B) \cap (A - C)$ . ■

(To see where this counterexample came from, draw Venn diagrams for  $A - (B \cap C)$  and  $(A - B) \cap (A - C)$ . You will see that the diagrams are different. The numbers 1, 2 and 3 can then be inserted into the regions of the diagrams in such a way as to create the above counterexample.)

## 9.2 Disproving Existence Statements

We have seen that we can disprove a universally quantified statement or a conditional statement simply by finding a counterexample. Now let's turn to the problem of disproving an existence statement such as

$$\exists x \in S, P(x).$$

Proving this would involve simply finding an example of an  $x$  that makes  $P(x)$  true. To *disprove* it, we have to prove its negation  $\sim (\exists x \in S, P(x)) = \forall x \in S, \sim P(x)$ . But this negation is universally quantified. Proving *it* involves showing that  $\sim P(x)$  is true for *all*  $x \in S$ , and for this an example does not suffice. Instead we must use direct, contrapositive or contradiction proof to prove the conditional statement "*If*  $x \in S$ , *then*  $\sim P(x)$ ." As an example, here is a conjecture to either prove or disprove.

**Example 9.3** Either prove or disprove the following conjecture.

**Conjecture** There is a real number  $x$  for which  $x^4 < x < x^2$ .

This may not seem like an unreasonable statement at first glance. After all, if the statement were asserting the existence of a real number for which  $x^3 < x < x^2$ , then it would be true: just take  $x = -2$ . But it asserts there is an  $x$  for which  $x^4 < x < x^2$ . When we apply some intelligent guessing to locate such an  $x$  we run into trouble. If  $x = \frac{1}{2}$ , then  $x^4 < x$ , but we don't have  $x < x^2$ ; similarly if  $x = 2$ , we have  $x < x^2$  but not  $x^4 < x$ . Since finding an  $x$  with  $x^4 < x < x^2$  seems problematic, we may begin to suspect that the given statement is false.

Let's see if we can disprove it. According to our strategy for disproof, to *disprove* it we must *prove* its negation. Symbolically, the statement is

$\exists x \in \mathbb{R}, x^4 < x < x^2$ , so its negation is

$$\sim (\exists x \in \mathbb{R}, x^4 < x < x^2) = \forall x \in \mathbb{R}, \sim (x^4 < x < x^2).$$

Thus, in words the negation is

*For every real number  $x$ , it is not the case that  $x^4 < x < x^2$ .*

This can be proved with contradiction, as follows. Suppose for the sake of contradiction that there is an  $x$  for which  $x^4 < x < x^2$ . Then  $x$  must be positive since it's greater than the non-negative number  $x^4$ . Dividing all parts of  $x^4 < x < x^2$  by the positive number  $x$  produces  $x^3 < 1 < x$ . Now subtract 1 from all parts of  $x^3 < 1 < x$  to obtain  $x^3 - 1 < 0 < x - 1$  and reason as follows:

$$\begin{aligned} x^3 - 1 &< 0 < x - 1 \\ (x - 1)(x^2 + x + 1) &< 0 < (x - 1) \\ x^2 + x + 1 &< 0 < 1 \end{aligned}$$

(Division by  $x - 1$  did not reverse the inequality  $<$  because the second line above shows  $0 < x - 1$ , that is,  $x - 1$  is positive.) Now we have  $x^2 + x + 1 < 0$ , which is a contradiction because  $x$  being positive forces  $x^2 + x + 1 > 0$ .

We summarize our work as follows.

The statement "*There is a real number  $x$  for which  $x^4 < x < x^2$* " is **false** because we have proved its negation "*For every real number  $x$ , it is not the case that  $x^4 < x < x^2$* ".

As you work the exercises, keep in mind that not every conjecture will be false. If one is true, then a disproof is impossible and you must produce a proof. Here is an example:

**Example 9.4** Either prove or disprove the following conjecture.

**Conjecture** There exist three integers  $x, y, z$ , all greater than 1 and no two equal, for which  $x^y = y^z$ .

This conjecture is true. It is an existence statement, so to prove it we just need to give an example of three integers  $x, y, z$ , all greater than 1 and no two equal, so that  $x^y = y^z$ . A proof follows.

**Proposition** There exist three integers  $x, y, z$ , all greater than 1 and no two equal, for which  $x^y = y^z$ .

*Proof.* Note that if  $x = 2$ ,  $y = 16$  and  $z = 4$ , then  $x^y = 2^{16} = (2^4)^4 = 16^4 = y^z$ . ■

### 9.3 Disproof by Contradiction

Contradiction can be a very useful way to disprove a statement. To see how this works, suppose we wish to disprove a statement  $P$ . We know that to disprove  $P$ , we must *prove*  $\sim P$ . To prove  $\sim P$  with contradiction, we assume  $\sim\sim P$  is true and deduce a contradiction. But since  $\sim\sim P = P$ , this boils down to assuming  $P$  is true and deducing a contradiction. Here is an outline:

**How to disprove  $P$  with contradiction:**

Assume  $P$  is true, and deduce a contradiction.

To illustrate this, let's revisit Example 9.3 but do the disproof with contradiction. You will notice that the work duplicates much of what we did in Example 9.3, but is it much more streamlined because here we do not have to negate the conjecture.

**Example 9.5** Disprove the following conjecture.

**Conjecture** There is a real number  $x$  for which  $x^4 < x < x^2$ .

*Disproof.* Suppose for the sake of contradiction that this conjecture is true. Let  $x$  be a real number for which  $x^4 < x < x^2$ . Then  $x$  is positive, since it is greater than the non-negative number  $x^4$ . Dividing all parts of  $x^4 < x < x^2$  by the positive number  $x$  produces  $x^3 < 1 < x$ . Now subtract 1 from all parts of  $x^3 < 1 < x$  to obtain  $x^3 - 1 < 0 < x - 1$  and reason as follows:

$$\begin{aligned} x^3 - 1 &< 0 < x - 1 \\ (x - 1)(x^2 + x + 1) &< 0 < (x - 1) \\ x^2 + x + 1 &< 0 < 1 \end{aligned}$$

Now we have  $x^2 + x + 1 < 0$ , which is a contradiction because  $x$  is positive. Thus the conjecture must be false. ■

### Exercises for Chapter 9

Each of the following statements is either true or false. If a statement is true, prove it. If a statement is false, disprove it. These exercises are cumulative, covering all topics addressed in Chapters 1–9.

1. If  $x, y \in \mathbb{R}$ , then  $|x + y| = |x| + |y|$ .
2. For every natural number  $n$ , the integer  $2n^2 - 4n + 31$  is prime.
3. If  $n \in \mathbb{Z}$  and  $n^5 - n$  is even, then  $n$  is even.

4. For every natural number  $n$ , the integer  $n^2 + 17n + 17$  is prime.
5. If  $A, B, C$  and  $D$  are sets, then  $(A \times B) \cup (C \times D) = (A \cup C) \times (B \cup D)$ .
6. If  $A, B, C$  and  $D$  are sets, then  $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$ .
7. If  $A, B$  and  $C$  are sets, and  $A \times C = B \times C$ , then  $A = B$ .
8. If  $A, B$  and  $C$  are sets, then  $A - (B \cup C) = (A - B) \cup (A - C)$ .
9. If  $A$  and  $B$  are sets, then  $\mathcal{P}(A) - \mathcal{P}(B) \subseteq \mathcal{P}(A - B)$ .
10. If  $A$  and  $B$  are sets and  $A \cap B = \emptyset$ , then  $\mathcal{P}(A) - \mathcal{P}(B) \subseteq \mathcal{P}(A - B)$ .
11. If  $a, b \in \mathbb{N}$ , then  $a + b < ab$ .
12. If  $a, b, c \in \mathbb{N}$  and  $ab, bc$  and  $ac$  all have the same parity, then  $a, b$  and  $c$  all have the same parity.
13. There exists a set  $X$  for which  $\mathbb{R} \subseteq X$  and  $\emptyset \in X$ .
14. If  $A$  and  $B$  are sets, then  $\mathcal{P}(A) \cap \mathcal{P}(B) = \mathcal{P}(A \cap B)$ .
15. Every odd integer is the sum of three odd integers.
16. If  $A$  and  $B$  are finite sets, then  $|A \cup B| = |A| + |B|$ .
17. For all sets  $A$  and  $B$ , if  $A - B = \emptyset$ , then  $B \neq \emptyset$ .
18. If  $a, b, c \in \mathbb{N}$ , then at least one of  $a - b$ ,  $a + c$  and  $b - c$  is even.
19. For every  $r, s \in \mathbb{Q}$  with  $r < s$ , there is an irrational number  $u$  for which  $r < u < s$ .
20. There exist prime numbers  $p$  and  $q$  for which  $p - q = 1000$ .
21. There exist prime numbers  $p$  and  $q$  for which  $p - q = 97$ .
22. If  $p$  and  $q$  are prime numbers for which  $p < q$ , then  $2p + q^2$  is odd.
23. If  $x, y \in \mathbb{R}$  and  $x^3 < y^3$ , then  $x < y$ .
24. The inequality  $2^x \geq x + 1$  is true for all positive real numbers  $x$ .
25. For all  $a, b, c \in \mathbb{Z}$ , if  $a \mid bc$ , then  $a \mid b$  or  $a \mid c$ .
26. Suppose  $A, B$  and  $C$  are sets. If  $A = B - C$ , then  $B = A \cup C$ .
27. The equation  $x^2 = 2^x$  has three real solutions.
28. Suppose  $a, b \in \mathbb{Z}$ . If  $a \mid b$  and  $b \mid a$ , then  $a = b$ .
29. If  $x, y \in \mathbb{R}$  and  $|x + y| = |x - y|$ , then  $y = 0$ .
30. There exist integers  $a$  and  $b$  for which  $42a + 7b = 1$ .
31. No number (other than 1) appears in Pascal's triangle more than four times.
32. If  $n, k \in \mathbb{N}$  and  $\binom{n}{k}$  is a prime number, then  $k = 1$  or  $k = n - 1$ .
33. Suppose  $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$  is a polynomial of degree 1 or greater, and for which each coefficient  $a_i$  is in  $\mathbb{N}$ . Then there is an  $n \in \mathbb{N}$  for which the integer  $f(n)$  is not prime.
34. If  $X \subseteq A \cup B$ , then  $X \subseteq A$  or  $X \subseteq B$ .

# CHAPTER 10

---

## Mathematical Induction

---

This chapter explains a powerful proof technique called **mathematical induction** (or just **induction** for short). To motivate the discussion, let's first examine the kinds of statements that induction is used to prove. Consider this statement:

**Conjecture** The sum of the first  $n$  odd natural numbers equals  $n^2$ .

The following table illustrates what this conjecture says. Each row is headed by a natural number  $n$ , followed by the sum of the first  $n$  odd natural numbers, followed by  $n^2$ .

$n$	sum of the first $n$ odd natural numbers	$n^2$
1	$1 = \dots$	1
2	$1+3 = \dots$	4
3	$1+3+5 = \dots$	9
4	$1+3+5+7 = \dots$	16
5	$1+3+5+7+9 = \dots$	25
$\vdots$	$\vdots$	$\vdots$
$n$	$1+3+5+7+9+11+\dots+(2n-1) = \dots$	$n^2$
$\vdots$	$\vdots$	$\vdots$

Note that in the first five lines of the table, the sum of the first  $n$  odd numbers really does add up to  $n^2$ . Notice also that these first five lines indicate that the  $n$ th odd natural number (the last number in each sum) is  $2n - 1$ . (For instance, when  $n = 2$ , the second odd natural number is  $2 \cdot 2 - 1 = 3$ ; when  $n = 3$ , the third odd natural number is  $2 \cdot 3 - 1 = 5$ , etc.)

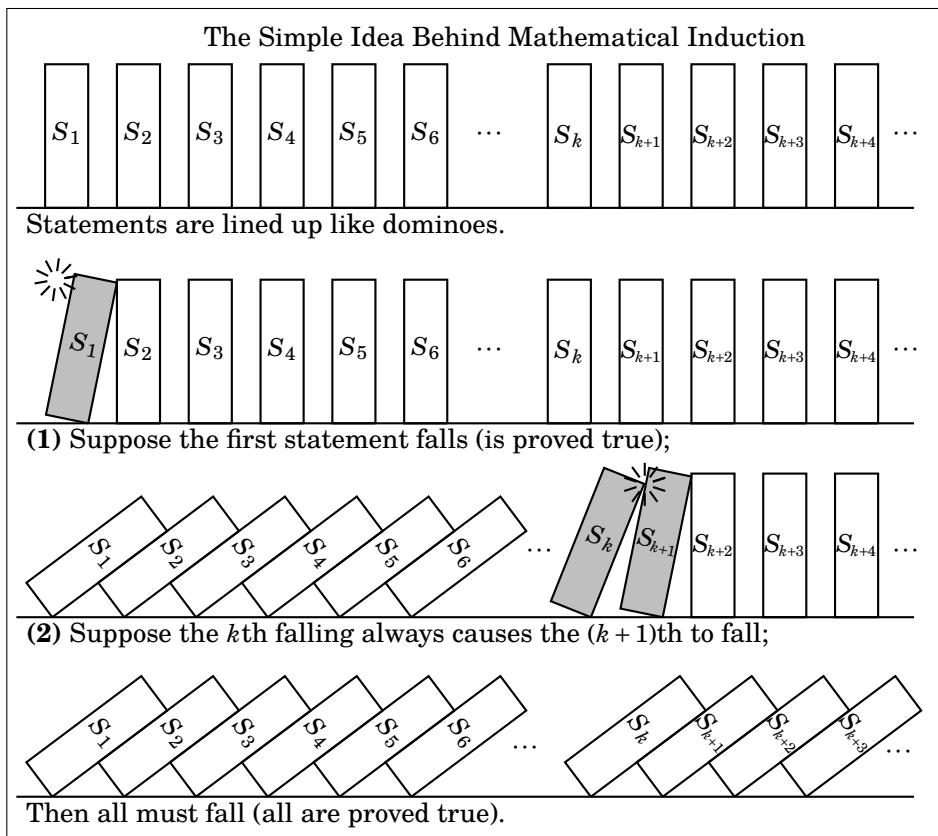
The table raises a question. Does the sum  $1+3+5+7+\dots+(2n-1)$  really always equal  $n^2$ ? In other words, is the conjecture true?

Let's rephrase this. For each natural number  $n$  (i.e., for each line of the table), we have a statement  $S_n$ , as follows:

$$\begin{aligned}
 S_1 &: 1 = 1^2 \\
 S_2 &: 1 + 3 = 2^2 \\
 S_3 &: 1 + 3 + 5 = 3^2 \\
 &\vdots \\
 S_n &: 1 + 3 + 5 + 7 + \cdots + (2n - 1) = n^2 \\
 &\vdots
 \end{aligned}$$

Our question is: Are all of these statements true?

Mathematical induction answers just this kind of question, where we have an infinite list of statements  $S_1, S_2, S_3, \dots$  that we want to prove true. The method is really quite simple. To visualize it, think of the statements as dominoes, lined up in a row. Suppose you can prove the first statement  $S_1$ , and symbolize this as domino  $S_1$  being knocked down. Also, say you can prove that any statement  $S_k$  being true (falling) forces the next statement  $S_{k+1}$  to be true (to fall). Then  $S_1$  falls, knocking down  $S_2$ . Next  $S_2$  falls, knocking down  $S_3$ , then  $S_3$  knocks down  $S_4$ , and so on. The inescapable conclusion is that all the statements are knocked down (proved true).



## 10.1 Proof by Induction

This domino analogy motivates an outline for our next major proof technique: *proof by mathematical induction*.

### Outline for Proof by Induction

**Proposition** The statements  $S_1, S_2, S_3, S_4, \dots$  are all true.

*Proof.* (Induction)

(1) Prove that the first statement  $S_1$  is true.

(2) Given any integer  $k \geq 1$ , prove that the statement  $S_k \Rightarrow S_{k+1}$  is true.

It follows by mathematical induction that every  $S_n$  is true. ■

In this setup, the first step (1) is called the **basis step**. Because  $S_1$  is usually a very simple statement, the basis step is often quite easy to do. The second step (2) is called the **inductive step**. In the inductive step direct proof is most often used to prove  $S_k \Rightarrow S_{k+1}$ , so this step is usually carried out by assuming  $S_k$  is true and showing this forces  $S_{k+1}$  to be true. The assumption that  $S_k$  is true is called the **inductive hypothesis**.

Now let's apply this technique to our original conjecture that the sum of the first  $n$  odd natural numbers equals  $n^2$ . Our goal is to show that for each  $n \in \mathbb{N}$ , the statement  $S_n : 1 + 3 + 5 + 7 + \dots + (2n - 1) = n^2$  is true. Before getting started, observe that  $S_k$  is obtained from  $S_n$  by plugging  $k$  in for  $n$ . Thus  $S_k$  is the statement  $S_k : 1 + 3 + 5 + 7 + \dots + (2k - 1) = k^2$ . Also, we get  $S_{k+1}$  by plugging in  $k + 1$  for  $n$ , so that  $S_{k+1} : 1 + 3 + 5 + 7 + \dots + (2(k + 1) - 1) = (k + 1)^2$ .

**Proposition** If  $n \in \mathbb{N}$ , then  $1 + 3 + 5 + 7 + \dots + (2n - 1) = n^2$ .

*Proof.* We will prove this with mathematical induction.

(1) Observe that if  $n = 1$ , this statement is  $1 = 1^2$ , which is obviously true.

(2) We must now prove  $S_k \Rightarrow S_{k+1}$  for any  $k \geq 1$ . That is, we must show that if  $1 + 3 + 5 + 7 + \dots + (2k - 1) = k^2$ , then  $1 + 3 + 5 + 7 + \dots + (2(k + 1) - 1) = (k + 1)^2$ .

We use direct proof. Suppose  $1 + 3 + 5 + 7 + \dots + (2k - 1) = k^2$ . Then

$$\begin{aligned} 1 + 3 + 5 + 7 + \dots + (2(k + 1) - 1) &= \\ 1 + 3 + 5 + 7 + \dots + (2k - 1) + (2(k + 1) - 1) &= \\ (1 + 3 + 5 + 7 + \dots + (2k - 1)) + (2(k + 1) - 1) &= \\ k^2 + (2(k + 1) - 1) &= k^2 + 2k + 1 \\ &= (k + 1)^2. \end{aligned}$$

Thus  $1 + 3 + 5 + 7 + \dots + (2(k + 1) - 1) = (k + 1)^2$ . This proves that  $S_k \Rightarrow S_{k+1}$ . It follows by induction that  $1 + 3 + 5 + 7 + \dots + (2n - 1) = n^2$  for every  $n \in \mathbb{N}$ . ■

In induction proofs it is usually the case that the first statement  $S_1$  is indexed by the natural number 1, but this need not always be so. Depending on the problem, the first statement could be  $S_0$ , or  $S_m$  for any other integer  $m$ . In the next example the statements are  $S_0, S_1, S_2, S_3, \dots$  The same outline is used, except that the basis step verifies  $S_0$ , not  $S_1$ .

**Proposition** If  $n$  is a non-negative integer, then  $5 | (n^5 - n)$ .

*Proof.* We will prove this with mathematical induction. Observe that the first non-negative integer is 0, so the basis step involves  $n = 0$ .

- (1) If  $n = 0$ , this statement is  $5 | (0^5 - 0)$  or  $5 | 0$ , which is obviously true.
- (2) Let  $k \geq 0$ . We need to prove that if  $5 | (k^5 - k)$ , then  $5 | ((k+1)^5 - (k+1))$ . We use direct proof. Suppose  $5 | (k^5 - k)$ . Thus  $k^5 - k = 5a$  for some  $a \in \mathbb{Z}$ . Observe that

$$\begin{aligned} (k+1)^5 - (k+1) &= k^5 + 5k^4 + 10k^3 + 10k^2 + 5k + 1 - k - 1 \\ &= (k^5 - k) + 5k^4 + 10k^3 + 10k^2 + 5k \\ &= 5a + 5k^4 + 10k^3 + 10k^2 + 5k \\ &= 5(a + k^4 + 2k^3 + 2k^2 + k). \end{aligned}$$

This shows  $(k+1)^5 - (k+1)$  is an integer multiple of 5, so  $5 | ((k+1)^5 - (k+1))$ . We have now shown that  $5 | (k^5 - k)$  implies  $5 | ((k+1)^5 - (k+1))$ .

It follows by induction that  $5 | (n^5 - n)$  for all non-negative integers  $n$ . ■

As noted, induction is used to prove statements of the form  $\forall n \in \mathbb{N}, S_n$ . But notice the outline does *not* work for statements of form  $\forall n \in \mathbb{Z}, S_n$  (where  $n$  is in  $\mathbb{Z}$ , not  $\mathbb{N}$ ). The reason is that if you are trying to prove  $\forall n \in \mathbb{Z}, S_n$  by induction, and you've shown  $S_1$  is true and  $S_k \Rightarrow S_{k+1}$ , then it only follows from this that  $S_n$  is true for  $n \geq 1$ . You haven't proved that any of the statements  $S_0, S_{-1}, S_{-2}, \dots$  are true. If you ever want to prove  $\forall n \in \mathbb{Z}, S_n$  by induction, you have to show that some  $S_a$  is true and  $S_k \Rightarrow S_{k+1}$  **and**  $S_k \Rightarrow S_{k-1}$ .

Unfortunately, the term *mathematical induction* is sometimes confused with *inductive reasoning*, which is the process of reaching the conclusion that something is likely to be true based on prior observations of similar circumstances. Please note that mathematical induction—as introduced in this chapter—is a rigorous technique that proves statements with absolute certainty.

To round out this section, we present four additional induction proofs.

**Proposition** If  $n \in \mathbb{Z}$  and  $n \geq 0$ , then  $\sum_{i=0}^n i \cdot i! = (n+1)! - 1$ .

*Proof.* We will prove this with mathematical induction.

- (1) If  $n = 0$ , this statement is  $\sum_{i=0}^0 i \cdot i! = (0+1)! - 1$ . The left-hand side is  $0 \cdot 0! = 0$ , and the right-hand side is  $1! - 1 = 0$ . Thus the equation holds, as both sides are zero.
- (2) Consider any integer  $k \geq 0$ . We must show that  $S_k$  implies  $S_{k+1}$ . That is, we must show that

$$\sum_{i=0}^k i \cdot i! = (k+1)! - 1 \quad \text{implies} \quad \sum_{i=0}^{k+1} i \cdot i! = ((k+1)+1)! - 1.$$

We use direct proof. Suppose  $\sum_{i=0}^k i \cdot i! = (k+1)! - 1$ . Observe that

$$\begin{aligned} \sum_{i=0}^{k+1} i \cdot i! &= \left( \sum_{i=0}^k i \cdot i! \right) + (k+1)(k+1)! \\ &= ((k+1)! - 1) + (k+1)(k+1)! \\ &= (k+1)! + (k+1)(k+1)! - 1 \\ &= (1 + (k+1))(k+1)! - 1 \\ &= (k+2)(k+1)! - 1 \\ &= (k+2)! - 1 \\ &= ((k+1)+1)! - 1. \end{aligned}$$

Therefore  $\sum_{i=0}^{k+1} i \cdot i! = ((k+1)+1)! - 1$ .

We have now proved by induction that  $\sum_{i=0}^n i \cdot i! = (n+1)! - 1$  for every integer  $n \geq 0$ . ■

In our outline for proof by induction, the inductive step (2) involved proving  $S_k \Rightarrow S_{k+1}$ . Obviously, you can prove  $S_n \Rightarrow S_{n+1}$  instead. (That is, assume the statement is true for  $n$ , and show that it is true for  $n+1$ .) Sometimes proving  $S_{n-1} \Rightarrow S_n$  is more convenient, and this too is valid. The proofs in the following examples will use the scheme  $S_k \Rightarrow S_{k+1}$ , but some solutions to odd-numbered exercises will be phrased as  $S_n \Rightarrow S_{n+1}$  or  $S_{n-1} \Rightarrow S_n$ . Fluency comes with reading and practice.

The next example illustrates a trick that is occasionally useful. You know that you can add equal quantities to both sides of an equation without violating equality. But don't forget that you can add *unequal* quantities to both sides of an *inequality*, as long as the quantity added to the bigger side is bigger than the quantity added to the smaller side. For example, if  $x \leq y$  and  $a \leq b$ , then  $x + a \leq y + b$ . Similarly, if  $x \leq y$  and  $b$  is positive, then  $x \leq y + b$ . This oft-neglected fact is used in the next proof.

**Proposition** The inequality  $2^n \leq 2^{n+1} - 2^{n-1} - 1$  holds for each  $n \in \mathbb{N}$ .

*Proof.* We will prove this with mathematical induction.

- (1) If  $n = 1$ , this statement is  $2^1 \leq 2^{1+1} - 2^{1-1} - 1$ , and this simplifies to  $2 \leq 4 - 1 - 1$ , which is obviously true.
- (2) Say  $k \geq 1$ . We use direct proof to show that  $2^k \leq 2^{k+1} - 2^{k-1} - 1$  implies  $2^{k+1} \leq 2^{(k+1)+1} - 2^{(k+1)-1} - 1$ . Suppose  $2^k \leq 2^{k+1} - 2^{k-1} - 1$ . Then

$$\begin{aligned} 2^k &\leq 2^{k+1} - 2^{k-1} - 1 \\ 2(2^k) &\leq 2(2^{k+1} - 2^{k-1} - 1) && \text{(multiply both sides by 2)} \\ 2^{k+1} &\leq 2^{k+2} - 2^k - 2 \\ 2^{k+1} &\leq 2^{k+2} - 2^k - 2 + 1 && \text{(add 1 to the bigger side)} \\ 2^{k+1} &\leq 2^{k+2} - 2^k - 1 \\ 2^{k+1} &\leq 2^{(k+1)+1} - 2^{(k+1)-1} - 1. \end{aligned}$$

It follows by induction that  $2^n \leq 2^{n+1} - 2^{n-1} - 1$  for each  $n \in \mathbb{N}$ . ■

We next prove that if  $n \in \mathbb{N}$ , then the inequality  $(1+x)^n \geq 1 + nx$  holds for all  $x \in \mathbb{R}$  with  $x > -1$ . Thus we will need to prove that the statement

$$S_n : (1+x)^n \geq 1 + nx \text{ for every } x \in \mathbb{R} \text{ with } x > -1$$

is true for every natural number  $n$ . This is (only) slightly different from our other examples, which proved statements of the form  $\forall n \in \mathbb{N}, P(n)$ , where  $P(n)$  is a statement about the number  $n$ . This time we are proving something of form

$$\forall n \in \mathbb{N}, \left( \forall x \in (-1, \infty), P(n, x) \right),$$

where the open sentence  $P(n, x) : (1+x)^n \geq 1 + nx$  involves not only  $n$ , but also a second variable  $x$ . (For the record, the inequality  $(1+x)^n \geq 1 + nx$  is known as *Bernoulli's inequality*.)

**Proposition** If  $n \in \mathbb{N}$ , then  $(1+x)^n \geq 1+nx$  for all  $x \in \mathbb{R}$  with  $x > -1$ .

*Proof.* We will prove this with mathematical induction.

- (1) For the basis step, notice that when  $n = 1$  the statement is  $(1+x)^1 \geq 1+1 \cdot x$ , and this is true because both sides equal  $1+x$ .
- (2) Assume that for some  $k \geq 1$ , the statement  $(1+x)^k \geq 1+kx$  is true for all  $x \in \mathbb{R}$  with  $x > -1$ . From this we need to prove  $(1+x)^{k+1} \geq 1+(k+1)x$ . Now,  $1+x$  is positive because  $x > -1$ , so we can multiply both sides of  $(1+x)^k \geq 1+kx$  by  $(1+x)$  without changing the direction of the  $\geq$ .

$$\begin{aligned}(1+x)^k(1+x) &\geq (1+kx)(1+x) \\ (1+x)^{k+1} &\geq 1+x+kx+kx^2 \\ (1+x)^{k+1} &\geq 1+(k+1)x+kx^2\end{aligned}$$

The above term  $kx^2$  is positive, so removing it from the right-hand side will only make that side smaller. Thus we get  $(1+x)^{k+1} \geq 1+(k+1)x$ . ■

Next, an example where the basis step involves more than routine checking. (It will be used later, so it is numbered for reference.)

**Proposition 10.1** Suppose  $a_1, a_2, \dots, a_n$  are  $n$  integers, where  $n \geq 2$ . If  $p$  is prime and  $p \mid (a_1 \cdot a_2 \cdot a_3 \cdots a_n)$ , then  $p \mid a_i$  for at least one of the  $a_i$ .

*Proof.* The proof is induction on  $n$ .

- (1) The basis step involves  $n = 2$ . Let  $p$  be prime and suppose  $p \mid (a_1 a_2)$ . We need to show that  $p \mid a_1$  or  $p \mid a_2$ , or equivalently, if  $p \nmid a_1$ , then  $p \mid a_2$ . Thus suppose  $p \nmid a_1$ . Since  $p$  is prime, it follows that  $\gcd(p, a_1) = 1$ . By Proposition 7.1 (on page 152), there are integers  $k$  and  $\ell$  for which  $1 = pk + a_1\ell$ . Multiplying this by  $a_2$  gives

$$a_2 = pka_2 + a_1a_2\ell.$$

As we are assuming that  $p$  divides  $a_1a_2$ , it is clear that  $p$  divides the expression  $pka_2 + a_1a_2\ell$  on the right; hence  $p \mid a_2$ . We've now proved that if  $p \mid (a_1a_2)$ , then  $p \mid a_1$  or  $p \mid a_2$ . This completes the basis step.

- (2) Suppose that  $k \geq 2$ , and  $p \mid (a_1 \cdot a_2 \cdots a_k)$  implies then  $p \mid a_i$  for some  $a_i$ . Now let  $p \mid (a_1 \cdot a_2 \cdots a_k \cdot a_{k+1})$ . Then  $p \mid ((a_1 \cdot a_2 \cdots a_k) \cdot a_{k+1})$ . By what we proved in the basis step, it follows that  $p \mid (a_1 \cdot a_2 \cdots a_k)$  or  $p \mid a_{k+1}$ . This and the inductive hypothesis imply that  $p$  divides one of the  $a_i$ . ■

Please test your understanding now by working a few exercises.

## 10.2 Proof by Strong Induction

Sometimes in an induction proof it is hard to show that  $S_k$  implies  $S_{k+1}$ . It may be easier to show some “lower”  $S_m$  (with  $m < k$ ) implies  $S_{k+1}$ . For such situations there is a slight variant of induction called strong induction. Strong induction works just like regular induction, except that in Step (2) instead of assuming  $S_k$  is true and showing this forces  $S_{k+1}$  to be true, we assume that *all* the statements  $S_1, S_2, \dots, S_k$  are true and show this forces  $S_{k+1}$  to be true. The idea is that if the first  $k$  dominoes falling always forces the  $(k+1)$ th domino to fall, then all the dominoes must fall.

### Outline for Proof by Strong Induction

**Proposition** The statements  $S_1, S_2, S_3, S_4, \dots$  are all true.

*Proof.* (Strong induction)

- (1) Prove the first statement  $S_1$ . (Or the first several  $S_n$ , if needed.)
- (2) Given any integer  $k \geq 1$ , prove  $(S_1 \wedge S_2 \wedge S_3 \wedge \dots \wedge S_k) \Rightarrow S_{k+1}$ . ■

This is useful when  $S_k$  does not easily imply  $S_{k+1}$ . You may be better served by showing some earlier statement ( $S_{k-1}$  or  $S_{k-2}$ , for instance.) implies  $S_k$ . In strong induction you can use any (or all) of  $S_1, S_2, \dots, S_k$  to prove  $S_{k+1}$ .

Here is a classic first example of a strong induction proof: The problem is to prove that you can achieve any postage of 8 cents or more, exactly, using only 3¢ and 5¢ stamps. For example, for a postage of 47 cents, you could use nine 3¢ stamps and four 5¢ stamps. Let  $S_n$  be the statement  $S_n$ : *You can get a postage of exactly  $n$ ¢ using only 3¢ and 5¢ stamps.* Thus we need to prove all the statements  $S_8, S_9, S_{10}, S_{11} \dots$  are true. In the proof, to show  $S_{k+1}$  is true we will need to “go back” three steps from  $S_{k+1}$ , so the basis step involves verifying the first **three** statements  $S_8, S_9$  and  $S_{10}$ .

**Proposition** Any postage of 8¢ or more is possible using 3¢ and 5¢ stamps.

*Proof.* We will use strong induction.

- (1) This holds for postages of 8, 9 and 10 cents: For 8¢, use one 3¢ stamp and one 5¢ stamp. For 9¢, three 3¢ stamps. For 10¢, two 5¢ stamps.
- (2) Let  $k \geq 10$ , and for each  $8 \leq m \leq k$ , assume a postage of  $m$  cents can be obtained exactly with 3¢ and 5¢ stamps. (That is, assume statements  $S_8, S_9, \dots, S_k$  are all true.) We must show that  $S_{k+1}$  is true, that is,  $(k+1)$ -cents postage can be achieved with 3¢ and 5¢ stamps. By assumption,  $S_{k-2}$  is true. Thus we can get  $(k-2)$ -cents postage with 3¢ and 5¢ stamps. Now just add one more 3¢ stamp, and we have  $(k-2)+3 = k+1$  cents postage with 3¢ and 5¢ stamps. ■

Our next example proves that  $12 \mid (n^4 - n^2)$  for any  $n \in \mathbb{N}$ . But first, let's see how regular induction is problematic. Regular induction starts by checking  $12 \mid (n^4 - n^2)$  for  $n = 1$ . This reduces to  $12 \mid 0$ , which is true. Next we assume  $12 \mid (k^4 - k^2)$  and try to show that this implies  $12 \mid ((k+1)^4 - (k+1)^2)$ . Now,  $12 \mid (k^4 - k^2)$  means  $k^4 - k^2 = 12a$  for some  $a \in \mathbb{Z}$ . We want to use this to get  $(k+1)^4 - (k+1)^2 = 12b$  for some integer  $b$ . Working it out,

$$\begin{aligned} (k+1)^4 - (k+1)^2 &= (k^4 + 4k^3 + 6k^2 + 4k + 1) - (k^2 + 2k + 1) \\ &= (k^4 - k^2) + 4k^3 + 6k^2 + 6k \\ &= 12a + 4k^3 + 6k^2 + 6k. \end{aligned}$$

At this point we're stuck because we can't factor out a 12.

Let's try strong induction. Say  $S_n$  is the statement  $S_n : 12 \mid (n^4 - n^2)$ . In strong induction, we assume each of  $S_1, S_2, \dots, S_k$  is true, and show that this makes  $S_{k+1}$  true. In particular, if  $S_1$  through  $S_k$  are true, then  $S_{k-5}$  is true, provided  $k - 5 \geq 1$ . We will show  $S_{k-5} \Rightarrow S_{k+1}$  instead of  $S_k \Rightarrow S_{k+1}$ . But as  $k - 5 \geq 1$ , we have  $k \geq 6$ . Thus our basis step must check that  $S_1, S_2, S_3, S_4, S_5, S_6$  are all true. Once this is done,  $S_{k-5} \Rightarrow S_{k+1}$  will imply that any other  $S_k$  is true. For example, if  $k = 6$ , then  $S_{k-5} \Rightarrow S_{k+1}$  is  $S_1 \Rightarrow S_7$ , so  $S_7$  is true. If  $k = 7$ , then  $S_{k-5} \Rightarrow S_{k+1}$  is  $S_2 \Rightarrow S_8$ , so  $S_8$  is true, etc.

**Proposition** If  $n \in \mathbb{N}$ , then  $12 \mid (n^4 - n^2)$ .

*Proof.* We will prove this with strong induction.

- (1) First note that the statement is true for the first six positive integers:

$$\begin{array}{ll} \text{For } n = 1, 12 \text{ divides } 1^4 - 1^2 = 0. & \text{For } n = 4, 12 \text{ divides } 4^4 - 4^2 = 240. \\ \text{For } n = 2, 12 \text{ divides } 2^4 - 2^2 = 12. & \text{For } n = 5, 12 \text{ divides } 5^4 - 5^2 = 600. \\ \text{For } n = 3, 12 \text{ divides } 3^4 - 3^2 = 72. & \text{For } n = 6, 12 \text{ divides } 6^4 - 6^2 = 1260. \end{array}$$

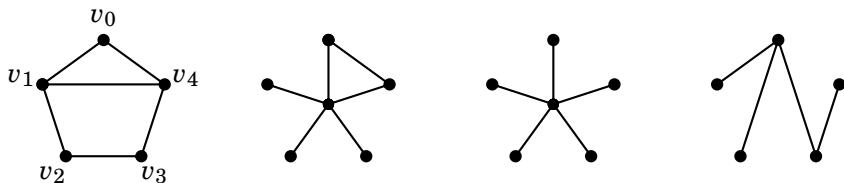
- (2) For  $k \geq 6$ , assume  $12 \mid (m^4 - m^2)$  for  $1 \leq m \leq k$  (i.e.,  $S_1, S_2, \dots, S_k$  are true).

We must show  $S_{k+1}$  is true, that is,  $12 \mid ((k+1)^4 - (k+1)^2)$ . Now,  $S_{k-5}$  being true means  $12 \mid ((k-5)^4 - (k-5)^2)$ . To simplify, put  $\boxed{k-5 = \ell}$  so  $12 \mid (\ell^4 - \ell^2)$ , meaning  $\boxed{\ell^4 - \ell^2 = 12a}$  for  $a \in \mathbb{Z}$ , and  $\boxed{k+1 = \ell+6}$ . Then:

$$\begin{aligned} (k+1)^4 - (k+1)^2 &= (\ell+6)^4 - (\ell+6)^2 \\ &= \ell^4 + 24\ell^3 + 216\ell^2 + 864\ell + 1296 - (\ell^2 + 12\ell + 36) \\ &= (\ell^4 - \ell^2) + 24\ell^3 + 216\ell^2 + 852\ell + 1260 \\ &= 12a + 24\ell^3 + 216\ell^2 + 852\ell + 1260 \\ &= 12(a + 2\ell^3 + 18\ell^2 + 71\ell + 105). \end{aligned}$$

Because  $(a + 2\ell^3 + 18\ell^2 + 71\ell + 105) \in \mathbb{Z}$ , we get  $12 \mid ((k+1)^4 - (k+1)^2)$ . ■

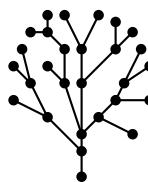
Our next example involves mathematical objects called *graphs*. The word *graph* has two meanings in mathematics. First, there are the graphs of equations and functions from algebra and calculus. But here we will be interested in the second meaning: A **graph** is a configuration consisting of points (called **vertices**) and **edges**, which are lines connecting the vertices. Following are pictures of four graphs. Let's agree that all of our graphs will be in "one piece," that is, you can travel from any vertex of a graph to any other vertex by traversing a route of edges from one vertex to the other.



**Figure 10.1.** Examples of Graphs

A **cycle** in a graph is a sequence of distinct edges in the graph that form a route that ends where it began. For example, the graph on the far left of Figure 10.1 has a cycle that starts at vertex  $v_1$ , then goes to  $v_2$ , then to  $v_3$ , then  $v_4$  and finally back to its starting point  $v_1$ . You can find cycles in both of the graphs on the left, but the two graphs on the right do not have cycles. There is a special name for a graph that has no cycles; it is called a **tree**. Thus the two graphs on the right of Figure 10.1 are trees, but the two graphs on the left are not trees. Note that a single vertex • has no cycle, so it is a tree (with one vertex and zero edges).

The two trees in Figure 10.1 both have one fewer edge than vertex. The tree on the far right has 5 vertices and 4 edges. The one next to it has 6 vertices and 5 edges. Draw any tree (like the one in Figure 10.2). If it has  $n$  vertices, then it will have  $n - 1$  edges. We now prove that this is always true.



**Figure 10.2.** A tree

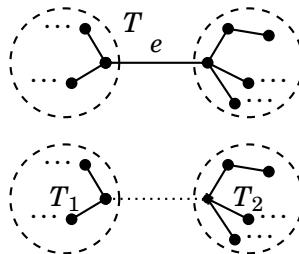
Our proof will use the following observation: If we remove an edge from a tree (but leave its two endpoints), then the tree is cut into two separate graphs, each a tree, and each smaller than the tree we began with.

**Proposition** If a tree has  $n$  vertices, then it has  $n - 1$  edges.

*Proof.* Notice that this theorem asserts that for any  $n \in \mathbb{N}$ , the following statement is true:  $S_n : A \text{ tree with } n \text{ vertices has } n - 1 \text{ edges}$ . We use strong induction to prove this.

- (1) Observe that if a tree has  $n = 1$  vertex then it has no edges. Thus it has  $n - 1 = 0$  edges, so the theorem is true when  $n = 1$ .
- (2) Now take an integer  $k \geq 1$ . We must show  $(S_1 \wedge S_2 \wedge \dots \wedge S_k) \Rightarrow S_{k+1}$ . In words, we must show that if it is true that any tree with  $m$  vertices has  $m - 1$  edges, where  $1 \leq m \leq k$ , then any tree with  $k + 1$  vertices has  $(k + 1) - 1 = k$  edges. We will use direct proof.

Suppose that for each integer  $m$  with  $1 \leq m \leq k$ , any tree with  $m$  vertices has  $m - 1$  edges. Now let  $T$  be a tree with  $k + 1$  vertices. Single out an edge of  $T$  and label it  $e$ , as illustrated below.



Now remove the edge  $e$  from  $T$ , but leave the two endpoints of  $e$ . This leaves two smaller trees that we call  $T_1$  and  $T_2$ . Let's say  $T_1$  has  $x$  vertices and  $T_2$  has  $y$  vertices. As each of these two smaller trees has fewer than  $k + 1$  vertices, our inductive hypothesis guarantees that  $T_1$  has  $x - 1$  edges, and  $T_2$  has  $y - 1$  edges. Think about our original tree  $T$ . It has  $x + y$  vertices. It has  $x - 1$  edges that belong to  $T_1$  and  $y - 1$  edges that belong to  $T_2$ , plus it has the additional edge  $e$  that belongs to neither  $T_1$  nor  $T_2$ . Thus, all together, the number of edges that  $T$  has is  $(x - 1) + (y - 1) + 1 = (x + y) - 1$ . In other words,  $T$  has one fewer edges than it has vertices. Thus it has  $(k + 1) - 1 = k$  edges.

It follows by strong induction that a tree with  $n$  vertices has  $n - 1$  edges. ■

Notice that it was absolutely essential that we used strong induction in the above proof because the two trees  $T_1$  and  $T_2$  will not both have  $k$  vertices. At least one will have fewer than  $k$  vertices. Thus the statement  $S_k$  is not enough to imply  $S_{k+1}$ . We need to use the assumption that  $S_m$  will be true whenever  $m \leq k$ , and strong induction allows us to do this.

### 10.3 Proof by Smallest Counterexample

This section introduces yet another proof technique, called **proof by smallest counterexample**. It is a hybrid of induction and proof by contradiction. It has the nice feature that it leads you straight to a contradiction. It is therefore more “automatic” than the proof by contradiction that was introduced in Chapter 6. Here is the outline:

#### Outline for Proof by Smallest Counterexample

**Proposition** The statements  $S_1, S_2, S_3, S_4, \dots$  are all true.

*Proof.* (Smallest counterexample)

- (1) Check that the first statement  $S_1$  is true.
- (2) For the sake of contradiction, suppose not every  $S_n$  is true.
- (3) Let  $k > 1$  be the smallest integer for which  $S_k$  is **false**.
- (4) Then  $S_{k-1}$  is true and  $S_k$  is false. Use this to get a contradiction. ■

This setup leads you to a point where  $S_{k-1}$  is true and  $S_k$  is false. It is here, where true and false collide, that you will find a contradiction. Let’s do an example.

**Proposition** If  $n \in \mathbb{N}$ , then  $4 \mid (5^n - 1)$ .

*Proof.* We use proof by smallest counterexample. (We will number the steps to match the outline, but that is not usually done in practice.)

- (1) If  $n = 1$ , then the statement is  $4 \mid (5^1 - 1)$ , or  $4 \mid 4$ , which is true.
- (2) For sake of contradiction, suppose it’s not true that  $4 \mid (5^n - 1)$  for all  $n$ .
- (3) Let  $k > 1$  be the smallest integer for which  $4 \nmid (5^k - 1)$ .
- (4) Then  $4 \mid (5^{k-1} - 1)$ , so there is an integer  $a$  for which  $5^{k-1} - 1 = 4a$ . Then

$$\begin{aligned} 5^{k-1} - 1 &= 4a \\ 5(5^{k-1} - 1) &= 5 \cdot 4a \\ 5^k - 5 &= 20a \\ 5^k - 1 &= 20a + 4 \\ 5^k - 1 &= 4(5a + 1). \end{aligned}$$

This means  $4 \mid (5^k - 1)$ , a contradiction, because  $4 \nmid (5^k - 1)$  in Step 3. Thus, we were wrong in Step 2 to assume that it is untrue that  $4 \mid (5^n - 1)$  for every  $n$ . Therefore  $4 \mid (5^n - 1)$  is true for every  $n$ . ■

## 10.4 The Fundamental Theorem of Arithmetic

The **fundamental theorem of arithmetic**, states that any integer greater than 1 has a unique prime factorization. For example, 12 factors into primes as  $12 = 2 \cdot 2 \cdot 3$ , and moreover *any* factorization of 12 into primes uses exactly the primes 2, 2 and 3. Our proof combines the techniques of induction, cases, minimum counterexample and the idea of uniqueness of existence outlined at the end of Section 7.3.

**Theorem 10.1** (Fundamental Theorem of Arithmetic) Any integer  $n > 1$  has a unique prime factorization. “Unique” means that if  $n = p_1 \cdot p_2 \cdot p_3 \cdots p_k$  and  $n = a_1 \cdot a_2 \cdot a_3 \cdots a_\ell$  are two prime factorizations of  $n$ , then  $k = \ell$ , and the primes  $p_i$  and  $a_i$  are the same, except that they may be in different orders.

*Proof.* Suppose  $n > 1$ . We first use strong induction to show that  $n$  has a prime factorization. For the basis step, if  $n = 2$ , it is prime, so it is already its own prime factorization. Let  $n \geq 2$  and assume every integer between 2 and  $n$  (inclusive) has a prime factorization. Consider  $n + 1$ . If it is prime, then it is its own prime factorization. If it is not prime, then it factors as  $n + 1 = ab$  with  $a, b > 1$ . Because  $a$  and  $b$  are both less than  $n + 1$  they have prime factorizations  $a = p_1 \cdot p_2 \cdot p_3 \cdots p_k$  and  $b = p'_1 \cdot p'_2 \cdot p'_3 \cdots p'_{\ell}$ . Then

$$n + 1 = ab = (p_1 \cdot p_2 \cdot p_3 \cdots p_k)(p'_1 \cdot p'_2 \cdot p'_3 \cdots p'_{\ell})$$

is a prime factorization of  $n + 1$ . This completes the proof by strong induction that every integer greater than 1 has a prime factorization.

Next we use proof by smallest counterexample to prove that the prime factorization of any  $n \geq 2$  is unique. If  $n = 2$ , then  $n$  clearly has only one prime factorization, namely itself. Assume for the sake of contradiction that there is an  $n > 2$  that has different prime factorizations  $n = p_1 \cdot p_2 \cdot p_3 \cdots p_k$  and  $n = a_1 \cdot a_2 \cdot a_3 \cdots a_\ell$ . Assume  $n$  is the smallest number with this property. From  $n = p_1 \cdot p_2 \cdot p_3 \cdots p_k$ , we see that  $p_1 \mid n$ , so  $p_1 \mid (a_1 \cdot a_2 \cdot a_3 \cdots a_\ell)$ . By Proposition 10.1 (page 186),  $p_1$  divides one of the primes  $a_i$ . As  $a_i$  is prime, we have  $p_1 = a_i$ . Dividing  $n = p_1 \cdot p_2 \cdot p_3 \cdots p_k = a_1 \cdot a_2 \cdot a_3 \cdots a_\ell$  by  $p_1 = a_i$  yields

$$p_2 \cdot p_3 \cdots p_k = a_1 \cdot a_2 \cdot a_3 \cdots a_{i-1} \cdot a_{i+1} \cdots a_\ell.$$

These two factorizations are different, because the two prime factorizations of  $n$  were different. (Remember: the primes  $p_1$  and  $a_i$  are equal, so the difference appears in the remaining factors, displayed above.) But also the above number  $p_2 \cdot p_3 \cdots p_k$  is smaller than  $n$ , and this contradicts the fact that  $n$  was the smallest number with two different prime factorizations. ■

A word of caution about induction and proof by smallest counterexample: In proofs in other textbooks or in mathematical papers, it often happens that the writer doesn't tell you up front that these techniques are being used. Instead, you will have to read through the proof to glean from context what technique is being used. In fact, the same warning applies to *all* of our proof techniques. If you continue with mathematics, you will gradually gain through experience the ability to analyze a proof and understand exactly what approach is being used when it is not stated explicitly. Frustrations await you, but do not be discouraged by them. Frustration is a natural part of anything that's worth doing.

## 10.5 Fibonacci Numbers

Leonardo Pisano, now known as Fibonacci, was a mathematician born around 1175 in what is now Italy. His most significant work was a book *Liber Abaci*, which is recognized as a catalyst in medieval Europe's slow transition from Roman numbers to the Hindu-Arabic number system. But he is best known today for a number sequence that he described in his book and that bears his name. The **Fibonacci sequence** is

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \dots$$

The numbers that appear in this sequence are called **Fibonacci numbers**. The first two numbers are 1 and 1, and thereafter any entry is the sum of the previous two entries. For example  $3 + 5 = 8$ , and  $5 + 8 = 13$ , etc. We denote the  $n$ th term of this sequence as  $F_n$ . Thus  $F_1 = 1$ ,  $F_2 = 1$ ,  $F_3 = 2$ ,  $F_4 = 3$ ,  $F_7 = 13$  and so on. Notice that the Fibonacci Sequence is entirely determined by the rules  $F_1 = 1$ ,  $F_2 = 1$ , and  $F_n = F_{n-1} + F_{n-2}$ .

We introduce Fibonacci's sequence here partly because it is something everyone should know about, but also because it is a great source of induction problems. This sequence, which appears with surprising frequency in nature, is filled with mysterious patterns and hidden structures. Some of these structures will be revealed in the examples and exercises.

We emphasize that the condition  $F_n = F_{n-1} + F_{n-2}$  (or equivalently  $F_{n+1} = F_n + F_{n-1}$ ) is the perfect setup for induction. It suggests that we can determine something about  $F_n$  by looking at earlier terms of the sequence. In using induction to prove something about the Fibonacci sequence, you should expect to use the equation  $F_n = F_{n-1} + F_{n-2}$  somewhere.

For our first example we will prove that  $F_{n+1}^2 - F_{n+1}F_n - F_n^2 = (-1)^n$  for any natural number  $n$ . For example, if  $n = 5$  we have  $F_6^2 - F_6F_5 - F_5^2 = 8^2 - 8 \cdot 5 - 5^2 = 64 - 40 - 25 = -1 = (-1)^5$ .

**Proposition** The Fibonacci sequence obeys  $F_{n+1}^2 - F_{n+1}F_n - F_n^2 = (-1)^n$ .

*Proof.* We will prove this with mathematical induction.

(1) If  $n = 1$  we have  $F_{n+1}^2 - F_{n+1}F_n - F_n^2 = F_2^2 - F_2F_1 - F_1^2 = 1^2 - 1 \cdot 1 - 1^2 = -1 = (-1)^1 = (-1)^n$ , so indeed  $F_{n+1}^2 - F_{n+1}F_n - F_n^2 = (-1)^n$  is true when  $n = 1$ .

(2) Let  $k \in \mathbb{N}$ . Using direct proof, we will show  $F_{k+1}^2 - F_{k+1}F_k - F_k^2 = (-1)^k$  implies  $F_{k+2}^2 - F_{k+2}F_{k+1} - F_{k+1}^2 = (-1)^{k+1}$ . Say  $F_{k+1}^2 - F_{k+1}F_k - F_k^2 = (-1)^k$ . Next we work out  $F_{k+2}^2 - F_{k+2}F_{k+1} - F_{k+1}^2$  and show that it equals  $(-1)^{k+1}$ . In doing this we will use the fact  $F_{k+2} = F_{k+1} + F_k$ . Observe that

$$\begin{aligned} F_{k+2}^2 - F_{k+2}F_{k+1} - F_{k+1}^2 &= (F_{k+1} + F_k)^2 - (F_{k+1} + F_k)F_{k+1} - F_{k+1}^2 \\ &= F_{k+1}^2 + 2F_{k+1}F_k + F_k^2 - F_{k+1}^2 - F_kF_{k+1} - F_{k+1}^2 \\ &= -F_{k+1}^2 + F_{k+1}F_k + F_k^2 \\ &= -(F_{k+1}^2 - F_{k+1}F_k - F_k^2) \\ &= -(-1)^k && \text{(inductive hypothesis)} \\ &= (-1)^1(-1)^k \\ &= (-1)^{k+1}. \end{aligned}$$

Therefore  $F_{k+2}^2 - F_{k+2}F_{k+1} - F_{k+1}^2 = (-1)^{k+1}$ .

It follows by induction that  $F_{n+1}^2 - F_{n+1}F_n - F_n^2 = (-1)^n$  for every  $n \in \mathbb{N}$ . ■

Let's pause for a moment and think about what the result we just proved means. Dividing both sides of  $F_{n+1}^2 - F_{n+1}F_n - F_n^2 = (-1)^n$  by  $F_n^2$  gives

$$\left(\frac{F_{n+1}}{F_n}\right)^2 - \frac{F_{n+1}}{F_n} - 1 = \frac{(-1)^n}{F_n^2}.$$

For large values of  $n$ , the right-hand side is very close to zero, and the left-hand side is  $F_{n+1}/F_n$  plugged into the polynomial  $x^2 - x - 1$ . Thus, as  $n$  increases, the ratio  $F_{n+1}/F_n$  approaches a root of  $x^2 - x - 1 = 0$ . By the quadratic formula, the roots of  $x^2 - x - 1$  are  $\frac{1 \pm \sqrt{5}}{2}$ . As  $F_{n+1}/F_n > 1$ , this ratio must be approaching the *positive* root  $\frac{1 + \sqrt{5}}{2}$ . Therefore

$$\lim_{n \rightarrow \infty} \frac{F_{n+1}}{F_n} = \frac{1 + \sqrt{5}}{2}. \quad (10.1)$$

For a quick spot check, note that  $F_{13}/F_{12} \approx 1.618025$ , while  $\frac{1 + \sqrt{5}}{2} \approx 1.618033$ . Even for the small value  $n = 12$ , the numbers match to four decimal places.

The number  $\phi = \frac{1+\sqrt{5}}{2}$  is sometimes called the **golden ratio**, and there has been much speculation about its occurrence in nature as well as in classical art and architecture. One theory holds that the Parthenon and the Great Pyramids of Egypt were designed in accordance with this number.

But we are here concerned with things that can be proved. We close by observing how the Fibonacci sequence in many ways resembles a geometric sequence. Recall that a **geometric sequence** with first term  $a$  and common ratio  $r$  has the form

$$a, ar, ar^2, ar^3, ar^4, ar^5, ar^6, ar^7, ar^8, \dots$$

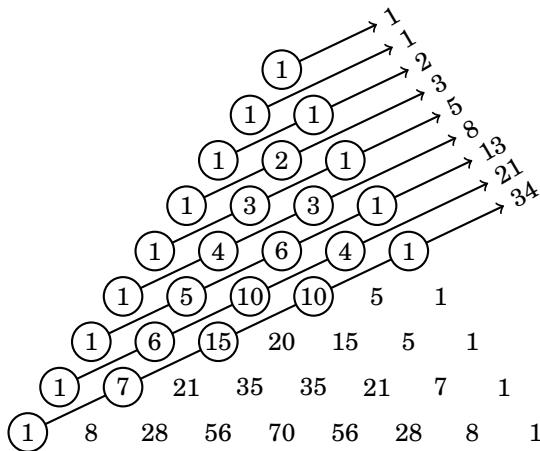
where any term is obtained by multiplying the previous term by  $r$ . In general its  $n$ th term is  $G_n = ar^n$ , and  $G_{n+1}/G_n = r$ . Equation (10.1) tells us that  $F_{n+1}/F_n \approx \phi$ . Thus even though it is not a geometric sequence, the Fibonacci sequence tends to behave like a geometric sequence with common ratio  $\phi$ , and the further “out” you go, the higher the resemblance.

## Exercises for Chapter 10

Prove the following statements with either induction, strong induction or proof by smallest counterexample.

1. Prove that  $1 + 2 + 3 + 4 + \dots + n = \frac{n^2 + n}{2}$  for every positive integer  $n$ .
2. Prove that  $1^2 + 2^2 + 3^2 + 4^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$  for every positive integer  $n$ .
3. Prove that  $1^3 + 2^3 + 3^3 + 4^3 + \dots + n^3 = \frac{n^2(n+1)^2}{4}$  for every positive integer  $n$ .
4. If  $n \in \mathbb{N}$ , then  $1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + 4 \cdot 5 + \dots + n(n+1) = \frac{n(n+1)(n+2)}{3}$ .
5. If  $n \in \mathbb{N}$ , then  $2^1 + 2^2 + 2^3 + \dots + 2^n = 2^{n+1} - 2$ .
6. Prove that  $\sum_{i=1}^n (8i - 5) = 4n^2 - n$  for every positive integer  $n$ .
7. If  $n \in \mathbb{N}$ , then  $1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 6 + \dots + n(n+2) = \frac{n(n+1)(2n+7)}{6}$ .
8. If  $n \in \mathbb{N}$ , then  $\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \dots + \frac{n}{(n+1)!} = 1 - \frac{1}{(n+1)!}$ .
9. Prove that  $24 | (5^{2n} - 1)$  for every integer  $n \geq 0$ .
10. Prove that  $3 | (5^{2n} - 1)$  for every integer  $n \geq 0$ .
11. Prove that  $3 | (n^3 + 5n + 6)$  for every integer  $n \geq 0$ .
12. Prove that  $9 | (4^{3n} + 8)$  for every integer  $n \geq 0$ .
13. Prove that  $6 | (n^3 - n)$  for every integer  $n \geq 0$ .

14. Suppose  $a \in \mathbb{Z}$ . Prove that  $5 | 2^n a$  implies  $5 | a$  for any  $n \in \mathbb{N}$ .
15. If  $n \in \mathbb{N}$ , then  $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \frac{1}{4 \cdot 5} + \cdots + \frac{1}{n(n+1)} = 1 - \frac{1}{n+1}$ .
16. Prove that  $2^n + 1 \leq 3^n$  for every positive integer  $n$ .
17. Suppose  $A_1, A_2, \dots, A_n$  are sets in some universal set  $U$ , and  $n \geq 2$ . Prove that  $\overline{A_1 \cap A_2 \cap \cdots \cap A_n} = \overline{A_1} \cup \overline{A_2} \cup \cdots \cup \overline{A_n}$ .
18. Suppose  $A_1, A_2, \dots, A_n$  are sets in some universal set  $U$ , and  $n \geq 2$ . Prove that  $\overline{A_1 \cup A_2 \cup \cdots \cup A_n} = \overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}$ .
19. Prove that  $\frac{1}{1} + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$  for every  $n \in \mathbb{N}$ .
20. Prove that  $(1+2+3+\cdots+n)^2 = 1^3+2^3+3^3+\cdots+n^3$  for every  $n \in \mathbb{N}$ .
21. If  $n \in \mathbb{N}$ , then  $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots + \frac{1}{2^n-1} + \frac{1}{2^n} \geq 1 + \frac{n}{2}$ .  
 (Note: This problem asserts that the sum of the first  $2^n$  terms of the harmonic series is at least  $1 + n/2$ . It thus implies that the harmonic series diverges.)
22. If  $n \in \mathbb{N}$ , then  $\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{8}\right)\cdots\left(1 - \frac{1}{2^n}\right) \geq \frac{1}{4} + \frac{1}{2^{n+1}}$ .
23. Use mathematical induction to prove the binomial theorem (Theorem 3.1 on page 92). You may find that you need Equation (3.3) on page 90.
24. Prove that  $\sum_{k=1}^n k \binom{n}{k} = n2^{n-1}$  for each natural number  $n$ .
25. Concerning the Fibonacci sequence, prove that  $F_1 + F_2 + F_3 + F_4 + \cdots + F_n = F_{n+2} - 1$ .
26. Concerning the Fibonacci sequence, prove that  $\sum_{k=1}^n F_k^2 = F_n F_{n+1}$ .
27. Concerning the Fibonacci sequence, prove that  $F_1 + F_3 + F_5 + F_7 + \cdots + F_{2n-1} = F_{2n}$ .
28. Concerning the Fibonacci sequence, prove that  $F_2 + F_4 + F_6 + F_8 + \cdots + F_{2n} = F_{2n+1} - 1$ .
29. The indicated diagonals of Pascal's triangle sum to Fibonacci numbers. Prove that this pattern continues forever.



- 30.** Here  $F_n$  is the  $n$ th Fibonacci number. Prove that

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

- 31.** Prove that  $\sum_{k=0}^n \binom{k}{r} = \binom{n+1}{r+1}$ , where  $1 \leq r \leq n$ .
- 32.** Prove that the number of  $n$ -digit binary numbers that have no consecutive 1's is the Fibonacci number  $F_{n+2}$ . For example, for  $n = 2$  there are three such numbers (00, 01, and 10), and  $3 = F_{2+2} = F_4$ . Also, for  $n = 3$  there are five such numbers (000, 001, 010, 100, 101), and  $5 = F_{3+2} = F_5$ .
- 33.** Suppose  $n$  (infinitely long) straight lines lie on a plane in such a way that no two of the lines are parallel, and no three of the lines intersect at a single point. Show that this arrangement divides the plane into  $\frac{n^2+n+2}{2}$  regions.
- 34.** Prove that  $3^1 + 3^2 + 3^3 + 3^4 + \cdots + 3^n = \frac{3^{n+1} - 3}{2}$  for every  $n \in \mathbb{N}$ .
- 35.** Prove that if  $n, k \in \mathbb{N}$ , and  $n$  is even and  $k$  is odd, then  $\binom{n}{k}$  is even.
- 36.** Prove that if  $n = 2^k - 1$  for some  $k \in \mathbb{N}$ , then every entry in the  $n$ th row of Pascal's triangle is odd.
- 37.** Prove that if  $m, n \in \mathbb{N}$ , then  $\sum_{k=0}^n k \binom{m+k}{m} = n \binom{m+n+1}{m+1} - \binom{m+n+1}{m+2}$ .
- 38.** Prove that  $\sum_{k=0}^p \binom{m}{k} \binom{n}{p-k} = \binom{m+n}{p}$  for non-negative integers  $m, n$  and  $p$ .  
(This equation is from Exercise 7 in Section 3.10. There we were asked to prove it by combinatorial proof. Here we are asked to prove it with induction.)
- 39.** Prove that  $\sum_{k=0}^m \binom{m}{k} \binom{n}{p+k} = \binom{m+n}{m+p}$  for non-negative integers  $m, n$  and  $p$ .  
(This equation is from Exercise 8 in Section 3.10. There we were asked to prove it by combinatorial proof. Here we are asked to prove it with induction.)
- 40.** Use Exercise 38 above to prove that if  $n, k \in \mathbb{N}$ , then  $\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}$ .  
(Note that this equality was also proved by combinatorial proof in Section 3.10.)
- 41.** If  $n$  and  $k$  are non-negative integers, then  $\binom{n+0}{0} + \binom{n+1}{1} + \binom{n+2}{2} + \cdots + \binom{n+k}{k} = \binom{n+k+1}{k}$ .



## *Part IV*

---

### *Relations, Functions and Cardinality*

---



# CHAPTER 11

---

## Relations

---

In mathematics there are endless ways that two entities can be related to each other. Consider the following mathematical statements.

$$\begin{array}{ccccccc} 5 < 10 & 5 \leq 5 & 6 = \frac{30}{5} & 5 | 80 & 7 > 4 & x \neq y & 8 \nmid 3 \\ a \equiv b \pmod{n} & 6 \in \mathbb{Z} & X \subseteq Y & \pi \approx 3.14 & 0 \geq -1 & \sqrt{2} \notin \mathbb{Z} & \mathbb{Z} \not\subseteq \mathbb{N} \end{array}$$

In each case two entities appear on either side of a symbol, and we interpret the symbol as expressing some relationship between the two entities. Symbols such as  $<$ ,  $\leq$ ,  $=$ ,  $\mid$ ,  $\nmid$ ,  $\geq$ ,  $>$ ,  $\in$  and  $\subseteq$ , etc., are called *relations* because they convey relationships among things.

Relations are significant. In fact, you would have to admit that there would be precious little left of mathematics if we took away all the relations. Therefore it is important to have a firm understanding of them, and this chapter is intended to develop that understanding.

Rather than focusing on each relation individually (an impossible task anyway since there are infinitely many different relations), we will develop a general theory that encompasses *all* relations. Understanding this general theory will give us the conceptual framework and language needed to understand and discuss any specific relation.

### 11.1 Relations

Before stating the theoretical definition of a relation, let's look at a motivational example. This example will lead naturally to our definition.

Consider the set  $A = \{1, 2, 3, 4, 5\}$ . (There's nothing special about this particular set; any set of numbers would do for this example.) Elements of  $A$  can be compared to each other by the symbol " $<$ ". For example,  $1 < 4$ ,  $2 < 3$ ,  $2 < 4$ , and so on. You have no trouble understanding this because the notion of numeric order is so ingrained. But imagine you had to explain it to an idiot savant, one with an obsession for detail but absolutely no understanding of the meaning of (or relationships between) integers. You might consider writing down for your student the following set:

$$R = \{(1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5)\}.$$

The set  $R$  encodes the meaning of the  $<$  relation for elements in  $A$ . An ordered pair  $(a, b)$  appears in the set if and only if  $a < b$ . If asked whether or not it is true that  $3 < 4$ , your student could look through  $R$  until he found the ordered pair  $(3, 4)$ ; then he would know  $3 < 4$  is true. If asked about  $5 < 2$ , he would see that  $(5, 2)$  does not appear in  $R$ , so  $5 \not< 2$ . The set  $R$ , which is a subset of  $A \times A$ , completely describes the relation  $<$  for  $A$ .

It seems simple-minded at first, but this is exactly the idea we will use for our main definition. This definition is general enough to describe not just the relation  $<$  for the set  $A = \{1, 2, 3, 4, 5\}$ , but *any* relation for *any* set  $A$ .

**Definition 11.1** A **relation** on a set  $A$  is a subset  $R \subseteq A \times A$ . We often abbreviate the statement  $(x, y) \in R$  as  $xRy$ . The statement  $(x, y) \notin R$  is abbreviated as  $x \not R y$ .

Notice that a relation is a set, so we can use what we know about sets to understand and explore relations. But before getting deeper into the theory of relations, let's look at some examples of Definition 11.1.

**Example 11.1** Let  $A = \{1, 2, 3, 4\}$ , and consider the following set:

$$R = \{(1, 1), (2, 1), (2, 2), (3, 3), (3, 2), (3, 1), (4, 4), (4, 3), (4, 2), (4, 1)\} \subseteq A \times A.$$

The set  $R$  is a relation on  $A$ , by Definition 11.1. Since  $(1, 1) \in R$ , we have  $1R1$ . Similarly  $2R1$  and  $2R2$ , and so on. However, notice that (for example)  $(3, 4) \notin R$ , so  $3 \not R 4$ . Observe that  $R$  is the familiar relation  $\geq$  for the set  $A$ .

Chapter 1 proclaimed that all of mathematics can be described with sets. Just look at how successful this program has been! The greater-than-or-equal-to relation is now a set  $R$ . (We might even express this in the rather cryptic form  $\geq = R$ .)

**Example 11.2** Let  $A = \{1, 2, 3, 4\}$ , and consider the following set:

$$S = \{(1, 1), (1, 3), (3, 1), (3, 3), (2, 2), (2, 4), (4, 2), (4, 4)\} \subseteq A \times A.$$

Here we have  $1S1$ ,  $1S3$ ,  $4S2$ , etc., but  $3 \not S 4$  and  $2 \not S 1$ . What does  $S$  mean? Think of it as meaning “*has the same parity as*.” Thus  $1S1$  reads “*1 has the same parity as 1*,” and  $4S2$  reads “*4 has the same parity as 2*.”

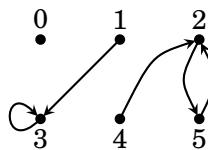
**Example 11.3** Consider relations  $R$  and  $S$  of the previous two examples. Note that  $R \cap S = \{(1, 1), (2, 2), (3, 3), (3, 1), (4, 4), (4, 2)\} \subseteq A \times A$  is a relation on  $A$ . The expression  $x(R \cap S)y$  means “*x  $\geq$  y, and x has the same parity as y*.”

**Example 11.4** Let  $B = \{0, 1, 2, 3, 4, 5\}$ , and consider the following set:

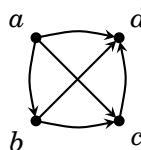
$$U = \{(1, 3), (3, 3), (5, 2), (2, 5), (4, 2)\} \subseteq B \times B.$$

Then  $U$  is a relation on  $B$  because  $U \subseteq B \times B$ . You may be hard-pressed to invent any “meaning” for this particular relation. A relation does not have to have any meaning. Any random subset of  $B \times B$  is a relation on  $B$ , whether or not it describes anything familiar.

Some relations can be described with pictures. For example, we can depict the above relation  $U$  on  $B$  by drawing points labeled by elements of  $B$ . The statement  $(x, y) \in U$  is then represented by an arrow pointing from  $x$  to  $y$ , a graphic symbol meaning “ $x$  relates to  $y$ .” Here is a picture of  $U$ :



The next picture illustrates the relation  $R$  on the set  $A = \{a, b, c, d\}$ , where  $xRy$  means  $x$  comes before  $y$  in the alphabet. According to Definition 11.1, as a set this relation is  $R = \{(a, b), (a, c), (a, d), (b, c), (b, d), (c, d)\}$ . You may feel that the picture conveys the relation better than the set does. They are two different ways of expressing the same thing. In some instances pictures are more convenient than sets for discussing relations.



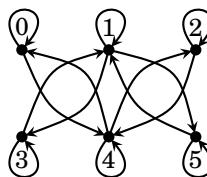
Although such diagrams can help us visualize relations, they do have their limitations. If  $A$  and  $R$  were infinite, then the diagram would be impossible to draw, but the set  $R$  might be easily expressed in set-builder notation. Here are some examples.

**Example 11.5** Consider the set  $R = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x - y \in \mathbb{N}\} \subseteq \mathbb{Z} \times \mathbb{Z}$ . This is the  $>$  relation on the set  $A = \mathbb{Z}$ . It is infinite because there are infinitely many ways to have  $x > y$  where  $x$  and  $y$  are integers.

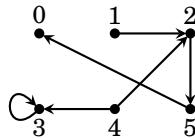
**Example 11.6** The set  $R = \{(x, x) : x \in \mathbb{R}\} \subseteq \mathbb{R} \times \mathbb{R}$  is the relation  $=$  on the set  $\mathbb{R}$ , because  $xRy$  means the same thing as  $x = y$ . Thus  $R$  is a set that expresses the notion of equality of real numbers.

**Exercises for Section 11.1**

- Let  $A = \{0, 1, 2, 3, 4, 5\}$ . Write out the relation  $R$  that expresses  $>$  on  $A$ . Then illustrate it with a diagram.
- Let  $A = \{1, 2, 3, 4, 5, 6\}$ . Write out the relation  $R$  that expresses  $|$  (divides) on  $A$ . Then illustrate it with a diagram.
- Let  $A = \{0, 1, 2, 3, 4, 5\}$ . Write out the relation  $R$  that expresses  $\geq$  on  $A$ . Then illustrate it with a diagram.
- Here is a diagram for a relation  $R$  on a set  $A$ . Write the sets  $A$  and  $R$ .

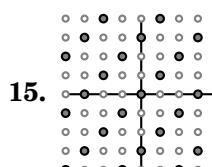
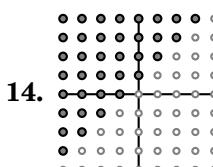
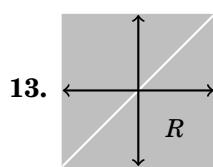
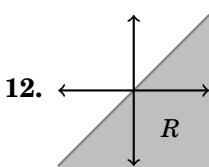


- Here is a diagram for a relation  $R$  on a set  $A$ . Write the sets  $A$  and  $R$ .



- Congruence modulo 5 is a relation on the set  $A = \mathbb{Z}$ . In this relation  $xRy$  means  $x \equiv y \pmod{5}$ . Write out the set  $R$  in set-builder notation.
- Write the relation  $<$  on the set  $A = \mathbb{Z}$  as a subset  $R$  of  $\mathbb{Z} \times \mathbb{Z}$ . This is an infinite set, so you will have to use set-builder notation.
- Let  $A = \{1, 2, 3, 4, 5, 6\}$ . Observe that  $\emptyset \subseteq A \times A$ , so  $R = \emptyset$  is a relation on  $A$ . Draw a diagram for this relation.
- Let  $A = \{1, 2, 3, 4, 5, 6\}$ . How many different relations are there on the set  $A$ ?
- Consider the subset  $R = (\mathbb{R} \times \mathbb{R}) - \{(x, x) : x \in \mathbb{R}\} \subseteq \mathbb{R} \times \mathbb{R}$ . What familiar relation on  $\mathbb{R}$  is this? Explain.
- Given a finite set  $A$ , how many different relations are there on  $A$ ?

In the following exercises, subsets  $R$  of  $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$  or  $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$  are indicated by gray shading. In each case,  $R$  is a familiar relation on  $\mathbb{R}$  or  $\mathbb{Z}$ . State it.



## 11.2 Properties of Relations

A relational expression  $xRy$  is an *open sentence*; it is either true or false. For example,  $5 < 10$  is true, and  $10 < 5$  is false. (Thus an operation like  $+$  is not a relation, because, for instance,  $5+10$  has a numeric value, not a T/F value.) Since relational expressions have T/F values, we can combine them with logical operators; for example,  $xRy \Rightarrow yRx$  is a statement or open sentence whose truth or falsity may depend on  $x$  and  $y$ .

With this in mind, note that some relations have properties that others don't have. For example, the relation  $\leq$  on  $\mathbb{Z}$  satisfies  $x \leq x$  for every  $x \in \mathbb{Z}$ . But this is not so for  $<$  because  $x < x$  is never true. The next definition lays out three particularly significant properties that relations may have.

**Definition 11.2** Suppose  $R$  is a relation on a set  $A$ .

1. Relation  $R$  is **reflexive** if  $xRx$  for every  $x \in A$ .  
That is,  $R$  is reflexive if  $\forall x \in A, xRx$ .
2. Relation  $R$  is **symmetric** if  $xRy$  implies  $yRx$  for all  $x, y \in A$   
That is,  $R$  is symmetric if  $\forall x, y \in A, xRy \Rightarrow yRx$ .
3. Relation  $R$  is **transitive** if whenever  $xRy$  and  $yRz$ , then also  $xRz$ .  
That is,  $R$  is transitive if  $\forall x, y, z \in A, ((xRy) \wedge (yRz)) \Rightarrow xRz$ .

To illustrate this, let's consider the set  $A = \mathbb{Z}$ . Examples of reflexive relations on  $\mathbb{Z}$  include  $\leq$ ,  $=$ , and  $|$ , because  $x \leq x$ ,  $x = x$  and  $x|x$  are all true for any  $x \in \mathbb{Z}$ . On the other hand,  $>$ ,  $<$ ,  $\neq$  and  $\nmid$  are not reflexive for none of the statements  $x < x$ ,  $x > x$ ,  $x \neq x$  and  $x \nmid x$  is ever true.

The relation  $\neq$  is symmetric, for if  $x \neq y$ , then surely  $y \neq x$  also. Also, the relation  $=$  is symmetric because  $x = y$  always implies  $y = x$ .

The relation  $\leq$  is not symmetric, as  $x \leq y$  does not necessarily imply  $y \leq x$ . For instance  $5 \leq 6$  is true, but  $6 \leq 5$  is false. Notice  $(x \leq y) \Rightarrow (y \leq x)$  is true for some  $x$  and  $y$  (for example, it is true when  $x = 2$  and  $y = 2$ ), but still  $\leq$  is not symmetric because it is not the case that  $(x \leq y) \Rightarrow (y \leq x)$  is true for all integers  $x$  and  $y$ .

The relation  $\leq$  is transitive because whenever  $x \leq y$  and  $y \leq z$ , it also is true that  $x \leq z$ . Likewise  $<, \geq, >$  and  $=$  are all transitive. Examine the following table and be sure you understand why it is labeled as it is.

Relations on $\mathbb{Z}$ :	$<$	$\leq$	$=$	$ $	$\nmid$	$\neq$
Reflexive	no	yes	yes	yes	no	no
Symmetric	no	no	yes	no	no	yes
Transitive	yes	yes	yes	yes	no	no

**Example 11.7** Here  $A = \{b, c, d, e\}$ , and  $R$  is the following relation on  $A$ :  
 $R = \{(b, b), (b, c), (c, b), (c, c), (d, d), (b, d), (d, b), (c, d), (d, c)\}$ .

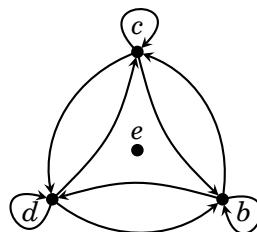
This relation is **not** reflexive, for although  $bRb$ ,  $cRc$  and  $dRd$ , it is **not** true that  $eRe$ . For a relation to be reflexive,  $xRx$  must be true for all  $x \in A$ .

The relation  $R$  is **symmetric**, because whenever we have  $xRy$ , it follows that  $yRx$  too. Observe that  $bRc$  and  $cRb$ ;  $bRd$  and  $dRb$ ;  $dRc$  and  $cRd$ . Take away the ordered pair  $(c, b)$  from  $R$ , and  $R$  is no longer symmetric.

The relation  $R$  is transitive, but it takes some work to check it. We must check that the statement  $(xRy \wedge yRz) \Rightarrow xRz$  is true for all  $x, y, z \in A$ . For example, taking  $x = b$ ,  $y = c$  and  $z = d$ , we have  $(bRc \wedge cRd) \Rightarrow bRd$ , which is the true statement  $(T \wedge T) \Rightarrow T$ . Likewise,  $(bRd \wedge dRc) \Rightarrow bRc$  is the true statement  $(T \wedge T) \Rightarrow T$ . Take note that if  $x = b$ ,  $y = e$  and  $z = c$ , then  $(bRe \wedge eRc) \Rightarrow bRc$  becomes  $(F \wedge F) \Rightarrow T$ , which is *still* true. It's not much fun, but going through all the combinations, you can verify that  $(xRy \wedge yRz) \Rightarrow xRz$  is true for all choices  $x, y, z \in A$ . (Try at least a few of them.)

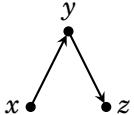
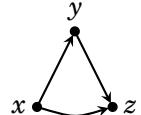
The relation  $R$  from Example 11.7 has a meaning. You can think of  $xRy$  as meaning that  $x$  and  $y$  are both consonants. Thus  $bRc$  because  $b$  and  $c$  are both consonants; but  $bRe$  because it's not true that  $b$  and  $e$  are both consonants. Once we look at it this way, it's immediately clear that  $R$  has to be transitive. If  $x$  and  $y$  are both consonants and  $y$  and  $z$  are both consonants, then surely  $x$  and  $z$  are both consonants. This illustrates a point that we will see again later in this section: Knowing the meaning of a relation can help us understand it and prove things about it.

Here is a picture of  $R$ . Notice that we can immediately spot several properties of  $R$  that may not have been so clear from its set description. For instance, we see that  $R$  is not reflexive because it lacks a loop at  $e$ , hence  $eR\neq e$ .



**Figure 11.1.** The relation  $R$  from Example 11.7

In what follows, we summarize how to spot the various properties of a relation from its diagram. Compare these with Figure 11.1.

1. A relation is <b>reflexive</b> if for each point $x$ ...		...there is a loop at $x$ : 
2. A relation is <b>symmetric</b> if whenever there is an arrow from $x$ to $y$ ...		...there is also an arrow from $y$ back to $x$ : 
3. A relation is <b>transitive</b> if whenever there are arrows from $x$ to $y$ and $y$ to $z$ ...		...there is also an arrow from $x$ to $z$ : 
(If $x = z$ , this means that if there are arrows from $x$ to $y$ and from $y$ to $x$ ...		...there is also a loop from $x$ back to $x$ .) 

Consider the bottom diagram in Box 3, above. The transitive property demands  $(xRy \wedge yRx) \Rightarrow xRx$ . Thus, if  $xRy$  and  $yRx$  in a transitive relation, then also  $xRx$ , so there is a loop at  $x$ . In this case  $(yRx \wedge xRy) \Rightarrow yRy$ , so there will be a loop at  $y$  too.

Although these visual aids can be illuminating, their use is limited because many relations are too large and complex to be adequately described as diagrams. For example, it would be impossible to draw a diagram for the relation  $\equiv (\text{mod } n)$ , where  $n \in \mathbb{N}$ . Such a relation would best be explained in a more theoretical (and less visual) way.

We next prove that  $\equiv (\text{mod } n)$  is reflexive, symmetric and transitive. Obviously we will not glean this from a drawing. Instead we will prove it from the properties of  $\equiv (\text{mod } n)$  and Definition 11.2. Pay attention to this example. It illustrates how to **prove** things about relations.

**Example 11.8** Prove the following proposition.

**Proposition** Let  $n \in \mathbb{N}$ . The relation  $\equiv (\text{mod } n)$  on the set  $\mathbb{Z}$  is reflexive, symmetric and transitive.

*Proof.* First we will show that  $\equiv (\text{mod } n)$  is reflexive. Take any integer  $x \in \mathbb{Z}$ , and observe that  $n \mid 0$ , so  $n \mid (x - x)$ . By definition of congruence modulo  $n$ , we have  $x \equiv x (\text{mod } n)$ . This shows  $x \equiv x (\text{mod } n)$  for every  $x \in \mathbb{Z}$ , so  $\equiv (\text{mod } n)$  is reflexive.

Next, we will show that  $\equiv (\text{mod } n)$  is symmetric. For this, we must show that for all  $x, y \in \mathbb{Z}$ , the condition  $x \equiv y (\text{mod } n)$  implies that  $y \equiv x (\text{mod } n)$ . We use direct proof. Suppose  $x \equiv y (\text{mod } n)$ . Thus  $n \mid (x - y)$  by definition of congruence modulo  $n$ . Then  $x - y = na$  for some  $a \in \mathbb{Z}$  by definition of divisibility. Multiplying both sides by  $-1$  gives  $y - x = n(-a)$ . Therefore  $n \mid (y - x)$ , and this means  $y \equiv x (\text{mod } n)$ . We've shown that  $x \equiv y (\text{mod } n)$  implies that  $y \equiv x (\text{mod } n)$ , and this means  $\equiv (\text{mod } n)$  is symmetric.

Finally we will show that  $\equiv (\text{mod } n)$  is transitive. For this we must show that if  $x \equiv y (\text{mod } n)$  and  $y \equiv z (\text{mod } n)$ , then  $x \equiv z (\text{mod } n)$ . Again we use direct proof. Suppose  $x \equiv y (\text{mod } n)$  and  $y \equiv z (\text{mod } n)$ . This means  $n \mid (x - y)$  and  $n \mid (y - z)$ . Therefore there are integers  $a$  and  $b$  for which  $x - y = na$  and  $y - z = nb$ . Adding these two equations, we obtain  $x - z = na + nb$ . Consequently,  $x - z = n(a + b)$ , so  $n \mid (x - z)$ , hence  $x \equiv z (\text{mod } n)$ . This completes the proof that  $\equiv (\text{mod } n)$  is transitive.

The past three paragraphs have shown that the relation  $\equiv (\text{mod } n)$  is reflexive, symmetric and transitive, so the proof is complete. ■

As you continue with mathematics the reflexive, symmetric and transitive properties will take on special significance in a variety of settings. In preparation for this, the next section explores further consequences of these properties. But first work some of the following exercises.

### Exercises for Section 11.2

- Consider the relation  $R = \{(a, a), (b, b), (c, c), (d, d), (a, b), (b, a)\}$  on set  $A = \{a, b, c, d\}$ . Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why.
- Consider the relation  $R = \{(a, b), (a, c), (c, c), (b, b), (c, b), (b, c)\}$  on the set  $A = \{a, b, c\}$ . Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why.
- Consider the relation  $R = \{(a, b), (a, c), (c, b), (b, c)\}$  on the set  $A = \{a, b, c\}$ . Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why.

4. Let  $A = \{a, b, c, d\}$ . Suppose  $R$  is the relation

$$\begin{aligned} R = & \{(a,a), (b,b), (c,c), (d,d), (a,b), (b,a), (a,c), (c,a), \\ & (a,d), (d,a), (b,c), (c,b), (b,d), (d,b), (c,d), (d,c)\}. \end{aligned}$$

Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why.

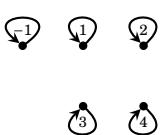
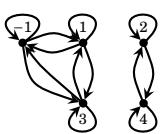
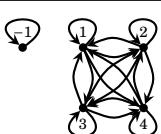
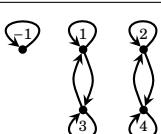
5. Consider the relation  $R = \{(0,0), (\sqrt{2},0), (0,\sqrt{2}), (\sqrt{2},\sqrt{2})\}$  on  $\mathbb{R}$ . Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why.
6. Consider the relation  $R = \{(x,x) : x \in \mathbb{Z}\}$  on  $\mathbb{Z}$ . Is this  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why. What familiar relation is this?
7. There are 16 possible different relations  $R$  on the set  $A = \{a, b\}$ . Describe all of them. (A picture for each one will suffice, but don't forget to label the nodes.) Which ones are reflexive? Symmetric? Transitive?
8. Define a relation on  $\mathbb{Z}$  as  $xRy$  if  $|x - y| < 1$ . Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why. What familiar relation is this?
9. Define a relation on  $\mathbb{Z}$  by declaring  $xRy$  if and only if  $x$  and  $y$  have the same parity. Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why. What familiar relation is this?
10. Suppose  $A \neq \emptyset$ . Since  $\emptyset \subseteq A \times A$ , the set  $R = \emptyset$  is a relation on  $A$ . Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why.
11. Let  $A = \{a, b, c, d\}$  and  $R = \{(a,a), (b,b), (c,c), (d,d)\}$ . Is  $R$  reflexive? Symmetric? Transitive? If a property does not hold, say why.
12. Prove that the relation  $|$  (divides) on the set  $\mathbb{Z}$  is reflexive and transitive. (Use Example 11.8 as a guide if you are unsure of how to proceed.)
13. Consider the relation  $R = \{(x,y) \in \mathbb{R} \times \mathbb{R} : x - y \in \mathbb{Z}\}$  on  $\mathbb{R}$ . Prove that this relation is reflexive, symmetric and transitive.
14. Suppose  $R$  is a symmetric and transitive relation on a set  $A$ , and there is an element  $a \in A$  for which  $aRx$  for every  $x \in A$ . Prove that  $R$  is reflexive.
15. Prove or disprove: If a relation is symmetric and transitive, then it is also reflexive.
16. Define a relation  $R$  on  $\mathbb{Z}$  by declaring that  $xRy$  if and only if  $x^2 \equiv y^2 \pmod{4}$ . Prove that  $R$  is reflexive, symmetric and transitive.
17. Modifying the above Exercise 8 (above) slightly, define a relation  $\sim$  on  $\mathbb{Z}$  as  $x \sim y$  if and only if  $|x - y| \leq 1$ . Say whether  $\sim$  is reflexive. Is it symmetric? Transitive?
18. The table on page 205 shows that relations on  $\mathbb{Z}$  may obey various combinations of the reflexive, symmetric and transitive properties. In all, there are  $2^3 = 8$  possible combinations, and the table shows 5 of them. (There is some redundancy, as  $\leq$  and  $|$  have the same type.) Complete the table by finding examples of relations on  $\mathbb{Z}$  for the three missing combinations.

### 11.3 Equivalence Relations

The relation  $=$  on the set  $\mathbb{Z}$  (or on any set  $A$ ) is reflexive, symmetric and transitive. There are many other relations that are also reflexive, symmetric and transitive. Relations that have all three of these properties occur very frequently in mathematics and often play quite significant roles. (For instance, this is certainly true of the relation  $=$ .) Such relations are given a special name. They are called *equivalence relations*.

**Definition 11.3** A relation  $R$  on a set  $A$  is an **equivalence relation** if it is reflexive, symmetric and transitive.

As an example, Figure 11.3 shows four different equivalence relations  $R_1$ ,  $R_2$ ,  $R_3$  and  $R_4$  on the set  $A = \{-1, 1, 2, 3, 4\}$ . Each one has its own meaning, as labeled. For example, in the second row the relation  $R_2$  literally means “*has the same parity as*.” So  $1R_23$  means “*1 has the same parity as 3*,” etc.

Relation $R$	Diagram	Equivalence classes (see next page)
“is equal to” ( $=$ ) $R_1 = \{(-1,-1), (1,1), (2,2), (3,3), (4,4)\}$		$\{-1\}, \{1\}, \{2\}, \{3\}, \{4\}$
“has same parity as” $R_2 = \{(-1,-1), (1,1), (2,2), (3,3), (4,4), (-1,1), (1,-1), (-1,3), (3,-1), (1,3), (3,1), (2,4), (4,2)\}$		$\{-1, 1, 3\}, \{2, 4\}$
“has same sign as” $R_3 = \{(-1,-1), (1,1), (2,2), (3,3), (4,4), (1,2), (2,1), (1,3), (3,1), (1,4), (4,1), (3,4), (4,3), (2,3), (3,2), (2,4), (4,2), (1,3), (3,1)\}$		$\{-1\}, \{1, 2, 3, 4\}$
“has same parity and sign as” $R_4 = \{(-1,-1), (1,1), (2,2), (3,3), (4,4), (1,3), (3,1), (2,4), (4,2)\}$		$\{-1\}, \{1, 3\}, \{2, 4\}$

**Figure 11.2.** Examples of equivalence relations on the set  $A = \{-1, 1, 2, 3, 4\}$

The above diagrams make it easy to check that each relation is reflexive, symmetric and transitive, i.e., that each is an equivalence relation. For example,  $R_1$  is symmetric because  $xR_1y \Rightarrow yR_1x$  is always true: When  $x = y$  it becomes  $T \Rightarrow T$  (true), and when  $x \neq y$  it becomes  $F \Rightarrow F$  (also true). In a similar fashion,  $R_1$  is transitive because  $(xR_1y \wedge yR_1z) \Rightarrow xR_1z$  is always true: It always works out to one of  $T \Rightarrow T$ ,  $F \Rightarrow T$  or  $F \Rightarrow F$ . (Check this.)

As you can see from the examples in Figure 11.3, equivalence relations on a set tend to express some measure of “sameness” among the elements of the set, whether it is true equality or something weaker (like having the same parity).

It's time to introduce an important definition. Whenever you have an equivalence relation  $R$  on a set  $A$ , it divides  $A$  into subsets called *equivalence classes*. Here is the definition:

**Definition 11.4** Suppose  $R$  is an equivalence relation on a set  $A$ . Given any element  $a \in A$ , the **equivalence class containing  $a$**  is the subset  $\{x \in A : xRa\}$  of  $A$  consisting of all the elements of  $A$  that relate to  $a$ . This set is denoted as  $[a]$ . Thus the equivalence class containing  $a$  is the set  $[a] = \{x \in A : xRa\}$ .

**Example 11.9** Consider the relation  $R_1$  in Figure 11.3. The equivalence class containing 2 is the set  $[2] = \{x \in A : xR_12\}$ . Because in this relation the only element that relates to 2 is 2 itself, we have  $[2] = \{2\}$ . Other equivalence classes for  $R_1$  are  $[-1] = \{-1\}$ ,  $[1] = \{1\}$ ,  $[3] = \{3\}$  and  $[4] = \{4\}$ . Thus this relation has five separate equivalence classes.

**Example 11.10** Consider the relation  $R_2$  in Figure 11.3. The equivalence class containing 2 is the set  $[2] = \{x \in A : xR_22\}$ . Because only 2 and 4 relate to 2, we have  $[2] = \{2, 4\}$ . Observe that we also have  $[4] = \{x \in A : xR_24\} = \{2, 4\}$ , so  $[2] = [4]$ . Another equivalence class for  $R_2$  is  $[1] = \{x \in A : xR_21\} = \{-1, 1, 3\}$ . In addition, note that  $[1] = [-1] = [3] = \{-1, 1, 3\}$ . Thus this relation has just two equivalence classes, namely  $\{2, 4\}$  and  $\{-1, 1, 3\}$ .

**Example 11.11** The relation  $R_4$  in Figure 11.3 has three equivalence classes. They are  $[-1] = \{-1\}$  and  $[1] = [3] = \{1, 3\}$  and  $[2] = [4] = \{2, 4\}$ .

Don't be misled by Figure 11.3. It's important to realize that not every equivalence relation can be drawn as a diagram involving nodes and arrows. Even the simple relation  $R = \{(x, x) : x \in \mathbb{R}\}$ , which expresses equality in the set  $\mathbb{R}$ , is too big to be drawn. Its picture would involve a point for every real number and a loop at each point. Clearly that's too many points and loops to draw.

We close this section with several other examples of equivalence relations on infinite sets.

**Example 11.12** Let  $P$  be the set of all polynomials with real coefficients. Define a relation  $R$  on  $P$  as follows. Given  $f(x), g(x) \in P$ , let  $f(x)Rg(x)$  mean that  $f(x)$  and  $g(x)$  have the same degree. Thus  $(x^2 + 3x - 4)R(3x^2 - 2)$  and  $(x^3 + 3x^2 - 4)R(3x^2 - 2)$ , for example. It takes just a quick mental check to see that  $R$  is an equivalence relation. (Do it.) It's easy to describe the equivalence classes of  $R$ . For example,  $[3x^2 + 2]$  is the set of all polynomials that have the same degree as  $3x^2 + 2$ , that is, the set of all polynomials of degree 2. We can write this as  $[3x^2 + 2] = \{ax^2 + bx + c : a, b, c \in \mathbb{R}, a \neq 0\}$ .

**Example 11.13** In Example 11.8 we proved that for a given  $n \in \mathbb{N}$  the relation  $\equiv (\text{mod } n)$  is reflexive, symmetric and transitive. Thus, in our new parlance,  $\equiv (\text{mod } n)$  is an equivalence relation on  $\mathbb{Z}$ . Consider the case  $n = 3$ . Let's find the equivalence classes of the equivalence relation  $\equiv (\text{mod } 3)$ . The equivalence class containing 0 seems like a reasonable place to start. Observe that

$$\begin{aligned}[0] &= \{x \in \mathbb{Z} : x \equiv 0 \pmod{3}\} = \\ &\quad \{x \in \mathbb{Z} : 3 \mid (x - 0)\} = \{x \in \mathbb{Z} : 3 \mid x\} = \{\dots, -3, 0, 3, 6, 9, \dots\}.\end{aligned}$$

Thus the class  $[0]$  consists of all the multiples of 3. (Or, said differently,  $[0]$  consists of all integers that have a remainder of 0 when divided by 3.) Note that  $[0] = [3] = [6] = [9]$ , etc. The number 1 does not show up in the set  $[0]$  so let's next look at the equivalence class  $[1]$ :

$$[1] = \{x \in \mathbb{Z} : x \equiv 1 \pmod{3}\} = \{x \in \mathbb{Z} : 3 \mid (x - 1)\} = \{\dots, -5, -2, 1, 4, 7, 10, \dots\}.$$

The equivalence class  $[1]$  consists of all integers that give a remainder of 1 when divided by 3. The number 2 is in neither of the sets  $[0]$  or  $[1]$ , so we next look at the equivalence class  $[2]$ :

$$[2] = \{x \in \mathbb{Z} : x \equiv 2 \pmod{3}\} = \{x \in \mathbb{Z} : 3 \mid (x - 2)\} = \{\dots, -4, -1, 2, 5, 8, 11, \dots\}.$$

The equivalence class  $[2]$  consists of all integers that give a remainder of 2 when divided by 3. Observe that any integer is in one of the sets  $[0]$ ,  $[1]$  or  $[2]$ , so we have listed all of the equivalence classes. Thus  $\equiv (\text{mod } 3)$  has exactly three equivalence classes, as described above.

Similarly, you can show that the equivalence relation  $\equiv (\text{mod } n)$  has  $n$  equivalence classes  $[0], [1], [2], \dots, [n - 1]$ .

The idea of an equivalence relation is fundamental. In a very real sense you have dealt with equivalence relations for much of your life, without being aware of it. In fact your conception of fractions is entwined with an intuitive notion of an equivalence relation. To see how this is so, consider the set of all fractions, *not necessarily reduced*:

$$F = \left\{ \frac{m}{n} : m, n \in \mathbb{Z}, n \neq 0 \right\}.$$

Interpret this set not as  $\mathbb{Q}$ , but rather as the set of all possible fractions. For example, we consider the fractions  $\frac{1}{2}$  and  $\frac{2}{4}$  as being distinct (unequal) elements of  $F$  because their numerators and denominators don't match. Of course  $\frac{1}{2}$  and  $\frac{2}{4}$  are equal numbers, but they are *different* fractions, so  $\frac{1}{2}, \frac{2}{4} \in F$ , but  $\frac{1}{2} \neq \frac{2}{4}$  (meaning they are distinct, unequal elements of  $F$ ).

Define a relation  $\doteq$  on  $F$  by saying  $\frac{a}{b} \doteq \frac{c}{d}$  provided that  $ad = bc$ . Thus  $\frac{1}{2} \doteq \frac{2}{4}$  because  $1 \cdot 4 = 2 \cdot 2$ . Similarly, notice that  $\frac{-15}{3} \doteq \frac{10}{2}$  because  $-15 \cdot 2 = -3 \cdot 10$ . We have defined  $\doteq$  so that  $\frac{a}{b} \doteq \frac{c}{d}$  if and only if  $\frac{a}{b}$  and  $\frac{c}{d}$  are equal numbers, so  $\doteq$  models your intuitive, ingrained understanding of when two different fractions are equal.

Observe that  $\doteq$  is an equivalence relation on the set  $F$  of all fractions: It is reflexive because for any  $\frac{a}{b} \in F$  the equation  $ab = ba$  guarantees  $\frac{a}{b} \doteq \frac{a}{b}$ . To see that  $\doteq$  is symmetric, suppose  $\frac{a}{b} \doteq \frac{c}{d}$ . This means  $ad = bc$ , so  $cb = da$ , which implies  $\frac{c}{d} \doteq \frac{a}{b}$ . Exercise 16 below asks you to confirm that  $\doteq$  is transitive.

This discussion shows that your everyday understanding of equality of fractions is an equivalence relation. The equivalence class containing, say,  $\frac{2}{3}$  is the set  $\{\frac{2n}{3n} : n \in \mathbb{Z}, n \neq 0\}$  of all fractions that are numerically equal to  $\frac{2}{3}$ . The takeaway is that you have for years lumped together equal fractions into equivalence classes under this equivalence relation.

Later, when you learned calculus, equivalence relations once again were just under the surface. The antiderivative  $\int f(x)dx$  of a function  $f(x)$  was defined to be the set of functions  $F(x) + C$  whose derivatives are  $f(x)$ . This set is an equivalence class in the set of integrable functions, where two functions are related if their difference is a constant. (We are glossing over some fine points that will be cleared up in an advanced calculus course.)

Such examples underscore an important point: Equivalence relations arise in many areas of mathematics. This is especially true in the advanced realms of mathematics, where equivalence relations are the right tool for important constructions, constructions as natural and far-reaching as fractions, or antiderivatives. Learning about equivalence relations now paves the way to a deeper understanding of later courses, and work.

---

**Exercises for Section 11.3**

1. Let  $A = \{1, 2, 3, 4, 5, 6\}$ , and consider the following equivalence relation on  $A$ :  
 $R = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (2, 3), (3, 2), (4, 5), (5, 4), (4, 6), (6, 4), (5, 6), (6, 5)\}$   
 List the equivalence classes of  $R$ .
  2. Let  $A = \{a, b, c, d, e\}$ . Suppose  $R$  is an equivalence relation on  $A$ . Suppose  $R$  has two equivalence classes. Also  $aRd$ ,  $bRc$  and  $eRd$ . Write out  $R$  as a set.
  3. Let  $A = \{a, b, c, d, e\}$ . Suppose  $R$  is an equivalence relation on  $A$ . Suppose  $R$  has three equivalence classes. Also  $aRd$  and  $bRc$ . Write out  $R$  as a set.
  4. Let  $A = \{a, b, c, d, e\}$ . Suppose  $R$  is an equivalence relation on  $A$ . Suppose also that  $aRd$  and  $bRc$ ,  $eRa$  and  $cRe$ . How many equivalence classes does  $R$  have?
  5. There are two different equivalence relations on the set  $A = \{a, b\}$ . Describe them. Diagrams will suffice.
  6. There are five different equivalence relations on the set  $A = \{a, b, c\}$ . Describe them all. Diagrams will suffice.
  7. Define a relation  $R$  on  $\mathbb{Z}$  as  $xRy$  if and only if  $3x - 5y$  is even. Prove  $R$  is an equivalence relation. Describe its equivalence classes.
  8. Define a relation  $R$  on  $\mathbb{Z}$  as  $xRy$  if and only if  $x^2 + y^2$  is even. Prove  $R$  is an equivalence relation. Describe its equivalence classes.
  9. Define a relation  $R$  on  $\mathbb{Z}$  as  $xRy$  if and only if  $4|(x+3y)$ . Prove  $R$  is an equivalence relation. Describe its equivalence classes.
  10. Suppose  $R$  and  $S$  are two equivalence relations on a set  $A$ . Prove that  $R \cap S$  is also an equivalence relation. (For an example of this, look at Figure 11.3. Observe that for the equivalence relations  $R_2, R_3$  and  $R_4$ , we have  $R_2 \cap R_3 = R_4$ .)
  11. Prove or disprove: If  $R$  is an equivalence relation on an infinite set  $A$ , then  $R$  has infinitely many equivalence classes.
  12. Prove or disprove: If  $R$  and  $S$  are two equivalence relations on a set  $A$ , then  $R \cup S$  is also an equivalence relation on  $A$ .
  13. Suppose  $R$  is an equivalence relation on a finite set  $A$ , and every equivalence class has the same cardinality  $m$ . Express  $|R|$  in terms of  $m$  and  $|A|$ .
  14. Suppose  $R$  is a reflexive and symmetric relation on a finite set  $A$ . Define a relation  $S$  on  $A$  by declaring  $xSy$  if and only if for some  $n \in \mathbb{N}$  there are elements  $x_1, x_2, \dots, x_n \in A$  satisfying  $xRx_1$ ,  $x_1Rx_2$ ,  $x_2Rx_3$ ,  $x_3Rx_4, \dots, x_{n-1}Rx_n$ , and  $x_nRy$ . Show that  $S$  is an equivalence relation and  $R \subseteq S$ . Prove that  $S$  is the unique smallest equivalence relation on  $A$  containing  $R$ .
  15. Suppose  $R$  is an equivalence relation on a set  $A$ , with four equivalence classes. How many different equivalence relations  $S$  on  $A$  are there for which  $R \subseteq S$ ?
  16. Show that the relation  $\doteq$  defined on page 213 is transitive.
-

## 11.4 Equivalence Classes and Partitions

This section collects several properties of equivalence classes.

Our first result proves that  $[a] = [b]$  if and only if  $aRb$ . This is useful because it assures us that whenever we are in a situation where  $[a] = [b]$ , we also have  $aRb$ , and vice versa. Being able to switch back and forth between these two pieces of information can be helpful in a variety of situations, and you may find yourself using this result a lot. Be sure to notice that the proof uses all three properties (reflexive, symmetric and transitive) of equivalence relations. Notice also that we have to use some techniques from Chapter 8 (Proofs Involving Sets) in dealing with the sets  $[a]$  and  $[b]$ .

**Theorem 11.1** Suppose  $R$  is an equivalence relation on a set  $A$ . Suppose also that  $a, b \in A$ . Then  $[a] = [b]$  if and only if  $aRb$ .

*Proof.* Suppose  $[a] = [b]$ . Note that  $aRa$  by the reflexive property of  $R$ , so  $a \in \{x \in A : xRa\} = [a] = [b] = \{x \in A : xRb\}$ . But  $a$  belonging to  $\{x \in A : xRb\}$  means  $aRb$ . This completes the first part of the if-and-only-if proof.

Conversely, suppose  $aRb$ . We need to show  $[a] = [b]$ . We will do this by showing  $[a] \subseteq [b]$  and  $[b] \subseteq [a]$ .

First we show  $[a] \subseteq [b]$ . Suppose  $c \in [a]$ . As  $c \in [a] = \{x \in A : xRa\}$ , we get  $cRa$ . Now we have  $cRa$  and  $aRb$ , so  $cRb$  because  $R$  is transitive. But  $cRb$  implies  $c \in \{x \in A : xRb\} = [b]$ . This demonstrates that  $c \in [a]$  implies  $c \in [b]$ , so  $[a] \subseteq [b]$ .

Next we show  $[b] \subseteq [a]$ . Suppose  $c \in [b]$ . As  $c \in [b] = \{x \in A : xRb\}$ , we get  $cRb$ . Remember that we are assuming  $aRb$ , so  $bRa$  because  $R$  is symmetric. Now we have  $cRb$  and  $bRa$ , so  $cRa$  because  $R$  is transitive. But  $cRa$  implies  $c \in \{x \in A : xRa\} = [a]$ . This demonstrates that  $c \in [b]$  implies  $c \in [a]$ ; hence  $[b] \subseteq [a]$ .

The previous two paragraphs imply that  $[a] = [b]$ . ■

To illustrate Theorem 11.1, recall that in Example 11.13 we worked out the equivalence classes of  $\equiv (\text{mod } 3)$ . We observed that

$$[-3] = [9] = \{\dots, -3, 0, 3, 6, 9, \dots\}.$$

Note that  $[-3] = [9]$  and  $-3 \equiv 9 \pmod{3}$ , just as Theorem 11.1 predicts. The theorem assures us that this will work for any equivalence relation. In the future you may find yourself using the result of Theorem 11.1 often. Over time it may become natural and familiar; you will use it automatically, without even thinking of it as a theorem.

Our next topic addresses the fact that an equivalence relation on a set  $A$  divides  $A$  into various equivalence classes. There is a special word for this kind of situation. We address it now, as you are likely to encounter it in subsequent mathematics classes.

**Definition 11.5** A **partition** of a set  $A$  is a set of non-empty subsets of  $A$ , such that the union of all the subsets equals  $A$ , and the intersection of any two different subsets is  $\emptyset$ .

**Example 11.14** Let  $A = \{a, b, c, d\}$ . One partition of  $A$  is  $\{\{a, b\}, \{c\}, \{d\}\}$ . This is a set of three subsets  $\{a, b\}$ ,  $\{c\}$  and  $\{d\}$  of  $A$ . The union of the three subsets equals  $A$ ; the intersection of any two subsets is  $\emptyset$ .

Other partitions of  $A$  are

$$\{\{a, b\}, \{c, d\}\}, \quad \{\{a, c\}, \{b\}, \{d\}\}, \quad \{\{a\}, \{b\}, \{c\}, \{d\}\}, \quad \{\{a, b, c, d\}\},$$

to name a few. Intuitively, a partition is just a dividing up of  $A$  into pieces.

**Example 11.15** Consider the equivalence relations in Figure 11.3. Each of these is a relation on the set  $A = \{-1, 1, 2, 3, 4\}$ . The equivalence classes of each relation are listed on the right side of the figure. Observe that, in each case, the set of equivalence classes forms a partition of  $A$ . For example, the relation  $R_1$  yields the partition  $\{\{-1\}, \{1\}, \{2\}, \{3\}, \{4\}\}$  of  $A$ . Likewise the equivalence classes of  $R_2$  form the partition  $\{\{-1, 1, 3\}, \{2, 4\}\}$ .

**Example 11.16** Recall that Example 11.13 worked out the equivalence classes of the equivalence relation  $\equiv (\text{mod } 3)$  on the set  $\mathbb{Z}$ . These equivalence classes give the following partition of  $\mathbb{Z}$ :

$$\left\{ \{\dots, -3, 0, 3, 6, 9, \dots\}, \{\dots, -2, 1, 4, 7, 10, \dots\}, \{\dots, -1, 2, 5, 8, 11, \dots\} \right\}.$$

We can write it more compactly as  $\{[0], [1], [2]\}$ .

Our examples and experience suggest that the equivalence classes of an equivalence relation on a set form a partition of that set. This is indeed the case, and we now prove it.

**Theorem 11.2** Suppose  $R$  is an equivalence relation on a set  $A$ . Then the set  $\{[a] : a \in A\}$  of equivalence classes of  $R$  forms a partition of  $A$ .

*Proof.* To show that  $\{[a] : a \in A\}$  is a partition of  $A$  we need to show two things: We need to show that the union of all the sets  $[a]$  equals  $A$ , and we need to show that if  $[a] \neq [b]$ , then  $[a] \cap [b] = \emptyset$ .

Notationally, the union of all the sets  $[a]$  is  $\bigcup_{a \in A} [a]$ , so we need to prove  $\bigcup_{a \in A} [a] = A$ . Suppose  $x \in \bigcup_{a \in A} [a]$ . This means  $x \in [a]$  for some  $a \in A$ . Since  $[a] \subseteq A$ , it then follows that  $x \in A$ . Thus  $\bigcup_{a \in A} [a] \subseteq A$ . On the other hand, suppose  $x \in A$ . As  $x \in [x]$ , we know  $x \in [a]$  for some  $a \in A$  (namely  $a = x$ ). Therefore  $x \in \bigcup_{a \in A} [a]$ , and this shows  $A \subseteq \bigcup_{a \in A} [a]$ . Since  $\bigcup_{a \in A} [a] \subseteq A$  and  $A \subseteq \bigcup_{a \in A} [a]$ , it follows that  $\bigcup_{a \in A} [a] = A$ .

Next we need to show that if  $[a] \neq [b]$  then  $[a] \cap [b] = \emptyset$ . Let's use contrapositive proof. Suppose it's not the case that  $[a] \cap [b] = \emptyset$ , so there is some element  $c$  with  $c \in [a] \cap [b]$ . Thus  $c \in [a]$  and  $c \in [b]$ . Now,  $c \in [a]$  means  $cRa$ , and then  $aRc$  since  $R$  is symmetric. Also  $c \in [b]$  means  $cRb$ . Now we have  $aRc$  and  $cRb$ , so  $aRb$  (because  $R$  is transitive). By Theorem 11.1,  $aRb$  implies  $[a] = [b]$ . Thus  $[a] \neq [b]$  is not true.

We've now shown that the union of all the equivalence classes is  $A$ , and the intersection of two different equivalence classes is  $\emptyset$ . Therefore the set of equivalence classes is a partition of  $A$ . ■

Theorem 11.2 says the equivalence classes of any equivalence relation on a set  $A$  form a partition of  $A$ . Conversely, any partition of  $A$  describes an equivalence relation  $R$  where  $xRy$  if and only if  $x$  and  $y$  belong to the same set in the partition. (See Exercise 4 for this section, below.) Thus equivalence relations and partitions are really just two different ways of looking at the same thing. In your future mathematical studies you may find yourself easily switching between these two points of view.

## Exercises for Section 11.4

1. List all the partitions of the set  $A = \{a, b\}$ . Compare your answer to the answer to Exercise 5 of Section 11.3.
2. List all the partitions of the set  $A = \{a, b, c\}$ . Compare your answer to the answer to Exercise 6 of Section 11.3.
3. Describe the partition of  $\mathbb{Z}$  resulting from the equivalence relation  $\equiv (\text{mod } 4)$ .
4. Suppose  $P$  is a partition of a set  $A$ . Define a relation  $R$  on  $A$  by declaring  $xRy$  if and only if  $x, y \in X$  for some  $X \in P$ . Prove  $R$  is an equivalence relation on  $A$ . Then prove that  $P$  is the set of equivalence classes of  $R$ .
5. Consider the partition  $P = \{\dots, -4, -2, 0, 2, 4, \dots, \}, \{\dots, -5, -3, -1, 1, 3, 5, \dots, \}\}$  of  $\mathbb{Z}$ . Let  $R$  be the equivalence relation whose equivalence classes are the two elements of  $P$ . What familiar equivalence relation is  $R$ ?
6. Consider the partition  $P = \{\{0\}, \{-1, 1\}, \{-2, 2\}, \{-3, 3\}, \{-4, 4\}, \dots\}$  of  $\mathbb{Z}$ . Describe the equivalence relation whose equivalence classes are the elements of  $P$ .

## 11.5 The Integers Modulo n

Example 11.8 proved that for any given  $n \in \mathbb{N}$ , the relation  $\equiv (\text{mod } n)$  is reflexive, symmetric and transitive, so it is an equivalence relation. This is a particularly significant equivalence relation in mathematics, and in the present section we deduce some of its properties.

To make matters simpler, let's pick a concrete  $n$ , say  $n = 5$ . Let's begin by looking at the equivalence classes of the relation  $\equiv (\text{mod } 5)$ . There are five equivalence classes, as follows:

$$\begin{aligned}[0] &= \{x \in \mathbb{Z} : x \equiv 0 \pmod{5}\} = \{x \in \mathbb{Z} : 5 | (x - 0)\} = \{\dots, -10, -5, 0, 5, 10, 15, \dots\}, \\ [1] &= \{x \in \mathbb{Z} : x \equiv 1 \pmod{5}\} = \{x \in \mathbb{Z} : 5 | (x - 1)\} = \{\dots, -9, -4, 1, 6, 11, 16, \dots\}, \\ [2] &= \{x \in \mathbb{Z} : x \equiv 2 \pmod{5}\} = \{x \in \mathbb{Z} : 5 | (x - 2)\} = \{\dots, -8, -3, 2, 7, 12, 17, \dots\}, \\ [3] &= \{x \in \mathbb{Z} : x \equiv 3 \pmod{5}\} = \{x \in \mathbb{Z} : 5 | (x - 3)\} = \{\dots, -7, -2, 3, 8, 13, 18, \dots\}, \\ [4] &= \{x \in \mathbb{Z} : x \equiv 4 \pmod{5}\} = \{x \in \mathbb{Z} : 5 | (x - 4)\} = \{\dots, -6, -1, 4, 9, 14, 19, \dots\}.\end{aligned}$$

Notice how these equivalence classes form a partition of the set  $\mathbb{Z}$ . We label the five equivalence classes as  $[0], [1], [2], [3]$  and  $[4]$ , but you know of course that there are other ways to label them. For example,  $[0] = [5] = [10] = [15]$ , and so on; and  $[1] = [6] = [-4]$ , etc. Still, for this discussion we denote the five classes as  $[0], [1], [2], [3]$  and  $[4]$ .

These five classes form a set, which we shall denote as  $\mathbb{Z}_5$ . Thus

$$\mathbb{Z}_5 = \{[0], [1], [2], [3], [4]\}$$

is a set of five sets. The interesting thing about  $\mathbb{Z}_5$  is that even though its elements are sets (and not numbers), it is possible to add and multiply them. In fact, we can define the following rules that tell how elements of  $\mathbb{Z}_5$  can be added and multiplied.

$$\begin{aligned}[a] + [b] &= [a + b] \\ [a] \cdot [b] &= [a \cdot b]\end{aligned}$$

For example,  $[2] + [1] = [2 + 1] = [3]$ , and  $[2] \cdot [2] = [2 \cdot 2] = [4]$ . We stress that in doing this we are adding and multiplying *sets* (more precisely equivalence classes), not numbers. We added (or multiplied) two elements of  $\mathbb{Z}_5$  and obtained another element of  $\mathbb{Z}_5$ .

Here is a trickier example. Observe that  $[2] + [3] = [5]$ . This time we added elements  $[2], [3] \in \mathbb{Z}_5$ , and got the element  $[5] \in \mathbb{Z}_5$ . That was easy, except where is our answer  $[5]$  in the set  $\mathbb{Z}_5 = \{[0], [1], [2], [3], [4]\}$ ? Since  $[5] = [0]$ , it is more appropriate to write  $[2] + [3] = [0]$ .

In a similar vein,  $[2] \cdot [3] = [6]$  would be written as  $[2] \cdot [3] = [1]$  because  $[6] = [1]$ . Test your skill with this by verifying the following addition and multiplication tables for  $\mathbb{Z}_5$ .

+	[0]	[1]	[2]	[3]	[4]	.	[0]	[1]	[2]	[3]	[4]
[0]	[0]	[1]	[2]	[3]	[4]	[0]	[0]	[0]	[0]	[0]	[0]
[1]	[1]	[2]	[3]	[4]	[0]	[1]	[0]	[1]	[2]	[3]	[4]
[2]	[2]	[3]	[4]	[0]	[1]	[2]	[0]	[2]	[4]	[1]	[3]
[3]	[3]	[4]	[0]	[1]	[2]	[3]	[0]	[3]	[1]	[4]	[2]
[4]	[4]	[0]	[1]	[2]	[3]	[4]	[0]	[4]	[3]	[2]	[1]

We call the set  $\mathbb{Z}_5 = \{[0], [1], [2], [3], [4]\}$  the **integers modulo 5**. As our tables suggest,  $\mathbb{Z}_5$  is more than just a set: It is a little number system with its own addition and multiplication. In this way it is like the familiar set  $\mathbb{Z}$  which also comes equipped with an addition and a multiplication.

Of course, there is nothing special about the number 5. We can also define  $\mathbb{Z}_n$  for any natural number  $n$ . Here is the definition:

**Definition 11.6** Let  $n \in \mathbb{N}$ . The equivalence classes of the equivalence relation  $\equiv (\text{mod } n)$  are  $[0], [1], [2], \dots, [n - 1]$ . The **integers modulo  $n$**  is the set  $\mathbb{Z}_n = \{[0], [1], [2], \dots, [n - 1]\}$ . Elements of  $\mathbb{Z}_n$  can be added by the rule  $[a] + [b] = [a + b]$  and multiplied by the rule  $[a] \cdot [b] = [ab]$ .

Given a natural number  $n$ , the set  $\mathbb{Z}_n$  is a number system containing  $n$  elements. It has many of the algebraic properties that  $\mathbb{Z}, \mathbb{R}$  and  $\mathbb{Q}$  possess. For example, it is probably obvious to you already that elements of  $\mathbb{Z}_n$  obey the commutative laws  $[a] + [b] = [b] + [a]$  and  $[a] \cdot [b] = [b] \cdot [a]$ . You can also verify the distributive law  $[a] \cdot ([b] + [c]) = [a] \cdot [b] + [a] \cdot [c]$ , as follows:

$$\begin{aligned}
 [a] \cdot ([b] + [c]) &= [a] \cdot [b + c] \\
 &= [a(b + c)] \\
 &= [ab + ac] \\
 &= [ab] + [ac] \\
 &= [a] \cdot [b] + [a] \cdot [c].
 \end{aligned}$$

The integers modulo  $n$  are significant because they more closely fit certain applications than do other number systems such as  $\mathbb{Z}$  or  $\mathbb{R}$ . If you go on to

take a course in abstract algebra, then you will work extensively with  $\mathbb{Z}_n$  as well as other, more exotic, number systems. (In such a course you will also use all of the proof techniques that we have discussed, as well as the ideas of equivalence relations.)

To close this section we take up an issue that may have bothered you earlier. It has to do with our definitions of addition  $[a] + [b] = [a + b]$  and multiplication  $[a] \cdot [b] = [ab]$ . These definitions define addition and multiplication of equivalence classes in terms of representatives  $a$  and  $b$  in the equivalence classes. Since there are many different ways to choose such representatives, we may well wonder if addition and multiplication are consistently defined. For example, suppose two people, Alice and Bob, want to multiply the elements  $[2]$  and  $[3]$  in  $\mathbb{Z}_5$ . Alice does the calculation as  $[2] \cdot [3] = [6] = [1]$ , so her final answer is  $[1]$ . Bob does it differently. Since  $[2] = [7]$  and  $[3] = [8]$ , he works out  $[2] \cdot [3]$  as  $[7] \cdot [8] = [56]$ . Since  $56 \equiv 1 \pmod{5}$ , Bob's answer is  $[56] = [1]$ , and that agrees with Alice's answer. Will their answers always agree or did they just get lucky (with the arithmetic)?

The fact is that no matter how they do the multiplication in  $\mathbb{Z}_n$ , their answers will agree. To see why, suppose Alice and Bob want to multiply the elements  $[a], [b] \in \mathbb{Z}_n$ , and suppose  $[a] = [a']$  and  $[b] = [b']$ . Alice and Bob do the multiplication as follows:

$$\begin{aligned} \text{Alice: } & [a] \cdot [b] = [ab], \\ \text{Bob: } & [a'] \cdot [b'] = [a'b']. \end{aligned}$$

We need to show that their answers agree, that is, we need to show  $[ab] = [a'b']$ . Since  $[a] = [a']$ , we know by Theorem 11.1 that  $a \equiv a' \pmod{n}$ . Thus  $n \mid (a - a')$ , so  $a - a' = nk$  for some integer  $k$ . Likewise, as  $[b] = [b']$ , we know  $b \equiv b' \pmod{n}$ , or  $n \mid (b - b')$ , so  $b - b' = n\ell$  for some integer  $\ell$ . Thus we get  $a = a' + nk$  and  $b = b' + n\ell$ . Therefore:

$$\begin{aligned} ab &= (a' + nk)(b' + n\ell) \\ &= a'b' + a'n\ell + nk b' + n^2 k\ell. \end{aligned}$$

Hence  $ab - a'b' = n(a'\ell + kb' + nk\ell)$ . This means  $n \mid (ab - a'b')$ , so  $ab \equiv a'b' \pmod{n}$ , and from that we conclude  $[ab] = [a'b']$ . Consequently Alice and Bob really do get the same answer, so we can be assured that the definition of multiplication in  $\mathbb{Z}_n$  is consistent.

Exercise 8 (below) asks you to prove that addition in  $\mathbb{Z}_n$  is similarly consistent.

---

### Exercises for Section 11.5

1. Write the addition and multiplication tables for  $\mathbb{Z}_2$ .
  2. Write the addition and multiplication tables for  $\mathbb{Z}_3$ .
  3. Write the addition and multiplication tables for  $\mathbb{Z}_4$ .
  4. Write the addition and multiplication tables for  $\mathbb{Z}_6$ .
  5. Suppose  $[a], [b] \in \mathbb{Z}_5$  and  $[a] \cdot [b] = [0]$ . Is it necessarily true that either  $[a] = [0]$  or  $[b] = [0]$ ?
  6. Suppose  $[a], [b] \in \mathbb{Z}_6$  and  $[a] \cdot [b] = [0]$ . Is it necessarily true that either  $[a] = [0]$  or  $[b] = [0]$ ? What if  $[a], [b] \in \mathbb{Z}_7$ ?
  7. Do the following calculations in  $\mathbb{Z}_9$ , in each case expressing your answer as  $[a]$  with  $0 \leq a \leq 8$ .
 

<b>(a)</b> $[8] + [8]$	<b>(b)</b> $[24] + [11]$	<b>(c)</b> $[21] \cdot [15]$	<b>(d)</b> $[8] \cdot [8]$
------------------------	--------------------------	------------------------------	----------------------------
  8. Suppose  $[a], [b] \in \mathbb{Z}_n$ , and  $[a] = [a']$  and  $[b] = [b']$ . Alice adds  $[a]$  and  $[b]$  as  $[a] + [b] = [a + b]$ . Bob adds them as  $[a'] + [b'] = [a' + b']$ . Show that their answers  $[a + b]$  and  $[a' + b']$  are the same.
- 

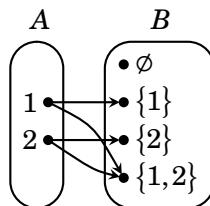
### 11.6 Relations Between Sets

In the beginning of this chapter, we defined a relation on a set  $A$  to be a subset  $R \subseteq A \times A$ . This created a framework that could model any situation in which elements of  $A$  are compared to themselves. In this setting, the statement  $xRy$  has elements  $x$  and  $y$  from  $A$  on either side of the  $R$  because  $R$  compares elements from  $A$ . But there are other relational symbols that don't work this way. Consider  $\in$ . The statement  $5 \in \mathbb{Z}$  expresses a relationship between 5 and  $\mathbb{Z}$  (namely that the element 5 is in the set  $\mathbb{Z}$ ) but 5 and  $\mathbb{Z}$  are not in any way naturally regarded as both elements of some set  $A$ . To overcome this difficulty, we generalize the idea of a relation on  $A$  to a *relation from  $A$  to  $B$* .

**Definition 11.7** A **relation** from a set  $A$  to a set  $B$  is a subset  $R \subseteq A \times B$ . We often abbreviate the statement  $(x, y) \in R$  as  $xRy$ . The statement  $(x, y) \notin R$  is abbreviated as  $xR\bar{y}$ .

**Example 11.17** Suppose  $A = \{1, 2\}$  and  $B = \mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ . Then  $R = \{(1, \{1\}), (2, \{2\}), (1, \{1, 2\}), (2, \{1, 2\})\} \subseteq A \times B$  is a relation from  $A$  to  $B$ . Note that  $1R\{1\}$ ,  $2R\{2\}$ ,  $1R\{1, 2\}$  and  $2R\{1, 2\}$ . The relation  $R$  is the familiar relation  $\in$  for the set  $A$ , that is,  $xR X$  means exactly the same thing as  $x \in X$ .

Diagrams for relations from  $A$  to  $B$  differ from diagrams for relations on  $A$ . Since there are two sets  $A$  and  $B$  in a relation from  $A$  to  $B$ , we have to draw labeled nodes for each of the two sets. Then we draw arrows from  $x$  to  $y$  whenever  $xRy$ . The following figure illustrates this for Example 11.17.



**Figure 11.3.** A relation from  $A$  to  $B$

The ideas from this chapter show that any relation (whether it is a familiar one like  $\geq$ ,  $\leq$ ,  $=$ ,  $|$ ,  $\in$  or  $\subseteq$ , or a more exotic one) is really just a set. Therefore the theory of relations is a part of the theory of sets. In the next chapter, we will see that this idea touches on another important mathematical construction, namely functions. We will define a function to be a special kind of relation from one set to another, and in this context we will see that any function is really just a set.

# CHAPTER 12

---

## Functions

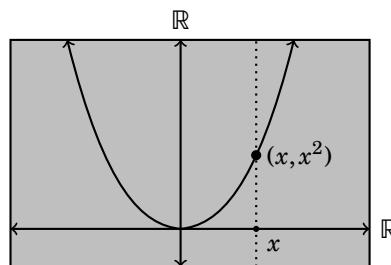
---

You know from algebra and calculus that functions play a fundamental role in mathematics. You may view a function as a kind of formula that describes a relationship between two (or more) quantities. You certainly understand and appreciate the fact that relationships between quantities are central to all scientific disciplines, so you do not need to be convinced that functions are important. Still, you may not be fully aware of the significance of functions. Functions are more than merely descriptions of numeric relationships. In a more general sense, functions can compare and relate different kinds of mathematical structures. You will see this as your understanding of mathematics deepens. In preparation of this, we will now explore a more general and versatile view of functions.

The concept of a relation between sets (Definition 11.7) plays a big role here, so you may want to quickly review it.

### 12.1 Functions

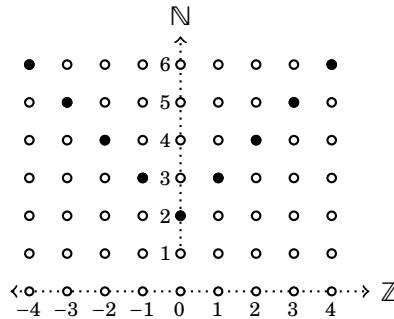
Let's start on familiar ground. Consider the function  $f(x) = x^2$  from  $\mathbb{R}$  to  $\mathbb{R}$ . Its graph is the set of points  $R = \{(x, x^2) : x \in \mathbb{R}\} \subseteq \mathbb{R} \times \mathbb{R}$ .



**Figure 12.1.** A familiar function

Having read Chapter 11, you may see  $f$  in a new light. Its graph  $R \subseteq \mathbb{R} \times \mathbb{R}$  is a relation on the set  $\mathbb{R}$ . In fact, as we shall see, functions are just special kinds of relations. Before stating the exact definition, we look at another

example. Consider the function  $f(n) = |n| + 2$  that converts integers  $n$  into natural numbers  $|n| + 2$ . Its graph is  $R = \{(n, |n| + 2) : n \in \mathbb{Z}\} \subseteq \mathbb{Z} \times \mathbb{N}$ .



**Figure 12.2.** The function  $f : \mathbb{Z} \rightarrow \mathbb{N}$ , where  $f(n) = |n| + 2$

Figure 12.2 shows the graph  $R$  as darkened dots in the grid of points  $\mathbb{Z} \times \mathbb{N}$ . Notice that in this example  $R$  is not a relation on a single set. The set of input values  $\mathbb{Z}$  is different from the set  $\mathbb{N}$  of output values, so the graph  $R \subseteq \mathbb{Z} \times \mathbb{N}$  is a *relation from  $\mathbb{Z}$  to  $\mathbb{N}$* .

This example illustrates three things. First, a function can be viewed as sending elements from one set  $A$  to another set  $B$ . (In the case of  $f$ ,  $A = \mathbb{Z}$  and  $B = \mathbb{N}$ .) Second, such a function can be regarded as a relation from  $A$  to  $B$ . Third, for every input value  $n$ , there is *exactly one* output value  $f(n)$ . In your high school algebra course, this was expressed by the *vertical line test*: Any vertical line intersects a function's graph at most once. It means that for any input value  $x$ , the graph contains exactly one point of form  $(x, f(x))$ . Our main definition, given below, incorporates all of these ideas.

**Definition 12.1** Suppose  $A$  and  $B$  are sets. A **function**  $f$  from  $A$  to  $B$  (denoted as  $f : A \rightarrow B$ ) is a relation  $f \subseteq A \times B$  from  $A$  to  $B$ , satisfying the property that for each  $a \in A$  the relation  $f$  contains exactly one ordered pair of form  $(a, b)$ . The statement  $(a, b) \in f$  is abbreviated  $f(a) = b$ .

**Example 12.1** Consider the function  $f : \mathbb{Z} \rightarrow \mathbb{N}$  graphed in Figure 12.2. According to Definition 12.1, we regard  $f$  as the set of points in its graph, that is,  $f = \{(n, |n| + 2) : n \in \mathbb{Z}\} \subseteq \mathbb{Z} \times \mathbb{N}$ . This is a relation from  $\mathbb{Z}$  to  $\mathbb{N}$ , and indeed given any  $a \in \mathbb{Z}$  the set  $f$  contains exactly one ordered pair  $(a, |a| + 2)$  whose first coordinate is  $a$ . Since  $(1, 3) \in f$ , we write  $f(1) = 3$ ; and since  $(-3, 5) \in f$  we write  $f(-3) = 5$ , etc. In general,  $(a, b) \in f$  means that  $f$  sends the input value  $a$  to the output value  $b$ , and we express this as  $f(a) = b$ . This

function can be expressed by a formula: For each input value  $n$ , the output value is  $|n| + 2$ , so we may write  $f(n) = |n| + 2$ . All this agrees with the way we thought of functions in algebra and calculus; the only difference is that now we also think of a function as a relation.

**Definition 12.2** For a function  $f : A \rightarrow B$ , the set  $A$  is called the **domain** of  $f$ . (Think of the domain as the set of possible “input values” for  $f$ .) The set  $B$  is called the **codomain** of  $f$ . The **range** of  $f$  is the set  $\{f(a) : a \in A\} = \{b : (a, b) \in f\}$ . (Think of the range as the set of all possible “output values” for  $f$ . Think of the codomain as a sort of “target” for the outputs.)

Consider the function  $f : \mathbb{Z} \rightarrow \mathbb{N}$ , where  $f(n) = |n| + 2$ , from Example 12.1. The domain is  $\mathbb{Z}$  and the codomain is  $\mathbb{N}$ . The range of this function is the set  $\{f(a) : a \in \mathbb{Z}\} = \{|a| + 2 : a \in \mathbb{Z}\} = \{2, 3, 4, 5, \dots\}$ . Notice that the range is a subset of the codomain  $\mathbb{N}$ , but it does not (in this case) equal the codomain. In general, the range of a function is a subset of the codomain. In this sense the codomain could have been *any* set that contains the range. We might just as well have said that this  $f$  is a function  $f : \mathbb{Z} \rightarrow \mathbb{Z}$ , or even  $f : \mathbb{Z} \rightarrow \mathbb{R}$ .

This illustrates an important point: the codomain of a function is not an intrinsic feature of the function; it is more a matter of choice or context. In Example 12.1 we chose  $\mathbb{N}$  as the codomain because all the output values of  $f$  are natural numbers. But in general, the codomain of a function can be any set that contains the function’s range as a subset.

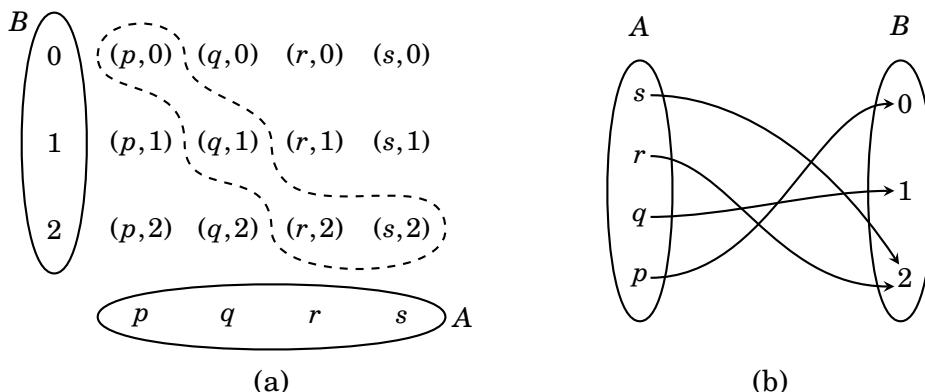
In our examples so far, the domains and codomains are sets of numbers, but this needn’t be the case in general, as the next example indicates.

**Example 12.2** Let  $A = \{p, q, r, s\}$  and  $B = \{0, 1, 2\}$ , and

$$f = \{(p, 0), (q, 1), (r, 2), (s, 2)\} \subseteq A \times B.$$

This is a function  $f : A \rightarrow B$  because each element of  $A$  occurs exactly once as a first coordinate of an ordered pair in  $f$ . Observe that we have  $f(p) = 0$ ,  $f(q) = 1$ ,  $f(r) = 2$  and  $f(s) = 2$ . The domain of this function is  $A = \{p, q, r, s\}$ . The codomain and range are both  $B = \{0, 1, 2\}$ .

If  $A$  and  $B$  are not both sets of numbers (as in this example), it can be difficult to draw a graph of  $f : A \rightarrow B$  in the traditional sense. Figure 12.3(a) is an attempt at a graph of  $f$ . The sets  $A$  and  $B$  are aligned as  $x$ - and  $y$ -axes, and the Cartesian product  $A \times B$  is filled in accordingly. The subset  $f \subseteq A \times B$  is indicated with dashed lines, and we can regard it as a “graph” of  $f$ . Figure 12.3(b) shows a more natural visual description of  $f$ . The sets  $A$  and  $B$  are drawn side-by-side, and arrows point from  $a$  to  $b$  whenever  $f(a) = b$ .



**Figure 12.3.** Two ways of drawing the function  $f = \{(p,0),(q,1),(r,2),(s,2)\}$

In general, if  $f : A \rightarrow B$  is the kind of function you may have encountered in algebra or calculus, then conventional graphing techniques offer the best visual description of it. On the other hand, if  $A$  and  $B$  are finite or if we are thinking of them as generic sets, then describing  $f$  with arrows is often a more appropriate way of visualizing it.

We emphasize that, according to Definition 12.1, a function is really just a special kind of set. Any function  $f : A \rightarrow B$  is a subset of  $A \times B$ . By contrast, your calculus text probably defined a function as a certain kind of “rule.” While that intuitive outlook is adequate for the first few semesters of calculus, it does not hold up well to the rigorous mathematical standards necessary for further progress. The problem is that words like “rule” are too vague. Defining a function as a set removes the ambiguity. It makes a function into a concrete mathematical object.

Still, in practice we tend to think of functions as rules. Given  $f : \mathbb{Z} \rightarrow \mathbb{N}$  where  $f(n) = |n| + 2$ , we think of this as a rule that associates any number  $n \in \mathbb{Z}$  to the number  $|n| + 2$  in  $\mathbb{N}$ , rather than a set containing ordered pairs  $(n, |n| + 2)$ . It is only when we have to understand or interpret the theoretical nature of functions (as we do in this text) that Definition 12.1 comes to bear. The definition is a foundation that gives us license to think about functions in a more informal way. For instance, suppose we are discussing a set of functions, such as the set  $S$  of all functions  $\mathbb{R} \rightarrow \mathbb{R}$ . Without Definition 12.1, it would be unclear just what kinds of objects the elements of  $S$  are. But with Definition 12.1, we know exactly what the elements are: each element of  $S$  is a subset  $f \subseteq \mathbb{R} \times \mathbb{R}$ . So we are free to think of  $S$  as a collection of “rules,” and can fall back on Definition 12.1 when greater scrutiny is required.

The next example brings up a point about notation. Consider a function such as  $f : \mathbb{Z}^2 \rightarrow \mathbb{Z}$ , whose domain is a Cartesian product. This function takes as input an ordered pair  $(m, n) \in \mathbb{Z}^2$  and sends it to a number  $f((m, n)) \in \mathbb{Z}$ . To simplify the notation, it is common to write  $f(m, n)$  instead of  $f((m, n))$ , even though this is like writing  $fx$  instead of  $f(x)$ . We also remark that although we've been using the letters  $f$ ,  $g$  and  $h$  to denote functions, any other reasonable symbol could be used. Greek letters such as  $\varphi$  and  $\theta$  are common.

**Example 12.3** Say a function  $\varphi : \mathbb{Z}^2 \rightarrow \mathbb{Z}$  is defined as  $\varphi(m, n) = 6m - 9n$ . Note that as a set, this function is  $\varphi = \{(m, n), 6m - 9n\} : (m, n) \in \mathbb{Z}^2\} \subseteq \mathbb{Z}^2 \times \mathbb{Z}$ . What is the range of  $\varphi$ ?

To answer this, first observe that for any  $(m, n) \in \mathbb{Z}^2$ , the value  $f(m, n) = 6m - 9n = 3(2m - 3n)$  is a multiple of 3. Thus every number in the range is a multiple of 3, so the range is a *subset* of the set of all multiples of 3. On the other hand if  $b = 3k$  is a multiple of 3 we have  $\varphi(-k, -k) = 6(-k) - 9(-k) = 3k = b$ , which means any multiple of 3 is in the range of  $\varphi$ . Therefore the range of  $\varphi$  is the set  $\{3k : k \in \mathbb{Z}\}$  of all multiples of 3.

To conclude this section, let's use Definition 12.1 to help us understand what it means for two functions  $f : A \rightarrow B$  and  $g : C \rightarrow D$  to be equal. The definition says  $f$  and  $g$  are subsets  $f \subseteq A \times B$  and  $g \subseteq C \times D$ . It makes sense to say that  $f$  and  $g$  are equal if  $f = g$ , that is, if they are equal as sets.

Thus the two functions  $f = \{(1, a), (2, a), (3, b)\}$  and  $g = \{(3, b), (2, a), (1, a)\}$  are equal because the sets  $f$  and  $g$  are equal. Notice that the domain of both functions is  $A = \{1, 2, 3\}$ , the set of first elements  $x$  in the ordered pairs  $(x, y) \in f = g$ . In general, equal functions must have equal domains.

Observe also that the equality  $f = g$  means  $f(x) = g(x)$  for every  $x \in A$ . We repackage these ideas in the following definition.

**Definition 12.3** Two functions  $f : A \rightarrow B$  and  $g : A \rightarrow D$  are **equal** if  $f = g$  (as sets). Equivalently,  $f = g$  if and only if  $f(x) = g(x)$  for every  $x \in A$ .

Observe that  $f$  and  $g$  can have different codomains and still be equal. Consider the functions  $f : \mathbb{Z} \rightarrow \mathbb{N}$  and  $g : \mathbb{Z} \rightarrow \mathbb{Z}$  defined as  $f(x) = |x| + 2$  and  $g(x) = |x| + 2$ . Even though their codomains are different, the functions are equal because  $f(x) = g(x)$  for every  $x$  in the domain. If you are bothered that these equal functions have different codomains, recall that we remarked on page 225 that a function's codomain is not really an intrinsic feature of the function, but more a matter of convenience. (Any set that contains the range as a subset is a valid choice of codomain.)

---

**Exercises for Section 12.1**

1. Suppose  $A = \{0, 1, 2, 3, 4\}$ ,  $B = \{2, 3, 4, 5\}$  and  $f = \{(0, 3), (1, 3), (2, 4), (3, 2), (4, 2)\}$ . State the domain and range of  $f$ . Find  $f(2)$  and  $f(1)$ .
  2. Suppose  $A = \{a, b, c, d\}$ ,  $B = \{2, 3, 4, 5, 6\}$  and  $f = \{(a, 2), (b, 3), (c, 4), (d, 5)\}$ . State the domain and range of  $f$ . Find  $f(b)$  and  $f(d)$ .
  3. There are four different functions  $f : \{a, b\} \rightarrow \{0, 1\}$ . List them. Diagrams suffice.
  4. There are eight different functions  $f : \{a, b, c\} \rightarrow \{0, 1\}$ . List them. Diagrams suffice.
  5. Give an example of a relation from  $\{a, b, c, d\}$  to  $\{d, e\}$  that is not a function.
  6. Suppose  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  is defined as  $f = \{(x, 4x + 5) : x \in \mathbb{Z}\}$ . State the domain, codomain and range of  $f$ . Find  $f(10)$ .
  7. Consider the set  $f = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : 3x + y = 4\}$ . Is this a function from  $\mathbb{Z}$  to  $\mathbb{Z}$ ? Explain.
  8. Consider the set  $f = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : x + 3y = 4\}$ . Is this a function from  $\mathbb{Z}$  to  $\mathbb{Z}$ ? Explain.
  9. Consider the set  $f = \{(x^2, x) : x \in \mathbb{R}\}$ . Is this a function from  $\mathbb{R}$  to  $\mathbb{R}$ ? Explain.
  10. Consider the set  $f = \{(x^3, x) : x \in \mathbb{R}\}$ . Is this a function from  $\mathbb{R}$  to  $\mathbb{R}$ ? Explain.
  11. Is the set  $\theta = \{(X, |X|) : X \subseteq \mathbb{Z}_5\}$  a function? If so, what is its domain and range?
  12. Is the set  $\theta = \{(x, y), (3y, 2x, x + y) : x, y \in \mathbb{R}\}$  a function? If so, what is its domain and range? What can be said about the codomain?
- 

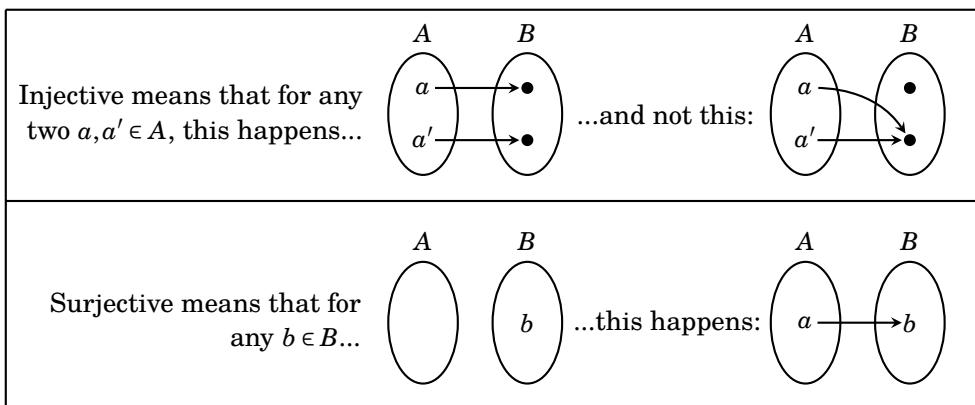
## 12.2 Injective and Surjective Functions

You may recall from algebra and calculus that a function may be *one-to-one* and *onto*, and these properties are related to whether or not the function is invertible. We now review these important ideas. In advanced mathematics, the word *injective* is often used instead of *one-to-one*, and *surjective* is used instead of *onto*. Here are the exact definitions:

**Definition 12.4** A function  $f : A \rightarrow B$  is:

1. **injective** (or *one-to-one*) if for all  $a, a' \in A$ ,  $a \neq a'$  implies  $f(a) \neq f(a')$ ;
2. **surjective** (or *onto B*) if for every  $b \in B$  there is an  $a \in A$  with  $f(a) = b$ ;
3. **bijective** if  $f$  is both injective and surjective.

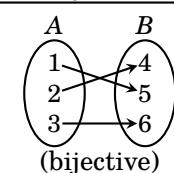
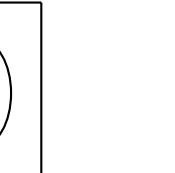
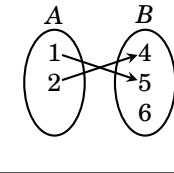
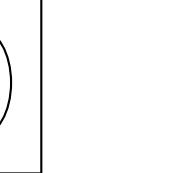
Below is a visual description of Definition 12.4. In essence, injective means that unequal elements in  $A$  always get sent to unequal elements in  $B$ . Surjective means that every element of  $B$  has an arrow pointing to it, that is, it equals  $f(a)$  for some  $a$  in the domain of  $f$ .



For more concrete examples, consider the following functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$ . The function  $f(x) = x^2$  is not injective because  $-2 \neq 2$ , but  $f(-2) = f(2)$ . Nor is it surjective, for if  $b = -1$  (or if  $b$  is any negative number), then there is no  $a \in \mathbb{R}$  with  $f(a) = b$ . On the other hand,  $g(x) = x^3$  is both injective and surjective, so it is also bijective.

Notice that whether or not  $f$  is surjective depends on its codomain. For example,  $f(x) = x^2$  is not surjective as a function  $\mathbb{R} \rightarrow \mathbb{R}$ , but it *is* surjective as a function  $\mathbb{R} \rightarrow [0, \infty)$ . When we speak of a function being surjective, we always have in mind a particular codomain.

There are four possible injective/surjective combinations that a function may possess. This is illustrated below for four functions  $A \rightarrow B$ . Functions in the first column are injective, those in the second column are not injective. Functions in the first row are surjective, those in the second row are not.

	Injective	Not injective
Surjective	 (bijective)	
Not surjective		

We note in passing that, according to the definitions, a function is surjective if and only if its codomain equals its range.

Often it is necessary to prove that a particular function  $f : A \rightarrow B$  is injective. For this, Definition 12.4 says we must prove that for any two elements  $a, a' \in A$ , the conditional statement  $(a \neq a') \Rightarrow (f(a) \neq f(a'))$  is true. The two main approaches for this are summarized below.

**How to show a function  $f : A \rightarrow B$  is injective:**

**Direct approach:**

Suppose  $a, a' \in A$  and  $a \neq a'$ .  
 $\vdots$   
Therefore  $f(a) \neq f(a')$ .

**Contrapositive approach:**

Suppose  $a, a' \in A$  and  $f(a) = f(a')$ .  
 $\vdots$   
Therefore  $a = a'$ .

Of these two approaches, the contrapositive is often the easiest to use, especially if  $f$  is defined by an algebraic formula. This is because the contrapositive approach starts with the *equation*  $f(a) = f(a')$  and proceeds to the *equation*  $a = a'$ . In algebra, as you know, it is usually easier to work with equations than inequalities.

To prove that a function is *not* injective, you must *disprove* the statement  $(a \neq a') \Rightarrow (f(a) \neq f(a'))$ . For this it suffices to find example of two elements  $a, a' \in A$  for which  $a \neq a'$  and  $f(a) = f(a')$ .

Next we examine how to prove that  $f : A \rightarrow B$  is *surjective*. According to Definition 12.4, we must prove the statement  $\forall b \in B, \exists a \in A, f(a) = b$ . In words, we must show that for any  $b \in B$ , there is at least one  $a \in A$  (which may depend on  $b$ ) having the property that  $f(a) = b$ . Here is an outline:

**How to show a function  $f : A \rightarrow B$  is surjective:**

Suppose  $b \in B$ .  
[Prove there exists  $a \in A$  for which  $f(a) = b$ .]

The second line involves proving the existence of an  $a$  for which  $f(a) = b$ . For this, just finding an example of such an  $a$  would suffice. (How to find such an example depends on how  $f$  is defined. If  $f$  is given as a formula, we may be able to find  $a$  by solving the equation  $f(a) = b$  for  $a$ . Sometimes you can find  $a$  by just plain common sense.) To show  $f$  is *not* surjective, we must prove the negation of  $\forall b \in B, \exists a \in A, f(a) = b$ , that is, we must prove  $\exists b \in B, \forall a \in A, f(a) \neq b$ .

The following examples illustrate these ideas. (For the first example, note that the set  $\mathbb{R} - \{0\}$  is  $\mathbb{R}$  with the number 0 removed.)

**Example 12.4** Show that the function  $f : \mathbb{R} - \{0\} \rightarrow \mathbb{R}$  defined as  $f(x) = \frac{1}{x} + 1$  is injective but not surjective.

We will use the contrapositive approach to show that  $f$  is injective. Suppose  $a, a' \in \mathbb{R} - \{0\}$  and  $f(a) = f(a')$ . This means  $\frac{1}{a} + 1 = \frac{1}{a'} + 1$ . Subtracting 1 from both sides and inverting produces  $a = a'$ . Therefore  $f$  is injective.

The function  $f$  is not surjective because there exists an element  $b = 1 \in \mathbb{R}$  for which  $f(x) = \frac{1}{x} + 1 \neq 1$  for every  $x \in \mathbb{R} - \{0\}$ .

**Example 12.5** Show that the function  $f : \mathbb{R} - \{0\} \rightarrow \mathbb{R} - \{1\}$  where  $f(x) = \frac{1}{x} + 1$  is injective and surjective (hence bijective).

This is just like the previous example, except that the codomain has been changed. The previous example shows  $f$  is injective. To show that it is surjective, take an arbitrary  $b \in \mathbb{R} - \{1\}$ . We seek an  $a \in \mathbb{R} - \{0\}$  for which  $f(a) = b$ , that is, for which  $\frac{1}{a} + 1 = b$ . Solving for  $a$  gives  $a = \frac{1}{b-1}$ , which is defined because  $b \neq 1$ . In summary, for any  $b \in \mathbb{R} - \{1\}$ , we have  $f(\frac{1}{b-1}) = b$ , so  $f$  is surjective.

**Example 12.6** Show that the function  $g : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined by the formula  $g(m, n) = (m + n, m + 2n)$ , is both injective and surjective.

We will use the contrapositive approach to show that  $g$  is injective. Thus we need to show that  $g(m, n) = g(k, \ell)$  implies  $(m, n) = (k, \ell)$ . Suppose  $(m, n), (k, \ell) \in \mathbb{Z} \times \mathbb{Z}$  and  $g(m, n) = g(k, \ell)$ . Then  $(m + n, m + 2n) = (k + \ell, k + 2\ell)$ . It follows that  $m + n = k + \ell$  and  $m + 2n = k + 2\ell$ . Subtracting the first equation from the second gives  $n = \ell$ . Next, subtract  $n = \ell$  from  $m + n = k + \ell$  to get  $m = k$ . Since  $m = k$  and  $n = \ell$ , it follows that  $(m, n) = (k, \ell)$ . Thus  $g$  is injective.

To see that  $g$  is surjective, consider an arbitrary element  $(b, c) \in \mathbb{Z} \times \mathbb{Z}$ . We need to show that there is some  $(x, y) \in \mathbb{Z} \times \mathbb{Z}$  for which  $g(x, y) = (b, c)$ . To find  $(x, y)$ , note that  $g(x, y) = (b, c)$  means  $(x + y, x + 2y) = (b, c)$ . This leads to the following system of equations:

$$\begin{aligned} x + y &= b \\ x + 2y &= c. \end{aligned}$$

Solving gives  $x = 2b - c$  and  $y = c - b$ . Then  $(x, y) = (2b - c, c - b)$ . We now have  $g(2b - c, c - b) = (b, c)$ , and it follows that  $g$  is surjective.

**Example 12.7** Consider function  $h : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Q}$  defined as  $h(m, n) = \frac{m}{|n| + 1}$ . Determine whether this is injective and whether it is surjective.

This function is *not* injective because of the unequal elements  $(1, 2)$  and  $(1, -2)$  in  $\mathbb{Z} \times \mathbb{Z}$  for which  $h(1, 2) = h(1, -2) = \frac{1}{3}$ . However,  $h$  is surjective: Take any element  $b \in \mathbb{Q}$ . Then  $b = \frac{c}{d}$  for some  $c, d \in \mathbb{Z}$ . Notice we may assume  $d$  is positive by making  $c$  negative, if necessary. Then  $h(c, d - 1) = \frac{c}{|d-1|+1} = \frac{c}{d} = b$ .

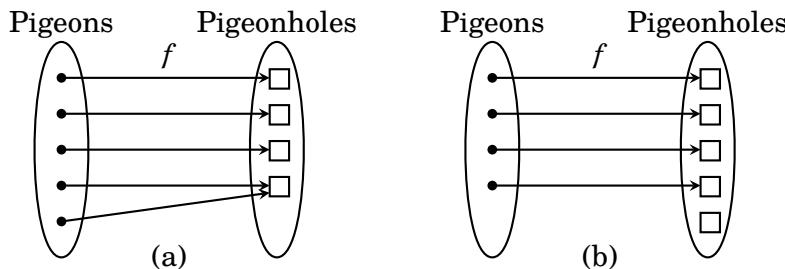
### Exercises for Section 12.2

1. Let  $A = \{1, 2, 3, 4\}$  and  $B = \{a, b, c\}$ . Give an example of a function  $f : A \rightarrow B$  that is neither injective nor surjective.
2. Consider the logarithm function  $\ln : (0, \infty) \rightarrow \mathbb{R}$ . Decide whether this function is injective and whether it is surjective.
3. Consider the cosine function  $\cos : \mathbb{R} \rightarrow \mathbb{R}$ . Decide whether this function is injective and whether it is surjective. What if it had been defined as  $\cos : \mathbb{R} \rightarrow [-1, 1]$ ?
4. A function  $f : \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  is defined as  $f(n) = (2n, n+3)$ . Verify whether this function is injective and whether it is surjective.
5. A function  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  is defined as  $f(n) = 2n + 1$ . Verify whether this function is injective and whether it is surjective.
6. A function  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$  is defined as  $f(m, n) = 3n - 4m$ . Verify whether this function is injective and whether it is surjective.
7. A function  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$  is defined as  $f(m, n) = 2n - 4m$ . Verify whether this function is injective and whether it is surjective.
8. A function  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  is defined as  $f(m, n) = (m+n, 2m+n)$ . Verify whether this function is injective and whether it is surjective.
9. Prove that the function  $f : \mathbb{R} - \{2\} \rightarrow \mathbb{R} - \{5\}$  defined by  $f(x) = \frac{5x+1}{x-2}$  is bijective.
10. Prove the function  $f : \mathbb{R} - \{1\} \rightarrow \mathbb{R} - \{1\}$  defined by  $f(x) = \left(\frac{x+1}{x-1}\right)^3$  is bijective.
11. Consider the function  $\theta : \{0, 1\} \times \mathbb{N} \rightarrow \mathbb{Z}$  defined as  $\theta(a, b) = (-1)^a b$ . Is  $\theta$  injective? Is it surjective? Bijective? Explain.
12. Consider the function  $\theta : \{0, 1\} \times \mathbb{N} \rightarrow \mathbb{Z}$  defined as  $\theta(a, b) = a - 2ab + b$ . Is  $\theta$  injective? Is it surjective? Bijective? Explain.
13. Consider the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  defined by the formula  $f(x, y) = (xy, x^3)$ . Is  $f$  injective? Is it surjective? Bijective? Explain.
14. Consider the function  $\theta : \mathcal{P}(\mathbb{Z}) \rightarrow \mathcal{P}(\mathbb{Z})$  defined as  $\theta(X) = \overline{X}$ . Is  $\theta$  injective? Is it surjective? Bijective? Explain.
15. This question concerns functions  $f : \{A, B, C, D, E, F, G\} \rightarrow \{1, 2, 3, 4, 5, 6, 7\}$ . How many such functions are there? How many of these functions are injective? How many are surjective? How many are bijective?
16. This question concerns functions  $f : \{A, B, C, D, E\} \rightarrow \{1, 2, 3, 4, 5, 6, 7\}$ . How many such functions are there? How many of these functions are injective? How many are surjective? How many are bijective?
17. This question concerns functions  $f : \{A, B, C, D, E, F, G\} \rightarrow \{1, 2\}$ . How many such functions are there? How many of these functions are injective? How many are surjective? How many are bijective?
18. Prove that the function  $f : \mathbb{N} \rightarrow \mathbb{Z}$  defined as  $f(n) = \frac{(-1)^n(2n-1)+1}{4}$  is bijective.

### 12.3 The Pigeonhole Principle Revisited

We first encountered a result called the *pigeonhole principle* in Section 3.9. It turns out that the pigeonhole principle has a useful phrasing in the language of injective and surjective functions, and we now discuss this. Our discussion will not use any material from Chapter 3, so it does not matter if you skipped it.

The pigeonhole principle is motivated by a simple thought experiment: Imagine there is a set  $A$  of pigeons and a set  $B$  of pigeonholes, and all the pigeons fly into the pigeonholes. You can think of this as describing a function  $f : A \rightarrow B$ , where pigeon  $p$  flies into pigeonhole  $f(p)$ . See Figure 12.4.



**Figure 12.4.** The pigeonhole principle

In Figure 12.4(a) there are more pigeons than pigeonholes, and it is obvious that in such a case at least two pigeons have to fly into the same pigeonhole, meaning that  $f$  is not injective. In Figure 12.4(b) there are fewer pigeons than pigeonholes, so clearly at least one pigeonhole remains empty, meaning that  $f$  is not surjective.

Although the underlying idea expressed by these figures has little to do with pigeons, it is nonetheless called the *pigeonhole principle*:

**The Pigeonhole Principle** (function version)

Suppose  $A$  and  $B$  are finite sets and  $f : A \rightarrow B$  is any function.

1. If  $|A| > |B|$ , then  $f$  is not injective.
2. If  $|A| < |B|$ , then  $f$  is not surjective.

Though the pigeonhole principle is obvious, it can be used to prove some things that are not so obvious. Two examples follow.

**Proposition** If  $A$  is any set of 10 integers between 1 and 100, then there exist two different subsets  $X \subseteq A$  and  $Y \subseteq A$  for which the sum of elements in  $X$  equals the sum of elements in  $Y$ .

To illustrate what this proposition is saying, consider the random set

$$A = \{5, 7, 12, 11, 17, 50, 51, 80, 90, 100\}$$

of 10 integers between 1 and 100. Notice that  $A$  has subsets  $X = \{5, 80\}$  and  $Y = \{7, 11, 17, 50\}$  for which the sum of the elements in  $X$  equals the sum of those in  $Y$ . If we tried to “mess up”  $A$  by changing the 5 to a 6, we get

$$A = \{6, 7, 12, 11, 17, 50, 51, 80, 90, 100\}$$

which has subsets  $X = \{7, 12, 17, 50\}$  and  $Y = \{6, 80\}$  both of whose elements add up to the same number (86). The proposition asserts that this is always possible, no matter what  $A$  is. Here is a proof:

*Proof.* Suppose  $A \subseteq \{1, 2, 3, 4, \dots, 99, 100\}$  and  $|A| = 10$ , as stated. Notice that if  $X \subseteq A$ , then  $X$  has no more than 10 elements, each between 1 and 100, and therefore the sum of all the elements of  $X$  is less than  $100 \cdot 10 = 1000$ . Consider the function

$$f : \mathcal{P}(A) \rightarrow \{0, 1, 2, 3, 4, \dots, 1000\},$$

where  $f(X)$  is the sum of the elements in  $X$ . (Examples:  $f(\{3, 7, 50\}) = 60$ ;  $f(\{1, 70, 80, 95\}) = 246$ .) As  $|\mathcal{P}(A)| = 2^{10} = 1024 > 1001 = |\{0, 1, 2, 3, \dots, 1000\}|$ , it follows from the pigeonhole principle that  $f$  is not injective. Therefore there are two unequal sets  $X, Y \in \mathcal{P}(A)$  for which  $f(X) = f(Y)$ . In other words, there are subsets  $X \subseteq A$  and  $Y \subseteq A$  for which the sum of elements in  $X$  equals the sum of elements in  $Y$ . ■

**Proposition** There are at least two Texans with the same number of hairs on their heads.

*Proof.* We will use two facts. First, the population of Texas is more than twenty million. Second, it is a biological fact that every human head has fewer than one million hairs. Let  $A$  be the set of all Texans, and let  $B = \{0, 1, 2, 3, 4, \dots, 1000000\}$ . Let  $f : A \rightarrow B$  be the function for which  $f(x)$  equals the number of hairs on the head of  $x$ . Since  $|A| > |B|$ , the pigeonhole principle asserts that  $f$  is not injective. Thus there are two Texans  $x$  and  $y$  for whom  $f(x) = f(y)$ , meaning that they have the same number of hairs on their heads. ■

Proofs that use the pigeonhole principle tend to be inherently non-constructive, in the sense discussed in Section 7.4. For example, the above proof does not explicitly give us two Texans with the same number of hairs on their heads; it only shows that two such people exist. If we were to make a constructive proof, we could find examples of two bald Texans. Then they have the same number of head hairs, namely zero.

---

### Exercises for Section 12.3

1. Prove that if six numbers are chosen at random, then at least two of them will have the same remainder when divided by 5.
  2. Prove that if  $a$  is a natural number, then there exist two unequal natural numbers  $k$  and  $\ell$  for which  $a^k - a^\ell$  is divisible by 10.
  3. Prove that if six natural numbers are chosen at random, then the sum or difference of two of them is divisible by 9.
  4. Consider a square whose side-length is one unit. Select any five points from inside this square. Prove that at least two of these points are within  $\frac{\sqrt{2}}{2}$  units of each other.
  5. Prove that any set of seven distinct natural numbers contains a pair of numbers whose sum or difference is divisible by 10.
  6. Given a sphere  $S$ , a *great circle* of  $S$  is the intersection of  $S$  with a plane through its center. Every great circle divides  $S$  into two parts. A *hemisphere* is the union of the great circle and one of these two parts. Prove that if five points are placed arbitrarily on  $S$ , then there is a hemisphere that contains four of them.
  7. Prove or disprove: Any subset  $X \subseteq \{1, 2, 3, \dots, 2n\}$  with  $|X| > n$  contains two (unequal) elements  $a, b \in X$  for which  $a \mid b$  or  $b \mid a$ .
- 

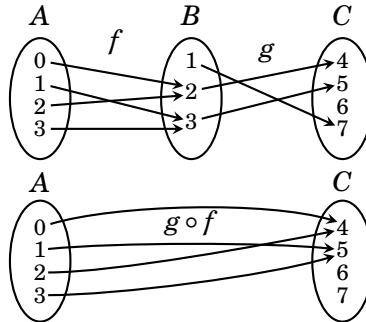
## 12.4 Composition

You are familiar with the notion of function composition from algebra and calculus. Still, it is worthwhile to revisit it now with our more sophisticated ideas about functions.

**Definition 12.5** Suppose  $f : A \rightarrow B$  and  $g : B \rightarrow C$  are functions with the property that the codomain of  $f$  equals the domain of  $g$ . The **composition** of  $f$  with  $g$  is another function, denoted as  $g \circ f$  and defined as follows: If  $x \in A$ , then  $g \circ f(x) = g(f(x))$ . Therefore  $g \circ f$  sends elements of  $A$  to elements of  $C$ , so  $g \circ f : A \rightarrow C$ .

The following figure illustrates the definition. Here  $f : A \rightarrow B$ ,  $g : B \rightarrow C$ , and  $g \circ f : A \rightarrow C$ . We have, for example,  $g \circ f(0) = g(f(0)) = g(2) = 4$ . Be very

careful with the order of the symbols. Even though  $g$  comes first in the symbol  $g \circ f$ , we work out  $g \circ f(x)$  as  $g(f(x))$ , with  $f$  acting on  $x$  first, followed by  $g$  acting on  $f(x)$ .



**Figure 12.5.** Composition of two functions

Notice that the composition  $g \circ f$  also makes sense if the range of  $f$  is a *subset* of the domain of  $g$ . You should take note of this fact, but to keep matters simple we will continue to emphasize situations where the codomain of  $f$  equals the domain of  $g$ .

**Example 12.8** Suppose  $A = \{a, b, c\}$ ,  $B = \{0, 1\}$ ,  $C = \{1, 2, 3\}$ . Let  $f : A \rightarrow B$  be the function  $f = \{(a, 0), (b, 1), (c, 0)\}$ , and let  $g : B \rightarrow C$  be  $g = \{(0, 3), (1, 1)\}$ . Then  $g \circ f = \{(a, 3), (b, 1), (c, 3)\}$ .

**Example 12.9** Say  $A = \{a, b, c\}$ ,  $B = \{0, 1\}$ ,  $C = \{1, 2, 3\}$ . Let  $f : A \rightarrow B$  be the function  $f = \{(a, 0), (b, 1), (c, 0)\}$ , and let  $g : C \rightarrow B$  be the function  $g = \{(1, 0), (2, 1), (3, 1)\}$ . In this situation the composition  $g \circ f$  is not defined because the codomain  $B$  of  $f$  is not the same set as the domain  $C$  of  $g$ . Remember: In order for  $g \circ f$  to make sense, the codomain of  $f$  must equal the domain of  $g$ . (Or at least be a subset of it.)

**Example 12.10** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be defined as  $f(x) = x^2 + x$ , and  $g : \mathbb{R} \rightarrow \mathbb{R}$  be defined as  $g(x) = x + 1$ . Then  $g \circ f : \mathbb{R} \rightarrow \mathbb{R}$  is the function defined by the formula  $g \circ f(x) = g(f(x)) = g(x^2 + x) = x^2 + x + 1$ .

Since the domains and codomains of  $g$  and  $f$  are the same, we can in this case do a composition in the other order. Note that  $f \circ g : \mathbb{R} \rightarrow \mathbb{R}$  is the function defined as  $f \circ g(x) = f(g(x)) = f(x + 1) = (x + 1)^2 + (x + 1) = x^2 + 3x + 2$ .

This example illustrates that even when  $g \circ f$  and  $f \circ g$  are both defined, they are not necessarily equal. We can express this fact by saying *function composition is not commutative*.

We close this section by proving several facts about composition that you are likely to encounter in your future study of mathematics. First, we note that, although it is not commutative, function composition *is* associative.

**Theorem 12.1** Composition of functions is associative. That is if  $f : A \rightarrow B$ ,  $g : B \rightarrow C$  and  $h : C \rightarrow D$ , then  $(h \circ g) \circ f = h \circ (g \circ f)$ .

*Proof.* Suppose  $f, g, h$  are as stated. It follows from Definition 12.5 that both  $(h \circ g) \circ f$  and  $h \circ (g \circ f)$  are functions from  $A$  to  $D$ . To show that they are equal, we just need to show

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

for every  $x \in A$ . Note that Definition 12.5 yields

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

Also

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

Thus

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

as both sides equal  $h(g(f(x)))$ . ■

**Theorem 12.2** Suppose  $f : A \rightarrow B$  and  $g : B \rightarrow C$ . If both  $f$  and  $g$  are injective, then  $g \circ f$  is injective. If both  $f$  and  $g$  are surjective, then  $g \circ f$  is surjective.

*Proof.* First suppose both  $f$  and  $g$  are injective. To see that  $g \circ f$  is injective, we must show that  $g \circ f(x) = g \circ f(y)$  implies  $x = y$ . Suppose  $g \circ f(x) = g \circ f(y)$ . This means  $g(f(x)) = g(f(y))$ . It follows that  $f(x) = f(y)$ . (For otherwise  $g$  wouldn't be injective.) But since  $f(x) = f(y)$  and  $f$  is injective, it must be that  $x = y$ . Therefore  $g \circ f$  is injective.

Next suppose both  $f$  and  $g$  are surjective. To see that  $g \circ f$  is surjective, we must show that for any element  $c \in C$ , there is a corresponding element  $a \in A$  for which  $g \circ f(a) = c$ . Thus consider an arbitrary  $c \in C$ . Because  $g$  is surjective, there is an element  $b \in B$  for which  $g(b) = c$ . And because  $f$  is surjective, there is an element  $a \in A$  for which  $f(a) = b$ . Therefore  $g(f(a)) = g(b) = c$ , which means  $g \circ f(a) = c$ . Thus  $g \circ f$  is surjective. ■

### Exercises for Section 12.4

1. Suppose  $A = \{5, 6, 8\}$ ,  $B = \{0, 1\}$ ,  $C = \{1, 2, 3\}$ . Let  $f : A \rightarrow B$  be the function  $f = \{(5, 1), (6, 0), (8, 1)\}$ , and  $g : B \rightarrow C$  be  $g = \{(0, 1), (1, 1)\}$ . Find  $g \circ f$ .
2. Suppose  $A = \{1, 2, 3, 4\}$ ,  $B = \{0, 1, 2\}$ ,  $C = \{1, 2, 3\}$ . Let  $f : A \rightarrow B$  be  $f = \{(1, 0), (2, 1), (3, 2), (4, 0)\}$ , and  $g : B \rightarrow C$  be  $g = \{(0, 1), (1, 1), (2, 3)\}$ . Find  $g \circ f$ .
3. Suppose  $A = \{1, 2, 3\}$ . Let  $f : A \rightarrow A$  be the function  $f = \{(1, 2), (2, 2), (3, 1)\}$ , and let  $g : A \rightarrow A$  be the function  $g = \{(1, 3), (2, 1), (3, 2)\}$ . Find  $g \circ f$  and  $f \circ g$ .
4. Suppose  $A = \{a, b, c\}$ . Let  $f : A \rightarrow A$  be the function  $f = \{(a, c), (b, c), (c, c)\}$ , and let  $g : A \rightarrow A$  be the function  $g = \{(a, a), (b, b), (c, a)\}$ . Find  $g \circ f$  and  $f \circ g$ .
5. Consider the functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = \sqrt[3]{x+1}$  and  $g(x) = x^3$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
6. Consider the functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = \frac{1}{x^2+1}$  and  $g(x) = 3x + 2$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
7. Consider the functions  $f, g : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined as  $f(m, n) = (mn, m^2)$  and  $g(m, n) = (m+1, m+n)$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
8. Consider the functions  $f, g : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined as  $f(m, n) = (3m - 4n, 2m + n)$  and  $g(m, n) = (5m + n, m)$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
9. Consider the functions  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$  defined as  $f(m, n) = m + n$  and  $g : \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined as  $g(m) = (m, m)$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
10. Consider the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  defined by the formula  $f(x, y) = (xy, x^3)$ . Find a formula for  $f \circ f$ .

### 12.5 Inverse Functions

You may recall from calculus that if a function  $f$  is injective and surjective, then it has an inverse function  $f^{-1}$  that “undoes” the effect of  $f$  in the sense that  $f^{-1}(f(x)) = x$  for every  $x$  in the domain. (For example, if  $f(x) = x^3$ , then  $f^{-1}(x) = \sqrt[3]{x}$ .) We now review these ideas. Our approach uses two ingredients, outlined in the following definitions.

**Definition 12.6** For a set  $A$ , the **identity function** on  $A$  is the function  $i_A : A \rightarrow A$  defined as  $i_A(x) = x$  for every  $x \in A$ .

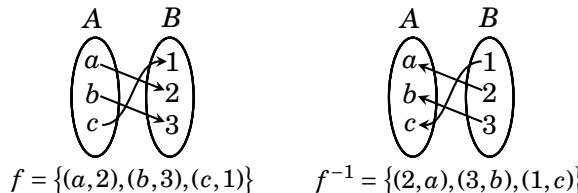
**Example:** If  $A = \{1, 2, 3\}$ , then  $i_A = \{(1, 1), (2, 2), (3, 3)\}$ . Also  $i_{\mathbb{Z}} = \{(n, n) : n \in \mathbb{Z}\}$ . The identity function on a set is the function that sends any element of the set to itself.

Notice that for any set  $A$ , the identity function  $i_A$  is bijective: It is injective because  $i_A(x) = i_A(y)$  immediately reduces to  $x = y$ . It is surjective

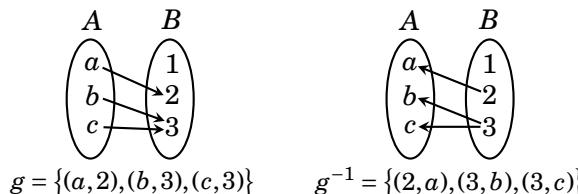
because if we take any element  $b$  in the codomain  $A$ , then  $b$  is also in the domain  $A$ , and  $i_A(b) = b$ .

**Definition 12.7** Given a relation  $R$  from  $A$  to  $B$ , the **inverse relation of  $R$**  is the relation from  $B$  to  $A$  defined as  $R^{-1} = \{(y, x) : (x, y) \in R\}$ . In other words, the inverse of  $R$  is the relation  $R^{-1}$  obtained by interchanging the elements in every ordered pair in  $R$ .

For example, let  $A = \{a, b, c\}$  and  $B = \{1, 2, 3\}$ , and suppose  $f$  is the relation  $f = \{(a, 2), (b, 3), (c, 1)\}$  from  $A$  to  $B$ . Then  $f^{-1} = \{(2, a), (3, b), (1, c)\}$  and this is a relation from  $B$  to  $A$ . Notice that  $f$  is actually a function from  $A$  to  $B$ , and  $f^{-1}$  is a function from  $B$  to  $A$ . These two relations are drawn below. Notice the drawing for relation  $f^{-1}$  is just the drawing for  $f$  with arrows reversed.



For another example, let  $A$  and  $B$  be the same sets as above, but consider the relation  $g = \{(a, 2), (b, 3), (c, 3)\}$  from  $A$  to  $B$ . Then  $g^{-1} = \{(2, a), (3, b), (3, c)\}$  is a relation from  $B$  to  $A$ . These two relations are sketched below.



This time, even though the relation  $g$  is a function, its inverse  $g^{-1}$  is not a function because the element 3 occurs twice as a first coordinate of an ordered pair in  $g^{-1}$ .

In the above examples, relations  $f$  and  $g$  are both functions, and  $f^{-1}$  is a function and  $g^{-1}$  is not. This raises a question: What properties does  $f$  have and  $g$  lack that makes  $f^{-1}$  a function and  $g^{-1}$  not a function? The answer is not hard to see. Function  $g$  is not injective because  $g(b) = g(c) = 3$ , and thus  $(b, 3)$  and  $(c, 3)$  are both in  $g$ . This causes a problem with  $g^{-1}$  because it means  $(3, b)$  and  $(3, c)$  are both in  $g^{-1}$ , so  $g^{-1}$  can't be a function. Thus, in order for  $g^{-1}$  to be a function, it would be necessary that  $g$  be injective.

But that is not enough. Function  $g$  also fails to be surjective because no element of  $A$  is sent to the element  $1 \in B$ . This means  $g^{-1}$  contains no ordered pair whose first coordinate is 1, so it can't be a function from  $B$  to  $A$ . If  $g^{-1}$  were to be a function it would be necessary that  $g$  be surjective.

The previous two paragraphs suggest that if  $g$  is a function, then it must be bijective in order for its inverse relation  $g^{-1}$  to be a function. Indeed, this is easy to verify. Conversely, if a function is bijective, then its inverse relation is easily seen to be a function. We summarize this in the following theorem.

**Theorem 12.3** Let  $f : A \rightarrow B$  be a function. Then  $f$  is bijective if and only if the inverse relation  $f^{-1}$  is a function from  $B$  to  $A$ .

Suppose  $f : A \rightarrow B$  is bijective, so according to the theorem  $f^{-1}$  is a function. Observe that the relation  $f$  contains all the pairs  $(x, f(x))$  for  $x \in A$ , so  $f^{-1}$  contains all the pairs  $(f(x), x)$ . But  $(f(x), x) \in f^{-1}$  means  $f^{-1}(f(x)) = x$ . Therefore  $f^{-1} \circ f(x) = x$  for every  $x \in A$ . From this we get  $f^{-1} \circ f = i_A$ . Similar reasoning produces  $f \circ f^{-1} = i_B$ . This leads to the following definitions.

**Definition 12.8** If  $f : A \rightarrow B$  is bijective then its **inverse** is the function  $f^{-1} : B \rightarrow A$ . The functions  $f$  and  $f^{-1}$  obey the equations  $f^{-1} \circ f = i_A$  and  $f \circ f^{-1} = i_B$ .

You probably recall from algebra at least one technique for computing the inverse of a bijective function  $f$ . To find  $f^{-1}$ , start with the equation  $y = f(x)$ . Then interchange variables to get  $x = f(y)$ . Solving this equation for  $y$  (if possible) produces  $y = f^{-1}(x)$ . The next two examples illustrate this.

**Example 12.11** The function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = x^3 + 1$  is bijective. Find its inverse.

We begin by writing  $y = x^3 + 1$ . Now interchange variables to obtain  $x = y^3 + 1$ . Solving for  $y$  produces  $y = \sqrt[3]{x - 1}$ . Thus

$$f^{-1}(x) = \sqrt[3]{x - 1}.$$

(You can check your answer by computing

$$f^{-1}(f(x)) = \sqrt[3]{f(x) - 1} = \sqrt[3]{x^3 + 1 - 1} = x.$$

Therefore  $f^{-1}(f(x)) = x$ . Any answer other than  $x$  indicates a mistake.)

**Example 12.12** Example 12.6 showed that the function  $g : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined by the formula  $g(m, n) = (m + n, m + 2n)$  is bijective. Find its inverse.

The approach outlined above should work, but we need to be careful to keep track of coordinates in  $\mathbb{Z} \times \mathbb{Z}$ . We begin by writing  $(x, y) = g(m, n)$ , then interchanging the variables  $(x, y)$  and  $(m, n)$  to get  $(m, n) = g(x, y)$ . This gives

$$(m, n) = (x + y, x + 2y),$$

from which we get the following system of equations:

$$\begin{aligned} x + y &= m \\ x + 2y &= n. \end{aligned}$$

Solving this system, we get

$$\begin{aligned} x &= 2m - n \\ y &= n - m. \end{aligned}$$

Then  $(x, y) = (2m - n, n - m)$ , so  $\boxed{g^{-1}(m, n) = (2m - n, n - m)}$ .

We can check this by confirming  $g^{-1}(g(m, n)) = (m, n)$ . Doing the math,

$$\begin{aligned} g^{-1}(g(m, n)) &= g^{-1}(m + n, m + 2n) \\ &= (2(m + n) - (m + 2n), (m + 2n) - (m + n)) \\ &= (m, n). \end{aligned}$$

### Exercises for Section 12.5

- Check that  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  defined by  $f(n) = 6 - n$  is bijective. Then compute  $f^{-1}$ .
- In Exercise 9 of Section 12.2 you proved that  $f : \mathbb{R} - \{2\} \rightarrow \mathbb{R} - \{5\}$  defined by  $f(x) = \frac{5x+1}{x-2}$  is bijective. Now find its inverse.
- Let  $B = \{2^n : n \in \mathbb{Z}\} = \{\dots, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, \dots\}$ . Show that the function  $f : \mathbb{Z} \rightarrow B$  defined as  $f(n) = 2^n$  is bijective. Then find  $f^{-1}$ .
- The function  $f : \mathbb{R} \rightarrow (0, \infty)$  defined as  $f(x) = e^{x^3+1}$  is bijective. Find its inverse.
- The function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = \pi x - e$  is bijective. Find its inverse.
- The function  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined by the formula  $f(m, n) = (5m + 4n, 4m + 3n)$  is bijective. Find its inverse.
- Show that the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  defined by the formula  $f(x, y) = ((x^2 + 1)y, x^3)$  is bijective. Then find its inverse.
- Is the function  $\theta : \mathcal{P}(\mathbb{Z}) \rightarrow \mathcal{P}(\mathbb{Z})$  defined as  $\theta(X) = \overline{X}$  bijective? If so, find  $\theta^{-1}$ .
- Consider the function  $f : \mathbb{R} \times \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{R}$  defined as  $f(x, y) = (y, 3xy)$ . Check that this is bijective; find its inverse.
- Consider  $f : \mathbb{N} \rightarrow \mathbb{Z}$  defined as  $f(n) = \frac{(-1)^n(2n-1)+1}{4}$ . This function is bijective by Exercise 18 in Section 12.2. Find its inverse.

## 12.6 Image and Preimage

It is time to take up a matter of notation that you will encounter in future mathematics classes. Suppose we have a function  $f : A \rightarrow B$ . If  $X \subseteq A$ , the expression  $f(X)$  has a special meaning. It stands for the set  $\{f(x) : x \in X\}$ . And if  $Y \subseteq B$ , then  $f^{-1}(Y)$  has a meaning *even if f is not invertible*: it stands for the set  $\{x \in A : f(x) \in Y\}$ . Here are the precise definitions.

**Definition 12.9** Suppose  $f : A \rightarrow B$  is a function.

1. If  $X \subseteq A$ , the **image** of  $X$  is the set  $f(X) = \{f(x) : x \in X\} \subseteq B$ .
2. If  $Y \subseteq B$ , the **preimage** of  $Y$  is the set  $f^{-1}(Y) = \{x \in A : f(x) \in Y\} \subseteq A$ .

In words, the image  $f(X)$  of  $X$  is the set of all things in  $B$  that  $f$  sends elements of  $X$  to. (Roughly speaking, you might think of  $f(X)$  as a kind of distorted “copy” or “image” of  $X$  in  $B$ .) The preimage  $f^{-1}(Y)$  of  $Y$  is the set of all things in  $A$  that  $f$  sends into  $Y$ .

Maybe you have already encountered these ideas in linear algebra, in a setting involving a linear transformation  $T : V \rightarrow W$  between two vector spaces. If  $X \subseteq V$  is a subspace of  $V$ , then its image  $T(X)$  is a subspace of  $W$ . If  $Y \subseteq W$  is a subspace of  $W$ , then its preimage  $T^{-1}(Y)$  is a subspace of  $V$ . (If this does not sound familiar, then ignore it.)

**Example 12.13** Let  $f : \{s, t, u, v, w, x, y, z\} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$  be

$$f = \{(s, 4), (t, 8), (u, 8), (v, 1), (w, 2), (x, 4), (y, 6), (z, 4)\}.$$

This  $f$  is neither injective nor surjective, so it certainly is not invertible. Be sure you understand the following statements.

- |   |   |
|---|---|
| 1. $f(\{s, t, u, z\}) = \{8, 4\}$       | 5. $f^{-1}(\{4\}) = \{s, x, z\}$                |
| 2. $f(\{s, x, z\}) = \{4\}$             | 6. $f^{-1}(\{4, 9\}) = \{s, x, z\}$             |
| 3. $f(\{s, v, w, y\}) = \{1, 2, 4, 6\}$ | 7. $f^{-1}(\{9\}) = \emptyset$                  |
| 4. $f(\emptyset) = \emptyset$           | 8. $f^{-1}(\{1, 4, 8\}) = \{s, t, u, v, x, z\}$ |

It is important to realize that the  $X$  and  $Y$  in Definition 12.9 are subsets (not elements!) of  $A$  and  $B$ . In Example 12.13 we had  $f^{-1}(\{4\}) = \{s, x, z\}$ , while  $f^{-1}(4)$  is meaningless because the inverse function  $f^{-1}$  does not exist. And there is a subtle difference between  $f(\{s\}) = \{4\}$  and  $f(s) = 4$ . Be careful.

**Example 12.14** Consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = x^2$ . Note that  $f(\{0, 1, 2\}) = \{0, 1, 4\}$  and  $f^{-1}(\{0, 1, 4\}) = \{-2, -1, 0, 1, 2\}$ . This shows that  $f^{-1}(f(X)) \neq X$  in general.

Using the same  $f$ , check your understanding of these statements about images and preimages of intervals:  $f([-2, 3]) = [0, 9]$ , and  $f^{-1}([0, 9]) = [-3, 3]$ . Also  $f(\mathbb{R}) = [0, \infty)$  and  $f^{-1}([-2, -1]) = \emptyset$ .

If you continue with mathematics you will likely encounter the following results. For now, you are asked to prove them in the exercises.

**Theorem 12.4** Given  $f : A \rightarrow B$ , let  $W, X \subseteq A$ , and  $Y, Z \subseteq B$ . Then

- |   |  |
|---|--|
| 1. $f(W \cap X) \subseteq f(W) \cap f(X)$ | 4. $f^{-1}(Y \cup Z) = f^{-1}(Y) \cup f^{-1}(Z)$ |
| 2. $f(W \cup X) = f(W) \cup f(X)$         | 5. $f^{-1}(Y \cap Z) = f^{-1}(Y) \cap f^{-1}(Z)$ |
| 3. $X \subseteq f^{-1}[f(X)]$             | 6. $f(f^{-1}(Y)) \subseteq Y$ .                  |

### Exercises for Section 12.6

1. Consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = x^2 + 3$ . Find  $f([-3, 5])$  and  $f^{-1}([12, 19])$ .
2. Consider the function  $f : \{1, 2, 3, 4, 5, 6, 7\} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$  given as

$$f = \{(1, 3), (2, 8), (3, 3), (4, 1), (5, 2), (6, 4), (7, 6)\}.$$

Find:  $f(\{1, 2, 3\})$ ,  $f(\{4, 5, 6, 7\})$ ,  $f(\emptyset)$ ,  $f^{-1}(\{0, 5, 9\})$  and  $f^{-1}(\{0, 3, 5, 9\})$ .

3. This problem concerns functions  $f : \{1, 2, 3, 4, 5, 6, 7\} \rightarrow \{0, 1, 2, 3, 4\}$ . How many such functions have the property that  $|f^{-1}(\{3\})| = 3$ ?
4. This problem concerns functions  $f : \{1, 2, 3, 4, 5, 6, 7, 8\} \rightarrow \{0, 1, 2, 3, 4, 5, 6\}$ . How many such functions have the property that  $|f^{-1}(\{2\})| = 4$ ?
5. Consider a function  $f : A \rightarrow B$  and a subset  $X \subseteq A$ . We observed in Example 12.14 that  $f^{-1}(f(X)) \neq X$  in general. However  $X \subseteq f^{-1}(f(X))$  is always true. Prove this.
6. Given a function  $f : A \rightarrow B$  and a subset  $Y \subseteq B$ , is  $f(f^{-1}(Y)) = Y$  always true? Prove or give a counterexample.
7. Given a function  $f : A \rightarrow B$  and subsets  $W, X \subseteq A$ , prove  $f(W \cap X) \subseteq f(W) \cap f(X)$ .
8. Given a function  $f : A \rightarrow B$  and subsets  $W, X \subseteq A$ , then  $f(W \cap X) = f(W) \cap f(X)$  is false in general. Produce a counterexample.
9. Given a function  $f : A \rightarrow B$  and subsets  $W, X \subseteq A$ , prove  $f(W \cup X) = f(W) \cup f(X)$ .
10. Given  $f : A \rightarrow B$  and subsets  $Y, Z \subseteq B$ , prove  $f^{-1}(Y \cap Z) = f^{-1}(Y) \cap f^{-1}(Z)$ .
11. Given  $f : A \rightarrow B$  and subsets  $Y, Z \subseteq B$ , prove  $f^{-1}(Y \cup Z) = f^{-1}(Y) \cup f^{-1}(Z)$ .
12. Consider  $f : A \rightarrow B$ . Prove that  $f$  is injective if and only if  $X = f^{-1}(f(X))$  for all  $X \subseteq A$ . Prove that  $f$  is surjective if and only if  $f(f^{-1}(Y)) = Y$  for all  $Y \subseteq B$ .
13. Let  $f : A \rightarrow B$  be a function, and  $X \subseteq A$ . Prove or disprove:  $f(f^{-1}(f(X))) = f(X)$ .
14. Let  $f : A \rightarrow B$  be a function, and  $Y \subseteq B$ . Prove or disprove:  $f^{-1}(f(f^{-1}(Y))) = f^{-1}(Y)$ .

# CHAPTER 13

---

## Proofs in Calculus

---

The proofs we have dealt with so far in this text have been largely proofs about integers, or about structures related to integers (divisibility, congruence modulo  $n$ , sets of integers, relations among integers, functions of integers, etc.).

Of course mathematics is not restricted to just integers. Calculus is built on the system of real numbers  $\mathbb{R}$ . Thus the main definitions in calculus cater to  $\mathbb{R}$ . Consequently the proofs in calculus (which use the definitions) have a distinct flavor that is quite different from proofs in other areas of mathematics. In reading and writing proofs in calculus you will still use the main proof techniques (direct, contrapositive, contradiction), but it can take some time to adjust your thinking to the idiosyncrasies of  $\mathbb{R}$ . This chapter is intended to ease that adjustment. It is an introduction to some of the ideas you will encounter in later courses in advanced calculus (also called *analysis*). This chapter is not needed for the remainder of the text, so it can be skipped without a loss of continuity.

Single-variable calculus (the first two semesters of a standard calculus sequence) deals with functions  $f : \mathbb{R} \rightarrow \mathbb{R}$ , or more generally  $f : X \rightarrow \mathbb{R}$  for  $X \subseteq \mathbb{R}$ . Usually the domain  $X$  is an interval or a union of intervals. For example,  $f(x) = \frac{x^2+5}{(x-1)(x-2)}$  is a function  $f : (-\infty, 1) \cup (1, 2) \cup (2, \infty) \rightarrow \mathbb{R}$ , whereas  $f(x) = \sqrt{x}$  has domain  $X = [0, \infty)$ , and  $f(x) = x^2 - x$  has domain  $X = (-\infty, \infty) = \mathbb{R}$ .

Calculus rests on the idea of a *limit*, and it is the limit that separates calculus from algebra and trigonometry. We will study limits in sections 13.2 through 13.6. It is assumed that you have had a prior course in calculus and already have some experience with limits. But our present treatment is more theoretical. It serves the double purpose of putting your earlier work on a firmer foundation while preparing you for more advanced studies.

Another calculus topic (typically from the second semester of a standard course) concerns sequences and series, where functions  $f : \mathbb{N} \rightarrow \mathbb{R}$  play a major role. We will turn to this in sections 13.7 and 13.8.

All of this requires a result called the *triangle inequality*, so we begin there.

### 13.1 The Triangle Inequality

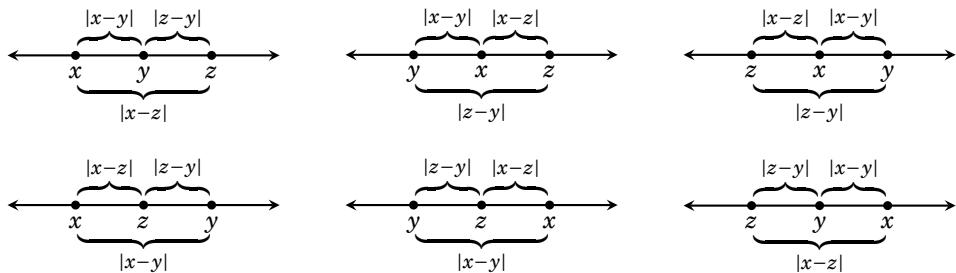
Definitions in calculus and analysis use absolute value extensively. As you know, the absolute value of a real number  $x$  is the non-negative number

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

Fundamental properties of absolute value include  $|xy| = |x| \cdot |y|$  and  $x \leq |x|$ . Another property—used often in proofs—is the *triangle inequality*:

**Theorem 13.1** (Triangle inequality) If  $x, y, z \in \mathbb{R}$ , then  $|x - y| \leq |x - z| + |z - y|$ .

*Proof.* The name *triangle inequality* comes from the fact that the theorem can be interpreted as asserting that for any “triangle” on the number line, the length of any side never exceeds the sum of the lengths of the other two sides. Indeed, the distance between any two numbers  $a, b \in \mathbb{R}$  is  $|a - b|$ . With this in mind, observe in the diagrams below that regardless of the order of  $x, y, z$  on the number line, the inequality  $|x - y| \leq |x - z| + |z - y|$  holds.



(These diagrams show  $x, y, z$  as distinct points. If  $x = y$ ,  $x = z$  or  $y = z$ , then  $|x - y| \leq |x - z| + |z - y|$  holds automatically.) ■

The triangle inequality says the shortest route from  $x$  to  $y$  avoids  $z$  unless  $z$  lies between  $x$  and  $y$ . Several useful results flow from it. Put  $z = 0$  to get

$$|x - y| \leq |x| + |y| \quad \text{for any } x, y \in \mathbb{R}. \quad (13.1)$$

Using the triangle inequality,  $|x + y| = |x - (-y)| \leq |x - 0| + |0 - (-y)| = |x| + |y|$ , so

$$|x + y| \leq |x| + |y| \quad \text{for any } x, y \in \mathbb{R}. \quad (13.2)$$

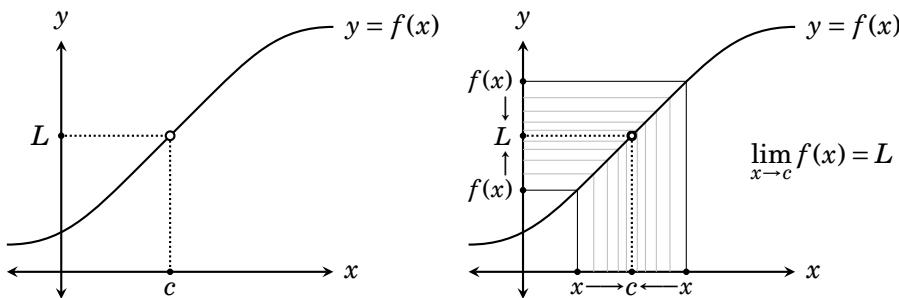
Also by the triangle inequality,  $|x - 0| \leq |x - (-y)| + |-y - 0|$ , which yields

$$|x| - |y| \leq |x + y| \quad \text{for any } x, y \in \mathbb{R}. \quad (13.3)$$

The three inequalities (13.1), (13.2) and (13.3) are very useful in proofs.

### 13.2 Definition of a Limit

Limits are designed to deal with the following type of problem: We need to know how a certain function  $f(x)$  behaves when  $x$  is close to some number  $c$ . Perhaps  $f(c)$  is not even defined, so the graph of  $f$  looks something as shown below; a curve with a hole at a point  $(c, L)$ .



In this picture, for any  $x \neq c$ , the corresponding value  $f(x)$  is either greater than  $L$  or less than  $L$ . But the closer  $x$  is to  $c$ , the closer  $f(x)$  is to  $L$ , as illustrated on the right. We express this as  $\lim_{x \rightarrow c} f(x) = L$ . That is, the symbols  $\lim_{x \rightarrow c} f(x)$  stand for the number that  $f(x)$  approaches as  $x$  approaches  $c$ .

Your calculus text probably presented an informal, intuitive definition of a limit that likely went something like this.

#### Definition 13.1 (Informal definition of a limit)

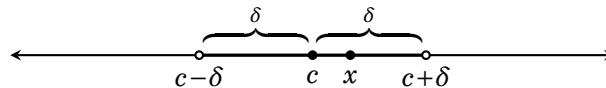
Suppose  $f$  is a function and  $c$  is a number. Then  $\lim_{x \rightarrow c} f(x) = L$  means that  $f(x)$  is arbitrarily close to  $L$  provided that  $x$  is sufficiently close to  $c$ .

The idea is that no matter how close we want to make  $f(x)$  to  $L$ , we can be assured that it will be that close (or closer) if  $x$  is close enough to  $c$ .

Definition 13.1 is sufficient for the first few semesters of calculus, but it is not adequate for deeper, more rigorous investigations. The problem is that it is too vague. What, exactly, is meant by *close*? Saying  $x$  is “close” to  $c$  is not much better than saying that an integer  $n$  is “sort of even.” No proof can be done in the presence of such ambiguity.

So this section’s first task is to motivate and develop a more rigorous and precise limit definition, the one used in advanced calculus. Achieving this goal forces us to grapple with the imprecise term *close*. What do we mean by *close*? Within 0.1 units? Within 0.001 or 0.00001 units, or even closer? We will make the definition precise by introducing a numeric, quantitative measure of closeness.

Standard practice uses the Greek letters  $\varepsilon$  (epsilon) and  $\delta$  (delta) for variables representing how close  $f(x)$  is to  $L$ , and  $x$  is to  $c$ . For instance,  $x$  is within a distance of  $\delta$  from  $c$  if and only if  $c - \delta < x < c + \delta$ , that is,  $-\delta < x - c < \delta$ , or  $|x - c| < \delta$ . So for any real number  $\delta > 0$  (no matter how small) the statement  $|x - c| < \delta$  means that  $x$  is within  $\delta$  units from  $c$ .



Likewise  $|f(x) - L| < \varepsilon$  means that  $f(x)$  is within  $\varepsilon$  units from  $L$ . Let's apply these ideas to Definition 13.1, and transform it line by line.

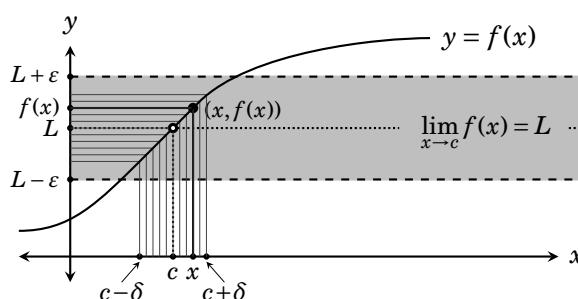
Informal definition	→	Precise definition
$\lim_{x \rightarrow c} f(x) = L$ means that	→	$\lim_{x \rightarrow c} f(x) = L$ means that
$f(x)$ is arbitrarily close to $L$	→	for any $\varepsilon > 0$ , $ f(x) - L  < \varepsilon$
provided that	→	provided that
$x$ is sufficiently close to $c$	→	$0 <  x - c  < \delta$ for some $\delta > 0$ .

We have arrived at a precise definition of a limit.

### Definition 13.2 (Precise definition of a limit)

Suppose  $f : X \rightarrow \mathbb{R}$  is a function, where  $X \subseteq \mathbb{R}$ , and  $c \in \mathbb{R}$ . Then  $\lim_{x \rightarrow c} f(x) = L$  means that for any real  $\varepsilon > 0$  (no matter how small), there is a real number  $\delta > 0$  for which  $|f(x) - L| \leq \varepsilon$  provided that  $0 < |x - c| \leq \delta$ .

Figure 13.1 illustrates this. For any  $\varepsilon > 0$ , no matter how small, consider the narrow shaded band of points on the plane whose  $y$ -coordinates are between  $y = L - \varepsilon$  and  $y = L + \varepsilon$ . Given this  $\varepsilon$ , we can find another number  $\delta > 0$  such that the point  $(x, f(x))$  is in the shaded band whenever  $x$  is within  $\delta$  units from  $c$ . In other words,  $|f(x) - L| \leq \varepsilon$  provided that  $0 < |x - c| \leq \delta$ .



**Figure 13.1.** A graphic description of the limit definition.

Three comments are in order. First, we have slipped into Definition 13.2 the expression  $0 < |x - c| < \delta$  instead of  $|x - c| < \delta$ . This is to rule out the possibility  $x = c$ , as  $f(c)$  may not be defined, depending of  $f$  and  $c$ .

Second, Definition 13.2 applies only if there is some  $\delta > 0$  for which  $(c - \delta, c) \cup (c, c + \delta)$  is a subset of the domain of  $f$ . Otherwise the statement “ $|f(x) - L|$  provided that  $0 < |x - c| < \delta$ ” is meaningless for some  $x$ , no matter how small  $\delta$  is. Thus  $\lim_{x \rightarrow c} f(x)$  makes sense only if  $f(x)$  is defined for all  $x \in \mathbb{R}$  that are “close to”  $c$  in the sense that  $x \in (c - \delta, c) \cup (c, c + \delta)$  for some  $\delta$ .

Third, in symbolic form Definition 13.2 says  $\lim_{x \rightarrow c} f(x) = L$  if and only if

$$\forall \varepsilon > 0, \exists \delta > 0, (0 < |x - c| < \delta) \Rightarrow (|f(x) - L| < \varepsilon). \quad (13.4)$$

Thus proving  $\lim_{x \rightarrow c} f(x) = L$  amounts to proving that Statement (13.4) is true.

One strategy for proving Statement (13.4) is the direct approach. Begin by assuming  $\varepsilon > 0$ . Then find a  $\delta$  for which  $(0 < |x - c| < \delta) \Rightarrow (|f(x) - L| < \varepsilon)$ . To find  $\delta$ , try to extract a factor of  $|x - c|$  from  $|f(x) - L|$ . If you can do this, inspection usually tells you how small  $|x - c|$  needs to be to make  $|f(x) - L| < \varepsilon$ .

We will use this strategy in Example 13.1, which proves  $\lim_{x \rightarrow 2} (3x + 4) = 10$ . Here  $f(x) = 3x + 4$  and  $L = 10$ , so  $|f(x) - L|$  is  $|(3x + 4) - 10|$ . Also  $|x - c|$  is  $|x - 2|$ .

**Example 13.1** Prove that  $\lim_{x \rightarrow 2} (3x + 4) = 10$ .

*Proof.* Suppose  $\varepsilon > 0$ . Note that  $|(3x + 4) - 10| = |3x - 6| = |3(x - 2)| = 3|x - 2|$ . So if  $\delta = \frac{\varepsilon}{3}$ , then  $0 < |x - 2| < \delta$  yields  $|(3x + 4) - 10| = 3|x - 2| < 3\delta = 3\frac{\varepsilon}{3} = \varepsilon$ .

In summary, for any  $\varepsilon > 0$ , there is a  $\delta = \frac{\varepsilon}{3}$  for which  $0 < |x - 2| < \delta$  implies  $|(3x + 4) - 10| < \varepsilon$ . By Definition 13.2,  $\lim_{x \rightarrow 2} (3x + 4) = 10$ . ■

**Example 13.2** Prove that  $\lim_{x \rightarrow 2} 5x^2 = 20$ .

*Proof.* Suppose  $\varepsilon > 0$ . Notice that

$$|f(x) - L| = |5x^2 - 20| = |5(x^2 - 4)| = |5(x - 2)(x + 2)| = 5 \cdot |x - 2| \cdot |x + 2|.$$

Now we have a factor of  $|x - 2|$  in  $|f(x) - L|$ , but it is accompanied with  $|x + 2|$ . But if  $|x - 2|$  is small, then  $x$  is close to 2, so  $|x + 2|$  should be close to 4. In fact, if  $|x - 2| \leq 1$ , then  $|x + 2| = |(x - 2) + 4| \leq |x - 2| + 4 \leq 1 + 4 = 5$ . (Here we applied the inequality (13.2) from page 245.) In other words, if  $|x - 2| \leq 1$ , then  $|x + 2| \leq 5$ , and the above equation yields

$$|f(x) - L| = |5x^2 - 20| = 5 \cdot |x - 2| \cdot |x + 2| < 5 \cdot |x - 2| \cdot 5 = 25|x - 2|.$$

Take  $\delta$  to be smaller than both 1 and  $\frac{\varepsilon}{25}$ . Then  $0 < |x - 2| < \delta$  implies  $|5x^2 - 20| < 25 \cdot |x - 2| < 25\delta < 25\frac{\varepsilon}{25} = \varepsilon$ . By Definition 13.2,  $\lim_{x \rightarrow 2} 5x^2 = 20$ . ■

The examples above (and the exercises below) involve limits that you probably regard as obvious. Our point is to illustrate Definition 13.2, not to compute difficult limits. Difficult limits come later (mostly in more advanced courses, not in this book) where Definition 13.2 will be used to great effect.

### Exercises for Section 13.2

1. Prove that  $\lim_{x \rightarrow 5} (8x - 3) = 37$ .
2. Prove that  $\lim_{x \rightarrow -1} (4x + 6) = 2$ .
3. Prove that  $\lim_{x \rightarrow 0} (x + 2) = 2$ .
4. Prove that  $\lim_{x \rightarrow 8} (2x - 7) = 9$ .
5. Prove that  $\lim_{x \rightarrow 3} (x^2 - 2) = 7$ .
6. Prove that  $\lim_{x \rightarrow 1} (4x^2 + 1) = 5$ .

### 13.3 Limits That Do Not Exist

Given a function  $f$  and a number  $c$ , there are two ways that  $\lim_{x \rightarrow c} f(x) = L$  can be false. First, there may be a different number  $M \neq L$  for which  $\lim_{x \rightarrow c} f(x) = M$ . Second, it may be that Statement (13.4) is false for all  $L \in \mathbb{R}$ . In such a case we say that  $\lim_{x \rightarrow c} f(x)$  **does not exist**. Contradiction is one way to prove that  $\lim_{x \rightarrow c} f(x)$  does not exist. Assume  $\lim_{x \rightarrow c} f(x) = L$  and produce a contradiction.

**Example 13.3** Suppose  $f(x) = \frac{x}{2} + \frac{|x-2|}{x-2} + 2$ . Prove  $\lim_{x \rightarrow 2} f(x)$  does not exist.

*Proof.* Notice that  $f(2)$  is not defined, as it involves division by zero. Also,  $f(x)$  behaves differently depending on whether  $x$  is to the right or left of 2.

If  $x > 2$ , then  $x-2$  is positive, so  $|x-2| = x-2$  and  $\frac{|x-2|}{x-2} = 1$ , so  $f(x) = \frac{x}{2} + 3$ .

If  $x < 2$ , then  $x-2$  is negative, so  $|x-2| = -(x-2)$  and  $\frac{|x-2|}{x-2} = -1$ , so  $f(x) = \frac{x}{2} + 1$ .

Therefore  $f$ , graphed below, is a piecewise function  $f(x) = \begin{cases} \frac{1}{2}x + 3 & \text{if } x > 2 \\ \frac{1}{2}x + 1 & \text{if } x < 2. \end{cases}$

Suppose for the sake of contradiction that  $\lim_{x \rightarrow 2} f(x) = L$ , where  $L$  is a real number. Let  $\varepsilon = \frac{1}{2}$ .

By Definition 13.2, there is a real number  $\delta > 0$  for which  $0 < |x-2| < \delta$  implies  $|f(x)-L| < \frac{1}{2}$ .

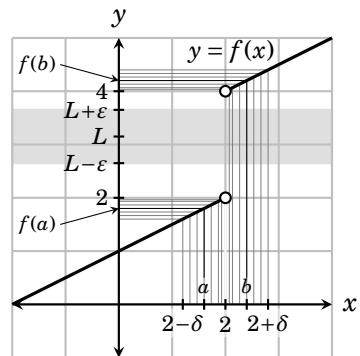
Put  $a = 2 - \frac{\delta}{2}$ , so  $0 < |a-2| < \delta$ . Hence  $|f(a)-L| < \frac{1}{2}$ .

Put  $b = 2 + \frac{\delta}{2}$ , so  $0 < |b-2| < \delta$ . Hence  $|f(b)-L| < \frac{1}{2}$ .

Further,  $f(a) < 2$  and  $f(b) > 4$ , so  $2 < |f(b)-f(a)|$ .

With this and the help of the inequality (13.1), we get a contradiction  $2 < 1$ , as follows:

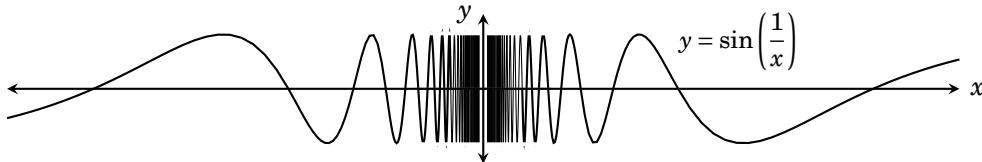
$$2 < |f(b)-f(a)| = |(f(b)-L) - (f(a)-L)| \leq |f(b)-L| + |f(a)-L| < \frac{1}{2} + \frac{1}{2} = 1. \blacksquare$$



Our next limit is a classic example of a non-existent limit. It often appears in first-semester calculus texts, where it is treated informally.

**Example 13.4** Prove that  $\lim_{x \rightarrow 0} \sin\left(\frac{1}{x}\right)$  does not exist.

As  $x$  approaches 0, the number  $\frac{1}{x}$  grows bigger, approaching infinity, so  $\sin\left(\frac{1}{x}\right)$  just bounces up and down, faster and faster the closer  $x$  gets to 0.



Intuitively, we would guess that the limit does not exist, because  $\sin\left(\frac{1}{x}\right)$  does not approach any single number as  $x$  approaches 0. Here is a proof.

*Proof.* Suppose for the sake of contradiction that  $\lim_{x \rightarrow 0} \sin\left(\frac{1}{x}\right) = L$  for  $L \in \mathbb{R}$ .

Definition 13.2 guarantees a number  $\delta$  for which  $0 < |x - 0| < \delta$  implies  $|\sin\left(\frac{1}{x}\right) - L| < \frac{1}{4}$ . Select  $k \in \mathbb{N}$  large enough so that  $\frac{1}{k\pi} < \delta$ . As  $0 < \left|\frac{1}{k\pi} - 0\right| < \delta$ , we have  $|\sin\left(\frac{1}{1/k\pi}\right) - L| < \frac{1}{4}$ , and this yields  $|\sin(k\pi) - L| = |0 - L| = |L| < \frac{1}{4}$ .

Next, take  $\ell \in \mathbb{N}$  large enough so that  $\frac{1}{\frac{\pi}{2} + 2\ell\pi} < \delta$ . Then  $0 < \left|\frac{1}{\frac{\pi}{2} + 2\ell\pi} - 0\right| < \delta$ , so we have  $\left|\sin\left(\frac{1}{\frac{\pi}{2} + 2\ell\pi}\right) - L\right| < \frac{1}{4}$ , which simplifies to  $|\sin\left(\frac{\pi}{2} + 2\ell\pi\right) - L| = |1 - L| < \frac{1}{4}$ .

Above we showed  $|L| \leq \frac{1}{4}$  and  $|1 - L| \leq \frac{1}{4}$ . Now apply the inequality (13.2) to get the contradiction  $1 < \frac{1}{2}$ , as  $1 = |L + (1 - L)| \leq |L| + |1 - L| < \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ . ■

**Example 13.5** Investigate  $\lim_{x \rightarrow 0} x \sin\left(\frac{1}{x}\right)$ .

This is like the previous example, except for the extra  $x$ . Because  $|\sin\left(\frac{1}{x}\right)| \leq 1$ , we expect  $x \sin\left(\frac{1}{x}\right)$  to go to 0 as  $x$  goes to 0. Indeed, we prove  $\lim_{x \rightarrow 0} x \sin\left(\frac{1}{x}\right) = 0$ .

*Proof.* Given  $\varepsilon > 0$ , let  $\delta = \varepsilon$ . Suppose  $0 < |x - 0| \leq \delta$ . Simplifying,  $|x| < \delta$ , which is the same as  $|x| < \varepsilon$ . We get  $|x \sin\left(\frac{1}{x}\right) - 0| = |x \sin\left(\frac{1}{x}\right)| = |x| \cdot |\sin\left(\frac{1}{x}\right)| < \varepsilon |\sin\left(\frac{1}{x}\right)| \leq \varepsilon \cdot 1 = \varepsilon$ . From this, Definition 13.2 gives  $\lim_{x \rightarrow 0} x \sin\left(\frac{1}{x}\right) = 0$ . ■

One final point. We remarked on page 248 that for  $\lim_{x \rightarrow c} f(x) = L$  to make sense, there must be a  $\delta$  for which  $f(x)$  is defined for all  $x \in (c - \delta, c) \cup (c, c + \delta)$ . Thus, for example, following Definition 13.2 to the letter, we have to say that  $\lim_{x \rightarrow 0} \sqrt{x}$  does not exist because  $\sqrt{x}$  is not defined for all  $x \in (-\delta, 0) \cup (0, \delta)$ . Your calculus text probably introduced a *right-hand limit*  $\lim_{x \rightarrow 0^+} \sqrt{x} = 0$ . Though this notion is not programmed into our Definition 13.2, you may revisit such embellishments in later courses.

---

**Exercises for Section 13.3**

Prove that the following limits do not exist.

1.  $\lim_{x \rightarrow 0} \log_{10} |x|$

2.  $\lim_{x \rightarrow 0} \frac{|x|}{x}$

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

4.  $\lim_{x \rightarrow \frac{\pi}{2}} \cos\left(\frac{1}{x}\right)$

5.  $\lim_{x \rightarrow 0} x \cot\left(\frac{1}{x}\right)$

6.  $\lim_{x \rightarrow 1} \frac{1}{x^2 - 2x + 1}$ 


---

**13.4 Limits Laws**

When you studied Calculus I your text presented a number of *limit laws*, such as  $\lim_{x \rightarrow c} f(x)g(x) = \left(\lim_{x \rightarrow c} f(x)\right) \cdot \left(\lim_{x \rightarrow c} g(x)\right)$ . These laws allowed you to compute complex limits by reducing them to simpler limits, until the answer was at hand. But your calculus text probably did not *prove* the laws. Rather, you were asked to accept them as intuitively plausible (and useful) facts.

Using Definition 13.2, we now present proofs of some limit laws. This serves two purposes. First, it puts your knowledge of calculus on a firmer foundation. Second, it highlights various strategies and thought patterns that are useful in limit proofs, which come to bear in later courses and work.

The inequalities (13.1), (13.2) and (13.3) from page 245 play a crucial role. For convenience we repeat them here. For any  $x, y \in \mathbb{R}$ ,

$$|x - y| \leq |x| + |y|, \quad |x + y| \leq |x| + |y|, \quad \text{and} \quad |x| - |y| \leq |x + y|.$$

We will use these frequently, usually without comment.

Our first limit law concerns the constant function  $f(x) = a$  where  $a \in \mathbb{R}$ . Its graph is a horizontal line with  $y$ -intercept  $a$ . It should be obvious that  $\lim_{x \rightarrow c} f(x) = a$  for any real number  $c$ . Nonetheless, let's prove this obvious fact.

**Theorem 13.2** (Constant function rule) If  $a \in \mathbb{R}$ , then  $\lim_{x \rightarrow c} a = a$ .

*Proof.* Suppose  $a \in \mathbb{R}$ . According to Definition 13.2, to prove  $\lim_{x \rightarrow c} a = a$ , we must show that for any  $\varepsilon > 0$ , there is a  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|a - a| < \varepsilon$ . This is almost too easy. Just let  $\delta = 1$  (or any other number). Then  $|a - a| < \varepsilon$  is automatic, because  $|a - a| = 0$ . ■

The *identity function*  $f : \mathbb{R} \rightarrow \mathbb{R}$  is  $f(x) = x$ . Next we prove  $\lim_{x \rightarrow c} f(x) = c$ .

**Theorem 13.3** (Identity function rule) If  $c \in \mathbb{R}$ , then  $\lim_{x \rightarrow c} x = c$ .

*Proof.* Given  $\varepsilon > 0$ , let  $\delta = \varepsilon$ . Then  $0 < |x - c| < \delta$  implies  $|x - c| < \varepsilon$ . By Definition 13.2, this means  $\lim_{x \rightarrow c} x = c$ . ■

**Theorem 13.4** (Constant multiple rule)

If  $\lim_{x \rightarrow c} f(x)$  exists, and  $a \in \mathbb{R}$ , then  $\lim_{x \rightarrow c} af(x) = a \lim_{x \rightarrow c} f(x)$ .

*Proof.* Suppose  $\lim_{x \rightarrow c} f(x) = L$  exists. We must show  $\lim_{x \rightarrow c} af(x) = a \lim_{x \rightarrow c} f(x)$ . If  $a = 0$ , then this reduces to  $\lim_{x \rightarrow c} 0 = 0$ , which is true by Theorem 13.2. Thus, for the remainder of the proof we can assume  $a \neq 0$ .

Suppose  $\lim_{x \rightarrow c} f(x) = L$ . We must prove  $\lim_{x \rightarrow c} af(x) = aL$ . By Definition 13.2, this means we must show that for any  $\varepsilon > 0$ , there is a  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|af(x) - aL| < \varepsilon$ . Let  $\varepsilon > 0$ . Because  $\lim_{x \rightarrow c} f(x) = L$ , there exists a  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|f(x) - L| < \frac{\varepsilon}{|a|}$ . So if  $0 < |x - c| < \delta$ , then  $|af(x) - aL| = |a(f(x) - L)| = |a| \cdot |f(x) - L| < |a| \frac{\varepsilon}{|a|} = \varepsilon$ .

In summary, we've shown that for any  $\varepsilon > 0$ , there is a  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|af(x) - aL| < \varepsilon$ . By Definition 13.2,  $\lim_{x \rightarrow c} af(x) = aL$ . ■

**Theorem 13.5** (Sum rule)

If both  $\lim_{x \rightarrow c} f(x)$  and  $\lim_{x \rightarrow c} g(x)$  exist, then  $\lim_{x \rightarrow c} (f(x) + g(x)) = \lim_{x \rightarrow c} f(x) + \lim_{x \rightarrow c} g(x)$ .

*Proof.* Let  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M$ . We must prove  $\lim_{x \rightarrow c} (f(x) + g(x)) = L + M$ . To prove this, take  $\varepsilon > 0$ . We need to find a corresponding  $\delta$  for which  $0 < |x - c| < \delta$  implies  $|(f(x) + g(x)) - (L + M)| < \varepsilon$ . With this in mind, notice that

$$\begin{aligned} |(f(x) + g(x)) - (L + M)| &= |(f(x) - L) + (g(x) - M)| \\ &\leq |f(x) - L| + |g(x) - M|. \end{aligned} \quad (\text{A})$$

As  $\lim_{x \rightarrow c} f(x) = L$ , there is a  $\delta' > 0$  such that  $0 < |x - c| < \delta'$  implies  $|f(x) - L| < \frac{\varepsilon}{2}$ .

As  $\lim_{x \rightarrow c} g(x) = M$ , there is a  $\delta'' > 0$  such that  $0 < |x - c| < \delta''$  implies  $|g(x) - M| < \frac{\varepsilon}{2}$ .

Now put  $\delta = \min\{\delta', \delta''\}$ , meaning that  $\delta$  equals the smaller of  $\delta'$  and  $\delta''$ .

If  $0 < |x - c| < \delta$ , then (A) gives  $|(f(x) + g(x)) - (L + M)| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ .

We've now shown that for any  $\varepsilon > 0$ , there is a  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|(f(x) + g(x)) - (L + M)| < \varepsilon$ . Thus  $\lim_{x \rightarrow c} (f(x) + g(x)) = L + M$ . ■

**Theorem 13.6** (Difference rule)

If both  $\lim_{x \rightarrow c} f(x)$  and  $\lim_{x \rightarrow c} g(x)$  exist, then  $\lim_{x \rightarrow c} (f(x) - g(x)) = \lim_{x \rightarrow c} f(x) - \lim_{x \rightarrow c} g(x)$ .

*Proof.* Combining the sum rule with the constant multiple rule gives

$$\begin{aligned} \lim_{x \rightarrow c} (f(x) - g(x)) &= \lim_{x \rightarrow c} (f(x) + (-1) \cdot g(x)) = \lim_{x \rightarrow c} f(x) + \lim_{x \rightarrow c} (-1) \cdot g(x) \\ &= \lim_{x \rightarrow c} f(x) + (-1) \lim_{x \rightarrow c} g(x) = \lim_{x \rightarrow c} f(x) - \lim_{x \rightarrow c} g(x). \end{aligned} \quad \blacksquare$$

**Theorem 13.7** (Multiplication rule)

If both  $\lim_{x \rightarrow c} f(x)$  and  $\lim_{x \rightarrow c} g(x)$  exist, then  $\lim_{x \rightarrow c} f(x)g(x) = \left(\lim_{x \rightarrow c} f(x)\right) \cdot \left(\lim_{x \rightarrow c} g(x)\right)$ .

*Proof.* Let  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M$ . We must prove  $\lim_{x \rightarrow c} f(x)g(x) = LM$ . To prove this, take  $\varepsilon > 0$ . We need to find a corresponding  $\delta$  for which  $0 < |x - c| < \delta$  implies  $|f(x)g(x) - LM| < \varepsilon$ . With this in mind, notice that

$$\begin{aligned} |f(x)g(x) - LM| &= |(f(x)g(x) - Lg(x)) + (Lg(x) - LM)| \\ &\leq |f(x)g(x) - Lg(x)| + |Lg(x) - LM| \\ &= |(f(x) - L)g(x)| + |L(g(x) - M)| \\ &= |f(x) - L| \cdot |g(x)| + |L| \cdot |g(x) - M|. \end{aligned} \quad (\text{A})$$

Because  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M$ , we can make the expressions  $|f(x) - L|$  and  $|L| \cdot |g(x) - M|$  in (A) arbitrarily small by making  $|x - c|$  sufficiently small. But the term  $|f(x) - L| \cdot |g(x)|$  is a problem. For all we know,  $|g(x)|$  could grow large as  $|f(x) - L|$  shrinks. To deal with this, choose some  $\delta' > 0$  small enough so that  $0 < |x - c| < \delta'$  implies  $|g(x) - M| < 1$ . Then as long as  $0 < |x - c| < \delta'$ ,

$$|g(x)| = |(g(x) - M) + M| \leq |g(x) - M| + |M| < 1 + |M|.$$

Replacing the factor of  $|g(x)|$  in (A) with the larger quantity  $1 + |M|$ , we get

$$|f(x)g(x) - LM| < |f(x) - L| \cdot (1 + |M|) + |L| \cdot |g(x) - M|, \quad (\text{B})$$

which holds provided  $0 < |x - c| < \delta'$ .

Choose  $\delta'' > 0$  such that  $0 < |x - c| < \delta''$  implies  $|f(x) - L| < \frac{\varepsilon}{2(1+|M|)}$ . Also choose  $\delta''' > 0$  such that  $0 < |x - c| < \delta'''$  implies  $|g(x) - M| < \frac{\varepsilon}{2|L|}$ . Now put  $\delta = \min\{\delta', \delta'', \delta'''\}$ . If  $0 < |x - c| < \delta$ , then (B) becomes

$$|f(x)g(x) - LM| < \frac{\varepsilon}{2(1+|M|)} \cdot (1 + |M|) + |L| \cdot \frac{\varepsilon}{2|L|} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

To summarize, we've shown that for any  $\varepsilon > 0$ , there is a  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|f(x)g(x) - LM| < \varepsilon$ . Therefore  $\lim_{x \rightarrow c} f(x)g(x) = LM$ . ■

Our final rule has proof similar to that of the multiplication rule. We just have to take a little extra care with the denominators.

**Theorem 13.8** (Division rule)

If both  $\lim_{x \rightarrow c} f(x)$  and  $\lim_{x \rightarrow c} g(x)$  exist, and  $\lim_{x \rightarrow c} g(x) \neq 0$ , then  $\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = \frac{\lim_{x \rightarrow c} f(x)}{\lim_{x \rightarrow c} g(x)}$ .

*Proof.* Suppose  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M \neq 0$ . We must prove  $\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = \frac{L}{M}$ . To prove this, take  $\varepsilon > 0$ . We need to find a corresponding  $\delta$  for which  $0 < |x - c| < \delta$  implies  $\left| \frac{f(x)}{g(x)} - \frac{L}{M} \right| < \varepsilon$ . With this in mind, notice that

$$\begin{aligned} \left| \frac{f(x)}{g(x)} - \frac{L}{M} \right| &= \left| \frac{Mf(x) - Lg(x)}{Mg(x)} \right| = \left| \frac{(Mf(x) - LM) - (Lg(x) - LM)}{Mg(x)} \right| \\ &= \left| \frac{1}{g(x)}(f(x) - L) - \frac{L}{Mg(x)}(g(x) - M) \right| \\ &\leq \left| \frac{1}{g(x)}(f(x) - L) \right| + \left| \frac{L}{Mg(x)}(g(x) - M) \right| \\ &= \frac{1}{|g(x)|} \cdot |f(x) - L| + \frac{1}{|g(x)|} \cdot \left| \frac{L}{M} \right| \cdot |g(x) - M|. \end{aligned} \quad (\text{A})$$

Because  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M$ , we can make the terms  $|f(x) - L|$  and  $\left| \frac{L}{M} \right| \cdot |g(x) - M|$  in (A) arbitrarily small by making  $|x - c|$  sufficiently small. To deal with the factor  $\frac{1}{|g(x)|}$ , choose  $\delta' > 0$  so that  $0 < |x - c| < \delta'$  implies  $|g(x) - M| < \frac{|M|}{2}$ . So if  $0 < |x - c| < \delta'$ , the inequality (13.3) assures us

$$|g(x)| = |M + (g(x) - M)| \geq |M| - |g(x) - M| > |M| - \frac{|M|}{2} = \frac{|M|}{2}.$$

That is,  $|g(x)| > \frac{|M|}{2}$ , and consequently  $\frac{1}{|g(x)|} < \frac{2}{|M|}$ . Replacing the occurrences of  $\frac{1}{|g(x)|}$  in (A) with the larger value  $\frac{2}{|M|}$ , we get

$$\left| \frac{f(x)}{g(x)} - \frac{L}{M} \right| < \frac{2}{|M|} \cdot |f(x) - L| + \left| \frac{2L}{M^2} \right| \cdot |g(x) - M|, \quad (\text{B})$$

which holds provided  $0 < |x - c| < \delta'$ . Two cases finish the proof.

**Case 1.** Suppose  $L \neq 0$ . Choose  $\delta'' > 0$  so  $0 < |x - c| < \delta''$  implies  $|f(x) - L| < \frac{\varepsilon|M|}{4}$ . Also choose  $\delta''' > 0$  so that  $0 < |x - c| < \delta'''$  implies  $|g(x) - M| < \varepsilon \left| \frac{M^2}{4L} \right|$ . Put  $\delta = \min \{\delta', \delta'', \delta'''\}$ . If  $0 < |x - c| < \delta$ , then (B) yields

$$\left| \frac{f(x)}{g(x)} - \frac{L}{M} \right| < \frac{2}{|M|} \cdot \frac{\varepsilon|M|}{4} + \left| \frac{2L}{M^2} \right| \cdot \varepsilon \left| \frac{M^2}{4L} \right| = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

**Case 2.** Suppose  $L = 0$ . Let  $\delta'' > 0$  be such that  $0 < |x - c| < \delta''$  implies  $|f(x) - L| < \frac{\varepsilon|M|}{2}$ . Putting  $\delta = \min \{\delta', \delta''\}$ , the inequality (B) becomes

$$\left| \frac{f(x)}{g(x)} - \frac{L}{M} \right| < \frac{2}{|M|} \frac{\varepsilon|M|}{2} = \varepsilon.$$

In each case we have shown that for any  $\varepsilon > 0$ , there is a  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $\left| \frac{f(x)}{g(x)} - \frac{L}{M} \right| < \varepsilon$ , so the proof is finished. ■

Though you may not have proved any limit laws in your calculus course, you *used* them extensively. A common situation involved  $\lim_{x \rightarrow c} f(x)$ , where  $f(c)$  was undefined because of a zero denominator. You learned to overcome this by algebraically canceling the offending part of the denominator.

**Example 13.6** Find  $\lim_{x \rightarrow 1} \frac{\frac{1}{x} - 1}{1 - x}$ .

Here  $x$  approaches 1, but simply plugging in  $x = 1$  gives  $\frac{\frac{1}{1} - 1}{1 - 1} = \frac{0}{0}$  (undefined). So we apply whatever algebra is needed to cancel the denominator  $1 - x$ , and follow this with limit laws:

$$\begin{aligned} \lim_{x \rightarrow 1} \frac{\frac{1}{x} - 1}{1 - x} &= \lim_{x \rightarrow 1} \frac{\frac{1}{x} - 1}{1 - x} \cdot \frac{x}{x} && \text{(multiply quotient by } 1 = \frac{x}{x}) \\ &= \lim_{x \rightarrow 1} \frac{(1-x)}{(1-x)x} && \text{(distribute } x \text{ on top)} \\ &= \lim_{x \rightarrow 1} \frac{1}{x} && \text{(cancel the } (1-x)) \\ &= \frac{\lim_{x \rightarrow 1} 1}{\lim_{x \rightarrow 1} x} = \frac{1}{1} = 1. && \text{(apply limit laws)} \end{aligned}$$

### Exercises for Section 13.4

- Given two or more functions  $f_1, f_2, \dots, f_n$ , suppose that  $\lim_{x \rightarrow c} f_i(x)$  exists for each  $1 \leq i \leq n$ . Prove that  $\lim_{x \rightarrow c} (f_1(x) + f_2(x) + \dots + f_n(x)) = \lim_{x \rightarrow c} f_1(x) + \lim_{x \rightarrow c} f_2(x) + \dots + \lim_{x \rightarrow c} f_n(x)$ . Use induction on  $n$ , with Theorem 13.5 serving as the base case.
- Given two or more functions  $f_1, f_2, \dots, f_n$ , suppose that  $\lim_{x \rightarrow c} f_i(x)$  exists for each  $1 \leq i \leq n$ . Prove that  $\lim_{x \rightarrow c} (f_1(x)f_2(x) \cdots f_n(x)) = (\lim_{x \rightarrow c} f_1(x)) \cdot (\lim_{x \rightarrow c} f_2(x)) \cdots (\lim_{x \rightarrow c} f_n(x))$ . Use induction on  $n$ , with Theorem 13.7 serving as the base case.
- Use the previous two exercises and the constant multiple rule (Theorem 13.4) to prove that if  $f(x)$  is a polynomial, then  $\lim_{x \rightarrow c} f(x) = f(c)$  for any  $c \in \mathbb{R}$ .
- Use Exercise 3 with a limit law to prove that if  $\frac{f(x)}{g(x)}$  is a rational function (a polynomial divided by a polynomial), and  $g(c) \neq 0$ , then  $\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = \frac{f(c)}{g(c)}$ .
- Use Definition 13.2 to prove that limits are unique in the sense that if  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} f(x) = M$ , then  $L = M$ .
- Prove the *squeeze theorem*: Suppose  $g(x) \leq f(x) \leq h(x)$  for all  $x \in \mathbb{R}$  satisfying  $0 < |x - c| < \delta$  for some  $\delta > 0$ . If  $\lim_{x \rightarrow c} g(x) = L = \lim_{x \rightarrow c} h(x)$ , then  $\lim_{x \rightarrow c} f(x) = L$ .

### 13.5 Continuity and Derivatives

A major purpose of limits is that they can give information about how a function behaves near a “bad point”  $x = c$ . Even if  $f(c)$  is not defined, it may be that  $\lim_{x \rightarrow c} f(x) = L$ , for some number  $L$ . In this event we know that  $f(x)$  becomes ever closer to  $L$  as  $x$  approaches the forbidden  $c$ .

Of course not every value  $x = c$  is a “bad point.” It could be that  $f(c)$  is defined, and, moreover,  $\lim_{x \rightarrow c} f(x) = f(c)$ . If this is the case for every  $c$  in the domain of  $f(x)$ , then we say that  $f$  is *continuous*. Issues concerning whether or not  $f$  is continuous are called issues of *continuity*.

In a first course in calculus it is easy to overlook the huge importance of continuity. And happily, we can (in a first course) almost ignore it. But in fact, the theoretical foundation of calculus rests on continuity. Roughly speaking, there are countless theorems having the form

If  $f$  is continuous, then  $f$  has some significant property.

Continuity allows us to draw certain important conclusions about a function. Here is its definition.

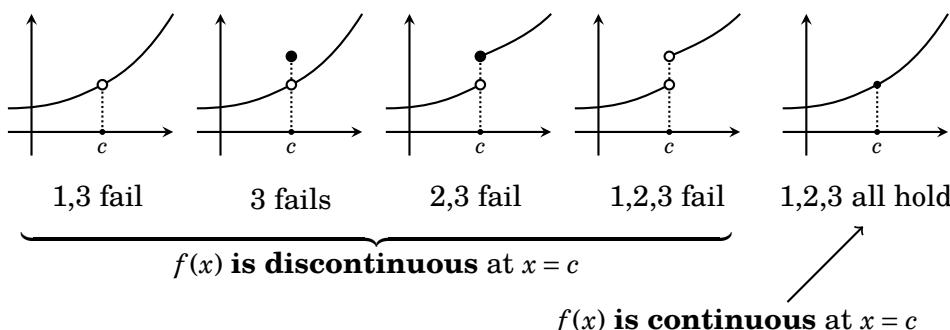
**Definition 13.3** A function  $f(x)$  is **continuous** at  $x = c$  if  $\lim_{x \rightarrow c} f(x) = f(c)$ .

Note that this means *all* of the following three conditions must be met:

1.  $f(c)$  is defined,
2.  $\lim_{x \rightarrow c} f(x)$  exists,
3.  $\lim_{x \rightarrow c} f(x) = f(c)$ .

If one or more of these conditions fail, then  $f(x)$  is **discontinuous** at  $c$ .

To illustrate this definition, five functions  $f(x)$  are graphed below. Only the function on the far right is continuous at  $x = c$ .



Most familiar functions are continuous at each point  $x = c$  in their domain. For instance, exercises 3 and 4 in the previous section imply that polynomials and rational functions are continuous at any number  $c$  in their domains.

One application of continuity is a limit law for composition. The previous section might prompt us to conjecture that  $\lim_{x \rightarrow c} f(g(x)) = f\left(\lim_{x \rightarrow c} g(x)\right)$ . However, this does not hold without an assumption of continuity.

**Theorem 13.9** (Composition rule)

If  $\lim_{x \rightarrow c} g(x) = L$  and  $f$  is continuous at  $x = L$ , then  $\lim_{x \rightarrow c} f(g(x)) = f\left(\lim_{x \rightarrow c} g(x)\right)$ .

*Proof.* Suppose  $\lim_{x \rightarrow c} g(x) = L$  and  $f$  is continuous at  $x = L$ . We need to show  $\lim_{x \rightarrow c} f(g(x)) = f(L)$ . According to Definition 13.2, for any  $\varepsilon > 0$  we must show there is a corresponding  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|f(g(x)) - f(L)| < \varepsilon$ .

So let  $\varepsilon > 0$ . As  $f$  is continuous at  $L$ , Definition 13.3 yields  $\lim_{x \rightarrow L} f(x) = f(L)$ . From this, we know there is a real number  $\delta' > 0$  for which

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

But also, from  $\lim_{x \rightarrow c} g(x) = L$ , we know that there is a real number  $\delta > 0$  for which  $0 < |x - c| < \delta$  implies  $|g(x) - L| < \delta'$ .

If  $0 < |x - c| < \delta$ , then we have  $|g(x) - L| < \delta'$ , and from this (A) yields  $|f(g(x)) - f(L)| < \varepsilon$ . Thus  $\lim_{x \rightarrow c} f(g(x)) = f(L)$ , and the proof is complete. ■

In calculus you learned that the **derivative** of a real-valued function  $f$  is another function  $f'$  for which  $f'(c)$  is defined as

$$f'(c) = \lim_{x \rightarrow c} \frac{f(x) - f(c)}{x - c},$$

provided the limit exists (in which case we say  $f$  is **differentiable** at  $c$ ).

You may recall that differentiability implies continuity.

**Theorem 13.10** If  $f$  is differentiable at  $c$ , then  $f$  is continuous at  $c$ .

*Proof.* Suppose  $f$  is differentiable at  $c$ , so  $\lim_{x \rightarrow c} \frac{f(x) - f(c)}{x - c} = f'(c)$ . Write  $f(x)$  as

$$f(x) = \frac{f(x) - f(c)}{x - c}(x - c) + f(c).$$

Taking limits of both sides and using limit laws,

$$\lim_{x \rightarrow c} f(x) = \left( \lim_{x \rightarrow c} \frac{f(x) - f(c)}{x - c} \right) \cdot \left( \lim_{x \rightarrow c} (x - c) \right) + \lim_{x \rightarrow c} f(c) = f'(c) \cdot 0 + f(c) = f(c).$$

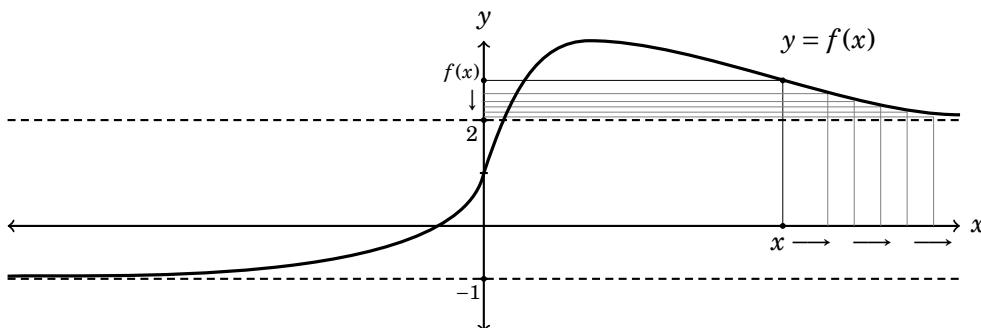
Thus  $\lim_{x \rightarrow c} f(x) = f(c)$ , which means  $f$  is continuous at  $c$ . ■

### Exercises for Section 13.5

- Prove that the function  $f(x) = \sqrt{x}$  is continuous at any number  $c > 0$ . Deduce that  $\lim_{x \rightarrow c} \sqrt{g(x)} = \sqrt{\lim_{x \rightarrow c} g(x)}$ , provided  $\lim_{x \rightarrow c} g(x)$  exists and is greater than zero.
- Show that the condition of continuity in Theorem 13.9 is necessary by finding functions  $f$  and  $g$  for which  $\lim_{x \rightarrow c} g(x) = L$ , and  $f$  is not continuous at  $x = L$ , and  $\lim_{x \rightarrow c} f(g(x)) \neq f\left(\lim_{x \rightarrow c} g(x)\right)$ .

### 13.6 Limits at Infinity

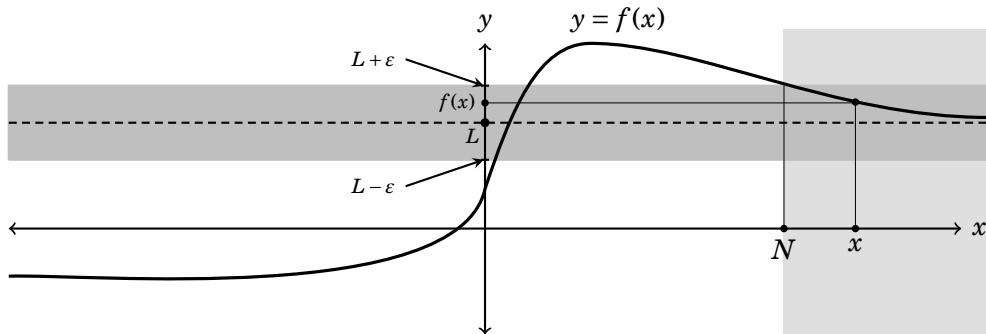
For some functions  $f(x)$ , limits such as  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow -\infty} f(x)$  make sense. Consider the function graphed below. As  $x$  moves to the right (towards positive infinity) the corresponding  $f(x)$  value approaches 2. We express this in symbols as  $\lim_{x \rightarrow \infty} f(x) = 2$ . Such a limit is called a *limit at infinity*, which is a bit of a misnomer because  $x$  is never “at” infinity, just moving toward it.



The graph squeezes in on the dashed horizontal line  $y = 2$  as  $x$  moves to  $\infty$ . This line is called a **horizontal asymptote** of the function  $f(x)$ . It is not a part of the graph, but it helps us visualize the behavior of  $f(x)$  as  $x$  grows.

Also, in this picture, as  $x$  moves to the left (toward negative infinity), the corresponding value  $f(x)$  approaches  $-1$ . We express this in symbols as  $\lim_{x \rightarrow -\infty} f(x) = -1$ . The horizontal line  $y = -1$  is a second horizontal asymptote of this function  $f(x)$ .

In general,  $\lim_{x \rightarrow \infty} f(x) = L$  means that  $f(x)$  is arbitrarily close to  $L$ , provided that  $x$  is sufficiently large (i.e., “provided that  $x$  is sufficiently close to  $\infty$ ”). In other words, given any  $\varepsilon > 0$ , there is a number  $N > 0$  (possibly quite large) such that  $x > N$  implies  $|f(x) - L| < \varepsilon$ . This is illustrated below.



Analogously, for  $x$  approaching  $-\infty$ , we say  $\lim_{x \rightarrow -\infty} f(x) = L$  means that  $f(x)$  is arbitrarily close to  $L$ , provided  $x$  is a sufficiently close to  $-\infty$ . In other words, given any  $\varepsilon > 0$ , there is a number  $N < 0$  such that  $x < N$  implies  $|f(x) - L| < \varepsilon$ . Here is a summary of these ideas.

#### Definition 13.4 (Limits at Infinity)

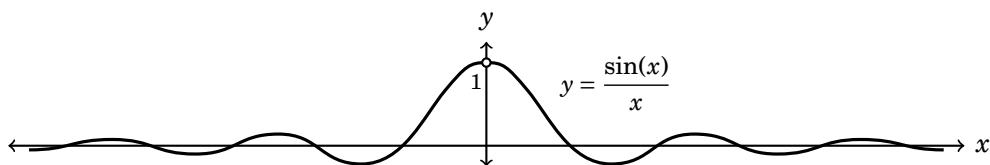
1. The statement  $\lim_{x \rightarrow \infty} f(x) = L$  means that for any real  $\varepsilon > 0$ , there is a number  $N > 0$  for which  $x > N$  implies  $|f(x) - L| \leq \varepsilon$ .
2. The statement  $\lim_{x \rightarrow -\infty} f(x) = L$  means that for any real  $\varepsilon > 0$ , there is a number  $N < 0$  for which  $x < N$  implies  $|f(x) - L| \leq \varepsilon$ .

**Example 13.7** Investigate  $\lim_{x \rightarrow \infty} \frac{\sin(x)}{x}$ .

For any  $x \in \mathbb{R}$ , we know that  $-1 \leq \sin(x) \leq 1$ . Consequently we would expect  $\frac{\sin(x)}{x}$  to be very small when  $x$  is large, that is, we expect  $\lim_{x \rightarrow \infty} \frac{\sin(x)}{x} = 0$ .

Let us use Definition 13.4 to prove this. Given  $\varepsilon > 0$ , put  $N = \frac{1}{\varepsilon}$ . If  $x > N$ , then  $x > \frac{1}{\varepsilon}$ , so  $\frac{1}{x} < \varepsilon$ , and hence  $-\varepsilon < \frac{1}{x} \sin(x) < \varepsilon$ , meaning  $\left| \frac{\sin(x)}{x} \right| < \varepsilon$ .

In summary, given  $\varepsilon > 0$ , there is an  $N > 0$  for which  $x > N$  implies  $\left| \frac{\sin(x)}{x} - 0 \right| < \varepsilon$ . By Definition 13.4,  $\lim_{x \rightarrow \infty} \frac{\sin(x)}{x} = 0$ .



In a similar manner we can prove  $\lim_{x \rightarrow -\infty} \frac{\sin(x)}{x} = 0$ . Thus the  $x$ -axis  $y = 0$  is a horizontal asymptote to  $\frac{\sin(x)}{x}$ , as illustrated above.

Of course, not every limit at infinity will exist. Consider  $\lim_{x \rightarrow \infty} x^2$ . As  $x$  goes to infinity, the quantity  $x^2$  approaches infinity too. Common sense says the limit does not exist because  $x^2$  eventually exceeds any finite number  $L$ . But it's good practice to prove this common-sensical statement.

Suppose for the sake of contradiction that  $\lim_{x \rightarrow \infty} x^2 = L$  for some  $L \in \mathbb{R}$ . Let  $\epsilon = 1$ , and apply Definition 13.4 to get a number  $N$  for which  $x > N$  implies  $|x^2 - L| < 1$ . The inequality (13.3) yields  $|x^2| - |L| = |x^2| - | - L| \leq |x^2 + (-L)| = |x^2 - L| < 1$ . In other words,  $x^2 - |L| < 1$ , or  $x^2 < 1 + |L|$  for all  $x > N$ . But this is false for those  $x$  that are bigger than both  $N$  and  $1 + |L|$ , a contradiction.

Even though  $\lim_{x \rightarrow \infty} x^2$  does not exist, we allow the notation  $\lim_{x \rightarrow \infty} x^2 = \infty$  to indicate that  $x^2$  grows without bound as  $x$  goes to infinity. In general,  $\lim_{x \rightarrow \infty} f(x) = \infty$  means that  $f(x)$  eventually exceeds any number  $L$ :

1.  $\lim_{x \rightarrow \infty} f(x) = \infty$  means that for any real number  $L$ , there is a positive  $N$  for which  $x > N$  implies  $f(x) > L$ .
2.  $\lim_{x \rightarrow \infty} f(x) = -\infty$  means that for any real number  $L$ , there is a positive  $N$  for which  $x > N$  implies  $f(x) < L$ .

Limits of the form  $\lim_{x \rightarrow \infty} f(x) = \pm\infty$  play a small role in the next section.

## Exercises for Section 13.6

Use Definition 13.4 to prove the following results. (Where appropriate, you may wish to adapt the corresponding proofs from Section 13.4.)

1.  $\lim_{x \rightarrow \infty} \frac{1}{x^n} = 0$  if  $n \in \mathbb{N}$ .
2.  $\lim_{x \rightarrow \infty} \frac{3x+2}{2x-1} = \frac{3}{2}$ .
3. If  $a \in \mathbb{R}$ , then  $\lim_{x \rightarrow \infty} a = a$ .
4. If  $\lim_{x \rightarrow \infty} f(x)$  exists, and  $a \in \mathbb{R}$ , then  $\lim_{x \rightarrow \infty} af(x) = a \lim_{x \rightarrow \infty} f(x)$ .
5. If both  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow \infty} g(x)$  exist, then  $\lim_{x \rightarrow \infty} (f(x) + g(x)) = \lim_{x \rightarrow \infty} f(x) + \lim_{x \rightarrow \infty} g(x)$ .
6. If both  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow \infty} g(x)$  exist, then  $\lim_{x \rightarrow \infty} f(x)g(x) = (\lim_{x \rightarrow \infty} f(x)) \cdot (\lim_{x \rightarrow \infty} g(x))$ .
7. If both  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow \infty} g(x)$  exist, then  $\lim_{x \rightarrow \infty} (f(x) - g(x)) = \lim_{x \rightarrow \infty} f(x) - \lim_{x \rightarrow \infty} g(x)$ .
8. If both  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow \infty} g(x)$  exist, and  $\lim_{x \rightarrow \infty} g(x) \neq 0$ , then  $\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = \frac{\lim_{x \rightarrow \infty} f(x)}{\lim_{x \rightarrow \infty} g(x)}$ .
9. If  $\lim_{x \rightarrow \infty} g(x) = L$  and  $f$  is continuous at  $x = L$ , then  $\lim_{x \rightarrow \infty} f(g(x)) = f(\lim_{x \rightarrow \infty} g(x))$ .
10. Prove that  $\lim_{x \rightarrow \infty} \sin(x)$  does not exist.

### 13.7 Sequences

Our final two sections treat sequences and series, topics usually covered in a second semester of calculus.

Recall that a **sequence** is an infinitely long list of real numbers

$$a_1, a_2, a_3, a_4, a_5, \dots$$

The number  $a_1$  is called the *first term*,  $a_2$  is the *second term*,  $a_3$  is the *third term*, and so on. For example, the sequence

$$2, \frac{3}{4}, \frac{4}{9}, \frac{5}{16}, \frac{6}{25}, \frac{7}{36}, \dots$$

has  $n$ th term  $a_n = \frac{n+1}{n^2}$ . The  $n$ th term is sometimes called the **general term**.

We can define a sequence by giving a formula for its general term. The sequence with general term  $a_n = \frac{(-1)^{n+1}(n+1)}{n}$  is

$$2, -\frac{3}{2}, \frac{4}{3}, -\frac{5}{4}, \frac{6}{5}, -\frac{7}{5}, \dots$$

We denote a sequence with  $n$ th term  $a_n$  as  $\{a_n\}$ . For example, the three sequences displayed above are denoted compactly as  $\{a_n\}$  and  $\{\frac{n+1}{n^2}\}$  and  $\{\frac{(-1)^{n+1}(n+1)}{n}\}$ , respectively. In this manner, the sequence  $\{n^2 + 1\}$  is

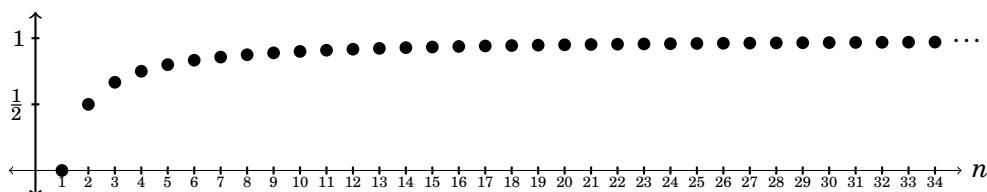
$$2, 5, 10, 17, 26, 37, \dots$$

Sometimes we define a sequence by writing down its first several terms, with the agreement that the general term is implied by the number pattern. For instance, the sequence

$$1, 4, 9, 16, 25, \dots$$

is understood to be  $\{n^2\}$  because  $n^2$  is the most obvious formula that matches the first six terms. But be alert to the fact that a finite number of terms can *never* completely and unambiguously specify an infinite sequence. For all we know, the  $n$ th term of 1, 4, 9, 16, 25, ... might not be  $a_n = n^2$ , but actually  $a_n = n^2 + (n-1)(n-2)(n-3)(n-4)(n-5)$ . This agrees with the first five listed terms, but the sixth term is  $a_6 = 156$ , not the expected  $a_6 = 36$ .

A sequence  $\{a_n\}$  can be regarded as a function  $f : \mathbb{N} \rightarrow \mathbb{R}$ , where  $f(n) = a_n$ . For example, the sequence  $\{1 - \frac{1}{n}\}$  is the function  $f(n) = 1 - \frac{1}{n}$ . In this sense we can graph a sequence; but the graph looks like a string of beads rather than a curve, because the domain is  $\mathbb{N}$ , not  $\mathbb{R}$ . Here is the graph of  $\{1 - \frac{1}{n}\}$ .



Roughly speaking, we say a sequence  $\{a_n\}$  *converges* to a number  $L$  if the numbers  $a_n$  get closer and closer to  $L$  as  $n$  gets bigger and bigger.

For example, the sequence  $\{1 - \frac{1}{n}\}$  from the previous page converges to  $L = 1$ , because as  $n$  gets big, the number  $1 - \frac{1}{n}$  approaches 1.

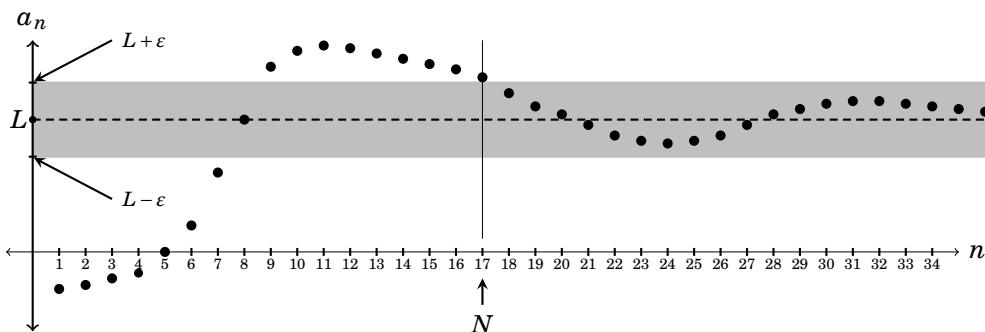
In general, proving facts about convergence requires a precise definition. For this, we can adapt the definition of a limit at infinity from Section 13.6. The sequence  $\{a_n\}$  converges to  $L$  if  $a_n$  can be made arbitrarily close to  $L$  by choosing  $n$  sufficiently large. Here is the exact definition.

**Definition 13.5** A sequence  $\{a_n\}$  **converges** to a number  $L \in \mathbb{R}$  provided that for any  $\varepsilon > 0$  there is an  $N \in \mathbb{N}$  for which  $n > N$  implies  $|a_n - L| < \varepsilon$ .

If  $\{a_n\}$  converges to  $L$ , we denote this state of affairs as  $\lim_{n \rightarrow \infty} a_n = L$ .

If  $\{a_n\}$  does not converge to any number  $L$ , we say it **diverges**.

Definition 13.5 is illustrated below. For any  $\varepsilon > 0$  (no matter how small), there is an integer  $N$  for which the terms  $a_n$  of the sequence lie between  $L - \varepsilon$  and  $L + \varepsilon$  provided  $n > N$ . Smaller values of  $\varepsilon$  require larger values of  $N$ . But no matter how small  $\varepsilon$  is, there is a (possibly quite large) number  $N$  for which  $a_n$  is within  $\varepsilon$  units from  $L$  when  $n > N$ .



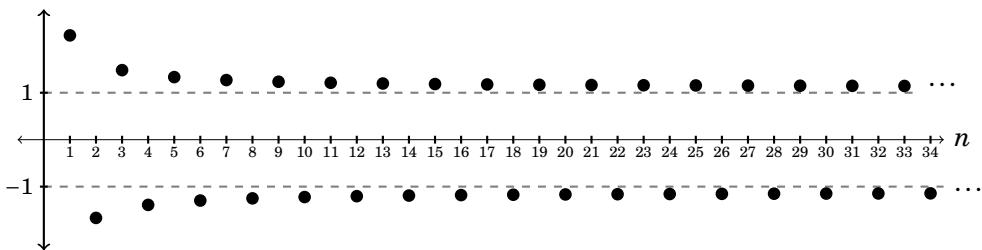
For our first example, let's return to the sequence  $\{1 - \frac{1}{n}\}$ , which is graphed on the previous page (page 261). Notice that as  $n$  gets large,  $\frac{1}{n}$  approaches 0, and  $1 - \frac{1}{n}$  approaches 1. So we can see that the sequence converges to 1. But let's prove this, in order to illustrate Definition 13.5.

**Example 13.8** Prove that the sequence  $\{1 - \frac{1}{n}\}$  converges to 1.

*Proof.* Suppose  $\varepsilon > 0$ . Choose an integer  $N > \frac{1}{\varepsilon}$ , so that  $\frac{1}{N} < \varepsilon$ . Then if  $n > N$  we have  $|a_n - 1| = |(1 - \frac{1}{n}) - 1| = \frac{1}{n} < \frac{1}{N} < \varepsilon$ . By Definition 13.5 the sequence  $\{1 - \frac{1}{n}\}$  converges to 1. ■

**Example 13.9** Investigate the sequence  $\left\{ \frac{(-1)^{n+1}(n+1)}{n} \right\}$ .

The first few terms of this sequence are  $2, -\frac{3}{2}, \frac{4}{3}, -\frac{5}{4}, \frac{6}{5}, -\frac{7}{6}, \dots$ . The terms alternate between positive and negative, with the odd terms positive and the even terms negative. Here is a graph of the sequence.



The picture suggests that as  $n$  increases, the terms bounce back and forth between values that are alternately close to 1 and  $-1$ . This is also evident by inspection of the general term  $a_n = \frac{(-1)^{n+1}(n+1)}{n}$ , because  $\frac{n+1}{n}$  approaches 1 as  $n$  grows toward infinity, while the power of  $-1$  alternates the sign. Because the general term does not approach any single number, it appears that this sequence diverges. Now let's set out to prove this. Our proof formalizes the idea that if the sequence *did* converge to a number  $L$ , then  $L$  would have to be within  $\varepsilon$  units of *both*  $-1$  and  $1$ , and this is impossible if  $\varepsilon < 1$ .

*Proof.* Suppose for the sake of contradiction that the sequence  $\left\{ \frac{(-1)^{n+1}(n+1)}{n} \right\}$  converges to a real number  $L$ . Let  $\varepsilon = 1$ . By Definition 13.5 there is an  $N \in \mathbb{N}$  for which  $n > N$  implies  $\left| \frac{(-1)^{n+1}(n+1)}{n} - L \right| < 1$ .

If  $n$  is odd, then the  $n$ th term of the sequence is  $a_n = \frac{(-1)^{n+1}(n+1)}{n} = \frac{n+1}{n} > 1$ . For  $n$  even, the  $n$ th term of the sequence is  $a_n = \frac{(-1)^{n+1}(n+1)}{n} = -\frac{n+1}{n} < -1$ . Take an odd number  $m > N$  and an even number  $n > N$ . The above three lines yield

$$\begin{aligned} 2 &= 1 - (-1) < a_m - a_n && (\text{because } 1 < a_m \text{ and } 1 < -a_n) \\ &= |a_m - a_n| && (a_m - a_n \text{ is positive}) \\ &= |(a_m - L) - (a_n - L)| && (\text{add } 0 = L - L \text{ to } a_m - a_n) \\ &\leq |a_m - L| + |a_n - L| && (\text{using } |x - y| < |x| + |y|) \\ &< 1 + 1 = 2. && (\text{because } |a_n - L| \leq 1 \text{ when } n > N) \end{aligned}$$

Thus  $2 < 2$ , which is a contradiction. Consequently the series diverges. ■

For another example of a sequence that diverges, consider  $1, 4, 9, 16, 25, \dots$  whose  $n$ th term is  $a_n = n^2$ . Clearly this diverges, because  $\lim_{n \rightarrow \infty} n^2 = \infty$ , which is not a number. In such a case we say that the sequence *diverges to  $\infty$* .

**Definition 13.6** (Divergence to infinity)

1. We say a sequence  $\{a_n\}$  **diverges to  $\infty$**  if  $\lim_{n \rightarrow \infty} a_n = \infty$ . This means that for any  $L > 0$ , there is a positive  $N$  for which  $n > N$  implies  $a_n > L$ .
2. We say a sequence  $\{a_n\}$  **diverges to  $-\infty$**  if  $\lim_{n \rightarrow \infty} a_n = -\infty$ . This means that for any  $L < 0$ , there is a positive  $N$  for which  $n > N$  implies  $a_n < L$ .

This definition spells out a condition called *divergence to  $\infty$* . But we haven't yet proved that a sequence meeting this condition actually diverges in the sense of Definition 13.5. Exercise 7 below asks you to do this.

---

**Exercises for Section 13.7**

1. Prove that  $\left\{\frac{2^n}{n!}\right\}$  converges to 0.
  2. Prove that  $\left\{5 + \frac{2}{n^2}\right\}$  converges to 5.
  3. Prove that  $\left\{\frac{2n^2+1}{3n-1}\right\}$  diverges to  $\infty$ .
  4. Prove that  $\left\{1 - \frac{1}{2^n}\right\}$  converges to 1.
  5. Prove that  $\left\{\frac{2n+1}{3n-1}\right\}$  converges to  $\frac{2}{3}$ .
  6. Prove that  $\left\{\frac{5n^2+n+1}{4n^2+2}\right\}$  converges to  $\frac{5}{4}$ .
  7. Prove that if a sequence diverges to infinity, then it diverges.
  8. Prove that the **constant sequence**  $c, c, c, c, \dots$  converges to  $c$ , for any  $c \in \mathbb{R}$ .
  9. Prove that if  $\{a_n\}$  converges to  $L$ , and  $c \in \mathbb{R}$ , then the sequence  $\{ca_n\}$  converges to  $cL$ .
  10. Prove that if  $\{a_n\}$  converges to  $L$  and  $\{b_n\}$  converges to  $M$ , then the sequence  $\{a_n + b_n\}$  converges to  $L + M$ .
  11. Prove that if  $\{a_n\}$  converges to  $L$  and  $\{b_n\}$  converges to  $M$ , then the sequence  $\{a_n b_n\}$  converges to  $LM$ .
  12. Prove that if  $\{a_n\}$  converges to  $L$  and  $\{b_n\}$  converges to  $M \neq 0$ , then the sequence  $\left\{\frac{a_n}{b_n}\right\}$  converges to  $\frac{L}{M}$ . (You may assume  $b_n \neq 0$  for each  $n \in \mathbb{N}$ .)
  13. For any sequence  $\{a_n\}$ , there is a corresponding sequence  $\{|a_n|\}$ . Prove that if  $\{|a_n|\}$  converges to 0, then  $\{a_n\}$  converges to 0. Give an example of a sequence  $\{a_n\}$  for which  $\{|a_n|\}$  converges to a number  $L \neq 0$ , but  $\{a_n\}$  diverges.
  14. Suppose that  $\{a_n\}$ ,  $\{b_n\}$ , and  $\{c_n\}$  are sequences for which  $a_n \leq b_n \leq c_n$  for all sufficiently large  $n$ . (That is,  $a_n \leq b_n \leq c_n$  for all  $n > M$  for some integer  $M$ .) Prove that if  $\{a_n\}$  and  $\{c_n\}$  converge to  $L$ , then  $\{b_n\}$  also converges to  $L$ .
-

### 13.8 Series

You may recall from your calculus course that there is a big difference between a *sequence* and a *series*.

A **sequence** is an infinite list  $a_1, a_2, a_3, a_4, a_5, a_6, \dots$ .

But a **series** is an infinite sum  $a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + \dots$ .

We use the notation  $\{a_n\}$  to denote the sequence  $a_1, a_2, a_3, a_4, \dots$ , but we use sigma notation to denote a series:

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + \dots$$

For example,

$$\sum_{k=1}^{\infty} \frac{1}{2^k} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \dots$$

You may have a sense that this series should sum to 1, whereas

$$\sum_{k=1}^{\infty} \frac{k+1}{k} = \frac{2}{1} + \frac{3}{2} + \frac{4}{3} + \frac{5}{4} + \frac{6}{5} + \frac{7}{6} + \dots$$

equals  $\infty$  because every fraction in the infinite sum is greater than 1.

Series are significant in calculus because many complex functions can be expressed as series involving terms built from simple algebraic operations. For example, your calculus course may have developed the *Maclaurin series* for various functions, such as

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^{10}}{10!} + \dots$$

But before we make any progress with series, it is essential that we clearly specify what it means to add up infinitely many numbers. We need to understand the situations in which this does and does not make sense.

The key to codifying whether or not a series

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + a_9 \dots$$

adds up to a finite number is to terminate it at an arbitrary  $n$ th term:

$$\sum_{k=1}^n a_k = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + \dots + a_n.$$

This is sum called the  $n$ th **partial sum** of the series, and is denoted as  $s_n$ .

The series has a partial sum  $s_n$  for each positive integer  $n$ :

$$\begin{aligned}s_1 &= a_1 \\ s_2 &= a_1 + a_2 \\ s_3 &= a_1 + a_2 + a_3 \\ s_4 &= a_1 + a_2 + a_3 + a_4 \\ s_5 &= a_1 + a_2 + a_3 + a_4 + a_5 \\ &\vdots \\ s_n &= a_1 + a_2 + a_3 + a_4 + a_5 + \cdots + a_n = \sum_{k=1}^n a_k. \\ &\vdots\end{aligned}$$

If indeed the infinite sum  $S = \sum_{k=1}^{\infty} a_k$  makes sense, then we expect that the partial sum  $s_n = \sum_{k=1}^n a_k$  is a very good approximation to  $S$  when  $n$  is large. Moreover, the larger  $n$  gets, the closer  $s_n$  should be to  $S$ . In other words, the sequence  $s_1, s_2, s_3, s_4, s_5, \dots$  of partial sums should converge to  $S$ . This leads to our main definition. We say that an infinite series *converges* if its sequence of partial sums converges.

**Definition 13.7** A series  $\sum_{k=1}^{\infty} a_k$  **converges** to a real number  $S$  if its sequence of partial sums  $\{s_n\}$  converges to  $S$ . In this case we say  $\sum_{k=1}^{\infty} a_k = S$ .

We say  $\sum_{k=1}^{\infty} a_k$  **diverges** if the sequence  $\{s_n\}$  diverges. In this case  $\sum_{k=1}^{\infty} a_k$  does not make sense as a sum or does not sum to a finite number.

**Example 13.10** Prove that  $\sum_{k=1}^{\infty} \frac{1}{2^k} = 1$ .

*Proof.* Consider the partial sum  $s_n = \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^n}$ . We can get a neat formula for  $s_n$  by noting  $s_n = 2s_n - s_n$ . Then simplify and cancel like terms:

$$\begin{aligned}s_n &= 2s_n - s_n = 2\left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{n-1}} + \frac{1}{2^n}\right) - \left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{n-1}} + \frac{1}{2^n}\right) \\ &= \left(\frac{2}{2^1} + \frac{2}{2^2} + \frac{2}{2^3} + \cdots + \frac{2}{2^{n-1}} + \frac{2}{2^n}\right) - \left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{n-1}} + \frac{1}{2^n}\right) \\ &= \left(1 + \frac{1}{2^1} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-2}} + \frac{1}{2^{n-1}}\right) - \left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{n-1}} + \frac{1}{2^n}\right) = 1 - \frac{1}{2^n}.\end{aligned}$$

Thus  $s_n = 1 - \frac{1}{2^n}$ , so the sequence of partial sums is  $\{s_n\} = \{1 - \frac{1}{2^n}\}$ , which converges to 1 by Exercise 13.7.4. Definition 13.7 yields  $\sum_{k=1}^{\infty} \frac{1}{2^k} = 1$ . ■

Despite the previous example, in practice definitions 13.7 and 13.5 are rarely used to prove that a particular sequence or series converges to a particular number. Instead we tend to use a multitude of convergence tests that are covered in a typical calculus course. Examples of such tests include the the *comparison test*, the *ratio test*, the *root test* and the *alternating series test*. You learned how to use these tests and techniques in your calculus course, though that course may not have actually *proved* that the tests were valid. The point of our present discussion is that definitions 13.7 and 13.5 can be used to prove the tests. To underscore this point, this section's exercises ask you to prove several convergence tests.

By way of illustration, we close with a proof of a theorem that leads to a test for divergence.

**Theorem 13.11** If  $\sum_{k=1}^{\infty} a_k$  converges, then the sequence  $\{a_n\}$  converges to 0.

*Proof.* We use direct proof. Suppose  $\sum_{k=1}^{\infty} a_k$  converges, and say  $\sum_{k=1}^{\infty} a_k = S$ .

Then by Definition 13.7, the sequence of partial sums  $\{s_n\}$  converges to  $S$ . From this, Definition 13.5 says that for any  $\varepsilon > 0$  there is an  $N \in \mathbb{N}$  for which  $n > N$  implies  $|s_n - S| < \varepsilon$ . Thus also  $n - 1 > N$  implies  $|s_{n-1} - S| < \varepsilon$ .

We need to show that  $\{a_n\}$  converges to 0. So take  $\varepsilon > 0$ . By the previous paragraph, there is an  $N' \in \mathbb{N}$  for which  $n > N'$  implies  $|s_n - S| < \frac{\varepsilon}{2}$  and  $|s_{n-1} - S| < \frac{\varepsilon}{2}$ . Notice that  $a_n = s_n - s_{n-1}$  for any  $n > 2$ . So if  $n > N'$  we have

$$\begin{aligned}|a_n - 0| &= |s_n - s_{n-1}| = |(s_n - S) - (s_{n-1} - S)| \\ &\leq |s_n - S| + |s_{n-1} - S| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Therefore, by Definition 13.5, the sequence  $\{a_n\}$  converges to 0. ■

The contrapositive of this theorem is a convenient test for divergence:

**Corollary 13.1** (Divergence test) If  $\{a_n\}$  diverges, or if it converges to a non-zero number, then  $\sum_{k=1}^{\infty} a_k$  diverges.

For example, according to the divergence test, the series  $\sum_{k=1}^{\infty} (1 - \frac{1}{k})$  diverges, because the sequence  $\{1 - \frac{1}{n}\}$  converges to 1. Also,  $\sum_{k=1}^{\infty} \frac{(-1)^{k+1}(k+1)}{k}$  diverges because  $\left\{ \frac{(-1)^{n+1}(n+1)}{n} \right\}$  diverges. (See Example 13.9 on page 263.)

The divergence test gives only a criterion for deciding if a series diverges. It says nothing about convergence. If  $\{a_n\}$  converges to 0, then  $\sum_{k=1}^{\infty} a_k$  may

or may not converge, depending on the particular series. Certainly if  $\sum_{k=1}^{\infty} a_k$  converges, then  $\{a_n\}$  converges to 0, by Theorem 13.11. But  $\{a_n\}$  converging to 0 does not necessarily mean that  $\sum_{k=1}^{\infty} a_k$  converges. A significant example of this is the so-called **harmonic series**:

$$\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \dots.$$

According to Exercise 21 in Chapter 10, if we go out as far as  $2^n$  terms, then the  $2^n$ th partial sum satisfies

$$s_{2^n} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots + \frac{1}{2^{n-1}} + \frac{1}{2^n} \geq 1 + \frac{n}{2}.$$

Because  $1 + \frac{n}{2}$  grows arbitrarily large as  $n$  increases, the sequence of partial sums diverges to  $\infty$ . Consequently the harmonic series diverges.

### Exercises for Section 13.8

Use Definition 13.7 (and Definition 13.5, as needed) to prove the following results. Solutions for these exercises are not included in the back of the book, for they can be found in most good calculus texts. In the exercises we abbreviate  $\sum_{k=1}^{\infty} a_k$  as  $\sum a_k$ .

1. A *geometric series* is one having the form  $a + ar + ar^2 + ar^3 \dots$ , where  $a, r \in \mathbb{R}$ . (The first term in the sum is  $a$ , and beyond that, the  $k$ th term is  $r$  times the previous term.) Prove that if  $|r| < 1$ , then the series converges to  $\frac{a}{1-r}$ . Also, if  $a \neq 0$  and  $|r| \geq 1$ , then the series diverges. (If you need guidance, you may draw inspiration from Example 13.10, which concerns a geometric series with  $a = r = \frac{1}{2}$ .)
2. Prove the *comparison test*: Suppose  $\sum a_k$  and  $\sum b_k$  are series. If  $0 \leq a_k \leq b_k$  for each  $k$ , and  $\sum b_k$  converges, then  $\sum a_k$  converges. Also, if  $0 \leq b_k \leq a_k$  for each  $k$ , and  $\sum b_k$  diverges, then  $\sum a_k$  diverges.
3. Prove the *limit comparison test*: Suppose  $\sum a_k$  and  $\sum b_k$  are series for which  $a_k, b_k > 0$  for each  $k$ . If  $\lim_{n \rightarrow \infty} \frac{a_k}{b_k} = 0$  and  $\sum b_k$  converges, then  $\sum a_k$  converges. (Your proof may use any of the above exercises.)
4. Prove the *absolute convergence test*: Let  $\sum a_k$  be a series. If  $\sum |a_k|$  converges, then  $\sum a_k$  converges. (Your proof may use any of the above exercises.)
5. Prove the *ratio test*: Given a series  $\sum a_k$  with each  $a_k$  positive, if  $\lim_{n \rightarrow \infty} \frac{a_{k+1}}{a_k} = L < 1$ , then  $\sum a_k$  converges. Also, if  $L > 1$ , then  $\sum a_k$  diverges. (Your proof may use any of the above exercises.)

---

## Cardinality of Sets

---

This chapter is all about cardinality of sets. At first this looks like a very simple concept. To find the cardinality of a set, just count its elements. If  $A = \{a, b, c, d\}$ , then  $|A| = 4$ ; if  $B = \{n \in \mathbb{Z} : -5 \leq n \leq 5\}$ , then  $|B| = 11$ . In this case  $|A| < |B|$ . What could be simpler than that?

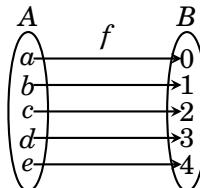
Actually, the idea of cardinality becomes quite subtle when the sets are infinite. The main point of this chapter is to explain how there are numerous different kinds of infinity, and some infinities are bigger than others. Two sets  $A$  and  $B$  can both have infinite cardinality, yet  $|A| < |B|$ .

### 14.1 Sets with Equal Cardinalities

We begin with a discussion of what it means for two sets to have the same cardinality. Up until this point we've said  $|A| = |B|$  if  $A$  and  $B$  have the same number of elements: Count the elements of  $A$ , then count the elements of  $B$ . If you get the same number, then  $|A| = |B|$ .

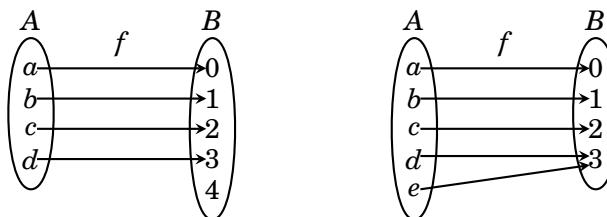
Although this is a fine strategy if the sets are finite (and not too big!), it doesn't apply to infinite sets because we'd never be done counting their elements. We need a new approach that applies to both finite and infinite sets. Here it is:

**Definition 14.1** Two sets  $A$  and  $B$  have the **same cardinality**, written  $|A| = |B|$ , if there exists a bijective function  $f : A \rightarrow B$ . If no such bijective  $f$  exists, then the sets have **unequal cardinalities**, written  $|A| \neq |B|$ .



The above picture illustrates our definition. There is a bijective function  $f : A \rightarrow B$ , so  $|A| = |B|$ . The function  $f$  matches up  $A$  with  $B$ . Think of  $f$  as describing how to overlay  $A$  onto  $B$  so that they fit together perfectly.

On the other hand, if  $A$  and  $B$  are as indicated in either of the following figures, then there can be no bijection  $f : A \rightarrow B$ . (The best we can do is a function that is either injective or surjective, but not both.) Therefore the definition says  $|A| \neq |B|$  in these cases.



**Example 14.1** The sets  $A = \{n \in \mathbb{Z} : 0 \leq n \leq 5\}$  and  $B = \{n \in \mathbb{Z} : -5 \leq n \leq 0\}$  have the same cardinality because there is a bijective function  $f : A \rightarrow B$  given by the rule  $f(n) = -n$ .

Several comments are in order. First, if  $|A| = |B|$ , there can be *lots* of bijective functions from  $A$  to  $B$ . We only need to find one of them in order to conclude  $|A| = |B|$ . Second, as bijective functions play such a big role here, we use the word **bijection** to mean *bijective function*. Thus the function  $f(n) = -n$  from Example 14.1 is a bijection. Also, an injective function is called an **injection** and a surjective function is called a **surjection**.

We emphasize and reiterate that Definition 14.1 applies to finite as well as infinite sets. If  $A$  and  $B$  are infinite, then  $|A| = |B|$  provided there exists a bijection  $f : A \rightarrow B$ . If no such bijection exists, then  $|A| \neq |B|$ .

**Example 14.2** This example shows that  $|\mathbb{N}| = |\mathbb{Z}|$ . To see why this is true, notice that the following table describes a bijection  $f : \mathbb{N} \rightarrow \mathbb{Z}$ .

$n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
$f(n)$	0	1	-1	2	-2	3	-3	4	-4	5	-5	6	-6	7	-7	...

Notice that  $f$  is described in such a way that it is both injective and surjective. Every integer appears exactly once on the infinitely long second row. Thus, according to the table, given any  $b \in \mathbb{Z}$  there is some natural number  $n$  with  $f(n) = b$ , so  $f$  is surjective. It is injective because the way the table is constructed forces  $f(m) \neq f(n)$  whenever  $m \neq n$ . Because of this bijection  $f : \mathbb{N} \rightarrow \mathbb{Z}$ , we must conclude from Definition 14.1 that  $|\mathbb{N}| = |\mathbb{Z}|$ .

Example 14.2 may seem slightly unsettling. On one hand it makes sense that  $|\mathbb{N}| = |\mathbb{Z}|$  because  $\mathbb{N}$  and  $\mathbb{Z}$  are both infinite, so their cardinalities are both “infinity.” On the other hand,  $\mathbb{Z}$  may seem twice as large as  $\mathbb{N}$  because  $\mathbb{Z}$

has all the negative integers as well as the positive ones. Definition 14.1 settles the issue. Because the bijection  $f : \mathbb{N} \rightarrow \mathbb{Z}$  matches up  $\mathbb{N}$  with  $\mathbb{Z}$ , it follows that  $|\mathbb{N}| = |\mathbb{Z}|$ . We summarize this with a theorem.

**Theorem 14.1** There exists a bijection  $f : \mathbb{N} \rightarrow \mathbb{Z}$ . Therefore  $|\mathbb{N}| = |\mathbb{Z}|$ .

The fact that  $\mathbb{N}$  and  $\mathbb{Z}$  have the same cardinality might prompt us compare the cardinalities of other infinite sets. How, for example, do  $\mathbb{N}$  and  $\mathbb{R}$  compare? Let's turn our attention to this.

In fact,  $|\mathbb{N}| \neq |\mathbb{R}|$ . This was first recognized by Georg Cantor (1845–1918), who devised an ingenious argument to show that there are no surjective functions  $f : \mathbb{N} \rightarrow \mathbb{R}$ . (This in turn implies that there can be no bijections  $f : \mathbb{N} \rightarrow \mathbb{R}$ , so  $|\mathbb{N}| \neq |\mathbb{R}|$  by Definition 14.1.)

We now describe Cantor's argument for why there are no surjections  $f : \mathbb{N} \rightarrow \mathbb{R}$ . We will reason informally, rather than writing out an exact proof. Take any arbitrary function  $f : \mathbb{N} \rightarrow \mathbb{R}$ . Here's why  $f$  can't be surjective:

Imagine making a table for  $f$ , where values of  $n$  in  $\mathbb{N}$  are in the left-hand column and the corresponding values  $f(n)$  are on the right. The first few entries might look something as follows. In this table, the real numbers  $f(n)$  are written with all their decimal places trailing off to the right. Thus, even though  $f(1)$  happens to be the real number 0.4, we write it as 0.40000000..., etc.

$n$	$f(n)$
1	0 . 4 0 0 0 0 0 0 0 0 0 0 0 0 0 ...
2	8 . 5 0 0 6 0 7 0 8 6 6 6 9 0 0 ...
3	7 . 5 0 5 0 9 4 0 0 4 4 1 0 1 ...
4	5 . 5 0 7 0 4 0 0 8 0 4 8 0 5 0 ...
5	6 . 9 0 0 2 6 0 0 0 0 0 0 5 0 6 ...
6	6 . 8 2 8 0 9 5 8 2 0 5 0 0 2 0 ...
7	6 . 5 0 5 0 5 5 0 6 5 5 8 0 8 ...
8	8 . 7 2 0 8 0 6 4 0 0 0 0 4 4 8 ...
9	0 . 5 5 0 0 0 0 8 8 8 8 0 0 7 7 ...
10	0 . 5 0 0 2 0 7 2 2 0 7 8 0 5 1 ...
11	2 . 9 0 0 0 0 8 8 0 0 0 0 9 0 0 ...
12	6 . 5 0 2 8 0 0 0 8 0 0 9 6 7 1 ...
13	8 . 8 9 0 0 8 0 2 4 0 0 8 0 5 0 ...
14	8 . 5 0 0 0 8 7 4 2 0 8 0 2 2 6 ...
:	:

There is a diagonal shaded band in the table. For each  $n \in \mathbb{N}$ , this band covers the  $n^{\text{th}}$  decimal place of  $f(n)$ :

- The 1st decimal place of  $f(1)$  is the 1st entry on the diagonal.
- The 2nd decimal place of  $f(2)$  is the 2nd entry on the diagonal.
- The 3rd decimal place of  $f(3)$  is the 3rd entry on the diagonal.
- The 4th decimal place of  $f(4)$  is the 4th entry on the diagonal, etc.

The diagonal helps us construct a number  $b \in \mathbb{R}$  that is unequal to any  $f(n)$ . Just let the  $n^{\text{th}}$  decimal place of  $b$  differ from the  $n^{\text{th}}$  entry of the diagonal. Then the  $n^{\text{th}}$  decimal place of  $b$  differs from the  $n^{\text{th}}$  decimal place of  $f(n)$ . In order to be definite, define  $b$  to be the positive number less than 1 whose  $n^{\text{th}}$  decimal place is 0 if the  $n^{\text{th}}$  decimal place of  $f(n)$  does not equal 0, and whose  $n^{\text{th}}$  decimal place is 1 if the  $n^{\text{th}}$  decimal place of  $f(n)$  equals 0. Thus, for the function  $f$  illustrated in the above table, we have

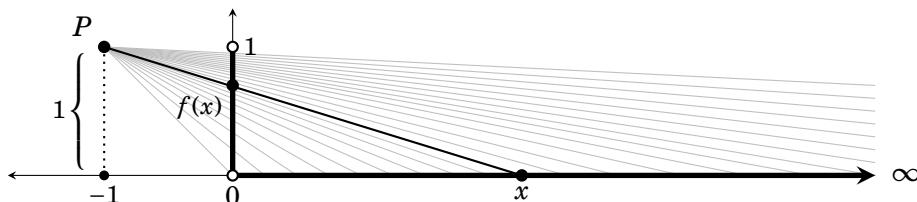
$$b = 0.01010001001000\dots$$

and  $b$  has been defined so that, for any  $n \in \mathbb{N}$ , its  $n^{\text{th}}$  decimal place is unequal to the  $n^{\text{th}}$  decimal place of  $f(n)$ . Therefore  $f(n) \neq b$  for every natural number  $n$ , meaning  $f$  is not surjective.

Since this argument applies to *any* function  $f : \mathbb{N} \rightarrow \mathbb{R}$  (not just the one in the above example) we conclude that there exist no bijections  $f : \mathbb{N} \rightarrow \mathbb{R}$ , so  $|\mathbb{N}| \neq |\mathbb{R}|$  by Definition 14.1. We summarize this as a theorem.

**Theorem 14.2** There exists no bijection  $f : \mathbb{N} \rightarrow \mathbb{R}$ . Therefore  $|\mathbb{N}| \neq |\mathbb{R}|$ .

This is our first indication of how there are different kinds of infinities. Both  $\mathbb{N}$  and  $\mathbb{R}$  are infinite sets, yet  $|\mathbb{N}| \neq |\mathbb{R}|$ . We will continue to develop this theme throughout this chapter. The next example shows that the intervals  $(0, \infty)$  and  $(0, 1)$  on  $\mathbb{R}$  have the same cardinality.



**Figure 14.1.** A bijection  $f : (0, \infty) \rightarrow (0, 1)$ . Imagine a light source at point  $P$ . Then  $f(x)$  is the point on the  $y$ -axis whose shadow is  $x$ .

**Example 14.3** Show that  $|(0, \infty)| = |(0, 1)|$ .

To accomplish this, we need to show that there is a bijection  $f : (0, \infty) \rightarrow (0, 1)$ . We describe this function geometrically. Consider the interval  $(0, \infty)$  as the positive  $x$ -axis of  $\mathbb{R}^2$ . Let the interval  $(0, 1)$  be on the  $y$ -axis as illustrated in Figure 14.1, so that  $(0, \infty)$  and  $(0, 1)$  are perpendicular to each other.

The figure also shows a point  $P = (-1, 1)$ . Define  $f(x)$  to be the point on  $(0, 1)$  where the line from  $P$  to  $x \in (0, \infty)$  intersects the  $y$ -axis. By similar triangles, we have

$$\frac{1}{x+1} = \frac{f(x)}{x},$$

and therefore

$$f(x) = \frac{x}{x+1}.$$

If it is not clear from the figure that  $f : (0, \infty) \rightarrow (0, 1)$  is bijective, then you can verify it using the techniques from Section 12.2. (Exercise 16, below.)

It is important to note that equality of cardinalities is an equivalence relation on sets: it is reflexive, symmetric and transitive. Let us confirm this. Given a set  $A$ , the identity function  $A \rightarrow A$  is a bijection, so  $|A| = |A|$ . (This is the reflexive property.) For the symmetric property, if  $|A| = |B|$ , then there is a bijection  $f : A \rightarrow B$ , and its inverse is a bijection  $f^{-1} : B \rightarrow A$ , so  $|B| = |A|$ . For transitivity, suppose  $|A| = |B|$  and  $|B| = |C|$ . Then there are bijections  $f : A \rightarrow B$  and  $g : B \rightarrow C$ . The composition  $g \circ f : A \rightarrow C$  is a bijection (Theorem 12.2), so  $|A| = |C|$ .

The transitive property can be useful. If, in trying to show two sets  $A$  and  $C$  have the same cardinality, we can produce a third set  $B$  for which  $|A| = |B|$  and  $|B| = |C|$ , then transitivity assures us that indeed  $|A| = |C|$ . The next example uses this idea.

**Example 14.4** Show that  $|\mathbb{R}| = |(0, 1)|$ .

Because of the bijection  $g : \mathbb{R} \rightarrow (0, \infty)$  where  $g(x) = 2^x$ , we have  $|\mathbb{R}| = |(0, \infty)|$ . Also, Example 14.3 shows that  $|(0, \infty)| = |(0, 1)|$ . Therefore  $|\mathbb{R}| = |(0, 1)|$ .

So far in this chapter we have declared that two sets have “the same cardinality” if there is a bijection between them. They have “different cardinalities” if there exists no bijection between them. Using this idea, we showed that  $|\mathbb{Z}| = |\mathbb{N}| \neq |\mathbb{R}| = |(0, \infty)| = |(0, 1)|$ . So, we have a means of determining when two sets have the same or different cardinalities. But we have neatly avoided saying exactly what cardinality *is*. For example, we can say that  $|\mathbb{Z}| = |\mathbb{N}|$ , but what exactly *is*  $|\mathbb{Z}|$ , or  $|\mathbb{N}|$ ? What exactly *are* these things that are equal? Certainly not numbers, for they are too big. And

saying they are “infinity” is not accurate, because we now know that there are different types of infinity. So just what kind of mathematical entity is  $|\mathbb{Z}|$ ? In general, given a set  $X$ , exactly what *is* its cardinality  $|X|$ ?

This is a lot like asking what a number is. A number, say 5, is an abstraction, not a physical thing. Early in life we instinctively grouped together certain sets of things (five apples, five oranges, etc.) and conceived of 5 as the thing common to all such sets. In a very real sense, the number 5 is an abstraction of the fact that any two of these sets can be matched up via a bijection. That is, it can be identified with a certain equivalence class of sets under the “*has the same cardinality as*” relation. (Recall that this is an equivalence relation.) This is easy to grasp because our sense of numeric quantity is so innate. But in exactly the same way we can say that the cardinality of a set  $X$  is what is common to all sets that can be matched to  $X$  via a bijection. This may be harder to grasp, but it is really no different from the idea of the magnitude of a (finite) number.

In fact, we could be concrete and define  $|X|$  to be the equivalence class of all sets whose cardinality is the same as that of  $X$ . This has the advantage of giving an explicit meaning to  $|X|$ . But there is no harm in taking the intuitive approach and just interpreting the cardinality  $|X|$  of a set  $X$  to be a measure the “size” of  $X$ . The point of this section is that we have a means of deciding whether two sets have the same size or different sizes.

### Exercises for Section 14.1

- A.** Show that the two given sets have equal cardinality by describing a bijection from one to the other. Describe your bijection with a formula (not as a table).

- |   |   |
|---|---|
| 1. $\mathbb{R}$ and $(0, \infty)$<br>2. $\mathbb{R}$ and $(\sqrt{2}, \infty)$<br>3. $\mathbb{R}$ and $(0, 1)$<br>4. The set of even integers and<br>the set of odd integers<br>5. $A = \{3k : k \in \mathbb{Z}\}$ and $B = \{7k : k \in \mathbb{Z}\}$ | 6. $\mathbb{N}$ and $S = \left\{ \frac{\sqrt{2}}{n} : n \in \mathbb{N} \right\}$<br>7. $\mathbb{Z}$ and $S = \left\{ \dots, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, \dots \right\}$<br>8. $\mathbb{Z}$ and $S = \{x \in \mathbb{R} : \sin x = 1\}$<br>9. $\{0, 1\} \times \mathbb{N}$ and $\mathbb{N}$<br>10. $\{0, 1\} \times \mathbb{N}$ and $\mathbb{Z}$<br>11. $[0, 1]$ and $(0, 1)$<br>12. $\mathbb{N}$ and $\mathbb{Z}$ (Suggestion: use Exercise 18 of Section 12.2.)<br>13. $\mathcal{P}(\mathbb{N})$ and $\mathcal{P}(\mathbb{Z})$ (Suggestion: use Exercise 12, above.)<br>14. $\mathbb{N} \times \mathbb{N}$ and $\{(n, m) \in \mathbb{N} \times \mathbb{N} : n \leq m\}$ |
|---|---|

- B.** Answer the following questions concerning bijections from this section.

15. Find a formula for the bijection  $f$  in Example 14.2 (page 270).  
 16. Verify that the function  $f$  in Example 14.3 (page 273) is a bijection.

## 14.2 Countable and Uncountable Sets

Let's summarize the main points from the previous section.

1.  $|A| = |B|$  if and only if there exists a bijection  $A \rightarrow B$ .
2.  $|\mathbb{N}| = |\mathbb{Z}|$  because there exists a bijection  $\mathbb{N} \rightarrow \mathbb{Z}$ .
3.  $|\mathbb{N}| \neq |\mathbb{R}|$  because there exists *no* bijection  $\mathbb{N} \rightarrow \mathbb{R}$ .

Thus, even though  $\mathbb{N}$ ,  $\mathbb{Z}$  and  $\mathbb{R}$  are all infinite sets, their cardinalities are not all the same. The sets  $\mathbb{N}$  and  $\mathbb{Z}$  have the same cardinality, but  $\mathbb{R}$ 's cardinality is different from that of both the other sets. This means infinite sets can have different sizes. We now make some definitions to put words and symbols to this phenomenon.

In a certain sense you can count the elements of  $\mathbb{N}$ ; you can count its elements off as  $1, 2, 3, 4, \dots$ , but you'd have to continue this process forever to count the whole set. Thus we will call  $\mathbb{N}$  a *countably infinite set*, and the same term is used for any set whose cardinality equals that of  $\mathbb{N}$ .

**Definition 14.2** Suppose  $A$  is a set. Then  $A$  is **countably infinite** if  $|\mathbb{N}| = |A|$ , that is, if there exists a bijection  $\mathbb{N} \rightarrow A$ . The set  $A$  is **countable** if it is finite or countably infinite. The set  $A$  is **uncountable** if it is infinite and  $|\mathbb{N}| \neq |A|$ , that is, if  $A$  is infinite and there is *no* bijection  $\mathbb{N} \rightarrow A$ .

Thus  $\mathbb{Z}$  is countably infinite but  $\mathbb{R}$  is uncountable. This section deals mainly with countably infinite sets. Uncountable sets are treated later.

If  $A$  is countably infinite, then  $|\mathbb{N}| = |A|$ , so there is a bijection  $f : \mathbb{N} \rightarrow A$ . Think of  $f$  as “counting” the elements of  $A$ . The first element of  $A$  is  $f(1)$ , followed by  $f(2)$ , then  $f(3)$  and so on. It makes sense to think of a countably infinite set as the smallest type of infinite set, because if the counting process stopped, the set would be finite, not infinite; a countably infinite set has the fewest elements that a set can have and still be infinite. We reserve the special symbol  $\aleph_0$  to stand for the cardinality of countably infinite sets.

**Definition 14.3** The cardinality of the natural numbers is denoted as  $\aleph_0$ . That is,  $|\mathbb{N}| = \aleph_0$ . Thus any countably infinite set has cardinality  $\aleph_0$ .

(The symbol  $\aleph$  is the first letter in the Hebrew alphabet, and is pronounced “aleph.” The symbol  $\aleph_0$  is pronounced “aleph naught.”) The summary of facts at the beginning of this section shows  $|\mathbb{Z}| = \aleph_0$  and  $|\mathbb{R}| \neq \aleph_0$ .

**Example 14.5** Let  $E = \{2k : k \in \mathbb{Z}\}$  be the set of even integers. The function  $f : \mathbb{Z} \rightarrow E$  defined as  $f(n) = 2n$  is easily seen to be a bijection, so we have  $|\mathbb{Z}| = |E|$ . Thus, as  $|\mathbb{N}| = |\mathbb{Z}| = |E|$ , the set  $E$  is countably infinite and  $|E| = \aleph_0$ .

Here is a significant fact: The elements of any countably infinite set  $A$  can be written in an infinitely long list  $a_1, a_2, a_3, a_4, \dots$  that begins with some element  $a_1 \in A$  and includes every element of  $A$ . For example, the set  $E$  in the above example can be written in list form as  $0, 2, -2, 4, -4, 6, -6, 8, -8, \dots$  The reason that this can be done is as follows. Since  $A$  is countably infinite, Definition 14.2 says there is a bijection  $f : \mathbb{N} \rightarrow A$ . This allows us to list out the set  $A$  as an infinite list  $f(1), f(2), f(3), f(4), \dots$  Conversely, if the elements of  $A$  can be written in list form as  $a_1, a_2, a_3, \dots$ , then the function  $f : \mathbb{N} \rightarrow A$  defined as  $f(n) = a_n$  is a bijection, so  $A$  is countably infinite. We summarize this as follows.

**Theorem 14.3** A set  $A$  is countably infinite if and only if its elements can be arranged in an infinite list  $a_1, a_2, a_3, a_4, \dots$

As an example of how this theorem might be used, let  $P$  denote the set of all prime numbers. Since we can list its elements as  $2, 3, 5, 7, 11, 13, \dots$ , it follows that the set  $P$  is countably infinite.

As another consequence of Theorem 14.3, note that we can interpret the fact that the set  $\mathbb{R}$  is not countably infinite as meaning that it is impossible to write out all the elements of  $\mathbb{R}$  in an infinite list. (After all, we tried to do that in the table on page 271, and failed!)

This raises a question. Is it also impossible to write out all the elements of  $\mathbb{Q}$  in an infinite list? In other words, is the set  $\mathbb{Q}$  of rational numbers countably infinite or uncountable? If you start plotting the rational numbers on the number line, they seem to mostly fill up  $\mathbb{R}$ . Sure, some numbers such as  $\sqrt{2}$ ,  $\pi$  and  $e$  will not be plotted, but the dots representing rational numbers seem to predominate. We might thus expect  $\mathbb{Q}$  to be uncountable. However, it is a surprising fact that  $\mathbb{Q}$  is countable. The proof presented below arranges all the rational numbers in an infinitely long list.

**Theorem 14.4** The set  $\mathbb{Q}$  of rational numbers is countably infinite.

*Proof.* To prove this, we just need to show how to write the set  $\mathbb{Q}$  in list form. Begin by arranging all rational numbers in an infinite array. This is done by making the following chart. The top row has a list of all integers, beginning with 0, then alternating signs as they increase. Each column headed by an integer  $k$  contains all the fractions (in reduced form) with numerator  $k$ . For example, the column headed by 2 contains the fractions  $\frac{2}{1}, \frac{2}{3}, \frac{2}{5}, \frac{2}{7}, \dots$ , and so on. It does not contain  $\frac{2}{2}, \frac{2}{4}, \frac{2}{6}$ , etc., because those are not reduced, and in fact their reduced forms appear in the column headed by 1. You should examine this table and convince yourself that it contains all rational numbers in  $\mathbb{Q}$ .

0	1	-1	2	-2	3	-3	4	-4	5	-5	...
$\frac{0}{1}$	$\frac{1}{1}$	$-\frac{1}{1}$	$\frac{2}{1}$	$-\frac{2}{1}$	$\frac{3}{1}$	$-\frac{3}{1}$	$\frac{4}{1}$	$-\frac{4}{1}$	$\frac{5}{1}$	$-\frac{5}{1}$	...
$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{2}{3}$	$-\frac{2}{3}$	$\frac{3}{2}$	$-\frac{3}{2}$	$\frac{4}{3}$	$-\frac{4}{3}$	$\frac{5}{2}$	$-\frac{5}{2}$	...	
$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{5}$	$-\frac{2}{5}$	$\frac{3}{4}$	$-\frac{3}{4}$	$\frac{4}{5}$	$-\frac{4}{5}$	$\frac{5}{3}$	$-\frac{5}{3}$	...	
$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{2}{7}$	$-\frac{2}{7}$	$\frac{3}{5}$	$-\frac{3}{5}$	$\frac{4}{7}$	$-\frac{4}{7}$	$\frac{5}{4}$	$-\frac{5}{4}$	...	
$\frac{1}{5}$	$-\frac{1}{5}$	$\frac{2}{9}$	$-\frac{2}{9}$	$\frac{3}{7}$	$-\frac{3}{7}$	$\frac{4}{9}$	$-\frac{4}{9}$	$\frac{5}{6}$	$-\frac{5}{6}$	...	
$\frac{1}{6}$	$-\frac{1}{6}$	$\frac{2}{11}$	$-\frac{2}{11}$	$\frac{3}{8}$	$-\frac{3}{8}$	$\frac{4}{11}$	$-\frac{4}{11}$	$\frac{5}{7}$	$-\frac{5}{7}$	...	
$\frac{1}{7}$	$-\frac{1}{7}$	$\frac{2}{13}$	$-\frac{2}{13}$	$\frac{3}{10}$	$-\frac{3}{10}$	$\frac{4}{13}$	$-\frac{4}{13}$	$\frac{5}{8}$	$-\frac{5}{8}$	...	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Next, draw an infinite path in this array, beginning at  $\frac{0}{1}$  and snaking back and forth as indicated below. Every rational number is on this path.

0	1	-1	2	-2	3	-3	4	-4	5	-5	...
$\frac{0}{1}$	$\frac{1}{1}$	$-\frac{1}{1}$	$\frac{2}{1}$	$-\frac{2}{1}$	$\frac{3}{1}$	$-\frac{3}{1}$	$\frac{4}{1}$	$-\frac{4}{1}$	$\frac{5}{1}$	$-\frac{5}{1}$	...
$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{2}{3}$	$-\frac{2}{3}$	$\frac{3}{2}$	$-\frac{3}{2}$	$\frac{4}{3}$	$-\frac{4}{3}$	$\frac{5}{2}$	$-\frac{5}{2}$	...	
$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{5}$	$-\frac{2}{5}$	$\frac{3}{4}$	$-\frac{3}{4}$	$\frac{4}{5}$	$-\frac{4}{5}$	$\frac{5}{3}$	$-\frac{5}{3}$	...	
$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{2}{7}$	$-\frac{2}{7}$	$\frac{3}{5}$	$-\frac{3}{5}$	$\frac{4}{7}$	$-\frac{4}{7}$	$\frac{5}{4}$	$-\frac{5}{4}$	...	
$\frac{1}{5}$	$-\frac{1}{5}$	$\frac{2}{9}$	$-\frac{2}{9}$	$\frac{3}{7}$	$-\frac{3}{7}$	$\frac{4}{9}$	$-\frac{4}{9}$	$\frac{5}{6}$	$-\frac{5}{6}$	...	
$\frac{1}{6}$	$-\frac{1}{6}$	$\frac{2}{11}$	$-\frac{2}{11}$	$\frac{3}{8}$	$-\frac{3}{8}$	$\frac{4}{11}$	$-\frac{4}{11}$	$\frac{5}{7}$	$-\frac{5}{7}$	...	
$\frac{1}{7}$	$-\frac{1}{7}$	$\frac{2}{13}$	$-\frac{2}{13}$	$\frac{3}{10}$	$-\frac{3}{10}$	$\frac{4}{13}$	$-\frac{4}{13}$	$\frac{5}{8}$	$-\frac{5}{8}$	...	
$\frac{1}{8}$	$-\frac{1}{8}$	$\frac{2}{15}$	$-\frac{2}{15}$	$\frac{3}{11}$	$-\frac{3}{11}$	$\frac{4}{15}$	$-\frac{4}{15}$	$\frac{5}{9}$	$-\frac{5}{9}$	...	
$\frac{1}{9}$	$-\frac{1}{9}$	$\frac{2}{17}$	$-\frac{2}{17}$	$\frac{3}{13}$	$-\frac{3}{13}$	$\frac{4}{17}$	$-\frac{4}{17}$	$\frac{5}{11}$	$-\frac{5}{11}$	...	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Beginning at  $\frac{0}{1}$  and following the path, we get an infinite list of all rational numbers:

$$0, 1, \frac{1}{2}, -\frac{1}{2}, -1, 2, \frac{2}{3}, \frac{2}{5}, -\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, -\frac{1}{4}, \frac{2}{7}, -\frac{2}{7}, -\frac{2}{5}, -\frac{2}{3}, -2, 3, \frac{3}{2}, \dots$$

By Theorem 14.3, it follows that  $\mathbb{Q}$  is countably infinite, that is,  $|\mathbb{Q}| = |\mathbb{N}|$ . ■

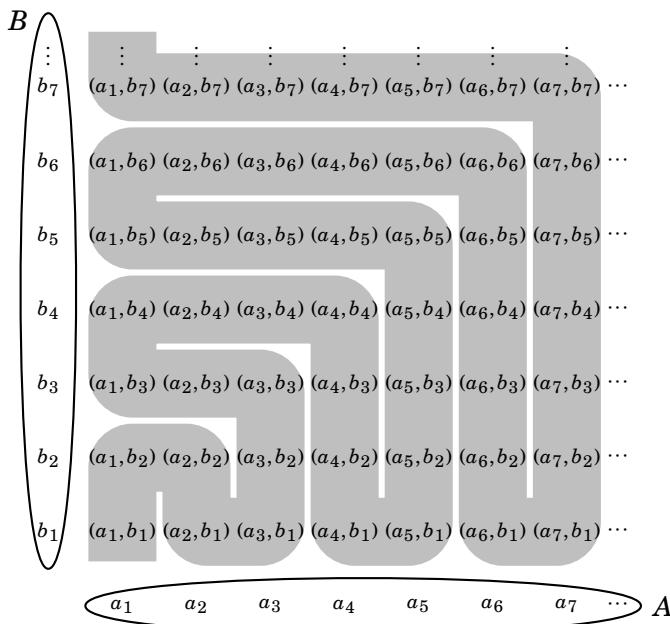
It is also true that the Cartesian product of two countably infinite sets is itself countably infinite, as our next theorem states.

**Theorem 14.5** If  $A$  and  $B$  are both countably infinite, then so is  $A \times B$ .

*Proof.* Suppose  $A$  and  $B$  are both countably infinite. By Theorem 14.3, we know we can write  $A$  and  $B$  in list form as

$$\begin{aligned} A &= \{a_1, a_2, a_3, a_4, \dots\}, \\ B &= \{b_1, b_2, b_3, b_4, \dots\}. \end{aligned}$$

Figure 14.2 shows how to form an infinite path winding through all of  $A \times B$ . Therefore  $A \times B$  can be written in list form, so it is countably infinite. ■



**Figure 14.2.** A product of two countably infinite sets is countably infinite

As an example of a consequence of this theorem, notice that since  $\mathbb{Q}$  is countably infinite, the set  $\mathbb{Q} \times \mathbb{Q}$  is also countably infinite.

Recall that the word “corollary” means a result that follows easily from some other result. We have the following corollary of Theorem 14.5.

**Corollary 14.1** Given  $n$  countably infinite sets  $A_1, A_2, \dots, A_n$ , with  $n \geq 2$ , the Cartesian product  $A_1 \times A_2 \times \dots \times A_n$  is also countably infinite.

*Proof.* The proof is by induction on  $n$ . For the basis step, notice that when  $n = 2$  the statement asserts that for countably infinite sets  $A_1$  and  $A_2$ , the product  $A_1 \times A_2$  is countably infinite, and this is true by Theorem 14.5.

Assume that for some  $k \geq 2$ , any product  $A_1 \times A_2 \times \dots \times A_k$  of countably infinite sets is countably infinite. Consider a product  $A_1 \times A_2 \times \dots \times A_k \times A_{k+1}$  of  $k + 1$  countably infinite sets. It is easy to confirm that the function

$$\begin{aligned} f : A_1 \times A_2 \times A_3 \times \dots \times A_k \times A_{k+1} &\longrightarrow (A_1 \times A_2 \times A_3 \times \dots \times A_k) \times A_{k+1} \\ f(x_1, x_2, \dots, x_k, x_{k+1}) &= ((x_1, x_2, \dots, x_k), x_{k+1}) \end{aligned}$$

is bijective, so  $|A_1 \times A_2 \times A_3 \times \dots \times A_k \times A_{k+1}| = |(A_1 \times A_2 \times A_3 \times \dots \times A_k) \times A_{k+1}|$ . By the induction hypothesis,  $(A_1 \times A_2 \times A_3 \times \dots \times A_k) \times A_{k+1}$  is a product of two countably infinite sets, so it is countably infinite by Theorem 14.5. As noted above,  $A_1 \times A_2 \times A_3 \times \dots \times A_k \times A_{k+1}$  has the same cardinality as the set  $(A_1 \times A_2 \times A_3 \times \dots \times A_k) \times A_{k+1}$ , so it too is countably infinite. ■

**Theorem 14.6** If  $A$  and  $B$  are both countably infinite, then their union  $A \cup B$  is countably infinite.

*Proof.* Suppose  $A$  and  $B$  are both countably infinite. By Theorem 14.3, we know we can write  $A$  and  $B$  in list form as

$$\begin{aligned} A &= \{a_1, a_2, a_3, a_4, \dots\}, \\ B &= \{b_1, b_2, b_3, b_4, \dots\}. \end{aligned}$$

We can “shuffle”  $A$  and  $B$  into one infinite list for  $A \cup B$  as follows.

$$A \cup B = \{a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4, \dots\}.$$

(We agree not to list an element twice if it belongs to both  $A$  and  $B$ .) Thus  $A \cup B$  is countably infinite by Theorem 14.3. ■

### Exercises for Section 14.2

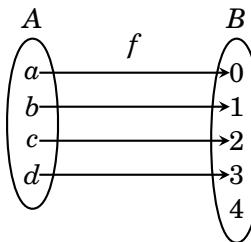
1. Prove that the set  $A = \{\ln(n) : n \in \mathbb{N}\} \subseteq \mathbb{R}$  is countably infinite.
2. Prove that the set  $A = \{(m, n) \in \mathbb{N} \times \mathbb{N} : m \leq n\}$  is countably infinite.
3. Prove that the set  $A = \{(5n, -3n) : n \in \mathbb{Z}\}$  is countably infinite.
4. Prove that the set of all irrational numbers is uncountable. (Suggestion: Consider proof by contradiction using Theorems 14.4 and 14.6.)
5. Prove or disprove: There exists a countably infinite subset of the set of irrational numbers.
6. Prove or disprove: There exists a bijective function  $f : \mathbb{Q} \rightarrow \mathbb{R}$ .
7. Prove or disprove: The set  $\mathbb{Q}^{100}$  is countably infinite.
8. Prove or disprove: The set  $\mathbb{Z} \times \mathbb{Q}$  is countably infinite.
9. Prove or disprove: The set  $\{0, 1\} \times \mathbb{N}$  is countably infinite.
10. Prove or disprove: The set  $A = \left\{ \frac{\sqrt{2}}{n} : n \in \mathbb{N} \right\}$  is countably infinite.
11. Describe a partition of  $\mathbb{N}$  that divides  $\mathbb{N}$  into eight countably infinite subsets.
12. Describe a partition of  $\mathbb{N}$  that divides  $\mathbb{N}$  into  $\aleph_0$  countably infinite subsets.
13. Prove or disprove: If  $A = \{X \subseteq \mathbb{N} : X \text{ is finite}\}$ , then  $|A| = \aleph_0$ .
14. Suppose  $A = \{(m, n) \in \mathbb{N} \times \mathbb{R} : n = \pi m\}$ . Is it true that  $|\mathbb{N}| = |A|$ ?
15. Theorem 14.5 implies that  $\mathbb{N} \times \mathbb{N}$  is countably infinite. Construct an alternate proof of this fact by showing that the function  $\varphi : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$  defined as  $\varphi(m, n) = 2^{n-1}(2m-1)$  is bijective.

### 14.3 Comparing Cardinalities

At this point we know that there are at least two different kinds of infinity. On one hand, there are countably infinite sets such as  $\mathbb{N}$ , of cardinality  $\aleph_0$ . Then there is the uncountable set  $\mathbb{R}$ . Are there other kinds of infinity beyond these two kinds? The answer is “yes,” but to see why we first need to introduce some new definitions and theorems.

Our first task will be to formulate a definition of what we mean by  $|A| < |B|$ . Of course if  $A$  and  $B$  are finite we know exactly what this means:  $|A| < |B|$  means that when the elements of  $A$  and  $B$  are counted,  $A$  is found to have fewer elements than  $B$ . But this process breaks down if  $A$  or  $B$  is infinite, for then the elements can’t be counted.

The language of functions helps us overcome this difficulty. Notice that for finite sets  $A$  and  $B$  it is intuitively clear that  $|A| < |B|$  if and only if there exists an injective function  $f : A \rightarrow B$  but there is no bijective function  $f : A \rightarrow B$ . The following diagram illustrates this:



We will use this idea to define what is meant by  $|A| < |B|$  and  $|A| \leq |B|$ . For emphasis, the following definition also restates what is meant by  $|A| = |B|$ .

**Definition 14.4** Suppose  $A$  and  $B$  are sets.

1.  $|A| = |B|$  means there is a bijection  $A \rightarrow B$ .
2.  $|A| < |B|$  means there is an injection  $A \rightarrow B$ , but no bijection  $A \rightarrow B$ .
3.  $|A| \leq |B|$  means there is an injection  $A \rightarrow B$ .

For example, consider  $\mathbb{N}$  and  $\mathbb{R}$ . The function  $f : \mathbb{N} \rightarrow \mathbb{R}$  defined as  $f(n) = n$  is clearly injective, but it is not surjective because given the element  $\frac{1}{2} \in \mathbb{R}$ , we have  $f(n) \neq \frac{1}{2}$  for every  $n \in \mathbb{N}$ . In fact, Theorem 14.2 of Section 14.1 asserts that there is no surjection  $\mathbb{N} \rightarrow \mathbb{R}$ , and hence no bijections either. Definition 14.4 yields

$$|\mathbb{N}| < |\mathbb{R}|. \quad (14.1)$$

Said differently,  $\aleph_0 < |\mathbb{R}|$ .

Is there a set  $X$  for which  $|\mathbb{R}| < |X|$ ? The answer is “yes.” The next theorem implies  $|\mathbb{R}| < |\mathcal{P}(\mathbb{R})|$ . (Recall that  $\mathcal{P}(A)$  denotes the power set of  $A$ .)

**Theorem 14.7** If  $A$  is any set, then  $|A| < |\mathcal{P}(A)|$ .

*Proof.* Before beginning the proof, we remark that this statement is obvious if  $A$  is finite, for then  $|A| < 2^{|A|} = |\mathcal{P}(A)|$ . But our proof must apply to *all* sets  $A$ , both finite and infinite, so it must use Definition 14.4.

We prove the theorem with direct proof. Let  $A$  be an arbitrary set. According to Definition 14.4, to prove  $|A| < |\mathcal{P}(A)|$  we must show that there is an injection  $f : A \rightarrow \mathcal{P}(A)$ , but no bijection  $f : A \rightarrow \mathcal{P}(A)$ .

To see that there is an injection  $f : A \rightarrow \mathcal{P}(A)$ , define  $f$  by the rule  $f(x) = \{x\}$ . In words,  $f$  sends any element  $x$  of  $A$  to the one-element set  $\{x\} \in \mathcal{P}(A)$ . Then  $f : A \rightarrow \mathcal{P}(A)$  is injective, as follows. Suppose  $f(x) = f(y)$ . Then  $\{x\} = \{y\}$ . Now, the only way that  $\{x\}$  and  $\{y\}$  can be equal is if  $x = y$ , so it follows that  $x = y$ . Thus  $f$  is injective.

Next we need to show that there exists no bijection  $A \rightarrow \mathcal{P}(A)$ . We will verify this by proving that there is no *surjection*  $A \rightarrow \mathcal{P}(A)$ . Take an

arbitrary function  $f : A \rightarrow \mathcal{P}(A)$ . To show  $f$  is not surjective we will produce a set  $B \in \mathcal{P}(A)$  for which  $f(a) \neq B$  for all  $a \in A$ . Notice that for any element  $x \in A$ , we have  $f(x) \in \mathcal{P}(A)$ , that is,  $f(x) \subseteq A$ . Thus  $f$  is a function sending elements of  $A$  to subsets of  $A$ . It follows that for any  $x \in A$ , either  $x \in f(x)$  or  $x \notin f(x)$ . Using this idea, define the following element  $B \in \mathcal{P}(A)$ :

$$B = \{x \in A : x \notin f(x)\} \subseteq A.$$

Take an arbitrary  $a \in A$ . The following two cases show that  $f(a) \neq B$ .

**Case 1.** If  $a \notin f(a)$ , then the definition of  $B$  implies  $a \in B$ . Consequently,  $f(a) = B$  is impossible, for it would mean  $a \notin B$  and  $a \in B$ .

**Case 2.** If  $a \in f(a)$ , then the definition of  $B$  implies  $a \notin B$ . Consequently,  $f(a) = B$  is impossible, for it would mean  $a \in B$  and  $a \notin B$ .

So  $f(a) \neq B$  for all  $a \in A$ , and hence  $f$  is not surjective. As this holds for any function  $f : A \rightarrow \mathcal{P}(A)$ , there are no surjective functions  $f : A \rightarrow \mathcal{P}(A)$ . Consequently there are no such bijections either.

In conclusion, we have seen that there exists an injection  $A \rightarrow \mathcal{P}(A)$  but no bijection  $A \rightarrow \mathcal{P}(A)$ , so Definition 14.4 implies that  $|A| < |\mathcal{P}(A)|$ . ■

Beginning with the set  $A = \mathbb{N}$  and applying Theorem 14.7 over and over again, we get the following chain of infinite cardinalities.

$$\aleph_0 = |\mathbb{N}| < |\mathcal{P}(\mathbb{N})| < |\mathcal{P}(\mathcal{P}(\mathbb{N}))| < |\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))| < \dots \quad (14.2)$$

Thus there is an infinite sequence of different types of infinity, starting with  $\aleph_0$  and becoming ever larger. The set  $\mathbb{N}$  is countable, and all the sets  $\mathcal{P}(\mathbb{N})$ ,  $\mathcal{P}(\mathcal{P}(\mathbb{N}))$ , etc., are uncountable.

In the next section we will prove that  $|\mathcal{P}(\mathbb{N})| = |\mathbb{R}|$ . Thus  $|\mathbb{N}|$  and  $|\mathbb{R}|$  are the first two entries in the chain (14.2) above. They are just two relatively tame infinities in a long list of other wild and exotic infinities.

Unless you plan on studying advanced set theory or the foundations of mathematics, you are unlikely to ever encounter any types of infinity beyond  $\aleph_0$  and  $|\mathbb{R}|$ . Still you will in future mathematics courses need to distinguish between countably infinite and uncountable sets, so we close with two final theorems that can help you do this.

**Theorem 14.8** An infinite subset of a countably infinite set is countably infinite.

*Proof.* Suppose  $A$  is an infinite subset of the countably infinite set  $B$ . As  $B$  is countably infinite, its elements can be written in a list  $b_1, b_2, b_3, b_4, \dots$

Then we can also write  $A$ 's elements in list form by proceeding through the elements of  $B$ , in order, and selecting those that belong to  $A$ . Thus  $A$  can be written in list form, and since  $A$  is infinite, its list will be infinite. Consequently  $A$  is countably infinite. ■

**Theorem 14.9** If  $U \subseteq A$ , and  $U$  is uncountable, then  $A$  is uncountable.

*Proof.* For the sake of contradiction say that  $U \subseteq A$ , and  $U$  is uncountable but  $A$  is not uncountable. Then since  $U \subseteq A$  and  $U$  is infinite, then  $A$  must be infinite too. Since  $A$  is infinite, and not uncountable, it must be countably infinite. Then  $U$  is an infinite subset of a countably infinite set  $A$ , so  $U$  is countably infinite by Theorem 14.8. Thus  $U$  is both uncountable and countably infinite, a contradiction. ■

Theorems 14.8 and 14.9 can be useful when we need to decide whether a set is countably infinite or uncountable. They sometimes allow us to decide its cardinality by comparing it to a set whose cardinality is known.

For example, suppose we want to decide whether or not the set  $A = \mathbb{R}^2$  is uncountable. Since the  $x$ -axis  $U = \{(x, 0) : x \in \mathbb{R}\} \subseteq \mathbb{R}^2$  has the same cardinality as  $\mathbb{R}$ , it is uncountable. Theorem 14.9 implies that  $\mathbb{R}^2$  is uncountable. Other examples can be found in the exercises.

---

### Exercises for Section 14.3

1. Suppose  $B$  is an uncountable set and  $A$  is a set. Given that there is a surjective function  $f : A \rightarrow B$ , what can be said about the cardinality of  $A$ ?
  2. Prove that the set  $\mathbb{C}$  of complex numbers is uncountable.
  3. Prove or disprove: If  $A$  is uncountable, then  $|A| = |\mathbb{R}|$ .
  4. Prove or disprove: If  $A \subseteq B \subseteq C$  and  $A$  and  $C$  are countably infinite, then  $B$  is countably infinite.
  5. Prove or disprove: The set  $\{0, 1\} \times \mathbb{R}$  is uncountable.
  6. Prove or disprove: Every infinite set is a subset of a countably infinite set.
  7. Prove or disprove: If  $A \subseteq B$  and  $A$  is countably infinite and  $B$  is uncountable, then  $B - A$  is uncountable.
  8. Prove or disprove: The set  $\{(a_1, a_2, a_3, \dots) : a_i \in \mathbb{Z}\}$  of infinite sequences of integers is countably infinite.
  9. Prove that if  $A$  and  $B$  are finite sets with  $|A| = |B|$ , then any injection  $f : A \rightarrow B$  is also a surjection. Show this is not necessarily true if  $A$  and  $B$  are not finite.
  10. Prove that if  $A$  and  $B$  are finite sets with  $|A| = |B|$ , then any surjection  $f : A \rightarrow B$  is also an injection. Show this is not necessarily true if  $A$  and  $B$  are not finite.
-

#### 14.4 The Cantor-Bernstein-Schröder Theorem

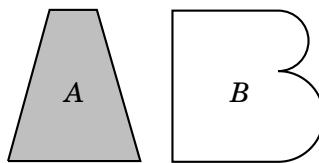
An often used property of numbers is that if  $a \leq b$  and  $b \leq a$ , then  $a = b$ . It is reasonable to ask if the same property applies to cardinality. If  $|A| \leq |B|$  and  $|B| \leq |A|$ , is it true that  $|A| = |B|$ ? This is in fact true, and this section's goal is to prove it. This will yield an alternate (and highly effective) method of proving that two sets have the same cardinality.

Recall (Definition 14.4) that  $|A| \leq |B|$  means that there is an injection  $f : A \rightarrow B$ . Likewise,  $|B| \leq |A|$  implies that there is an injection  $g : B \rightarrow A$ .

Our aim is to show that if  $|A| \leq |B|$  and  $|B| \leq |A|$ , then  $|A| = |B|$ . In other words, we aim to show that if there are injections  $f : A \rightarrow B$  and  $g : B \rightarrow A$ , then there is a bijection  $h : A \rightarrow B$ . The proof of this fact, though not particularly difficult, is not entirely trivial, either. The fact that  $f$  and  $g$  guarantee that such an  $h$  exists is called the **Cantor-Bernstein-Schröder theorem**. This theorem is very useful for proving two sets  $A$  and  $B$  have the same cardinality: it says that instead of finding a bijection  $A \rightarrow B$ , it suffices to find injections  $A \rightarrow B$  and  $B \rightarrow A$ . This is useful because injections are often easier to find than bijections.

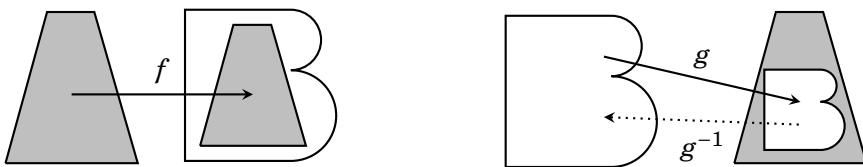
We will prove the Cantor-Bernstein-Schröder theorem, but before doing so let's work through an informal visual argument that will guide us through (and illustrate) the proof.

Suppose there are injections  $f : A \rightarrow B$  and  $g : B \rightarrow A$ . We want to use them to produce a bijection  $h : A \rightarrow B$ . Sets  $A$  and  $B$  are sketched below. For clarity, each has the shape of the letter that denotes it, and to help distinguish them the set  $A$  is shaded.



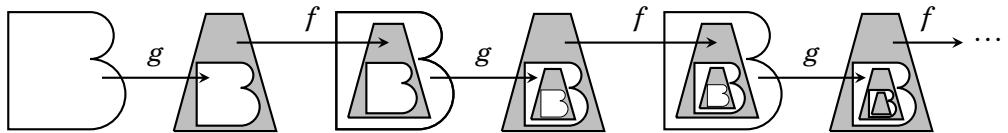
**Figure 14.3.** The sets  $A$  and  $B$

The injections  $f : A \rightarrow B$  and  $g : B \rightarrow A$  are illustrated in Figure 14.4. Think of  $f$  as putting a “copy”  $f(A) = \{f(x) : x \in A\}$  of  $A$  into  $B$ , as illustrated. This copy, the range of  $f$ , does not fill up all of  $B$  (unless  $f$  happens to be surjective). Likewise,  $g$  puts a “copy”  $g(B)$  of  $B$  into  $A$ . Because they are not necessarily bijective, neither  $f$  nor  $g$  is guaranteed to have an inverse. But the map  $g : B \rightarrow g(B)$  from  $B$  to  $g(B) = \{g(x) : x \in B\}$  is bijective, so there is an inverse  $g^{-1} : g(B) \rightarrow B$ . (We will need this inverse soon.)



**Figure 14.4.** The injections  $f : A \rightarrow B$  and  $g : B \rightarrow A$

Consider the chain of injections illustrated in Figure 14.5. On the left,  $g$  puts a copy of  $B$  into  $A$ . Then  $f$  puts a copy of  $A$  (containing the copy of  $B$ ) into  $B$ . Next,  $g$  puts a copy of this  $B$ -containing- $A$ -containing- $B$  into  $A$ , and so on, always alternating  $g$  and  $f$ .



**Figure 14.5.** An infinite chain of injections

Let's analyze our infinite sequence  $B \rightarrow A \rightarrow B \rightarrow A \rightarrow B \rightarrow A \rightarrow \dots$

The first time  $A$  occurs in this sequence, it has a shaded region  $A - g(B)$ . In the second occurrence of  $A$ , the shaded region is  $(A - g(B)) \cup (g \circ f)(A - g(B))$ . In the third occurrence of  $A$ , the shaded region is

$$(A - g(B)) \cup (g \circ f)(A - g(B)) \cup (g \circ f \circ g \circ f)(A - g(B)).$$

To tame the notation, let's say  $(g \circ f)^2 = (g \circ f) \circ (g \circ f)$ , and  $(g \circ f)^3 = (g \circ f) \circ (g \circ f) \circ (g \circ f)$ , and so on. Let's also agree that  $(g \circ f)^0 = \iota_A$ , that is, it is the identity function on  $A$ . Then the shaded region of the  $n$ th occurrence of  $A$  in the sequence is

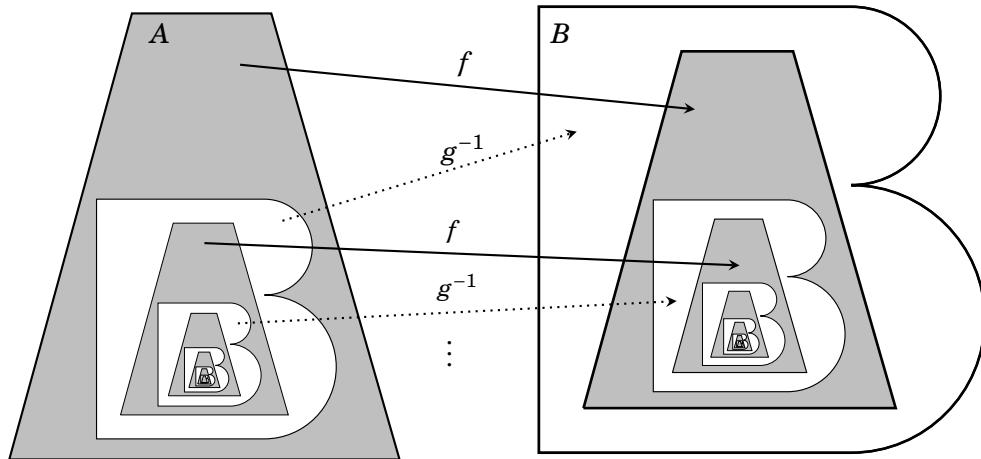
$$\bigcup_{k=0}^{n-1} (g \circ f)^k (A - g(B)).$$

This process divides  $A$  into gray and white regions: the gray region is

$$G = \bigcup_{k=0}^{\infty} (g \circ f)^k (A - g(B)),$$

and the white region is  $A - G$ . (See Figure 14.6.)

Figure 14.6 suggests our desired bijection  $h : A \rightarrow B$ . The injection  $f$  sends the gray areas on the left bijectively to the gray areas on the right. The injection  $g^{-1} : g(B) \rightarrow B$  sends the white areas on the left bijectively to the white areas on the right. We can thus define  $h : A \rightarrow B$  so that  $h(x) = f(x)$  if  $x$  is a gray point, and  $h(x) = g^{-1}(x)$  if  $x$  is a white point.



**Figure 14.6.** The bijection  $h : A \rightarrow B$

This informal argument suggests that given injections  $f : A \rightarrow B$  and  $g : B \rightarrow A$ , there is a bijection  $h : A \rightarrow B$ . But it is not a proof. We now present this as a theorem and tighten up our reasoning in a careful proof, with the above diagrams and ideas as a guide.

### Theorem 14.10 (The Cantor-Bernstein-Schröder Theorem)

If  $|A| \leq |B|$  and  $|B| \leq |A|$ , then  $|A| = |B|$ . In other words, if there are injections  $f : A \rightarrow B$  and  $g : B \rightarrow A$ , then there is a bijection  $h : A \rightarrow B$ .

*Proof.* (Direct) Suppose there are injections  $f : A \rightarrow B$  and  $g : B \rightarrow A$ . Then, in particular,  $g : B \rightarrow g(B)$  is a bijection from  $B$  onto the range of  $g$ , so it has an inverse  $g^{-1} : g(B) \rightarrow B$ . (Note that  $g : B \rightarrow A$  itself has no inverse  $g^{-1} : A \rightarrow B$  unless  $g$  is surjective.) Consider the subset

$$G = \bigcup_{k=0}^{\infty} (g \circ f)^k (A - g(B)) \subseteq A.$$

Let  $W = A - G$ , so  $A = G \cup W$  is partitioned into two sets  $G$  (think gray) and  $W$  (think white). Define a function  $h : A \rightarrow B$  as

$$h(x) = \begin{cases} f(x) & \text{if } x \in G \\ g^{-1}(x) & \text{if } x \in W. \end{cases}$$

Notice that this makes sense: if  $x \in W$ , then  $x \notin G$ , so  $x \notin A - g(B) \subseteq G$ , hence  $x \in g(B)$ , so  $g^{-1}(x)$  is defined.

To finish the proof, we must show that  $h$  is both injective and surjective.

For injective, we assume  $h(x) = h(y)$ , and deduce  $x = y$ . There are three cases to consider. First, if  $x$  and  $y$  are both in  $G$ , then  $h(x) = h(y)$  means  $f(x) = f(y)$ , so  $x = y$  because  $f$  is injective. Second, if  $x$  and  $y$  are both in  $W$ , then  $h(x) = h(y)$  means  $g^{-1}(x) = g^{-1}(y)$ , and applying  $g$  to both sides gives  $x = y$ . In the third case, one of  $x$  and  $y$  is in  $G$  and the other is in  $W$ . Say  $x \in G$  and  $y \in W$ . The definition of  $G$  gives  $x = (g \circ f)^k(z)$  for some  $k \geq 0$  and  $z \in A - g(B)$ . Note  $h(x) = h(y)$  now implies  $f(x) = g^{-1}(y)$ , that is,  $f((g \circ f)^k(z)) = g^{-1}(y)$ . Applying  $g$  to both sides gives  $(g \circ f)^{k+1}(z) = y$ , which means  $y \in G$ . But this is impossible, as  $y \in W$ . Thus this third case cannot happen. But in the first two cases  $h(x) = h(y)$  implies  $x = y$ , so  $h$  is injective.

To see that  $h$  is surjective, take any  $b \in B$ . We will find an  $x \in A$  with  $h(x) = b$ . Note that  $g(b) \in A$ , so either  $g(b) \in W$  or  $g(b) \in G$ . In the first case,  $h(g(b)) = g^{-1}(g(b)) = b$ , so we have an  $x = g(b) \in A$  for which  $h(x) = b$ . In the second case,  $g(b) \in G$ . The definition of  $G$  shows

$$g(b) = (g \circ f)^k(z)$$

for some  $z \in A - g(B)$  and  $k \geq 0$ . In fact we have  $k > 0$ , because  $k = 0$  would give  $g(b) = (g \circ f)^0(z) = z \in A - g(B)$ , but clearly  $g(b) \notin A - g(B)$ . Thus

$$\begin{aligned} g(b) &= (g \circ f) \circ (g \circ f)^{k-1}(z) \\ &= g\left(f\left((g \circ f)^{k-1}(z)\right)\right). \end{aligned}$$

Because  $g$  is injective, this implies

$$b = f\left((g \circ f)^{k-1}(z)\right).$$

Let  $x = (g \circ f)^{k-1}(z)$ , so  $x \in G$  by definition of  $G$ . Observe that  $h(x) = f(x) = f\left((g \circ f)^{k-1}(z)\right) = b$ . We have now seen that for any  $b \in B$ , there is an  $x \in A$  for which  $h(x) = b$ . Thus  $h$  is surjective.

Since  $h : A \rightarrow B$  is both injective and surjective, it is also bijective. ■

Here are some examples illustrating how the Cantor-Bernstein-Schröder theorem can be used. This includes a proof that  $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$ .

**Example 14.6** The intervals  $[0, 1)$  and  $(0, 1)$  in  $\mathbb{R}$  have equal cardinalities.

Surely this fact is plausible, for the two intervals are identical except for the endpoint 0. Yet concocting a bijection  $[0, 1) \rightarrow (0, 1)$  is tricky. (Though not particularly difficult: see the solution of Exercise 11 of Section 14.1.)

For a simpler approach, note that  $f(x) = \frac{1}{4} + \frac{1}{2}x$  is an injection  $[0, 1) \rightarrow (0, 1)$ . Also,  $g(x) = x$  is an injection  $(0, 1) \rightarrow [0, 1)$ . The Cantor-Bernstein-Schröder theorem guarantees a bijection  $h : [0, 1) \rightarrow (0, 1)$ , so  $|[0, 1)| = |(0, 1)|$ .

**Theorem 14.11** The sets  $\mathbb{R}$  and  $\mathcal{P}(\mathbb{N})$  have the same cardinality.

*Proof.* Example 14.4 shows that  $|\mathbb{R}| = |(0, 1)|$ , and Example 14.6 shows  $|(0, 1)| = |[0, 1)|$ . Thus  $|\mathbb{R}| = |[0, 1)|$ , so to prove the theorem we just need to show that  $|[0, 1)| = |\mathcal{P}(\mathbb{N})|$ . By the Cantor-Bernstein-Schröder theorem, it suffices to find injections  $f : [0, 1) \rightarrow \mathcal{P}(\mathbb{N})$  and  $g : \mathcal{P}(\mathbb{N}) \rightarrow [0, 1)$ .

To define  $f : [0, 1) \rightarrow \mathcal{P}(\mathbb{N})$ , we use the fact that any number in  $[0, 1)$  has a unique decimal representation  $0.b_1b_2b_3b_4\dots$ , where each  $b_i$  one of the digits  $0, 1, 2, \dots, 9$ , and there is not a repeating sequence of 9's at the end. (Recall that, e.g.,  $0.35999\bar{9} = 0.36\bar{0}$ , etc.) Define  $f : [0, 1) \rightarrow \mathcal{P}(\mathbb{N})$  as

$$f(0.b_1b_2b_3b_4\dots) = \{10b_1, 10^2b_2, 10^3b_3, \dots\}.$$

For example,  $f(0.1212\bar{12}) = \{10, 200, 1000, 20000, 100000, \dots\}$ , and  $f(0.05) = \{0, 500\}$ . Also  $f(0.5) = f(0.\bar{5}) = \{0, 50\}$ . To see that  $f$  is injective, take two unequal numbers  $0.b_1b_2b_3b_4\dots$  and  $0.d_1d_2d_3d_4\dots$  in  $[0, 1)$ . Then  $b_i \neq d_i$  for some index  $i$ . Now, either  $b_i \neq 0$  or  $d_i \neq 0$ ; without loss of generality say  $b_i \neq 0$ . Then  $b_i 10^i \in f(0.b_1b_2b_3b_4\dots)$  but  $b_i 10^i \notin f(0.d_1d_2d_3d_4\dots)$ , so  $f(0.b_1b_2b_3b_4\dots) \neq f(0.d_1d_2d_3d_4\dots)$ . Consequently  $f$  is injective.

Next, define  $g : \mathcal{P}(\mathbb{N}) \rightarrow [0, 1)$ , where  $g(X) = 0.b_1b_2b_3b_4\dots$  is the number for which  $b_i = 1$  if  $i \in X$  and  $b_i = 0$  if  $i \notin X$ . For example,  $g(\{1, 3\}) = 0.10100\bar{0}$ , and  $g(\{2, 4, 6, 8, \dots\}) = 0.010101\bar{01}$ . Also  $g(\emptyset) = 0$  and  $g(\mathbb{N}) = 0.111\bar{1}$ . To see that  $g$  is injective, suppose  $X \neq Y$ . Then there is at least one integer  $i$  that belongs to one of  $X$  or  $Y$ , but not the other. Consequently  $g(X) \neq g(Y)$  because they differ in the  $i$ th decimal place. This shows  $g$  is injective.

From the injections  $f : [0, 1) \rightarrow \mathcal{P}(\mathbb{N})$  and  $g : \mathcal{P}(\mathbb{N}) \rightarrow [0, 1)$ , the Cantor-Bernstein-Schröder theorem guarantees a bijection  $h : [0, 1) \rightarrow \mathcal{P}(\mathbb{N})$ . Hence  $|[0, 1)| = |\mathcal{P}(\mathbb{N})|$ . As  $|\mathbb{R}| = |[0, 1)|$ , we conclude  $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$ . ■

We know that  $|\mathbb{R}| \neq |\mathbb{N}|$ . But we just proved  $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$ . This suggests that the cardinality of  $\mathbb{R}$  is not “too far” from  $|\mathbb{N}| = \aleph_0$ . We close with a few informal remarks on this mysterious relationship between  $\aleph_0$  and  $|\mathbb{R}|$ .

We established earlier in this chapter that  $\aleph_0 < |\mathbb{R}|$ . For nearly a century after Cantor formulated his theories on infinite sets, mathematicians struggled with the question of whether or not there exists a set  $A$  for which

$$\aleph_0 < |A| < |\mathbb{R}|.$$

It was commonly suspected that no such set exists, but no one was able to prove or disprove this. The assertion that no such  $A$  exists came to be called the **continuum hypothesis**.

Theorem 14.11 states that  $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$ . Placing this in the context of the chain (14.2) on page 282, we have the following relationships.

$$\begin{array}{ccc} \aleph_0 & & |\mathbb{R}| \\ \parallel & & \parallel \\ |\mathbb{N}| & < & |\mathcal{P}(\mathbb{N})| & < & |\mathcal{P}(\mathcal{P}(\mathbb{N}))| & < & |\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))| & < & \dots \end{array}$$

From this, we can see that the continuum hypothesis asserts that no set has a cardinality between that of  $\mathbb{N}$  and its power set.

Though this may seem intuitively plausible, it eluded proof since Cantor first posed it in the 1880s. In fact, the real state of affairs seems almost paradoxical. In 1931, the logician Kurt Gödel proved that for any sufficiently strong and consistent axiomatic system, there exist statements which can neither be proved nor disproved within the system.

Later he proved that the negation of the continuum hypothesis cannot be proved within the standard axioms of set theory (i.e., the Zermelo-Fraenkel axioms, mentioned in Section 1.10). This meant that either the continuum hypothesis is false and cannot be proven false, or it is true.

In 1964, Paul Cohen discovered another startling truth: Given the laws of logic and the axioms of set theory, no proof can deduce the continuum hypothesis. In essence he proved that the continuum hypothesis cannot be *proved*.

Taken together, Gödel and Cohens' results mean that the standard axioms of mathematics cannot “decide” whether the continuum hypothesis is true or false, and that no logical conflict can arise from either asserting or denying the continuum hypothesis. We are free to either accept it as true or accept it as false, and the two choices lead to different—but equally consistent—versions of set theory.

On the face of it, this seems to undermine the foundation of logic, and everything we have done in this book. The continuum hypothesis should be a *statement*—it should be either true or false. How could it be both?

Here is an analogy that may help make sense of this. Consider the number systems  $\mathbb{Z}_n$ . What if we asked whether  $[2] = [0]$  is true or false? Of course the answer depends on  $n$ . The expression  $[2] = [0]$  is true in  $\mathbb{Z}_2$  and false in  $\mathbb{Z}_3$ . Moreover, if we assert that  $[2] = [0]$  is true, we are logically forced to the conclusion that this is taking place in the system  $\mathbb{Z}_2$ . If we assert that  $[2] = [0]$  is false, then we are dealing with some other  $\mathbb{Z}_n$ . The fact that  $[2] = [0]$  can be either true or false does not necessarily mean that there is some inherent inconsistency within the individual number systems  $\mathbb{Z}_n$ . The equation  $[2] = [0]$  is a true statement in the “universe” of  $\mathbb{Z}_2$  and a false statement in the universe of (say)  $\mathbb{Z}_3$ .

It is the same with the continuum hypothesis. Saying it’s true leads to one system of set theory. Saying it’s false leads to some other system of set theory. Gödel and Cohens’ discoveries mean that these two types of set theory, although different, are equally consistent and valid mathematical universes.

So what should you believe? Fortunately, it does not make much difference, because most important mathematical results do not hinge on the continuum hypothesis. (They are true in both universes.) Unless you undertake a deep study of the foundations of mathematics, you will be fine accepting the continuum hypothesis as true. Most mathematicians are agnostics on this issue, but they tend to prefer the version of set theory in which the continuum hypothesis holds.

The situation with the continuum hypothesis is a testament to the immense complexity of mathematics. It is a reminder of the importance of rigor and careful, systematic methods of reasoning that **begin** with the ideas introduced in this book.

### Exercises for Section 14.4

1. Show that if  $A \subseteq B$  and there is an injection  $g : B \rightarrow A$ , then  $|A| = |B|$ .
2. Show that  $|\mathbb{R}^2| = |\mathbb{R}|$ . Suggestion: Begin by showing  $|(0, 1) \times (0, 1)| = |(0, 1)|$ .
3. Let  $\mathcal{F}$  be the set of all functions  $\mathbb{N} \rightarrow \{0, 1\}$ . Show that  $|\mathbb{R}| = |\mathcal{F}|$ .
4. Let  $\mathcal{F}$  be the set of all functions  $\mathbb{R} \rightarrow \{0, 1\}$ . Show that  $|\mathbb{R}| < |\mathcal{F}|$ .
5. Consider the subset  $B = \{(x, y) : x^2 + y^2 \leq 1\} \subseteq \mathbb{R}^2$ . Show that  $|B| = |\mathbb{R}^2|$ .
6. Show that  $|\mathcal{P}(\mathbb{N} \times \mathbb{N})| = |\mathcal{P}(\mathbb{N})|$ .
7. Prove or disprove: If there is a injection  $f : A \rightarrow B$  and a surjection  $g : A \rightarrow B$ , then there is a bijection  $h : A \rightarrow B$ .

---

## **Conclusion**

---

If you have internalized the ideas in this book, then you have gained a set of rhetorical tools for deciphering and communicating mathematics. These tools are indispensable at the advanced levels. But of course it takes more than mere tools to build something. Creativity, inspiration, skill, talent, intuition, passion, planning and persistence are also vitally important. It is safe to say that if you have come this far, then you probably possess a sufficient measure of these traits.

The quest to understand mathematics has no end, but you are well equipped for the journey. It is my hope that the things you have learned from this book will lead you to a higher plane of understanding, creativity and expression.

Good luck and best wishes.

R.H.

---

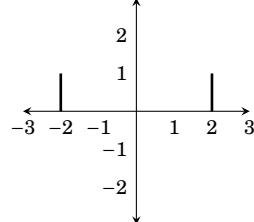
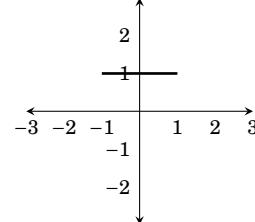
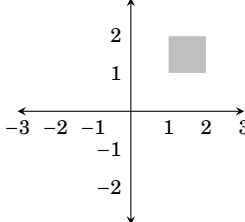
# Solutions

---

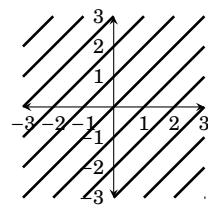
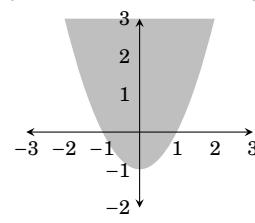
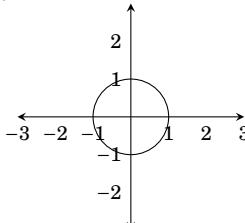
## Chapter 1 Exercises

### Section 1.1

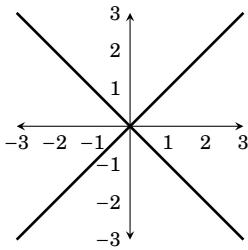
1.  $\{5x - 1 : x \in \mathbb{Z}\} = \{\dots, -11, -6, -1, 4, 9, 14, 19, 24, 29, \dots\}$
3.  $\{x \in \mathbb{Z} : -2 \leq x < 7\} = \{-2, -1, 0, 1, 2, 3, 4, 5, 6\}$
5.  $\{x \in \mathbb{R} : x^2 = 3\} = \{-\sqrt{3}, \sqrt{3}\}$
7.  $\{x \in \mathbb{R} : x^2 + 5x = -6\} = \{-2, -3\}$
9.  $\{x \in \mathbb{R} : \sin \pi x = 0\} = \{\dots, -2, -1, 0, 1, 2, 3, 4, \dots\} = \mathbb{Z}$
11.  $\{x \in \mathbb{Z} : |x| < 5\} = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$
13.  $\{x \in \mathbb{Z} : |6x| < 5\} = \{0\}$
15.  $\{5a + 2b : a, b \in \mathbb{Z}\} = \{\dots, -2, -1, 0, 1, 2, 3, \dots\} = \mathbb{Z}$
17.  $\{2, 4, 8, 16, 32, 64, \dots\} = \{2^x : x \in \mathbb{N}\}$
19.  $\{\dots, -6, -3, 0, 3, 6, 9, 12, 15, \dots\} = \{3x : x \in \mathbb{Z}\}$
21.  $\{0, 1, 4, 9, 16, 25, 36, \dots\} = \{x^2 : x \in \mathbb{Z}\}$
23.  $\{3, 4, 5, 6, 7, 8\} = \{x \in \mathbb{Z} : 3 \leq x \leq 8\} = \{x \in \mathbb{N} : 3 \leq x \leq 8\}$
25.  $\{\dots, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, \dots\} = \{2^n : n \in \mathbb{Z}\}$
27.  $\{\dots, -\pi, -\frac{\pi}{2}, 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi, \frac{5\pi}{2}, \dots\} = \left\{ \frac{k\pi}{2} : k \in \mathbb{Z} \right\}$
29.  $|\{(1), \{2, \{3, 4\}\}, \emptyset\}| = 3$     31.  $|\{\{(1), \{2, \{3, 4\}\}, \emptyset\}\}| = 1$     33.  $|\{x \in \mathbb{Z} : |x| < 10\}| = 19$
35.  $|\{x \in \mathbb{Z} : x^2 < 10\}| = 7$     37.  $|\{x \in \mathbb{N} : x^2 < 0\}| = 0$
39.  $\{(x, y) : x \in [1, 2], y \in [1, 2]\}$     41.  $\{(x, y) : x \in [-1, 1], y = 1\}$     43.  $\{(x, y) : |x| = 2, y \in [0, 1]\}$



45.  $\{(x, y) : x, y \in \mathbb{R}, x^2 + y^2 = 1\}$     47.  $\{(x, y) : x, y \in \mathbb{R}, y \geq x^2 - 1\}$     49.  $\{(x, x+y) : x \in \mathbb{R}, y \in \mathbb{Z}\}$



51.  $\{(x, y) \in \mathbb{R}^2 : (y - x)(y + x) = 0\}$



## Section 1.2

1. Suppose  $A = \{1, 2, 3, 4\}$  and  $B = \{a, c\}$ .

- (a)  $A \times B = \{(1, a), (1, c), (2, a), (2, c), (3, a), (3, c), (4, a), (4, c)\}$
- (b)  $B \times A = \{(a, 1), (a, 2), (a, 3), (a, 4), (c, 1), (c, 2), (c, 3), (c, 4)\}$
- (c)  $A \times A = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (2, 4), (3, 1), (3, 2), (3, 3), (3, 4), (4, 1), (4, 2), (4, 3), (4, 4)\}$
- (d)  $B \times B = \{(a, a), (a, c), (c, a), (c, c)\}$
- (e)  $\emptyset \times B = \{(a, b) : a \in \emptyset, b \in B\} = \emptyset$  (There are no ordered pairs  $(a, b)$  with  $a \in \emptyset$ .)
- (f)  $(A \times B) \times B =$   
 $\{(1, a), a, ((1, c), a), ((2, a), a), ((2, c), a), ((3, a), a), ((3, c), a), ((4, a), a), ((4, c), a), ((1, a), c), ((1, c), c), ((2, a), c), ((2, c), c), ((3, a), c), ((3, c), c), ((4, a), c), ((4, c), c)\}$
- (g)  $A \times (B \times B) =$   
 $\{(1, (a, a)), (1, (a, c)), (1, (c, a)), (1, (c, c)), (2, (a, a)), (2, (a, c)), (2, (c, a)), (2, (c, c)), (3, (a, a)), (3, (a, c)), (3, (c, a)), (3, (c, c)), (4, (a, a)), (4, (a, c)), (4, (c, a)), (4, (c, c))\}$
- (h)  $B^3 = \{(a, a, a), (a, a, c), (a, c, a), (a, c, c), (c, a, a), (c, a, c), (c, c, a), (c, c, c)\}$

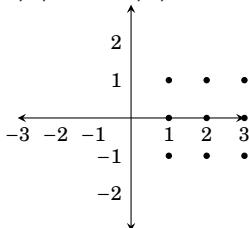
3.  $\{x \in \mathbb{R} : x^2 = 2\} \times \{a, c, e\} = \{(-\sqrt{2}, a), (\sqrt{2}, a), (-\sqrt{2}, c), (\sqrt{2}, c), (-\sqrt{2}, e), (\sqrt{2}, e)\}$

5.  $\{x \in \mathbb{R} : x^2 = 2\} \times \{x \in \mathbb{R} : |x| = 2\} = \{(-\sqrt{2}, -2), (\sqrt{2}, 2), (-\sqrt{2}, 2), (\sqrt{2}, -2)\}$

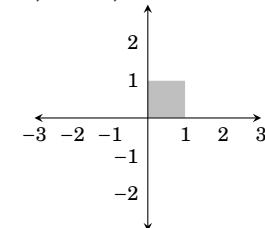
7.  $\{\emptyset\} \times \{0, \emptyset\} \times \{0, 1\} = \{(\emptyset, 0, 0), (\emptyset, 0, 1), (\emptyset, \emptyset, 0), (\emptyset, \emptyset, 1)\}$

Sketch the following Cartesian products on the  $x$ - $y$  plane.

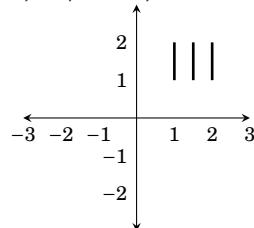
9.  $\{1, 2, 3\} \times \{-1, 0, 1\}$

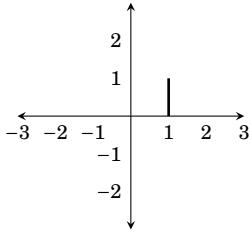
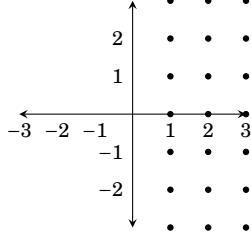
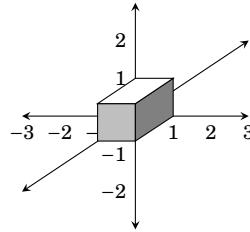


11.  $[0, 1] \times [0, 1]$



13.  $\{1, 1.5, 2\} \times [1, 2]$



**15.**  $\{1\} \times [0, 1]$ **17.**  $\mathbb{N} \times \mathbb{Z}$ **19.**  $[0, 1] \times [0, 1] \times [0, 1]$ 

### Section 1.3

**A.** List all the subsets of the following sets.

1. The subsets of  $\{1, 2, 3, 4\}$  are:  $\{\}, \{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}$ .
3. The subsets of  $\{\{\mathbb{R}\}\}$  are:  $\{\}$  and  $\{\{\mathbb{R}\}\}$ .
5. The subsets of  $\{\emptyset\}$  are  $\{\}$  and  $\{\emptyset\}$ .
7. The subsets of  $\{\mathbb{R}, \{\mathbb{Q}, \mathbb{N}\}\}$  are  $\{\}, \{\mathbb{R}\}, \{\{\mathbb{Q}, \mathbb{N}\}\}, \{\mathbb{R}, \{\mathbb{Q}, \mathbb{N}\}\}$ .

**B.** Write out the following sets by listing their elements between braces.

9.  $\{X : X \subseteq \{3, 2, a\} \text{ and } |X| = 2\} = \{\{3, 2\}, \{3, a\}, \{2, a\}\}$
11.  $\{X : X \subseteq \{3, 2, a\} \text{ and } |X| = 4\} = \{\} = \emptyset$

**C.** Decide if the following statements are true or false.

13.  $\mathbb{R}^3 \subseteq \mathbb{R}^3$  is **true** because any set is a subset of itself.
15.  $\{(x, y) : x - 1 = 0\} \subseteq \{(x, y) : x^2 - x = 0\}$ . This is true. (The even-numbered ones are both false. You have to explain why.)

### Section 1.4

**A.** Find the indicated sets.

1.  $\mathcal{P}(\{a, b\}, \{c\}) = \{\emptyset, \{\{a, b\}\}, \{\{c\}\}, \{\{a, b\}, \{c\}\}\}$
3.  $\mathcal{P}(\{\emptyset\}, 5) = \{\emptyset, \{\{\emptyset\}\}, \{5\}, \{\{\emptyset\}, 5\}\}$
5.  $\mathcal{P}(\mathcal{P}(\{2\})) = \{\emptyset, \{\emptyset\}, \{\{2\}\}, \{\emptyset, \{2\}\}\}$
7.  $\mathcal{P}(\{a, b\}) \times \mathcal{P}(\{0, 1\}) =$   

$$\{ (\emptyset, \emptyset), (\emptyset, \{0\}), (\emptyset, \{1\}), (\emptyset, \{0, 1\}),$$

$$(\{a\}, \emptyset), (\{a\}, \{0\}), (\{a\}, \{1\}), (\{a\}, \{0, 1\}),$$

$$(\{b\}, \emptyset), (\{b\}, \{0\}), (\{b\}, \{1\}), (\{b\}, \{0, 1\}),$$

$$(\{a, b\}, \emptyset), (\{a, b\}, \{0\}), (\{a, b\}, \{1\}), (\{a, b\}, \{0, 1\}) \}$$
9.  $\mathcal{P}(\{a, b\} \times \{0\}) = \{\emptyset, \{(a, 0)\}, \{(b, 0)\}, \{(a, 0), (b, 0)\}\}$
11.  $\{X \subseteq \mathcal{P}(\{1, 2, 3\}) : |X| \leq 1\} =$   

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

**B.** Suppose that  $|A| = m$  and  $|B| = n$ . Find the following cardinalities:

13.  $|\mathcal{P}(\mathcal{P}(\mathcal{P}(A)))| = 2^{(2^{2^m})}$

15.  $|\mathcal{P}(A \times B)| = 2^{mn}$

17.  $|\{X \in \mathcal{P}(A) : |X| \leq 1\}| = m + 1$

19.  $|\mathcal{P}(\mathcal{P}(\mathcal{P}(A \times \emptyset)))| = |\mathcal{P}(\mathcal{P}(\mathcal{P}(\emptyset)))| = 4$

### Section 1.5

1. Suppose  $A = \{4, 3, 6, 7, 1, 9\}$ ,  $B = \{5, 6, 8, 4\}$  and  $C = \{5, 8, 4\}$ . Find:

(a)  $A \cup B = \{1, 3, 4, 5, 6, 7, 8, 9\}$

(f)  $A \cap C = \{4\}$

(b)  $A \cap B = \{4, 6\}$

(g)  $B \cap C = \{5, 8, 4\}$

(c)  $A - B = \{3, 7, 1, 9\}$

(h)  $B \cup C = \{5, 6, 8, 4\}$

(d)  $A - C = \{3, 6, 7, 1, 9\}$

(i)  $C - B = \emptyset$

(e)  $B - A = \{5, 8\}$

3. Suppose  $A = \{0, 1\}$  and  $B = \{1, 2\}$ . Find:

(a)  $(A \times B) \cap (B \times B) = \{(1, 1), (1, 2)\}$

(b)  $(A \times B) \cup (B \times B) = \{(0, 1), (0, 2), (1, 1), (1, 2), (2, 1), (2, 2)\}$

(c)  $(A \times B) - (B \times B) = \{(0, 1), (0, 2)\}$

(f)  $\mathcal{P}(A) \cap \mathcal{P}(B) = \{\emptyset, \{1\}\}$

(d)  $(A \cap B) \times A = \{(1, 0), (1, 1)\}$

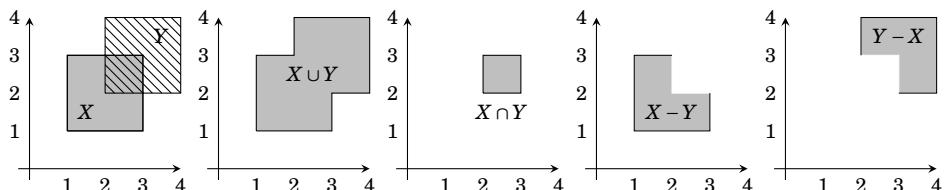
(g)  $\mathcal{P}(A) - \mathcal{P}(B) = \{\{0\}, \{0, 1\}\}$

(e)  $(A \times B) \cap B = \emptyset$

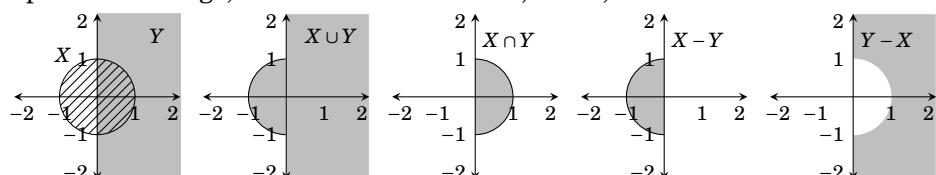
(h)  $\mathcal{P}(A \cap B) = \{\emptyset, \{1\}\}$

(i)  $\{\emptyset, \{(0, 1)\}, \{(0, 2)\}, \{(1, 1)\}, \{(1, 2)\}, \{(0, 1), (0, 2)\}, \{(0, 1), (1, 1)\}, \{(0, 1), (1, 2)\}, \{(0, 2), (1, 1)\}, \{(0, 2), (1, 2)\}, \{(1, 1), (1, 2)\}, \{(0, 2), (1, 1), (1, 2)\}, \{(0, 1), (1, 1), (1, 2)\}, \{(0, 1), (0, 2), (1, 1), (1, 2)\}, \{(0, 1), (0, 2), (1, 1)\}, \{(0, 1), (0, 2), (1, 1), (1, 2)\}\}$

5. Sketch the sets  $X = [1, 3] \times [1, 3]$  and  $Y = [2, 4] \times [2, 4]$  on the plane  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $X \cup Y$ ,  $X \cap Y$ ,  $X - Y$  and  $Y - X$ . (Hint:  $X$  and  $Y$  are Cartesian products of intervals. You may wish to review how you drew sets like  $[1, 3] \times [1, 3]$  in the Section 1.2.)



7. Sketch the sets  $X = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$  and  $Y = \{(x, y) \in \mathbb{R}^2 : x \geq 0\}$  on  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $X \cup Y$ ,  $X \cap Y$ ,  $X - Y$  and  $Y - X$ .



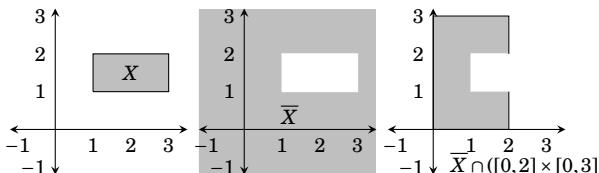
9. The first statement is true. (A picture should convince you; draw one if necessary.) The second statement is false: Notice for instance that  $(0.5, 0.5)$  is in the right-hand set, but not the left-hand set.

### Section 1.6

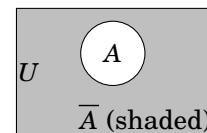
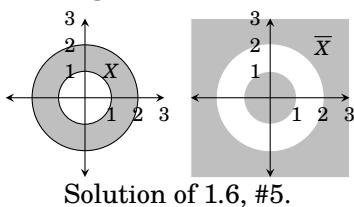
1. Suppose  $A = \{4, 3, 6, 7, 1, 9\}$  and  $B = \{5, 6, 8, 4\}$  have universal set  $U = \{n \in \mathbb{Z} : 0 \leq n \leq 10\}$ .

- |  |   |
|--|---|
| (a) $\overline{A} = \{0, 2, 5, 8, 10\}$                              | (f) $A - \overline{B} = \{4, 6\}$                                     |
| (b) $\overline{B} = \{0, 1, 2, 3, 7, 9, 10\}$                        | (g) $\overline{A} - \overline{B} = \{5, 8\}$                          |
| (c) $A \cap \overline{A} = \emptyset$                                | (h) $\overline{A} \cap B = \{5, 8\}$                                  |
| (d) $A \cup \overline{A} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} = U$ | (i) $\overline{\overline{A} \cap B} = \{0, 1, 2, 3, 4, 6, 7, 9, 10\}$ |
| (e) $A - \overline{A} = A$   |   |

3. Sketch the set  $X = [1, 3] \times [1, 2]$  on the plane  $\mathbb{R}^2$ . On separate drawings, shade in the sets  $\overline{X}$ , and  $\overline{X} \cap ([0, 2] \times [0, 3])$ .



5. Sketch the set  $X = \{(x, y) \in \mathbb{R}^2 : 1 \leq x^2 + y^2 \leq 4\}$  on the plane  $\mathbb{R}^2$ . On a separate drawing, shade in the set  $\overline{X}$ .



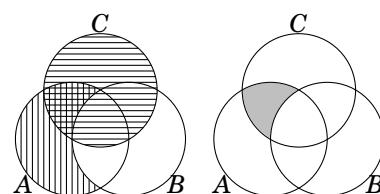
Solution of 1.7, #1.

### Section 1.7

1. Draw a Venn diagram for  $\overline{A}$  (solution above right).

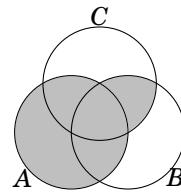
3. Draw a Venn diagram for  $(A - B) \cap C$ .

Scratch work is shown on the right. The set  $A - B$  is indicated with vertical shading. The set  $C$  is indicated with horizontal shading. The intersection of  $A - B$  and  $C$  is thus the overlapping region that is shaded with both vertical and horizontal lines. The final answer is drawn on the far right, where the set  $(A - B) \cap C$  is shaded in gray.



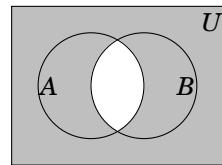
5. Draw Venn diagrams for  $A \cup (B \cap C)$  and  $(A \cup B) \cap (A \cup C)$ . Based on your drawings, do you think  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ ?

If you do the drawings carefully, you will find that your Venn diagrams are the same for both  $A \cup (B \cap C)$  and  $(A \cup B) \cap (A \cup C)$ . Each looks as illustrated on the right. Based on this, we are inclined to say that the equation  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$  holds for all sets  $A$ ,  $B$  and  $C$ .

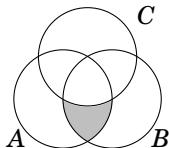


7. Suppose sets  $A$  and  $B$  are in a universal set  $U$ . Draw Venn diagrams for  $\overline{A \cap B}$  and  $\overline{A} \cup \overline{B}$ . Based on your drawings, do you think it's true that  $\overline{A \cap B} = \overline{A} \cup \overline{B}$ ?

The diagrams for  $\overline{A \cap B}$  and  $\overline{A} \cup \overline{B}$  look exactly alike. In either case the diagram is the shaded region illustrated on the right. Thus we would expect that the equation  $\overline{A \cap B} = \overline{A} \cup \overline{B}$  is true for any sets  $A$  and  $B$ .



9. Venn diagram for  $(A \cap B) - C$ :  
 11. The simplest answer is  $(B \cap C) - A$ .  
 13. One answer is  $(A \cup B \cup C) - (A \cap B \cap C)$ .



## Section 1.8

1. Suppose  $A_1 = \{a, b, d, e, g, f\}$ ,  $A_2 = \{a, b, c, d\}$ ,  $A_3 = \{b, d, a\}$  and  $A_4 = \{a, b, h\}$ .

(a)  $\bigcup_{i=1}^4 A_i = \{a, b, c, d, e, f, g, h\}$       (b)  $\bigcap_{i=1}^4 A_i = \{a, b\}$

3. For each  $n \in \mathbb{N}$ , let  $A_n = \{0, 1, 2, 3, \dots, n\}$ .

(a)  $\bigcup_{i \in \mathbb{N}} A_i = \{0\} \cup \mathbb{N}$       (b)  $\bigcap_{i \in \mathbb{N}} A_i = \{0, 1\}$

5. (a)  $\bigcup_{i \in \mathbb{N}} [i, i+1] = [1, \infty)$

(b)  $\bigcap_{i \in \mathbb{N}} [i, i+1] = \emptyset$

7. (a)  $\bigcup_{i \in \mathbb{N}} \mathbb{R} \times [i, i+1] = \{(x, y) : x, y \in \mathbb{R}, y \geq 1\}$

(b)  $\bigcap_{i \in \mathbb{N}} \mathbb{R} \times [i, i+1] = \emptyset$

9. (a)  $\bigcup_{X \in \mathcal{P}(\mathbb{N})} X = \mathbb{N}$

(b)  $\bigcap_{X \in \mathcal{P}(\mathbb{N})} X = \emptyset$

11. Yes, this is always true.

13. The first is true, the second is false.

## Chapter 2 Exercises

### Section 2.1

- Every real number is an even integer. (Statement, False)
- If  $x$  and  $y$  are real numbers and  $5x = 5y$ , then  $x = y$ . (Statement, True)
- Sets  $\mathbb{Z}$  and  $\mathbb{N}$  are infinite. (Statement, True)

- 7.** The derivative of any polynomial of degree 5 is a polynomial of degree 6. (Statement, False)

- 9.**  $\cos(x) = -1$

This is not a statement. It is an open sentence because whether it's true or false depends on the value of  $x$ .

- 11.** The integer  $x$  is a multiple of 7.

This is an open sentence, and not a statement.

- 13.** Either  $x$  is a multiple of 7, or it is not.

This is a statement, for the sentence is true no matter what  $x$  is.

- 15.** In the beginning God created the heaven and the earth.

This is a statement, for it is either definitely true or definitely false. There is some controversy over whether it's true or false, but no one claims that it is neither true nor false.

## Section 2.2

Express each statement as one of the forms  $P \wedge Q$ ,  $P \vee Q$ , or  $\sim P$ . Be sure to also state exactly what statements  $P$  and  $Q$  stand for.

- 1.** The number 8 is both even and a power of 2.

$$P \wedge Q$$

$P$ : 8 is even

$Q$ : 8 is a power of 2

Note: Do not say “ $Q$ : a power of 2,” because that is not a statement.

**3.**  $x \neq y$        $\sim(x = y)$       (Also  $\sim P$  where  $P : x = y$ .)

**5.**  $y \geq x$        $\sim(y < x)$       (Also  $\sim P$  where  $P : y < x$ .)

- 7.** The number  $x$  equals zero, but the number  $y$  does not.

$$P \wedge \sim Q$$

$P : x = 0$

$Q : y \neq 0$

- 9.**  $x \in A - B$

$(x \in A) \wedge \sim(x \in B)$

- 11.**  $A \in \{X \in \mathcal{P}(\mathbb{N}) : |\overline{X}| < \infty\}$

$(A \subseteq \mathbb{N}) \wedge (|\overline{A}| < \infty)$ .

- 13.** Human beings want to be good, but not too good, and not all the time.

$$P \wedge \sim Q \wedge \sim R$$

$P$  : Human beings want to be good.

$Q$  : Human beings want to be too good.

$R$  : Human beings want to be good all the time.

## Section 2.3

Without changing their meanings, convert each of the following sentences into a sentence having the form “*If P, then Q*.”

1. A matrix is invertible provided that its determinant is not zero.  
Answer: If a matrix has a determinant not equal to zero, then it is invertible.
3. For a function to be continuous, it is necessary that it is integrable.  
Answer: If a function is continuous, then it is integrable.
5. An integer is divisible by 8 only if it is divisible by 4.  
Answer: If an integer is divisible by 8, then it is divisible by 4.
7. A series converges whenever it converges absolutely.  
Answer: If a series converges absolutely, then it converges.
9. A function is integrable provided the function is continuous.  
Answer: If a function is continuous, then that function is integrable.
11. You fail only if you stop writing.  
Answer: If you fail, then you have stopped writing.
13. Whenever people agree with me I feel I must be wrong.  
Answer: If people agree with me, then I feel I must be wrong.

### Section 2.4

Without changing their meanings, convert each of the following sentences into a sentence having the form “*P if and only if Q*.”

1. For a matrix to be invertible, it is necessary and sufficient that its determinant is not zero.  
Answer: A matrix is invertible if and only if its determinant is not zero.
3. If  $xy = 0$  then  $x = 0$  or  $y = 0$ , and conversely.  
Answer:  $xy = 0$  if and only if  $x = 0$  or  $y = 0$
5. For an occurrence to become an adventure, it is necessary and sufficient for one to recount it.  
Answer: An occurrence becomes an adventure if and only if one recounts it.

### Section 2.5

1. Write a truth table for  $P \vee (Q \Rightarrow R)$
3. Write a truth table for  $\sim(P \Rightarrow Q)$

$P$	$Q$	$R$	$Q \Rightarrow R$	$P \vee (Q \Rightarrow R)$
T	T	T	T	T
T	T	F	F	T
T	F	T	T	T
T	F	F	T	T
F	T	T	T	T
F	T	F	F	F
F	F	T	T	T
F	F	F	T	T

$P$	$Q$	$P \Rightarrow Q$	$\sim(P \Rightarrow Q)$
T	T	T	F
T	F	F	T
F	T	T	F
F	F	T	F

5. Write a truth table for  $(P \wedge \sim P) \vee Q$     7. Write a truth table for  $(P \wedge \sim P) \Rightarrow Q$

$P$	$Q$	$(P \wedge \sim P)$	$(P \wedge \sim P) \vee Q$
$T$	$T$	$F$	$T$
$T$	$F$	$F$	$F$
$F$	$T$	$F$	$T$
$F$	$F$	$F$	$F$

$P$	$Q$	$(P \wedge \sim P)$	$(P \wedge \sim P) \Rightarrow Q$
$T$	$T$	$F$	$T$
$T$	$F$	$F$	$T$
$F$	$T$	$F$	$T$
$F$	$F$	$F$	$T$

9. Write a truth table for  $\sim(\sim P \vee \sim Q)$ .

$P$	$Q$	$\sim P$	$\sim Q$	$\sim P \vee \sim Q$	$\sim(\sim P \vee \sim Q)$
$T$	$T$	$F$	$F$	$F$	$T$
$T$	$F$	$F$	$T$	$T$	$F$
$F$	$T$	$T$	$F$	$T$	$F$
$F$	$F$	$T$	$T$	$T$	$F$

11. Suppose  $P$  is false and that the statement  $(R \Rightarrow S) \Leftrightarrow (P \wedge Q)$  is true. Find the truth values of  $R$  and  $S$ . (This can be done without a truth table.)

Answer: Since  $P$  is false, it follows that  $(P \wedge Q)$  is false also. But then in order for  $(R \Rightarrow S) \Leftrightarrow (P \wedge Q)$  to be true, it must be that  $(R \Rightarrow S)$  is false. The only way for  $(R \Rightarrow S)$  to be false is if  $R$  is true and  $S$  is false.

## Section 2.6

1.  $P \wedge (Q \vee R) = (P \wedge Q) \vee (P \wedge R)$

$P$	$Q$	$R$	$Q \vee R$	$P \wedge Q$	$P \wedge R$	$P \wedge (Q \vee R)$	$(P \wedge Q) \vee (P \wedge R)$
$T$	$T$	$T$	$T$	$T$	$T$	$T$	$T$
$T$	$T$	$F$	$T$	$T$	$F$	$T$	$T$
$T$	$F$	$T$	$T$	$F$	$T$	$T$	$T$
$T$	$F$	$F$	$F$	$F$	$F$	$F$	$F$
$F$	$T$	$T$	$T$	$F$	$F$	$F$	$F$
$F$	$T$	$F$	$T$	$F$	$F$	$F$	$F$
$F$	$F$	$T$	$T$	$F$	$F$	$F$	$F$
$F$	$F$	$F$	$F$	$F$	$F$	$F$	$F$

Thus since the columns agree, the two statements are logically equivalent.

3.  $P \Rightarrow Q = (\sim P) \vee Q$

$P$	$Q$	$\sim P$	$(\sim P) \vee Q$	$P \Rightarrow Q$
$T$	$T$	$F$	$T$	$T$
$T$	$F$	$F$	$F$	$F$
$F$	$T$	$T$	$T$	$T$
$F$	$F$	$T$	$T$	$T$

Since the columns agree, the two statements are logically equivalent.

5.  $\sim(P \vee Q \vee R) = (\sim P) \wedge (\sim Q) \wedge (\sim R)$

$P$	$Q$	$R$	$P \vee Q \vee R$	$\sim P$	$\sim Q$	$\sim R$	$\sim(P \vee Q \vee R)$	$(\sim P) \wedge (\sim Q) \wedge (\sim R)$
$T$	$T$	$T$	$T$	$F$	$F$	$F$	$F$	$F$
$T$	$T$	$F$	$T$	$F$	$F$	$T$	$F$	$F$
$T$	$F$	$T$	$T$	$F$	$T$	$F$	$F$	$F$
$T$	$F$	$F$	$T$	$F$	$T$	$T$	$F$	$F$
$F$	$T$	$T$	$T$	$T$	$F$	$F$	$F$	$F$
$F$	$T$	$F$	$T$	$T$	$F$	$T$	$F$	$F$
$F$	$F$	$T$	$T$	$T$	$T$	$F$	$F$	$F$
$F$	$F$	$F$	$F$	$T$	$T$	$T$	$T$	$T$

Since the columns agree, the two statements are logically equivalent.

7.  $P \Rightarrow Q = (P \wedge \sim Q) \Rightarrow (Q \wedge \sim Q)$

$P$	$Q$	$\sim Q$	$P \wedge \sim Q$	$Q \wedge \sim Q$	$(P \wedge \sim Q) \Rightarrow (Q \wedge \sim Q)$	$P \Rightarrow Q$
$T$	$T$	$F$	$F$	$F$	$T$	$T$
$T$	$F$	$T$	$T$	$F$	$F$	$F$
$F$	$T$	$F$	$F$	$F$	$T$	$T$
$F$	$F$	$T$	$F$	$F$	$T$	$T$

Since the columns agree, the two statements are logically equivalent.

9. By DeMorgan's law, we have  $\sim(\sim P \vee \sim Q) = \sim\sim P \wedge \sim\sim Q = P \wedge Q$ . Thus the two statements are logically equivalent.

11.  $(\sim P) \wedge (P \Rightarrow Q)$  and  $\sim(Q \Rightarrow P)$

$P$	$Q$	$\sim P$	$P \Rightarrow Q$	$Q \Rightarrow P$	$(\sim P) \wedge (P \Rightarrow Q)$	$\sim(Q \Rightarrow P)$
$T$	$T$	$F$	$T$	$T$	$F$	$F$
$T$	$F$	$F$	$F$	$T$	$F$	$F$
$F$	$T$	$T$	$T$	$F$	$T$	$T$
$F$	$F$	$T$	$T$	$T$	$T$	$F$

The columns for the two statements do not quite agree, thus the two statements are **not logically equivalent**.

## Section 2.7

Write the following as English sentences. Say if the statements are true or false.

1.  $\forall x \in \mathbb{R}, x^2 > 0$

Answer: For every real number  $x$ ,  $x^2 > 0$ .

Also: For every real number  $x$ , it follows that  $x^2 > 0$ .

Also: The square of any real number is positive. (etc.)

Statement is **false**. Reason: 0 is a real number, but it's not true that  $0^2 > 0$ .

3.  $\exists a \in \mathbb{R}, \forall x \in \mathbb{R}, ax = x.$

Answer: There exists a real number  $a$  for which  $ax = x$  for every real number  $x$ .  
This statement is TRUE. Reason: Consider  $a = 1$ .

5.  $\forall n \in \mathbb{N}, \exists X \in \mathcal{P}(\mathbb{N}), |X| < n$

Answer: For every natural number  $n$ , there is a subset  $X$  of  $\mathbb{N}$  with  $|X| < n$ .  
This statement is TRUE. Reason: Suppose  $n \in \mathbb{N}$ . Let  $X = \emptyset$ . Then  $|X| = 0 < n$ .

7.  $\forall X \subseteq \mathbb{N}, \exists n \in \mathbb{Z}, |X| = n$

Answer: For any subset  $X$  of  $\mathbb{N}$ , there exists an integer  $n$  for which  $|X| = n$ .  
This statement is FALSE. For example, the set  $X = \{2, 4, 6, 8, \dots\}$  of all even natural numbers is infinite, so there does not exist any integer  $n$  for which  $|X| = n$ .

9.  $\forall n \in \mathbb{Z}, \exists m \in \mathbb{Z}, m = n + 5$

Answer: For every integer  $n$  there is another integer  $m$  such that  $m = n + 5$ .  
This statement is TRUE.

## Section 2.9

Translate each of the following sentences into symbolic logic.

1. If  $f$  is a polynomial and its degree is greater than 2, then  $f'$  is not constant.

Translation:  $(P \wedge Q) \Rightarrow R$ , where

$P : f$  is a polynomial,

$Q : f$  has degree greater than 2,

$R : f'$  is not constant.

3. If  $x$  is prime then  $\sqrt{x}$  is not a rational number.

Translation:  $P \Rightarrow \neg Q$ , where

$P : x$  is prime,

$Q : \sqrt{x}$  is a rational number.

5. For every positive number  $\varepsilon$ , there is a positive number  $\delta$  for which  $|x - a| < \delta$  implies  $|f(x) - f(a)| < \varepsilon$ .

Translation:  $\forall \varepsilon \in \mathbb{R}, \varepsilon > 0, \exists \delta \in \mathbb{R}, \delta > 0, (|x - a| < \delta) \Rightarrow (|f(x) - f(a)| < \varepsilon)$

7. There exists a real number  $a$  for which  $a + x = x$  for every real number  $x$ .

Translation:  $\exists a \in \mathbb{R}, \forall x \in \mathbb{R}, a + x = x$

9. If  $x$  is a rational number and  $x \neq 0$ , then  $\tan(x)$  is not a rational number.

Translation:  $((x \in \mathbb{Q}) \wedge (x \neq 0)) \Rightarrow (\tan(x) \notin \mathbb{Q})$

11. There is a Providence that protects idiots, drunkards, children and the United States of America.

One translation is as follows. Let  $R$  be union of the set of idiots, the set of drunkards, the set of children, and the set consisting of the USA. Let  $P$  be the open sentence  $P(x)$ :  $x$  is a Providence. Let  $S$  be the open sentence  $S(x, y)$ :  $x$  protects  $y$ . Then the translation is  $\exists x, \forall y \in R, P(x) \wedge S(x, y)$ .

(Notice that, although this is mathematically correct, some humor has been lost in the translation.)

13. Everything is funny as long as it is happening to somebody else.

Translation:  $\forall x, (\sim M(x) \wedge S(x)) \Rightarrow F(x)$ ,

where  $M(x)$ : *x is happening to me*,  $S(x)$ : *x is happening to someone*, and  $F(x)$ : *x is funny*.

## Section 2.10

Negate the following sentences.

1. The number  $x$  is positive, but the number  $y$  is not positive.

The “but” can be interpreted as “and.” Using DeMorgan’s law, the negation is: *The number x is not positive or the number y is positive.*

3. For every prime number  $p$ , there is another prime number  $q$  with  $q > p$ .

Negation: *There is a prime number p such that for every prime number q,  $q \leq p$ .*  
Also: *There exists a prime number p for which  $q \leq p$  for every prime number q.* (etc.)

5. For every positive number  $\varepsilon$  there is a positive number  $M$  for which  $|f(x) - b| < \varepsilon$  whenever  $x > M$ .

To negate this, it may be helpful to first write it in symbolic form. The statement is  $\forall \varepsilon \in (0, \infty), \exists M \in (0, \infty), (x > M) \Rightarrow (|f(x) - b| < \varepsilon)$ .

Working out the negation, we have

$$\begin{aligned} \sim (\forall \varepsilon \in (0, \infty), \exists M \in (0, \infty), (x > M) \Rightarrow (|f(x) - b| < \varepsilon)) &= \\ \exists \varepsilon \in (0, \infty), \sim (\exists M \in (0, \infty), (x > M) \Rightarrow (|f(x) - b| < \varepsilon)) &= \\ \exists \varepsilon \in (0, \infty), \forall M \in (0, \infty), \sim ((x > M) \Rightarrow (|f(x) - b| < \varepsilon)). \end{aligned}$$

Finally, using the idea from Example 2.15, we can negate the conditional statement that appears here to get

$$\exists \varepsilon \in (0, \infty), \forall M \in (0, \infty), \exists x, (x > M) \wedge \sim (|f(x) - b| < \varepsilon).$$

Negation: *There exists a positive number  $\varepsilon$  with the property that for every positive number  $M$ , there is a number  $x$  for which  $x > M$  and  $|f(x) - b| \geq \varepsilon$ .*

7. I don’t eat anything that has a face.

Negation: *I will eat some things that have a face.*

(Note: If your answer was “I will eat anything that has a face.” then that is wrong, both morally and mathematically.)

9. If  $\sin(x) < 0$ , then it is not the case that  $0 \leq x \leq \pi$ .

Negation: *There exists a number  $x$  for which  $\sin(x) < 0$  and  $0 \leq x \leq \pi$ .*

11. You can fool all of the people all of the time.

There are several ways to negate this, including:

*There is a person that you can’t fool all the time.* or

*There is a person x and a time y for which x is not fooled at time y.*

(But Abraham Lincoln said it better.)

## Chapter 3 Exercises

### Section 3.2

1. Consider lists made from the letters  $T, H, E, O, R, Y$ , with repetition allowed.
  - (a) How many length-4 lists are there? Answer:  $6 \cdot 6 \cdot 6 \cdot 6 = 1296$ .
  - (b) How many length-4 lists are there that begin with  $T$ ?  
Answer:  $1 \cdot 6 \cdot 6 \cdot 6 = 216$ .
  - (c) How many length-4 lists are there that do not begin with  $T$ ?  
Answer:  $5 \cdot 6 \cdot 6 \cdot 6 = 1080$ .
3. How many ways can you make a list of length 3 from symbols  $A, B, C, D, E, F$  if...
  - (a) ... repetition is allowed. Answer:  $6 \cdot 6 \cdot 6 = 216$ .
  - (b) ... repetition is not allowed. Answer:  $6 \cdot 5 \cdot 4 = 120$ .
  - (c) ... repetition is not allowed and the list must contain the letter  $A$ .  
Answer:  $5 \cdot 4 + 5 \cdot 4 + 5 \cdot 4 = 60$ .
  - (d) ... repetition is allowed and the list must contain the letter  $A$ .  
Answer:  $6 \cdot 6 \cdot 6 - 5 \cdot 5 \cdot 5 = 91$ .

(Note: See Example 3.3 if a more detailed explanation is required.)
5. This problem involves 8-digit binary strings such as 10011011 or 00001010. (i.e., 8-digit numbers composed of 0's and 1's.)
  - (a) How many such strings are there? Answer:  $2 \cdot 2 = 256$ .
  - (b) How many such strings end in 0? Answer:  $2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 1 = 128$ .
  - (c) How many such strings have the property that their second and fourth digits are 1's? Answer:  $2 \cdot 1 \cdot 2 \cdot 1 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 64$ .
  - (d) How many such strings are such that their second **or** fourth digits are 1's?  
Solution: These strings can be divided into three types. Type 1 consists of those strings of form  $*1*0****$ , Type 2 consist of strings of form  $*0*1****$ , and Type 3 consists of those of form  $*1*1****$ . By the multiplication principle there are  $2^6 = 64$  strings of each type, so **there are  $3 \cdot 64 = 192$  8-digit binary strings whose second or fourth digits are 1's.**
7. This problem concerns 4-letter codes made from the letters  $A, B, C, D, \dots, Z$ .
  - (a) How many such codes can be made? Answer:  $26 \cdot 26 \cdot 26 \cdot 26 = 456,976$
  - (b) How many such codes have no two consecutive letters the same?  
Solution: We use the multiplication principle. There are 26 choices for the first letter. The second letter can't be the same as the first letter, so there are only 25 choices for it. The third letter can't be the same as the second letter, so there are only 25 choices for it. The fourth letter can't be the same as the third letter, so there are only 25 choices for it. **Thus there are  $26 \cdot 25 \cdot 25 \cdot 25 = 406,250$  codes with no two consecutive letters the same.**
9. A new car comes in a choice of five colors, three engine sizes and two transmissions. How many different combinations are there? Answer  $5 \cdot 3 \cdot 2 = 30$ .

### Section 3.3

1. Five cards are dealt off of a standard 52-card deck and lined up in a row. How many such lineups are there that have at least one red card?

Solution: All together there are  $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 = 311,875,200$  possible lineups. The number of lineups that **do not** have any red cards (i.e. are made up only of black cards) is  $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7,893,600$ . By the subtraction principle, the answer to the question is  $311,875,200 - 7,893,600 = \mathbf{303,981,600}$ .

How many such lineups are there in which the cards are all black or all hearts?

Solution: The number of lineups that are all black is  $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7,893,600$ . The number of lineups that are hearts (which are red) is  $13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 = 154,440$ . By the addition principle, the answer to the question is  $7,893,600 + 154,440 = \mathbf{8,048,040}$ .

3. Five cards are dealt off of a standard 52-card deck and lined up in a row. How many such lineups are there in which all 5 cards are of the same color (i.e., all black or all red)?

Solution: There are  $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7,893,600$  possible black-card lineups and  $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7,893,600$  possible red-card lineups, so by the addition principle the answer is  $7,893,600 + 7,893,600 = \mathbf{15,787,200}$ .

5. How many integers between 1 and 9999 have no repeated digits?

Solution: Consider the 1-digit, 2-digit, 3-digit and 4-digit number separately. The number of 1-digit numbers that have no repeated digits is 9 (i.e., all of them). The number of 2-digit numbers that have no repeated digits is  $9 \cdot 9 = 81$ . (The number can't begin in 0, so there are only 9 choices for its first digit.) The number of 3-digit numbers that have no repeated digits is  $9 \cdot 9 \cdot 8 = 648$ . The number of 4-digit numbers that have no repeated digits is  $9 \cdot 9 \cdot 8 \cdot 7 = 4536$ . By the addition principle, the answer to the question is  $9 + 81 + 648 + 4536 = \mathbf{5274}$ .

How many integers between 1 and 9999 have at least one repeated digit?

Solution: The total number of integers between 1 and 9999 is 9999. Using the subtraction principle, we can subtract from this the number of digits that have *no* repeated digits, which is 5274, as above. Therefore the answer to the question is  $9999 - 5274 = \mathbf{4725}$ .

7. A password on a certain site must have five characters made from letters of the alphabet, and there must be at least one upper case letter. How many different passwords are there?

Solution: Let  $U$  be the set of all possible passwords made from a choice of upper and lower case letters. Let  $X$  be the set of all possible passwords made from lower case letters. Then  $U - X$  is the set of passwords that have at least one lower case letter. By the subtraction principle our answer will be  $|U - X| = |U| - |X|$ .

All together, there are  $26 + 26 = 52$  upper and lower case letters, so by the multiplication principle  $|U| = 52 \cdot 52 \cdot 52 \cdot 52 \cdot 52 = 52^5 = 380,204,032$ .

Likewise  $|X| = 26 \cdot 26 \cdot 26 \cdot 26 \cdot 26 = 26^5 = 11,881,376$ .

Thus the answer is  $|U| - |X| = 380,204,032 - 11,881,376 = \mathbf{368,322,656}$ .

What if there must be a mix of upper and lower case?

Solution: The number of passwords using only upper case letters is  $26^5 = 11,881,376$ , and, as calculated above, this is also the number of passwords that use only lower case letters. By the addition principle, the number of passwords that use only lower case or only upper case is  $11,881,376 + 11,881,376 = 23,762,752$ . By the subtraction principle, the number of passwords that use a mix of upper and lower case is the total number of possible passwords minus the number that use only lower case or only upper case, namely  $380,204,032 - 23,762,752 = \mathbf{356,441,280}$ .

- 9.** This problem concerns lists of length 6 made from the letters  $A, B, C, D, E, F, G, H$ . How many such lists are possible if repetition is not allowed and the list contains two consecutive vowels?

Solution: There are just two vowels  $A$  and  $E$  to choose from. The lists we want to make can be divided into five types. They have one of the forms  $VV****$ , or  $*VV***$ , or  $**V*V*$ , or  $***VV*$ , or  $****VV$ , where  $V$  indicates a vowel and  $*$  indicates a consonant. By the multiplication principle, there are  $2 \cdot 1 \cdot 6 \cdot 5 \cdot 4 \cdot 3 = 720$  lists of form  $VV****$ . In fact, that for the same reason there are 720 lists of each form. Thus by the addition principle, the answer to the question is  $720 + 720 + 720 + 720 + 720 = \mathbf{3600}$

- 11.** How many integers between 1 and 1000 are divisible by 5? How many are not?

Solution: The integers that are divisible by 5 are  $5, 10, 15, 20, \dots, 995, 1000$ . There are  $1000/5 = \mathbf{200}$  such numbers. By the subtraction principle, the number that are **not** divisible by 5 is  $1000 - 200 = \mathbf{800}$ .

### Sections 3.4

- 1.** Answer  $n = 14$ .      **5.**  $\frac{120!}{118!} = \frac{120 \cdot 119 \cdot 118!}{118!} = 120 \cdot 119 = \mathbf{14,280}$ .  
**3.** Answer:  $5! = \mathbf{120}$ .      **7.** Answer:  $5!4! = \mathbf{2880}$ .

- 9.** How many permutations of the letters  $A, B, C, D, E, F, G$  are there in which the three letters ABC appear consecutively, in alphabetical order?

Solution: Regard  $ABC$  as a single symbol  $\boxed{ABC}$ . Then we are looking for the number of permutations of the five symbols  $\boxed{ABC}, D, E, F, G$ . The number of such permutations is  $5! = 120$ .

- 11.** You deal 7 cards off of a 52-card deck and line them up in a row. How many possible lineups are there in which not all cards are red?

Solution: All together, there are  $P(52, 7)$  7-card lineups with cards selected from the entire deck. And there are  $P(26, 7)$  7-card lineups with red cards selected from the 26 red cards in the deck. By the subtraction principle, the number of lineups that are not all red is  $P(52, 7) - P(26, 7) = \mathbf{670,958,870,400}$ .

- 13.**  $P(26, 6) = 165,765,600$     **15.**  $P(15, 4) = 32,760$       **17.**  $P(10, 3) = 720$

### Section 3.5

1. Suppose a set  $A$  has 37 elements. How many subsets of  $A$  have 10 elements? How many subsets have 30 elements? How many have 0 elements?  
Answers:  $\binom{37}{10} = 348,330,136$ ;  $\binom{37}{30} = 10,295,472$ ;  $\binom{37}{0} = 1$ .
3. A set  $X$  has exactly 56 subsets with 3 elements. What is the cardinality of  $X$ ?  
Solution: The answer will be the  $n$  for which  $\binom{n}{3} = 56$ . After some trial and error, you will discover  $\binom{8}{3} = 56$ , so  $|X| = 8$ .
5. How many 16-digit binary strings contain exactly seven 1's?  
Solution: Make such a string as follows. Start with a list of 16 blank spots. Choose 7 of the blank spots for the 1's and put 0's in the other spots. There are  $\binom{16}{7} = 11,440$  ways to do this.
7.  $|\{X \in \mathcal{P}(\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}) : |X| < 4\}| = \binom{10}{0} + \binom{10}{1} + \binom{10}{2} + \binom{10}{3} = 1 + 10 + 45 + 120 = 176$ .
9. This problem concerns lists of length six made from the letters  $A, B, C, D, E, F$ , without repetition. How many such lists have the property that the  $D$  occurs before the  $A$ ?  
Solution: Make such a list as follows. Begin with six blank spaces and select two of these spaces. Put the  $D$  in the first selected space and the  $A$  in the second. There are  $\binom{6}{2} = 15$  ways of doing this. For each of these 15 choices there are  $4! = 24$  ways of filling in the remaining spaces. Thus the answer to the question is  $15 \times 24 = 360$  such lists.
11. How many 10-digit integers contain no 0's and exactly three 6's?  
Solution: Make such a number as follows: Start with 10 blank spaces and choose three of these spaces for the 6's. There are  $\binom{10}{3} = 120$  ways of doing this. For each of these 120 choices we can fill in the remaining seven blanks with choices from the digits 1, 2, 3, 4, 5, 7, 8, 9, and there are  $8^7$  to do this. Thus the answer to the question is  $\binom{10}{3} \cdot 8^7 = 251,658,240$ .
13. Assume  $n, k \in \mathbb{Z}$  with  $0 \leq k \leq n$ . Then  $\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n!}{k!(n-k)!} = \frac{n!}{(n-(n-k))!(n-k)!} = \binom{n}{n-k}$ .
15. How many 10-digit binary strings are there that do not have exactly four 1's?  
Solution: All together, there are  $2^{10}$  different binary strings. The number of 10-digit binary strings with exactly four 1's is  $\binom{10}{4}$ , because to make one we need to choose 4 out of 10 positions for the 1's and fill the rest in with 0's. By the subtraction principle, the answer to our questions is  $2^{10} - \binom{10}{4}$ .
17. How many 10-digit binary numbers are there that have exactly four 1's or exactly five 1's?  
Solution: By the addition principle the answer is  $\binom{10}{4} + \binom{10}{5}$ .  
How many do not have exactly four 1's or exactly five 1's?  
Solution: By the subtraction principle combined with the answer to the first part of this problem, the answer is  $2^{10} - \binom{10}{4} - \binom{10}{5}$
19. A 5-card poker hand is called a *flush* if all cards are the same suit. How many different flushes are there?

**Solution:** There are  $\binom{13}{5} = 1287$  5-card hands that are all hearts. Similarly, there are  $\binom{13}{5} = 1287$  5-card hands that are all diamonds, or all clubs, or all spades. By the addition principle, there are then  $1287 + 1287 + 1287 + 1287 = \mathbf{5148}$  flushes.

### Section 3.6

1. Write out Row 11 of Pascal's triangle.

Answer: 1 11 55 165 330 462 462 330 165 55 11 1

3. Use the binomial theorem to find the coefficient of  $x^8$  in  $(x+2)^{13}$ .

Answer: According to the binomial theorem, the coefficient of  $x^8y^5$  in  $(x+y)^{13}$  is  $\binom{13}{5}x^8y^5 = 1287x^8y^5$ . Now plug in  $y=2$  to get the final answer of  $41184x^8$ .

5. Use the binomial theorem to show  $\sum_{k=0}^n \binom{n}{k} = 2^n$ . Hint: Observe that  $2^n = (1+1)^n$ . Now use the binomial theorem to work out  $(x+y)^n$  and plug in  $x=1$  and  $y=1$ .

7. Use the binomial theorem to show  $\sum_{k=0}^n 3^k \binom{n}{k} = 4^n$ .

Hint: Observe that  $4^n = (1+3)^n$ . Now look at the hint for the previous problem.

9. Use the binomial theorem to show  $\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \binom{n}{4} - \binom{n}{5} + \dots \pm \binom{n}{n} = 0$ . Hint: Observe that  $0 = 0^n = (1+(-1))^n$ . Now use the binomial theorem.

11. Use the binomial theorem to show  $9^n = \sum_{k=0}^n (-1)^k \binom{n}{k} 10^{n-k}$ .

Hint: Observe that  $9^n = (10+(-1))^n$ . Now use the binomial theorem.

13. Assume  $n \geq 3$ . Then  $\binom{n}{3} = \binom{n-1}{3} + \binom{n-1}{2} = \binom{n-2}{3} + \binom{n-2}{2} + \binom{n-1}{2} = \dots = \binom{2}{3} + \binom{3}{2} + \dots + \binom{n-1}{2}$ .

### Section 3.7

1. At a certain university 523 of the seniors are history majors or math majors (or both). There are 100 senior math majors, and 33 seniors are majoring in both history and math. How many seniors are majoring in history?

**Solution:** Let  $A$  be the set of senior math majors and  $B$  be the set of senior history majors. From  $|A \cup B| = |A| + |B| - |A \cap B|$  we get  $523 = 100 + |B| - 33$ , so  $|B| = 523 + 33 - 100 = 456$ . **There are 456 history majors.**

3. How many 4-digit positive integers are there that are even or contain no 0's?

**Solution:** Let  $A$  be the set of 4-digit even positive integers, and let  $B$  be the set of 4-digit positive integers that contain no 0's. We seek  $|A \cup B|$ . By the multiplication principle  $|A| = 9 \cdot 10 \cdot 10 \cdot 5 = 4500$ . (Note the first digit cannot be 0 and the last digit must be even.) Also  $|B| = 9 \cdot 9 \cdot 9 \cdot 9 = 6561$ . Further,  $A \cap B$  consists of all even 4-digit integers that have no 0's. It follows that  $|A \cap B| = 9 \cdot 9 \cdot 9 \cdot 4 = 2916$ . Then the answer to our question is  $|A \cup B| = |A| + |B| - |A \cap B| = 4500 + 6561 - 2916 = \mathbf{8145}$ .

5. How many 7-digit binary strings begin in 1 or end in 1 or have exactly four 1's?

**Solution:** Let  $A$  be the set of such strings that begin in 1. Let  $B$  be the set of such strings that end in 1. Let  $C$  be the set of such strings that have exactly four 1's. Then the answer to our question is  $|A \cup B \cup C|$ . Using Equation (3.5) to compute this number, we have  $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C| = 2^6 + 2^6 + \binom{7}{4} - 2^5 - \binom{6}{3} - \binom{6}{2} + \binom{5}{2} = 64 + 64 + 35 - 32 - 20 - 20 + 10 = \mathbf{101}$ .

7. This problem concerns 4-card hands dealt off of a standard 52-card deck. How many 4-card hands are there for which all four cards are of the same suit or all four cards are red?

**Solution:** Let  $A$  be the set of 4-card hands for which all four cards are of the same suit. Let  $B$  be the set of 4-card hands for which all four cards are red. Then  $A \cap B$  is the set of 4-card hands for which the four cards are either all hearts or all diamonds. The answer to our question is  $|A \cup B| = |A| + |B| - |A \cap B| = 4\binom{13}{4} + \binom{26}{4} - 2\binom{13}{4} = 2\binom{13}{4} + \binom{26}{4} = 1430 + 14,950 = \mathbf{16,380}$ .

- 9.** A 4-letter list is made from the letters  $L,I,S,T,E,D$  according to the following rule: Repetition is allowed, and the first two letters on the list are vowels or the list ends in  $D$ . How many such lists are possible?

**Solution:** Let  $A$  be the set of such lists for which the first two letters are vowels, so  $|A| = 2 \cdot 2 \cdot 6 \cdot 6 = 144$ . Let  $B$  be the set of such lists that end in  $D$ , so  $|B| = 6 \cdot 6 \cdot 6 \cdot 1 = 216$ . Then  $A \cap B$  is the set of such lists for which the first two entries are vowels and the list ends in  $D$ . Thus  $|A \cap B| = 2 \cdot 2 \cdot 6 \cdot 1 = 24$ . The answer to our question is  $|A \cup B| = |A| + |B| - |A \cap B| = 144 + 216 - 24 = \mathbf{336}$ .

- 11.** How many 7-digit numbers are even or have exactly three digits equal to 0?

**Solution:** Let  $A$  be the set of 7-digit numbers that are even. By the multiplication principle,  $|A| = 9 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 10 \cdot 5 = 4,500,000$ . Let  $B$  be the set of 7-digit numbers that have exactly three digits equal to 0. Then  $|B| = 9 \cdot \binom{6}{3} \cdot 9 \cdot 9 \cdot 9$ . (First digit is anything but 0. Then choose 3 of 6 of the remaining places in the number for the 0's. Finally the remaining 3 places can be anything but 0.)

Note  $A \cap B$  is the set of 7-digit numbers that are even and contain exactly three 0's. We can compute  $|A \cap B|$  with the addition principle, by dividing  $A \cap B$  into two parts: the even 7-digit numbers with three digits 0 and the last digit **is not** 0, and the even 7-digit numbers with three digits 0 and the last digit **is** 0. The first part has  $9 \cdot \binom{5}{3} \cdot 9 \cdot 9 \cdot 4$  elements. The second part has  $9 \cdot \binom{5}{2} \cdot 9 \cdot 9 \cdot 9 \cdot 1$  elements. Thus  $|A \cap B| = 9 \cdot \binom{5}{3} \cdot 9 \cdot 9 \cdot 4 + 9 \cdot \binom{5}{2} \cdot 9 \cdot 9 \cdot 9$ .

By the inclusion-exclusion formula, the answer to our question is  $|A \cup B| = |A| + |B| - |A \cap B| = 4,500,000 + 9^4 \binom{6}{3} - 9^3 \binom{5}{3} \cdot 4 - 9^4 \binom{5}{2} = 4,405,230$ .

- 13.** How many 8-digit binary strings end in 1 or have exactly four 1's?

**Solution:** Let  $A$  be the set of strings that end in 1. By the multiplication principle  $|A| = 2^7$ . Let  $B$  be the number of strings with exactly four 1's. Then  $|B| = \binom{8}{4}$  because we can make such a string by choosing 4 of 8 spots for the 1's and filling the remaining spots with 0's. Then  $A \cap B$  is the set of strings that end with 1 and have exactly four 1's. Note that  $|A \cap B| = \binom{7}{4}$  (make the last entry a 1 and choose 3 of the remaining 7 spots for 1's). By the inclusion-exclusion formula, the number 8-digit binary strings that end in 1 or have exactly four 1's is  $|A \cup B| = |A| + |B| - |A \cap B| = 2^7 + \binom{8}{4} - \binom{7}{4} = 163$ .

- 15.** How many 10-digit binary strings begin in 1 or end in 1?

**Solution:** Let  $A$  be the set of strings that begin with 1. By the multiplication principle  $|A| = 2^9$ . Let  $B$  be the number of strings that end with 1. By the multiplication principle  $|B| = 2^9$ . Then  $A \cap B$  is the set of strings that begin and end with 1. By the multiplication principle  $|A \cap B| = 2^8$ . By the inclusion-exclusion formula, the number 10-digit binary strings begin in 1 or end in 1 is  $|A \cup B| = |A| + |B| - |A \cap B| = 2^9 + 2^9 - 2^8 = 768$ .

### Section 3.8

1. How many 10-element multisets can be made from the symbols {1, 2, 3, 4}?

Answer:  $\binom{10+4-1}{10} = \binom{13}{10} = 286$ .

3. You have a dollar in pennies, a dollar in nickels, a dollar in dimes and a dollar in quarters. You give four coins to a friend. In how many ways can this be done?

Solution: In giving your friend four coins, you are giving her a 4-element multiset made from elements in {1, 5, 10, 25}. There are  $\binom{4+4-1}{4} = \binom{7}{4} = 35$  such multisets.

5. A bag contains 20 identical red balls, 20 identical blue balls, 20 identical green balls, and one white ball. You reach in and grab 15 balls. How many different outcomes are possible?

Solution: First we count the number of outcomes that don't have a white ball. Modeling this with stars and bars, we are looking at length-17 lists of the form

$$\overbrace{\ast \ast \ast \dots \ast}^{\text{red}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{blue}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{green}},$$

where there are 15 stars and two bars. Therefore there are  $\binom{17}{15}$  outcomes without the white ball. Next we count the outcomes that do have the white ball. Then there are 14 remaining balls in the grab. In counting the ways that they can be selected we can use the same stars-and-bars model above, but this time the list is of length 16 and has 14 stars. There are  $\binom{16}{14}$  outcomes. Finally, by the addition principle, the answer to the question is  $\binom{17}{15} + \binom{16}{14} = 256$ .

7. In how many ways can you place 20 identical balls into five different boxes?

Solution: Let's model this with stars and bars. Doing this we get a list of length 24 with 20 stars and 4 bars, where the first grouping of stars has as many stars as balls in Box 1, the second grouping has as many stars as balls in Box 2, and so on.

$$\overbrace{\ast \ast \ast \dots \ast}^{\text{Box 1}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{Box 2}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{Box 3}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{Box 4}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{Box 5}},$$

The number of ways to place 20 balls in the five boxes equals the number of such lists, which is  $\binom{24}{20} = 10,626$ .

9. A bag contains 50 pennies, 50 nickels, 50 dimes and 50 quarters. You reach in and grab 30 coins. How many different outcomes are possible?

Solution: The stars-and-bars model is

$$\overbrace{\ast \ast \ast \dots \ast}^{\text{pennies}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{nickels}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{dimes}} | \overbrace{\ast \ast \ast \dots \ast}^{\text{quarters}},$$

so there are  $\binom{33}{30} = 5456$  outcomes.

11. How many integer solutions does the equation  $w + x + y + z = 100$  have if  $w \geq 4$ ,  $x \geq 2$ ,  $y \geq 0$  and  $z \geq 0$ ?

**Solution:** Imagine a bag containing 100 red balls, 100 blue balls, 100 green balls and 100 white balls. Each solution of the equation corresponds to an outcome in selecting 100 balls from the bag, where the selection includes  $w \geq 4$  red balls,  $x \geq 2$  blue balls,  $y \geq 0$  green balls and  $z \geq 0$  white balls.

Now let's consider making such a selection. Pre-select 4 red balls and 2 blue balls, so 94 balls remain in the bag. Next the remaining 94 balls are selected. We can calculate the number of ways that this selection can be made with stars and bars, where there are 94 stars and 3 bars.

$$\overbrace{***\dots*}^{\text{red}} | \overbrace{***\dots*}^{\text{blue}} | \overbrace{***\dots*}^{\text{green}} | \overbrace{***\dots*}^{\text{white}},$$

The number of outcomes is thus  $\binom{94}{3} = 134,044$ .

13. How many length-6 lists can be made from the symbols {A, B, C, D, E, F, G}, if repetition is allowed and the list is in alphabetical order?

**Solution:** Any such list corresponds to a 6-element multiset made from the symbols {A, B, C, D, E, F, G}. For example, the list AACDDG corresponds to the multiset [A,A,C,D,D,G]. Thus the number of lists equals the number of multisets, which is  $\binom{6+7-1}{6} = \binom{12}{6} = 924$ .

15. How many permutations are there of the letters in the word “TENNESSEE”?

**Solution:** By Fact 3.8, the answer is  $\frac{9!}{4!2!2!} = 3,780$ .

17. You roll a dice six times in a row. How many possible outcomes are there that have two 1's three 5's and one 6?

**Solution:** This is the number of permutations of the “word”  $\square\square\square\square\square\square$ . By Fact 3.8, the answer is  $\frac{6!}{2!3!1!} = 60$ .

19. In how many ways can you place 15 identical balls into 20 different boxes if each box can hold at most one ball?

**Solution:** Regard each such distribution as a binary string of length 20, where there is a 1 in the  $i$ th position precisely if the  $i$ th box contains a ball (and zeros elsewhere). The answer is the number of permutations of such a string, which by Fact 3.8 is  $\frac{20!}{15!5!} = 15,504$ . Alternatively, the answer is the number of ways to choose 15 positions out of 20, which is  $\binom{20}{15} = 15,504$ .

21. How many numbers between 10,000 and 99,999 contain one or more of the digits 3, 4 and 8, but no others?

**Solution:** First count the numbers that have three 3's, one 4, and one 8, like 33,348. By Fact 3.8, the number of permutations of this is  $\frac{5!}{3!1!1!} = 20$ .

By the same reasoning there are 20 numbers that contain three 4's, one 3, and one 8, and 20 numbers that contain three 8's, one 3, and one 4.

Next, consider the numbers that have two 3's, two 4's and one 8, like 33,448. By Fact 3.8, the number of permutations of this is  $\frac{5!}{2!2!1!} = 30$ .

By the same reasoning there are 30 numbers that contain two 3's, two 8's and one 4, and 30 numbers that contain two 4's, two 8's and one 3. This exhausts all possibilities. By the addition principle the answer is  $20+20+20+30+30+30 = 150$ .

### Section 3.9

1. Show that if 6 integers are chosen at random, at least two will have the same remainder when divided by 5.

Solution: Pick six integers  $n_1, n_2, n_3, n_4, n_5$  and  $n_6$  at random. Imagine five boxes, labeled Box 0, Box 1, Box 2, Box 3, Box 4. Each of the picked integers has a remainder when divided by 5, and that remainder is 0, 1, 2, 3 or 4. For each  $n_i$ , let  $r_i$  be its remainder when divided by 5. Put  $n_i$  in Box  $r_i$ . We have now put six numbers in five boxes, so by the pigeonhole principle one of the boxes has two or more of the picked numbers in it. Those two numbers have the same remainder when divided by 5.

3. What is the fewest number of times you must roll a six-sided dice before you can be assured that 10 or more of the rolls resulted in the same number?

Solution: Imagine six boxes, labeled 1 through 6. Every time you roll a  $\square$ , put an object in Box 1. Every time you roll a  $\square$ , put an object in Box 2, etc. After  $n$  rolls, the division principle says that one box contains  $\lceil \frac{n}{6} \rceil$  objects, and this means you rolled the same number  $\lceil \frac{n}{6} \rceil$  times. We seek the smallest  $n$  for which  $\lceil \frac{n}{6} \rceil \geq 10$ . This is the smallest  $n$  for which  $\frac{n}{6} > 9$ , that is  $n > 9 \cdot 6 = 36$ . Thus the answer is  $n = 37$ . You need to roll the dice 37 times.

5. Prove that any set of 7 distinct natural numbers contains a pair of numbers whose sum or difference is divisible by 10.

Solution: Let  $S = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$  be any set of 7 natural numbers. Let's say that  $a_1 < a_2 < a_3 < \dots < a_7$ . Consider the set

$$\begin{aligned} A = & \{a_1 - a_2, a_1 - a_3, a_1 - a_4, a_1 - a_5, a_1 - a_6, a_1 - a_7, \\ & a_1 + a_2, a_1 + a_3, a_1 + a_4, a_1 + a_5, a_1 + a_6, a_1 + a_7\} \end{aligned}$$

Thus  $|A| = 12$ . Now imagine 10 boxes numbered 0, 1, 2, ..., 9. For each number  $a_1 \pm a_i \in A$ , put it in the box whose number is the one's digit of  $a_1 \pm a_i$ . (For example, if  $a_1 \pm a_i = 4$ , put it in Box 4. If  $a_1 \pm a_i = 8$ , put it in Box 8, etc.) Now we have placed the 12 numbers in  $A$  into 10 boxes, so the pigeonhole principle says at least one box contains two elements  $a_1 \pm a_i$  and  $a_1 \pm a_j$  from  $A$ . This means the last digit of  $a_1 \pm a_i$  is the same as the last digit of  $a_1 \pm a_j$ . Thus the last digit of the difference  $(a_1 \pm a_i) - (a_1 \pm a_j) = \pm a_i \pm a_j$  is 0. Hence  $\pm a_i \pm a_j$  is a sum or difference of elements of  $S$  that is divisible by 10.

### Section 3.10

1. Show that  $1(n-0) + 2(n-1) + 3(n-2) + 4(n-3) + \dots + (n-1)2 + (n-0)1 = \binom{n+2}{3}$ .

Solution: Let  $S = \{0, 1, 2, 3, \dots, n, n+1\}$ , which is a set with  $n+2$  elements. The right-hand side  $\binom{n+2}{3}$  of our equations is the number of 3-element subsets of  $S$ .

Let's now count these 3-element subsets in a different way. Any such subset  $X$  can be written as  $X = \{j, k, \ell\}$ , where  $0 \leq j < k < \ell \leq n+1$ . Note that this forces the middle element  $k$  to be in the range  $1 \leq k \leq n$ . Given a fixed middle element  $k$ ,

there are  $k$  choices for the smallest element  $j$  and  $n+1-k$  choices for the largest element  $\ell$ .

$$\overbrace{0 \quad 1 \quad 2 \quad \cdots \quad k-1}^{k \text{ choices for } j} \quad \begin{matrix} k \\ \uparrow \\ \text{middle} \end{matrix} \quad \overbrace{k+1 \quad k+2 \quad k+3 \quad \cdots \quad n \quad n+1}^{n+1-k \text{ choices for } \ell}$$

By the multiplication principle, there are  $k(n+1-k)$  possible 3-element sets  $X$  with middle element  $k$ . For example, if  $k=1$ , there are  $1(n-0)$  sets  $X$  with middle element 1. If  $k=2$ , there are  $2(n-1)$  sets  $X$  with middle element 2. If  $k=3$ , there are  $3(n-2)$  sets  $X$  with middle element 3. Thus the left-hand side of our equation counts up the number of 3-element subsets of  $S$ , so it is equal to the right-hand side.

3. Show that  $\binom{n}{2}\binom{n-2}{k-2} = \binom{n}{k}\binom{k}{2}$ .

**Solution:** Consider the following problem. From a group of  $n$  people, you need to select  $k$  people to serve on a committee, and you also need to select 2 of these  $k$  people to lead the committee's discussion. In how many ways can this be done?

One approach is to first select  $k$  people from  $n$ , and then select 2 of these  $k$  people to lead the discussion. By the multiplication principle, there are  $\binom{n}{k}\binom{k}{2}$  ways to make this selection.

Another approach is to first select 2 of the  $n$  people to be the discussion leaders, and there are  $\binom{n}{2}$  ways to do this. Next we need to fill out the committee by selecting  $k-2$  people from the remaining  $n-2$  people, and there are  $\binom{n-2}{k-2}$  ways to do this. By the multiplication principle, there are  $\binom{n}{2}\binom{n-2}{k-2}$  ways to make the selection.

By the previous two paragraphs,  $\binom{n}{2}\binom{n-2}{k-2}$  and  $\binom{n}{k}\binom{k}{2}$  are both answers to the same counting problem, so they are equal.

5. Show that  $\binom{2n}{2} = 2\binom{n}{2} + n^2$ .

**Solution:** Let  $S$  be a set with  $2n$  elements. Then the left-hand side counts the number of 2-element subsets of  $S$ .

Let's now count this in a different way. Split  $S$  as  $S = A \cup B$ , where  $|A| = n = |B|$ . We can choose a 2-element subset of  $S$  in three ways: We could choose both elements from  $A$ , and there are  $\binom{n}{2}$  ways to do this. We could choose both elements from  $B$ , and there are  $\binom{n}{2}$  ways to do this. Or we could choose one element from  $A$  and then another element from  $B$ , and by the multiplication principle there are  $n \cdot n = n^2$  ways to do this. Thus the number of 2-element subsets of  $S$  is  $\binom{n}{2} + \binom{n}{2} + n^2 = 2\binom{n}{2} + n^2$ , and this is the right-hand side. Therefore the equation holds because both sides count the same thing.

7. Show that  $\sum_{k=0}^p \binom{m}{k} \binom{n}{p-k} = \binom{m+n}{p}$ .

**Solution:** Take three non-negative integers  $m, n$  and  $p$ . Let  $S$  be a set with  $|S| = m+n$ , so the right-hand side counts the number of  $p$ -element subsets of  $S$ .

Now let's count this in a different way. Split  $S$  as  $S = A \cup B$ , where  $|A| = m$  and  $|B| = n$ . We can make any  $p$ -element subset of  $S$  by choosing  $k$  of its elements from  $A$  in and  $p - k$  of its elements from  $B$ , for any  $0 \leq k \leq p$ . There are  $\binom{m}{k}$  ways to choose  $k$  elements from  $A$ , and  $\binom{n}{p-k}$  ways to choose  $p - k$  elements from  $B$ , so there are  $\binom{m}{k} \binom{n}{p-k}$  ways to make a  $p$ -element subset of  $S$  that has  $k$  elements from  $A$ . As  $k$  could be any number between 0 and  $p$ , the left-hand side of our equation counts up the  $p$ -element subsets of  $S$ . Thus the left- and right-hand sides count the same thing, so they are equal.

9. Show that  $\sum_{k=m}^n \binom{k}{m} = \binom{n+1}{m+1}$ .

**Solution:** Let  $S = \{0, 1, 2, \dots, n\}$ , so  $|S| = n + 1$ . The right-hand side of our equation is the number of subsets  $X$  of  $S$  with  $m + 1$  elements.

Now let's think of a way to make such an  $X \subseteq S$  with  $|X| = m + 1$ . We could begin by selecting a largest element  $k$  for  $X$ . Now, once we have chosen  $k$ , there are  $k$  elements in  $S$  to the left of  $k$ , and we need to choose  $m$  of them to go in  $X$  (so these, along with  $k$ , form the set  $X$ ).

$$S = \{ \underbrace{0, 1, 2, 3, 4, 5, \dots, k-1}_{\text{choose } m \text{ of these } k \text{ numbers for } X}, \quad k, \quad k+1, \quad k+2, \quad k+3, \quad \dots, \quad n \}$$

↑  
 largest  
 number  
 in  $X$

There are  $\binom{k}{m}$  ways to choose these  $m$  numbers, so there are  $\binom{k}{m}$  subsets of  $S$  whose largest element is  $k$ . Notice that we must have  $m \leq k \leq n$ . (The largest element  $k$  of  $X$  cannot be smaller than  $m$  because we need at least  $m$  elements on its left.) Summing over all possible largest values in  $X$ , we see that  $\sum_{k=m}^n \binom{k}{m}$  equals the number of subsets of  $S$  with  $m + 1$  elements.

The previous two paragraphs show that  $\sum_{k=m}^n \binom{k}{m}$  and  $\binom{n+1}{m+1}$  are answers to the same counting question, so they are equal.

11. Show that  $\sum_{k=0}^n 2^k \binom{n}{k} = 3^n$ .

**Solution:** Consider the problem of counting the number of length- $n$  lists made from the symbols  $\{a, b, c\}$ , with repetition allowed. There are  $3^n$  such lists, so the right-hand side counts the number of such lists.

On the other hand, given  $k$  with  $0 \leq k \leq n$ , let's count the lists that have exactly  $k$  entries unequal to  $a$ . There are  $2^k \binom{n}{k}$  such lists. (First choose  $k$  of  $n$  list positions to be filled with  $b$  or  $c$ , in  $\binom{n}{k}$  ways. Then fill these  $k$  positions with  $b$ 's and  $c$ 's in  $2^k$  ways. Fill any remaining positions with  $a$ 's.) As  $k$  could be any number between 0 and  $n$ , the left-hand side of our equation counts up the number of length- $n$  lists made from the symbols  $\{a, b, c\}$ . Thus the right- and left-hand sides count the same thing, so they are equal.

## Chapter 4 Exercises

1. If  $x$  is an even integer, then  $x^2$  is even.

*Proof.* Suppose  $x$  is even. Thus  $x = 2a$  for some  $a \in \mathbb{Z}$ .

$$\text{Consequently } x^2 = (2a)^2 = 4a^2 = 2(2a^2).$$

Therefore  $x^2 = 2b$ , where  $b$  is the integer  $2a^2$ .

Thus  $x^2$  is even by definition of an even number. ■

3. If  $a$  is an odd integer, then  $a^2 + 3a + 5$  is odd.

*Proof.* Suppose  $a$  is odd.

Thus  $a = 2c + 1$  for some integer  $c$ , by definition of an odd number.

$$\begin{aligned} \text{Then } a^2 + 3a + 5 &= (2c + 1)^2 + 3(2c + 1) + 5 = 4c^2 + 4c + 1 + 6c + 3 + 5 = 4c^2 + 10c + 9 \\ &= 4c^2 + 10c + 8 + 1 = 2(2c^2 + 5c + 4) + 1. \end{aligned}$$

This shows  $a^2 + 3a + 5 = 2b + 1$ , where  $b = 2c^2 + 5c + 4 \in \mathbb{Z}$ .

Therefore  $a^2 + 3a + 5$  is odd. ■

5. Suppose  $x, y \in \mathbb{Z}$ . If  $x$  is even, then  $xy$  is even.

*Proof.* Suppose  $x, y \in \mathbb{Z}$  and  $x$  is even.

Then  $x = 2a$  for some integer  $a$ , by definition of an even number.

$$\text{Thus } xy = (2a)(y) = 2(ay).$$

Therefore  $xy = 2b$  where  $b$  is the integer  $ay$ , so  $xy$  is even. ■

7. Suppose  $a, b \in \mathbb{Z}$ . If  $a \mid b$ , then  $a^2 \mid b^2$ .

*Proof.* Suppose  $a \mid b$ .

By definition of divisibility, this means  $b = ac$  for some integer  $c$ .

Squaring both sides of this equation produces  $b^2 = a^2c^2$ .

Then  $b^2 = a^2d$ , where  $d = c^2 \in \mathbb{Z}$ .

By definition of divisibility, this means  $a^2 \mid b^2$ . ■

9. Suppose  $a$  is an integer. If  $7 \mid 4a$ , then  $7 \mid a$ .

*Proof.* Suppose  $7 \mid 4a$ .

By definition of divisibility, this means  $4a = 7c$  for some integer  $c$ .

Since  $4a = 2(2a)$  it follows that  $4a$  is even, and since  $4a = 7c$ , we know  $7c$  is even.

But then  $c$  can't be odd, because that would make  $7c$  odd, not even.

Thus  $c$  is even, so  $c = 2d$  for some integer  $d$ .

Now go back to the equation  $4a = 7c$  and plug in  $c = 2d$ . We get  $4a = 14d$ .

Dividing both sides by 2 gives  $2a = 7d$ .

Now, since  $2a = 7d$ , it follows that  $7d$  is even, and thus  $d$  cannot be odd.

Then  $d$  is even, so  $d = 2e$  for some integer  $e$ .

Plugging  $d = 2e$  back into  $2a = 7d$  gives  $2a = 14e$ .

Dividing both sides of  $2a = 14e$  by 2 produces  $a = 7e$ .

Finally, the equation  $a = 7e$  means that  $7 \mid a$ , by definition of divisibility. ■

- 11.** Suppose  $a, b, c, d \in \mathbb{Z}$ . If  $a | b$  and  $c | d$ , then  $ac | bd$ .

*Proof.* Suppose  $a | b$  and  $c | d$ .

As  $a | b$ , the definition of divisibility means there is an integer  $x$  for which  $b = ax$ . As  $c | d$ , the definition of divisibility means there is an integer  $y$  for which  $d = cy$ . Since  $b = ax$ , we can multiply one side of  $d = cy$  by  $b$  and the other by  $ax$ . This gives  $bd = axcy$ , or  $bd = (ac)(xy)$ .

Since  $xy \in \mathbb{Z}$ , the definition of divisibility applied to  $bd = (ac)(xy)$  gives  $ac | bd$ . ■

- 13.** Suppose  $x, y \in \mathbb{R}$ . If  $x^2 + 5y = y^2 + 5x$ , then  $x = y$  or  $x + y = 5$ .

*Proof.* Suppose  $x^2 + 5y = y^2 + 5x$ .

Then  $x^2 - y^2 = 5x - 5y$ , and factoring gives  $(x - y)(x + y) = 5(x - y)$ .

Now consider two cases.

**Case 1.** If  $x - y \neq 0$  we can divide both sides of  $(x - y)(x + y) = 5(x - y)$  by the non-zero quantity  $x - y$  to get  $x + y = 5$ .

**Case 2.** If  $x - y = 0$ , then  $x = y$ . (By adding  $y$  to both sides.)

Thus  $x = y$  or  $x + y = 5$ . ■

- 15.** If  $n \in \mathbb{Z}$ , then  $n^2 + 3n + 4$  is even.

*Proof.* Suppose  $n \in \mathbb{Z}$ . We consider two cases.

**Case 1.** Suppose  $n$  is even. Then  $n = 2a$  for some  $a \in \mathbb{Z}$ .

Therefore  $n^2 + 3n + 4 = (2a)^2 + 3(2a) + 4 = 4a^2 + 6a + 4 = 2(2a^2 + 3a + 2)$ .

So  $n^2 + 3n + 4 = 2b$  where  $b = 2a^2 + 3a + 2 \in \mathbb{Z}$ , so  $n^2 + 3n + 4$  is even.

**Case 2.** Suppose  $n$  is odd. Then  $n = 2a + 1$  for some  $a \in \mathbb{Z}$ .

Therefore  $n^2 + 3n + 4 = (2a + 1)^2 + 3(2a + 1) + 4 = 4a^2 + 4a + 1 + 6a + 3 + 4 = 4a^2 + 10a + 8 = 2(2a^2 + 5a + 4)$ . So  $n^2 + 3n + 4 = 2b$  where  $b = 2a^2 + 5a + 4 \in \mathbb{Z}$ , so  $n^2 + 3n + 4$  is even.

In either case  $n^2 + 3n + 4$  is even. ■

- 17.** If two integers have opposite parity, then their product is even.

*Proof.* Suppose  $a$  and  $b$  are two integers with opposite parity. Thus one is even and the other is odd. Without loss of generality, suppose  $a$  is even and  $b$  is odd. Therefore there are integers  $c$  and  $d$  for which  $a = 2c$  and  $b = 2d + 1$ . Then the product of  $a$  and  $b$  is  $ab = 2c(2d + 1) = 2(2cd + c)$ . Therefore  $ab = 2k$  where  $k = 2cd + c \in \mathbb{Z}$ . Therefore the product  $ab$  is even. ■

- 19.** Suppose  $a, b, c \in \mathbb{Z}$ . If  $a^2 | b$  and  $b^3 | c$  then  $a^6 | c$ .

*Proof.* Since  $a^2 | b$  we have  $b = ka^2$  for some  $k \in \mathbb{Z}$ . Since  $b^3 | c$  we have  $c = hb^3$  for some  $h \in \mathbb{Z}$ . Thus  $c = h(ka^2)^3 = hk^3a^6$ . Hence  $a^6 | c$ . ■

- 21.** If  $p$  is prime and  $0 < k < p$  then  $p | \binom{p}{k}$ .

*Proof.* From the formula  $\binom{p}{k} = \frac{p!}{(p-k)!k!}$ , we get  $p! = \binom{p}{k}(p-k)!k!$ . Now, since the prime number  $p$  is a factor of  $p!$  on the left, it must also be a factor of  $\binom{p}{k}(p-k)!k!$

on the right. Thus the prime number  $p$  appears in the prime factorization of  $\binom{p}{k}(p-k)!k!$ .

As  $k!$  is a product of numbers smaller than  $p$ , its prime factorization contains no  $p$ 's. Similarly the prime factorization of  $(p-k)!$  contains no  $p$ 's. But we noted that the prime factorization of  $\binom{p}{k}(p-k)!k!$  must contain a  $p$ , so the prime factorization of  $\binom{p}{k}$  contains a  $p$ . Thus  $\binom{p}{k}$  is a multiple of  $p$ , so  $p$  divides  $\binom{p}{k}$ . ■

- 23.** If  $n \in \mathbb{N}$  then  $\binom{2n}{n}$  is even.

*Proof.* By definition,  $\binom{2n}{n}$  is the number of  $n$ -element subsets of a set  $A$  with  $2n$  elements. For each subset  $X \subseteq A$  with  $|X| = n$ , the complement  $\overline{X}$  is a different set, but it also has  $2n - n = n$  elements. Imagine listing out all the  $n$ -elements subset of a set  $A$ . It could be done in such a way that the list has form

$$X_1, \overline{X_1}, X_2, \overline{X_2}, X_3, \overline{X_3}, X_4, \overline{X_4}, X_5, \overline{X_5} \dots$$

This list has an even number of items, for they are grouped in pairs. Thus  $\binom{2n}{n}$  is even. ■

- 25.** If  $a, b, c \in \mathbb{N}$  and  $c \leq b \leq a$  then  $\binom{a}{b}\binom{b}{c} = \binom{a}{b-c}\binom{a-b+c}{c}$ .

*Proof.* Assume  $a, b, c \in \mathbb{N}$  with  $c \leq b \leq a$ . Then we have  $\binom{a}{b}\binom{b}{c} = \frac{a!}{(a-b)!b!} \frac{b!}{(b-c)!c!} = \frac{a!}{(a-b+c)!(a-b)!} \frac{(a-b+c)!}{(b-c)!(a-b+c)!} \frac{(b-c)!c!}{(a-b)!c!} = \binom{a}{b-c}\binom{a-b+c}{c}$ . ■

- 27.** Suppose  $a, b \in \mathbb{N}$ . If  $\gcd(a, b) > 1$ , then  $b \mid a$  or  $b$  is not prime.

*Proof.* Suppose  $\gcd(a, b) > 1$ . Let  $c = \gcd(a, b) > 1$ . Then since  $c$  is a divisor of both  $a$  and  $b$ , we have  $a = cx$  and  $b = cy$  for integers  $x$  and  $y$ . We divide into two cases according to whether or not  $b$  is prime.

**Case I.** Suppose  $b$  is prime. Then the above equation  $b = cy$  with  $c > 1$  forces  $c = b$  and  $y = 1$ . Then  $a = cx$  becomes  $a = bx$ , which means  $b \mid a$ . We conclude that the statement “ $b \mid a$  or  $b$  is not prime,” is true.

**Case II.** Suppose  $b$  is not prime. Then the statement “ $b \mid a$  or  $b$  is not prime,” is automatically true. ■

## Chapter 5 Exercises

1. Suppose  $n \in \mathbb{Z}$ . If  $n^2$  is even, then  $n$  is even.

*Proof.* (Contrapositive) Suppose  $n$  is not even. Then  $n$  is odd, so  $n = 2a + 1$  for some integer  $a$ , by definition of an odd number. Thus  $n^2 = (2a + 1)^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + 1$ . Consequently  $n^2 = 2b + 1$ , where  $b$  is the integer  $2a^2 + 2a$ , so  $n^2$  is odd. Therefore  $n^2$  is not even. ■

3. Suppose  $a, b \in \mathbb{Z}$ . If  $a^2(b^2 - 2b)$  is odd, then  $a$  and  $b$  are odd.

*Proof.* (Contrapositive) Suppose it is not the case that  $a$  and  $b$  are odd. Then, by DeMorgan's law, at least one of  $a$  and  $b$  is even. Let us look at these cases separately.

**Case 1.** Suppose  $a$  is even. Then  $a = 2c$  for some integer  $c$ . Thus  $a^2(b^2 - 2b) = (2c)^2(b^2 - 2b) = 2(2c^2(b^2 - 2b))$ , which is even.

**Case 2.** Suppose  $b$  is even. Then  $b = 2c$  for some integer  $c$ . Thus  $a^2(b^2 - 2b) = a^2((2c)^2 - 2(2c)) = 2(a^2(2c^2 - 2c))$ , which is even.

(A third case involving  $a$  and  $b$  both even is unnecessary, for either of the two cases above cover this case.) Thus in either case  $a^2(b^2 - 2b)$  is even, so it is not odd. ■

5. Suppose  $x \in \mathbb{R}$ . If  $x^2 + 5x < 0$  then  $x < 0$ .

*Proof.* (Contrapositive) Suppose it is not the case that  $x < 0$ , so  $x \geq 0$ . Then neither  $x^2$  nor  $5x$  is negative, so  $x^2 + 5x \geq 0$ . Thus it is not true that  $x^2 + 5x < 0$ . ■

7. Suppose  $a, b \in \mathbb{Z}$ . If both  $ab$  and  $a + b$  are even, then both  $a$  and  $b$  are even.

*Proof.* (Contrapositive) Suppose it is not the case that both  $a$  and  $b$  are even. Then at least one of them is odd. There are three cases to consider.

**Case 1.** Suppose  $a$  is even and  $b$  is odd. Then there are integers  $c$  and  $d$  for which  $a = 2c$  and  $b = 2d + 1$ . Then  $ab = 2c(2d + 1)$ , which is even; and  $a + b = 2c + 2d + 1 = 2(c + d) + 1$ , which is odd. Thus it is not the case that both  $ab$  and  $a + b$  are even.

**Case 2.** Suppose  $a$  is odd and  $b$  is even. Then there are integers  $c$  and  $d$  for which  $a = 2c + 1$  and  $b = 2d$ . Then  $ab = (2c + 1)(2d) = 2(d(2c + 1))$ , which is even; and  $a + b = 2c + 1 + 2d = 2(c + d) + 1$ , which is odd. Thus it is not the case that both  $ab$  and  $a + b$  are even.

**Case 3.** Suppose  $a$  is odd and  $b$  is odd. Then there are integers  $c$  and  $d$  for which  $a = 2c + 1$  and  $b = 2d + 1$ . Then  $ab = (2c + 1)(2d + 1) = 4cd + 2c + 2d + 1 = 2(2cd + c + d) + 1$ , which is odd; and  $a + b = 2c + 1 + 2d + 1 = 2(c + d + 1)$ , which is even. Thus it is not the case that both  $ab$  and  $a + b$  are even.

These cases show that it is not the case that  $ab$  and  $a + b$  are both even. (Note that unlike Exercise 3 above, we really did need all three cases here, for each case involved specific parities for **both**  $a$  and  $b$ .) ■

9. Suppose  $n \in \mathbb{Z}$ . If  $3 \nmid n^2$ , then  $3 \nmid n$ .

*Proof.* (Contrapositive) Suppose it is not the case that  $3 \nmid n$ , so  $3 \mid n$ . This means that  $n = 3a$  for some integer  $a$ . Consequently  $n^2 = 9a^2$ , from which we get  $n^2 = 3(3a^2)$ . This shows that there is an integer  $b = 3a^2$  for which  $n^2 = 3b$ , which means  $3 \mid n^2$ . Therefore it is not the case that  $3 \nmid n^2$ . ■

11. Suppose  $x, y \in \mathbb{Z}$ . If  $x^2(y + 3)$  is even, then  $x$  is even or  $y$  is odd.

*Proof.* (Contrapositive) Suppose it is not the case that  $x$  is even or  $y$  is odd. Using DeMorgan's law, this means  $x$  is not even and  $y$  is not odd, which is to

say  $x$  is odd and  $y$  is even. Thus there are integers  $a$  and  $b$  for which  $x = 2a + 1$  and  $y = 2b$ . Consequently  $x^2(y+3) = (2a+1)^2(2b+3) = (4a^2+4a+1)(2b+3) = 8a^2b+8ab+2b+12a^2+12a+3 = 8a^2b+8ab+2b+12a^2+12a+2+1 = 2(4a^2b+4ab+b+6a^2+6a+1)+1$ . This shows  $x^2(y+3) = 2c+1$  for  $c = 4a^2b+4ab+b+6a^2+6a+1 \in \mathbb{Z}$ . Consequently,  $x^2(y+3)$  is not even. ■

13. Suppose  $x \in \mathbb{R}$ . If  $x^5 + 7x^3 + 5x \geq x^4 + x^2 + 8$ , then  $x \geq 0$ .

*Proof.* (Contrapositive) Suppose it is not true that  $x \geq 0$ . Then  $x < 0$ , that is  $x$  is negative. Consequently, the expressions  $x^5$ ,  $7x^3$  and  $5x$  are all negative (note the odd powers) so  $x^5 + 7x^3 + 5x < 0$ . Similarly the terms  $x^4$ ,  $x^2$  and 8 are all positive (note the even powers), so  $0 < x^4 + x^2 + 8$ . From this we get  $x^5 + 7x^3 + 5x < x^4 + x^2 + 8$ , so it is not true that  $x^5 + 7x^3 + 5x \geq x^4 + x^2 + 8$ . ■

15. Suppose  $x \in \mathbb{Z}$ . If  $x^3 - 1$  is even, then  $x$  is odd.

*Proof.* (Contrapositive) Suppose  $x$  is not odd. Thus  $x$  is even, so  $x = 2a$  for some integer  $a$ . Then  $x^3 - 1 = (2a)^3 - 1 = 8a^3 - 1 = 8a^3 - 2 + 1 = 2(4a^3 - 1) + 1$ . Therefore  $x^3 - 1 = 2b + 1$  where  $b = 4a^3 - 1 \in \mathbb{Z}$ , so  $x^3 - 1$  is odd. Thus  $x^3 - 1$  is not even. ■

17. If  $n$  is odd, then  $8 \mid (n^2 - 1)$ .

*Proof.* (Direct) Suppose  $n$  is odd, so  $n = 2a + 1$  for some integer  $a$ . Then  $n^2 - 1 = (2a+1)^2 - 1 = 4a^2 + 4a = 4(a^2 + a) = 4a(a+1)$ . So far we have  $n^2 - 1 = 4a(a+1)$ , but we want a factor of 8, not 4. But notice that one of  $a$  or  $a+1$  must be even, so  $a(a+1)$  is even and hence  $a(a+1) = 2c$  for some integer  $c$ . Now we have  $n^2 - 1 = 4a(a+1) = 4(2c) = 8c$ . But  $n^2 - 1 = 8c$  means  $8 \mid (n^2 - 1)$ . ■

19. Let  $a, b, c \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$  and  $a \equiv c \pmod{n}$ , then  $c \equiv b \pmod{n}$ .

*Proof.* (Direct) Suppose  $a \equiv b \pmod{n}$  and  $a \equiv c \pmod{n}$ .

This means  $n \mid (a - b)$  and  $n \mid (a - c)$ .

Thus there are integers  $d$  and  $e$  for which  $a - b = nd$  and  $a - c = ne$ .

Subtracting the second equation from the first gives  $c - b = nd - ne$ .

Thus  $c - b = n(d - e)$ , so  $n \mid (c - b)$  by definition of divisibility.

Therefore  $c \equiv b \pmod{n}$  by definition of congruence modulo  $n$ . ■

21. Let  $a, b \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$ , then  $a^3 \equiv b^3 \pmod{n}$ .

*Proof.* (Direct) Suppose  $a \equiv b \pmod{n}$ . This means  $n \mid (a - b)$ , so there is an integer  $c$  for which  $a - b = nc$ . Then:

$$\begin{aligned} a - b &= nc \\ (a - b)(a^2 + ab + b^2) &= nc(a^2 + ab + b^2) \\ a^3 + a^2b + ab^2 - ba^2 - ab^2 - b^3 &= nc(a^2 + ab + b^2) \\ a^3 - b^3 &= nc(a^2 + ab + b^2). \end{aligned}$$

Since  $a^2 + ab + b^2 \in \mathbb{Z}$ , the equation  $a^3 - b^3 = nc(a^2 + ab + b^2)$  implies  $n \mid (a^3 - b^3)$ , and therefore  $a^3 \equiv b^3 \pmod{n}$ . ■

- 23.** Let  $a, b \in \mathbb{Z}$  and  $n \in \mathbb{N}$ . If  $a \equiv b \pmod{n}$ , then  $a^2 \equiv ab \pmod{n}$ .

*Proof.* (Direct) Suppose  $a \equiv b \pmod{n}$ . This means  $n \mid (a - b)$ , so there is an integer  $d$  for which  $a - b = nd$ . Multiply both sides of this by  $a$  to get  $a^2 - ab = and$ . Consequently, there is an integer  $e = da$  for which  $a^2 - ab = ne$ , so  $n \mid (a^2 - ab)$  and consequently  $a^2 \equiv ab \pmod{n}$ . ■

- 25.** If  $n \in \mathbb{N}$  and  $2^n - 1$  is prime, then  $n$  is prime.

*Proof.* Assume  $n$  is not prime. Write  $n = ab$  for some  $a, b > 1$ . Then  $2^n - 1 = 2^{ab} - 1 = (2^b - 1)(2^{ab-b} + 2^{ab-2b} + 2^{ab-3b} + \dots + 2^{ab-ab})$ . Hence  $2^n - 1$  is composite. ■

- 27.** If  $a \equiv 0 \pmod{4}$  or  $a \equiv 1 \pmod{4}$  then  $\binom{a}{2}$  is even.

*Proof.* We prove this directly. Assume  $a \equiv 0 \pmod{4}$ . Then  $\binom{a}{2} = \frac{a(a-1)}{2}$ . Since  $a = 4k$  for some  $k \in \mathbb{N}$ , we have  $\binom{a}{2} = \frac{4k(4k-1)}{2} = 2k(4k-1)$ . Hence  $\binom{a}{2}$  is even.

Now assume  $a \equiv 1 \pmod{4}$ . Then  $a = 4k+1$  for some  $k \in \mathbb{N}$ . Hence  $\binom{a}{2} = \frac{(4k+1)(4k)}{2} = 2k(4k+1)$ . Hence,  $\binom{a}{2}$  is even. This proves the result. ■

- 29.** If integers  $a$  and  $b$  are not both zero, then  $\gcd(a, b) = \gcd(a - b, b)$ .

*Proof.* (Direct) Suppose integers  $a$  and  $b$  are not both zero. Let  $d = \gcd(a, b)$ . Because  $d$  is a divisor of both  $a$  and  $b$ , we have  $a = dx$  and  $b = dy$  for some integers  $x$  and  $y$ . Then  $a - b = dx - dy = d(x - y)$ , so it follows that  $d$  is also a common divisor of  $a - b$  and  $b$ . Therefore it can't be greater than the greatest common divisor of  $a - b$  and  $b$ , which is to say  $\gcd(a, b) = d \leq \gcd(a - b, b)$ .

Now let  $e = \gcd(a - b, b)$ . Then  $e$  divides both  $a - b$  and  $b$ , that is,  $a - b = ex$  and  $b = ey$  for integers  $x$  and  $y$ . Then  $a = (a - b) + b = ex + ey = e(x + y)$ , so now we see that  $e$  is a divisor of both  $a$  and  $b$ . Thus it is not more than their greatest common divisor, that is,  $\gcd(a - b, b) = e \leq \gcd(a, b)$ .

The above two paragraphs have given  $\gcd(a, b) \leq \gcd(a - b, b)$  and  $\gcd(a - b, b) \leq \gcd(a, b)$ . Thus  $\gcd(a, b) = \gcd(a - b, b)$ . ■

- 31.** Suppose the division algorithm applied to  $a$  and  $b$  yields  $a = qb + r$ . Then  $\gcd(a, b) = \gcd(r, b)$ .

*Proof.* Suppose  $a = qb + r$ . Let  $d = \gcd(a, b)$ , so  $d$  is a common divisor of  $a$  and  $b$ ; thus  $a = dx$  and  $b = dy$  for some integers  $x$  and  $y$ . Then  $dx = a = qb + r = qdy + r$ , hence  $dx = qdy + r$ , and so  $r = dx - qdy = d(x - qy)$ . Thus  $d$  is a divisor of  $r$  (and also of  $b$ ), so  $\gcd(a, b) = d \leq \gcd(r, b)$ .

On the other hand, let  $e = \gcd(r, b)$ , so  $r = ex$  and  $b = ey$  for some integers  $x$  and  $y$ . Then  $a = qb + r = qey + ex = e(qy + x)$ . Hence  $e$  is a divisor of  $a$  (and of course also of  $b$ ) so  $\gcd(r, b) = e \leq \gcd(a, b)$ .

We've shown  $\gcd(a, b) \leq \gcd(r, b)$  and  $\gcd(r, b) \leq \gcd(a, b)$ , so  $\gcd(r, b) = \gcd(a, b)$ . ■

## Chapter 6 Exercises

1. Suppose  $n$  is an integer. If  $n$  is odd, then  $n^2$  is odd.

*Proof.* Suppose for the sake of contradiction that  $n$  is odd and  $n^2$  is not odd. Then  $n^2$  is even. Now, since  $n$  is odd, we have  $n = 2a + 1$  for some integer  $a$ . Thus  $n^2 = (2a + 1)^2 = 4a^2 + 4a + 1 = 2(2a^2 + 2a) + 1$ . This shows  $n^2 = 2b + 1$ , where  $b$  is the integer  $b = 2a^2 + 2a$ . Therefore we have  $n^2$  is odd and  $n^2$  is even, a contradiction. ■

3. Prove that  $\sqrt[3]{2}$  is irrational.

*Proof.* Suppose for the sake of contradiction that  $\sqrt[3]{2}$  is not irrational. Therefore it is rational, so there exist integers  $a$  and  $b$  for which  $\sqrt[3]{2} = \frac{a}{b}$ . Let us assume that this fraction is reduced, so  $a$  and  $b$  are not both even. Now we have  $\sqrt[3]{2}^3 = \left(\frac{a}{b}\right)^3$ , which gives  $2 = \frac{a^3}{b^3}$ , or  $2b^3 = a^3$ . From this we see that  $a^3$  is even, from which we deduce that  $a$  is even. (For if  $a$  were odd, then  $a^3 = (2c + 1)^3 = 8c^3 + 12c^2 + 6c + 1 = 2(4c^3 + 6c^2 + 3c) + 1$  would be odd, not even.) Since  $a$  is even, it follows that  $a = 2d$  for some integer  $d$ . The equation  $2b^3 = a^3$  from above then becomes  $2b^3 = (2d)^3$ , or  $2b^3 = 8d^3$ . Dividing by 2, we get  $b^3 = 4d^3$ , and it follows that  $b^3$  is even. Thus  $b$  is even also. (Using the same argument we used when  $a^3$  was even.) At this point we have discovered that both  $a$  and  $b$  are even, contradicting the fact (observed above) that the  $a$  and  $b$  are not both even. ■

Here is an alternative proof.

*Proof.* Suppose for the sake of contradiction that  $\sqrt[3]{2}$  is not irrational. Therefore there exist integers  $a$  and  $b$  for which  $\sqrt[3]{2} = \frac{a}{b}$ . Cubing both sides, we get  $2 = \frac{a^3}{b^3}$ . From this,  $a^3 = b^3 + b^3$ , which contradicts Fermat's last theorem. ■

5. Prove that  $\sqrt{3}$  is irrational.

*Proof.* Suppose for the sake of contradiction that  $\sqrt{3}$  is not irrational. Therefore it is rational, so there exist integers  $a$  and  $b$  for which  $\sqrt{3} = \frac{a}{b}$ . Let us assume that this fraction is reduced, so  $a$  and  $b$  have no common factor. Notice that  $\sqrt{3}^2 = \left(\frac{a}{b}\right)^2$ , so  $3 = \frac{a^2}{b^2}$ , or  $3b^2 = a^2$ . This means  $3 \mid a^2$ .

Now we are going to show that if  $a \in \mathbb{Z}$  and  $3 \mid a^2$ , then  $3 \mid a$ . (This is a proof-within-a-proof.) We will use contrapositive proof to prove this conditional statement. Suppose  $3 \nmid a$ . Then there is a remainder of either 1 or 2 when 3 is divided into  $a$ .

**Case 1.** There is a remainder of 1 when 3 is divided into  $a$ . Then  $a = 3m + 1$  for some integer  $m$ . Consequently,  $a^2 = 9m^2 + 6m + 1 = 3(3m^2 + 2m) + 1$ , and this means 3 divides into  $a^2$  with a remainder of 1. Thus  $3 \nmid a^2$ .

**Case 2.** There is a remainder of 2 when 3 is divided into  $a$ . Then  $a = 3m + 2$  for some integer  $m$ . Consequently,  $a^2 = 9m^2 + 12m + 4 = 9m^2 + 12m + 3 + 1 = 3(3m^2 + 4m + 1) + 1$ , and this means 3 divides into  $a^2$  with a remainder of 1. Thus  $3 \nmid a^2$ . In either case we have  $3 \nmid a^2$ , so we've shown  $3 \nmid a$  implies  $3 \nmid a^2$ . Therefore, if  $3 \mid a^2$ , then  $3 \mid a$ .

Now go back to  $3 \mid a^2$  in the first paragraph. This combined with the result of the second paragraph implies  $3 \mid a$ , so  $a = 3d$  for some integer  $d$ . Now also in the first paragraph we had  $3b^2 = a^2$ , which now becomes  $3b^2 = (3d)^2$  or  $3b^2 = 9d^2$ , so  $b^2 = 3d^2$ . But this means  $3 \mid b^2$ , and the second paragraph implies  $3 \mid b$ . Thus we have concluded that  $3 \mid a$  and  $3 \mid b$ , but this contradicts the fact that the fraction  $\frac{a}{b}$  is reduced. ■

7. If  $a, b \in \mathbb{Z}$ , then  $a^2 - 4b - 3 \neq 0$ .

*Proof.* Suppose for the sake of contradiction that  $a, b \in \mathbb{Z}$  but  $a^2 - 4b - 3 = 0$ . Then we have  $a^2 = 4b + 3 = 2(2b + 1) + 1$ , which means  $a^2$  is odd. Therefore  $a$  is odd also, so  $a = 2c + 1$  for some integer  $c$ . Plugging this back into  $a^2 - 4b - 3 = 0$  gives us

$$\begin{aligned} (2c+1)^2 - 4b - 3 &= 0 \\ 4c^2 + 4c + 1 - 4b - 3 &= 0 \\ 4c^2 + 4c - 4b &= 2 \\ 2c^2 + 2c - 2b &= 1 \\ 2(c^2 + c - b) &= 1. \end{aligned}$$

From this last equation, we see that 1 is an even number, a contradiction. ■

9. Suppose  $a, b \in \mathbb{R}$  and  $a \neq 0$ . If  $a$  is rational and  $ab$  is irrational, then  $b$  is irrational.

*Proof.* Suppose for the sake of contradiction that  $a$  is rational and  $ab$  is irrational and  $b$  is **not** irrational. Thus we have  $a$  and  $b$  rational, and  $ab$  irrational. Since  $a$  and  $b$  are rational, we know there are integers  $c, d, e, f$  for which  $a = \frac{e}{d}$  and  $b = \frac{f}{e}$ . Then  $ab = \frac{ce}{df}$ , and since both  $ce$  and  $df$  are integers, it follows that  $ab$  is rational. But this is a contradiction because we started out with  $ab$  irrational. ■

11. There exist no integers  $a$  and  $b$  for which  $18a + 6b = 1$ .

*Proof.* Suppose for the sake of contradiction that there do exist integers  $a$  and  $b$  with  $18a + 6b = 1$ . Then  $1 = 2(9a + 3b)$ , which means 1 is even, a contradiction. ■

13. For every  $x \in [\pi/2, \pi]$ ,  $\sin x - \cos x \geq 1$ .

*Proof.* Suppose for the sake of contradiction that  $x \in [\pi/2, \pi]$ , but  $\sin x - \cos x < 1$ . Since  $x \in [\pi/2, \pi]$ , we know  $\sin x \geq 0$  and  $\cos x \leq 0$ , so  $\sin x - \cos x \geq 0$ . Therefore we have  $0 \leq \sin x - \cos x < 1$ . Now the square of any number between 0 and 1 is still a number between 0 and 1, so we have  $0 \leq (\sin x - \cos x)^2 < 1$ , or  $0 \leq \sin^2 x - 2\sin x \cos x + \cos^2 x < 1$ . Using the fact that  $\sin^2 x + \cos^2 x = 1$ , this becomes  $0 \leq -2\sin x \cos x + 1 < 1$ . Subtracting 1, we obtain  $-2\sin x \cos x < 0$ . But above we remarked that  $\sin x \geq 0$  and  $\cos x \leq 0$ , and hence  $-2\sin x \cos x \geq 0$ . We now have the contradiction  $-2\sin x \cos x < 0$  and  $-2\sin x \cos x \geq 0$ . ■

15. If  $b \in \mathbb{Z}$  and  $b \nmid k$  for every  $k \in \mathbb{N}$ , then  $b = 0$ .

*Proof.* Suppose for the sake of contradiction that  $b \in \mathbb{Z}$  and  $b \nmid k$  for every  $k \in \mathbb{N}$ , but  $b \neq 0$ .

Case 1. Suppose  $b > 0$ . Then  $b \in \mathbb{N}$ , so  $b|b$ , contradicting  $b \nmid k$  for every  $k \in \mathbb{N}$ .

Case 2. Suppose  $b < 0$ . Then  $-b \in \mathbb{N}$ , so  $b|(-b)$ , again a contradiction ■

17. For every  $n \in \mathbb{Z}$ ,  $4 \nmid (n^2 + 2)$ .

*Proof.* Assume there exists  $n \in \mathbb{Z}$  with  $4|(n^2 + 2)$ . Then for some  $k \in \mathbb{Z}$ ,  $4k = n^2 + 2$  or  $2k = n^2 + 2(1 - k)$ . If  $n$  is odd, this means  $2k$  is odd, and we've reached a contradiction. If  $n$  is even then  $n = 2j$  and we get  $k = 2j^2 + 1 - k$  for some  $j \in \mathbb{Z}$ . Hence  $2(k - j^2) = 1$ , so 1 is even, a contradiction. ■

*Remark.* It is fairly easy to see that two more than a perfect square is always either 2 (mod 4) or 3 (mod 4). This would end the proof immediately.

19. The product of 5 consecutive integers is a multiple of 120.

*Proof.* Starting from 0, every fifth integer is a multiple of 5, every fourth integer is a multiple of 4, every third integer is a multiple of 3, and every other integer is a multiple of 2. It follows that any set of 5 consecutive integers must contain a multiple of 5, a multiple of 4, at least one multiple of 3, and at least two multiples of 2 (possibly one of which is a multiple of 4). It follows that the product of five consecutive integers is a multiple of  $5 \cdot 4 \cdot 3 \cdot 2 = 120$ . ■

For another approach, consider a product  $n(n-1)(n-2)(n-3)(n-4)$  of five consecutive integers (the largest of which is  $n$ ). Now, we know that  $\binom{n}{5}$  is an integer, and  $\binom{n}{5} = \frac{n!}{5!(n-5)!} = \frac{n!}{120(n-5)!} = \frac{n(n-1)(n-2)(n-3)(n-4)}{120}$ , so 120 divides the product.

21. Hints for Exercises 20–23. For Exercises 20, first show that the equation  $a^2 + b^2 = 3c^2$  has no solutions (other than the trivial solution  $(a, b, c) = (0, 0, 0)$ ) in the integers. To do this, investigate the remainders of a sum of squares (mod 4). After you've done this, prove that the only solution is indeed the trivial solution. Next assume that the equation  $x^2 + y^2 - 3 = 0$  has a rational solution. Use the definition of rational numbers to yield a contradiction.

## Chapter 7 Exercises

1. Suppose  $x \in \mathbb{Z}$ . Then  $x$  is even if and only if  $3x + 5$  is odd.

*Proof.* We first use direct proof to show that if  $x$  is even, then  $3x + 5$  is odd. If  $x$  is even, then  $x = 2n$  for some integer  $n$ , so  $3x + 5 = 3(2n) + 5 = 6n + 5 = 6n + 4 + 1 = 2(3n + 2) + 1$ . Thus  $3x + 5$  is odd because it has form  $2k + 1$ , where  $k = 3n + 2 \in \mathbb{Z}$ .

Conversely, we need to show that if  $3x + 5$  is odd, then  $x$  is even. We will prove this using contrapositive proof. Suppose  $x$  is *not* even. Then  $x$  is odd, so  $x = 2n + 1$  for some integer  $n$ . Thus  $3x + 5 = 3(2n + 1) + 5 = 6n + 8 = 2(3n + 4)$ . This means says  $3x + 5$  is twice the integer  $3n + 4$ , so  $3x + 5$  is even, not odd. ■

3. Given an integer  $a$ , then  $a^3 + a^2 + a$  is even if and only if  $a$  is even.

*Proof.* First we will prove that if  $a^3 + a^2 + a$  is even then  $a$  is even. This is done with contrapositive proof. Suppose  $a$  is not even. Then  $a$  is odd, so there is an integer  $n$  for which  $a = 2n + 1$ . Then

$$\begin{aligned} a^3 + a^2 + a &= (2n+1)^3 + (2n+1)^2 + (2n+1) \\ &= 8n^3 + 12n^2 + 6n + 1 + 4n^2 + 4n + 1 + 2n + 1 \\ &= 8n^3 + 16n^2 + 12n + 2 + 1 \\ &= 2(4n^3 + 8n^2 + 6n + 1) + 1. \end{aligned}$$

This expresses  $a^3 + a^2 + a$  as twice an integer plus 1, so  $a^3 + a^2 + a$  is odd, not even. We have now shown that if  $a^3 + a^2 + a$  is even then  $a$  is even.

Conversely, we need to show that if  $a$  is even, then  $a^3 + a^2 + a$  is even. We will use direct proof. Suppose  $a$  is even, so  $a = 2n$  for some integer  $n$ . Then  $a^3 + a^2 + a = (2n)^3 + (2n)^2 + 2n = 8n^3 + 4n^2 + 2n = 2(4n^3 + 2n^2 + n)$ . Therefore,  $a^3 + a^2 + a$  is even because it's twice an integer. ■

5. An integer  $a$  is odd if and only if  $a^3$  is odd.

*Proof.* Suppose that  $a$  is odd. Then  $a = 2n + 1$  for some integer  $n$ , and  $a^3 = (2n+1)^3 = 8n^3 + 12n^2 + 6n + 1 = 2(4n^3 + 6n^2 + 3n) + 1$ . This shows that  $a^3$  is twice an integer, plus 1, so  $a^3$  is odd. Thus we've proved that if  $a$  is odd then  $a^3$  is odd.

Conversely we need to show that if  $a^3$  is odd, then  $a$  is odd. For this we employ contrapositive proof. Suppose  $a$  is not odd. Thus  $a$  is even, so  $a = 2n$  for some integer  $n$ . Then  $a^3 = (2n)^3 = 8n^3 = 2(4n^3)$  is even (not odd). ■

7. Suppose  $x, y \in \mathbb{R}$ . Then  $(x+y)^2 = x^2 + y^2$  if and only if  $x = 0$  or  $y = 0$ .

*Proof.* First we prove with direct proof that if  $(x+y)^2 = x^2 + y^2$ , then  $x = 0$  or  $y = 0$ . Suppose  $(x+y)^2 = x^2 + y^2$ . From this we get  $x^2 + 2xy + y^2 = x^2 + y^2$ , so  $2xy = 0$ , and hence  $xy = 0$ . Thus  $x = 0$  or  $y = 0$ .

Conversely, we need to show that if  $x = 0$  or  $y = 0$ , then  $(x+y)^2 = x^2 + y^2$ . This will be done with cases.

**Case 1.** If  $x = 0$  then  $(x+y)^2 = (0+y)^2 = y^2 = 0^2 + y^2 = x^2 + y^2$ .

**Case 2.** If  $y = 0$  then  $(x+y)^2 = (x+0)^2 = x^2 = x^2 + 0^2 = x^2 + y^2$ .

Either way, we have  $(x+y)^2 = x^2 + y^2$ . ■

9. Suppose  $a \in \mathbb{Z}$ . Prove that  $14 | a$  if and only if  $7 | a$  and  $2 | a$ .

*Proof.* First we prove that if  $14 | a$ , then  $7 | a$  and  $2 | a$ . Direct proof is used. Suppose  $14 | a$ . This means  $a = 14m$  for some integer  $m$ . Therefore  $a = 7(2m)$ , which means  $7 | a$ , and also  $a = 2(7m)$ , which means  $2 | a$ . Thus  $7 | a$  and  $2 | a$ .

Conversely, we need to prove that if  $7 | a$  and  $2 | a$ , then  $14 | a$ . Once again direct proof is used. Suppose  $7 | a$  and  $2 | a$ . Since  $2 | a$  it follows that  $a = 2m$  for some

integer  $m$ , and that in turn implies that  $a$  is even. Since  $7 \mid a$  it follows that  $a = 7n$  for some integer  $n$ . Now, since  $a$  is known to be even, and  $a = 7n$ , it follows that  $n$  is even (if it were odd, then  $a = 7n$  would be odd). Thus  $n = 2p$  for an appropriate integer  $p$ , and plugging  $n = 2p$  back into  $a = 7n$  gives  $a = 7(2p)$ , so  $a = 14p$ . Therefore  $14 \mid a$ . ■

11. Suppose  $a, b \in \mathbb{Z}$ . Prove that  $(a - 3)b^2$  is even if and only if  $a$  is odd or  $b$  is even.

*Proof.* First we will prove that if  $(a - 3)b^2$  is even, then  $a$  is odd or  $b$  is even. For this we use contrapositive proof. Suppose it is not the case that  $a$  is odd or  $b$  is even. Then by DeMorgan's law,  $a$  is even and  $b$  is odd. Thus there are integers  $m$  and  $n$  for which  $a = 2m$  and  $b = 2n + 1$ . Now observe  $(a - 3)b^2 = (2m - 3)(2n + 1)^2 = (2m - 3)(4n^2 + 4n + 1) = 8mn^2 + 8mn + 2m - 12n^2 - 12n - 3 = 8mn^2 + 8mn + 2m - 12n^2 - 12n - 4 + 1 = 2(4mn^2 + 4mn + m - 6n^2 - 6n - 2) + 1$ . This shows  $(a - 3)b^2$  is odd, so it's not even.

Conversely, we need to show that if  $a$  is odd or  $b$  is even, then  $(a - 3)b^2$  is even. For this we use direct proof, with cases.

**Case 1.** Suppose  $a$  is odd. Then  $a = 2m + 1$  for some integer  $m$ . Thus  $(a - 3)b^2 = (2m + 1 - 3)b^2 = (2m - 2)b^2 = 2(m - 1)b^2$ . Thus in this case  $(a - 3)b^2$  is even.

**Case 2.** Suppose  $b$  is even. Then  $b = 2n$  for some integer  $n$ . Thus  $(a - 3)b^2 = (a - 3)(2n)^2 = (a - 3)4n^2 = 2(a - 3)2n^2$ . Thus in this case  $(a - 3)b^2$  is even.

Therefore, in any event,  $(a - 3)b^2$  is even. ■

13. Suppose  $a, b \in \mathbb{Z}$ . If  $a + b$  is odd, then  $a^2 + b^2$  is odd.

Hint: Use direct proof. Suppose  $a + b$  is odd. Argue that this means  $a$  and  $b$  have opposite parity. Then use cases.

15. Suppose  $a, b \in \mathbb{Z}$ . Prove that  $a + b$  is even if and only if  $a$  and  $b$  have the same parity.

*Proof.* First we will show that if  $a + b$  is even, then  $a$  and  $b$  have the same parity. For this we use contrapositive proof. Suppose it is not the case that  $a$  and  $b$  have the same parity. Then one of  $a$  and  $b$  is even and the other is odd. Without loss of generality, let's say that  $a$  is even and  $b$  is odd. Thus there are integers  $m$  and  $n$  for which  $a = 2m$  and  $b = 2n + 1$ . Then  $a + b = 2m + 2n + 1 = 2(m + n) + 1$ , so  $a + b$  is odd, not even.

Conversely, we need to show that if  $a$  and  $b$  have the same parity, then  $a + b$  is even. For this, we use direct proof with cases. Suppose  $a$  and  $b$  have the same parity.

**Case 1.** Both  $a$  and  $b$  are even. Then there are integers  $m$  and  $n$  for which  $a = 2m$  and  $b = 2n$ , so  $a + b = 2m + 2n = 2(m + n)$  is clearly even.

**Case 2.** Both  $a$  and  $b$  are odd. Then there are integers  $m$  and  $n$  for which  $a = 2m + 1$  and  $b = 2n + 1$ , so  $a + b = 2m + 1 + 2n + 1 = 2(m + n + 1)$  is clearly even.

Either way,  $a + b$  is even. This completes the proof. ■

- 17.** There is a prime number between 90 and 100.

*Proof.* Simply observe that 97 is prime. ■

- 19.** If  $n \in \mathbb{N}$ , then  $2^0 + 2^1 + 2^2 + 2^3 + 2^4 + \cdots + 2^n = 2^{n+1} - 1$ .

*Proof.* We use direct proof. Suppose  $n \in \mathbb{N}$ . Let  $S$  be the number

$$S = 2^0 + 2^1 + 2^2 + 2^3 + 2^4 + \cdots + 2^{n-1} + 2^n. \quad (1)$$

In what follows, we will solve for  $S$  and show  $S = 2^{n+1} - 1$ . Multiplying both sides of (1) by 2 gives

$$2S = 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + \cdots + 2^n + 2^{n+1}. \quad (2)$$

Now subtract Equation (1) from Equation (2) to obtain  $2S - S = -2^0 + 2^{n+1}$ , which simplifies to  $S = 2^{n+1} - 1$ . Combining this with Equation (1) produces  $2^0 + 2^1 + 2^2 + 2^3 + 2^4 + \cdots + 2^n = 2^{n+1} - 1$ , so the proof is complete. ■

- 21.** Every real solution of  $x^3 + x + 3 = 0$  is irrational.

*Proof.* Suppose for the sake of contradiction that this polynomial has a rational solution  $\frac{a}{b}$ . We may assume that this fraction is fully reduced, so  $a$  and  $b$  are not both even. We have  $\left(\frac{a}{b}\right)^3 + \frac{a}{b} + 3 = 0$ . Clearing the denominator gives

$$a^3 + ab^2 + 3b^3 = 0.$$

Consider two cases: First, if both  $a$  and  $b$  are odd, the left-hand side is a sum of three odds, which is odd, meaning 0 is odd, a contradiction. Second, if one of  $a$  and  $b$  is odd and the other is even, then the middle term of  $a^3 + ab^2 + 3b^3$  is even, while  $a^3$  and  $3b^3$  have opposite parity. Then  $a^3 + ab^2 + 3b^3$  is the sum of two evens and an odd, which is odd, again contradicting the fact that 0 is even. ■

- 23.** Suppose  $a, b$  and  $c$  are integers. If  $a \mid b$  and  $a \mid (b^2 - c)$ , then  $a \mid c$ .

*Proof.* (Direct) Suppose  $a \mid b$  and  $a \mid (b^2 - c)$ . This means that  $b = ad$  and  $b^2 - c = ae$  for some integers  $d$  and  $e$ . Squaring the first equation produces  $b^2 = a^2d^2$ . Subtracting  $b^2 - c = ae$  from  $b^2 = a^2d^2$  gives  $c = a^2d^2 - ae = a(ad^2 - e)$ . As  $ad^2 - e \in \mathbb{Z}$ , it follows that  $a \mid c$ . ■

- 25.** If  $p > 1$  is an integer and  $n \nmid p$  for each integer  $n$  for which  $2 \leq n \leq \sqrt{p}$ , then  $p$  is prime.

*Proof.* (Contrapositive) Suppose that  $p$  is not prime, so it factors as  $p = mn$  for  $1 < m, n < p$ .

Observe that it is not the case that both  $m > \sqrt{p}$  and  $n > \sqrt{p}$ , because if this were true the inequalities would multiply to give  $mn > \sqrt{p}\sqrt{p} = p$ , which contradicts  $p = mn$ .

Therefore  $m \leq \sqrt{p}$  or  $n \leq \sqrt{p}$ . Without loss of generality, say  $n \leq \sqrt{p}$ . Then the equation  $p = mn$  gives  $n \mid p$ , with  $1 < n \leq \sqrt{p}$ . Therefore it is not true that  $n \nmid p$  for each integer  $n$  for which  $2 \leq n \leq \sqrt{p}$ . ■

27. Suppose  $a, b \in \mathbb{Z}$ . If  $a^2 + b^2$  is a perfect square, then  $a$  and  $b$  are not both odd.

*Proof.* (Contradiction) Suppose  $a^2 + b^2$  is a perfect square, and  $a$  and  $b$  are both odd. As  $a^2 + b^2$  is a perfect square, say  $c$  is the integer for which  $c^2 = a^2 + b^2$ . As  $a$  and  $b$  are odd, we have  $a = 2m + 1$  and  $b = 2n + 1$  for integers  $m$  and  $n$ . Then

$$c^2 = a^2 + b^2 = (2m+1)^2 + (2n+1)^2 = 4(m^2 + n^2 + m + n) + 2.$$

This is even, so  $c$  is even also; let  $c = 2k$ . Now the above equation results in  $(2k)^2 = 4(m^2 + n^2 + m + n) + 2$ , which simplifies to  $2k^2 = 2(m^2 + n^2 + m + n) + 1$ . Thus  $2k^2$  is both even and odd, a contradiction. ■

29. If  $a | bc$  and  $\gcd(a, b) = 1$ , then  $a | c$ .

*Proof.* (Direct) Suppose  $a | bc$  and  $\gcd(a, b) = 1$ . The fact that  $a | bc$  means  $bc = az$  for some integer  $z$ . The fact that  $\gcd(a, b) = 1$  means that  $ax + by = 1$  for some integers  $x$  and  $y$  (by Proposition 7.1 on page 152). From this we get  $acx + bcy = c$ ; substituting  $bc = az$  yields  $acx + azy = c$ , that is,  $a(cx + zy) = c$ . Therefore  $a | c$ . ■

31. If  $n \in \mathbb{Z}$ , then  $\gcd(n, n + 1) = 1$ .

*Proof.* Suppose  $d$  is a positive integer that is a common divisor of  $n$  and  $n + 1$ . Then  $n = dx$  and  $n + 1 = dy$  for integers  $x$  and  $y$ . Then  $1 = (n + 1) - n = dy - dx = d(y - x)$ . Now,  $1 = d(y - x)$  is only possible if  $d = \pm 1$  and  $y - x = \pm 1$ . Thus the greatest common divisor of  $n$  and  $n + 1$  can be no greater than 1. But 1 does divide both  $n$  and  $n + 1$ , so  $\gcd(n, n + 1) = 1$ . ■

33. If  $n \in \mathbb{Z}$ , then  $\gcd(2n + 1, 4n^2 + 1) = 1$ .

*Proof.* Note that  $4n^2 + 1 = (2n + 1)(2n - 1) + 2$ . Therefore, it suffices to show that  $\gcd(2n + 1, (2n + 1)(2n - 1) + 2) = 1$ . Let  $d$  be a common positive divisor of both  $2n + 1$  and  $(2n + 1)(2n - 1) + 2$ , so  $2n + 1 = dx$  and  $(2n + 1)(2n - 1) + 2 = dy$  for integers  $x$  and  $y$ . Substituting the first equation into the second gives  $dx(2n - 1) + 2 = dy$ , so  $2 = dy - dx(2n - 1) = d(y - 2nx + x)$ . This means  $d$  divides 2, so  $d$  equals 1 or 2. But the equation  $2n + 1 = dx$  means  $d$  must be odd. Therefore  $d = 1$ , that is,  $\gcd(2n + 1, (2n + 1)(2n - 1) + 2) = 1$ . ■

35. Suppose  $a, b \in \mathbb{N}$ . Then  $a = \gcd(a, b)$  if and only if  $a | b$ .

*Proof.* Suppose  $a = \gcd(a, b)$ . This means  $a$  is a divisor of both  $a$  and  $b$ . In particular  $a | b$ .

Conversely, suppose  $a | b$ . Then  $a$  divides both  $a$  and  $b$ , so  $a \leq \gcd(a, b)$ . On the other hand, since  $\gcd(a, b)$  divides  $a$ , we have  $a = \gcd(a, b) \cdot x$  for some integer  $x$ . As all integers involved are positive, it follows that  $a \geq \gcd(a, b)$ .

It has been established that  $a \leq \gcd(a, b)$  and  $a \geq \gcd(a, b)$ . Thus  $a = \gcd(a, b)$ . ■

## Chapter 8 Exercises

1. Prove that  $\{12n : n \in \mathbb{Z}\} \subseteq \{2n : n \in \mathbb{Z}\} \cap \{3n : n \in \mathbb{Z}\}$ .

*Proof.* Suppose  $a \in \{12n : n \in \mathbb{Z}\}$ . This means  $a = 12n$  for some  $n \in \mathbb{Z}$ . Therefore  $a = 2(6n)$  and  $a = 3(4n)$ . From  $a = 2(6n)$ , it follows that  $a$  is multiple of 2, so  $a \in \{2n : n \in \mathbb{Z}\}$ . From  $a = 3(4n)$ , it follows that  $a$  is multiple of 3, so  $a \in \{3n : n \in \mathbb{Z}\}$ . Thus by definition of the intersection of two sets, we have  $a \in \{2n : n \in \mathbb{Z}\} \cap \{3n : n \in \mathbb{Z}\}$ . Thus  $\{12n : n \in \mathbb{Z}\} \subseteq \{2n : n \in \mathbb{Z}\} \cap \{3n : n \in \mathbb{Z}\}$ . ■

3. If  $k \in \mathbb{Z}$ , then  $\{n \in \mathbb{Z} : n | k\} \subseteq \{n \in \mathbb{Z} : n | k^2\}$ .

*Proof.* Suppose  $k \in \mathbb{Z}$ . We now need to show  $\{n \in \mathbb{Z} : n | k\} \subseteq \{n \in \mathbb{Z} : n | k^2\}$ .

Suppose  $a \in \{n \in \mathbb{Z} : n | k\}$ . Then it follows that  $a | k$ , so there is an integer  $c$  for which  $k = ac$ . Then  $k^2 = a^2c^2$ . Therefore  $k^2 = a(ac^2)$ , and from this the definition of divisibility gives  $a | k^2$ . But  $a | k^2$  means that  $a \in \{n \in \mathbb{Z} : n | k^2\}$ . We have now shown  $\{n \in \mathbb{Z} : n | k\} \subseteq \{n \in \mathbb{Z} : n | k^2\}$ . ■

5. If  $p$  and  $q$  are integers, then  $\{pn : n \in \mathbb{N}\} \cap \{qn : n \in \mathbb{N}\} \neq \emptyset$ .

*Proof.* Suppose  $p$  and  $q$  are integers. Consider the integer  $pq$ . Observe that  $pq \in \{pn : n \in \mathbb{N}\}$  and  $pq \in \{qn : n \in \mathbb{N}\}$ , so  $pq \in \{pn : n \in \mathbb{N}\} \cap \{qn : n \in \mathbb{N}\}$ . Therefore  $\{pn : n \in \mathbb{N}\} \cap \{qn : n \in \mathbb{N}\} \neq \emptyset$ . ■

7. Suppose  $A, B$  and  $C$  are sets. If  $B \subseteq C$ , then  $A \times B \subseteq A \times C$ .

*Proof.* This is a conditional statement, and we'll prove it with direct proof. Suppose  $B \subseteq C$ . (Now we need to prove  $A \times B \subseteq A \times C$ .)

Suppose  $(a, b) \in A \times B$ . Then by definition of the Cartesian product we have  $a \in A$  and  $b \in B$ . But since  $b \in B$  and  $B \subseteq C$ , we have  $b \in C$ . Since  $a \in A$  and  $b \in C$ , it follows that  $(a, b) \in A \times C$ . Now we've shown  $(a, b) \in A \times B$  implies  $(a, b) \in A \times C$ , so  $A \times B \subseteq A \times C$ .

In summary, we've shown that if  $B \subseteq C$ , then  $A \times B \subseteq A \times C$ . This completes the proof. ■

9. If  $A, B$  and  $C$  are sets then  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ .

*Proof.* We use the distributive law  $P \wedge (Q \vee R) = (P \wedge Q) \vee (P \wedge R)$  from page 52.

$$\begin{aligned} A \cap (B \cup C) &= \{x : x \in A \wedge x \in B \cup C\} && (\text{def. of intersection}) \\ &= \{x : x \in A \wedge (x \in B \vee x \in C)\} && (\text{def. of union}) \\ &= \{x : (x \in A \wedge x \in B) \vee (x \in A \wedge x \in C)\} && (\text{distributive law}) \\ &= \{x : (x \in A \cap B) \vee (x \in A \cap C)\} && (\text{def. of intersection}) \\ &= (A \cap B) \cup (A \cap C) && (\text{def. of union}) \end{aligned}$$

The proof is complete. ■

11. If  $A$  and  $B$  are sets in a universal set  $U$ , then  $\overline{A \cup B} = \overline{A} \cap \overline{B}$ .

*Proof.* Just observe the following sequence of equalities.

$$\begin{aligned}
 \overline{A \cup B} &= U - (A \cup B) && \text{(def. of complement)} \\
 &= \{x : (x \in U) \wedge (x \notin A \cup B)\} && \text{(def. of } -\text{)} \\
 &= \{x : (x \in U) \wedge \sim(x \in A \cup B)\} \\
 &= \{x : (x \in U) \wedge \sim((x \in A) \vee (x \in B))\} && \text{(def. of } \cup\text{)} \\
 &= \{x : (x \in U) \wedge (\sim(x \in A) \wedge \sim(x \in B))\} && \text{(DeMorgan)} \\
 &= \{x : (x \in U) \wedge (x \notin A) \wedge (x \notin B)\} \\
 &= \{x : (x \in U) \wedge (x \in U) \wedge (x \notin A) \wedge (x \notin B)\} && (x \in U) = (x \in U) \wedge (x \in U) \\
 &= \{x : ((x \in U) \wedge (x \notin A)) \wedge ((x \in U) \wedge (x \notin B))\} && \text{(regroup)} \\
 &= \{x : (x \in U) \wedge (x \notin A)\} \cap \{x : (x \in U) \wedge (x \notin B)\} && \text{(def. of } \cap\text{)} \\
 &= (U - A) \cap (U - B) && \text{(def. of } -\text{)} \\
 &= \overline{A} \cap \overline{B} && \text{(def. of complement)}
 \end{aligned}$$

The proof is complete. ■

13. If  $A, B$  and  $C$  are sets, then  $A - (B \cup C) = (A - B) \cap (A - C)$ .

*Proof.* Just observe the following sequence of equalities.

$$\begin{aligned}
 A - (B \cup C) &= \{x : (x \in A) \wedge (x \notin B \cup C)\} && \text{(def. of } -\text{)} \\
 &= \{x : (x \in A) \wedge \sim(x \in B \cup C)\} \\
 &= \{x : (x \in A) \wedge \sim((x \in B) \vee (x \in C))\} && \text{(def. of } \cup\text{)} \\
 &= \{x : (x \in A) \wedge (\sim(x \in B) \wedge \sim(x \in C))\} && \text{(DeMorgan)} \\
 &= \{x : (x \in A) \wedge (x \notin B) \wedge (x \notin C)\} \\
 &= \{x : (x \in A) \wedge (x \in A) \wedge (x \notin B) \wedge (x \notin C)\} && (x \in A) = (x \in A) \wedge (x \in A) \\
 &= \{x : ((x \in A) \wedge (x \notin B)) \wedge ((x \in A) \wedge (x \notin C))\} && \text{(regroup)} \\
 &= \{x : (x \in A) \wedge (x \notin B)\} \cap \{x : (x \in A) \wedge (x \notin C)\} && \text{(def. of } \cap\text{)} \\
 &= (A - B) \cap (A - C) && \text{(def. of } -\text{)}
 \end{aligned}$$

The proof is complete. ■

15. If  $A, B$  and  $C$  are sets, then  $(A \cap B) - C = (A - C) \cap (B - C)$ .

*Proof.* Just observe the following sequence of equalities.

$$\begin{aligned}
 (A \cap B) - C &= \{x : (x \in A \cap B) \wedge (x \notin C)\} && \text{(def. of } -\text{)} \\
 &= \{x : (x \in A) \wedge (x \in B) \wedge (x \notin C)\} && \text{(def. of } \cap\text{)} \\
 &= \{x : (x \in A) \wedge (x \notin C) \wedge (x \in B) \wedge (x \notin C)\} && \text{(regroup)} \\
 &= \{x : ((x \in A) \wedge (x \notin C)) \wedge ((x \in B) \wedge (x \notin C))\} && \text{(regroup)} \\
 &= \{x : (x \in A) \wedge (x \notin C)\} \cap \{x : (x \in B) \wedge (x \notin C)\} && \text{(def. of } \cap\text{)} \\
 &= (A - C) \cap (B - C) && \text{(def. of } -\text{)}
 \end{aligned}$$

The proof is complete. ■

17. If  $A, B$  and  $C$  are sets, then  $A \times (B \cap C) = (A \times B) \cap (A \times C)$ .

*Proof.* See Example 8.12. ■

- 19.** Prove that  $\{9^n : n \in \mathbb{Z}\} \subseteq \{3^n : n \in \mathbb{Z}\}$ , but  $\{9^n : n \in \mathbb{Z}\} \neq \{3^n : n \in \mathbb{Z}\}$ .

*Proof.* Suppose  $a \in \{9^n : n \in \mathbb{Z}\}$ . This means  $a = 9^n$  for some integer  $n \in \mathbb{Z}$ . Thus  $a = 9^n = (3^2)^n = 3^{2n}$ . This shows  $a$  is an integer power of 3, so  $a \in \{3^n : n \in \mathbb{Z}\}$ . Therefore  $a \in \{9^n : n \in \mathbb{Z}\}$  implies  $a \in \{3^n : n \in \mathbb{Z}\}$ , so  $\{9^n : n \in \mathbb{Z}\} \subseteq \{3^n : n \in \mathbb{Z}\}$ .

But notice  $\{9^n : n \in \mathbb{Z}\} \neq \{3^n : n \in \mathbb{Z}\}$  as  $3 \in \{3^n : n \in \mathbb{Z}\}$ , but  $3 \notin \{9^n : n \in \mathbb{Z}\}$ . ■

- 21.** Suppose  $A$  and  $B$  are sets. Prove  $A \subseteq B$  if and only if  $A - B = \emptyset$ .

*Proof.* First we will prove that if  $A \subseteq B$ , then  $A - B = \emptyset$ . Contrapositive proof is used. Suppose that  $A - B \neq \emptyset$ . Thus there is an element  $a \in A - B$ , which means  $a \in A$  but  $a \notin B$ . Since not every element of  $A$  is in  $B$ , we have  $A \not\subseteq B$ .

Conversely, we will prove that if  $A - B = \emptyset$ , then  $A \subseteq B$ . Again, contrapositive proof is used. Suppose  $A \not\subseteq B$ . This means that it is not the case that every element of  $A$  is an element of  $B$ , so there is an element  $a \in A$  with  $a \notin B$ . Therefore we have  $a \in A - B$ , so  $A - B \neq \emptyset$ . ■

- 23.** For each  $a \in \mathbb{R}$ , let  $A_a = \{(x, a(x^2 - 1)) \in \mathbb{R}^2 : x \in \mathbb{R}\}$ . Prove that  $\bigcap_{a \in \mathbb{R}} A_a = \{(-1, 0), (1, 0)\}$ .

*Proof.* First we will show that  $\{(-1, 0), (1, 0)\} \subseteq \bigcap_{a \in \mathbb{R}} A_a$ . Notice that for any  $a \in \mathbb{R}$ , we have  $(-1, 0) \in A_a$  because  $A_a$  contains the ordered pair  $(-1, a((-1)^2 - 1)) = (-1, 0)$ . Similarly  $(1, 0) \in A_a$ . Thus each element of  $\{(-1, 0), (1, 0)\}$  belongs to every set  $A_a$ , so  $\{(-1, 0), (1, 0)\} \subseteq \bigcap_{a \in \mathbb{R}} A_a$ .

Now we will show  $\bigcap_{a \in \mathbb{R}} A_a \subseteq \{(-1, 0), (1, 0)\}$ . Suppose  $(c, d) \in \bigcap_{a \in \mathbb{R}} A_a$ . This means  $(c, d)$  is in every set  $A_a$ . In particular  $(c, d) \in A_0 = \{(x, 0(x^2 - 1)) : x \in \mathbb{R}\} = \{(x, 0) : x \in \mathbb{R}\}$ . It follows that  $d = 0$ . Then also we have  $(c, d) = (c, 0) \in A_1 = \{(x, 1(x^2 - 1)) : x \in \mathbb{R}\} = \{(x, x^2 - 1) : x \in \mathbb{R}\}$ . Therefore  $(c, 0)$  has the form  $(c, c^2 - 1)$ , that is  $(c, 0) = (c, c^2 - 1)$ . From this we get  $c^2 - 1 = 0$ , so  $c = \pm 1$ . Therefore  $(c, d) = (1, 0)$  or  $(c, d) = (-1, 0)$ , so  $(c, d) \in \{(-1, 0), (1, 0)\}$ . This completes the demonstration that  $(c, d) \in \bigcap_{a \in \mathbb{R}} A_a$  implies  $(c, d) \in \{(-1, 0), (1, 0)\}$ , so it follows that  $\bigcap_{a \in \mathbb{R}} A_a \subseteq \{(-1, 0), (1, 0)\}$ .

Now it's been shown that  $\{(-1, 0), (1, 0)\} \subseteq \bigcap_{a \in \mathbb{R}} A_a$  and  $\bigcap_{a \in \mathbb{R}} A_a \subseteq \{(-1, 0), (1, 0)\}$ , so it follows that  $\bigcap_{a \in \mathbb{R}} A_a = \{(-1, 0), (1, 0)\}$ . ■

- 25.** Suppose  $A, B, C$  and  $D$  are sets. Prove that  $(A \times B) \cup (C \times D) \subseteq (A \cup C) \times (B \cup D)$ .

*Proof.* Suppose  $(a, b) \in (A \times B) \cup (C \times D)$ .

By definition of union, this means  $(a, b) \in (A \times B)$  or  $(a, b) \in (C \times D)$ .

We examine these two cases individually.

**Case 1.** Suppose  $(a, b) \in (A \times B)$ . By definition of  $\times$ , it follows that  $a \in A$  and  $b \in B$ . From this, it follows from the definition of  $\cup$  that  $a \in A \cup C$  and  $b \in B \cup D$ . Again from the definition of  $\times$ , we get  $(a, b) \in (A \cup C) \times (B \cup D)$ .

**Case 2.** Suppose  $(a, b) \in (C \times D)$ . By definition of  $\times$ , it follows that  $a \in C$  and  $b \in D$ . From this, it follows from the definition of  $\cup$  that  $a \in A \cup C$  and  $b \in B \cup D$ . Again from the definition of  $\times$ , we get  $(a, b) \in (A \cup C) \times (B \cup D)$ .

In either case, we obtained  $(a, b) \in (A \cup C) \times (B \cup D)$ , so we've proved that  $(a, b) \in (A \times B) \cup (C \times D)$  implies  $(a, b) \in (A \cup C) \times (B \cup D)$ . Therefore  $(A \times B) \cup (C \times D) \subseteq (A \cup C) \times (B \cup D)$ . ■

**27.** Prove  $\{12a + 4b : a, b \in \mathbb{Z}\} = \{4c : c \in \mathbb{Z}\}$ .

*Proof.* First we show  $\{12a + 4b : a, b \in \mathbb{Z}\} \subseteq \{4c : c \in \mathbb{Z}\}$ . Suppose  $x \in \{12a + 4b : a, b \in \mathbb{Z}\}$ . Then  $x = 12a + 4b$  for some integers  $a$  and  $b$ . From this we get  $x = 4(3a + b)$ , so  $x = 4c$  where  $c$  is the integer  $3a + b$ . Consequently  $x \in \{4c : c \in \mathbb{Z}\}$ . This establishes that  $\{12a + 4b : a, b \in \mathbb{Z}\} \subseteq \{4c : c \in \mathbb{Z}\}$ .

Next we show  $\{4c : c \in \mathbb{Z}\} \subseteq \{12a + 4b : a, b \in \mathbb{Z}\}$ . Suppose  $x \in \{4c : c \in \mathbb{Z}\}$ . Then  $x = 4c$  for some  $c \in \mathbb{Z}$ . Thus  $x = (12 + 4(-2))c = 12c + 4(-2c)$ , and since  $c$  and  $-2c$  are integers we have  $x \in \{12a + 4b : a, b \in \mathbb{Z}\}$ .

This proves that  $\{12a + 4b : a, b \in \mathbb{Z}\} = \{4c : c \in \mathbb{Z}\}$ . ■

**29.** Suppose  $A \neq \emptyset$ . Prove that  $A \times B \subseteq A \times C$ , if and only if  $B \subseteq C$ .

*Proof.* First we will prove that if  $A \times B \subseteq A \times C$ , then  $B \subseteq C$ . Using contrapositive, suppose that  $B \not\subseteq C$ . This means there is an element  $b \in B$  with  $b \notin C$ . Since  $A \neq \emptyset$ , there exists an element  $a \in A$ . Now consider the ordered pair  $(a, b)$ . Note that  $(a, b) \in A \times B$ , but  $(a, b) \notin A \times C$ . This means  $A \times B \not\subseteq A \times C$ .

Conversely, we will now show that if  $B \subseteq C$ , then  $A \times B \subseteq A \times C$ . We use direct proof. Suppose  $B \subseteq C$ . Assume that  $(a, b) \in A \times B$ . This means  $a \in A$  and  $b \in B$ . But, as  $B \subseteq C$ , we also have  $b \in C$ . From  $a \in A$  and  $b \in C$ , we get  $(a, b) \in A \times C$ . We've now shown  $(a, b) \in A \times B$  implies  $(a, b) \in A \times C$ , so  $A \times B \subseteq A \times C$ . ■

**31.** Suppose  $B \neq \emptyset$  and  $A \times B \subseteq B \times C$ . Prove  $A \subseteq C$ .

*Proof.* Suppose  $B \neq \emptyset$  and  $A \times B \subseteq B \times C$ . In what follows, we show that  $A \subseteq C$ .

Let  $x \in A$ . Because  $B$  is not empty, it contains some element  $b$ . Observe that  $(x, b) \in A \times B$ . But as  $A \times B \subseteq B \times C$ , we also have  $(x, b) \in B \times C$ , so, in particular,  $x \in B$ . As  $x \in A$  and  $x \in B$ , we have  $(x, x) \in A \times B$ . But as  $A \times B \subseteq B \times C$ , it follows that  $(x, x) \in B \times C$ . This implies  $x \in C$ . We've shown  $x \in A$  implies  $x \in C$ , so  $A \subseteq C$ . ■

## Chapter 9 Exercises

1. If  $x, y \in \mathbb{R}$ , then  $|x + y| = |x| + |y|$ .

This is **false**.

*Disproof:* Here is a counterexample: Let  $x = 1$  and  $y = -1$ . Then  $|x + y| = 0$  and  $|x| + |y| = 2$ , so it's not true that  $|x + y| = |x| + |y|$ .

3. If  $n \in \mathbb{Z}$  and  $n^5 - n$  is even, then  $n$  is even.

This is **false**.

*Disproof:* Here is a counterexample: Let  $n = 3$ . Then  $n^5 - n = 3^5 - 3 = 240$ , but  $n$  is not even.

5. If  $A, B, C$  and  $D$  are sets, then  $(A \times B) \cup (C \times D) = (A \cup C) \times (B \cup D)$ .

This is **false**.

*Disproof:* Here is a counterexample: Let  $A = \{1, 2\}$ ,  $B = \{1, 2\}$ ,  $C = \{2, 3\}$  and  $D = \{2, 3\}$ . Then  $(A \times B) \cup (C \times D) = \{(1, 1), (1, 2), (2, 1), (2, 2)\} \cup \{(2, 2), (2, 3), (3, 2), (3, 3)\} = \{(1, 1), (1, 2), (2, 1), (2, 2), (2, 3), (3, 2), (3, 3)\}$ . Also  $(A \cup C) \times (B \cup D) = \{1, 2, 3\} \times \{1, 2, 3\} = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}$ , so you can see that  $(A \times B) \cup (C \times D) \neq (A \cup C) \times (B \cup D)$ .

7. If  $A, B$  and  $C$  are sets, and  $A \times C = B \times C$ , then  $A = B$ .

This is **false**.

*Disproof:* Here is a counterexample: Let  $A = \{1\}$ ,  $B = \{2\}$  and  $C = \emptyset$ . Then  $A \times C = B \times C = \emptyset$ , but  $A \neq B$ .

9. If  $A$  and  $B$  are sets, then  $\mathcal{P}(A) - \mathcal{P}(B) \subseteq \mathcal{P}(A - B)$ .

This is **false**.

*Disproof:* Here is a counterexample: Let  $A = \{1, 2\}$  and  $B = \{1\}$ . Then  $\mathcal{P}(A) - \mathcal{P}(B) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\} - \{\emptyset, \{1\}\} = \{\{2\}, \{1, 2\}\}$ . Also  $\mathcal{P}(A - B) = \mathcal{P}(\{2\}) = \{\emptyset, \{2\}\}$ . In this example we have  $\mathcal{P}(A) - \mathcal{P}(B) \not\subseteq \mathcal{P}(A - B)$ .

11. If  $a, b \in \mathbb{N}$ , then  $a + b < ab$ .

This is **false**.

*Disproof:* Here is a counterexample: Let  $a = 1$  and  $b = 1$ . Then  $a + b = 2$  and  $ab = 1$ , so it's not true that  $a + b < ab$ .

13. There exists a set  $X$  for which  $\mathbb{R} \subseteq X$  and  $\emptyset \in X$ .

This is **true**.

*Proof.* Simply let  $X = \mathbb{R} \cup \{\emptyset\}$ . If  $x \in \mathbb{R}$ , then  $x \in \mathbb{R} \cup \{\emptyset\} = X$ , so  $\mathbb{R} \subseteq X$ . Likewise,  $\emptyset \in \mathbb{R} \cup \{\emptyset\} = X$  because  $\emptyset \in \{\emptyset\}$ . ■

15. Every odd integer is the sum of three odd integers.

This is **true**.

*Proof.* If  $n$  is odd, then  $n = n + 1 + (-1)$ . Thus  $n$  is the sum of three odd integers. ■

17. For all sets  $A$  and  $B$ , if  $A - B = \emptyset$ , then  $B \neq \emptyset$ .

This is **false**.

*Disproof:* Here is a counterexample: Just let  $A = \emptyset$  and  $B = \emptyset$ . Then  $A - B = \emptyset$ , but it's not true that  $B \neq \emptyset$ .

19. For every  $r, s \in \mathbb{Q}$  with  $r < s$ , there is an irrational number  $u$  for which  $r < u < s$ . This is **true**.

*Proof.* (Direct) Suppose  $r, s \in \mathbb{Q}$  with  $r < s$ . Consider the number  $u = r + \sqrt{2} \frac{s-r}{2}$ . In what follows we will show that  $u$  is irrational and  $r < u < s$ . Certainly since  $s-r$  is positive, it follows that  $r < r + \sqrt{2} \frac{s-r}{2} = u$ . Also, since  $\sqrt{2} < 2$  we have

$$u = r + \sqrt{2} \frac{s-r}{2} < r + 2 \frac{s-r}{2} = s,$$

and therefore  $u < s$ . Thus we can conclude  $r < u < s$ .

Now we just need to show that  $u$  is irrational. Suppose for the sake of contradiction that  $u$  is rational. Then  $u = \frac{a}{b}$  for some integers  $a$  and  $b$ . Since  $r$  and  $s$  are rational, we have  $r = \frac{c}{d}$  and  $s = \frac{e}{f}$  for some  $c, d, e, f \in \mathbb{Z}$ . Now we have

$$\begin{aligned} u &= r + \sqrt{2} \frac{s-r}{2} \\ \frac{a}{b} &= \frac{c}{d} + \sqrt{2} \frac{\frac{e}{f} - \frac{c}{d}}{2} \\ \frac{ad - bc}{bd} &= \sqrt{2} \frac{ed - cf}{2df} \\ \frac{(ad - bc)2df}{bd(ed - cf)} &= \sqrt{2} \end{aligned}$$

This expresses  $\sqrt{2}$  as a quotient of two integers, so  $\sqrt{2}$  is rational, a contradiction. Thus  $u$  is irrational.

In summary, we have produced an irrational number  $u$  with  $r < u < s$ , so the proof is complete. ■

21. There exist two prime numbers  $p$  and  $q$  for which  $p - q = 97$ .  
This statement is **false**.

*Disproof:* Suppose for the sake of contradiction that this is true. Let  $p$  and  $q$  be prime numbers for which  $p - q = 97$ . Now, since their difference is odd,  $p$  and  $q$  must have opposite parity, so one of  $p$  and  $q$  is even and the other is odd. But there exists only one even prime number (namely 2), so either  $p = 2$  or  $q = 2$ . If  $p = 2$ , then  $p - q = 97$  implies  $q = 2 - 97 = -95$ , which is not prime. On the other hand if  $q = 2$ , then  $p - q = 97$  implies  $p = 99$ , but that's not prime either. Thus one of  $p$  or  $q$  is not prime, a contradiction.

23. If  $x, y \in \mathbb{R}$  and  $x^3 < y^3$ , then  $x < y$ . This is **true**.

*Proof.* (Contrapositive) Suppose  $x \geq y$ . We need to show  $x^3 \geq y^3$ .

**Case 1.** Suppose  $x$  and  $y$  have opposite signs, that is one of  $x$  and  $y$  is positive and the other is negative. Then since  $x \geq y$ ,  $x$  is positive and  $y$  is negative. Then, since the powers are odd,  $x^3$  is positive and  $y^3$  is negative, so  $x^3 \geq y^3$ .

**Case 2.** Suppose  $x$  and  $y$  do not have opposite signs. Then  $x^2 + xy + y^2 \geq 0$  and

also  $x - y \geq 0$  because  $x \geq y$ . Thus we have  $x^3 - y^3 = (x - y)(x^2 + xy + y^2) \geq 0$ . From this we get  $x^3 - y^3 \geq 0$ , so  $x^3 \geq y^3$ .

In either case we have  $x^3 \geq y^3$ . ■

- 25.** For all  $a, b, c \in \mathbb{Z}$ , if  $a \mid bc$ , then  $a \mid b$  or  $a \mid c$ .

This is **false**.

*Disproof:* Let  $a = 6$ ,  $b = 3$  and  $c = 4$ . Note that  $a \mid bc$ , but  $a \nmid b$  and  $a \nmid c$ .

- 27.** The equation  $x^2 = 2^x$  has three real solutions.

*Proof.* By inspection, the numbers  $x = 2$  and  $x = 4$  are two solutions of this equation. But there is a third solution. Let  $m$  be a positive real number for which  $m2^m = \frac{1}{2}$ . (The existence of such an  $m$  is guaranteed by the intermediate value theorem of calculus.) Then negative number  $x = -2m$  is a solution, as follows.

$$x^2 = (-2m)^2 = 4m^2 = 4\left(\frac{m2^m}{2^m}\right)^2 = 4\left(\frac{\frac{1}{2}}{2^m}\right)^2 = \frac{1}{2^{2m}} = 2^{-2m} = 2^x.$$

Therefore we have three solutions 2, 4 and  $m$ . ■

- 29.** If  $x, y \in \mathbb{R}$  and  $|x + y| = |x - y|$ , then  $y = 0$ .

This is **false**.

*Disproof:* Let  $x = 0$  and  $y = 1$ . Then  $|x + y| = |x - y|$ , but  $y = 1$ .

- 31.** No number appears in Pascal's triangle more than four times.

This is **false**.

*Disproof:* The number 120 appears six times. Check that  $\binom{10}{3} = \binom{10}{7} = \binom{16}{2} = \binom{16}{14} = \binom{120}{1} = \binom{120}{119} = 120$ .

- 33.** Suppose  $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$  is a polynomial of degree 1 or greater, and for which each coefficient  $a_i$  is in  $\mathbb{N}$ . Then there is an  $n \in \mathbb{N}$  for which the integer  $f(n)$  is not prime.

*Proof.* (Outline) Note that, because the coefficients are all positive and the degree is greater than 1, we have  $f(1) > 1$ . Let  $b = f(1) > 1$ . Now, the polynomial  $f(x) - b$  has a root 1, so  $f(x) - b = (x - 1)g(x)$  for some polynomial  $g$ . Then  $f(x) = (x - 1)g(x) + b$ . Now note that  $f(b + 1) = bg(b) + b = b(g(b) + 1)$ . If we can now show that  $g(b) + 1$  is an integer, then we have a nontrivial factoring  $f(b + 1) = b(g(b) + 1)$ , and  $f(b + 1)$  is not prime. To complete the proof, use the fact that  $f(x) - b = (x - 1)g(x)$  has integer coefficients, and deduce that  $g(x)$  must also have integer coefficients. ■

## Chapter 10 Exercises

- 1.** Prove that  $1 + 2 + 3 + 4 + \cdots + n = \frac{n^2 + n}{2}$  for every integer  $n \in \mathbb{N}$ .

*Proof.* We will prove this with mathematical induction.

- (1) Observe that if  $n = 1$ , this statement is  $1 = \frac{1^2 + 1}{2}$ , which is obviously true.

- (2) Consider any integer  $k \geq 1$ . We must show that  $S_k$  implies  $S_{k+1}$ . In other words, we must show that if  $1 + 2 + 3 + 4 + \dots + k = \frac{k^2+k}{2}$  is true, then

$$1 + 2 + 3 + 4 + \dots + k + (k+1) = \frac{(k+1)^2 + (k+1)}{2}$$

is also true. We use direct proof.

Suppose  $k \geq 1$  and  $1 + 2 + 3 + 4 + \dots + k = \frac{k^2+k}{2}$ . Observe that

$$\begin{aligned} 1 + 2 + 3 + 4 + \dots + k + (k+1) &= \\ (1 + 2 + 3 + 4 + \dots + k) + (k+1) &= \\ \frac{k^2+k}{2} + (k+1) &= \frac{k^2+k+2(k+1)}{2} \\ &= \frac{k^2+2k+1+k+1}{2} \\ &= \frac{(k+1)^2+(k+1)}{2}. \end{aligned}$$

Therefore we have shown that  $1 + 2 + 3 + 4 + \dots + k + (k+1) = \frac{(k+1)^2+(k+1)}{2}$ . ■

3. Prove that  $1^3 + 2^3 + 3^3 + 4^3 + \dots + n^3 = \frac{n^2(n+1)^2}{4}$  for every positive integer  $n$ .

*Proof.* We will prove this with mathematical induction.

- (1) When  $n = 1$  the statement is  $1^3 = \frac{1^2(1+1)^2}{4} = \frac{4}{4} = 1$ , and this is true.  
 (2) Now assume the statement is true for some integer  $n = k \geq 1$ , that is assume  $1^3 + 2^3 + 3^3 + 4^3 + \dots + k^3 = \frac{k^2(k+1)^2}{4}$ . Observe that this implies the statement is true for  $n = k + 1$ :

$$\begin{aligned} 1^3 + 2^3 + 3^3 + 4^3 + \dots + k^3 + (k+1)^3 &= \\ (1^3 + 2^3 + 3^3 + 4^3 + \dots + k^3) + (k+1)^3 &= \\ \frac{k^2(k+1)^2}{4} + (k+1)^3 &= \frac{k^2(k+1)^2}{4} + \frac{4(k+1)^3}{4} \\ &= \frac{k^2(k+1)^2 + 4(k+1)^3}{4} \\ &= \frac{(k+1)^2(k^2 + 4(k+1)^1)}{4} \\ &= \frac{(k+1)^2(k^2 + 4k + 4)}{4} \\ &= \frac{(k+1)^2(k+2)^2}{4} \\ &= \frac{(k+1)^2((k+1)+1)^2}{4}. \end{aligned}$$

Therefore  $1^3 + 2^3 + 3^3 + 4^3 + \dots + k^3 + (k+1)^3 = \frac{(k+1)^2((k+1)+1)^2}{4}$ , which means the statement is true for  $n = k + 1$ . ■

5. If  $n \in \mathbb{N}$ , then  $2^1 + 2^2 + 2^3 + \cdots + 2^n = 2^{n+1} - 2$ .

*Proof.* The proof is by mathematical induction.

- (1) When  $n = 1$ , this statement is  $2^1 = 2^{1+1} - 2$ , or  $2 = 4 - 2$ , which is true.
- (2) Now assume the statement is true for some integer  $n = k \geq 1$ , that is assume  $2^1 + 2^2 + 2^3 + \cdots + 2^k = 2^{k+1} - 2$ . Observe this implies that the statement is true for  $n = k + 1$ , as follows:

$$\begin{aligned} 2^1 + 2^2 + 2^3 + \cdots + 2^k + 2^{k+1} &= \\ (2^1 + 2^2 + 2^3 + \cdots + 2^k) + 2^{k+1} &= \\ 2^{k+1} - 2 + 2^{k+1} &= 2 \cdot 2^{k+1} - 2 \\ &= 2^{k+2} - 2 \\ &= 2^{(k+1)+1} - 2. \end{aligned}$$

Thus we have  $2^1 + 2^2 + 2^3 + \cdots + 2^k + 2^{k+1} = 2^{(k+1)+1} - 2$ , so the statement is true for  $n = k + 1$ .

Thus the result follows by mathematical induction. ■

7. If  $n \in \mathbb{N}$ , then  $1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 6 + \cdots + n(n+2) = \frac{n(n+1)(2n+7)}{6}$ .

*Proof.* The proof is by mathematical induction.

- (1) When  $n = 1$ , we have  $1 \cdot 3 = \frac{1(1+1)(2+7)}{6}$ , which is the true statement  $3 = \frac{18}{6}$ .
- (2) Now assume the statement is true for some integer  $n = k \geq 1$ , that is assume  $1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 6 + \cdots + k(k+2) = \frac{k(k+1)(2k+7)}{6}$ . Now observe that

$$\begin{aligned} 1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 6 + \cdots + k(k+2) + (k+1)((k+1)+2) &= \\ (1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 6 + \cdots + k(k+2)) + (k+1)((k+1)+2) &= \\ \frac{k(k+1)(2k+7)}{6} + (k+1)((k+1)+2) &= \\ \frac{k(k+1)(2k+7)}{6} + \frac{6(k+1)(k+3)}{6} &= \\ \frac{k(k+1)(2k+7) + 6(k+1)(k+3)}{6} &= \\ \frac{(k+1)(k(2k+7) + 6(k+3))}{6} &= \\ \frac{(k+1)(2k^2 + 13k + 18)}{6} &= \\ \frac{(k+1)(k+2)(2k+9)}{6} &= \\ \frac{(k+1)((k+1)+1)(2(k+1)+7)}{6}. \end{aligned}$$

Thus we have  $1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 6 + \cdots + k(k+2) + (k+1)((k+1)+2) = \frac{(k+1)((k+1)+1)(2(k+1)+7)}{6}$ , and this means the statement is true for  $n = k + 1$ .

Thus the result follows by mathematical induction. ■

- 9.** Prove that  $24 \mid (5^{2n} - 1)$  for any integer  $n \geq 0$ .

*Proof.* The proof is by mathematical induction.

- (1) For  $n = 0$ , the statement is  $24 \mid (5^{2 \cdot 0} - 1)$ . This is  $24 \mid 0$ , which is true.
- (2) Now assume the statement is true for some integer  $n = k \geq 0$ , that is assume  $24 \mid (5^{2k} - 1)$ . This means  $5^{2k} - 1 = 24a$  for some integer  $a$ , and from this we get  $5^{2k} = 24a + 1$ . Now observe that

$$\begin{aligned} 5^{2(k+1)} - 1 &= \\ 5^{2k+2} - 1 &= \\ 5^2 5^{2k} - 1 &= \\ 5^2(24a + 1) - 1 &= \\ 25(24a + 1) - 1 &= \\ 25 \cdot 24a + 25 - 1 &= 24(25a + 1). \end{aligned}$$

This shows  $5^{2(k+1)} - 1 = 24(25a + 1)$ , which means  $24 \mid 5^{2(k+1)} - 1$ .

This completes the proof by mathematical induction. ■

- 11.** Prove that  $3 \mid (n^3 + 5n + 6)$  for any integer  $n \geq 0$ .

*Proof.* The proof is by mathematical induction.

- (1) When  $n = 0$ , the statement is  $3 \mid (0^3 + 5 \cdot 0 + 6)$ , or  $3 \mid 6$ , which is true.
- (2) Now assume the statement is true for some integer  $n = k \geq 0$ , that is assume  $3 \mid (k^3 + 5k + 6)$ . This means  $k^3 + 5k + 6 = 3a$  for some integer  $a$ . We need to show that  $3 \mid ((k+1)^3 + 5(k+1) + 6)$ . Observe that

$$\begin{aligned} (k+1)^3 + 5(k+1) + 6 &= k^3 + 3k^2 + 3k + 1 + 5k + 5 + 6 \\ &= (k^3 + 5k + 6) + 3k^2 + 3k + 6 \\ &= 3a + 3k^2 + 3k + 6 \\ &= 3(a + k^2 + k + 2). \end{aligned}$$

Thus we have deduced  $(k+1)^3 - (k+1) = 3(a + k^2 + k + 2)$ . Since  $a + k^2 + k + 2$  is an integer, it follows that  $3 \mid ((k+1)^3 + 5(k+1) + 6)$ .

It follows by mathematical induction that  $3 \mid (n^3 + 5n + 6)$  for every  $n \geq 0$ . ■

- 13.** Prove that  $6 \mid (n^3 - n)$  for any integer  $n \geq 0$ .

*Proof.* The proof is by mathematical induction.

- (1) When  $n = 0$ , the statement is  $6 \mid (0^3 - 0)$ , or  $6 \mid 0$ , which is true.

- (2) Now assume the statement is true for some integer  $n = k \geq 0$ , that is, assume  $6 | (k^3 - k)$ . This means  $k^3 - k = 6a$  for some integer  $a$ . We need to show that  $6 | ((k+1)^3 - (k+1))$ . Observe that

$$\begin{aligned}(k+1)^3 - (k+1) &= k^3 + 3k^2 + 3k + 1 - k - 1 \\&= (k^3 - k) + 3k^2 + 3k \\&= 6a + 3k^2 + 3k \\&= 6a + 3k(k+1).\end{aligned}$$

Thus we have deduced  $(k+1)^3 - (k+1) = 6a + 3k(k+1)$ . Since one of  $k$  or  $(k+1)$  must be even, it follows that  $k(k+1)$  is even, so  $k(k+1) = 2b$  for some integer  $b$ . Consequently  $(k+1)^3 - (k+1) = 6a + 3k(k+1) = 6a + 3(2b) = 6(a+b)$ . Since  $(k+1)^3 - (k+1) = 6(a+b)$  it follows that  $6 | ((k+1)^3 - (k+1))$ .

Thus the result follows by mathematical induction. ■

- 15.** If  $n \in \mathbb{N}$ , then  $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \frac{1}{4 \cdot 5} + \cdots + \frac{1}{n(n+1)} = 1 - \frac{1}{n+1}$ .

*Proof.* The proof is by mathematical induction.

- (1) When  $n = 1$ , the statement is  $\frac{1}{1(1+1)} = 1 - \frac{1}{1+1}$ , which simplifies to  $\frac{1}{2} = \frac{1}{2}$ .
- (2) Now assume the statement is true for some integer  $n = k \geq 1$ , that is assume  $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \frac{1}{4 \cdot 5} + \cdots + \frac{1}{k(k+1)} = 1 - \frac{1}{k+1}$ . Next we show that the statement for  $n = k+1$  is true. Observe that

$$\begin{aligned}\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \frac{1}{4 \cdot 5} + \cdots + \frac{1}{k(k+1)} + \frac{1}{(k+1)((k+1)+1)} &= \\ \left( \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \frac{1}{4 \cdot 5} + \cdots + \frac{1}{k(k+1)} \right) + \frac{1}{(k+1)(k+2)} &= \\ \left( 1 - \frac{1}{k+1} \right) + \frac{1}{(k+1)(k+2)} &= \\ 1 - \frac{1}{k+1} + \frac{1}{(k+1)(k+2)} &= \\ 1 - \frac{k+2}{(k+1)(k+2)} + \frac{1}{(k+1)(k+2)} &= \\ 1 - \frac{k+1}{(k+1)(k+2)} &= \\ 1 - \frac{1}{k+2} &= \\ 1 - \frac{1}{(k+1)+1}.\end{aligned}$$

This establishes  $\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \frac{1}{4 \cdot 5} + \cdots + \frac{1}{(k+1)((k+1)+1)} = 1 - \frac{1}{(k+1)+1}$ , which is to say that the statement is true for  $n = k+1$ .

This completes the proof by mathematical induction. ■

17. Suppose  $A_1, A_2, \dots, A_n$  are sets in some universal set  $U$ , and  $n \geq 2$ . Prove that  $\overline{A_1 \cap A_2 \cap \dots \cap A_n} = \overline{A_1} \cup \overline{A_2} \cup \dots \cup \overline{A_n}$ .

*Proof.* The proof is by strong induction.

- (1) When  $n = 2$  the statement is  $\overline{A_1 \cap A_2} = \overline{A_1} \cup \overline{A_2}$ . This is not an entirely obvious statement, so we have to prove it. Observe that

$$\begin{aligned}\overline{A_1 \cap A_2} &= \{x : (x \in U) \wedge (x \notin A_1 \cap A_2)\} \text{ (definition of complement)} \\ &= \{x : (x \in U) \wedge \sim(x \in A_1 \cap A_2)\} \\ &= \{x : (x \in U) \wedge \sim((x \in A_1) \wedge (x \in A_2))\} \text{ (definition of } \cap\text{)} \\ &= \{x : (x \in U) \wedge (\sim(x \in A_1) \vee \sim(x \in A_2))\} \text{ (DeMorgan)} \\ &= \{x : (x \in U) \wedge ((x \notin A_1) \vee (x \notin A_2))\} \\ &= \{x : (x \in U) \wedge (x \notin A_1) \vee (x \in U) \wedge (x \notin A_2)\} \text{ (distributive prop.)} \\ &= \{x : ((x \in U) \wedge (x \notin A_1)) \cup \{x : ((x \in U) \wedge (x \notin A_2))\} \text{ (def. of } \cup\text{)} \\ &= \overline{A_1} \cup \overline{A_2} \text{ (definition of complement).}\end{aligned}$$

- (2) Let  $k \geq 2$ . Assume the statement is true if it involves  $k$  or fewer sets. Then

$$\begin{aligned}\frac{\overline{A_1 \cap A_2 \cap \dots \cap A_{k-1} \cap A_k \cap A_{k+1}}}{A_1 \cap A_2 \cap \dots \cap A_{k-1} \cap (A_k \cap A_{k+1})} &= \\ \overline{\overline{A_1} \cup \overline{A_2} \cup \dots \cup \overline{A_{k-1}} \cup \overline{A_k \cap A_{k+1}}} &= \\ \overline{\overline{A_1} \cup \overline{A_2} \cup \dots \cup \overline{A_{k-1}} \cup \overline{A_k} \cup \overline{A_{k+1}}}.\end{aligned}$$

Thus the statement is true when it involves  $k + 1$  sets.

This completes the proof by strong induction. ■

19. Prove  $\sum_{k=1}^n \frac{1}{k^2} \leq 2 - \frac{1}{n}$  for every  $n$ .

*Proof.* This clearly holds for  $n = 1$ . Assume it holds for some  $n \geq 1$ . Then  $\sum_{k=1}^{n+1} \frac{1}{k^2} = \sum_{k=1}^n \frac{1}{k^2} + \frac{1}{(n+1)^2} \leq 2 - \frac{1}{n} + \frac{1}{(n+1)^2} = 2 - \frac{(n+1)^2 - n}{n(n+1)^2} = 2 - \frac{n^2 + n + 1}{n(n+1)^2} < 2 - \frac{n^2 + n}{n(n+1)^2} = 2 - \frac{1}{(n+1)}$ . ■

21. If  $n \in \mathbb{N}$ , then  $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{2^n} \geq 1 + \frac{n}{2}$ .

*Proof.* If  $n = 1$ , the result is obvious.

Assume the proposition holds for some  $n > 1$ . Then

$$\begin{aligned}\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{2^{n+1}} &= \left( \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{2^n} \right) + \left( \frac{1}{2^{n+1}} + \frac{1}{2^{n+2}} + \frac{1}{2^{n+3}} + \dots + \frac{1}{2^{n+1}} \right) \\ &\geq \left( 1 + \frac{n}{2} \right) + \left( \frac{1}{2^{n+1}} + \frac{1}{2^{n+2}} + \frac{1}{2^{n+3}} + \dots + \frac{1}{2^{n+1}} \right).\end{aligned}$$

Now, the sum  $\left( \frac{1}{2^{n+1}} + \frac{1}{2^{n+2}} + \frac{1}{2^{n+3}} + \dots + \frac{1}{2^{n+1}} \right)$  on the right has  $2^{n+1} - 2^n = 2^n$  terms, all greater than or equal to  $\frac{1}{2^{n+1}}$ , so the sum is greater than  $2^n \cdot \frac{1}{2^{n+1}} = \frac{1}{2}$ . Therefore we get  $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{2^{n+1}} \geq \left( 1 + \frac{n}{2} \right) + \left( \frac{1}{2^{n+1}} + \frac{1}{2^{n+2}} + \frac{1}{2^{n+3}} + \dots + \frac{1}{2^{n+1}} \right) \geq \left( 1 + \frac{n}{2} \right) + \frac{1}{2} = 1 + \frac{n+1}{2}$ . This means the result is true for  $n + 1$ , so the theorem is proved. ■

- 23.** Use induction to prove the binomial theorem  $(x+y)^n = \sum_{i=0}^n \binom{n}{i} x^{n-i} y^i$ .

*Proof.* Notice that when  $n = 1$ , the formula is  $(x+y)^1 = \binom{1}{0} x^1 y^0 + \binom{1}{1} x^0 y^1 = x+y$ , which is true.

Now assume the theorem is true for some  $n > 1$ . We will show that this implies that it is true for the power  $n+1$ . Just observe that

$$\begin{aligned} (x+y)^{n+1} &= (x+y)(x+y)^n \\ &= (x+y) \sum_{i=0}^n \binom{n}{i} x^{n-i} y^i \\ &= \sum_{i=0}^n \binom{n}{i} x^{(n+1)-i} y^i + \sum_{i=0}^n \binom{n}{i} x^{n-i} y^{i+1} \\ &= \sum_{i=0}^n \left[ \binom{n}{i} + \binom{n}{i-1} \right] x^{(n+1)-i} y^i + y^{n+1} \\ &= \sum_{i=0}^n \binom{n+1}{i} x^{(n+1)-i} y^i + \binom{n+1}{n+1} y^{n+1} \\ &= \sum_{i=0}^{n+1} \binom{n+1}{i} x^{(n+1)-i} y^i. \end{aligned}$$

This shows that the formula is true for  $(x+y)^{n+1}$ , so the theorem is proved. ■

- 25.** Concerning the Fibonacci sequence, prove that  $F_1 + F_2 + F_3 + F_4 + \dots + F_n = F_{n+2} - 1$ .

*Proof.* The proof is by induction.

- (1) When  $n = 1$  the statement is  $F_1 = F_{1+2} - 1 = F_3 - 1 = 2 - 1 = 1$ , which is true. Also when  $n = 2$  the statement is  $F_1 + F_2 = F_{2+2} - 1 = F_4 - 1 = 3 - 1 = 2$ , which is true, as  $F_1 + F_2 = 1 + 1 = 2$ .
- (2) Now assume  $k \geq 1$  and  $F_1 + F_2 + F_3 + F_4 + \dots + F_k = F_{k+2} - 1$ . We need to show  $F_1 + F_2 + F_3 + F_4 + \dots + F_k + F_{k+1} = F_{k+3} - 1$ . Observe that

$$\begin{aligned} F_1 + F_2 + F_3 + F_4 + \dots + F_k + F_{k+1} &= \\ (F_1 + F_2 + F_3 + F_4 + \dots + F_k) + F_{k+1} &= \\ F_{k+2} - 1 + F_{k+1} &= (F_{k+1} + F_{k+2}) - 1 \\ &= F_{k+3} - 1. \end{aligned}$$

This completes the proof by induction. ■

- 27.** Concerning the Fibonacci sequence, prove that  $F_1 + F_3 + \dots + F_{2n-1} = F_{2n}$ .

*Proof.* If  $n = 1$ , the result is clear. Assume for some  $n > 1$  we have  $\sum_{i=1}^n F_{2i-1} = F_{2n}$ .

Then  $\sum_{i=1}^{n+1} F_{2i-1} = F_{2n+1} + \sum_{i=1}^n F_{2i-1} = F_{2n+1} + F_{2n} = F_{2n+2} = F_{2(n+1)}$  as desired. ■

- 29.** Notice that the sum of elements on the  $n$ th diagonal has the form

$$\binom{n}{0} + \binom{n-1}{1} + \binom{n-2}{2} + \binom{n-3}{3} + \cdots + \binom{0}{n}.$$

(For example,  $\binom{6}{0} + \binom{5}{1} + \binom{4}{2} + \binom{3}{3} + \binom{2}{4} + \binom{1}{5} + \binom{0}{6} = 1+5+6+1+0+0+0 = 13 = F_{6+1}$ .) Therefore, we need to prove that  $\binom{n}{0} + \binom{n-1}{1} + \binom{n-2}{2} + \binom{n-3}{3} + \cdots + \binom{1}{n-1} + \binom{0}{n} = F_{n+1}$  for each  $n \geq 0$ .

*Proof.* (Strong Induction) For  $n = 1$  this is  $\binom{1}{0} + \binom{0}{1} = 1 + 0 = 1 = F_2 = F_{1+1}$ . Thus the assertion is true when  $n = 1$ .

Now fix  $n$  and assume that  $\binom{k}{0} + \binom{k-1}{1} + \binom{k-2}{2} + \binom{k-3}{3} + \cdots + \binom{1}{k-1} + \binom{0}{k} = F_{k+1}$  whenever  $k < n$ . In what follows we use the identity  $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ . We also often use  $\binom{a}{b} = 0$  whenever it is untrue that  $0 \leq b \leq a$ .

$$\begin{aligned} & \binom{n}{0} + \binom{n-1}{1} + \binom{n-2}{2} + \cdots + \binom{1}{n-1} + \binom{0}{n} \\ &= \binom{n}{0} + \binom{n-1}{1} + \binom{n-2}{2} + \cdots + \binom{1}{n-1} \\ &= \binom{n-1}{-1} + \binom{n-1}{0} + \binom{n-2}{0} + \binom{n-2}{1} + \binom{n-3}{1} + \binom{n-3}{2} + \cdots + \binom{0}{n-1} + \binom{0}{n} \\ &= \binom{n-1}{0} + \binom{n-2}{0} + \binom{n-2}{1} + \binom{n-3}{1} + \binom{n-3}{2} + \cdots + \binom{0}{n-1} + \binom{0}{n} \\ &= \left[ \binom{n-1}{0} + \binom{n-2}{1} + \cdots + \binom{0}{n-1} \right] + \left[ \binom{n-2}{0} + \binom{n-3}{1} + \cdots + \binom{0}{n-2} \right] \\ &= F_n + F_{n-1} = F_n \end{aligned}$$

This completes the proof. ■

- 31.** Prove that  $\sum_{k=0}^n \binom{k}{r} = \binom{n+1}{r+1}$ , where  $r \in \mathbb{N}$ .

Hint: Use induction on  $n$ . If  $n = 0$ , the equation is  $\binom{0}{r} = \binom{0+1}{r+1}$ , which is  $0 = 0$ .

For the inductive step, we must show that  $\sum_{k=0}^n \binom{k}{r} = \binom{n+1}{r+1}$  implies  $\sum_{k=0}^{n+1} \binom{k}{r} = \binom{(n+1)+1}{r+1}$ .

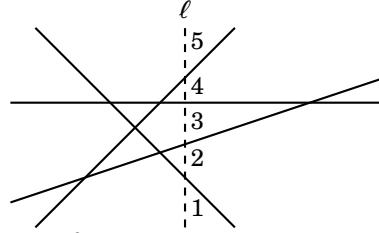
Thus assume  $\sum_{k=0}^n \binom{k}{r} = \binom{n+1}{r+1}$ . By Pascal's formula,  $\binom{(n+1)+1}{r+1} = \binom{n+1}{r} + \binom{n+1}{r+1}$ . Now use the inductive hypothesis and Pascal's formula again to transform this to  $\sum_{k=0}^{n+1} \binom{k}{r}$ .

- 33.** Suppose that  $n$  infinitely long straight lines lie on the plane in such a way that no two are parallel, and no three intersect at a single point. Show that this arrangement divides the plane into  $\frac{n^2+n+2}{2}$  regions.

*Proof.* The proof is by induction. For the basis step, suppose  $n = 1$ . Then there is one line, and it clearly divides the plane into 2 regions, one on either side of the line. As  $2 = \frac{1^2+1+2}{2} = \frac{n^2+n+2}{2}$ , the formula is correct when  $n = 1$ .

Now suppose there are  $n + 1$  lines on the plane, and that the formula is correct for when there are  $n$  lines on the plane. Single out one of the  $n + 1$  lines on the plane, and call it  $\ell$ . Remove line  $\ell$ , so that there are now  $n$  lines on the plane.

By the induction hypothesis, these  $n$  lines divide the plane into  $\frac{n^2+n+2}{2}$  regions. Now add line  $\ell$  back. Doing this adds an additional  $n + 1$  regions. (The diagram illustrates the case where  $n + 1 = 5$ . Without  $\ell$ , there are  $n = 4$  lines. Adding  $\ell$  back produces  $n + 1 = 5$  new regions.)



Thus, with  $n + 1$  lines there are all together  $(n + 1) + \frac{n^2+n+2}{2}$  regions. Observe

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

Thus, with  $n + 1$  lines, we have  $\frac{(n+1)^2+(n+1)+2}{2}$  regions, which means that the formula is true for when there are  $n + 1$  lines. We have shown that if the formula is true for  $n$  lines, it is also true for  $n + 1$  lines. This completes the proof. ■

- 35.** If  $n, k \in \mathbb{N}$ , and  $n$  is even and  $k$  is odd, then  $\binom{n}{k}$  is even.

*Proof.* Notice that if  $k$  is not a value between 0 and  $n$ , then  $\binom{n}{k} = 0$  is even; thus from here on we can assume that  $0 < k < n$ . We will use strong induction.

For the basis case, notice that the assertion is true for the even values  $n = 2$  and  $n = 4$ :  $\binom{2}{1} = 2$ ;  $\binom{4}{1} = 4$ ;  $\binom{4}{3} = 4$  (even in each case).

Now fix and even  $n$  assume that  $\binom{m}{k}$  is even whenever  $m$  is even,  $k$  is odd, and  $m < n$ . Using the identity  $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$  three times, we get

$$\begin{aligned} \binom{n}{k} &= \binom{n-1}{k-1} + \binom{n-1}{k} \\ &= \binom{n-2}{k-2} + \binom{n-2}{k-1} + \binom{n-2}{k-1} + \binom{n-2}{k} \\ &= \binom{n-2}{k-2} + 2\binom{n-2}{k-1} + \binom{n-2}{k}. \end{aligned}$$

Now,  $n - 2$  is even, and  $k$  and  $k - 2$  are odd. By the inductive hypothesis, the outer terms of the above expression are even, and the middle is clearly even; thus we have expressed  $\binom{n}{k}$  as the sum of three even integers, so it is even. ■

- 37.** Prove that if  $m, n \in \mathbb{N}$ , then  $\sum_{k=0}^n k \binom{m+k}{m} = n \binom{m+n+1}{m+1} - \binom{m+n+1}{m+2}$ .

*Proof.* We will use induction on  $n$ . Let  $m$  be any integer.

- (1) If  $n = 1$ , then the equation is  $\sum_{k=0}^1 k \binom{m+k}{m} = 1 \binom{m+1+1}{m+1} - \binom{m+1+1}{m+2}$ , and this is  $0 \binom{m}{m} + 1 \binom{m+1}{m} = 1 \binom{m+2}{m+1} - \binom{m+2}{m+2}$ , which yields the true statement  $m + 1 = m + 2 - 1$ .

- (2) Now let  $n > 1$  and assume the equation holds for  $n$ . (This is the inductive hypothesis.) Now we will confirm that it holds for  $n + 1$ . Observe that

$$\begin{aligned}
 \sum_{k=0}^{n+1} k \binom{m+k}{m} &= && \text{(left-hand side for } n+1\text{)} \\
 \sum_{k=0}^n k \binom{m+k}{m} + (n+1) \binom{m+(n+1)}{m} &= && \text{(split off final term)} \\
 n \binom{m+n+1}{m+1} - \binom{m+n+1}{m+2} + (n+1) \binom{m+n+1}{m} &= && \text{(apply inductive hypothesis)} \\
 n \binom{m+n+1}{m+1} + \binom{m+n+1}{m+1} - \binom{m+n+2}{m+2} + (n+1) \binom{m+n+1}{m} &= && \text{(Pascal's formula)} \\
 (n+1) \binom{m+n+1}{m+1} - \binom{m+n+2}{m+2} + (n+1) \binom{m+n+1}{m} &= && \text{(factor)} \\
 (n+1) \left[ \binom{m+n+1}{m+1} + \binom{m+n+1}{m} \right] - \binom{m+n+2}{m+2} &= && \text{(factor again)} \\
 (n+1) \binom{m+n+2}{m+1} - \binom{m+n+2}{m+2} &= && \text{(Pascal's formula)} \\
 (n+1) \binom{m+(n+1)+1}{m+1} - \binom{m+(n+1)+1}{m+2}. &= && \text{(right-hand side for } n+1\text{)} \blacksquare
 \end{aligned}$$

39. Prove that  $\sum_{k=0}^m \binom{m}{k} \binom{n}{p+k} = \binom{m+n}{m+p}$  for non-negative integers  $m, n$  and  $p$ .

*Proof.* We will use induction on  $n$ . Let  $m$  and  $p$  be any non-negative integers.

- (1) If  $n = 0$ , then the equation is  $\sum_{k=0}^m \binom{m}{k} \binom{0}{p+k} = \binom{m+0}{m+p}$ . This holds if  $p > 0$ , because then  $\binom{0}{p+k} = 0 = \binom{m}{m+p}$ , and both sides of the equation are zero. If  $p = 0$ , the equation is  $\sum_{k=0}^m \binom{m}{k} \binom{0}{k} = \binom{m}{m}$ , and both sides equal 1.
- (2) Now take  $n \geq 1$  and suppose the equation holds for  $n$ . (This is the inductive hypothesis.) Next we confirm that the equation holds for  $n + 1$ .

$$\begin{aligned}
 &\binom{m+(n+1)}{m+p} && \text{(right-hand side for } n+1\text{)} \\
 &= \binom{m+n}{m+(p-1)} + \binom{m+n}{m+p} && \text{(Pascal's formula)} \\
 &= \sum_{k=0}^m \binom{m}{k} \binom{n}{(p-1)+k} + \sum_{k=0}^m \binom{m}{k} \binom{n}{p+k} && \text{(apply inductive hypothesis)} \\
 &= \sum_{k=0}^m \binom{m}{k} \left[ \binom{n}{(p-1)+k} + \binom{n}{p+k} \right] && \text{(combine)} \\
 &= \sum_{k=0}^m \binom{m}{k} \binom{n+1}{p+k} && \text{(Pascal's formula)}
 \end{aligned}$$

This final expression is left-hand side for  $n + 1$ , so the proof is finished. ■

41. If  $n$  and  $k$  are non-negative integers, then  $\binom{n+0}{0} + \binom{n+1}{1} + \binom{n+2}{2} + \cdots + \binom{n+k}{k} = \binom{n+k+1}{k}$ .

*Proof.* We will use induction on  $k$ . Let  $n$  be any non-negative integer.

- (1) If  $k = 0$ , then the equation is  $\binom{n+0}{0} = \binom{n+0+1}{0}$ , which reduces to  $1 = 1$ .
- (2) Assume the equation holds for some  $k \geq 1$ . (This is the inductive hypothesis.) Now we will show that it holds for  $k + 1$ . Note that

$$\begin{aligned} & \binom{n+0}{0} + \binom{n+1}{1} + \binom{n+2}{2} + \cdots + \binom{n+k}{k} + \binom{n+(k+1)}{k+1} && \text{(left side for } k+1\text{)} \\ &= \binom{n+k+1}{k} + \binom{n+k+1}{k+1} && \text{(apply inductive hypothesis)} \\ &= \binom{n+k+2}{k+1} && \text{(Pascal's formula)} \\ &= \binom{n+(k+1)+1}{k+1}. && \text{(right-hand side for } k+1\text{)} \end{aligned}$$

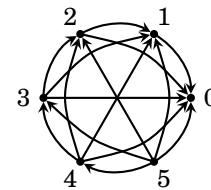
The proof is complete. ■

## Chapter 11 Exercises

### Section 11.1

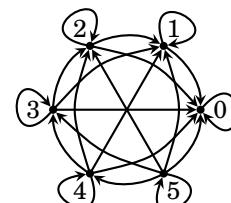
1. Let  $A = \{0, 1, 2, 3, 4, 5\}$ . Write out the relation  $R$  that expresses  $>$  on  $A$ . Then illustrate it with a diagram.

$$R = \{(5, 4), (5, 3), (5, 2), (5, 1), (5, 0), (4, 3), (4, 2), (4, 1), (4, 0), (3, 2), (3, 1), (3, 0), (2, 1), (2, 0), (1, 0)\}$$



3. Let  $A = \{0, 1, 2, 3, 4, 5\}$ . Write out the relation  $R$  that expresses  $\geq$  on  $A$ . Then illustrate it with a diagram.

$$\begin{aligned} R = & \{(5, 5), (5, 4), (5, 3), (5, 2), (5, 1), (5, 0), \\ & (4, 4), (4, 3), (4, 2), (4, 1), (4, 0), \\ & (3, 3), (3, 2), (3, 1), (3, 0), \\ & (2, 2), (2, 1), (2, 0), (1, 1), (1, 0), (0, 0)\} \end{aligned}$$



5. The following diagram represents a relation  $R$  on a set  $A$ . Write the sets  $A$  and  $R$ . Answer:  $A = \{0, 1, 2, 3, 4, 5\}$ ;  $R = \{(3, 3), (4, 3), (4, 2), (1, 2), (2, 5), (5, 0)\}$
7. Write the relation  $<$  on the set  $A = \mathbb{Z}$  as a subset  $R$  of  $\mathbb{Z} \times \mathbb{Z}$ . This is an infinite set, so you will have to use set-builder notation.

Answer:  $R = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : y - x \in \mathbb{N}\}$

9. How many different relations are there on the set  $A = \{1, 2, 3, 4, 5, 6\}$ ?

Consider forming a relation  $R \subseteq A \times A$  on  $A$ . For each ordered pair  $(x, y) \in A \times A$ , we have two choices: we can either include  $(x, y)$  in  $R$  or not include it. There are  $6 \cdot 6 = 36$  ordered pairs in  $A \times A$ . By the multiplication principle, there are thus  $2^{36}$  different subsets  $R$  and hence also this many relations on  $A$ .

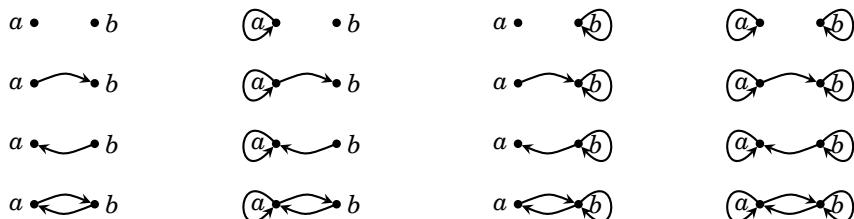
11. Answer:  $2^{|A|^2}$

13. Answer:  $\neq$

15. Answer:  $\equiv (\text{mod } 3)$

## Section 11.2

- Consider the relation  $R = \{(a, a), (b, b), (c, c), (d, d), (a, b), (b, a)\}$  on the set  $A = \{a, b, c, d\}$ . Which of the properties reflexive, symmetric and transitive does  $R$  possess and why? If a property does not hold, say why.  
This **is reflexive** because  $(x, x) \in R$  (i.e.,  $xRx$ ) for every  $x \in A$ .  
It **is symmetric** because it is impossible to find an  $(x, y) \in R$  for which  $(y, x) \notin R$ .  
It **is transitive** because  $(xRy \wedge yRz) \Rightarrow xRz$  always holds.
- Consider the relation  $R = \{(a, b), (a, c), (c, b), (b, c)\}$  on the set  $A = \{a, b, c\}$ . Which of the properties reflexive, symmetric and transitive does  $R$  possess and why? If a property does not hold, say why.  
This **is not reflexive** because  $(a, a) \notin R$  (for example).  
It **is not symmetric** because  $(a, b) \in R$  but  $(b, a) \notin R$ .  
It **is not transitive** because  $cRb$  and  $bRc$  are true, but  $cRc$  is false.
- Consider the relation  $R = \{(0, 0), (\sqrt{2}, 0), (0, \sqrt{2}), (\sqrt{2}, \sqrt{2})\}$  on  $\mathbb{R}$ . Say whether this relation is reflexive, symmetric and transitive. If a property does not hold, say why.  
This **is not reflexive** because  $(1, 1) \notin R$  (for example).  
It **is symmetric** because it is impossible to find an  $(x, y) \in R$  for which  $(y, x) \notin R$ .  
It **is transitive** because  $(xRy \wedge yRz) \Rightarrow xRz$  always holds.
- There are 16 possible different relations  $R$  on the set  $A = \{a, b\}$ . Describe all of them. (A picture for each one will suffice, but don't forget to label the nodes.) Which ones are reflexive? Symmetric? Transitive?



Only the four in the right column are reflexive. Only the eight in the first and fourth rows are symmetric. All of them are transitive **except** the first three on the fourth row.

9. Define a relation on  $\mathbb{Z}$  by declaring  $xRy$  if and only if  $x$  and  $y$  have the same parity. Say whether this relation is reflexive, symmetric and transitive. If a property does not hold, say why. What familiar relation is this?

This is **reflexive** because  $xRx$  since  $x$  always has the same parity as  $x$ .

It is **symmetric** because if  $x$  and  $y$  have the same parity, then  $y$  and  $x$  must have the same parity (that is,  $xRy \Rightarrow yRx$ ).

It is **transitive** because if  $x$  and  $y$  have the same parity and  $y$  and  $z$  have the same parity, then  $x$  and  $z$  must have the same parity. (That is  $(xRy \wedge yRz) \Rightarrow xRz$  always holds.)

The relation is congruence modulo 2.

11. Suppose  $A = \{a, b, c, d\}$  and  $R = \{(a, a), (b, b), (c, c), (d, d)\}$ . Say whether this relation is reflexive, symmetric and transitive. If a property does not hold, say why.

This is **reflexive** because  $(x, x) \in R$  for every  $x \in A$ .

It is **symmetric** because it is impossible to find an  $(x, y) \in R$  for which  $(y, x) \notin R$ .

It is **transitive** because  $(xRy \wedge yRz) \Rightarrow xRz$  always holds.

(For example  $(aRa \wedge aRa) \Rightarrow aRa$  is true, etc.)

13. Consider the relation  $R = \{(x, y) \in \mathbb{R} \times \mathbb{R} : x - y \in \mathbb{Z}\}$  on  $\mathbb{R}$ . Prove that this relation is reflexive and symmetric, and transitive.

*Proof.* In this relation,  $xRy$  means  $x - y \in \mathbb{Z}$ .

To see that  $R$  is reflexive, take any  $x \in \mathbb{R}$  and observe that  $x - x = 0 \in \mathbb{Z}$ , so  $xRx$ . Therefore  $R$  is reflexive.

To see that  $R$  is symmetric, we need to prove  $xRy \Rightarrow yRx$  for all  $x, y \in \mathbb{R}$ . We use direct proof. Suppose  $xRy$ . This means  $x - y \in \mathbb{Z}$ . Then it follows that  $-(x - y) = y - x$  is also in  $\mathbb{Z}$ . But  $y - x \in \mathbb{Z}$  means  $yRx$ . We've shown  $xRy$  implies  $yRx$ , so  $R$  is symmetric.

To see that  $R$  is transitive, we need to prove  $(xRy \wedge yRz) \Rightarrow xRz$  is always true. We prove this conditional statement with direct proof. Suppose  $xRy$  and  $yRz$ . Since  $xRy$ , we know  $x - y \in \mathbb{Z}$ . Since  $yRz$ , we know  $y - z \in \mathbb{Z}$ . Thus  $x - y$  and  $y - z$  are both integers; by adding these integers we get another integer  $(x - y) + (y - z) = x - z$ . Thus  $x - z \in \mathbb{Z}$ , and this means  $xRz$ . We've now shown that if  $xRy$  and  $yRz$ , then  $xRz$ . Therefore  $R$  is transitive. ■

15. Prove or disprove: If a relation is symmetric and transitive, then it is also reflexive.

This is **false**. For a counterexample, consider the relation  $R = \{(a, a), (a, b), (b, a), (b, b)\}$  on the set  $A = \{a, b, c\}$ . This is symmetric and transitive but it is not reflexive.

17. Define a relation  $\sim$  on  $\mathbb{Z}$  as  $x \sim y$  if and only if  $|x - y| \leq 1$ . Say whether  $\sim$  is reflexive, symmetric and transitive.

This is reflexive because  $|x - x| = 0 \leq 1$  for all integers  $x$ . It is symmetric because  $x \sim y$  if and only if  $|x - y| \leq 1$ , if and only if  $|y - x| \leq 1$ , if and only if  $y \sim x$ . It is not transitive because, for example,  $0 \sim 1$  and  $1 \sim 2$ , but is not the case that  $0 \sim 2$ .

### Section 11.3

1. Let  $A = \{1, 2, 3, 4, 5, 6\}$ , and consider the following equivalence relation on  $A$ :  $R = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (2, 3), (3, 2), (4, 5), (5, 4), (4, 6), (6, 4), (5, 6), (6, 5)\}$ . List the equivalence classes of  $R$ .

The equivalence classes are:  $[1] = \{1\}$ ;  $[2] = [3] = \{2, 3\}$ ;  $[4] = [5] = [6] = \{4, 5, 6\}$ .

3. Let  $A = \{a, b, c, d, e\}$ . Suppose  $R$  is an equivalence relation on  $A$ . Suppose  $R$  has three equivalence classes. Also  $aRd$  and  $bRc$ . Write out  $R$  as a set.

Answer:  $R = \{(a, a), (b, b), (c, c), (d, d), (e, e), (a, d), (d, a), (b, c), (c, b)\}$ .

5. There are two equivalence relations on the set  $A = \{a, b\}$ . Describe them.

Answer:  $R = \{(a, a), (b, b)\}$  and  $R = \{(a, a), (b, b), (a, b), (b, a)\}$

7. Define a relation  $R$  on  $\mathbb{Z}$  as  $xRy$  if and only if  $3x - 5y$  is even. Prove  $R$  is an equivalence relation. Describe its equivalence classes.

We must prove that  $R$  is reflexive, symmetric and transitive.

The relation  $R$  is reflexive for the following reason. If  $x \in \mathbb{Z}$ , then  $3x - 5x = -2x$  is even. But then since  $3x - 5x$  is even, we have  $xRx$ . Thus  $R$  is reflexive.

To see that  $R$  is symmetric, suppose  $xRy$ . We must show  $yRx$ . Since  $xRy$ , we know  $3x - 5y$  is even, so  $3x - 5y = 2a$  for some integer  $a$ . Now reason as follows:

$$\begin{aligned} 3x - 5y &= 2a \\ 3x - 5y + 8y - 8x &= 2a + 8y - 8x \\ 3y - 5x &= 2(a + 4y - 4x). \end{aligned}$$

From this it follows that  $3y - 5x$  is even, so  $yRx$ . We've now shown  $xRy$  implies  $yRx$ , so  $R$  is symmetric.

To prove that  $R$  is transitive, assume that  $xRy$  and  $yRz$ . (We will show that this implies  $xRz$ .) Since  $xRy$  and  $yRz$ , it follows that  $3x - 5y$  and  $3y - 5z$  are both even, so  $3x - 5y = 2a$  and  $3y - 5z = 2b$  for some integers  $a$  and  $b$ . Adding these equations, we get  $(3x - 5y) + (3y - 5z) = 2a + 2b$ , and this simplifies to  $3x - 5z = 2(a + b + y)$ . Therefore  $3x - 5z$  is even, so  $xRz$ . We've now shown that if  $xRy$  and  $yRz$ , then  $xRz$ , so  $R$  is transitive.

We've shown  $R$  is reflexive, symmetric and transitive, so it's an equivalence relation.

The completes the first part of the problem. Now we move on the second part. To find the equivalence classes, first note that

$$[0] = \{x \in \mathbb{Z} : xR0\} = \{x \in \mathbb{Z} : 3x - 5 \cdot 0 \text{ is even}\} = \{x \in \mathbb{Z} : 3x \text{ is even}\} = \{x \in \mathbb{Z} : x \text{ is even}\}.$$

Thus the equivalence class  $[0]$  consists of all even integers. Next, note that

$$[1] = \{x \in \mathbb{Z} : xR1\} = \{x \in \mathbb{Z} : 3x - 5 \cdot 1 \text{ is even}\} = \{x \in \mathbb{Z} : 3x - 5 \text{ is even}\} = \{x \in \mathbb{Z} : x \text{ is odd}\}.$$

Thus the equivalence class  $[1]$  consists of all odd integers.

Consequently there are just two equivalence classes  $\{\dots, -4, -2, 0, 2, 4, \dots\}$  and  $\{\dots, -3, -1, 1, 3, 5, \dots\}$ .

9. Define a relation  $R$  on  $\mathbb{Z}$  as  $xRy$  if and only if  $4 \mid (x+3y)$ . Prove  $R$  is an equivalence relation. Describe its equivalence classes.

This is reflexive, because for any  $x \in \mathbb{Z}$  we have  $4 \mid (x+3x)$ , so  $xRx$ .

To prove that  $R$  is symmetric, suppose  $xRy$ . Then  $4 \mid (x+3y)$ , so  $x+3y = 4a$  for some integer  $a$ . Multiplying by 3, we get  $3x+9y = 12a$ , which becomes  $y+3x = 12a - 8y$ . Then  $y+3x = 4(3a - 2y)$ , so  $4 \mid (y+3x)$ , hence  $yRx$ . Thus we've shown  $xRy$  implies  $yRx$ , so  $R$  is symmetric.

To prove transitivity, suppose  $xRy$  and  $yRz$ . Then  $4 \mid (x+3y)$  and  $4 \mid (y+3z)$ , so  $x+3y = 4a$  and  $y+3z = 4b$  for some integers  $a$  and  $b$ . Adding these two equations produces  $x+4y+3z = 4a+4b$ , or  $x+3z = 4a+4b-4y = 4(a+b-y)$ . Consequently  $4 \mid (x+3z)$ , so  $xRz$ , and  $R$  is transitive.

As  $R$  is reflexive, symmetric and transitive, it is an equivalence relation.

Now let's compute its equivalence classes.

$$[0] = \{x \in \mathbb{Z} : xR0\} = \{x \in \mathbb{Z} : 4 \mid (x+3 \cdot 0)\} = \{x \in \mathbb{Z} : 4 \mid x\} = \{\dots -4, 0, 4, 8, 12, 16, \dots\}$$

$$[1] = \{x \in \mathbb{Z} : xR1\} = \{x \in \mathbb{Z} : 4 \mid (x+3 \cdot 1)\} = \{x \in \mathbb{Z} : 4 \mid (x+3)\} = \{\dots -3, 1, 5, 9, 13, 17, \dots\}$$

$$[2] = \{x \in \mathbb{Z} : xR2\} = \{x \in \mathbb{Z} : 4 \mid (x+3 \cdot 2)\} = \{x \in \mathbb{Z} : 4 \mid (x+6)\} = \{\dots -2, 2, 6, 10, 14, 18, \dots\}$$

$$[3] = \{x \in \mathbb{Z} : xR3\} = \{x \in \mathbb{Z} : 4 \mid (x+3 \cdot 3)\} = \{x \in \mathbb{Z} : 4 \mid (x+9)\} = \{\dots -1, 3, 7, 11, 15, 19, \dots\}$$

11. Prove or disprove: If  $R$  is an equivalence relation on an infinite set  $A$ , then  $R$  has infinitely many equivalence classes.

This is **False**. Counterexample: consider the relation of congruence modulo 2. It is a relation on the infinite set  $\mathbb{Z}$ , but it has only two equivalence classes.

13. Answer:  $m|A$

15. Answer: 15

## Section 11.4

1. List all the partitions of the set  $A = \{a, b\}$ . Compare your answer to the answer to Exercise 5 of Section 11.3.

There are just two partitions  $\{\{a\}, \{b\}\}$  and  $\{\{a, b\}\}$ . These correspond to the two equivalence relations  $R_1 = \{(a, a), (b, b)\}$  and  $R_2 = \{(a, a), (a, b), (b, a), (b, b)\}$ , respectively, on  $A$ .

3. Describe the partition of  $\mathbb{Z}$  resulting from the equivalence relation  $\equiv (\text{mod } 4)$ .

Answer: The partition is  $\{[0], [1], [2], [3]\} =$

$$\{\dots, -4, 0, 4, 8, 12, \dots\}, \{\dots, -3, 1, 5, 9, 13, \dots\}, \{\dots, -2, 2, 6, 10, 14, \dots\}, \{\dots, -1, 3, 7, 11, 15, \dots\}\}$$

5. Answer: Congruence modulo 2, or “same parity.”

## Section 11.5

1. Write the addition and multiplication tables for  $\mathbb{Z}_2$ .

+	[0]	[1]
[0]	[0]	[1]
[1]	[1]	[0]

·	[0]	[1]
[0]	[0]	[0]
[1]	[0]	[1]

3. Write the addition and multiplication tables for  $\mathbb{Z}_4$ .

+	[0]	[1]	[2]	[3]		[0]	[1]	[2]	[3]
[0]	[0]	[1]	[2]	[3]	[0]	[0]	[0]	[0]	[0]
[1]	[1]	[2]	[3]	[0]	[1]	[0]	[1]	[2]	[3]
[2]	[2]	[3]	[0]	[1]	[2]	[0]	[2]	[0]	[2]
[3]	[3]	[0]	[1]	[2]	[3]	[0]	[3]	[2]	[1]

5. Suppose  $[a], [b] \in \mathbb{Z}_5$  and  $[a] \cdot [b] = [0]$ . Is it necessarily true that either  $[a] = [0]$  or  $[b] = [0]$ ?

The multiplication table for  $\mathbb{Z}_5$  is shown in Section 11.5. In the body of that table, the only place that [0] occurs is in the first row or the first column. That row and column are both headed by [0]. It follows that if  $[a] \cdot [b] = [0]$ , then either  $[a]$  or  $[b]$  must be [0].

7. Do the following calculations in  $\mathbb{Z}_9$ , in each case expressing your answer as  $[a]$  with  $0 \leq a \leq 8$ .

(a)  $[8] + [8] = [7]$       (b)  $[24] + [11] = [8]$       (c)  $[21] \cdot [15] = [0]$       (d)  $[8] \cdot [8] = [1]$

## Chapter 12

### Section 12.1

1. Suppose  $A = \{0, 1, 2, 3, 4\}$ ,  $B = \{2, 3, 4, 5\}$  and  $f = \{(0, 3), (1, 3), (2, 4), (3, 2), (4, 2)\}$ . State the domain and range of  $f$ . Find  $f(2)$  and  $f(1)$ .

Domain is  $A$ ; Range is  $\{2, 3, 4\}$ ;  $f(2) = 4$ ;  $f(1) = 3$ .

3. There are four different functions  $f : \{a, b\} \rightarrow \{0, 1\}$ . List them all.

$f_1 = \{(a, 0), (b, 0)\}$     $f_2 = \{(a, 1), (b, 0)\}$ ,    $f_3 = \{(a, 0), (b, 1)\}$     $f_4 = \{(a, 1), (b, 1)\}$

5. Give an example of a relation from  $\{a, b, c, d\}$  to  $\{d, e\}$  that is not a function.

One example is  $\{(a, d), (a, e), (b, d), (c, d), (d, d)\}$ .

7. Consider the set  $f = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} : 3x + y = 4\}$ . Is this a function from  $\mathbb{Z}$  to  $\mathbb{Z}$ ? Explain.

Yes, since  $3x + y = 4$  if and only if  $y = 4 - 3x$ , this is the function  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  defined as  $f(x) = 4 - 3x$ .

9. Consider the set  $f = \{(x^2, x) : x \in \mathbb{R}\}$ . Is this a function from  $\mathbb{R}$  to  $\mathbb{R}$ ? Explain.

No. This is not a function. Observe that  $f$  contains the ordered pairs  $(4, 2)$  and  $(4, -2)$ . Thus the real number 4 occurs as the first coordinate of more than one element of  $f$ .

11. Is the set  $\theta = \{(X, |X|) : X \subseteq \mathbb{Z}_5\}$  a function? If so, what is its domain and range?

Yes, this is a function. The domain is  $\mathcal{P}(\mathbb{Z}_5)$ . The range is  $\{0, 1, 2, 3, 4, 5\}$ .

## Section 12.2

1. Let  $A = \{1, 2, 3, 4\}$  and  $B = \{a, b, c\}$ . Give an example of a function  $f : A \rightarrow B$  that is neither injective nor surjective.

Consider  $f = \{(1, a), (2, a), (3, a), (4, a)\}$ .

Then  $f$  is not injective because  $f(1) = f(2)$ .

Also  $f$  is not surjective because it sends no element of  $A$  to the element  $c \in B$ .

3. Consider the cosine function  $\cos : \mathbb{R} \rightarrow \mathbb{R}$ . Decide whether this function is injective and whether it is surjective. What if it had been defined as  $\cos : \mathbb{R} \rightarrow [-1, 1]$ ?

The function  $\cos : \mathbb{R} \rightarrow \mathbb{R}$  is **not injective** because, for example,  $\cos(0) = \cos(2\pi)$ . It is **not surjective** because if  $b = 5 \in \mathbb{R}$  (for example), there is no real number for which  $\cos(x) = b$ . The function  $\cos : \mathbb{R} \rightarrow [-1, 1]$  is **surjective but not injective**.

5. A function  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  is defined as  $f(n) = 2n + 1$ . Verify whether this function is injective and whether it is surjective.

**This function is injective.** To see this, suppose  $m, n \in \mathbb{Z}$  and  $f(m) = f(n)$ .

This means  $2m + 1 = 2n + 1$ , from which we get  $2m = 2n$ , and then  $m = n$ .

Thus  $f$  is injective.

**This function is not surjective.** To see this notice that  $f(n)$  is odd for all  $n \in \mathbb{Z}$ . So given the (even) number 2 in the codomain  $\mathbb{Z}$ , there is no  $n$  with  $f(n) = 2$ .

7. A function  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$  is defined as  $f((m, n)) = 2n - 4m$ . Verify whether this function is injective and whether it is surjective.

This is **not injective** because  $(0, 2) \neq (-1, 0)$ , yet  $f((0, 2)) = f((-1, 0)) = 4$ . This is **not surjective** because  $f((m, n)) = 2n - 4m = 2(n - 2m)$  is always even. If  $b \in \mathbb{Z}$  is odd, then  $f((m, n)) \neq b$ , for all  $(m, n) \in \mathbb{Z} \times \mathbb{Z}$ .

9. Prove that the function  $f : \mathbb{R} - \{2\} \rightarrow \mathbb{R} - \{5\}$  defined by  $f(x) = \frac{5x+1}{x-2}$  is bijective.

*Proof.* First, let's check that  $f$  is injective. Suppose  $f(x) = f(y)$ . Then

$$\begin{aligned} \frac{5x+1}{x-2} &= \frac{5y+1}{y-2} \\ (5x+1)(y-2) &= (5y+1)(x-2) \\ 5xy - 10x + y - 2 &= 5yx - 10y + x - 2 \\ -10x + y &= -10y + x \\ 11y &= 11x \\ y &= x. \end{aligned}$$

Since  $f(x) = f(y)$  implies  $x = y$ , it follows that  $f$  is injective.

Next we check that  $f$  is surjective. Take an arbitrary element  $b \in \mathbb{R} - \{5\}$ . We seek an  $x \in \mathbb{R} - \{2\}$  for which  $f(x) = b$ , or  $\frac{5x+1}{x-2} = b$ . Solving this for  $x$ , we get:

$$\begin{aligned} 5x+1 &= b(x-2) \\ 5x+1 &= bx-2b \\ 5x-xb &= -2b-1 \\ x(5-b) &= -2b-1. \end{aligned}$$

Since we have assumed  $b \in \mathbb{R} - \{5\}$ , the term  $(5 - b)$  is not zero, and we can divide with impunity to get  $x = \frac{-2b - 1}{5 - b}$ . This is an  $x$  for which  $f(x) = b$ , so  $f$  is surjective. Since  $f$  is both injective and surjective, it is bijective. ■

11. Consider the function  $\theta : \{0, 1\} \times \mathbb{N} \rightarrow \mathbb{Z}$  defined as  $\theta(a, b) = (-1)^a b$ . Is  $\theta$  injective? Is it surjective? Explain.

First we show that  $\theta$  is injective. Suppose  $\theta(a, b) = \theta(c, d)$ . Then  $(-1)^a b = (-1)^c d$ . As  $b$  and  $d$  are both in  $\mathbb{N}$ , they are both positive. Then because  $(-1)^a b = (-1)^c d$ , it follows that  $(-1)^a$  and  $(-1)^c$  have the same sign. Since each of  $(-1)^a$  and  $(-1)^c$  equals  $\pm 1$ , we have  $(-1)^a = (-1)^c$ , so then  $(-1)^a b = (-1)^c d$  implies  $b = d$ . But also  $(-1)^a = (-1)^c$  means  $a$  and  $c$  have the same parity, and because  $a, c \in \{0, 1\}$ , it follows  $a = c$ . Thus  $(a, b) = (c, d)$ , so  $\theta$  is injective.

Next note that  $\theta$  is **not surjective** because  $\theta(a, b) = (-1)^a b$  is either positive or negative, but never zero. Therefore there exist no element  $(a, b) \in \{0, 1\} \times \mathbb{N}$  for which  $\theta(a, b) = 0 \in \mathbb{Z}$ .

13. Consider the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  defined by the formula  $f(x, y) = (xy, x^3)$ . Is  $f$  injective? Is it surjective?

Notice that  $f(0, 1) = (0, 0)$  and  $f(0, 0) = (0, 0)$ , so  $f$  is **not injective**. To show that  $f$  is also **not surjective**, we will show that it's impossible to find an ordered pair  $(x, y)$  with  $f(x, y) = (1, 0)$ . If there were such a pair, then  $f(x, y) = (xy, x^3) = (1, 0)$ , which yields  $xy = 1$  and  $x^3 = 0$ . From  $x^3 = 0$  we get  $x = 0$ , so  $xy = 0$ , a contradiction.

15. This question concerns functions  $f : \{A, B, C, D, E, F, G\} \rightarrow \{1, 2, 3, 4, 5, 6, 7\}$ . How many such functions are there? How many of these functions are injective? How many are surjective? How many are bijective?

Function  $f$  can be described as a list  $(f(A), f(B), f(C), f(D), f(E), f(F), f(G))$ , where there are seven choices for each entry. By the multiplication principle, the total number of functions  $f$  is  $7^7 = 823543$ .

If  $f$  is injective, then this list can't have any repetition, so there are  $7! = 5040$  injective functions. Since any injective function sends the seven elements of the domain to seven distinct elements of the codomain, all of the injective functions are surjective, and vice versa. Thus there are 5040 surjective functions and 5040 bijective functions.

17. This question concerns functions  $f : \{A, B, C, D, E, F, G\} \rightarrow \{1, 2\}$ . How many such functions are there? How many of these functions are injective? How many are surjective? How many are bijective?

Function  $f$  can be described as a list  $(f(A), f(B), f(C), f(D), f(E), f(F), f(G))$ , where there are two choices for each entry. Therefore the total number of functions is  $2^7 = 128$ . It is impossible for any function to send all seven elements of  $\{A, B, C, D, E, F, G\}$  to seven distinct elements of  $\{1, 2\}$ , so none of these 128 functions is injective, hence none are bijective.

How many are surjective? Only two of the 128 functions are not surjective, and they are the “constant” functions  $\{(A, 1), (B, 1), (C, 1), (D, 1), (E, 1), (F, 1), (G, 1)\}$  and  $\{(A, 2), (B, 2), (C, 2), (D, 2), (E, 2), (F, 2), (G, 2)\}$ . So there are 126 surjective functions.

### Section 12.3

1. If 6 integers are chosen at random, at least two will have the same remainder when divided by 5.

*Proof.* Write  $\mathbb{Z}$  as follows:  $\mathbb{Z} = \bigcup_{j=0}^4 \{5k + j : k \in \mathbb{Z}\}$ . This is a partition of  $\mathbb{Z}$  into 5 sets. If six integers are picked at random, by the pigeonhole principle, at least two will be in the same set. However, each set corresponds to the remainder of a number after being divided by 5 (for example,  $\{5k + 1 : k \in \mathbb{Z}\}$  are all those integers that leave a remainder of 1 after being divided by 5). ■

3. Given any six positive integers, there are two for which their sum or difference is divisible by 9.

*Proof.* Let  $A$  be a set of six positive integers. Let  $B = \{\{0\}, \{1, 8\}, \{2, 7\}, \{3, 6\}, \{4, 5\}\}$ . Notice that every element of  $B$  is a set that either has one element or has two elements whose sum is 9. Define  $f : A \rightarrow B$  so that  $f(x)$  is the set that contains the remainder when  $x$  is divided by 9. For example,  $f(12) = \{3, 6\}$  and  $f(18) = \{0\}$ . Since  $6 = |A| > |B| = 5$ , the pigeonhole principle implies that  $f$  is not injective. Thus there exist  $x, y \in A$  for which  $f(x) = f(y)$ . Then either  $x$  and  $y$  have the same remainder  $r$  when divided by 9, or the remainders  $r$  and  $s$  add to 9. In the first case  $x = 9m + r$  and  $y = 9n + r$  (for  $m, n \in \mathbb{Z}$ ), so  $x - y = 9(m - n)$  is divisible by 9. In the second case  $x = 9m + r$  and  $y = 9n + s$ , so  $x + y = 9m + 9n + r + s = 9(m + n + 1)$  is divisible by 9. ■

5. Prove that any set of 7 distinct natural numbers contains a pair of numbers whose sum or difference is divisible by 10.

*Proof.* Let  $S = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$  be any set of 7 natural numbers. Let's say that  $a_1 < a_2 < a_3 < \dots < a_7$ . Consider the set

$$\begin{aligned} A = & \{a_1 - a_2, a_1 - a_3, a_1 - a_4, a_1 - a_5, a_1 - a_6, a_1 - a_7, \\ & a_1 + a_2, a_1 + a_3, a_1 + a_4, a_1 + a_5, a_1 + a_6, a_1 + a_7\} \end{aligned}$$

Thus  $|A| = 12$ . Now let  $B = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ , so  $|B| = 10$ . Let  $f : A \rightarrow B$  be the function for which  $f(n)$  equals the last digit of  $n$ . (That is  $f(97) = 7$ ,  $f(12) = 2$ ,  $f(230) = 0$ , etc.) Then, since  $|A| > |B|$ , the pigeonhole principle guarantees that  $f$  is not injective. Thus  $A$  contains elements  $a_1 \pm a_i$  and  $a_1 \pm a_j$  for which  $f(a_1 \pm a_i) = f(a_1 \pm a_j)$ . This means the last digit of  $a_1 \pm a_i$  is the same as the last digit of  $a_1 \pm a_j$ . Thus the last digit of the difference  $(a_1 \pm a_i) - (a_1 \pm a_j) = \pm a_i \pm a_j$  is 0. Hence  $\pm a_i \pm a_j$  is a sum or difference of elements of  $S$  that is divisible by 10. ■

### Section 12.4

1. Suppose  $A = \{5, 6, 8\}$ ,  $B = \{0, 1\}$ ,  $C = \{1, 2, 3\}$ . Let  $f : A \rightarrow B$  be the function  $f = \{(5, 1), (6, 0), (8, 1)\}$ , and  $g : B \rightarrow C$  be  $g = \{(0, 1), (1, 1)\}$ . Find  $g \circ f$ .  
 $g \circ f = \{(5, 1), (6, 1), (8, 1)\}$

3. Suppose  $A = \{1, 2, 3\}$ . Let  $f : A \rightarrow A$  be the function  $f = \{(1, 2), (2, 2), (3, 1)\}$ , and let  $g : A \rightarrow A$  be the function  $g = \{(1, 3), (2, 1), (3, 2)\}$ . Find  $g \circ f$  and  $f \circ g$ .
- $$g \circ f = \{(1, 1), (2, 1), (3, 3)\}; \quad f \circ g = \{(1, 1), (2, 2), (3, 2)\}.$$
5. Consider the functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = \sqrt[3]{x+1}$  and  $g(x) = x^3$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
- $$g \circ f(x) = x + 1; \quad f \circ g(x) = \sqrt[3]{x^3 + 1}$$
7. Consider the functions  $f, g : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined as  $f(m, n) = (mn, m^2)$  and  $g(m, n) = (m+1, m+n)$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
- Note  $g \circ f(m, n) = g(f(m, n)) = g(mn, m^2) = (mn+1, mn+m^2)$ .
- Thus  $\boxed{g \circ f(m, n) = (mn+1, mn+m^2)}$ .
- Note  $f \circ g(m, n) = f(g(m, n)) = f(m+1, m+n) = ((m+1)(m+n), (m+1)^2)$ .
- Thus  $\boxed{f \circ g(m, n) = (m^2 + mn + m + n, m^2 + 2m + 1)}$ .
9. Consider the functions  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$  defined as  $f(m, n) = m+n$  and  $g : \mathbb{Z} \rightarrow \mathbb{Z} \times \mathbb{Z}$  defined as  $g(m) = (m, m)$ . Find the formulas for  $g \circ f$  and  $f \circ g$ .
- $$g \circ f(m, n) = (m+n, m+n)$$
- $$f \circ g(m) = 2m$$

## Section 12.5

1. Check that  $f : \mathbb{Z} \rightarrow \mathbb{Z}$  defined by  $f(n) = 6 - n$  is bijective. Then compute  $f^{-1}$ .
- It is injective as follows. Suppose  $f(m) = f(n)$ . Then  $6 - m = 6 - n$ , which reduces to  $m = n$ .
- It is surjective as follows. If  $b \in \mathbb{Z}$ , then  $f(6 - b) = 6 - (6 - b) = b$ .
- Inverse:  $f^{-1}(n) = 6 - n$ .
3. Let  $B = \{2^n : n \in \mathbb{Z}\} = \{\dots, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, \dots\}$ . Show that the function  $f : \mathbb{Z} \rightarrow B$  defined as  $f(n) = 2^n$  is bijective. Then find  $f^{-1}$ .
- It is injective as follows. Suppose  $f(m) = f(n)$ , which means  $2^m = 2^n$ . Taking  $\log_2$  of both sides gives  $\log_2(2^m) = \log_2(2^n)$ , which simplifies to  $m = n$ .
- The function  $f$  is surjective as follows. Suppose  $b \in B$ . By definition of  $B$  this means  $b = 2^n$  for some  $n \in \mathbb{Z}$ . Then  $f(n) = 2^n = b$ .
- Inverse:  $f^{-1}(n) = \log_2(n)$ .
5. The function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = \pi x - e$  is bijective. Find its inverse.
- Inverse:  $f^{-1}(x) = \frac{x + e}{\pi}$ .
7. Show that the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  defined by the formula  $f((x, y)) = ((x^2 + 1)y, x^3)$  is bijective. Then find its inverse.
- First we prove the function is injective. Assume  $f(x_1, y_1) = f(x_2, y_2)$ . Then  $(x_1^2 + 1)y_1 = (x_2^2 + 1)y_2$  and  $x_1^3 = x_2^3$ . Since the real-valued function  $f(x) = x^3$  is one-to-one, it follows that  $x_1 = x_2$ . Since  $x_1 = x_2$ , and  $x_1^2 + 1 > 0$  we may divide both sides of  $(x_1^2 + 1)y_1 = (x_2^2 + 1)y_2$  by  $(x_1^2 + 1)$  to get  $y_1 = y_2$ . Hence  $(x_1, y_1) = (x_2, y_2)$ .
- Now we prove the function is surjective. Let  $(a, b) \in \mathbb{R}^2$ . Set  $x = b^{1/3}$  and  $y = a/(b^{2/3} + 1)$ . Then  $f(x, y) = ((b^{2/3} + 1)\frac{a}{b^{2/3} + 1}, (b^{1/3})^3) = (a, b)$ . It now follows that  $f$  is bijective.

Finally, we compute the inverse. Write  $f(x, y) = (u, v)$ . Interchange variables to get  $(x, y) = f(u, v) = ((u^2 + 1)v, u^3)$ . Thus  $x = (u^2 + 1)v$  and  $y = u^3$ . Hence  $u = y^{1/3}$  and  $v = \frac{x}{y^{2/3} + 1}$ . Therefore  $f^{-1}(x, y) = (u, v) = \left(y^{1/3}, \frac{x}{y^{2/3} + 1}\right)$ .

- 9.** Consider the function  $f : \mathbb{R} \times \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{R}$  defined as  $f(x, y) = (y, 3xy)$ . Check that this is bijective; find its inverse.

To see that this is injective, suppose  $f(a, b) = f(c, d)$ . This means  $(b, 3ab) = (d, 3cd)$ . Since the first coordinates must be equal, we get  $b = d$ . As the second coordinates are equal, we get  $3ab = 3cd$ , which becomes  $3ab = 3bc$ . Note that, from the definition of  $f$ ,  $b \in \mathbb{N}$ , so  $b \neq 0$ . Thus we can divide both sides of  $3ab = 3bc$  by the non-zero quantity  $3b$  to get  $a = c$ . Now we have  $a = c$  and  $b = d$ , so  $(a, b) = (c, d)$ . It follows that  $f$  is injective.

Next we check that  $f$  is surjective. Given any  $(b, c)$  in the codomain  $\mathbb{N} \times \mathbb{R}$ , notice that  $(\frac{c}{3b}, b)$  belongs to the domain  $\mathbb{R} \times \mathbb{N}$ , and  $f(\frac{c}{3b}, b) = (b, c)$ . Thus  $f$  is surjective. As it is both injective and surjective, it is bijective; thus the inverse exists.

To find the inverse, recall that we obtained  $f(\frac{c}{3b}, b) = (b, c)$ . Then  $f^{-1}f(\frac{c}{3b}, b) = f^{-1}(b, c)$ , which reduces to  $(\frac{c}{3b}, b) = f^{-1}(b, c)$ . Replacing  $b$  and  $c$  with  $x$  and  $y$ , respectively, we get  $f^{-1}(x, y) = (\frac{y}{3x}, x)$ .

## Section 12.6

- Consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined as  $f(x) = x^2 + 3$ . Find  $f([-3, 5])$  and  $f^{-1}([12, 19])$ . Answers:  $f([-3, 5]) = [3, 28]$ ;  $f^{-1}([12, 19]) = [-4, -3] \cup [3, 4]$ .
- This problem concerns functions  $f : \{1, 2, 3, 4, 5, 6, 7\} \rightarrow \{0, 1, 2, 3, 4\}$ . How many such functions have the property that  $|f^{-1}(\{3\})| = 3$ ? Answer:  $4^4 \binom{7}{3}$ .
- Consider a function  $f : A \rightarrow B$  and a subset  $X \subseteq A$ . We observed in Section 12.6 that  $f^{-1}(f(X)) \neq X$  in general. However  $X \subseteq f^{-1}(f(X))$  is always true. Prove this.

*Proof.* Suppose  $a \in X$ . Thus  $f(a) \in \{f(x) : x \in X\} = f(X)$ , that is  $f(a) \in f(X)$ . Now, by definition of preimage, we have  $f^{-1}(f(X)) = \{x \in A : f(x) \in f(X)\}$ . Since  $a \in A$  and  $f(a) \in f(X)$ , it follows that  $a \in f^{-1}(f(X))$ . This proves  $X \subseteq f^{-1}(f(X))$ . ■

- Given a function  $f : A \rightarrow B$  and subsets  $W, X \subseteq A$ , prove  $f(W \cap X) \subseteq f(W) \cap f(X)$ .

*Proof.* Suppose  $b \in f(W \cap X)$ . This means  $b \in \{f(x) : x \in W \cap X\}$ , that is  $b = f(a)$  for some  $a \in W \cap X$ . Since  $a \in W$  we have  $b = f(a) \in \{f(x) : x \in W\} = f(W)$ . Since  $a \in X$  we have  $b = f(a) \in \{f(x) : x \in X\} = f(X)$ . Thus  $b$  is in both  $f(W)$  and  $f(X)$ , so  $b \in f(W) \cap f(X)$ . This completes the proof that  $f(W \cap X) \subseteq f(W) \cap f(X)$ . ■

- Given a function  $f : A \rightarrow B$  and subsets  $W, X \subseteq A$ , prove  $f(W \cup X) = f(W) \cup f(X)$ .

*Proof.* First we will show  $f(W \cup X) \subseteq f(W) \cup f(X)$ . Suppose  $b \in f(W \cup X)$ . This means  $b \in \{f(x) : x \in W \cup X\}$ , that is,  $b = f(a)$  for some  $a \in W \cup X$ . Thus  $a \in W$  or  $a \in X$ . If  $a \in W$ , then  $b = f(a) \in \{f(x) : x \in W\} = f(W)$ . If  $a \in X$ , then  $b = f(a) \in \{f(x) : x \in X\} = f(X)$ . Thus  $b$  is in  $f(W)$  or  $f(X)$ , so  $b \in f(W) \cup f(X)$ . This completes the proof that  $f(W \cup X) \subseteq f(W) \cup f(X)$ .

Next we will show  $f(W) \cup f(X) \subseteq f(W \cup X)$ . Suppose  $b \in f(W) \cup f(X)$ . This means  $b \in f(W)$  or  $b \in f(X)$ . If  $b \in f(W)$ , then  $b = f(a)$  for some  $a \in W$ . If  $b \in f(X)$ , then  $b = f(a)$  for some  $a \in X$ . Either way,  $b = f(a)$  for some  $a$  that is in  $W$  or  $X$ . That is,  $b = f(a)$  for some  $a \in W \cup X$ . But this means  $b \in f(W \cup X)$ . This completes the proof that  $f(W) \cup f(X) \subseteq f(W \cup X)$ .

The previous two paragraphs show  $f(W \cup X) = f(W) \cup f(X)$ . ■

- 11.** Given  $f : A \rightarrow B$  and subsets  $Y, Z \subseteq B$ , prove  $f^{-1}(Y \cup Z) = f^{-1}(Y) \cup f^{-1}(Z)$ .

*Proof.* First we will show  $f^{-1}(Y \cup Z) \subseteq f^{-1}(Y) \cup f^{-1}(Z)$ . Suppose  $a \in f^{-1}(Y \cup Z)$ . By Definition 12.9, this means  $f(a) \in Y \cup Z$ . Thus,  $f(a) \in Y$  or  $f(a) \in Z$ . If  $f(a) \in Y$ , then  $a \in f^{-1}(Y)$ , by Definition 12.9. Similarly, if  $f(a) \in Z$ , then  $a \in f^{-1}(Z)$ . Hence  $a \in f^{-1}(Y)$  or  $a \in f^{-1}(Z)$ , so  $a \in f^{-1}(Y) \cup f^{-1}(Z)$ . Consequently  $f^{-1}(Y \cup Z) \subseteq f^{-1}(Y) \cup f^{-1}(Z)$ .

Next we show  $f^{-1}(Y) \cup f^{-1}(Z) \subseteq f^{-1}(Y \cup Z)$ . Suppose  $a \in f^{-1}(Y) \cup f^{-1}(Z)$ . This means  $a \in f^{-1}(Y)$  or  $a \in f^{-1}(Z)$ . Hence, by Definition 12.9,  $f(a) \in Y$  or  $f(a) \in Z$ , which means  $f(a) \in Y \cup Z$ . But by Definition 12.9,  $f(a) \in Y \cup Z$  means  $a \in f^{-1}(Y \cup Z)$ . Consequently  $f^{-1}(Y) \cup f^{-1}(Z) \subseteq f^{-1}(Y \cup Z)$ .

The previous two paragraphs show  $f^{-1}(Y \cup Z) = f^{-1}(Y) \cup f^{-1}(Z)$ . ■

- 13.** Let  $f : A \rightarrow B$  be a function, and  $X \subseteq A$ . Prove or disprove:  $f(f^{-1}(f(X))) = f(X)$ .

*Proof.* First we will show  $f(f^{-1}(f(X))) \subseteq f(X)$ . Suppose  $y \in f(f^{-1}(f(X)))$ . By definition of image, this means  $y = f(x)$  for some  $x \in f^{-1}(f(X))$ . But by definition of preimage,  $x \in f^{-1}(f(X))$  means  $f(x) \in f(X)$ . Thus we have  $y = f(x) \in f(X)$ , as desired.

Next we show  $f(X) \subseteq f(f^{-1}(f(X)))$ . Suppose  $y \in f(X)$ . This means  $y = f(x)$  for some  $x \in X$ . Then  $f(x) = y \in f(X)$ , which means  $x \in f^{-1}(f(X))$ . Then by definition of image,  $f(x) \in f(f^{-1}(f(X)))$ . Now we have  $y = f(x) \in f(f^{-1}(f(X)))$ , as desired.

The previous two paragraphs show  $f(f^{-1}(f(X))) = f(X)$ . ■

## Chapter 13 Exercises

### Section 13.2

- 1.** Prove that  $\lim_{x \rightarrow 5} (8x - 3) = 37$ .

*Proof.* Take  $\varepsilon > 0$ . Note that  $|(8x - 3) - 37| = |8x - 40| = |8(x - 5)| = 8|x - 5|$ . So if  $\delta = \frac{\varepsilon}{8}$ , then  $0 < |x - 5| < \delta$  implies  $|(8x - 3) - 37| = 8|x - 5| < 8\delta = 8\frac{\varepsilon}{8} = \varepsilon$ . By Definition 13.2,  $\lim_{x \rightarrow 5} (8x - 3) = 37$ . ■

- 3.** Prove that  $\lim_{x \rightarrow 0} (x + 2) = 2$ .

*Proof.* Given  $\varepsilon > 0$ , let  $\delta = \varepsilon$ . Then  $0 < |x - 0| < \delta$  implies  $|(x + 2) - 2| = |x - 0| < \delta = \varepsilon$ . By Definition 13.2,  $\lim_{x \rightarrow 0} (x + 2) = 2$ . ■

5. Prove that  $\lim_{x \rightarrow 3} (x^2 - 2) = 7$ .

*Proof.* Suppose  $\varepsilon > 0$ . In what follows we will produce a corresponding  $\delta$  for which  $0 < |x - 3| < \delta$  implies  $|x^2 - 2 - 7| < \varepsilon$ . Notice that

$$|(x^2 - 2) - 7| = |x^2 - 9| = |(x - 3)(x + 3)| = |x - 3| \cdot |x + 3|.$$

If  $|x - 3| \leq 1$ , then  $|x + 3| = |(x - 3) + 6| \leq |x - 3| + 6 \leq 1 + 6 = 7$  (using the inequality (13.2) from page 245). So if  $|x - 3| \leq 1$ , then  $|x + 3| \leq 7$  and the above equation yields

$$|(x^2 - 2) - 7| = |x - 3| \cdot |x + 3| < |x - 3| \cdot 7 = 7|x - 3|.$$

Take  $\delta$  to be smaller than both 1 and  $\frac{\varepsilon}{7}$ . Then  $0 < |x - 3| < \delta$  implies  $|x^2 - 2 - 7| < 7 \cdot |x - 3| < 7\delta < 7\frac{\varepsilon}{7} = \varepsilon$ . By Definition 13.2, we have  $\lim_{x \rightarrow 3} (x^2 - 2) = 7$ . ■

### Section 13.3

1. Prove that  $\lim_{x \rightarrow 0} \log_{10} |x|$  does not exist.

*Proof.* Suppose for the sake of contradiction that  $\lim_{x \rightarrow 0} \log_{10} |x| = L$ , for some  $L \in \mathbb{R}$ .

Let  $\varepsilon = 1$ , so there is a  $\delta > 0$  for which  $0 < |x - 0| < \delta$  implies  $|\log_{10}(|x|) - L| < 1$ . Choose an  $x \neq 0$  for which  $|x|$  is smaller than both  $\delta$  and  $10^{L-1}$ . Then  $0 < |x - 0| < \delta$ , so  $|\log_{10}(|x|) - L| < 1$ . But also  $|x| < 10^{L-1}$ , so  $\log_{10} |x| < L - 1$ . Consequently  $\log_{10} |x| - L < -1$ , and thus  $|\log_{10} |x| - L| > 1$ . This is a contradiction. ■

3. Prove that  $\lim_{x \rightarrow 0} \frac{1}{x^2}$  does not exist.

*Proof.* Suppose for the sake of contradiction that  $\lim_{x \rightarrow 0} \frac{1}{x^2} = L$ , for some  $L \in \mathbb{R}$ . Fix an  $\varepsilon > 0$  for which  $L + \varepsilon > 0$ . Choose a real number  $\delta > 0$  for which  $0 < |x - 0| < \delta$  implies  $|\frac{1}{x^2} - L| < \varepsilon$ . Choose an  $x > 0$  that is smaller than both  $\delta$  and  $\sqrt{\frac{1}{L+\varepsilon}}$ . Then  $0 < |x - 0| < \delta$ , so  $|\frac{1}{x^2} - L| < \varepsilon$ . But also,  $x < \sqrt{\frac{1}{L+\varepsilon}}$ , so  $x^2 < \frac{1}{L+\varepsilon}$  and hence  $\frac{1}{x^2} > L + \varepsilon$ . Consequently  $\frac{1}{x^2} - L > \varepsilon$ , and thus  $|\frac{1}{x^2} - L| > \varepsilon$ . This is a contradiction. ■

5. Prove that  $\lim_{x \rightarrow 0} x \cot(\frac{1}{x})$  does not exist.

*Proof.* Note that  $\cot(x) = \frac{1}{\sin(x)}$ . Because  $\sin(k\pi) = 0$  for any  $k \in \mathbb{Z}$ , it follows that  $\cot(x)$  is undefined for any  $x = k\pi$ . Hence  $x \cot(\frac{1}{x})$  is undefined for any  $x = \frac{1}{k\pi}$ . Given any  $\delta > 0$ , there exist values of  $x = \frac{1}{k\pi}$  that satisfy  $0 < |x - 0| < \delta$ . The statement  $(0 < |x - 0| < \delta) \Rightarrow |x \cot(\frac{1}{x}) - L| < \varepsilon$  is meaningless for such  $x$ , so the limit cannot exist. (See the remark following Example 13.5 on page 250.) ■

### Section 13.4

1. Given two or more functions  $f_1, f_2, \dots, f_n$ , suppose that  $\lim_{x \rightarrow c} f_i(x)$  exists for each  $1 \leq i \leq n$ . Prove that  $\lim_{x \rightarrow c} (f_1(x) + f_2(x) + \dots + f_n(x)) = \lim_{x \rightarrow c} f_1(x) + \lim_{x \rightarrow c} f_2(x) + \dots + \lim_{x \rightarrow c} f_n(x)$ .

*Proof.* The proof is by induction. For the basis case  $n = 2$ , and the result follows from the multiplication rule (Theorem 13.7).

Now let  $k > 2$  assume that the theorem holds for  $k$  functions  $f_1, f_2, \dots, f_k$ . That is,  $\lim_{x \rightarrow c} (f_1(x) + f_2(x) + \dots + f_k(x)) = \lim_{x \rightarrow c} f_1(x) + \lim_{x \rightarrow c} f_2(x) + \dots + \lim_{x \rightarrow c} f_k(x)$ . We must show  $\lim_{x \rightarrow c} (f_1(x) + f_2(x) + \dots + f_k(x) + f_{k+1}(x)) = \lim_{x \rightarrow c} f_1(x) + \lim_{x \rightarrow c} f_2(x) + \dots + \lim_{x \rightarrow c} f_k(x) + \lim_{x \rightarrow c} f_{k+1}(x)$ . Just note that

$$\begin{aligned} & \lim_{x \rightarrow c} (f_1(x) + f_2(x) + \dots + f_k(x) + f_{k+1}(x)) \\ &= \lim_{x \rightarrow c} ((f_1(x) + f_2(x) + \dots + f_k(x)) + f_{k+1}(x)) && \text{(group)} \\ &= \lim_{x \rightarrow c} (f_1(x) + f_2(x) + \dots + f_k(x)) + \lim_{x \rightarrow c} f_{k+1}(x) && \text{(Theorem 13.7)} \\ &= \lim_{x \rightarrow c} f_1(x) + \lim_{x \rightarrow c} f_2(x) + \dots + \lim_{x \rightarrow c} f_k(x) + \lim_{x \rightarrow c} f_{k+1}(x) && \text{(inductive hypothesis).} \end{aligned}$$

This completes the proof by induction. ■

3. Use the previous two exercises and the constant multiple rule (Theorem 13.4) to prove that if  $f(x)$  is a polynomial, then  $\lim_{x \rightarrow c} f(x) = f(c)$  for any  $c \in \mathbb{R}$ .

*Proof.* First note that by Exercise 2 and the identity function rule, we have  $\lim_{x \rightarrow c} x^n = \lim_{x \rightarrow c} (x \cdot x \cdots x) = (\lim_{x \rightarrow c} x) \cdot (\lim_{x \rightarrow c} x) \cdots (\lim_{x \rightarrow c} x) = c \cdot c \cdots c = c^n$ . Thus  $\lim_{x \rightarrow c} x^n = c^n$ .

Now consider an arbitrary polynomial  $f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$ , where each  $a_i$  is a constant real number. Then

$$\begin{aligned} \lim_{x \rightarrow c} f(x) &= \lim_{x \rightarrow c} (a_0 + a_1x + a_2x^2 + \dots + a_nx^n) \\ &= \lim_{x \rightarrow c} a_0 + \lim_{x \rightarrow c} a_1x + \lim_{x \rightarrow c} a_2x^2 + \dots + \lim_{x \rightarrow c} a_nx^n && \text{(Exercise 1)} \\ &= \lim_{x \rightarrow c} a_0 + a_1 \lim_{x \rightarrow c} x + a_2 \lim_{x \rightarrow c} x^2 + \dots + a_n \lim_{x \rightarrow c} x^n && \text{(constant multiple rule)} \\ &= a_0 + a_1c + a_2c^2 + \dots + a_nc^n = f(c). \end{aligned}$$

5. Prove that if  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M$ , then  $L = M$ .

*Proof.* Suppose  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M$ . Then by limit laws,  $L - M = \lim_{x \rightarrow c} f(x) - \lim_{x \rightarrow c} g(x) = \lim_{x \rightarrow c} (f(x) - g(x)) = \lim_{x \rightarrow c} 0 = 0$ . This shows  $L - M = 0$ , so  $L = M$ . ■

## Section 13.5

1. Prove that the function  $f(x) = \sqrt{x}$  is continuous at any number  $c > 0$ . Deduce that  $\lim_{x \rightarrow c} \sqrt{g(x)} = \sqrt{\lim_{x \rightarrow c} g(x)}$ , provided  $\lim_{x \rightarrow c} g(x)$  exists and is greater than zero.

*Proof.* Suppose  $c > 0$ . Proving  $\sqrt{x}$  is continuous at  $c$  amounts to proving that  $\lim_{x \rightarrow c} \sqrt{x} = \sqrt{c}$ . Here is a proof of this limit: For any  $\varepsilon > 0$  let  $\delta$  be smaller than both  $c$  and  $\varepsilon\sqrt{c}$ . Now suppose  $0 < |x - c| < \delta$ . Because  $\delta < c$  it follows that  $|x - c| < c$ , and

hence  $-c < x - c < c$ . From this,  $0 < x$ , so  $\sqrt{x}$  exists. Also, because  $\delta < \epsilon\sqrt{c}$ , we have

$$\begin{aligned} |\sqrt{x} - \sqrt{c}| &= \left| (\sqrt{x} - \sqrt{c}) \frac{\sqrt{x} + \sqrt{c}}{\sqrt{x} + \sqrt{c}} \right| = \left| (x - c) \frac{1}{\sqrt{x} + \sqrt{c}} \right| = |x - c| \frac{1}{\sqrt{x} + \sqrt{c}} \\ &< |x - c| \cdot \frac{1}{\sqrt{c}} < \delta \frac{1}{\sqrt{c}} = \epsilon \sqrt{c} \frac{1}{\sqrt{c}} = \epsilon. \end{aligned}$$

(Note: above we used the fact  $\sqrt{x} + \sqrt{c} > \sqrt{c}$  to get  $\frac{1}{\sqrt{x}-\sqrt{c}} < \frac{1}{\sqrt{c}}$ .) We have now shown that  $0 < |x - c| < \delta$  implies  $|\sqrt{x} - \sqrt{c}| < \epsilon$ , so  $\lim_{x \rightarrow c} \sqrt{x} = \sqrt{c}$ . This means  $\sqrt{x}$  is continuous at any number  $x = c$ , by Definition 13.3.

Applying Theorem 13.9, we get  $\lim_{x \rightarrow c} \sqrt{g(x)} = \sqrt{\lim_{x \rightarrow c} g(x)}$ . ■

## Section 13.6

1. If  $n \in \mathbb{N}$ , then  $\lim_{x \rightarrow \infty} \frac{1}{x^n} = 0$ .

*Proof.* Suppose  $\epsilon > 0$ . Let  $N = \frac{1}{\sqrt[n]{\epsilon}}$ . If  $x > N$ , then  $x^n > N^n = \frac{1}{\epsilon}$ , so  $0 < \frac{1}{x^n} < \epsilon$ . Thus  $|\frac{1}{x^n} - 0| = |\frac{1}{x^n}| < \epsilon$ . In summary,  $x > N$  implies  $|\frac{1}{x^n} - 0| < \epsilon$ , so  $\lim_{x \rightarrow \infty} \frac{1}{x^n} = 0$  by Definition 13.4. ■

3. If  $a \in \mathbb{R}$ , then  $\lim_{x \rightarrow \infty} a = a$ .

*Proof.* Suppose  $\epsilon > 0$ . Let  $N = 1$ . Then  $x > N$  implies  $|a - a| < 0$ , which means  $\lim_{x \rightarrow \infty} a = a$ . (Note: The implication  $x > N \Rightarrow |a - a| < \epsilon$  is actually true no matter what value  $x$  has, because  $|a - a| < \epsilon$  is automatically true.) ■

5. If both  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow \infty} g(x)$  exist, then  $\lim_{x \rightarrow \infty} (f(x) + g(x)) = \lim_{x \rightarrow \infty} f(x) + \lim_{x \rightarrow \infty} g(x)$ .

*Proof.* Say  $\lim_{x \rightarrow \infty} f(x) = L$  and  $\lim_{x \rightarrow \infty} g(x) = M$ . We must prove  $\lim_{x \rightarrow \infty} (f(x) + g(x)) = L + M$ . Take  $\epsilon > 0$ . We need to find an  $N$  for which  $x > N$  implies  $|(f(x) + g(x)) - (L + M)| < \epsilon$ . Because  $\lim_{x \rightarrow \infty} f(x) = L$ , there is a  $N' > 0$  such that  $x > N'$  implies  $|f(x) - L| < \frac{\epsilon}{2}$ . Because  $\lim_{x \rightarrow \infty} g(x) = M$ , there is a  $N'' > 0$  such that  $x > N''$  implies  $|g(x) - M| < \frac{\epsilon}{2}$ . Put  $N = \max\{N', N''\}$ . If  $x > N$ , then

$$|(f(x) + g(x)) - (L + M)| = |(f(x) - L) + (g(x) - M)| \leq |f(x) - L| + |g(x) - M| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

We've now shown that for any  $\epsilon > 0$ , there is a  $N > 0$  for which  $x > N$  implies  $|(f(x) + g(x)) - (L + M)| < \epsilon$ . Thus  $\lim_{x \rightarrow \infty} (f(x) + g(x)) = L + M$ . ■

7. If both  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow \infty} g(x)$  exist, then  $\lim_{x \rightarrow \infty} (f(x) - g(x)) = \lim_{x \rightarrow \infty} f(x) - \lim_{x \rightarrow \infty} g(x)$ .

*Proof.* Suppose both  $\lim_{x \rightarrow \infty} f(x)$  and  $\lim_{x \rightarrow \infty} g(x)$  exist. Using exercises 4 and 5 above,

$$\begin{aligned} \lim_{x \rightarrow \infty} (f(x) - g(x)) &= \lim_{x \rightarrow \infty} (f(x) + (-1) \cdot g(x)) = \lim_{x \rightarrow \infty} f(x) + \lim_{x \rightarrow \infty} (-1) \cdot g(x) \\ &= \lim_{x \rightarrow \infty} f(x) + (-1) \cdot \lim_{x \rightarrow \infty} g(x) = \lim_{x \rightarrow \infty} f(x) - \lim_{x \rightarrow \infty} g(x). \end{aligned}$$

9. If  $\lim_{x \rightarrow \infty} g(x) = L$  and  $f$  is continuous at  $x = L$ , then  $\lim_{x \rightarrow \infty} f(g(x)) = f\left(\lim_{x \rightarrow \infty} g(x)\right)$ .

*Proof.* Suppose  $\lim_{x \rightarrow \infty} g(x) = L$  and  $f$  is continuous at  $x = L$ . We need to prove  $\lim_{x \rightarrow \infty} f(g(x)) = f(L)$ . Definition 13.4 says we must prove that for any  $\varepsilon > 0$ , there is a corresponding  $N > 0$  for which  $x > N$  implies  $|f(g(x)) - f(L)| < \varepsilon$ .

So let  $\varepsilon > 0$ . As  $f$  is continuous at  $L$ , Definition 13.3 yields  $\lim_{x \rightarrow L} f(x) = f(L)$ . From this, we know there is a real number  $\delta > 0$  for which

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

But also, from  $\lim_{x \rightarrow \infty} g(x) = L$ , we know that there is a real number  $N > 0$  for which  $x > N$  implies  $|g(x) - L| < \delta$ . If  $x > N$ , then we have  $|g(x) - L| < \delta$ , and from this  $(*)$  yields  $|f(g(x)) - f(L)| < \varepsilon$ . Thus  $\lim_{x \rightarrow \infty} f(g(x)) = f(L)$ , and the proof is complete. ■

## Section 13.7

1. Prove that  $\left\{\frac{2^n}{n!}\right\}$  converges to 0.

*Proof.* Observe that  $0 < \frac{2^n}{n!} < \frac{4}{n}$  for any  $n \in \mathbb{N}$  because

$$\begin{aligned} \frac{2^n}{n!} &= \frac{2 \cdot 2 \cdot 2 \cdots 2 \cdot 2 \cdot 2}{n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1} = \frac{2}{n} \cdot \frac{2}{n-1} \cdot \frac{2}{n-2} \cdots \frac{2}{3} \cdot \frac{2}{2} \cdot \frac{2}{1} \\ &\leq \frac{2}{n} \cdot 1 \cdot 1 \cdots 1 \cdot 1 \cdot 2 = \frac{4}{n}. \end{aligned}$$

Thus  $\left|\frac{2^n}{n!}\right| < \frac{4}{n}$  for any  $n \in \mathbb{N}$ . Given  $\varepsilon > 0$ , choose an integer  $N > \frac{4}{\varepsilon}$ . If  $n > N$ , then  $\left|\frac{2^n}{n!} - 0\right| = \left|\frac{2^n}{n!}\right| < \frac{4}{n} < \frac{4}{N} < \frac{4}{4/\varepsilon} = \varepsilon$ . By Definition 13.5,  $\left\{\frac{2^n}{n!}\right\}$  converges to 0. ■

3. Prove that  $\left\{\frac{2n^2+1}{3n-1}\right\}$  diverges to  $\infty$ .

*Proof.* Note that  $\frac{2n^2+1}{3n-1} > \frac{2n^2}{3n-1} > \frac{2n^2}{3n} = \frac{2n}{3}$ . For any  $L > 0$ , let  $N = \frac{3L}{2}$ . Then for  $n > N$  we have  $\frac{2n^2+1}{3n-1} > \frac{2n}{3} > \frac{2N}{3} = L$ . By Definition 13.6, the sequence diverges to  $\infty$ . ■

5. Prove that  $\left\{\frac{2n+1}{3n-1}\right\}$  converges to  $\frac{2}{3}$ .

*Proof.* For  $n \geq 1$  we have  $\left|\frac{2n+1}{3n-1} - \frac{2}{3}\right| = \left|\frac{3(2n+1)}{3(3n-1)} - \frac{2(3n-1)}{3(3n-1)}\right| = \left|\frac{5}{9n-3}\right| = \frac{5}{9n-3}$ . Given  $\varepsilon > 0$ , we will have  $\frac{5}{9n-3} < \varepsilon$  provided that  $\frac{9n-3}{5} > \frac{1}{\varepsilon}$ , or  $n > \frac{5}{9\varepsilon} + \frac{1}{3}$ .

Therefore, given any  $\varepsilon > 0$ , take an integer  $N > \frac{5}{9\varepsilon} + \frac{1}{3}$ . If  $n > N$ , then  $\left|\frac{2n+1}{3n-1} - \frac{2}{3}\right| = \frac{5}{9n-3} < \frac{5}{9N-3} < \frac{5}{9(\frac{5}{9\varepsilon} + \frac{1}{3})-3} = \varepsilon$ . By Definition 13.5,  $\left\{\frac{2n+1}{3n-1}\right\}$  converges to  $\frac{2}{3}$ . ■

7. Prove that if a sequence diverges to infinity, then it diverges.

*Proof.* For the sake of contradiction, suppose that  $\{a_n\}$  diverges to  $\infty$ , and  $\{a_n\}$  converges to a number  $L$ . Definition 13.5 says that for  $\varepsilon = 1$  there is a number

$N > 0$  for which  $n > N$  implies  $|a_n - L| < 1$ . Also, Definition 13.6 guarantees an  $N' > 0$  for which  $n > N'$  implies  $a_n > L + 1$ , that is,  $a_n - L > 1$ .

Let  $n$  be larger than both  $N$  and  $N'$ . Then  $|a_n - L| < 1$  and  $a_n - L > 1$ . Thus  $|a_n - L| < 1$  and  $|a_n - L| > 1$ , a contradiction. ■

9. Prove that if  $\{a_n\}$  converges to  $L$ , and  $c \in \mathbb{R}$ , then  $\{ca_n\}$  converges to  $cL$ .

*Proof.* Suppose  $\{a_n\}$  converges to  $L$ , and  $c \in \mathbb{R}$ . If  $c = 0$ , then  $\{ca_n\}$  is the sequence  $0, 0, 0, \dots$ , and this converges to  $0 = cL$ . Thus the theorem is true if  $c = 0$ , so for the remainder of the proof we treat the case  $c \neq 0$ .

Let  $\varepsilon > 0$ . Because  $\{a_n\}$  converges to  $L$ , there exists an  $N > 0$  for which  $n > N$  implies  $|a_n - L| < \frac{\varepsilon}{|c|}$ . So if  $n > N$ , then  $|ca_n - cL| = |c(a_n - L)| = |c| \cdot |a_n - L| < |c| \frac{\varepsilon}{|c|} = \varepsilon$ . In summary, we've shown that for any  $\varepsilon > 0$ , there is a  $N > 0$  for which  $n > N$  implies  $|ca_n - cL| < \varepsilon$ . By Definition 13.5,  $\{ca_n\}$  converges to  $cL$ . ■

11. Prove that if  $\{a_n\}$  converges to  $L$  and  $\{b_n\}$  converges to  $M$ , then the sequence  $\{a_n b_n\}$  converges to  $LM$ .

*Proof.* Suppose  $\{a_n\}$  converges to  $L$  and  $\{b_n\}$  converges to  $M$ . We must prove  $\{a_n b_n\}$  converges to  $LM$ . To prove this, take  $\varepsilon > 0$ . We need to find an  $N$  for which  $n > N$  implies  $|a_n b_n - LM| < \varepsilon$ . With this in mind, notice that

$$\begin{aligned} |a_n b_n - LM| &= |(a_n b_n - Lb_n) + (Lb_n - LM)| \\ &\leq |a_n b_n - Lb_n| + |Lb_n - LM| \\ &= |(a_n - L)b_n| + |L(b_n - M)| \\ &= |a_n - L| \cdot |b_n| + |L| \cdot |b_n - M|. \end{aligned} \tag{*}$$

Take  $N' > 0$  large enough so that  $n > N'$  implies  $|b_n - M| < 1$ . If  $n > N'$ , then

$$|b_n| = |(b_n - M) + M| \leq |b_n - M| + |M| < 1 + |M|.$$

Replacing  $|b_n|$  in  $(*)$  with the larger quantity  $1 + |M|$ , we get

$$|a_n b_n - LM| < |a_n - L| \cdot (1 + |M|) + |L| \cdot |b_n - M| \tag{**}$$

for all  $n > N'$ . Now take  $N'' > 0$  such that  $n > N''$  implies  $|a_n - L| < \frac{\varepsilon}{2(1+|M|)}$ . Take  $N''' > 0$  such that  $n > N'''$  implies  $|b_n - M| < \frac{\varepsilon}{2|L|}$ . Put  $N = \max\{N', N'', N'''\}$ . If  $n > N$ , then  $(**)$  becomes

$$|a_n b_n - LM| < \frac{\varepsilon}{2(1+|M|)} \cdot (1 + |M|) + |L| \cdot \frac{\varepsilon}{2|L|} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

To summarize, we've shown that for any  $\varepsilon > 0$ , there is a  $N > 0$  for which  $n > N$  implies  $|a_n b_n - LM| < \varepsilon$ . Therefore  $\{a_n b_n\}$  converges to  $LM$ . ■

13. Prove that if  $\{|a_n|\}$  converges to 0, then  $\{a_n\}$  converges to 0. Give an example of a sequence  $\{a_n\}$  for which  $\{|a_n|\}$  converges to a number  $L \neq 0$ , but  $\{a_n\}$  diverges.

*Proof.* Suppose  $\{|a_n|\}$  converges to 0. This means that for any  $\varepsilon > 0$ , there is an  $N > 0$  for which  $n > N$  implies  $||a_n| - 0| < \varepsilon$ . But  $||a_n| - 0| = |a_n - 0|$ . Thus  $n > N$  implies  $|a_n - 0| < \varepsilon$ . Therefore  $\{a_n\}$  converges to 0.

Consider the sequence  $\{(-1)^n\}$ , which is  $-1, 1, -1, 1, -1, \dots$ . This sequence diverges. But  $\{|(-1)^n|\}$  is the sequence  $1, 1, 1, 1, \dots$ , which converges to 1. ■

## Chapter 14 Exercises

### Section 14.1

#### 1. $\mathbb{R}$ and $(0, \infty)$

Observe that the function  $f(x) = e^x$  sends  $\mathbb{R}$  to  $(0, \infty)$ . It is injective because  $f(x) = f(y)$  implies  $e^x = e^y$ , and taking  $\ln$  of both sides gives  $x = y$ . It is surjective because if  $b \in (0, \infty)$ , then  $f(\ln(b)) = b$ . Therefore, because of the bijection  $f : \mathbb{R} \rightarrow (0, \infty)$ , it follows that  $|\mathbb{R}| = |(0, \infty)|$ .

#### 3. $\mathbb{R}$ and $(0, 1)$

Observe that the function  $\frac{1}{\pi}f(x) = \cot^{-1}(x)$  sends  $\mathbb{R}$  to  $(0, 1)$ . It is injective and surjective by elementary trigonometry. Therefore, because of the bijection  $f : \mathbb{R} \rightarrow (0, 1)$ , it follows that  $|\mathbb{R}| = |(0, 1)|$ .

#### 5. $A = \{3k : k \in \mathbb{Z}\}$ and $B = \{7k : k \in \mathbb{Z}\}$

Observe that the function  $f(x) = \frac{7}{3}x$  sends  $A$  to  $B$ . It is injective because  $f(x) = f(y)$  implies  $\frac{7}{3}x = \frac{7}{3}y$ , and multiplying both sides by  $\frac{3}{7}$  gives  $x = y$ . It is surjective because if  $b \in B$ , then  $b = 7k$  for some integer  $k$ . Then  $3k \in A$ , and  $f(3k) = 7k = b$ . Therefore, because of the bijection  $f : A \rightarrow B$ , it follows that  $|A| = |B|$ .

#### 7. $\mathbb{Z}$ and $S = \{\dots, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, \dots\}$

Observe that the function  $f : \mathbb{Z} \rightarrow S$  defined as  $f(n) = 2^n$  is bijective: It is injective because  $f(m) = f(n)$  implies  $2^m = 2^n$ , and taking  $\log_2$  of both sides produces  $m = n$ . It is surjective because any element  $b$  of  $S$  has form  $b = 2^n$  for some integer  $n$ , and therefore  $f(n) = 2^n = b$ . Because of the bijection  $f : \mathbb{Z} \rightarrow S$ , it follows that  $|\mathbb{Z}| = |S|$ .

#### 9. $\{0, 1\} \times \mathbb{N}$ and $\mathbb{N}$

Consider the function  $f : \{0, 1\} \times \mathbb{N} \rightarrow \mathbb{N}$  defined as  $f(a, n) = 2n - a$ . This is injective because if  $f(a, n) = f(b, m)$ , then  $2n - a = 2m - b$ . Now if  $a$  were unequal to  $b$ , one of  $a$  or  $b$  would be 0 and the other would be 1, and one side of  $2n - a = 2m - b$  would be odd and the other even, a contradiction. Therefore  $a = b$ . Then  $2n - a = 2m - b$  becomes  $2n - a = 2m - a$ ; add  $a$  to both sides and divide by 2 to get  $m = n$ . Thus we have  $a = b$  and  $m = n$ , so  $(a, n) = (b, m)$ , so  $f$  is injective. To see that  $f$  is surjective, take any  $b \in \mathbb{N}$ . If  $b$  is even, then  $b = 2n$  for some integer  $n$ , and  $f(0, n) = 2n - 0 = b$ . If  $b$  is odd, then  $b = 2n+1$  for some integer  $n$ . Then  $f(1, n+1) = 2(n+1) - 1 = 2n+1 = b$ . Therefore  $f$  is surjective. Then  $f$  is a bijection, so  $|\{0, 1\} \times \mathbb{N}| = |\mathbb{N}|$ .

#### 11. $[0, 1]$ and $(0, 1)$

*Proof.* Consider the subset  $X = \left\{ \frac{1}{n} : n \in \mathbb{N} \right\} \subseteq [0, 1]$ . Let  $f : [0, 1] \rightarrow (0, 1)$  be defined as  $f(x) = x$  if  $x \in [0, 1] - X$  and  $f\left(\frac{1}{n}\right) = \frac{1}{n+1}$  for any  $\frac{1}{n} \in X$ . It is easy to check that  $f$  is a bijection. Next let  $Y = \left\{ 1 - \frac{1}{n} : n \in \mathbb{N} \right\} \subseteq [0, 1]$ , and define  $g : [0, 1] \rightarrow (0, 1)$  as

$g(x) = x$  if  $x \in [0, 1] - Y$  and  $g(1 - \frac{1}{n}) = 1 - \frac{1}{n+1}$  for any  $1 - \frac{1}{n} \in Y$ . As in the case of  $f$ , it is easy to check that  $g$  is a bijection. Therefore the composition  $g \circ f : [0, 1] \rightarrow (0, 1)$  is a bijection. (See Theorem 12.2.) We conclude that  $|[0, 1]| = |(0, 1)|$ . ■

### 13. $\mathcal{P}(\mathbb{N})$ and $\mathcal{P}(\mathbb{Z})$

Outline: By Exercise 18 of Section 12.2, we have a bijection  $f : \mathbb{N} \rightarrow \mathbb{Z}$  defined as  $f(n) = \frac{(-1)^n(2n-1)+1}{4}$ . Now define a function  $\Phi : \mathcal{P}(\mathbb{N}) \rightarrow \mathcal{P}(\mathbb{Z})$  as  $\Phi(X) = \{f(x) : x \in X\}$ . Check that  $\Phi$  is a bijection.

### 15. Find a formula for the bijection $f$ in Example 14.2.

Hint: Consider the function  $f$  from Exercise 18 of Section 12.2.

## Section 14.2

### 1. Prove that the set $A = \{\ln(n) : n \in \mathbb{N}\} \subseteq \mathbb{R}$ is countably infinite.

Note that its elements can be written in infinite list form as  $\ln(1), \ln(2), \ln(3), \dots$ . Thus  $A$  is countably infinite.

### 3. Prove that the set $A = \{(5n, -3n) : n \in \mathbb{Z}\}$ is countably infinite.

Consider the function  $f : \mathbb{Z} \rightarrow A$  defined as  $f(n) = (5n, -3n)$ . This is clearly surjective, and it is injective because  $f(n) = f(m)$  gives  $(5n, -3n) = (5m, -3m)$ , so  $5n = 5m$ , hence  $m = n$ . Thus, because  $f$  is surjective,  $|\mathbb{Z}| = |A|$ , and  $|A| = |\mathbb{Z}| = \aleph_0$ . Therefore  $A$  is countably infinite.

### 5. Prove or disprove: There exists a countably infinite subset of the set of irrational numbers.

This is true. Just consider the set consisting of the irrational numbers  $\frac{\pi}{1}, \frac{\pi}{2}, \frac{\pi}{3}, \frac{\pi}{4}, \dots$ .

### 7. Prove or disprove: The set $\mathbb{Q}^{100}$ is countably infinite.

This is true. Note  $\mathbb{Q}^{100} = \mathbb{Q} \times \mathbb{Q} \times \dots \times \mathbb{Q}$  (100 times), and since  $\mathbb{Q}$  is countably infinite, it follows from the corollary of Theorem 14.5 that this product is countably infinite.

### 9. Prove or disprove: The set $\{0, 1\} \times \mathbb{N}$ is countably infinite.

This is true. Note that  $\{0, 1\} \times \mathbb{N}$  can be written in infinite list form as  $(0, 1), (1, 1), (0, 2), (1, 2), (0, 3), (1, 3), (0, 4), (1, 4), \dots$ . Thus the set is countably infinite.

### 11. Partition $\mathbb{N}$ into 8 countably infinite sets.

For each  $i \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ , let  $X_i$  be those natural numbers that are congruent to  $i$  modulo 8, that is,

$$\begin{aligned} X_1 &= \{1, 9, 17, 25, 33, \dots\} \\ X_2 &= \{2, 10, 18, 26, 34, \dots\} \\ X_3 &= \{3, 11, 19, 27, 35, \dots\} \\ X_4 &= \{4, 12, 20, 28, 36, \dots\} \\ X_5 &= \{5, 13, 21, 29, 37, \dots\} \\ X_6 &= \{6, 14, 22, 30, 38, \dots\} \end{aligned}$$

$$\begin{aligned} X_7 &= \{7, 15, 13, 31, 39, \dots\} \\ X_8 &= \{8, 16, 24, 32, 40, \dots\} \end{aligned}$$

13. If  $A = \{X \subset \mathbb{N} : X \text{ is finite}\}$ , then  $|A| = \aleph_0$ .

*Proof.* This is **true**. To show this we will describe how to arrange the items of  $A$  in an infinite list  $X_1, X_2, X_3, X_4, \dots$

For each natural number  $n$ , let  $p_n$  be the  $n$ th prime number. Thus  $p_1 = 2$ ,  $p_2 = 3$ ,  $p_3 = 5$ ,  $p_4 = 7$ ,  $p_5 = 11$ , and so on. Now consider any element  $X \in A$ . If  $X \neq \emptyset$ , then  $X = \{n_1, n_2, n_3, \dots, n_k\}$ , where  $k = |X|$  and  $n_i \in \mathbb{N}$  for each  $1 \leq i \leq k$ . Define a function  $f : A \rightarrow \mathbb{N} \cup \{0\}$  as follows:  $f(\{n_1, n_2, n_3, \dots, n_k\}) = p_{n_1}p_{n_2} \cdots p_{n_k}$ . For example,  $f(\{1, 2, 3\}) = p_1p_2p_3 = 2 \cdot 3 \cdot 5 = 30$ , and  $f(\{3, 5\}) = p_3p_5 = 5 \cdot 11 = 55$ , etc. Also, we should not forget that  $\emptyset \in A$ , and we define  $f(\emptyset) = 0$ .

Note  $f : A \rightarrow \mathbb{N} \cup \{0\}$  is injective: Let  $X = \{n_1, n_2, n_3, \dots, n_k\}$  and  $Y = \{m_1, m_2, m_3, \dots, m_\ell\}$ , and  $X \neq Y$ . Then there is an integer  $a$  that belongs to one of  $X$  or  $Y$  but not the other. Then the prime factorization of one of the numbers  $f(X)$  and  $f(Y)$  uses the prime number  $p_a$  but the prime factorization of the other does not use  $p_a$ . It follows that  $f(X) \neq f(Y)$  by the fundamental theorem of arithmetic. Thus  $f$  is injective.

So each set  $X \in A$  is associated with an integer  $f(X) \geq 0$ , and no two different sets are associated with the same number. Thus we can list the elements in  $X \in A$  in increasing order of the numbers  $f(X)$ . The list begins as

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{4\}, \{1, 3\}, \{5\}, \{6\}, \{1, 4\}, \{2, 3\}, \{7\}, \dots$$

It follows that  $A$  is countably infinite. ■

15. Hint: Use the fundamental theorem of arithmetic.

### Section 14.3

1. Suppose  $B$  is an uncountable set and  $A$  is a set. Given that there is a surjective function  $f : A \rightarrow B$ , what can be said about the cardinality of  $A$ ?

The set  $A$  must be uncountable, as follows. For each  $b \in B$ , let  $a_b$  be an element of  $A$  for which  $f(a_b) = b$ . (Such an element must exist because  $f$  is surjective.) Now form the set  $U = \{a_b : b \in B\}$ . Then the function  $f : U \rightarrow B$  is bijective, by construction. Then since  $B$  is uncountable, so is  $U$ . Therefore  $U$  is an uncountable subset of  $A$ , so  $A$  is uncountable by Theorem 14.9.

3. Prove or disprove: If  $A$  is uncountable, then  $|A| = |\mathbb{R}|$ .

This is false. Let  $A = \mathcal{P}(\mathbb{R})$ . Then  $A$  is uncountable, and by Theorem 14.7,  $|\mathbb{R}| < |\mathcal{P}(\mathbb{R})| = |A|$ .

5. Prove or disprove: The set  $\{0, 1\} \times \mathbb{R}$  is uncountable.

This is true. To see why, first note that the function  $f : \mathbb{R} \rightarrow \{0\} \times \mathbb{R}$  defined as  $f(x) = (0, x)$  is a bijection. Thus  $|\mathbb{R}| = |\{0\} \times \mathbb{R}|$ , and since  $\mathbb{R}$  is uncountable, so is

$\{0\} \times \mathbb{R}$ . Then  $\{0\} \times \mathbb{R}$  is an uncountable subset of the set  $\{0, 1\} \times \mathbb{R}$ , so  $\{0, 1\} \times \mathbb{R}$  is uncountable by Theorem 14.9.

7. Prove or disprove: If  $A \subseteq B$  and  $A$  is countably infinite and  $B$  is uncountable, then  $B - A$  is uncountable.

This is true. To see why, suppose to the contrary that  $B - A$  is countably infinite. Then  $B = A \cup (B - A)$  is a union of countably infinite sets, and thus countable, by Theorem 14.6. This contradicts the fact that  $B$  is uncountable.

## Section 14.4

1. Show that if  $A \subseteq B$  and there is an injection  $g : B \rightarrow A$ , then  $|A| = |B|$ .

Just note that the map  $f : A \rightarrow B$  defined as  $f(x) = x$  is an injection. Now apply the Cantor-Bernstein-Schröeder theorem.

3. Let  $\mathcal{F}$  be the set of all functions  $\mathbb{N} \rightarrow \{0, 1\}$ . Show that  $|\mathbb{R}| = |\mathcal{F}|$ .

Because  $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$ , it suffices to show that  $|\mathcal{F}| = |\mathcal{P}(\mathbb{N})|$ . To do this, we will exhibit a bijection  $f : \mathcal{F} \rightarrow \mathcal{P}(\mathbb{N})$ . Define  $f$  as follows. Given a function  $\varphi \in \mathcal{F}$ , let  $f(\varphi) = \{n \in \mathbb{N} : \varphi(n) = 1\}$ . To see that  $f$  is injective, suppose  $f(\varphi) = f(\theta)$ . Then  $\{n \in \mathbb{N} : \varphi(n) = 1\} = \{n \in \mathbb{N} : \theta(n) = 1\}$ . Put  $X = \{n \in \mathbb{N} : \varphi(n) = 1\}$ . Now we see that if  $n \in X$ , then  $\varphi(n) = 1 = \theta(n)$ . And if  $n \in \mathbb{N} - X$ , then  $\varphi(n) = 0 = \theta(n)$ . Consequently  $\varphi(n) = \theta(n)$  for any  $n \in \mathbb{N}$ , so  $\varphi = \theta$ . Thus  $f$  is injective. To see that  $f$  is surjective, take any  $X \in \mathcal{P}(\mathbb{N})$ . Consider the function  $\varphi \in \mathcal{F}$  for which  $\varphi(n) = 1$  if  $n \in X$  and  $\varphi(n) = 0$  if  $n \notin X$ . Then  $f(\varphi) = X$ , so  $f$  is surjective.

5. Consider the subset  $B = \{(x, y) : x^2 + y^2 \leq 1\} \subseteq \mathbb{R}^2$ . Show that  $|B| = |\mathbb{R}^2|$ .

This will follow from the Cantor-Bernstein-Schröeder theorem provided that we can find injections  $f : B \rightarrow \mathbb{R}^2$  and  $g : \mathbb{R}^2 \rightarrow B$ . The function  $f : B \rightarrow \mathbb{R}^2$  defined as  $f(x, y) = (x, y)$  is clearly injective. For  $g : \mathbb{R}^2 \rightarrow B$ , consider the function

$$g(x, y) = \left( \frac{x}{\sqrt{x^2 + y^2 + 1}}, \frac{y}{\sqrt{x^2 + y^2 + 1}} \right).$$

Verify that this is an injective function  $g : \mathbb{R}^2 \rightarrow B$ .

7. Prove or disprove: If there is a injection  $f : A \rightarrow B$  and a surjection  $g : A \rightarrow B$ , then there is a bijection  $h : A \rightarrow B$ .

This is true. Here is an outline of a proof. Define a function  $g' : B \rightarrow A$  as follows. For each  $b \in B$ , choose an element  $x_b \in g^{-1}(\{b\})$ . (That is, choose an element  $x_b \in A$  for which  $g(x_b) = b$ .) Now let  $g' : B \rightarrow A$  be the function defined as  $g'(b) = x_b$ . Check that  $g'$  is injective and apply the the Cantor-Bernstein-Schröeder theorem.

---

# Index

---

- absolute convergence test, 268  
absolute value, 6, 245  
addition principle, 74  
and, 39  
axiom of foundation, 32  
  
basis step, 182  
biconditional statement, 46  
bijection, 270  
bijective function, 228  
byte, 66  
  
 $C(n, k)$ , 85  
Cantor, Georg, 271  
Cantor-Bernstein-Schröder theorem, 286  
cardinality, 4, 269  
Cartesian plane, 10  
Cartesian power, 10  
Cartesian product, 8  
ceiling of a number, 104  
closed interval, 7  
codomain of a function, 225  
Cohen, Paul, 289  
combinatorial proof, 108  
comparison test, 268  
complement of a set, 20  
composite number, 116  
composition of functions, 235  
conditional statement, 42  
conjecture, 173  
constructive proof, 154  
continuity, 256  
continuous function, 256  
continuum hypothesis, 289  
contrapositive, 128  
convergence of a sequence, 262  
convergence of a series, 266  
converse of a statement, 46, 128  
  
corollary, 114  
countable set, 275  
counterexample, 175  
counting, 65  
  
definition, 113  
DeMorgan's laws, 51, 59  
difference of sets, 18  
differentiability, 257  
disproof, 172  
divergence of a sequence, 262  
divergence of a series, 266  
divergence test, 267  
divergence to  $\infty$ , 264  
divides, 116  
division algorithm, 30, 117  
division principle, 104  
divisor, 116  
domain of a function, 225  
Doxiadis, Apostolos, 33  
  
element of a set, 3  
elimination, 63  
empty set, 4  
entries of a list, 65  
equality of functions, 227  
equality of lists, 66  
equality of sets, 3  
equivalence class, 211  
equivalence relation, 210  
equivalent statements, 149  
Euclid, 140, 166  
Euler, Leonhard, 133, 168  
existence theorem, 151  
existential quantifier, 54  
existential statement, 151  
  
factorial, 78

- false, 35
- Fermat's last theorem, 37
- Fermat, Pierre de, 37
- Fibonacci sequence, 193
- floor of a number, 104
- function, 224
  - range of, 225
  - bijective, 228, 269
  - codomain of, 225
  - composition of, 235
  - continuous, 256
  - derivative of, 257
  - differentiable, 257
  - domain of, 225
  - equality, 227
  - injective, 228
  - inverse, 238
  - notation, 227
  - one-to-one, 228
  - onto, 228
  - surjective, 228
- function notation, 227
- fundamental theorem of arithmetic, 192
- fundamental theorem of calculus, viii
- gamma function, 84
- gcd, 116
- general term, 261
- geometric sequence, 195
- geometric series, 268
- Goldbach's conjecture, 37, 58
- Goldbach, Christian, 37
- golden ratio, 195
- graph, 189
  - cycle, 189
  - edges, 189
  - vertices, 189
- greatest common divisor, 116
- Hagy, Jessica, 33
- half-open interval, 7
- harmonic series, 268
- if-and-only-if theorem, 147
- image, 242
- inclusion-exclusion formula, 93
- index set, 26
- indexed set, 25
- induction, 180
  - strong, 187
- inductive hypothesis, 182
- inductive step, 182
- infinite interval, 7
- injection, 270
- injective function, 228
- integers, 3, 4
  - congruence, 131, 207
  - modulo  $n$ , 219
- intersection of sets, 18
- interval, 6
- inverse of a function, 238
- inverse relation, 239
- irrational number, 139
- lcm, 116
- least common multiple, 116
- lemma, 114
- length of a list, 65
- limit, 247
  - at infinity, 259
  - composition rule for, 257
  - constant function rule for, 251
  - constant multiple rule for, 252
  - difference rule for, 252
  - division rule for, 253
  - identity function rule for, 251
  - informal definition of, 246
  - multiplication rule for, 252
  - non-existence of, 249
  - precise definition of, 247
  - squeeze theorem for, 255
  - sum rule for, 252
- limit comparison test, 268
- list, 65
  - empty, 66
  - entries, 65
  - equal, 66
  - length, 65
  - non-repetitive, 69
  - order, 65
  - repetition, 69
- logic, 34
  - contradiction, 137
  - equivalence, 51
  - inference, 63
  - quantifier, 54
    - existential, 54
    - universal, 54

- symbols, 48, 53
- logical equivalence, 51
- logical inference, 63
- mean value theorem, 57
- Mersenne prime, 170
- modus ponens, 63
- modus tollens, 63
- multiple, 116
- multiplication principle, 69
- multiplicity, 96
- multiset, 96
- natural numbers, 4
- necessary condition, 44
- negation of a statement, 41
- non-constructive proof, 154
- one-to-one function, 228
- onto function, 228
- open interval, 7
- open sentence, 36, 56
- or, 40
- ordered pair, 8
- ordered triple, 10
- $P(n, k)$ , 81, 83
- Papadimitriou, Christos, 33
- parity, 115
- partial sum of a series, 265
- partition, 216
- Pascal's triangle, 91
- Pascal, Blaise, 91
- perfect number, 165, 168
- permutation, 80
  - $k$ -permutation, 81, 83
- pigeonhole principle, 104, 233
  - strong form, 104
- Pisano, Leonardo, 193
- power set, 15
- power, Cartesian, 10
- preimage, 242
- prime number, 37, 116
- proof
  - by cases, 124
  - by contradiction, 137
  - by induction, 180
  - by smallest counterexample, 191
  - by strong induction, 187
- combinatorial, 108
- constructive, 154
- contrapositive, 128
- direct, 113, 118
- existence, 150
- involving sets, 157
- non-constructive, 154
- uniqueness, 150, 153
- within-a-proof, 143
- proposition, 114, 118
- Pythagorean theorem, 37
- quadratic formula, 37
- quantifier, 54
- quotient, 30, 117
- range of a function, 225
- ratio test, 268
- rational numbers, 6, 139
- real numbers, 4
- reflexive property of a relation, 205
- relations, 201
  - between sets, 221
  - equivalence, 210
    - class, 211
    - inverse, 239
    - reflexive, 205
    - symmetric, 205
    - transitive, 205
  - remainder, 30, 117, 131
  - Russell's paradox, 32
  - Russell, Bertrand, 32
- sequence, 261
  - convergence of, 262
  - divergence of, 262
  - divergence to  $\infty$ , 264
  - general term, 261
- series, 265
  - absolute convergence test for, 268
  - comparison test for, 268
  - convergence of, 266
  - divergence of, 266
  - divergence test for, 267
  - geometric, 268
  - harmonic, 268
  - limit comparison test for, 268
  - Maclaurin, 265
  - partial sum of, 265

- ratio test for, 268
- set(s)
  - builder-notation, 5, 157
  - cardinalities of
    - comparison of, 280
    - equal, 269
    - unequal, 269
  - cardinality of, 3, 269
  - complement, 20
  - countable, 275
  - element of, 3
  - empty, 4
  - equal, 3
  - partition of, 216
  - subset of, 12
  - uncountable, 275
- sigma notation, 25
- size, *see* cardinality
- statement, 35
  - biconditional, 46
  - conditional, 42
    - necessary, 44
    - sufficient, 44
  - converse, 46
  - equivalent, 149
  - existential, 151
  - negation, 59
- string, 66
- strong form of pigeonhole principle, 104
- strong induction, 187
- subset, 12
- subtraction principle, 76
- sufficient condition, 44
- surjection, 270
- surjective function, 228
- symmetric property of a relation, 205
- theorem, 113
  - existence, 151
  - if-and-only-if, 147
- three-dimensional space, 10
- transitive property of a relation, 205
- tree, 189
- triangle inequality, 245
- triple, ordered, 10
- true, 35
- truth
  - table, 39
  - value, 39

Undergraduate Texts in Mathematics

# Paul R. Halmos

# Naive Set Theory

Undergraduate Texts in Mathematics

## Paul R. Halmos

## Naive Set Theory

Undergraduate Texts in Mathematics

### Paul R. Halmos

### Naive Set Theory

Undergraduate Texts in Mathematics  
Paul R. Halmos  
Naive Set Theory

Springer

Springer



Springer

# **Undergraduate Texts in Mathematics**

*Editors*

S. Axler  
F.W. Gehring  
K.A. Ribet

**Springer Science+Business Media, LLC**

**Paul R. Halmos**

**Naive Set Theory**



P.R. Halmos  
Department of Mathematics  
Santa Clara University  
Santa Clara, CA 95128  
USA

*Editorial Board*

S. Axler  
Mathematics Department  
San Francisco State  
University  
San Francisco, CA 94132  
USA

F.W. Gehring  
Mathematics Department  
East Hall  
University of Michigan  
Ann Arbor, MI 48109  
USA

K.A. Ribet  
Mathematics Department  
University of California  
at Berkeley  
Berkeley, CA 94720-3840  
USA

---

Mathematics Subject Classifications (2000): 03-01, 03EXX

---

Halmos, Paul Richard, 1914—

Naive set theory.

(Undergraduate texts in mathematics)

Reprint of the ed. published by Van Nostrand,  
Princeton, N.J., in series: The University series  
in undergraduate mathematics.

1. Set theory. 2. Arithmetic—Foundations.

I. Title.

[QA248.H26 1974] 511'.3 74-10687  
ISBN 978-0-387-90104-6 ISBN 978-1-4757-1645-0 (eBook)  
DOI 10.1007/978-1-4757-1645-0

Printed on acid-free paper.

All rights reserved.

No part of this book may be translated or reproduced in  
any form without written permission from Springer-Verlag.

© Springer Science+Business Media New York 1974

Originally published by Springer-Verlag New York Inc. in 1974

Softcover reprint of the hardcover 1st edition 1974

19 18 17 16 15 14 13 12

## PREFACE

---

Every mathematician agrees that every mathematician must know some set theory; the disagreement begins in trying to decide how much is some. This book contains my answer to that question. The purpose of the book is to tell the beginning student of advanced mathematics the basic set-theoretic facts of life, and to do so with the minimum of philosophical discourse and logical formalism. The point of view throughout is that of a prospective mathematician anxious to study groups, or integrals, or manifolds. From this point of view the concepts and methods of this book are merely some of the standard mathematical tools; the expert specialist will find nothing new here.

Scholarly bibliographical credits and references are out of place in a purely expository book such as this one. The student who gets interested in set theory for its own sake should know, however, that there is much more to the subject than there is in this book. One of the most beautiful sources of set-theoretic wisdom is still Hausdorff's *Set theory*. A recent and highly readable addition to the literature, with an extensive and up-to-date bibliography, is *Axiomatic set theory* by Suppes.

In set theory "naive" and "axiomatic" are contrasting words. The present treatment might best be described as axiomatic set theory from the naive point of view. It is axiomatic in that some axioms for set theory are stated and used as the basis of all subsequent proofs. It is naive in that the language and notation are those of ordinary informal (but formalizable) mathematics. A more important way in which the naive point of view predominates is that set theory is regarded as a body of facts, of which the axioms are a brief and convenient summary; in the orthodox axiomatic view the logical relations among various axioms are the central objects of study. Analogously, a study of geometry might be regarded as purely naive if it proceeded on the paper-folding kind of intuition alone; the other extreme, the purely axiomatic one, is the one in which axioms for the various non-Euclidean geometries are studied with the same amount of attention as Euclid's. The analogue of the point of view of this book

is the study of just one sane set of axioms with the intention of describing Euclidean geometry only.

Instead of *Naive set theory* a more honest title for the book would have been *An outline of the elements of naive set theory*. “Elements” would warn the reader that not everything is here; “outline” would warn him that even what is here needs filling in. The style is usually informal to the point of being conversational. There are very few displayed theorems; most of the facts are just stated and followed by a sketch of a proof, very much as they might be in a general descriptive lecture. There are only a few exercises, officially so labelled, but, in fact, most of the book is nothing but a long chain of exercises with hints. The reader should continually ask himself whether he knows how to jump from one hint to the next, and, accordingly, he should not be discouraged if he finds that his reading rate is considerably slower than normal.

This is not to say that the contents of this book are unusually difficult or profound. What is true is that the concepts are very general and very abstract, and that, therefore, they may take some getting used to. It is a mathematical truism, however, that the more generally a theorem applies, the less deep it is. The student’s task in learning set theory is to steep himself in unfamiliar but essentially shallow generalities till they become so familiar that they can be used with almost no conscious effort. In other words, general set theory is pretty trivial stuff really, but, if you want to be a mathematician, you need some, and here it is; read it, absorb it, and forget it.

P. R. H.

## CONTENTS

---

SECTION	PAGE
PREFACE	v
1 THE AXIOM OF EXTENSION	1
2 THE AXIOM OF SPECIFICATION	4
3 UNORDERED PAIRS	8
4 UNIONS AND INTERSECTIONS	12
5 COMPLEMENTS AND POWERS	17
6 ORDERED PAIRS	22
7 RELATIONS	26
8 FUNCTIONS	30
9 FAMILIES	34
10 INVERSES AND COMPOSITES	38
11 NUMBERS	42
12 THE PEANO AXIOMS	46
13 ARITHMETIC	50
14 ORDER	54
15 THE AXIOM OF CHOICE	59
16 ZORN'S LEMMA	62
17 WELL ORDERING	66
18 TRANSFINITE RECURSION	70
19 ORDINAL NUMBERS	74
20 SETS OF ORDINAL NUMBERS	78
21 ORDINAL ARITHMETIC	81
22 THE SCHRÖDER-BERNSTEIN THEOREM	86
23 COUNTABLE SETS	90
24 CARDINAL ARITHMETIC	94
25 CARDINAL NUMBERS	99
INDEX	102

## SECTION 1

---

### THE AXIOM OF EXTENSION

---

---

A pack of wolves, a bunch of grapes, or a flock of pigeons are all examples of sets of things. The mathematical concept of a set can be used as the foundation for all known mathematics. The purpose of this little book is to develop the basic properties of sets. Incidentally, to avoid terminological monotony, we shall sometimes say *collection* instead of *set*. The word “class” is also used in this context, but there is a slight danger in doing so. The reason is that in some approaches to set theory “class” has a special technical meaning. We shall have occasion to refer to this again a little later.

One thing that the development will not include is a definition of sets. The situation is analogous to the familiar axiomatic approach to elementary geometry. That approach does not offer a definition of points and lines; instead it describes what it is that one can do with those objects. The semi-axiomatic point of view adopted here assumes that the reader has the ordinary, human, intuitive (and frequently erroneous) understanding of what sets are; the purpose of the exposition is to delineate some of the many things that one can correctly do with them.

Sets, as they are usually conceived, have *elements* or *members*. An element of a set may be a wolf, a grape, or a pigeon. It is important to know that a set itself may also be an element of some other set. Mathematics is full of examples of sets of sets. A line, for instance, is a set of points; the set of all lines in the plane is a natural example of a set of sets (of points). What may be surprising is not so much that sets may occur as elements, but that for mathematical purposes no other elements need ever be considered. In this book, in particular, we shall study sets, and sets of sets, and similar towers of sometimes frightening height and complexity—and nothing else. By way of examples we might occasionally speak of sets of

cabbages, and kings, and the like, but such usage is always to be construed as an illuminating parable only, and not as a part of the theory that is being developed.

The principal concept of set theory, the one that in completely axiomatic studies is the principal primitive (undefined) concept, is that of *belonging*. If  $x$  belongs to  $A$  ( $x$  is an element of  $A$ ,  $x$  is *contained* in  $A$ ), we shall write

$$x \in A.$$

This version of the Greek letter epsilon is so often used to denote belonging that its use to denote anything else is almost prohibited. Most authors relegate  $\epsilon$  to its set-theoretic use forever and use  $\epsilon$  when they need the fifth letter of the Greek alphabet.

Perhaps a brief digression on alphabetic etiquette in set theory might be helpful. There is no compelling reason for using small and capital letters as in the preceding paragraph; we might have written, and often will write, things like  $x \epsilon y$  and  $A \epsilon B$ . Whenever possible, however, we shall informally indicate the status of a set in a particular hierarchy under consideration by means of the convention that letters at the beginning of the alphabet denote elements, and letters at the end denote sets containing them; similarly letters of a relatively simple kind denote elements, and letters of the larger and gaudier fonts denote sets containing them. Examples:  $x \in A$ ,  $A \in X$ ,  $X \in C$ .

A possible relation between sets, more elementary than belonging, is *equality*. The equality of two sets  $A$  and  $B$  is universally denoted by the familiar symbol

$$A = B;$$

the fact that  $A$  and  $B$  are not equal is expressed by writing

$$A \neq B.$$

The most basic property of belonging is its relation to equality, which can be formulated as follows.

**Axiom of extension.** *Two sets are equal if and only if they have the same elements.*

With greater pretentiousness and less clarity: a set is determined by its extension.

It is valuable to understand that the axiom of extension is not just a logically necessary property of equality but a non-trivial statement about belonging. One way to come to understand the point is to consider a partially analogous situation in which the analogue of the axiom of extension

does not hold. Suppose, for instance, that we consider human beings instead of sets, and that, if  $x$  and  $A$  are human beings, we write  $x \in A$  whenever  $x$  is an ancestor of  $A$ . (The ancestors of a human being are his parents, his parents' parents, their parents, etc., etc.) The analogue of the axiom of extension would say here that if two human beings are equal, then they have the same ancestors (this is the "only if" part, and it is true), and also that if two human beings have the same ancestors, then they are equal (this is the "if" part, and it is false).

If  $A$  and  $B$  are sets and if every element of  $A$  is an element of  $B$ , we say that  $A$  is a *subset* of  $B$ , or  $B$  *includes*  $A$ , and we write

$$A \subset B$$

or

$$B \supset A.$$

The wording of the definition implies that each set must be considered to be included in itself ( $A \subset A$ ); this fact is described by saying that set inclusion is *reflexive*. (Note that, in the same sense of the word, equality also is reflexive.) If  $A$  and  $B$  are sets such that  $A \subset B$  and  $A \neq B$ , the word *proper* is used (proper subset, proper inclusion). If  $A$ ,  $B$ , and  $C$  are sets such that  $A \subset B$  and  $B \subset C$ , then  $A \subset C$ ; this fact is described by saying that set inclusion is *transitive*. (This property is also shared by equality.)

If  $A$  and  $B$  are sets such that  $A \subset B$  and  $B \subset A$ , then  $A$  and  $B$  have the same elements and therefore, by the axiom of extension,  $A = B$ . This fact is described by saying that set inclusion is *antisymmetric*. (In this respect set inclusion behaves differently from equality. Equality is *symmetric*, in the sense that if  $A = B$ , then necessarily  $B = A$ .) The axiom of extension can, in fact, be reformulated in these terms: if  $A$  and  $B$  are sets, then a necessary and sufficient condition that  $A = B$  is that both  $A \subset B$  and  $B \subset A$ . Correspondingly, almost all proofs of equalities between two sets  $A$  and  $B$  are split into two parts; first show that  $A \subset B$ , and then show that  $B \subset A$ .

Observe that belonging ( $\in$ ) and inclusion ( $\subset$ ) are conceptually very different things indeed. One important difference has already manifested itself above: inclusion is always reflexive, whereas it is not at all clear that belonging is ever reflexive. That is:  $A \subset A$  is always true; is  $A \in A$  ever true? It is certainly not true of any reasonable set that anyone has ever seen. Observe, along the same lines, that inclusion is transitive, whereas belonging is not. Everyday examples, involving, for instance, super-organizations whose members are organizations, will readily occur to the interested reader.

## SECTION 2

---

### THE AXIOM OF SPECIFICATION

---

All the basic principles of set theory, except only the axiom of extension, are designed to make new sets out of old ones. The first and most important of these basic principles of set manufacture says, roughly speaking, that anything intelligent one can assert about the elements of a set specifies a subset, namely, the subset of those elements about which the assertion is true.

Before formulating this principle in exact terms, we look at a heuristic example. Let  $A$  be the set of all men. The sentence " $x$  is married" is true for some of the elements  $x$  of  $A$  and false for others. The principle we are illustrating is the one that justifies the passage from the given set  $A$  to the subset (namely, the set of all married men) specified by the given sentence. To indicate the generation of the subset, it is usually denoted by

$$\{x \in A : x \text{ is married}\}.$$

Similarly

$$\{x \in A : x \text{ is not married}\}$$

is the set of all bachelors;

$$\{x \in A : \text{the father of } x \text{ is Adam}\}$$

is the set that contains Cain and Abel and nothing else; and

$$\{x \in A : x \text{ is the father of Abel}\}$$

is the set that contains Adam and nothing else. Warning: a box that contains a hat and nothing else is not the same thing as a hat, and, in the same way, the last set in this list of examples is not to be confused with

Adam. The analogy between sets and boxes has many weak points, but sometimes it gives a helpful picture of the facts.

All that is lacking for the precise general formulation that underlies the examples above is a definition of *sentence*. Here is a quick and informal one. There are two basic types of sentences, namely, assertions of belonging,

$$x \in A,$$

and assertions of equality,

$$A = B;$$

all other sentences are obtained from such *atomic* sentences by repeated applications of the usual logical operators, subject only to the minimal courtesies of grammar and unambiguity. To make the definition more explicit (and longer) it is necessary to append to it a list of the “usual logical operators” and the rules of syntax. An adequate (and, in fact, redundant) list of the former contains seven items:

*and*,  
*or* (in the sense of “either—or—or both”),  
*not*,  
*if—then*—(or *implies*),  
*if and only if*,  
*for some* (or *there exists*),  
*for all*.

As for the rules of sentence construction, they can be described as follows.  
(i) Put “not” before a sentence and enclose the result between parentheses. (The reason for parentheses, here and below, is to guarantee unambiguity. Note, incidentally, that they make all other punctuation marks unnecessary. The complete parenthetical equipment that the definition of sentences calls for is rarely needed. We shall always omit as many parentheses as it seems safe to omit without leading to confusion. In normal mathematical practice, to be followed in this book, several different sizes and shapes of parentheses are used, but that is for visual convenience only.)  
(ii) Put “and” or “or” or “if and only if” between two sentences and enclose the result between parentheses. (iii) Replace the dashes in “if—then —” by sentences and enclose the result in parentheses. (iv) Replace the dash in “for some—” or in “for all—” by a letter, follow the result by a sentence, and enclose the whole in parentheses. (If the letter used does not occur in the sentence, no harm is done. According to the usual and natural convention “for some  $y$  ( $x \in A$ )” just means “ $x \in A$ ”. It is equally

harmless if the letter used has already been used with “for some—” or “for all—.” Recall that “for some  $x$  ( $x \in A$ )” means the same as “for some  $y$  ( $y \in A$ )”; it follows that a judicious change of notation will always avert alphabetic collisions.)

We are now ready to formulate the major principle of set theory, often referred to by its German name *Aussonderungsaxiom*.

**Axiom of specification.** *To every set  $A$  and to every condition  $S(x)$  there corresponds a set  $B$  whose elements are exactly those elements  $x$  of  $A$  for which  $S(x)$  holds.*

A “condition” here is just a sentence. The symbolism is intended to indicate that the letter  $x$  is *free* in the sentence  $S(x)$ ; that means that  $x$  occurs in  $S(x)$  at least once without being introduced by one of the phrases “for some  $x$ ” or “for all  $x$ .” It is an immediate consequence of the axiom of extension that the axiom of specification determines the set  $B$  uniquely. To indicate the way  $B$  is obtained from  $A$  and from  $S(x)$  it is customary to write

$$B = \{x \in A : S(x)\}.$$

To obtain an amusing and instructive application of the axiom of specification, consider, in the role of  $S(x)$ , the sentence

$$\text{not } (x \in x).$$

It will be convenient, here and throughout, to write “ $x \notin A$ ” (alternatively “ $x \not\in A$ ”) instead of “not ( $x \in A$ )”; in this notation, the role of  $S(x)$  is now played by

$$x \notin x.$$

It follows that, whatever the set  $A$  may be, if  $B = \{x \in A : x \notin x\}$ , then, for all  $y$ ,

$$(*) \quad y \in B \text{ if and only if } (y \in A \text{ and } y \notin y).$$

Can it be that  $B \in A$ ? We proceed to prove that the answer is no. Indeed, if  $B \in A$ , then either  $B \in B$  also (unlikely, but not obviously impossible), or else  $B \notin B$ . If  $B \in B$ , then, by (\*), the assumption  $B \in A$  yields  $B \notin B$ —a contradiction. If  $B \notin B$ , then, by (\*) again, the assumption  $B \in A$  yields  $B \in B$ —a contradiction again. This completes the proof that  $B \in A$  is impossible, so that we must have  $B \notin A$ . The most interesting part of this conclusion is that there exists something (namely  $B$ ) that does not belong to  $A$ . The set  $A$  in this argument was quite arbitrary. We have proved, in other words, that

$$\text{nothing contains everything,}$$

or, more spectacularly,

*there is no universe.*

“Universe” here is used in the sense of “universe of discourse,” meaning, in any particular discussion, a set that contains all the objects that enter into that discussion.

In older (pre-axiomatic) approaches to set theory, the existence of a universe was taken for granted, and the argument in the preceding paragraph was known as the *Russell paradox*. The moral is that it is impossible, especially in mathematics, to get something for nothing. To specify a set, it is not enough to pronounce some magic words (which may form a sentence such as “ $x \in x$ ”); it is necessary also to have at hand a set to whose elements the magic words apply.

## SECTION 3

---

### UNORDERED PAIRS

---

For all that has been said so far, we might have been operating in a vacuum. To give the discussion some substance, let us now officially assume that

*there exists a set.*

Since later on we shall formulate a deeper and more useful existential assumption, this assumption plays a temporary role only. One consequence of this innocuous seeming assumption is that there exists a set without any elements at all. Indeed, if  $A$  is a set, apply the axiom of specification to  $A$  with the sentence " $x \neq x$ " (or, for that matter, with any other universally false sentence). The result is the set  $\{x \in A : x \neq x\}$ , and that set, clearly, has no elements. The axiom of extension implies that there can be only one set with no elements. The usual symbol for that set is

$\emptyset$ ;

the set is called the *empty set*.

The empty set is a subset of every set, or, in other words,  $\emptyset \subset A$  for every  $A$ . To establish this, we might argue as follows. It is to be proved that every element in  $\emptyset$  belongs to  $A$ ; since there are no elements in  $\emptyset$ , the condition is automatically fulfilled. The reasoning is correct but perhaps unsatisfying. Since it is a typical example of a frequent phenomenon, a condition holding in the "vacuous" sense, a word of advice to the inexperienced reader might be in order. To prove that something is true about the empty set, prove that it cannot be false. How, for instance, could it be false that  $\emptyset \subset A$ ? It could be false only if  $\emptyset$  had an element that did not belong to  $A$ . Since  $\emptyset$  has no elements at all, this is absurd. Conclusion:  $\emptyset \subset A$  is not false, and therefore  $\emptyset \subset A$  for every  $A$ .

The set theory developed so far is still a pretty poor thing; for all we know there is only one set and that one is empty. Are there enough sets to ensure that every set is an element of some set? Is it true that for any two sets there is a third one that they both belong to? What about three sets, or four, or any number? We need a new principle of set construction to resolve such questions. The following principle is a good beginning.

**Axiom of pairing.** *For any two sets there exists a set that they both belong to.*

Note that this is just the affirmative answer to the second question above.

To reassure worriers, let us hasten to observe that words such as "two," "three," and "four," used above, do not refer to the mathematical concepts bearing those names, which will be defined later; at present such words are merely the ordinary linguistic abbreviations for "something and then something else" repeated an appropriate number of times. Thus, for instance, the axiom of pairing, in unabbreviated form, says that if  $a$  and  $b$  are sets, then there exists a set  $A$  such that  $a \in A$  and  $b \in A$ .

One consequence (in fact an equivalent formulation) of the axiom of pairing is that for any two sets there exists a set that contains both of them and nothing else. Indeed, if  $a$  and  $b$  are sets, and if  $A$  is a set such that  $a \in A$  and  $b \in A$ , then we can apply the axiom of specification to  $A$  with the sentence " $x = a$  or  $x = b$ ." The result is the set

$$\{x \in A : x = a \text{ or } x = b\},$$

and that set, clearly, contains just  $a$  and  $b$ . The axiom of extension implies that there can be only one set with this property. The usual symbol for that set is

$$\{a, b\};$$

the set is called the *pair* (or, by way of emphatic comparison with a subsequent concept, the *unordered pair*) formed by  $a$  and  $b$ .

If, temporarily, we refer to the sentence " $x = a$  or  $x = b$ " as  $S(x)$ , we may express the axiom of pairing by saying that there exists a set  $B$  such that

$$(*) \quad x \in B \text{ if and only if } S(x).$$

The axiom of specification, applied to a set  $A$ , asserts the existence of a set  $B$  such that

$$(**) \quad x \in B \text{ if and only if } (x \in A \text{ and } S(x)).$$

The relation between (\*) and (\*\*) typifies something that occurs quite frequently. All the remaining principles of set construction are pseudo-special cases of the axiom of specification in the sense in which (\*) is a pseudo-special case of (\*\*). They all assert the existence of a set specified by a certain condition; if it were known in advance that there exists a set containing all the specified elements, then the existence of a set containing just them would indeed follow as a special case of the axiom of specification.

If  $a$  is a set, we may form the unordered pair  $\{a, a\}$ . That unordered pair is denoted by

$$\{a\}$$

and is called the *singleton* of  $a$ ; it is uniquely characterized by the statement that it has  $a$  as its only element. Thus, for instance,  $\emptyset$  and  $\{\emptyset\}$  are very different sets; the former has no elements, whereas the latter has the unique element  $\emptyset$ . To say that  $a \in A$  is equivalent to saying that  $\{a\} \subset A$ .

The axiom of pairing ensures that every set is an element of some set and that any two sets are simultaneously elements of some one and the same set. (The corresponding questions for three and four and more sets will be answered later.) Another pertinent comment is that from the assumptions we have made so far we can infer the existence of very many sets indeed. For examples consider the sets  $\emptyset$ ,  $\{\emptyset\}$ ,  $\{\{\emptyset\}\}$ ,  $\{\{\{\emptyset\}\}\}$ , etc.; consider the pairs, such as  $\{\emptyset, \{\emptyset\}\}$ , formed by any two of them; consider the pairs formed by any two such pairs, or else the mixed pairs formed by any singleton and any pair; and proceed so on ad infinitum.

**EXERCISE.** Are all the sets obtained in this way distinct from one another?

Before continuing our study of set theory, we pause for a moment to discuss a notational matter. It seems natural to denote the set  $B$  described in (\*) by  $\{x: S(x)\}$ ; in the special case that was there considered

$$\{x: x = a \text{ or } x = b\} = \{a, b\}.$$

We shall use this symbolism whenever it is convenient and permissible to do so. If, that is,  $S(x)$  is a condition on  $x$  such that the  $x$ 's that  $S(x)$  specifies constitute a set, then we may denote that set by

$$\{x: S(x)\}.$$

In case  $A$  is a set and  $S(x)$  is  $(x \in A)$ , then it is permissible to form  $\{x: S(x)\}$ ; in fact

$$\{x: x \in A\} = A.$$

If  $A$  is a set and  $S(x)$  is an arbitrary sentence, it is permissible to form  $\{x: x \in A \text{ and } S(x)\}$ ; this set is the same as  $\{x \in A : S(x)\}$ . As further examples, we note that

$$\{x: x \neq x\} = \emptyset$$

and

$$\{x: x = a\} = \{a\}.$$

In case  $S(x)$  is  $(x \neq x)$ , or in case  $S(x)$  is  $(x = x)$ , the specified  $x$ 's do not constitute a set.

Despite the maxim about never getting something for nothing, it seems a little harsh to be told that certain sets are not really sets and even their names must never be mentioned. Some approaches to set theory try to soften the blow by making systematic use of such illegal sets but just not calling them sets; the customary word is "class." A precise explanation of what classes really are and how they are used is irrelevant in the present approach. Roughly speaking, a class may be identified with a condition (sentence), or, rather, with the "extension" of a condition.

## SECTION 4

---

### UNIONS AND INTERSECTIONS

---

If  $A$  and  $B$  are sets, it is sometimes natural to wish to unite their elements into one comprehensive set. One way of describing such a comprehensive set is to require it to contain all the elements that belong to at least one of the two members of the pair  $\{A, B\}$ . This formulation suggests a sweeping generalization of itself; surely a similar construction should apply to arbitrary collections of sets and not just to pairs of them. What is wanted, in other words, is the following principle of set construction.

**Axiom of unions.** *For every collection of sets there exists a set that contains all the elements that belong to at least one set of the given collection.*

Here it is again: for every collection  $\mathcal{C}$  there exists a set  $U$  such that if  $x \in X$  for some  $X$  in  $\mathcal{C}$ , then  $x \in U$ . (Note that “at least one” is the same as “some.”)

The comprehensive set  $U$  described above may be too comprehensive; it may contain elements that belong to none of the sets  $X$  in the collection  $\mathcal{C}$ . This is easy to remedy; just apply the axiom of specification to form the set

$$\{x \in U: x \in X \text{ for some } X \text{ in } \mathcal{C}\}.$$

(The condition here is a translation into idiomatic usage of the mathematically more acceptable “*for some  $X$  ( $x \in X$  and  $X \in \mathcal{C}$ )*.”) It follows that, for every  $x$ , a necessary and sufficient condition that  $x$  belong to this set is that  $x$  belong to  $X$  for some  $X$  in  $\mathcal{C}$ . If we change notation and call the new set  $U$  again, then

$$U = \{x: x \in X \text{ for some } X \text{ in } \mathcal{C}\}.$$

This set  $U$  is called the *union* of the collection  $\mathcal{C}$  of sets; note that the

axiom of extension guarantees its uniqueness. The simplest symbol for  $U$  that is in use at all is not very popular in mathematical circles; it is

$$\text{U c.}$$

Most mathematicians prefer something like

$$\text{U } \{X: X \in \mathcal{C}\}$$

or

$$\text{U}_{x \in \mathcal{C}} X.$$

Further alternatives are available in certain important special cases; they will be described in due course.

For the time being we restrict our study of the theory of unions to the simplest facts only. The simplest fact of all is that

$$\text{U } \{X: X \in \emptyset\} = \emptyset,$$

and the next simplest fact is that

$$\text{U } \{X: X \in \{A\}\} = A.$$

In the brutally simple notation mentioned above these facts are expressed by

$$\text{U } \emptyset = \emptyset$$

and

$$\text{U } \{A\} = A.$$

The proofs are immediate from the definitions.

There is a little more substance in the union of pairs of sets (which is what started this whole discussion anyway). In that case special notation is used:

$$\text{U } \{X: X \in \{A, B\}\} = A \cup B.$$

The general definition of unions implies in the special case that  $x \in A \cup B$  if and only if  $x$  belongs to either  $A$  or  $B$  or both; it follows that

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

Here are some easily proved facts about the unions of pairs:

$$A \cup \emptyset = A,$$

$$A \cup B = B \cup A \text{ (commutativity),}$$

$$A \cup (B \cup C) = (A \cup B) \cup C \text{ (associativity),}$$

$$A \cup A = A \text{ (idempotence),}$$

$$A \subset B \text{ if and only if } A \cup B = B.$$

Every student of mathematics should prove these things for himself at least once in his life. The proofs are based on the corresponding elementary properties of the logical operator *or*.

An equally simple but quite suggestive fact is that

$$\{a\} \cup \{b\} = \{a, b\}.$$

What this suggests is the way to generalize pairs. Specifically, we write

$$\{a, b, c\} = \{a\} \cup \{b\} \cup \{c\}.$$

The equation defines its left side. The right side should by rights have at least one pair of parentheses in it, but, in view of the associative law, their omission can lead to no misunderstanding. Since it is easy to prove that

$$\{a, b, c\} = \{x: x = a \text{ or } x = b \text{ or } x = c\},$$

we know now that for every three sets there exists a set that contains them and nothing else; it is natural to call that uniquely determined set the (*unordered*) *triple* formed by them. The extension of the notation and terminology thus introduced to more terms (*quadruples*, etc.) is obvious.

The formation of unions has many points of similarity with another set-theoretic operation. If  $A$  and  $B$  are sets, the *intersection* of  $A$  and  $B$  is the set

$$A \cap B$$

defined by

$$A \cap B = \{x \in A: x \in B\}.$$

The definition is symmetric in  $A$  and  $B$  even if it looks otherwise; we have

$$A \cap B = \{x \in B: x \in A\},$$

and, in fact, since  $x \in A \cap B$  if and only if  $x$  belongs to both  $A$  and  $B$ , it follows that

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

The basic facts about intersections, as well as their proofs, are similar to the basic facts about unions:

$$A \cap \emptyset = \emptyset,$$

$$A \cap B = B \cap A,$$

$$A \cap (B \cap C) = (A \cap B) \cap C,$$

$$A \cap A = A,$$

$$A \subset B \text{ if and only if } A \cap B = A.$$

Pairs of sets with an empty intersection occur frequently enough to justify the use of a special word: if  $A \cap B = \emptyset$ , the sets  $A$  and  $B$  are called *disjoint*. The same word is sometimes applied to a collection of sets to indicate that any two distinct sets of the collection are disjoint; alternatively we may speak in such a situation of a *pairwise disjoint* collection.

Two useful facts about unions and intersections involve both the operations at the same time:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$

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

These identities are called the *distributive laws*. By way of a sample of a set-theoretic proof, we prove the second one. If  $x$  belongs to the left side, then  $x$  belongs either to  $A$  or to both  $B$  and  $C$ ; if  $x$  is in  $A$ , then  $x$  is in both  $A \cup B$  and  $A \cup C$ , and if  $x$  is in both  $B$  and  $C$ , then, again,  $x$  is in both  $A \cup B$  and  $A \cup C$ ; it follows that, in any case,  $x$  belongs to the right side. This proves that the right side includes the left. To prove the reverse inclusion, just observe that if  $x$  belongs to both  $A \cup B$  and  $A \cup C$ , then  $x$  belongs either to  $A$  or to both  $B$  and  $C$ .

The formation of the intersection of two sets  $A$  and  $B$ , or, we might as well say, the formation of the intersection of a pair  $\{A, B\}$  of sets, is a special case of a much more general operation. (This is another respect in which the theory of intersections imitates that of unions.) The existence of the general operation of intersection depends on the fact that for each non-empty collection of sets there exists a set that contains exactly those elements that belong to every set of the given collection. In other words: for each collection  $\mathcal{C}$ , other than  $\emptyset$ , there exists a set  $V$  such that  $x \in V$  if and only if  $x \in X$  for every  $X$  in  $\mathcal{C}$ . To prove this assertion, let  $A$  be any particular set in  $\mathcal{C}$  (this step is justified by the fact that  $\mathcal{C} \neq \emptyset$ ) and write

$$V = \{x \in A : x \in X \text{ for every } X \text{ in } \mathcal{C}\}.$$

(The condition means “*for all  $X$  (if  $X \in \mathcal{C}$ , then  $x \in X$ .)*”) The dependence of  $V$  on the arbitrary choice of  $A$  is illusory; in fact

$$V = \{x : x \in X \text{ for every } X \text{ in } \mathcal{C}\}.$$

The set  $V$  is called the *intersection* of the collection  $\mathcal{C}$  of sets; the axiom of extension guarantees its uniqueness. The customary notation is similar to the one for unions: instead of the unobjectionable but unpopular

$$\bigcap \mathcal{C},$$

the set  $V$  is usually denoted by

$$\bigcap \{X: X \in \mathcal{C}\}$$

or

$$\bigcap_{X \in \mathcal{C}} X.$$

EXERCISE. A necessary and sufficient condition that  $(A \cap B) \cup C = A \cap (B \cup C)$  is that  $C \subset A$ . Observe that the condition has nothing to do with the set  $B$ .

## SECTION 5

---

### COMPLEMENTS AND POWERS

---

---

If  $A$  and  $B$  are sets, the *difference* between  $A$  and  $B$ , more often known as the *relative complement* of  $B$  in  $A$ , is the set  $A - B$  defined by

$$A - B = \{x \in A : x \notin B\}.$$

Note that in this definition it is not necessary to assume that  $B \subset A$ . In order to record the basic facts about complementation as simply as possible, we assume nevertheless (in this section only) that all the sets to be mentioned are subsets of one and the same set  $E$  and that all complements (unless otherwise specified) are formed relative to that  $E$ . In such situations (and they are quite common) it is easier to remember the underlying set  $E$  than to keep writing it down, and this makes it possible to simplify the notation. An often used symbol for the temporarily absolute (as opposed to relative) complement of  $A$  is  $A'$ . In terms of this symbol the basic facts about complementation can be stated as follows:

$$\begin{aligned}(A')' &= A, \\ \emptyset' &= E, \quad E' = \emptyset, \\ A \cap A' &= \emptyset, \quad A \cup A' = E, \\ A \subset B &\text{ if and only if } B' \subset A'.\end{aligned}$$

The most important statements about complements are the so-called *De Morgan laws*:

$$(A \cup B)' = A' \cap B', \quad (A \cap B)' = A' \cup B'.$$

(We shall see presently that the De Morgan laws hold for the unions and intersections of larger collections of sets than just pairs.) These facts about

complementation imply that the theorems of set theory usually come in pairs. If in an inclusion or equation involving unions, intersections, and complements of subsets of  $E$  we replace each set by its complement, interchange unions and intersections, and reverse all inclusions, the result is another theorem. This fact is sometimes referred to as the *principle of duality* for sets.

Here are some easy exercises on complementation.

$$A - B = A \cap B'.$$

$$A \subset B \text{ if and only if } A - B = \emptyset.$$

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

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

$$A \cap B \subset (A \cap C) \cup (B \cap C').$$

$$(A \cup C) \cap (B \cup C') \subset A \cup B.$$

If  $A$  and  $B$  are sets, the *symmetric difference* (or *Boolean sum*) of  $A$  and  $B$  is the set  $A + B$  defined by

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

This operation is commutative ( $A + B = B + A$ ) and associative ( $A + (B + C) = (A + B) + C$ ), and is such that  $A + \emptyset = A$  and  $A + A = \emptyset$ .

This may be the right time to straighten out a trivial but occasionally puzzling part of the theory of intersections. Recall, to begin with, that intersections were defined for non-empty collections only. The reason is that the same approach to the empty collection does not define a set. Which  $x$ 's are specified by the sentence

$$x \in X \text{ for every } X \text{ in } \emptyset?$$

As usual for questions about  $\emptyset$  the answer is easier to see for the corresponding negative question. Which  $x$ 's do *not* satisfy the stated condition? If it is not true that  $x \in X$  for every  $X$  in  $\emptyset$ , then there must exist an  $X$  in  $\emptyset$  such that  $x \notin X$ ; since, however, there do not exist any  $X$ 's in  $\emptyset$  at all, this is absurd. Conclusion: no  $x$  fails to satisfy the stated condition, or, equivalently, every  $x$  does satisfy it. In other words, the  $x$ 's that the condition specifies exhaust the (nonexistent) universe. There is no profound problem here; it is merely a nuisance to be forced always to be making

qualifications and exceptions just because some set somewhere along some construction might turn out to be empty. There is nothing to be done about this; it is just a fact of life.

If we restrict our attention to subsets of a particular set  $E$ , as we have temporarily agreed to do, then the unpleasantness described in the preceding paragraph appears to go away. The point is that in that case we can define the intersection of a collection  $\mathcal{C}$  (of subsets of  $E$ ) to be the set

$$\{x \in E : x \in X \text{ for every } X \text{ in } \mathcal{C}\}.$$

This is nothing revolutionary; for each non-empty collection, the new definition agrees with the old one. The difference is in the way the old and the new definitions treat the empty collection; according to the new definition  $\bigcap_{X \in \emptyset} X$  is equal to  $E$ . (For which elements  $x$  of  $E$  can it be false that  $x \in X$  for every  $X$  in  $\emptyset$ ?) The difference is just a matter of language. A little reflection reveals that the “new” definition offered for the intersection of a collection  $\mathcal{C}$  of subsets of  $E$  is really the same as the old definition of the intersection of the collection  $\mathcal{C} \cup \{E\}$ , and the latter is never empty.

We have been considering the subsets of a set  $E$ ; do those subsets themselves constitute a set? The following principle guarantees that the answer is yes.

**Axiom of powers.** *For each set there exists a collection of sets that contains among its elements all the subsets of the given set.*

In other words, if  $E$  is a set, then there exists a set (collection)  $\mathcal{P}$  such that if  $X \subset E$ , then  $X \in \mathcal{P}$ .

The set  $\mathcal{P}$  described above may be larger than wanted; it may contain elements other than the subsets of  $E$ . This is easy to remedy; just apply the axiom of specification to form the set  $\{X \in \mathcal{P} : X \subset E\}$ . (Recall that “ $X \subset E$ ” says the same thing as “for all  $x$  (if  $x \in X$  then  $x \in E$ ).”) Since, for every  $X$ , a necessary and sufficient condition that  $X$  belong to this set is that  $X$  be a subset of  $E$ , it follows that if we change notation and call this set  $\mathcal{P}$  again, then

$$\mathcal{P} = \{X : X \subset E\}.$$

The set  $\mathcal{P}$  is called the *power set* of  $E$ ; the axiom of extension guarantees its uniqueness. The dependence of  $\mathcal{P}$  on  $E$  is denoted by writing  $\mathcal{P}(E)$  instead of just  $\mathcal{P}$ .

Because the set  $\mathcal{P}(E)$  is very big in comparison with  $E$ , it is not easy to give examples. If  $E = \emptyset$ , the situation is clear enough; the set  $\mathcal{P}(\emptyset)$  is

the singleton  $\{\emptyset\}$ . The power sets of singletons and pairs are also easily describable; we have

$$\wp(\{a\}) = \{\emptyset, \{a\}\}$$

and

$$\wp(\{a, b\}) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}.$$

The power set of a triple has eight elements. The reader can probably guess (and is hereby challenged to prove) the generalization that includes all these statements: the power set of a finite set with, say,  $n$  elements has  $2^n$  elements. (Of course concepts like “finite” and “ $2^n$ ” have no official standing for us yet; this should not prevent them from being unofficially understood.) The occurrence of  $n$  as an exponent (the  $n$ -th power of 2) has something to do with the reason why a power set bears its name.

If  $\mathcal{C}$  is a collection of subsets of a set  $E$  (that is,  $\mathcal{C}$  is a subcollection of  $\wp(E)$ ), then write

$$\mathfrak{D} = \{X \in \wp(E) : X' \in \mathcal{C}\}.$$

(To be certain that the condition used in the definition of  $\mathfrak{D}$  is a sentence in the precise technical sense, it must be rewritten in something like the form

*for some  $Y$  [ $Y \in \mathcal{C}$  and for all  $x$  ( $x \in X$  if and only if  $(x \in E$  and  $x \notin Y)$ )].*

Similar comments often apply when we wish to use defined abbreviations instead of logical and set-theoretic primitives only. The translation rarely requires any ingenuity and we shall usually omit it.) It is customary to denote the union and the intersection of the collection  $\mathfrak{D}$  by the symbols

$$\bigcup_{X \in \mathfrak{D}} X' \quad \text{and} \quad \bigcap_{X \in \mathfrak{D}} X'.$$

In this notation the general forms of the De Morgan laws become

$$(\bigcup_{X \in \mathfrak{D}} X)' = \bigcap_{X \in \mathfrak{D}} X'$$

and

$$(\bigcap_{X \in \mathfrak{D}} X)' = \bigcup_{X \in \mathfrak{D}} X'.$$

The proofs of these equations are immediate consequences of the appropriate definitions.

**EXERCISE.** Prove that  $\wp(E) \cap \wp(F) = \wp(E \cap F)$  and  $\wp(E) \cup \wp(F) \subset \wp(E \cup F)$ . These assertions can be generalized to

$$\bigcap_{X \in \mathfrak{D}} \wp(X) = \wp(\bigcap_{X \in \mathfrak{D}} X)$$

and

$$\bigcup_{X \in \mathfrak{D}} \wp(X) \subset \wp(\bigcup_{X \in \mathfrak{D}} X);$$

find a reasonable interpretation of the notation in which these generalizations were here expressed and then prove them. Further elementary facts:

$$\bigcap_{X \in \wp(E)} X = \emptyset,$$

and

$$\text{if } E \subset F, \text{ then } \wp(E) \subset \wp(F).$$

A curious question concerns the commutativity of the operators  $\wp$  and  $\bigcup$ . Show that  $E$  is always equal to  $\bigcup_{X \in \wp(E)} X$  (that is  $E = \bigcup \wp(E)$ ), but that the result of applying  $\wp$  and  $\bigcup$  to  $E$  in the other order is a set that includes  $E$  as a subset, typically a proper subset.

## SECTION 6

---

### ORDERED PAIRS

---

---

What does it mean to arrange the elements of a set  $A$  in some order? Suppose, for instance, that the set  $A$  is the quadruple  $\{a, b, c, d\}$  of distinct elements, and suppose that we want to consider its elements in the order

$$c \ b \ d \ a.$$

Even without a precise definition of what this means, we can do something set-theoretically intelligent with it. We can, namely, consider, for each particular spot in the ordering, the set of all those elements that occur at or before that spot; we obtain in this way the sets

$$\{c\} \quad \{c, b\} \quad \{c, b, d\} \quad \{c, b, d, a\}.$$

We can go on then to consider the set (or collection, if that sounds better)

$$C = \{\{a, b, c, d\}, \{b, c\}, \{b, c, d\}, \{c\}\}$$

that has exactly those sets for its elements. In order to emphasize that the intuitively based and possibly unclear concept of order has succeeded in producing something solid and simple, namely a plain, unembellished set  $C$ , the elements of  $C$ , and *their* elements, are presented above in a scrambled manner. (The lexicographically inclined reader might be able to see a method in the manner of scrambling.)

Let us continue to pretend for a while that we do know what order means. Suppose that in a hasty glance at the preceding paragraph all we could catch is the set  $C$ ; can we use it to recapture the order that gave rise to it? The answer is easily seen to be yes. Examine the elements of  $C$  (they themselves are sets, of course) to find one that is included in all the others; since  $\{c\}$  fills the bill (and nothing else does) we know that  $c$  must have been the first element. Look next for the next smallest element of  $C$ ,

i.e., the one that is included in all the ones that remain after  $\{c\}$  is removed; since  $\{b, c\}$  fills the bill (and nothing else does), we know that  $b$  must have been the second element. Proceeding thus (only two more steps are needed) we pass from the set  $\mathcal{C}$  to the given ordering of the given set  $A$ .

The moral is this: we may not know precisely what it means to order the elements of a set  $A$ , but with each order we can associate a set  $\mathcal{C}$  of subsets of  $A$  in such a way that the given order can be uniquely recaptured from  $\mathcal{C}$ . (Here is a non-trivial exercise: find an intrinsic characterization of those sets of subsets of  $A$  that correspond to some order in  $A$ . Since "order" has no official meaning for us yet, the whole problem is officially meaningless. Nothing that follows depends on the solution, but the reader would learn something valuable by trying to find it.) The passage from an order in  $A$  to the set  $\mathcal{C}$ , and back, was illustrated above for a quadruple; for a pair everything becomes at least twice as simple. If  $A = \{a, b\}$  and if, in the desired order,  $a$  comes first, then  $\mathcal{C} = \{\{a\}, \{a, b\}\}$ ; if, however,  $b$  comes first, then  $\mathcal{C} = \{\{b\}, \{a, b\}\}$ .

The *ordered pair* of  $a$  and  $b$ , with *first coordinate*  $a$  and *second coordinate*  $b$ , is the set  $(a, b)$  defined by

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

However convincing the motivation of this definition may be, we must still prove that the result has the main property that an ordered pair must have to deserve its name. We must show that if  $(a, b)$  and  $(x, y)$  are ordered pairs and if  $(a, b) = (x, y)$ , then  $a = x$  and  $b = y$ . To prove this, we note first that if  $a$  and  $b$  happen to be equal, then the ordered pair  $(a, b)$  is the same as the singleton  $\{\{a\}\}$ . If, conversely,  $(a, b)$  is a singleton, then  $\{a\} = \{a, b\}$ , so that  $b \in \{a\}$ , and therefore  $a = b$ . Suppose now that  $(a, b) = (x, y)$ . If  $a = b$ , then both  $(a, b)$  and  $(x, y)$  are singletons, so that  $x = y$ ; since  $\{x\} \in (a, b)$  and  $\{a\} \in (x, y)$ , it follows that  $a, b, x$ , and  $y$  are all equal. If  $a \neq b$ , then both  $(a, b)$  and  $(x, y)$  contain exactly one singleton, namely  $\{a\}$  and  $\{x\}$  respectively, so that  $a = x$ . Since in this case it is also true that both  $(a, b)$  and  $(x, y)$  contain exactly one unordered pair that is not a singleton, namely  $\{a, b\}$  and  $\{x, y\}$  respectively, it follows that  $\{a, b\} = \{x, y\}$ , and therefore, in particular,  $b \in \{x, y\}$ . Since  $b$  cannot be  $x$  (for then we should have  $a = x$  and  $b = x$ , and, therefore,  $a = b$ ), we must have  $b = y$ , and the proof is complete.

If  $A$  and  $B$  are sets, does there exist a set that contains all the ordered pairs  $(a, b)$  with  $a$  in  $A$  and  $b$  in  $B$ ? It is quite easy to see that the answer is yes. Indeed, if  $a \in A$  and  $b \in B$ , then  $\{a\} \subset A$  and  $\{b\} \subset B$ , and therefore  $\{a, b\} \subset A \cup B$ . Since also  $\{a\} \subset A \cup B$ , it follows that both  $\{a\}$

and  $\{a, b\}$  are elements of  $\mathcal{P}(A \cup B)$ . This implies that  $\{\{a\}, \{a, b\}\}$  is a subset of  $\mathcal{P}(A \cup B)$ , and hence that it is an element of  $\mathcal{P}(\mathcal{P}(A \cup B))$ ; in other words  $(a, b) \in \mathcal{P}(\mathcal{P}(A \cup B))$  whenever  $a \in A$  and  $b \in B$ . Once this is known, it is a routine matter to apply the axiom of specification and the axiom of extension to produce the unique set  $A \times B$  that consists exactly of the ordered pairs  $(a, b)$  with  $a$  in  $A$  and  $b$  in  $B$ . This set is called the *Cartesian product* of  $A$  and  $B$ ; it is characterized by the fact that

$$A \times B = \{x : x = (a, b) \text{ for some } a \text{ in } A \text{ and for some } b \text{ in } B\}.$$

The Cartesian product of two sets is a set of ordered pairs (that is, a set each of whose elements is an ordered pair), and the same is true of every subset of a Cartesian product. It is of technical importance to know that we can go in the converse direction also: every set of ordered pairs is a subset of the Cartesian product of two sets. In other words: if  $R$  is a set such that every element of  $R$  is an ordered pair, then there exist two sets  $A$  and  $B$  such that  $R \subset A \times B$ . The proof is elementary. Suppose indeed that  $x \in R$ , so that  $x = \{\{a\}, \{a, b\}\}$  for some  $a$  and for some  $b$ . The problem is to dig out  $a$  and  $b$  from under the braces. Since the elements of  $R$  are sets, we can form the union of the sets in  $R$ ; since  $x$  is one of the sets in  $R$ , the elements of  $x$  belong to that union. Since  $\{a, b\}$  is one of the elements of  $x$ , we may write, in what has been called the brutal notation above,  $\{a, b\} \in \bigcup R$ . One set of braces has disappeared; let us do the same thing again to make the other set go away. Form the union of the sets in  $\bigcup R$ . Since  $\{a, b\}$  is one of those sets, it follows that the elements of  $\{a, b\}$  belong to that union, and hence both  $a$  and  $b$  belong to  $\bigcup \bigcup R$ . This fulfills the promise made above; to exhibit  $R$  as a subset of some  $A \times B$ , we may take both  $A$  and  $B$  to be  $\bigcup \bigcup R$ . It is often desirable to take  $A$  and  $B$  as small as possible. To do so, just apply the axiom of specification to produce the sets

$$A = \{a : \text{for some } b ((a, b) \in R)\}$$

and

$$B = \{b : \text{for some } a ((a, b) \in R)\}.$$

These sets are called the *projections* of  $R$  onto the first and second coordinates respectively.

However important set theory may be now, when it began some scholars considered it a disease from which, it was to be hoped, mathematics would soon recover. For this reason many set-theoretic considerations were called pathological, and the word lives on in mathematical usage; it often refers to something the speaker does not like. The explicit definition of an

ordered pair  $((a, b) = \{\{a\}, \{a, b\}\})$  is frequently relegated to pathological set theory. For the benefit of those who think that in this case the name is deserved, we note that the definition has served its purpose by now and will never be used again. We need to know that ordered pairs are determined by and uniquely determine their first and second coordinates, that Cartesian products can be formed, and that every set of ordered pairs is a subset of some Cartesian product; which particular approach is used to achieve these ends is immaterial.

It is easy to locate the source of the mistrust and suspicion that many mathematicians feel toward the explicit definition of ordered pair given above. The trouble is not that there is anything wrong or anything missing; the relevant properties of the concept we have defined are all correct (that is, in accord with the demands of intuition) and all the correct properties are present. The trouble is that the concept has some irrelevant properties that are accidental and distracting. The theorem that  $(a, b) = (x, y)$  if and only if  $a = x$  and  $b = y$  is the sort of thing we expect to learn about ordered pairs. The fact that  $\{a, b\} \in (a, b)$ , on the other hand, seems accidental; it is a freak property of the definition rather than an intrinsic property of the concept.

The charge of artificiality is true, but it is not too high a price to pay for conceptual economy. The concept of an ordered pair could have been introduced as an additional primitive, axiomatically endowed with just the right properties, no more and no less. In some theories this is done. The mathematician's choice is between having to remember a few more axioms and having to forget a few accidental facts; the choice is pretty clearly a matter of taste. Similar choices occur frequently in mathematics; in this book, for instance, we shall encounter them again in connection with the definitions of numbers of various kinds.

EXERCISE. If  $A, B, X$ , and  $Y$  are sets, then

- (i)  $(A \cup B) \times X = (A \times X) \cup (B \times X)$ ,
- (ii)  $(A \cap B) \times (X \cap Y) = (A \times X) \cap (B \times Y)$ ,
- (iii)  $(A - B) \times X = (A \times X) - (B \times X)$ .

If either  $A = \emptyset$  or  $B = \emptyset$ , then  $A \times B = \emptyset$ , and conversely. If  $A \subset X$  and  $B \subset Y$ , then  $A \times B \subset X \times Y$ , and (provided  $A \times B \neq \emptyset$ ) conversely.

## SECTION 7

---

### RELATIONS

---

Using ordered pairs, we can formulate the mathematical theory of relations in set-theoretic language. By a relation we mean here something like marriage (between men and women) or belonging (between elements and sets). More explicitly, what we shall call a relation is sometimes called a *binary* relation. An example of a ternary relation is parenthood for people (Adam and Eve are the parents of Cain). In this book we shall have no occasion to treat the theory of relations that are ternary, quaternary, or worse.

Looking at any specific relation, such as marriage for instance, we might be tempted to consider certain ordered pairs  $(x, y)$ , namely just those for which  $x$  is a man,  $y$  is a woman, and  $x$  is married to  $y$ . We have not yet seen the definition of the general concept of a relation, but it seems plausible that, just as in this marriage example, every relation should uniquely determine the set of all those ordered pairs for which the first coordinate does stand in that relation to the second. If we know the relation, we know the set, and, better yet, if we know the set, we know the relation. If, for instance, we were presented with the set of ordered pairs of people that corresponds to marriage, then, even if we forgot the definition of marriage, we could always tell when a man  $x$  is married to a woman  $y$  and when not; we would just have to see whether the ordered pair  $(x, y)$  does or does not belong to the set.

We may not know what a relation is, but we do know what a set is, and the preceding considerations establish a close connection between relations and sets. The precise set-theoretic treatment of relations takes advantage of that heuristic connection; the simplest thing to do is to define a relation to be the corresponding set. This is what we do; we hereby define a *relation* as a set of ordered pairs. Explicitly: a set  $R$  is a relation if each ele-

ment of  $R$  is an ordered pair; this means, of course, that if  $z \in R$ , then there exist  $x$  and  $y$  so that  $z = (x, y)$ . If  $R$  is a relation, it is sometimes convenient to express the fact that  $(x, y) \in R$  by writing

$$x R y$$

and saying, as in everyday language, that  $x$  stands in the relation  $R$  to  $y$ .

The least exciting relation is the empty one. (To prove that  $\emptyset$  is a set of ordered pairs, look for an element of  $\emptyset$  that is not an ordered pair.) Another dull example is the Cartesian product of any two sets  $X$  and  $Y$ . Here is a slightly more interesting example: let  $X$  be any set, and let  $R$  be the set of all those pairs  $(x, y)$  in  $X \times X$  for which  $x = y$ . The relation  $R$  is just the relation of equality between elements of  $X$ ; if  $x$  and  $y$  are in  $X$ , then  $x R y$  means the same as  $x = y$ . One more example will suffice for now: let  $X$  be any set, and let  $R$  be the set of all those pairs  $(x, A)$  in  $X \times \wp(X)$  for which  $x \in A$ . This relation  $R$  is just the relation of belonging between elements of  $X$  and subsets of  $X$ ; if  $x \in X$  and  $A \in \wp(X)$ , then  $x R A$  means the same as  $x \in A$ .

In the preceding section we saw that associated with every set  $R$  of ordered pairs there are two sets called the projections of  $R$  onto the first and second coordinates. In the theory of relations these sets are known as the *domain* and the *range* of  $R$  (abbreviated  $\text{dom } R$  and  $\text{ran } R$ ); we recall that they are defined by

$$\text{dom } R = \{x: \text{for some } y (x R y)\}$$

and

$$\text{ran } R = \{y: \text{for some } x (x R y)\}.$$

If  $R$  is the relation of marriage, so that  $x R y$  means that  $x$  is a man,  $y$  is a woman, and  $x$  and  $y$  are married to one another, then  $\text{dom } R$  is the set of married men and  $\text{ran } R$  is the set of married women. Both the domain and the range of  $\emptyset$  are equal to  $\emptyset$ . If  $R = X \times Y$ , then  $\text{dom } R = X$  and  $\text{ran } R = Y$ . If  $R$  is equality in  $X$ , then  $\text{dom } R = \text{ran } R = X$ . If  $R$  is belonging, between  $X$  and  $\wp(X)$ , then  $\text{dom } R = X$  and  $\text{ran } R = \wp(X) - \{\emptyset\}$ .

If  $R$  is a relation included in a Cartesian product  $X \times Y$  (so that  $\text{dom } R \subset X$  and  $\text{ran } R \subset Y$ ), it is sometimes convenient to say that  $R$  is a relation from  $X$  to  $Y$ ; instead of a relation from  $X$  to  $X$  we may speak of a relation in  $X$ . A relation  $R$  in  $X$  is *reflexive* if  $x R x$  for every  $x$  in  $X$ ; it is *symmetric* if  $x R y$  implies that  $y R x$ ; and it is *transitive* if  $x R y$  and  $y R z$  imply that  $x R z$ . (Exercise: for each of these three possible properties, find a relation that does not have that property but does have the other two.) A relation

in a set is an *equivalence relation* if it is reflexive, symmetric, and transitive. The smallest equivalence relation in a set  $X$  is the relation of equality in  $X$ ; the largest equivalence relation in  $X$  is  $X \times X$ .

There is an intimate connection between equivalence relations in a set  $X$  and certain collections (called partitions) of subsets of  $X$ . A *partition* of  $X$  is a disjoint collection  $\mathcal{C}$  of non-empty subsets of  $X$  whose union is  $X$ . If  $R$  is an equivalence relation in  $X$ , and if  $x$  is in  $X$ , the *equivalence class* of  $x$  with respect to  $R$  is the set of all those elements  $y$  in  $X$  for which  $x R y$ . (The weight of tradition makes the use of the word "class" at this point unavoidable.) Examples: if  $R$  is equality in  $X$ , then each equivalence class is a singleton; if  $R = X \times X$ , then the set  $X$  itself is the only equivalence class. There is no standard notation for the equivalence class of  $x$  with respect to  $R$ ; we shall usually denote it by  $x/R$ , and we shall write  $X/R$  for the set of all equivalence classes. (Pronounce  $X/R$  as " $X$  modulo  $R$ ," or, in abbreviated form, " $X$  mod  $R$ ." Exercise: show that  $X/R$  is indeed a set by exhibiting a condition that specifies exactly the subset  $X/R$  of the power set  $\mathcal{P}(X)$ .) Now forget  $R$  for a moment and begin anew with a partition  $\mathcal{C}$  of  $X$ . A relation, which we shall call  $X/\mathcal{C}$ , is defined in  $X$  by writing

$$x \quad X/\mathcal{C} \quad y$$

just in case  $x$  and  $y$  belong to the same set of the collection  $\mathcal{C}$ . We shall call  $X/\mathcal{C}$  the relation *induced* by the partition  $\mathcal{C}$ .

In the preceding paragraph we saw how to associate a set of subsets of  $X$  with every equivalence relation in  $X$  and how to associate a relation in  $X$  with every partition of  $X$ . The connection between equivalence relations and partitions can be described by saying that the passage from  $\mathcal{C}$  to  $X/\mathcal{C}$  is exactly the reverse of the passage from  $R$  to  $X/R$ . More explicitly: if  $R$  is an equivalence relation in  $X$ , then the set of equivalence classes is a partition of  $X$  that induces the relation  $R$ , and if  $\mathcal{C}$  is a partition of  $X$ , then the induced relation is an equivalence relation whose set of equivalence classes is exactly  $\mathcal{C}$ .

For the proof, let us start with an equivalence relation  $R$ . Since each  $x$  belongs to some equivalence class (for instance  $x \in x/R$ ), it is clear that the union of the equivalence classes is all  $X$ . If  $z \in x/R \cap y/R$ , then  $x R z$  and  $z R y$ , and therefore  $x R y$ . This implies that if two equivalence classes have an element in common, then they are identical, or, in other words, that two distinct equivalence classes are always disjoint. The set of equivalence classes is therefore a partition. To say that two elements belong to the same set (equivalence class) of this partition means, by defini-

tion, that they stand in the relation  $R$  to one another. This proves the first half of our assertion.

The second half is easier. Start with a partition  $\mathcal{C}$  and consider the induced relation. Since every element of  $X$  belongs to some set of  $\mathcal{C}$ , reflexivity just says that  $x$  and  $x$  are in the same set of  $\mathcal{C}$ . Symmetry says that if  $x$  and  $y$  are in the same set of  $\mathcal{C}$ , then  $y$  and  $x$  are in the same set of  $\mathcal{C}$ , and this is obviously true. Transitivity says that if  $x$  and  $y$  are in the same set of  $\mathcal{C}$  and if  $y$  and  $z$  are in the same set of  $\mathcal{C}$ , then  $x$  and  $z$  are in the same set of  $\mathcal{C}$ , and this too is obvious. The equivalence class of each  $x$  in  $X$  is just the set of  $\mathcal{C}$  to which  $x$  belongs. This completes the proof of everything that was promised.

## SECTION 8

---

### FUNCTIONS

---

If  $X$  and  $Y$  are sets, a *function* from (or *on*)  $X$  to (or *into*)  $Y$  is a relation  $f$  such that  $\text{dom } f = X$  and such that for each  $x$  in  $X$  there is a unique element  $y$  in  $Y$  with  $(x, y) \in f$ . The uniqueness condition can be formulated explicitly as follows: if  $(x, y) \in f$  and  $(x, z) \in f$ , then  $y = z$ . For each  $x$  in  $X$ , the unique  $y$  in  $Y$  such that  $(x, y) \in f$  is denoted by  $f(x)$ . For functions this notation and its minor variants supersede the others used for more general relations; from now on, if  $f$  is a function, we shall write  $f(x) = y$  instead of  $(x, y) \in f$  or  $x f y$ . The element  $y$  is called the *value* that the function  $f$  *assumes* (or *takes on*) at the *argument*  $x$ ; equivalently we may say that  $f$  *sends* or *maps* or *transforms*  $x$  onto  $y$ . The words *map* or *mapping*, *transformation*, *correspondence*, and *operator* are among some of the many that are sometimes used as synonyms for *function*. The symbol

$$f: X \rightarrow Y$$

is sometimes used as an abbreviation for “ $f$  is a function from  $X$  to  $Y$ .” The set of all functions from  $X$  to  $Y$  is a subset of the power set  $\mathcal{P}(X \times Y)$ ; it will be denoted by  $Y^X$ .

The connotations of activity suggested by the synonyms listed above make some scholars dissatisfied with the definition according to which a function does not *do* anything but merely *is*. This dissatisfaction is reflected in a different use of the vocabulary: *function* is reserved for the undefined object that is somehow active, and the set of ordered pairs that we have called the function is then called the *graph* of the function. It is easy to find examples of functions in the precise set-theoretic sense of the word in both mathematics and everyday life; all we have to look for is information, not necessarily numerical, in tabulated form. One example

is a city directory; the arguments of the function are, in this case, the inhabitants of the city, and the values are their addresses.

For relations in general, and hence for functions in particular, we have defined the concepts of domain and range. The domain of a function  $f$  from  $X$  into  $Y$  is, by definition, equal to  $X$ , but its range need not be equal to  $Y$ ; the range consists of those elements  $y$  of  $Y$  for which there exists an  $x$  in  $X$  such that  $f(x) = y$ . If the range of  $f$  is equal to  $Y$ , we say that  $f$  maps  $X$  onto  $Y$ . If  $A$  is a subset of  $X$ , we may want to consider the set of all those elements  $y$  of  $Y$  for which there exists an  $x$  in the subset  $A$  such that  $f(x) = y$ . This subset of  $Y$  is called the *image* of  $A$  under  $f$  and is frequently denoted by  $f(A)$ . The notation is bad but not catastrophic. What is bad about it is that if  $A$  happens to be both an element of  $X$  and a subset of  $X$  (an unlikely situation, but far from an impossible one), then the symbol  $f(A)$  is ambiguous. Does it mean the value of  $f$  at  $A$  or does it mean the set of values of  $f$  at the elements of  $A$ ? Following normal mathematical custom, we shall use the bad notation, relying on context, and, on the rare occasions when it is necessary, adding verbal stipulations, to avoid confusion. Note that the image of  $X$  itself is the range of  $f$ ; the "onto" character of  $f$  can be expressed by writing  $f(X) = Y$ .

If  $X$  is a subset of a set  $Y$ , the function  $f$  defined by  $f(x) = x$  for each  $x$  in  $X$  is called the *inclusion map* (or the *embedding*, or the *injection*) of  $X$  into  $Y$ . The phrase "the function  $f$  defined by . . ." is a very common one in such contexts. It is intended to imply, of course, that there does indeed exist a unique function satisfying the stated condition. In the special case at hand this is obvious enough; we are being invited to consider the set of all those ordered pairs  $(x, y)$  in  $X \times Y$  for which  $x = y$ . Similar considerations apply in every case, and, following normal mathematical practice, we shall usually describe a function by describing its value  $y$  at each argument  $x$ . Such a description is sometimes longer and more cumbersome than a direct description of the set (of ordered pairs) involved, but, nevertheless, most mathematicians regard the argument-value description as more perspicuous than any other.

The inclusion map of  $X$  into  $X$  is called the *identity map* on  $X$ . (In the language of relations, the identity map on  $X$  is the same as the relation of equality in  $X$ .) If, as before,  $X \subset Y$ , then there is a connection between the inclusion map of  $X$  into  $Y$  and the identity map on  $Y$ ; that connection is a special case of a general procedure for making small functions out of large ones. If  $f$  is a function from  $Y$  to  $Z$ , say, and if  $X$  is a subset of  $Y$ , then there is a natural way of constructing a function  $g$  from  $X$  to  $Z$ ; define  $g(x)$  to be equal to  $f(x)$  for each  $x$  in  $X$ . The function  $g$  is called the

*restriction* of  $f$  to  $X$ , and  $f$  is called an *extension* of  $g$  to  $Y$ ; it is customary to write  $g = f|X$ . The definition of restriction can be expressed by writing  $(f|X)(x) = f(x)$  for each  $x$  in  $X$ ; observe also that  $\text{ran}(f|X) = f(X)$ . The inclusion map of a subset of  $Y$  is the restriction to that subset of the identity map on  $Y$ .

Here is a simple but useful example of a function. Consider any two sets  $X$  and  $Y$ , and define a function  $f$  from  $X \times Y$  onto  $X$  by writing  $f(x, y) = x$ . (The purist will have noted that we should have written  $f((x, y))$  instead of  $f(x, y)$ , but nobody ever does.) The function  $f$  is called the *projection* from  $X \times Y$  onto  $X$ ; if, similarly,  $g(x, y) = y$ , then  $g$  is the projection from  $X \times Y$  onto  $Y$ . The terminology here is at variance with an earlier one, but not too badly. If  $R = X \times Y$ , then what was earlier called the projection of  $R$  onto the first coordinate is, in the present language, the range of the projection  $f$ .

A more complicated and correspondingly more valuable example of a function can be obtained as follows. Suppose  $R$  is an equivalence relation in  $X$ , and let  $f$  be the function from  $X$  onto  $X/R$  defined by  $f(x) = x/R$ . The function  $f$  is sometimes called the *canonical map* from  $X$  to  $X/R$ .

If  $f$  is an arbitrary function, from  $X$  onto  $Y$ , then there is a natural way of defining an equivalence relation  $R$  in  $X$ ; write  $a R b$  (where  $a$  and  $b$  are in  $X$ ) in case  $f(a) = f(b)$ . For each element  $y$  of  $Y$ , let  $g(y)$  be the set of all those elements  $x$  in  $X$  for which  $f(x) = y$ . The definition of  $R$  implies that  $g(y)$  is, for each  $y$ , an equivalence class of the relation  $R$ ; in other words,  $g$  is a function from  $Y$  onto the set  $X/R$  of all equivalence classes of  $R$ . The function  $g$  has the following special property: if  $u$  and  $v$  are distinct elements of  $Y$ , then  $g(u)$  and  $g(v)$  are distinct elements of  $X/R$ . A function that always maps distinct elements onto distinct elements is called *one-to-one* (usually a *one-to-one correspondence*). Among the examples above the inclusion maps are one-to-one, but, except in some trivial special cases, the projections are not. (Exercise: what special cases?)

To introduce the next aspect of the elementary theory of functions we must digress for a moment and anticipate a tiny fragment of our ultimate definition of natural numbers. We shall not find it necessary to define all the natural numbers now; all we need is the first three of them. Since this is not the appropriate occasion for lengthy heuristic preliminaries, we shall proceed directly to the definition, even at the risk of temporarily shocking or worrying some readers. Here it is: we define 0, 1, and 2 by writing

$$0 = \emptyset, \quad 1 = \{\emptyset\}, \quad \text{and} \quad 2 = \{\emptyset, \{\emptyset\}\}.$$

In other words, 0 is empty, 1 is the singleton  $\{0\}$ , and 2 is the pair  $\{0, 1\}$ .

Observe that there is some method in this apparent madness; the number of elements in the sets 0, 1, or 2 (in the ordinary everyday sense of the word) is, respectively, zero, one, or two.

If  $A$  is a subset of a set  $X$ , the *characteristic function* of  $A$  is the function  $\chi$  from  $X$  to 2 such that  $\chi(x) = 1$  or 0 according as  $x \in A$  or  $x \in X - A$ . The dependence of the characteristic function of  $A$  on the set  $A$  may be indicated by writing  $\chi_A$  instead of  $\chi$ . The function that assigns to each subset  $A$  of  $X$  (that is, to each element of  $\wp(X)$ ) the characteristic function of  $A$  (that is, an element of  $2^X$ ) is a one-to-one correspondence between  $\wp(X)$  and  $2^X$ . (Parenthetically: instead of the phrase "the function that assigns to each  $A$  in  $\wp(X)$  the element  $\chi_A$  in  $2^X$ " it is customary to use the abbreviation "the function  $A \rightarrow \chi_A$ ." In this language, the projection from  $X \times Y$  onto  $X$ , for instance, may be called the function  $(x, y) \rightarrow x$ , and the canonical map from a set  $X$  with a relation  $R$  onto  $X/R$  may be called the function  $x \rightarrow x/R$ .)

**EXERCISE.** (i)  $Y^\emptyset$  has exactly one element, namely  $\emptyset$ , whether  $Y$  is empty or not, and (ii) if  $X$  is not empty, then  $\emptyset^X$  is empty.

## SECTION 9

---

### FAMILIES

---

There are occasions when the range of a function is deemed to be more important than the function itself. When that is the case, both the terminology and the notation undergo radical alterations. Suppose, for instance, that  $x$  is a function from a set  $I$  to a set  $X$ . (The very choice of letters indicates that something strange is afoot.) An element of the domain  $I$  is called an *index*,  $I$  is called the *index set*, the range of the function is called an *indexed set*, the function itself is called a *family*, and the value of the function  $x$  at an index  $i$ , called a *term* of the family, is denoted by  $x_i$ . (This terminology is not absolutely established, but it is one of the standard choices among related slight variants; in the sequel it and it alone will be used.) An unacceptable but generally accepted way of communicating the notation and indicating the emphasis is to speak of a family  $\{x_i\}$  in  $X$ , or of a family  $\{x_i\}$  of whatever the elements of  $X$  may be; when necessary, the index set  $I$  is indicated by some such parenthetical expression as  $(i \in I)$ . Thus, for instance, the phrase “a family  $\{A_i\}$  of subsets of  $X$ ” is usually understood to refer to a function  $A$ , from some set  $I$  of indices, into  $\mathcal{P}(X)$ .

If  $\{A_i\}$  is a family of subsets of  $X$ , the union of the range of the family is called the union of the family  $\{A_i\}$ , or the union of the sets  $A_i$ ; the standard notation for it is

$$\bigcup_{i \in I} A_i \text{ or } \bigcup_i A_i,$$

according as it is or is not important to emphasize the index set  $I$ . It follows immediately from the definition of unions that  $x \in \bigcup_i A_i$  if and only if  $x$  belongs to  $A_i$  for at least one  $i$ . If  $I = 2$ , so that the range of the family  $\{A_i\}$  is the unordered pair  $\{A_0, A_1\}$ , then  $\bigcup_i A_i = A_0 \cup A_1$ . Observe that there is no loss of generality in considering families of sets instead of arbitrary collections of sets; every collection of sets is the range

of some family. If, indeed,  $\mathcal{C}$  is a collection of sets, let  $\mathcal{C}$  itself play the role of the index set, and consider the identity mapping on  $\mathcal{C}$  in the role of the family.

The algebraic laws satisfied by the operation of union for pairs can be generalized to arbitrary unions. Suppose, for instance, that  $\{I_j\}$  is a family of sets with domain  $J$ , say; write  $K = \bigcup_j I_j$ , and let  $\{A_k\}$  be a family of sets with domain  $K$ . It is then not difficult to prove that

$$\bigcup_{k \in K} A_k = \bigcup_{j \in J} (\bigcup_{i \in I_j} A_i);$$

this is the generalized version of the associative law for unions. Exercise: formulate and prove a generalized version of the commutative law.

An empty union makes sense (and is empty), but an empty intersection does not make sense. Except for this triviality, the terminology and notation for intersections parallels that for unions in every respect. Thus, for instance, if  $\{A_i\}$  is a non-empty family of sets, the intersection of the range of the family is called the intersection of the family  $\{A_i\}$ , or the intersection of the sets  $A_i$ ; the standard notation for it is

$$\bigcap_{i \in I} A_i \text{ or } \bigcap_i A_i,$$

according as it is or is not important to emphasize the index set  $I$ . (By a “non-empty family” we mean a family whose domain  $I$  is not empty.) It follows immediately from the definition of intersections that if  $I \neq \emptyset$ , then a necessary and sufficient condition that  $x$  belong to  $\bigcap_i A_i$  is that  $x$  belong to  $A_i$  for all  $i$ .

The generalized commutative and associative laws for intersections can be formulated and proved the same way as for unions, or, alternatively, De Morgan’s laws can be used to derive them from the facts for unions. This is almost obvious, and, therefore, it is not of much interest. The interesting algebraic identities are the ones that involve both unions and intersections. Thus, for instance, if  $\{A_i\}$  is a family of subsets of  $X$  and  $B \subset X$ , then

$$B \cap \bigcup_i A_i = \bigcup_i (B \cap A_i)$$

and

$$B \cup \bigcap_i A_i = \bigcap_i (B \cup A_i);$$

these equations are a mild generalization of the distributive laws.

**EXERCISE.** If both  $\{A_i\}$  and  $\{B_j\}$  are families of sets, then

$$(\bigcup_i A_i) \cap (\bigcup_j B_j) = \bigcup_{i,j} (A_i \cap B_j)$$

and

$$(\bigcap_i A_i) \cup (\bigcap_j B_j) = \bigcap_{i,j} (A_i \cup B_j).$$

Explanation of notation: a symbol such as  $\bigcup_{(i,j) \in I \times J}$  is an abbreviation for  $\bigcup_{(i,j) \in I \times J} z_j$ .

The notation of families is the one normally used in generalizing the concept of Cartesian product. The Cartesian product of two sets  $X$  and  $Y$  was defined as the set of all ordered pairs  $(x, y)$  with  $x$  in  $X$  and  $y$  in  $Y$ . There is a natural one-to-one correspondence between this set and a certain set of families. Consider, indeed, any particular unordered pair  $\{a, b\}$ , with  $a \neq b$ , and consider the set  $Z$  of all families  $z$ , indexed by  $\{a, b\}$ , such that  $z_a \in X$  and  $z_b \in Y$ . If the function  $f$  from  $Z$  to  $X \times Y$  is defined by  $f(z) = (z_a, z_b)$ , then  $f$  is the promised one-to-one correspondence. The difference between  $Z$  and  $X \times Y$  is merely a matter of notation. The generalization of Cartesian products generalizes  $Z$  rather than  $X \times Y$  itself. (As a consequence there is a little terminological friction in the passage from the special case to the general. There is no help for it; that is how mathematical language is in fact used nowadays.) The generalization is now straightforward. If  $\{X_i\}$  is a family of sets ( $i \in I$ ), the *Cartesian product* of the family is, by definition, the set of all families  $\{x_i\}$  with  $x_i \in X_i$  for each  $i$  in  $I$ . There are several symbols for the Cartesian product in more or less current usage; in this book we shall denote it by

$$\bigtimes_{i \in I} X_i \text{ or } X_i X_i.$$

It is clear that if every  $X_i$  is equal to one and the same set  $X$ , then  $\bigtimes_i X_i = X^I$ . If  $I$  is a pair  $\{a, b\}$ , with  $a \neq b$ , then it is customary to identify  $\bigtimes_{i \in I} X_i$  with the Cartesian product  $X_a \times X_b$  as defined earlier, and if  $I$  is a singleton  $\{a\}$ , then, similarly, we identify  $\bigtimes_{i \in I} X_i$  with  $X_a$  itself. *Ordered triples*, *ordered quadruples*, etc., may be defined as families whose index sets are unordered triples, quadruples, etc.

Suppose that  $\{X_i\}$  is a family of sets ( $i \in I$ ) and let  $X$  be its Cartesian product. If  $J$  is a subset of  $I$ , then to each element of  $X$  there corresponds in a natural way an element of the partial Cartesian product  $\bigtimes_{i \in J} X_i$ . To define the correspondence, recall that each element  $x$  of  $X$  is itself a family  $\{x_i\}$ , that is, in the last analysis, a function on  $I$ ; the corresponding element, say  $y$ , of  $\bigtimes_{i \in J} X_i$  is obtained by simply restricting that function to  $J$ . Explicitly, we write  $y_i = x_i$  whenever  $i \in J$ . The correspondence  $x \rightarrow y$  is called the projection from  $X$  onto  $\bigtimes_{i \in J} X_i$ ; we shall temporarily denote it by  $f_J$ . If, in particular,  $J$  is a singleton, say  $J = \{j\}$ , then we shall write  $f_j$  (instead of  $f_{\{j\}}$ ) for  $f_J$ . The word "projection" has a multiple use; if  $x \in X$ , the value of  $f_j$  at  $x$ , that is  $x_j$ , is also called the projection of  $x$  onto  $X_j$ , or, alternatively, the *j-coordinate* of  $x$ . A function on a Car-

sian product such as  $X$  is called a function of *several variables*, and, in particular, a function on a Cartesian product  $X_a \times X_b$  is called a function of two variables.

**EXERCISE.** Prove that  $(\bigcup_i A_i) \times (\bigcup_j B_j) = \bigcup_{i,j} (A_i \times B_j)$ , and that the same equation holds for intersections (provided that the domains of the families involved are not empty). Prove also (with appropriate provisos about empty families) that  $\bigcap_i X_i \subset X_j \subset \bigcup_i X_i$  for each index  $j$  and that intersection and union can in fact be characterized as the extreme solutions of these inclusions. This means that if  $X_j \subset Y$  for each index  $j$ , then  $\bigcup_i X_i \subset Y$ , and that  $\bigcup_i X_i$  is the only set satisfying this minimality condition; the formulation for intersections is similar.

## SECTION 10

---

### INVERSES AND COMPOSITES

---

Associated with every function  $f$ , from  $X$  to  $Y$ , say, there is a function from  $\mathcal{P}(X)$  to  $\mathcal{P}(Y)$ , namely the function (frequently called  $f$  also) that assigns to each subset  $A$  of  $X$  the image subset  $f(A)$  of  $Y$ . The algebraic behavior of the mapping  $A \rightarrow f(A)$  leaves something to be desired. It is true that if  $\{A_i\}$  is a family of subsets of  $X$ , then  $f(\bigcup_i A_i) = \bigcup_i f(A_i)$  (proof?), but the corresponding equation for intersections is false in general (example?), and the connection between images and complements is equally unsatisfactory.

A correspondence between the elements of  $X$  and the elements of  $Y$  does always induce a well-behaved correspondence between the subsets of  $X$  and the subsets of  $Y$ , not forward, by the formation of images, but backward, by the formation of inverse images. Given a function  $f$  from  $X$  to  $Y$ , let  $f^{-1}$ , the *inverse* of  $f$ , be the function from  $\mathcal{P}(Y)$  to  $\mathcal{P}(X)$  such that if  $B \subset Y$ , then

$$f^{-1}(B) = \{x \in X : f(x) \in B\}.$$

In words:  $f^{-1}(B)$  consists of exactly those elements of  $X$  that  $f$  maps into  $B$ ; the set  $f^{-1}(B)$  is called the *inverse image* of  $B$  under  $f$ . A necessary and sufficient condition that  $f$  map  $X$  onto  $Y$  is that the inverse image under  $f$  of each non-empty subset of  $Y$  be a non-empty subset of  $X$ . (Proof?) A necessary and sufficient condition that  $f$  be one-to-one is that the inverse image under  $f$  of each singleton in the range of  $f$  be a singleton in  $X$ .

If the last condition is satisfied, then the symbol  $f^{-1}$  is frequently assigned a second interpretation, namely as the function whose domain is the range of  $f$ , and whose value for each  $y$  in the range of  $f$  is the unique  $x$  in  $X$  for which  $f(x) = y$ . In other words, for one-to-one functions  $f$  we may write  $f^{-1}(y) = x$  if and only if  $f(x) = y$ . This use of the notation is

mildly inconsistent with our first interpretation of  $f^{-1}$ , but the double meaning is not likely to lead to any confusion.

The connection between images and inverse images is worth a moment's consideration.

If  $B \subset Y$ , then

$$f(f^{-1}(B)) \subset B.$$

**Proof.** If  $y \in f(f^{-1}(B))$ , then  $y = f(x)$  for some  $x$  in  $f^{-1}(B)$ ; this means that  $y = f(x)$  and  $f(x) \in B$ , and therefore  $y \in B$ .

If  $f$  maps  $X$  onto  $Y$ , then

$$f(f^{-1}(B)) = B.$$

**Proof.** If  $y \in B$ , then  $y = f(x)$  for some  $x$  in  $X$ , and therefore for some  $x$  in  $f^{-1}(B)$ ; this means that  $y \in f(f^{-1}(B))$ .

If  $A \subset X$ , then

$$A \subset f^{-1}(f(A)).$$

**Proof.** If  $x \in A$ , then  $f(x) \in f(A)$ ; this means that  $x \in f^{-1}(f(A))$ .

If  $f$  is one-to-one, then

$$A = f^{-1}(f(A)).$$

**Proof.** If  $x \in f^{-1}(f(A))$ , then  $f(x) \in f(A)$ , and therefore  $f(x) = f(u)$  for some  $u$  in  $A$ ; this implies that  $x = u$  and hence that  $x \in A$ .

The algebraic behavior of  $f^{-1}$  is unexceptionable. If  $\{B_i\}$  is a family of subsets of  $Y$ , then

$$f^{-1}(\bigcup_i B_i) = \bigcup_i f^{-1}(B_i)$$

and

$$f^{-1}(\bigcap_i B_i) = \bigcap_i f^{-1}(B_i).$$

The proofs are straightforward. If, for instance,  $x \in f^{-1}(\bigcap_i B_i)$ , then  $f(x) \in B_i$  for all  $i$ , so that  $x \in f^{-1}(B_i)$  for all  $i$ , and therefore  $x \in \bigcap_i f^{-1}(B_i)$ ; all the steps in this argument are reversible. The formation of inverse images commutes with complementation also; i.e.,

$$f^{-1}(Y - B) = X - f^{-1}(B)$$

for each subset  $B$  of  $Y$ . Indeed: if  $x \in f^{-1}(Y - B)$ , then  $f(x) \in Y - B$ , so that  $x \notin f^{-1}(B)$ , and therefore  $x \in X - f^{-1}(B)$ ; the steps are reversible. (Observe that the last equation is indeed a kind of commutative law: it says that complementation followed by inversion is the same as inversion followed by complementation.)

The discussion of inverses shows that what a function does can in a cer-

tain sense be undone; the next thing we shall see is that what two functions do can sometimes be done in one step. If, to be explicit,  $f$  is a function from  $X$  to  $Y$  and  $g$  is a function from  $Y$  to  $Z$ , then every element in the range of  $f$  belongs to the domain of  $g$ , and, consequently,  $g(f(x))$  makes sense for each  $x$  in  $X$ . The function  $h$  from  $X$  to  $Z$ , defined by  $h(x) = g(f(x))$  is called the *composite* of the functions  $f$  and  $g$ ; it is denoted by  $g \circ f$  or, more simply, by  $gf$ . (Since we shall not have occasion to consider any other kind of multiplication for functions, in this book we shall use the latter, simpler notation only.)

Observe that the order of events is important in the theory of functional composition. In order that  $gf$  be defined, the range of  $f$  must be included in the domain of  $g$ , and this can happen without it necessarily happening in the other direction at the same time. Even if both  $fg$  and  $gf$  are defined, which happens if, for instance,  $f$  maps  $X$  into  $Y$  and  $g$  maps  $Y$  into  $X$ , the functions  $fg$  and  $gf$  need not be the same; in other words, functional composition is not necessarily commutative.

Functional composition may not be commutative, but it is always associative. If  $f$  maps  $X$  into  $Y$ , if  $g$  maps  $Y$  into  $Z$ , and if  $h$  maps  $Z$  into  $U$ , then we can form the composite of  $h$  with  $gf$  and the composite of  $hg$  with  $f$ ; it is a simple exercise to show that the result is the same in either case.

The connection between inversion and composition is important; something like it crops up all over mathematics. If  $f$  maps  $X$  into  $Y$  and  $g$  maps  $Y$  into  $Z$ , then  $f^{-1}$  maps  $\wp(Y)$  into  $\wp(X)$  and  $g^{-1}$  maps  $\wp(Z)$  into  $\wp(Y)$ . In this situation, the composites that are formable are  $gf$  and  $f^{-1}g^{-1}$ ; the assertion is that the latter is the inverse of the former. Proof: if  $x \in (gf)^{-1}(C)$ , where  $x \in X$  and  $C \subset Z$ , then  $g(f(x)) \in C$ , so that  $f(x) \in g^{-1}(C)$ , and therefore  $x \in f^{-1}(g^{-1}(C))$ ; the steps of the argument are reversible.

Inversion and composition for functions are special cases of similar operations for relations. Thus, in particular, associated with every relation  $R$  from  $X$  to  $Y$  there is the *inverse* (or *converse*) relation  $R^{-1}$  from  $Y$  to  $X$ ; by definition  $y R^{-1} x$  means that  $x R y$ . Example: if  $R$  is the relation of belonging, from  $X$  to  $\wp(X)$ , then  $R^{-1}$  is the relation of containing, from  $\wp(X)$  to  $X$ . It is an immediate consequence of the definitions involved that  $\text{dom } R^{-1} = \text{ran } R$  and  $\text{ran } R^{-1} = \text{dom } R$ . If the relation  $R$  is a function, then the equivalent assertions  $x R y$  and  $y R^{-1} x$  can be written in the equivalent forms  $R(x) = y$  and  $x \in R^{-1}(\{y\})$ .

Because of difficulties with commutativity, the generalization of functional composition has to be handled with care. The composite of the relations  $R$  and  $S$  is defined in case  $R$  is a relation from  $X$  to  $Y$  and  $S$  is a rela-

tion from  $Y$  to  $Z$ . The composite relation  $T$ , from  $X$  to  $Z$ , is denoted by  $S \circ R$ , or, simply, by  $SR$ ; it is defined so that  $x T z$  if and only if there exists an element  $y$  in  $Y$  such that  $x R y$  and  $y S z$ . For an instructive example, let  $R$  mean “son” and let  $S$  mean “brother” in the set of human males, say. In other words,  $x R y$  means that  $x$  is a son of  $y$ , and  $y S z$  means that  $y$  is a brother of  $z$ . In this case the composite relation  $SR$  means “nephew.” (Query: what do  $R^{-1}$ ,  $S^{-1}$ ,  $RS$ , and  $R^{-1}S^{-1}$  mean?) If both  $R$  and  $S$  are functions, then  $x R y$  and  $y S z$  can be rewritten as  $R(x) = y$  and  $S(y) = z$ , respectively. It follows that  $S(R(x)) = z$  if and only if  $x T z$ , so that functional composition is indeed a special case of what is sometimes called the *relative product*.

The algebraic properties of inversion and composition are the same for relations as for functions. Thus, in particular, composition is commutative by accident only, but it is always associative, and it is always connected with inversion via the equation  $(SR)^{-1} = R^{-1}S^{-1}$ . (Proofs?)

The algebra of relations provides some amusing formulas. Suppose that, temporarily, we consider relations in one set  $X$  only, and, in particular, let  $I$  be the relation of equality in  $X$  (which is the same as the identity mapping on  $X$ ). The relation  $I$  acts as a multiplicative unit; this means that  $IR = RI = R$  for every relation  $R$  in  $X$ . Query: is there a connection among  $I$ ,  $RR^{-1}$ , and  $R^{-1}R$ ? The three defining properties of an equivalence relation can be formulated in algebraic terms as follows: reflexivity means  $I \subset R$ , symmetry means  $R \subset R^{-1}$ , and transitivity means  $RR \subset R$ .

**EXERCISE.** (Assume in each case that  $f$  is a function from  $X$  to  $Y$ .)

- (i) If  $g$  is a function from  $Y$  to  $X$  such that  $gf$  is the identity on  $X$ , then  $f$  is one-to-one and  $g$  maps  $Y$  onto  $X$ .
- (ii) A necessary and sufficient condition that  $f(A \cap B) = f(A) \cap f(B)$  for all subsets  $A$  and  $B$  of  $X$  is that  $f$  be one-to-one.
- (iii) A necessary and sufficient condition that  $f(X - A) \subset Y - f(A)$  for all subsets  $A$  of  $X$  is that  $f$  be one-to-one.
- (iv) A necessary and sufficient condition that  $Y - f(A) \subset f(X - A)$  for all subsets  $A$  of  $X$  is that  $f$  map  $X$  onto  $Y$ .

## SECTION 11

---

### NUMBERS

---

---

How much is two? How, more generally, are we to define numbers? To prepare for the answer, let us consider a set  $X$  and let us form the collection  $P$  of all unordered pairs  $\{a, b\}$ , with  $a$  in  $X$ ,  $b$  in  $X$ , and  $a \neq b$ . It seems clear that all the sets in the collection  $P$  have a property in common, namely the property of consisting of two elements. It is tempting to try to define "twoness" as the common property of all the sets in the collection  $P$ , but the temptation must be resisted; such a definition is, after all, mathematical nonsense. What is a "property"? How do we know that there is only one property in common to all the sets in  $P$ ?

After some cogitation we might hit upon a way of saving the idea behind the proposed definition without using vague expressions such as "the common property." It is ubiquitous mathematical practice to identify a property with a set, namely with the set of all objects that possess the property; why not do it here? Why not, in other words, define "two" as the set  $P$ ? Something like this is done at times, but it is not completely satisfying. The trouble is that our present modified proposal depends on  $P$ , and hence ultimately on  $X$ . At best the proposal defines twoness for subsets of  $X$ ; it gives no hint as to when we may attribute twoness to a set that is not included in  $X$ .

There are two ways out. One way is to abandon the restriction to a particular set and to consider instead all possible unordered pairs  $\{a, b\}$  with  $a \neq b$ . These unordered pairs do not constitute a set; in order to base the definition of "two" on them, the entire theory under consideration would have to be extended to include the "unsets" (classes) of another theory. This can be done, but it will not be done here; we shall follow a different route.

How would a mathematician define a meter? The procedure analogous

to the one sketched above would involve the following two steps. First, select an object that is one of the intended models of the concept being defined—an object, in other words, such that on intuitive or practical grounds it deserves to be called one meter long if anything does. Second, form the set of all objects in the universe that are of the same length as the selected one (note that this does not depend on knowing what a meter is), and define a meter as the set so formed.

How in fact is a meter defined? The example was chosen so that the answer to this question should suggest an approach to the definition of numbers. The point is that in the customary definition of a meter the second step is omitted. By a more or less arbitrary convention an object is selected and its length is called a meter. If the definition is accused of circularity (what does "length" mean?), it can easily be converted into an unexceptionable demonstrative definition; there is after all nothing to stop us from defining a meter as equal to the selected object. If this demonstrative approach is adopted, it is just as easy to explain as before when "one-meter-ness" shall be attributed to some other object, namely, just in case the new object has the same length as the selected standard. We comment again that to determine whether two objects have the same length depends on a simple act of comparison only, and does not depend on having a precise definition of length.

Motivated by the considerations described above, we have earlier defined 2 as some particular set with (intuitively speaking) exactly two elements. How was that standard set selected? How should other such standard sets for other numbers be selected? There is no compelling mathematical reason for preferring one answer to this question to another; the whole thing is largely a matter of taste. The selection should presumably be guided by considerations of simplicity and economy. To motivate the particular selection that is usually made, suppose that a number, say 7, has already been defined as a set (with seven elements). How, in this case, should we define 8? Where, in other words, can we find a set consisting of exactly eight elements? We can find seven elements in the set 7; what shall we use as an eighth to adjoin to them? A reasonable answer to the last question is the number (set) 7 itself; the proposal is to define 8 to be the set consisting of the seven elements of 7, together with 7. Note that according to this proposal each number will be equal to the set of its own predecessors.

The preceding paragraph motivates a set-theoretic construction that makes sense for every set, but that is of interest in the construction of numbers only. For every set  $x$  we define the *successor*  $x^+$  of  $x$  to be the set obtained by adjoining  $x$  to the elements of  $x$ ; in other words,

$$x^+ = x \cup \{x\}.$$

(The successor of  $x$  is frequently denoted by  $x'$ .)

We are now ready to define the natural numbers. In defining 0 to be a set with zero elements, we have no choice; we must write (as we did)

$$0 = \emptyset.$$

If every natural number is to be equal to the set of its predecessors, we have no choice in defining 1, or 2, or 3 either; we must write

$$1 = 0^+ (= \{0\}),$$

$$2 = 1^+ (= \{0, 1\}),$$

$$3 = 2^+ (= \{0, 1, 2\}),$$

etc. The "etc." means that we hereby adopt the usual notation, and, in what follows, we shall feel free to use numerals such as "4" or "956" without any further explanation or apology.

From what has been said so far it does not follow that the construction of successors can be carried out ad infinitum within one and the same set. What we need is a new set-theoretic principle.

**Axiom of infinity.** *There exists a set containing 0 and containing the successor of each of its elements.*

The reason for the name of the axiom should be clear. We have not yet given a precise definition of infinity, but it seems reasonable that sets such as the ones that the axiom of infinity describes deserve to be called infinite.

We shall say, temporarily, that a set  $A$  is a *successor set* if  $0 \in A$  and if  $x^+ \in A$  whenever  $x \in A$ . In this language the axiom of infinity simply says that there exists a successor set  $A$ . Since the intersection of every (non-empty) family of successor sets is a successor set itself (proof?), the intersection of all the successor sets included in  $A$  is a successor set  $\omega$ . The set  $\omega$  is a subset of every successor set. If, indeed,  $B$  is an arbitrary successor set, then so is  $A \cap B$ . Since  $A \cap B \subset A$ , the set  $A \cap B$  is one of the sets that entered into the definition of  $\omega$ ; it follows that  $\omega \subset A \cap B$ , and, consequently, that  $\omega \subset B$ . The minimality property so established uniquely characterizes  $\omega$ ; the axiom of extension guarantees that there can be only one successor set that is included in every other successor set. A *natural number* is, by definition, an element of the minimal successor set  $\omega$ . This definition of natural numbers is the rigorous counterpart of the intuitive description according to which they consist of 0, 1, 2, 3, "and

so on." Incidentally, the symbol we are using for the set of all natural numbers ( $\omega$ ) has a plurality of the votes of the writers on the subject, but nothing like a clear majority. In this book that symbol will be used systematically and exclusively in the sense defined above.

The slight feeling of discomfort that the reader may experience in connection with the definition of natural numbers is quite common and in most cases temporary. The trouble is that here, as once before (in the definition of ordered pairs), the object defined has some irrelevant structure, which seems to get in the way (but is in fact harmless). We want to be told that the successor of 7 is 8, but to be told that 7 is a subset of 8 or that 7 is an element of 8 is disturbing. We shall make use of this superstructure of natural numbers just long enough to derive their most important natural properties; after that the superstructure may safely be forgotten.

A family  $\{x_i\}$  whose index set is either a natural number or else the set of all natural numbers is called a *sequence* (*finite* or *infinite*, respectively). If  $\{A_i\}$  is a sequence of sets, where the index set is the natural number  $n^+$ , then the union of the sequence is denoted by

$$\bigcup_{i=0}^n A_i \text{ or } A_0 \cup A_1 \cup A_2 \cup \dots$$

If the index set is  $\omega$ , the notation is

$$\bigcup_{i=0}^\infty A_i \text{ or } A_0 \cup A_1 \cup A_2 \cup \dots$$

Intersections and Cartesian products of sequences are denoted similarly by

$$\bigcap_{i=0}^n A_i, \quad A_0 \cap A_1 \cap A_2 \cap \dots$$

$$\bigtimes_{i=0}^n A_i, \quad A_0 \times A_1 \times A_2 \times \dots$$

and

$$\bigcap_{i=0}^\infty A_i, \quad A_0 \cap A_1 \cap A_2 \cap \dots$$

$$\bigtimes_{i=0}^\infty A_i, \quad A_0 \times A_1 \times A_2 \times \dots$$

The word "sequence" is used in a few different ways in the mathematical literature, but the differences among them are more notational than conceptual. The most common alternative starts at 1 instead of 0; in other words, it refers to a family whose index set is  $\omega - \{0\}$  instead of  $\omega$ .

## SECTION 12

---

### THE PEANO AXIOMS

---

---

We enter now into a minor digression. The purpose of the digression is to make fleeting contact with the arithmetic theory of natural numbers. From the set-theoretic point of view this is a pleasant luxury.

The most important thing we know about the set  $\omega$  of all natural numbers is that it is the unique successor set that is a subset of every successor set. To say that  $\omega$  is a successor set means that

$$(I) \quad 0 \in \omega$$

(where, of course,  $0 = \emptyset$ ), and that

$$(II) \quad \text{if } n \in \omega, \text{ then } n^+ \in \omega$$

(where  $n^+ = n \cup \{n\}$ ). The minimality property of  $\omega$  can be expressed by saying that if a subset  $S$  of  $\omega$  is a successor set, then  $S = \omega$ . Alternatively, and in more primitive terms,

$$(III) \quad \text{if } S \subset \omega, \text{ if } 0 \in S, \text{ and if } n^+ \in S \text{ whenever } n \in S, \text{ then } S = \omega.$$

Property (III) is known as the **principle of mathematical induction**. We shall now add to this list of properties of  $\omega$  two others:

$$(IV) \quad n^+ \neq 0 \text{ for all } n \text{ in } \omega,$$

and

$$(V) \quad \text{if } n \text{ and } m \text{ are in } \omega, \text{ and if } n^+ = m^+, \text{ then } n = m.$$

The proof of (IV) is trivial; since  $n^+$  always contains  $n$ , and since 0 is empty, it is clear that  $n^+$  is different from 0. The proof of (V) is not trivial; it depends on a couple of auxiliary propositions. The first one asserts that something that ought not to happen indeed does not happen. Even

if the considerations that the proof involves seem to be pathological and foreign to the arithmetic spirit that we expect to see in the theory of natural numbers, the end justifies the means. The second proposition refers to behavior that is quite similar to the one just excluded. This time, however, the apparently artificial considerations end in an affirmative result: something mildly surprising always does happen. The statements are as follows: (i) *no natural number is a subset of any of its elements*, and (ii) *every element of a natural number is a subset of it*. Sometimes a set with the property that it includes ( $\subset$ ) everything that it contains ( $\epsilon$ ) is called a *transitive* set. More precisely, to say that  $E$  is transitive means that if  $x \epsilon y$  and  $y \epsilon E$ , then  $x \epsilon E$ . (Recall the slightly different use of the word that we encountered in the theory of relations.) In this language, (ii) says that every natural number is transitive.

The proof of (i) is a typical application of the principle of mathematical induction. Let  $S$  be the set of all those natural numbers  $n$  that are not included in any of their elements. (Explicitly:  $n \epsilon S$  if and only if  $n \epsilon \omega$  and  $n$  is not a subset of  $x$  for any  $x$  in  $n$ .) Since 0 is not a subset of any of its elements, it follows that  $0 \epsilon S$ . Suppose now that  $n \epsilon S$ . Since  $n$  is a subset of  $n$ , we may infer that  $n$  is not an element of  $n$ , and hence that  $n^+$  is not a subset of  $n$ . What can  $n^+$  be a subset of? If  $n^+ \subset x$ , then  $n \subset x$ , and therefore (since  $n \epsilon S$ )  $x \epsilon n$ . It follows that  $n^+$  cannot be a subset of  $n$ , and  $n^+$  cannot be a subset of any element of  $n$ . This means that  $n^+$  cannot be a subset of any element of  $n^+$ , and hence that  $n^+ \epsilon S$ . The desired conclusion (i) is now a consequence of (III).

The proof of (ii) is also inductive. This time let  $S$  be the set of all transitive natural numbers. (Explicitly:  $n \epsilon S$  if and only if  $n \epsilon \omega$  and  $x$  is a subset of  $n$  for every  $x$  in  $n$ .) The requirement that  $0 \epsilon S$  is vacuously satisfied. Suppose now that  $n \epsilon S$ . If  $x \epsilon n^+$ , then either  $x \epsilon n$  or  $x = n$ . In the first case  $x \subset n$  (since  $n \epsilon S$ ) and therefore  $x \subset n^+$ ; in the second case  $x \subset n^+$  for even more trivial reasons. It follows that every element of  $n^+$  is a subset of  $n^+$ , or, in other words, that  $n^+ \epsilon S$ . The desired conclusion (ii) is a consequence of (III).

We are now ready to prove (V). Suppose indeed that  $n$  and  $m$  are natural numbers and that  $n^+ = m^+$ . Since  $n \epsilon n^+$ , it follows that  $n \epsilon m^+$ , and hence that either  $n \epsilon m$  or  $n = m$ . Similarly, either  $m \epsilon n$  or  $m = n$ . If  $n \neq m$ , then we must have  $n \epsilon m$  and  $m \epsilon n$ . Since, by (ii),  $n$  is transitive, it follows that  $n \epsilon n$ . Since, however,  $n \subset n$ , this contradicts (i), and the proof is complete.

The assertions (I)–(V) are known as the Peano axioms; they used to be considered as the fountainhead of all mathematical knowledge. From

them (together with the set-theoretic principles we have already met) it is possible to define integers, rational numbers, real numbers, and complex numbers, and to derive their usual arithmetic and analytic properties. Such a program is not within the scope of this book; the interested reader should have no difficulty in locating and studying it elsewhere.

Induction is often used not only to prove things but also to define things. Suppose, to be specific, that  $f$  is a function from a set  $X$  into the same set  $X$ , and suppose that  $a$  is an element of  $X$ . It seems natural to try to define an infinite sequence  $\{u(n)\}$  of elements of  $X$  (that is, a function  $u$  from  $\omega$  to  $X$ ) in some such way as this: write  $u(0) = a$ ,  $u(1) = f(u(0))$ ,  $u(2) = f(u(1))$ , and so on. If the would-be definer were pressed to explain the "and so on," he might lean on induction. What it all means, he might say, is that we define  $u(0)$  as  $a$ , and then, inductively, we define  $u(n^+)$  as  $f(u(n))$  for every  $n$ . This may sound plausible, but, as justification for an existential assertion, it is insufficient. The principle of mathematical induction does indeed prove, easily, that there can be at most one function satisfying all the stated conditions, but it does not establish the existence of such a function. What is needed is the following result.

**Recursion theorem.** *If  $a$  is an element of a set  $X$ , and if  $f$  is a function from  $X$  into  $X$ , then there exists a function  $u$  from  $\omega$  into  $X$  such that  $u(0) = a$  and such that  $u(n^+) = f(u(n))$  for all  $n$  in  $\omega$ .*

**PROOF.** Recall that a function from  $\omega$  to  $X$  is a certain kind of subset of  $\omega \times X$ ; we shall construct  $u$  explicitly as a set of ordered pairs. Consider, for this purpose, the collection  $\mathcal{C}$  of all those subsets  $A$  of  $\omega \times X$  for which  $(0, a) \in A$  and for which  $(n^+, f(x)) \in A$  whenever  $(n, x) \in A$ . Since  $\omega \times X$  has these properties, the collection  $\mathcal{C}$  is not empty. We may, therefore, form the intersection  $u$  of all the sets of the collection  $\mathcal{C}$ . Since it is easy to see that  $u$  itself belongs to  $\mathcal{C}$ , it remains only to prove that  $u$  is a function. We are to prove, in other words, that for each natural number  $n$  there exists at most one element  $x$  of  $X$  such that  $(n, x) \in u$ . (Explicitly: if both  $(n, x)$  and  $(n, y)$  belong to  $u$ , then  $x = y$ .) The proof is inductive. Let  $S$  be the set of all those natural numbers  $n$  for which it is indeed true that  $(n, x) \in u$  for at most one  $x$ . We shall prove that  $0 \in S$  and that if  $n \in S$ , then  $n^+ \in S$ .

Does  $0$  belong to  $S$ ? If not, then  $(0, b) \in u$  for some  $b$  distinct from  $a$ . Consider, in this case, the set  $u - \{(0, b)\}$ . Observe that this diminished set still contains  $(0, a)$  (since  $a \neq b$ ), and that if the diminished set contains  $(n, x)$ , then it contains  $(n^+, f(x))$  also. The reason for the second assertion is that since  $n^+ \neq 0$ , the discarded element is not equal to

$(n^+, f(x))$ . In other words,  $u - \{(0, b)\} \in \mathcal{C}$ . This contradicts the fact that  $u$  is the smallest set in  $\mathcal{C}$ , and we may conclude that  $0 \in S$ .

Suppose now that  $n \in S$ ; this means that there exists a unique element  $x$  in  $X$  such that  $(n, x) \in u$ . Since  $(n, x) \in u$ , it follows that  $(n^+, f(x)) \in u$ . If  $n^+$  does not belong to  $S$ , then  $(n^+, y) \in u$  for some  $y$  different from  $f(x)$ . Consider, in this case, the set  $u - \{(n^+, y)\}$ . Observe that this diminished set contains  $(0, a)$  (since  $n^+ \neq 0$ ), and that if the diminished set contains  $(m, t)$ , say, then it contains  $(m^+, f(t))$  also. Indeed, if  $m = n$ , then  $t$  must be  $x$ , and the reason the diminished set contains  $(n^+, f(x))$  is that  $f(x) \neq y$ ; if, on the other hand,  $m \neq n$ , then the reason the diminished set contains  $(m^+, f(t))$  is that  $m^+ \neq n^+$ . In other words,  $u - \{(n^+, y)\} \in \mathcal{C}$ . This again contradicts the fact that  $u$  is the smallest set in  $\mathcal{C}$ , and we may conclude that  $n^+ \in S$ .

The proof of the recursion theorem is complete. An application of the recursion theorem is called *definition by induction*.

**EXERCISE.** Prove that if  $n$  is a natural number, then  $n \neq n^+$ ; if  $n \neq 0$ , then  $n = m^+$  for some natural number  $m$ . Prove that  $\omega$  is transitive. Prove that if  $E$  is a non-empty subset of some natural number, then there exists an element  $k$  in  $E$  such that  $k \in m$  whenever  $m$  is an element of  $E$  distinct from  $k$ .

## SECTION 13

---

### ARITHMETIC

---

---

The introduction of addition for natural numbers is a typical example of definition by induction. Indeed, it follows from the recursion theorem that for each natural number  $m$  there exists a function  $s_m$  from  $\omega$  to  $\omega$  such that  $s_m(0) = m$  and such that  $s_m(n^+) = (s_m(n))^+$  for every natural number  $n$ ; the value  $s_m(n)$  is, by definition, the *sum*  $m + n$ . The general arithmetic properties of addition are proved by repeated applications of the principle of mathematical induction. Thus, for instance, addition is associative. This means that

$$(k + m) + n = k + (m + n)$$

whenever  $k$ ,  $m$ , and  $n$  are natural numbers. The proof goes by induction on  $n$  as follows. Since  $(k + m) + 0 = k + m$  and  $k + (m + 0) = k + m$ , the equation is true if  $n = 0$ . If the equation is true for  $n$ , then  $(k + m) + n^+ = ((k + m) + n)^+$  (by definition) =  $(k + (m + n))^+$  (by the induction hypothesis) =  $k + (m + n)^+$  (again by the definition of addition) =  $k + (m + n^+)$  (ditto), and the argument is complete. The proof that addition is commutative (i.e.,  $m + n = n + m$  for all  $m$  and  $n$ ) is a little tricky; a straightforward attack might fail. The trick is to prove, by induction on  $n$ , that (i)  $0 + n = n$  and (ii)  $m^+ + n = (m + n)^+$ , and then to prove the desired commutativity equation by induction on  $m$ , via (i) and (ii).

Similar techniques are applied in the definitions of products and exponents and in the derivations of their basic arithmetic properties. To define multiplication, apply the recursion theorem to produce functions  $p_m$  such that  $p_m(0) = 0$  and such that  $p_m(n^+) = p_m(n) + m$  for every natural number  $n$ ; then the value  $p_m(n)$  is, by definition, the *product*  $m \cdot n$ . (The dot is frequently omitted.) Multiplication is associative and commutative; the

proofs are straightforward adaptations of the ones that worked for addition. The distributive law (i.e., the assertion that  $k \cdot (m + n) = k \cdot m + k \cdot n$  whenever  $k$ ,  $m$ , and  $n$  are natural numbers) is another easy consequence of the principle of mathematical induction. (Use induction on  $n$ .) Anyone who has worked through sums and products in this way should have no trouble with exponents. The recursion theorem yields functions  $e_m$  such that  $e_m(0) = 1$  and such that  $e_m(n^+) = e_m(n) \cdot m$  for every natural number  $n$ ; the value  $e_m(n)$  is, by definition, the *power*  $m^n$ . The discovery and establishment of the properties of powers, as well as the detailed proofs of the statements about products, can safely be left as exercises for the reader.

The next topic that deserves some attention is the theory of order in the set of natural numbers. For this purpose we proceed to examine with some care the question of which natural numbers belong to which others. Formally, we say that two natural numbers  $m$  and  $n$  are comparable if  $m \in n$ , or  $m = n$ , or  $n \in m$ . Assertion: two natural numbers are always comparable. The proof of this assertion consists of several steps; it will be convenient to introduce some notation. For each  $n$  in  $\omega$ , write  $S(n)$  for the set of all  $m$  in  $\omega$  that are comparable with  $n$ , and let  $S$  be the set of all those  $n$  for which  $S(n) = \omega$ . In these terms, the assertion is that  $S = \omega$ . We begin the proof by showing that  $S(0) = \omega$  (i.e., that  $0 \in S$ ). Clearly  $S(0)$  contains 0. If  $m \in S(0)$ , then, since  $m \in 0$  is impossible, either  $m = 0$  (in which case  $0 \in m^+$ ), or  $0 \in m$  (in which case, again,  $0 \in m^+$ ). Hence, in all cases, if  $m \in S(0)$ , then  $m^+ \in S(0)$ ; this proves that  $S(0) = \omega$ . We complete the proof by showing that if  $S(n) = \omega$ , then  $S(n^+) = \omega$ . The fact that  $0 \in S(n^+)$  is immediate (since  $n^+ \in S(0)$ ); it remains to prove that if  $m \in S(n^+)$ , then  $m^+ \in S(n^+)$ . Since  $m \in S(n^+)$ , therefore either  $n^+ \in m$  (in which case  $n^+ \in m^+$ ), or  $n^+ = m$  (ditto), or  $m \in n^+$ . In the latter case, either  $m = n$  (in which case  $m^+ = n^+$ ), or  $m \in n$ . The last case, in turn, splits according to the behavior of  $m^+$  and  $n$ : since  $m^+ \in S(n)$ , we must have either  $n \in m^+$ , or  $n = m^+$ , or  $m^+ \in n$ . The first possibility is incompatible with the present situation (i.e., with  $m \in n$ ). The reason is that if  $n \in m^+$ , then either  $n \in m$  or  $n = m$ , so that, in any case,  $n \subset m$ , and we know that no natural number is a subset of one of its elements. Both the remaining possibilities imply that  $m^+ \in n^+$ , and the proof is complete.

The preceding paragraph implies that if  $m$  and  $n$  are in  $\omega$ , then at least one of the three possibilities ( $m \in n$ ,  $m = n$ ,  $n \in m$ ) must hold; it is easy to see that, in fact, always exactly one of them holds. (The reason is another application of the fact that a natural number is not a subset of one of its elements.) Another consequence of the preceding paragraph is that if  $n$  and  $m$  are distinct natural numbers, then a necessary and sufficient condi-

tion that  $m \in n$  is that  $m \subset n$ . Indeed, the implication from  $m \in n$  to  $m \subset n$  is just the transitivity of  $n$ . If, conversely,  $m \subset n$  and  $m \neq n$ , then  $n \in m$  cannot happen (for then  $m$  would be a subset of one of its elements), and therefore  $m \in n$ . If  $m \in n$ , or if, equivalently,  $m$  is a proper subset of  $n$ , we shall write  $m < n$  and we shall say that  $m$  is *less than*  $n$ . If  $m$  is known to be either less than  $n$  or else equal to  $n$ , we write  $m \leq n$ . Note that  $\leq$  and  $<$  are relations in  $\omega$ . The former is reflexive, but the latter is not; neither is symmetric; both are transitive. If  $m \leq n$  and  $n \leq m$ , then  $m = n$ .

**EXERCISE.** Prove that if  $m < n$ , then  $m + k < n + k$ , and prove that if  $m < n$  and  $k \neq 0$ , then  $m \cdot k < n \cdot k$ . Prove that if  $E$  is a non-empty set of natural numbers, then there exists an element  $k$  in  $E$  such that  $k \leq m$  for all  $m$  in  $E$ .

Two sets  $E$  and  $F$  (not necessarily subsets of  $\omega$ ) are called *equivalent*, in symbols  $E \sim F$ , if there exists a one-to-one correspondence between them. It is easy to verify that equivalence in this sense, for subsets of some particular set  $X$ , is an equivalence relation in the power set  $\mathcal{P}(X)$ .

Every proper subset of a natural number  $n$  is equivalent to some smaller natural number (i.e., to some element of  $n$ ). The proof of this assertion is inductive. For  $n = 0$  it is trivial. If it is true for  $n$ , and if  $E$  is a proper subset of  $n^+$ , then either  $E$  is a proper subset of  $n$  and the induction hypothesis applies, or  $E = n$  and the result is trivial, or  $n \in E$ . In the latter case, find a number  $k$  in  $n$  but not in  $E$  and define a function  $f$  on  $E$  by writing  $f(i) = i$  when  $i \neq n$  and  $f(n) = k$ . Clearly  $f$  is one-to-one and  $f$  maps  $E$  into  $n$ . It follows that the image of  $E$  under  $f$  is either equal to  $n$  or (by the induction hypothesis) equivalent to some element of  $n$ , and, consequently,  $E$  itself is always equivalent to some element of  $n^+$ .

It is a mildly shocking fact that a set can be equivalent to a proper subset of itself. If, for instance, a function  $f$  from  $\omega$  to  $\omega$  is defined by writing  $f(n) = n^+$  for all  $n$  in  $\omega$ , then  $f$  is a one-to-one correspondence between the set of all natural numbers and the proper subset consisting of the non-zero natural numbers. It is nice to know that even though the set of all natural numbers has this peculiar property, sanity prevails for each particular natural number. In other words, if  $n \in \omega$ , then  $n$  is not equivalent to a proper subset of  $n$ . For  $n = 0$  this is clear. Suppose now that it is true for  $n$ , and suppose that  $f$  is a one-to-one correspondence from  $n^+$  to a proper subset  $E$  of  $n^+$ . If  $n \notin E$ , then the restriction of  $f$  to  $n$  is a one-to-one correspondence between  $n$  and a proper subset of  $n$ , which contradicts the induction hypothesis. If  $n \in E$ , then  $n$  is equivalent to  $E - \{n\}$ , so that, by the in-

duction hypothesis,  $n = E - \{n\}$ . This implies that  $E = n^+$ , which contradicts the assumption that  $E$  is a proper subset of  $n^+$ .

A set  $E$  is called *finite* if it is equivalent to some natural number; otherwise  $E$  is *infinite*.

**EXERCISE.** Use this definition to prove that  $\omega$  is infinite.

A set can be equivalent to at most one natural number. (Proof: we know that for any two distinct natural numbers one must be an element and therefore a proper subset of the other; it follows from the preceding paragraph that they cannot be equivalent.) We may infer that a finite set is never equivalent to a proper subset; in other words, as long as we stick to finite sets, the whole is always greater than any of its parts.

**EXERCISE.** Use this consequence of the definition of finiteness to prove that  $\omega$  is infinite.

Since every subset of a natural number is equivalent to a natural number, it follows also that every subset of a finite set is finite.

The *number of elements* in a finite set  $E$  is, by definition, the unique natural number equivalent to  $E$ ; we shall denote it by  $\#(E)$ . It is clear that if the correspondence between  $E$  and  $\#(E)$  is restricted to the finite subsets of some set  $X$ , the result is a function from a subset of the power set  $\wp(X)$  to  $\omega$ . This function is pleasantly related to the familiar set-theoretic relations and operations. Thus, for example, if  $E$  and  $F$  are finite sets such that  $E \subset F$ , then  $\#(E) \leq \#(F)$ . (The reason is that since  $E \sim \#(E)$  and  $F \sim \#(F)$ , it follows that  $\#(E)$  is equivalent to a subset of  $\#(F)$ .) Another example is the assertion that if  $E$  and  $F$  are finite sets, then  $E \cup F$  is finite, and, moreover, if  $E$  and  $F$  are disjoint, then  $\#(E \cup F) = \#(E) + \#(F)$ . The crucial step in the proof is the fact that if  $m$  and  $n$  are natural numbers, then the complement of  $m$  in the sum  $m + n$  is equivalent to  $n$ ; the proof of this auxiliary fact is achieved by induction on  $n$ . Similar techniques prove that if  $E$  and  $F$  are finite sets, then so also are  $E \times F$  and  $E^F$ , and, moreover,  $\#(E \times F) = \#(E) \cdot \#(F)$  and  $\#(E^F) = \#(E)^{\#(F)}$ .

**EXERCISE.** The union of a finite set of finite sets is finite. If  $E$  is finite, then  $\wp(E)$  is finite and, moreover,  $\#(\wp(E)) = 2^{\#(E)}$ . If  $E$  is a non-empty finite set of natural numbers, then there exists an element  $k$  in  $E$  such that  $m \leq k$  for all  $m$  in  $E$ .

## SECTION 14

---

### ORDER

---

Throughout mathematics, and, in particular, for the generalization to infinite sets of the counting process appropriate to finite sets, the theory of order plays an important role. The basic definitions are simple. The only thing to remember is that the primary motivation comes from the familiar properties of “less than or equal to” and not “less than.” There is no profound reason for this; it just happens that the generalization of “less than or equal to” occurs more frequently and is more amenable to algebraic treatment.

A relation  $R$  in a set  $X$  is called *antisymmetric* if, for every  $x$  and  $y$  in  $X$ , the simultaneous validity of  $x R y$  and  $y R x$  implies that  $x = y$ . A *partial order* (or sometimes simply an *order*) in a set  $X$  is a reflexive, antisymmetric, and transitive relation in  $X$ . It is customary to use only one symbol (or some typographically close relative of it) for most partial orders in most sets; the symbol in common use is the familiar inequality sign. Thus a partial order in  $X$  may be defined as a relation  $\leq$  in  $X$  such that, for all  $x$ ,  $y$ , and  $z$  in  $X$ , we have (i)  $x \leq x$ , (ii) if  $x \leq y$  and  $y \leq x$ , then  $x = y$ , and (iii) if  $x \leq y$  and  $y \leq z$ , then  $x \leq z$ . The reason for the qualifying “partial” is that some questions about order may be left unanswered. If for every  $x$  and  $y$  in  $X$  either  $x \leq y$  or  $y \leq x$ , then  $\leq$  is called a *total* (sometimes also *simple* or *linear*) order. A totally ordered set is frequently called a *chain*.

**EXERCISE.** Express the conditions of antisymmetry and totality for a relation  $R$  by means of equations involving  $R$  and its inverse.

The most natural example of a partial (and not total) order is inclusion. Explicitly: for each set  $X$ , the relation  $\subset$  is a partial order in the power set  $\wp(X)$ ; it is a total order if and only if  $X$  is empty or  $X$  is a singleton. A

well known example of a total order is the relation “less than or equal to” in the set of natural numbers. An interesting and frequently seen partial order is the relation of extension for functions. Explicitly: for given sets  $X$  and  $Y$ , let  $F$  be the set of all those functions whose domain is included in  $X$  and whose range is included in  $Y$ . Define a relation  $R$  in  $F$  by writing  $f R g$  in case  $\text{dom } f \subset \text{dom } g$  and  $f(x) = g(x)$  for all  $x$  in  $\text{dom } f$ ; in other words,  $f R g$  means that  $f$  is a restriction of  $g$ , or, equivalently, that  $g$  is an extension of  $f$ . If we recall that the functions in  $F$  are, after all, certain subsets of the Cartesian product  $X \times Y$ , we recognize that  $f R g$  means the same as  $f \subset g$ ; extension is a special case of inclusion.

A *partially ordered set* is a set together with a partial order in it. A precise formulation of this “togetherness” goes as follows: a partially ordered set is an ordered pair  $(X, \leq)$ , where  $X$  is a set and  $\leq$  is a partial order in  $X$ . This kind of definition is very common in mathematics; a mathematical structure is almost always a set “together” with some specified other sets, functions, and relations. The accepted way of making such definitions precise is by reference to ordered pairs, triples, or whatever is appropriate. That is not the only way. Observe, for instance, that knowledge of a partial order implies knowledge of its domain. If, therefore, we describe a partially ordered set as an ordered pair, we are being quite redundant; the second coordinate alone would have conveyed the same amount of information. In matters of language and notation, however, tradition always conquers pure reason. The accepted mathematical behavior (for structures in general, illustrated here for partially ordered sets) is to admit that ordered pairs are the right approach, to forget that the second coordinate is the important one, and to speak as if the first coordinate were all that mattered. Following custom, we shall often say something like “let  $X$  be a partially ordered set,” when what we really mean is “let  $X$  be the domain of a partial order.” The same linguistic conventions apply to totally ordered sets, i.e., to partially ordered sets whose order is in fact total.

The theory of partially ordered sets uses many words whose technical meaning is so near to their everyday connotation that they are almost self-explanatory. Suppose, to be specific, that  $X$  is a partially ordered set and that  $x$  and  $y$  are elements of  $X$ . We write  $y \geq x$  in case  $x \leq y$ ; in other words,  $\geq$  is the inverse of the relation  $\leq$ . If  $x \leq y$  and  $x \neq y$ , we write  $x < y$  and we say that  $x$  is *less than* or *smaller than*  $y$ , or that  $x$  is a *predecessor* of  $y$ . Alternatively, under the same circumstances, we write  $y > x$  and we say that  $y$  is *greater* or *larger* than  $x$ , or  $y$  is a *successor* of  $x$ . The relation  $<$  is such that (i) for no elements  $x$  and  $y$  do  $x < y$  and  $y < x$  hold simul-

taneously, and (ii) if  $x < y$  and  $y < z$ , then  $x < z$  (i.e.,  $<$  is transitive). If, conversely,  $<$  is a relation in  $X$  satisfying (i) and (ii), and if  $x \leq y$  is defined to mean that either  $x < y$  or  $x = y$ , then  $\leq$  is a partial order in  $X$ .

The connection between  $\leq$  and  $<$  can be generalized to arbitrary relations. That is, given any relation  $R$  in a set  $X$ , we can define a relation  $S$  in  $X$  by writing  $x S y$  in case  $x R y$  but  $x \neq y$ , and, vice versa, given any relation  $S$  in  $X$ , we can define a relation  $R$  in  $X$  by writing  $x R y$  in case either  $x S y$  or  $x = y$ . To have an abbreviated way of referring to the passage from  $R$  to  $S$  and back, we shall say that  $S$  is the *strict* relation corresponding to  $R$ , and  $R$  is the *weak* relation corresponding to  $S$ . We shall say of a relation in a set  $X$  that it "partially orders  $X$ " in case either it is a partial order in  $X$  or else the corresponding weak relation is one.

If  $X$  is a partially ordered set, and if  $a \in X$ , the set  $\{x \in X : x < a\}$  is the *initial segment* determined by  $a$ ; we shall usually denote it by  $s(a)$ . The set  $\{x \in X : x \leq a\}$  is the *weak initial segment* determined by  $a$ , and will be denoted by  $\bar{s}(a)$ . When it is important to emphasize the distinction between initial segments and weak initial segments, the former will be called *strict initial segments*. In general the words "strict" and "weak" refer to  $<$  and  $\leq$  respectively. Thus, for instance, the initial segment determined by  $a$  may be described as the set of all predecessors of  $a$ , or, for emphasis, as the set of all *strict predecessors* of  $a$ ; similarly the weak initial segment determined by  $a$  consists of all *weak predecessors* of  $a$ . If  $x \leq y$  and  $y \leq z$ , we may say that  $y$  is *between*  $x$  and  $z$ ; if  $x < y$  and  $y < z$ , then  $y$  is *strictly between*  $x$  and  $z$ . If  $x < y$  and if there is no element strictly between  $x$  and  $y$ , we say that  $x$  is an *immediate predecessor* of  $y$ , or  $y$  is an *immediate successor* of  $x$ .

If  $X$  is a partially ordered set (which may in particular be totally ordered), then it could happen that  $X$  has an element  $a$  such that  $a \leq x$  for every  $x$  in  $X$ . In that case we say that  $a$  is the *least* (*smallest, first*) element of  $X$ . The antisymmetry of an order implies that if  $X$  has a least element, then it has only one. If, similarly,  $X$  has an element  $a$  such that  $x \leq a$  for every  $x$  in  $X$ , then  $a$  is the *greatest* (*largest, last*) element of  $X$ ; it too is unique (if it exists at all). The set  $\omega$  of all natural numbers (with its customary ordering by magnitude) is an example of a partially ordered set with a first element (namely 0) but no last. The same set, but this time with the inverse ordering, has a last element but no first.

In partially ordered sets there is an important distinction between least elements and minimal ones. If, as before,  $X$  is a partially ordered set, an element  $a$  of  $X$  is called a *minimal* element of  $X$  in case there is no element in  $X$  strictly smaller than  $a$ . Equivalently,  $a$  is minimal if  $x \leq a$  implies

that  $x = a$ . For an example, consider the collection  $\mathcal{C}$  of non-empty subsets of a non-empty set  $X$ , with ordering by inclusion. Each singleton is a minimal element of  $\mathcal{C}$ , but clearly  $\mathcal{C}$  has no least element (unless  $X$  itself is a singleton). We distinguish similarly between greatest and maximal elements; a *maximal* element of  $X$  is an element  $a$  such that  $X$  contains nothing strictly greater than  $a$ . Equivalently,  $a$  is maximal if  $a \leq x$  implies that  $x = a$ .

An element  $a$  of a partially ordered set is said to be a *lower bound* of a subset  $E$  of  $X$  in case  $a \leq x$  for every  $x$  in  $E$ ; similarly  $a$  is an *upper bound* of  $E$  in case  $x \leq a$  for every  $x$  in  $E$ . A set  $E$  may have no lower bounds or upper bounds at all, or it may have many; in the latter case it could happen that none of them belongs to  $E$ . (Examples?) Let  $E_*$  be the set of all lower bounds of  $E$  in  $X$  and let  $E^*$  be the set of all upper bounds of  $E$  in  $X$ . What was just said is that  $E_*$  may be empty, or  $E_* \cap E$  may be empty. If  $E_* \cap E$  is not empty, then it is a singleton consisting of the unique least element of  $E$ . Similar remarks apply, of course, to  $E^*$ . If it happens that the set  $E_*$  contains a greatest element  $a$  (necessarily unique), then  $a$  is called the *greatest lower bound* or *infimum* of  $E$ . The abbreviations *g.l.b.* and *inf* are in common use. Because of the difficulties in pronouncing the former, and even in remembering whether g.l.b. is up (greatest) or down (lower), we shall use the latter notation only. Thus  $\inf E$  is the unique element in  $X$  (possibly not in  $E$ ) that is a lower bound of  $E$  and that dominates (i.e., is greater than) every other lower bound of  $E$ . The definitions at the other end are completely parallel. If  $E^*$  has a least element  $a$  (necessarily unique), then  $a$  is called the *least upper bound* (*l.u.b.*) or *supremum* (*sup*) of  $E$ .

The ideas connected with partially ordered sets are easy to express but they take some time to assimilate. The reader is advised to manufacture many examples to illustrate the various possibilities in the behavior of partially ordered sets and their subsets. To aid him in this enterprise, we proceed to describe three special partially ordered sets with some amusing properties. (i) The set is  $\omega \times \omega$ . To avoid any possible confusion, we shall denote the order we are about to introduce by the neutral symbol  $R$ . If  $(a, b)$  and  $(x, y)$  are ordered pairs of natural numbers, then  $(a, b) R (x, y)$  means, by definition, that  $(2a + 1) \cdot 2^y \leq (2x + 1) \cdot 2^b$ . (Here the inequality sign refers to the customary ordering of natural numbers.) The reader who is not willing to pretend ignorance of fractions will recognize that,

except for notation, what we just defined is the usual order for  $\frac{2a+1}{2^b}$  and  $\frac{2x+1}{2^y}$ . (ii) The set is  $\omega \times \omega$  again. Once more we use a neutral symbol

for the order; say  $S$ . If  $(a, b)$  and  $(x, y)$  are ordered pairs of natural numbers, then  $(a, b) S (x, y)$  means, by definition, that either  $a$  is strictly less than  $x$  (in the customary sense), or else  $a = x$  and  $b \leq y$ . Because of its resemblance to the way words are arranged in a dictionary, this is called the *lexicographical* order of  $\omega \times \omega$ . (iii) Once more the set is  $\omega \times \omega$ . The present order relation, say  $T$ , is such that  $(a, b) T (x, y)$  means, by definition, that  $a \leq x$  and  $b \leq y$ .

## SECTION 15

---

### THE AXIOM OF CHOICE

---

For the deepest results about partially ordered sets we need a new set-theoretic tool; we interrupt the development of the theory of order long enough to pick up that tool.

We begin by observing that a set is either empty or it is not, and, if it is not, then, by the definition of the empty set, there is an element in it. This remark can be generalized. If  $X$  and  $Y$  are sets, and if one of them is empty, then the Cartesian product  $X \times Y$  is empty. If neither  $X$  nor  $Y$  is empty, then there is an element  $x$  in  $X$ , and there is an element  $y$  in  $Y$ ; it follows that the ordered pair  $(x, y)$  belongs to the Cartesian product  $X \times Y$ , so that  $X \times Y$  is not empty. The preceding remarks constitute the cases  $n = 1$  and  $n = 2$  of the following assertion: if  $\{X_i\}$  is a finite sequence of sets, for  $i$  in  $n$ , say, then a necessary and sufficient condition that their Cartesian product be empty is that at least one of them be empty. The assertion is easy to prove by induction on  $n$ . (The case  $n = 0$  leads to a slippery argument about the empty function; the uninterested reader may start his induction at 1 instead of 0.)

The generalization to infinite families of the non-trivial part of the assertion in the preceding paragraph (necessity) is the following important principle of set theory.

**Axiom of choice.** *The Cartesian product of a non-empty family of non-empty sets is non-empty.*

In other words: if  $\{X_i\}$  is a family of non-empty sets indexed by a non-empty set  $I$ , then there exists a family  $\{x_i\}$ ,  $i \in I$ , such that  $x_i \in X_i$  for each  $i$  in  $I$ .

Suppose that  $\mathcal{C}$  is a non-empty collection of non-empty sets. We may regard  $\mathcal{C}$  as a family, or, to say it better, we can convert  $\mathcal{C}$  into an indexed set, just by using the collection  $\mathcal{C}$  itself in the role of the index set and using the identity mapping on  $\mathcal{C}$  in the role of the indexing. The axiom

of choice then says that the Cartesian product of the sets of  $\mathcal{C}$  has at least one element. An element of such a Cartesian product is, by definition, a function (family, indexed set) whose domain is the index set (in this case  $\mathcal{C}$ ) and whose value at each index belongs to the set bearing that index. Conclusion: there exists a function  $f$  with domain  $\mathcal{C}$  such that if  $A \in \mathcal{C}$ , then  $f(A) \in A$ . This conclusion applies, in particular, in case  $\mathcal{C}$  is the collection of all non-empty subsets of a non-empty set  $X$ . The assertion in that case is that there exists a function  $f$  with domain  $\wp(X) - \{\emptyset\}$  such that if  $A$  is in that domain, then  $f(A) \in A$ . In intuitive language the function  $f$  can be described as a simultaneous choice of an element from each of many sets; this is the reason for the name of the axiom. (A function that in this sense “chooses” an element out of each non-empty subset of a set  $X$  is called a *choice function* for  $X$ .) We have seen that if the collection of sets we are choosing from is finite, then the possibility of simultaneous choice is an easy consequence of what we knew before the axiom of choice was even stated; the role of the axiom is to guarantee that possibility in infinite cases.

The two consequences of the axiom of choice in the preceding paragraph (one for the power set of a set and the other for more general collections of sets) are in fact just reformulations of that axiom. It used to be considered important to examine, for each consequence of the axiom of choice, the extent to which the axiom is needed in the proof of the consequence. An alternative proof without the axiom of choice spelled victory; a converse proof, showing that the consequence is equivalent to the axiom of choice (in the presence of the remaining axioms of set theory) meant honorable defeat. Anything in between was considered exasperating. As a sample (and an exercise) we mention the assertion that every relation includes a function with the same domain. Another sample: if  $\mathcal{C}$  is a collection of pairwise disjoint non-empty sets, then there exists a set  $A$  such that  $A \cap C$  is a singleton for each  $C$  in  $\mathcal{C}$ . Both these assertions are among the many known to be equivalent to the axiom of choice.

As an illustration of the use of the axiom of choice, consider the assertion that if a set is infinite, then it has a subset equivalent to  $\omega$ . An informal argument might run as follows. If  $X$  is infinite, then, in particular, it is not empty (that is, it is not equivalent to 0); hence it has an element, say  $x_0$ . Since  $X$  is not equivalent to 1, the set  $X - \{x_0\}$  is not empty; hence it has an element, say  $x_1$ . Repeat this argument ad infinitum; the next step, for instance, is to say that  $X - \{x_0, x_1\}$  is not empty, and, therefore, it has an element, say  $x_2$ . The result is an infinite sequence  $\{x_n\}$  of distinct elements of  $X$ ; q.e.d. This sketch of a proof at least has the virtue of being

honest about the most important idea behind it; the act of choosing an element from a non-empty set was repeated infinitely often. The mathematician experienced in the ways of the axiom of choice will often offer such an informal argument; his experience enables him to see at a glance how to make it precise. For our purposes it is advisable to take a longer look.

Let  $f$  be a choice function for  $X$ ; that is,  $f$  is a function from the collection of all non-empty subsets of  $X$  to  $X$  such that  $f(A) \in A$  for all  $A$  in the domain of  $f$ . Let  $\mathcal{C}$  be the collection of all finite subsets of  $X$ . Since  $X$  is infinite, it follows that if  $A \in \mathcal{C}$ , then  $X - A$  is not empty, and hence that  $X - A$  belongs to the domain of  $f$ . Define a function  $g$  from  $\mathcal{C}$  to  $\mathcal{C}$  by writing  $g(A) = A \cup \{f(X - A)\}$ . In words:  $g(A)$  is obtained by adjoining to  $A$  the element that  $f$  chooses from  $X - A$ . We apply the recursion theorem to the function  $g$ ; we may start it rolling with, for instance, the set  $\emptyset$ . The result is that there exists a function  $U$  from  $\omega$  into  $\mathcal{C}$  such that  $U(0) = \emptyset$  and  $U(n^+) = U(n) \cup \{f(X - U(n))\}$  for every natural number  $n$ . Assertion: if  $v(n) = f(X - U(n))$ , then  $v$  is a one-to-one correspondence from  $\omega$  to  $X$ , and hence, indeed,  $\omega$  is equivalent to some subset of  $X$  (namely the range of  $v$ ). To prove the assertion, we make a series of elementary observations; their proofs are easy consequences of the definitions. First:  $v(n) \in U(n)$  for all  $n$ . Second:  $v(n) \in U(n^+)$  for all  $n$ . Third: if  $n$  and  $m$  are natural numbers and  $n \leq m$ , then  $U(n) \subset U(m)$ . Fourth: if  $n$  and  $m$  are natural numbers and  $n < m$ , then  $v(n) \neq v(m)$ . (Reason:  $v(n) \in U(m)$  but  $v(m) \notin U(m)$ .) The last observation implies that  $v$  maps distinct natural numbers onto distinct elements of  $X$ ; all we have to remember is that of any two distinct natural numbers one of them is strictly smaller than the other.

The proof is complete; we know now that every infinite set has a subset equivalent to  $\omega$ . This result, proved here not so much for its intrinsic interest as for an example of the proper use of the axiom of choice, has an interesting corollary. The assertion is that a set is infinite if and only if it is equivalent to a proper subset of itself. The “if” we already know; it says merely that a finite set cannot be equivalent to a proper subset. To prove the “only if,” suppose that  $X$  is infinite, and let  $v$  be a one-to-one correspondence from  $\omega$  into  $X$ . If  $x$  is in the range of  $v$ , say  $x = v(n)$ , write  $h(x) = v(n^+)$ ; if  $x$  is not in the range of  $v$ , write  $h(x) = x$ . It is easy to verify that  $h$  is a one-to-one correspondence from  $X$  into itself. Since the range of  $h$  is a proper subset of  $X$  (it does not contain  $v(0)$ ), the proof of the corollary is complete. The assertion of the corollary was used by Dedekind as the very definition of infinity.

## SECTION 16

---

### ZORN'S LEMMA

---

An existence theorem asserts the existence of an object belonging to a certain set and possessing certain properties. Many existence theorems can be formulated (or, if need be, reformulated) so that the underlying set is a partially ordered set and the crucial property is maximality. Our next purpose is to state and prove the most important theorem of this kind.

**Zorn's lemma.** *If  $X$  is a partially ordered set such that every chain in  $X$  has an upper bound, then  $X$  contains a maximal element.*

DISCUSSION. Recall that a chain is a totally ordered set. By a chain “in  $X$ ” we mean a subset of  $X$  such that the subset, considered as a partially ordered set on its own right, turns out to be totally ordered. If  $A$  is a chain in  $X$ , the hypothesis of Zorn's lemma guarantees the existence of an upper bound for  $A$  in  $X$ ; it does not guarantee the existence of an upper bound for  $A$  in  $A$ . The conclusion of Zorn's lemma is the existence of an element  $a$  in  $X$  with the property that if  $a \leq x$ , then necessarily  $a = x$ .

The basic idea of the proof is similar to the one used in our preceding discussion of infinite sets. Since, by hypothesis,  $X$  is not empty, it has an element, say  $x_0$ . If  $x_0$  is maximal, stop here. If it is not, then there exists an element, say  $x_1$ , strictly greater than  $x_0$ . If  $x_1$  is maximal, stop here; otherwise continue. Repeat this argument ad infinitum; ultimately it must lead to a maximal element.

The last sentence is probably the least convincing part of the argument; it hides a multitude of difficulties. Observe, for instance, the following possibility. It could happen that the argument, repeated ad infinitum, leads to a whole infinite sequence of non-maximal elements; what are we to do in that case? The answer is that the range of such an infinite sequence is a chain in  $X$ , and, consequently, has an upper bound; the thing to do is to start the whole argument all over again, beginning with that

upper bound. Just exactly when and how all this comes to an end is obscure, to say the least. There is no help for it; we must look at the precise proof. The structure of the proof is an adaptation of one originally given by Zermelo.

**PROOF.** The first step is to replace the abstract partial ordering by the inclusion order in a suitable collection of sets. More precisely, we consider, for each element  $x$  in  $X$ , the weak initial segment  $\bar{s}(x)$  consisting of  $x$  and all its predecessors. The range  $\bar{s}$  of the function  $\bar{s}$  (from  $X$  to  $\mathcal{P}(X)$ ) is a certain collection of subsets of  $X$ , which we may, of course, regard as (partially) ordered by inclusion. The function  $\bar{s}$  is one-to-one, and a necessary and sufficient condition that  $\bar{s}(x) \subset \bar{s}(y)$  is that  $x \leq y$ . In view of this, the task of finding a maximal element in  $X$  is the same as the task of finding a maximal set in  $\bar{s}$ . The hypothesis about chains in  $X$  implies (and is, in fact, equivalent to) the corresponding statement about chains in  $\bar{s}$ .

Let  $\mathfrak{X}$  be the set of all chains in  $X$ ; every member of  $\mathfrak{X}$  is included in  $\bar{s}(x)$  for some  $x$  in  $X$ . The collection  $\mathfrak{X}$  is a non-empty collection of sets, partially ordered by inclusion, and such that if  $C$  is a chain in  $\mathfrak{X}$ , then the union of the sets in  $C$  (i.e.,  $\bigcup_{A \in C} A$ ) belongs to  $\mathfrak{X}$ . Since each set in  $\mathfrak{X}$  is dominated by some set in  $\bar{s}$ , the passage from  $\bar{s}$  to  $\mathfrak{X}$  cannot introduce any new maximal elements. One advantage of the collection  $\mathfrak{X}$  is the slightly more specific form that the chain hypothesis assumes; instead of saying that each chain  $C$  has some upper bound in  $\bar{s}$ , we can say explicitly that the union of the sets of  $C$ , which is clearly an upper bound of  $C$ , is an element of the collection  $\mathfrak{X}$ . Another technical advantage of  $\mathfrak{X}$  is that it contains all the subsets of each of its sets; this makes it possible to enlarge non-maximal sets in  $\mathfrak{X}$  slowly, one element at a time.

Now we can forget about the given partial order in  $X$ . In what follows we consider a non-empty collection  $\mathfrak{X}$  of subsets of a non-empty set  $X$ , subject to two conditions: every subset of each set in  $\mathfrak{X}$  is in  $\mathfrak{X}$ , and the union of each chain of sets in  $\mathfrak{X}$  is in  $\mathfrak{X}$ . Note that the first condition implies that  $\emptyset \in \mathfrak{X}$ . Our task is to prove that there exists in  $\mathfrak{X}$  a maximal set.

Let  $f$  be a choice function for  $X$ , that is,  $f$  is a function from the collection of all non-empty subsets of  $X$  to  $X$  such that  $f(A) \in A$  for all  $A$  in the domain of  $f$ . For each set  $A$  in  $\mathfrak{X}$ , let  $\hat{A}$  be the set of all those elements  $x$  of  $X$  whose adjunction to  $A$  produces a set in  $\mathfrak{X}$ ; in other words,  $\hat{A} = \{x \in X : A \cup \{x\} \in \mathfrak{X}\}$ . Define a function  $g$  from  $\mathfrak{X}$  to  $\mathfrak{X}$  as follows: if  $\hat{A} - A \neq \emptyset$ , then  $g(A) = A \cup \{f(\hat{A} - A)\}$ ; if  $\hat{A} - A = \emptyset$ , then  $g(A) = A$ . It follows from the definition of  $\hat{A}$  that  $\hat{A} - A = \emptyset$  if and only if  $A$  is maximal. In these terms, therefore, what we must prove is that there exists in  $\mathfrak{X}$  a set  $A$  such that  $g(A) = A$ . It turns out that the crucial prop-

erty of  $g$  is the fact that  $g(A)$  (which always includes  $A$ ) contains at most one more element than  $A$ .

Now, to facilitate the exposition, we introduce a temporary definition. We shall say that a subcollection  $\mathfrak{I}$  of  $\mathfrak{X}$  is a *tower* if

- (i)  $\emptyset \in \mathfrak{I}$ ,
- (ii) if  $A \in \mathfrak{I}$ , then  $g(A) \in \mathfrak{I}$ ,
- (iii) if  $\mathfrak{C}$  is a chain in  $\mathfrak{I}$ , then  $\bigcup_{A \in \mathfrak{C}} A \in \mathfrak{I}$ .

Towers surely exist; the whole collection  $\mathfrak{X}$  is one. Since the intersection of a collection of towers is again a tower, it follows, in particular, that if  $\mathfrak{J}_0$  is the intersection of all towers, then  $\mathfrak{J}_0$  is the smallest tower. Our immediate purpose is to prove that the tower  $\mathfrak{J}_0$  is a chain.

Let us say that a set  $C$  in  $\mathfrak{J}_0$  is *comparable* if it is comparable with every set in  $\mathfrak{J}_0$ ; this means that if  $A \in \mathfrak{J}_0$ , then either  $A \subset C$  or  $C \subset A$ . To say that  $\mathfrak{J}_0$  is a chain means that all the sets in  $\mathfrak{J}_0$  are comparable. Comparable sets surely exist;  $\emptyset$  is one of them. In the next couple of paragraphs we concentrate our attention on an arbitrary but temporarily fixed comparable set  $C$ .

Suppose that  $A \in \mathfrak{J}_0$  and  $A$  is a proper subset of  $C$ . Assertion:  $g(A) \subset C$ . The reason is that since  $C$  is comparable, either  $g(A) \subset C$  or  $C$  is a proper subset of  $g(A)$ . In the latter case  $A$  is a proper subset of a proper subset of  $g(A)$ , and this contradicts the fact that  $g(A) - A$  cannot be more than a singleton.

Consider next the collection  $\mathfrak{U}$  of all those sets  $A$  in  $\mathfrak{J}_0$  for which either  $A \subset C$  or  $g(C) \subset A$ . The collection  $\mathfrak{U}$  is somewhat smaller than the collection of sets in  $\mathfrak{J}_0$  comparable with  $g(C)$ ; indeed if  $A \in \mathfrak{U}$ , then, since  $C \subset g(C)$ , either  $A \subset g(C)$  or  $g(C) \subset A$ . Assertion:  $\mathfrak{U}$  is a tower. Since  $\emptyset \subset C$ , the first condition on towers is satisfied. To prove the second condition, i.e., that if  $A \in \mathfrak{U}$ , then  $g(A) \in \mathfrak{U}$ , split the discussion into three cases. First:  $A$  is a proper subset of  $C$ . Then  $g(A) \subset C$  by the preceding paragraph, and therefore  $g(A) \in \mathfrak{U}$ . Second:  $A = C$ . Then  $g(A) = g(C)$ , so that  $g(C) \subset g(A)$ , and therefore  $g(A) \in \mathfrak{U}$ . Third:  $g(C) \subset A$ . Then  $g(C) \subset g(A)$ , and therefore  $g(A) \in \mathfrak{U}$ . The third condition on towers, i.e., that the union of a chain in  $\mathfrak{U}$  belongs to  $\mathfrak{U}$ , is immediate from the definition of  $\mathfrak{U}$ . Conclusion:  $\mathfrak{U}$  is a tower included in  $\mathfrak{J}_0$ , and therefore, since  $\mathfrak{J}_0$  is the smallest tower,  $\mathfrak{U} = \mathfrak{J}_0$ .

The preceding considerations imply that for each comparable set  $C$  the set  $g(C)$  is comparable also. Reason: given  $C$ , form  $\mathfrak{U}$  as above; the fact that  $\mathfrak{U} = \mathfrak{J}_0$  means that if  $A \in \mathfrak{J}_0$ , then either  $A \subset C$  (in which case  $A \subset g(C)$ ) or  $g(C) \subset A$ .

We now know that  $\emptyset$  is comparable and that  $g$  maps comparable sets onto comparable sets. Since the union of a chain of comparable sets is comparable, it follows that the comparable sets (in  $\mathfrak{I}_0$ ) constitute a tower, and hence that they exhaust  $\mathfrak{I}_0$ ; this is what we set out to prove about  $\mathfrak{I}_0$ .

Since  $\mathfrak{I}_0$  is a chain, the union, say  $A$ , of all the sets in  $\mathfrak{I}_0$  is itself a set in  $\mathfrak{I}_0$ . Since the union includes all the sets in  $\mathfrak{I}_0$ , it follows that  $g(A) \subset A$ . Since always  $A \subset g(A)$ , it follows that  $A = g(A)$ , and the proof of Zorn's lemma is complete.

**EXERCISE.** Zorn's lemma is equivalent to the axiom of choice. [Hint for the proof: given a set  $X$ , consider functions  $f$  such that  $\text{dom } f \subset \mathcal{P}(X)$ ,  $\text{ran } f \subset X$ , and  $f(A) \in A$  for all  $A$  in  $\text{dom } f$ ; order these functions by extension, use Zorn's lemma to find a maximal one among them, and prove that if  $f$  is maximal, then  $\text{dom } f = \mathcal{P}(X) - \{\emptyset\}$ .] Consider each of the following statements and prove that they too are equivalent to the axiom of choice. (i) Every partially ordered set has a maximal chain (i.e., a chain that is not a proper subset of any other chain). (ii) Every chain in a partially ordered set is included in some maximal chain. (iii) Every partially ordered set in which each chain has a least upper bound has a maximal element.

## SECTION 17

---

### WELL ORDERING

---

A partially ordered set may not have a smallest element; and, even if it has one, it is perfectly possible that some subset will fail to have one. A partially ordered set is called *well ordered* (and its ordering is called a *well ordering*) if every non-empty subset of it has a smallest element. One consequence of this definition, worth noting even before we look at any examples and counterexamples, is that every well ordered set is totally ordered. Indeed, if  $x$  and  $y$  are elements of a well ordered set, then  $\{x, y\}$  is a non-empty subset of that well ordered set and has therefore a first element; according as that first element is  $x$  or  $y$ , we have  $x \leq y$  or  $y \leq x$ .

For each natural number  $n$ , the set of all predecessors of  $n$  (that is, in accordance with our definitions, the set  $n$ ) is a well ordered set (ordered by magnitude), and the same is true of the set  $\omega$  of all natural numbers. The set  $\omega \times \omega$ , with  $(a, b) \leq (x, y)$  defined to mean  $(2a + 1)2^y \leq (2x + 1)2^b$  is not well ordered. One way to see this is to note that  $(a, b + 1) \leq (a, b)$  for all  $a$  and  $b$ ; it follows that the entire set  $\omega \times \omega$  has no least element. Some subsets of  $\omega \times \omega$  do have a least element. Consider, for example, the set  $E$  of all those pairs  $(a, b)$  for which  $(1, 1) \leq (a, b)$ ; the set  $E$  has  $(1, 1)$  for its least element. Caution:  $E$ , considered as a partially ordered set on its own right, is still not well ordered. The trouble is that even though  $E$  has a least element, many subsets of  $E$  fail to have one; for an example consider the set of all those pairs  $(a, b)$  in  $E$  for which  $(a, b) \not\leq (1, 1)$ . One more example:  $\omega \times \omega$  is well ordered by its lexicographical ordering.

One of the pleasantest facts about well ordered sets is that we can prove things about their elements by a process similar to mathematical induction. Precisely speaking, suppose that  $S$  is a subset of a well ordered set  $X$ , and suppose that whenever an element  $x$  of  $X$  is such that the entire initial segment  $s(x)$  is included in  $S$ , then  $x$  itself belongs to  $S$ ; the **principle of transfinite induction** asserts that under these circumstances we must

have  $S = X$ . Equivalently: if the presence in a set of all the strict predecessors of an element always implies the presence of the element itself, then the set must contain everything.

A few remarks are in order before we look at the proof. The statement of the ordinary principle of mathematical induction differs from that of transfinite induction in two conspicuous respects. One: the latter, instead of passing to each element from its predecessor, passes to each element from the set of all its predecessors. Two: in the latter there is no assumption about a starting element (such as zero). The first difference is important: an element in a well ordered set may fail to have an immediate predecessor. The present statement when applied to  $\omega$  is easily proved to be equivalent to the principle of mathematical induction; that principle, however, when applied to an arbitrary well ordered set, is not equivalent to the principle of transfinite induction. To put it differently: the two statements are in general not equivalent to each other; their equivalence in  $\omega$  is a happy but special circumstance.

Here is an example. Let  $X$  be  $\omega^+$ , i.e.,  $X = \omega \cup \{\omega\}$ . Define order in  $X$  by ordering the elements of  $\omega$  as usual and by requiring that  $n < \omega$  for all  $n$  in  $\omega$ . The result is a well ordered set. Question: does there exist a proper subset  $S$  of  $X$  such that  $0 \in S$  and such that  $n + 1 \in S$  whenever  $n \in S$ ? Answer: yes, namely  $S = \omega$ .

The second difference between ordinary induction and transfinite induction (no starting element required for the latter) is more linguistic than conceptual. If  $x_0$  is the smallest element of  $X$ , then  $s(x_0)$  is empty, and, consequently,  $s(x_0) \subset S$ ; the hypothesis of the principle of transfinite induction requires therefore that  $x_0$  belong to  $S$ .

The proof of the principle of transfinite induction is almost trivial. If  $X - S$  is not empty, then it has a smallest element, say  $x$ . This implies that every element of the initial segment  $s(x)$  belongs to  $S$ , and hence, by the induction hypothesis, that  $x$  belongs to  $S$ . This is a contradiction ( $x$  cannot belong to both  $S$  and  $X - S$ ); the conclusion is that  $X - S$  is empty after all.

We shall say that a well ordered set  $A$  is a *continuation* of a well ordered set  $B$ , if, in the first place,  $B$  is a subset of  $A$ , if, in fact,  $B$  is an initial segment of  $A$ , and if, finally, the ordering of the elements in  $B$  is the same as their ordering in  $A$ . Thus if  $X$  is a well ordered set and if  $a$  and  $b$  are elements of  $X$  with  $b < a$ , then  $s(a)$  is a continuation of  $s(b)$ , and, of course,  $X$  is a continuation of both  $s(a)$  and  $s(b)$ .

If  $\mathcal{C}$  is an arbitrary collection of initial segments of a well ordered set, then  $\mathcal{C}$  is a chain with respect to continuation; this means that  $\mathcal{C}$  is a collec-

tion of well ordered sets with the property that of any two distinct members of the collection one is a continuation of the other. A sort of converse of this comment is also true and is frequently useful. If a collection  $\mathcal{C}$  of well ordered sets is a chain with respect to continuation, and if  $U$  is the union of the sets of  $\mathcal{C}$ , then there is a unique well ordering of  $U$  such that  $U$  is a continuation of each set (distinct from  $U$  itself) in the collection  $\mathcal{C}$ . Roughly speaking, the union of a chain of well ordered sets is well ordered. This abbreviated formulation is dangerous because it does not explain that “chain” is meant with respect to continuation. If the ordering implied by the word “chain” is taken to be simply order-preserving inclusion, then the conclusion is not valid.

The proof is straightforward. If  $a$  and  $b$  are in  $U$ , then there exist sets  $A$  and  $B$  in  $\mathcal{C}$  with  $a \in A$  and  $b \in B$ . Since either  $A = B$  or one of  $A$  and  $B$  is a continuation of the other, it follows that in every case both  $a$  and  $b$  belong to some one set in  $\mathcal{C}$ ; the order of  $U$  is defined by ordering each pair  $\{a, b\}$  the way it is ordered in any set of  $\mathcal{C}$  that contains both  $a$  and  $b$ . Since  $\mathcal{C}$  is a chain, this order is unambiguously determined. (An alternative way of defining the promised order in  $U$  is to recall that the given orders, in the sets of  $\mathcal{C}$ , are sets of ordered pairs, and to form the union of all those sets of ordered pairs.)

A direct verification shows that the relation defined in the preceding paragraph is indeed an order, and that, moreover, its construction was forced on us at every step (i.e., that the final order is uniquely determined by the given orders). The proof that the result is actually a well ordering is equally direct. Each non-empty subset of  $U$  must have a non-empty intersection with some set in  $\mathcal{C}$ , and hence it must have a first element in that set; the fact that  $\mathcal{C}$  is a continuation chain implies that that first element is necessarily the first element of  $U$  also.

**EXERCISE.** A subset  $A$  of a partially ordered set  $X$  is *cofinal* in  $X$  in case for each element  $x$  of  $X$  there exists an element  $a$  of  $A$  such that  $x \leq a$ . Prove that every totally ordered set has a cofinal well ordered subset.

The importance of well ordering stems from the following result, from which we may infer, among other things, that the principle of transfinite induction is much more widely applicable than a casual glance might indicate.

**Well ordering theorem.** *Every set can be well ordered.*

**DISCUSSION.** A better (but less traditional) statement is this: for each set  $X$ , there is a well ordering with domain  $X$ . Warning: the well ordering

is not promised to have any relation whatsoever to any other structure that the given set might already possess. If, for instance, the reader knows of some partially or totally ordered sets whose ordering is very definitely not a well ordering, he should not jump to the conclusion that he has discovered a paradox. The only conclusion to be drawn is that some sets can be ordered in many ways, some of which are well orderings and others are not, and we already knew that.

PROOF. We apply Zorn's lemma. Given the set  $X$ , consider the collection  $\mathbb{W}$  of all well ordered subsets of  $X$ . Explicitly: an element of  $\mathbb{W}$  is a subset  $A$  of  $X$  together with a well ordering of  $A$ . We partially order  $\mathbb{W}$  by continuation.

The collection  $\mathbb{W}$  is not empty, because, for instance,  $\emptyset \in \mathbb{W}$ . If  $X \neq \emptyset$ , less annoying elements of  $\mathbb{W}$  can be exhibited; one such is  $\{(x, x)\}$ , for any particular element  $x$  of  $X$ . If  $C$  is a chain in  $\mathbb{W}$ , then the union  $U$  of the sets in  $C$  has a unique well ordering that makes  $U$  "larger" than (or equal to) each set in  $C$ ; this is exactly what our preceding discussion of continuation has accomplished. This means that the principal hypothesis of Zorn's lemma has been verified; the conclusion is that there exists a maximal well ordered set, say  $M$ , in  $\mathbb{W}$ . The set  $M$  must be equal to the entire set  $X$ . Reason: if  $x$  is an element of  $X$  not in  $M$ , then  $M$  can be enlarged by putting  $x$  after all the elements of  $M$ . The rigorous formulation of this unambiguous but informal description is left as an exercise for the reader. With that out of the way, the proof of the well ordering theorem is complete.

EXERCISE. Prove that a totally ordered set is well ordered if and only if the set of strict predecessors of each element is well ordered. Does any such condition apply to partially ordered sets? Prove that the well ordering theorem implies the axiom of choice (and hence is equivalent to that axiom and to Zorn's lemma). Prove that if  $R$  is a partial order in a set  $X$ , then there exists a total order  $S$  in  $X$  such that  $R \subset S$ ; in other words, every partial order can be extended to a total order without enlarging the domain.

## SECTION 18

---

### TRANSFINITE RECURSION

---

The process of “definition by induction” has a transfinite analogue. The ordinary recursion theorem constructs a function on  $\omega$ ; the raw material is a way of getting the value of the function at each non-zero element  $n$  of  $\omega$  from its value at the element preceding  $n$ . The transfinite analogue constructs a function on any well ordered set  $W$ ; the raw material is a way of getting the value of the function at each element  $a$  of  $W$  from its values at all the predecessors of  $a$ .

To be able to state the result concisely, we introduce some auxiliary concepts. If  $a$  is an element of a well ordered set  $W$ , and if  $X$  is an arbitrary set, then by a *sequence of type a in X* we shall mean a function from the initial segment of  $a$  in  $W$  into  $X$ . The sequences of type  $a$ , for  $a$  in  $\omega^+$ , are just what we called sequences before, finite or infinite according as  $a < \omega$  or  $a = \omega$ . If  $U$  is a function from  $W$  to  $X$ , then the restriction of  $U$  to the initial segment  $s(a)$  of  $a$  is an example of a sequence of type  $a$  for each  $a$  in  $W$ ; in what follows we shall find it convenient to denote that sequence by  $U^a$  (instead of  $U|s(a)$ ).

A *sequence function of type W in X* is a function  $f$  whose domain consists of all sequences of type  $a$  in  $X$ , for all elements  $a$  in  $W$ , and whose range is included in  $X$ . Roughly speaking, a sequence function tells us how to “lengthen” a sequence; given a sequence that stretches up to (but not including) some element of  $W$  we can use a sequence function to tack on one more term.

**Transfinite recursion theorem.** *If  $W$  is a well ordered set, and if  $f$  is a sequence function of type  $W$  in a set  $X$ , then there exists a unique function  $U$  from  $W$  into  $X$  such that  $U(a) = f(U^a)$  for each  $a$  in  $W$ .*

**PROOF.** The proof of uniqueness is an easy transfinite induction. To prove existence, recall that a function from  $W$  to  $X$  is a certain kind of subset of  $W \times X$ ; we shall construct  $U$  explicitly as a set of ordered pairs. Call a subset  $A$  of  $W \times X$  *f-closed* if it has the following property: whenever  $a \in W$  and  $t$  is a sequence of type  $a$  included in  $A$  (that is,  $(c, t(c)) \in A$  for all  $c$  in the initial segment  $s(a)$ ), then  $(a, f(t)) \in A$ . Since  $W \times X$  itself is *f*-closed, such sets do exist; let  $U$  be the intersection of them all. Since  $U$  itself is *f*-closed, it remains only to prove that  $U$  is a function. We are to prove, in other words, that for each  $c$  in  $W$  there exists at most one element  $x$  in  $X$  such that  $(c, x) \in U$ . (Explicitly: if both  $(c, x)$  and  $(c, y)$  belong to  $U$ , then  $x = y$ .) The proof is inductive. Let  $S$  be the set of all those elements  $c$  of  $W$  for which it is indeed true that  $(c, x) \in U$  for at most one  $x$ . We shall prove that if  $s(a) \subset S$ , then  $a \in S$ .

To say that  $s(a) \subset S$  means that if  $c < a$  in  $W$ , then there exists a unique element  $x$  in  $X$  such that  $(c, x) \in U$ . The correspondence  $c \rightarrow x$  thereby defined is a sequence of type  $a$ , say  $t$ , and  $t \subset U$ . If  $a$  does not belong to  $S$ , then  $(a, y) \in U$  for some  $y$  different from  $f(t)$ . Assertion: the set  $U - \{(a, y)\}$  is *f*-closed. This means that if  $b \in W$  and if  $r$  is a sequence of type  $b$  included in  $U - \{(a, y)\}$ , then  $(b, f(r)) \in U - \{(a, y)\}$ . Indeed, if  $b = a$ , then  $r$  must be  $t$  (by the uniqueness assertion of the theorem), and the reason the diminished set contains  $(b, f(r))$  is that  $f(t) \neq y$ ; if, on the other hand,  $b \neq a$ , then the reason the diminished set contains  $(b, f(r))$  is that  $U$  is *f*-closed (and  $b \neq a$ ). This contradicts the fact that  $U$  is the smallest *f*-closed set, and we may conclude that  $a \in S$ .

The proof of the existence assertion of the transfinite recursion theorem is complete. An application of the transfinite recursion theorem is called *definition by transfinite induction*.

We continue with an important part of the theory of order, which, incidentally, will also serve as an illustration of how the transfinite recursion theorem can be applied.

Two partially ordered sets (which may in particular be totally ordered and even well ordered) are called *similar* if there exists an order-preserving one-to-one correspondence between them. More explicitly: to say of the partially ordered sets  $X$  and  $Y$  that they are similar (in symbols  $X \cong Y$ ) means that there exists a one-to-one correspondence, say  $f$ , from  $X$  onto  $Y$ , such that if  $a$  and  $b$  are in  $X$ , then a necessary and sufficient condition that  $f(a) \leq f(b)$  (in  $Y$ ) is that  $a \leq b$  (in  $X$ ). A correspondence such as  $f$  is sometimes called a *similarity*.

**EXERCISE.** Prove that a similarity preserves  $<$  (in the same sense in which the definition demands the preservation of  $\leq$ ) and that, in fact,  $\varepsilon$

one-to-one function that maps one partially ordered set onto another is a similarity if and only if it preserves  $<$ .

The identity mapping on a partially ordered set  $X$  is a similarity from  $X$  onto  $X$ . If  $X$  and  $Y$  are partially ordered sets and if  $f$  is a similarity from  $X$  onto  $Y$ , then (since  $f$  is one-to-one) there exists an unambiguously determined inverse function  $f^{-1}$  from  $Y$  onto  $X$ , and  $f^{-1}$  is a similarity. If, moreover,  $g$  is a similarity from  $Y$  onto a partially ordered set  $Z$ , then the composite  $gf$  is a similarity from  $X$  onto  $Z$ . It follows from these comments that if we restrict attention to some particular set  $E$ , and if, accordingly, we consider only such partial orders whose domain is a subset of  $E$ , then similarity is an equivalence relation in the set of partially ordered sets so obtained. The same is true if we narrow the field even further and consider only well orderings whose domain is included in  $E$ ; similarity is an equivalence relation in the set of well ordered sets so obtained. Although similarity was defined for partially ordered sets in complete generality, and the subject can be studied on that level, our interest in what follows will be in similarity for well ordered sets only.

It is easily possible for a well ordered set to be similar to a proper subset; for an example consider the set of all natural numbers and the set of all even numbers. (As always, a natural number  $m$  is defined to be even if there exists a natural number  $n$  such that  $m = 2n$ . The mapping  $n \rightarrow 2n$  is a similarity from the set of all natural numbers onto the set of all even numbers.) A similarity of a well ordered set with a part of itself is, however, a very special kind of mapping. If, in fact,  $f$  is a similarity of a well ordered set  $X$  into itself, then  $a \leq f(a)$  for each  $a$  in  $X$ . The proof is based directly on the definition of well ordering. If there are elements  $b$  such that  $f(b) < b$ , then there is a least one among them. If  $a < b$ , where  $b$  is that least one, then  $a \leq f(a)$ ; it follows, in particular, with  $a = f(b)$ , that  $f(b) \leq f(f(b))$ . Since, however,  $f(b) < b$ , the order-preserving character of  $f$  implies that  $f(f(b)) < f(b)$ . The only way out of the contradiction is to admit the impossibility of  $f(b) < b$ .

The result of the preceding paragraph has three especially useful consequences. The first of these is the fact that if two well ordered sets,  $X$  and  $Y$  say, are similar at all, then there is just one similarity between them. Suppose indeed that both  $g$  and  $h$  are similarities from  $X$  onto  $Y$ , and write  $f = g^{-1}h$ . Since  $f$  is a similarity of  $X$  onto itself, it follows that  $a \leq f(a)$  for each  $a$  in  $X$ . This means that  $a \leq g^{-1}(h(a))$  for each  $a$  in  $X$ . Applying  $g$ , we infer that  $g(a) \leq h(a)$  for each  $a$  in  $X$ . The situation is symmetric in  $g$  and  $h$ , so that we may also infer that  $h(a) \leq g(a)$  for each  $a$  in  $X$ . Conclusion:  $g = h$ .

A second consequence is the fact that a well ordered set is never similar to one of its initial segments. If, indeed,  $X$  is a well ordered set,  $a$  is an element of  $X$ , and  $f$  is a similarity from  $X$  onto  $s(a)$ , then, in particular,  $f(a) \in s(a)$ , so that  $f(a) < a$ , and that is impossible.

The third and chief consequence is the comparability theorem for well ordered sets. The assertion is that if  $X$  and  $Y$  are well ordered sets, then either  $X$  and  $Y$  are similar, or one of them is similar to an initial segment of the other. Just for practice we shall use the transfinite recursion theorem in the proof, although it is perfectly easy to avoid it. We assume that  $X$  and  $Y$  are non-empty well ordered sets such that neither is similar to an initial segment of the other; we proceed to prove that under these circumstances  $X$  must be similar to  $Y$ . Suppose that  $a \in X$  and that  $t$  is a sequence of type  $a$  in  $Y$ ; in other words  $t$  is a function from  $s(a)$  into  $Y$ . Let  $f(t)$  be the least of the proper upper bounds of the range of  $t$  in  $Y$ , if there are any; in the contrary case, let  $f(t)$  be the least element of  $Y$ . In the terminology of the transfinite recursion theorem, the function  $f$  thereby determined is a sequence function of type  $X$  in  $Y$ . Let  $U$  be the function that the transfinite recursion theorem associates with this situation. An easy argument (by transfinite induction) shows that, for each  $z$  in  $X$ , the function  $U$  maps the initial segment determined by  $a$  in  $X$  one-to-one onto the initial segment determined by  $U(a)$  in  $Y$ . This implies that  $U$  is a similarity, and the proof is complete.

Here is a sketch of an alternative proof that does not use the transfinite recursion theorem. Let  $X_0$  be the set of those elements  $a$  of  $X$  for which there exists an element  $b$  of  $Y$  such that  $s(a)$  is similar to  $s(b)$ . For each  $a$  in  $X_0$ , write  $U(a)$  for the corresponding (uniquely determined)  $b$  in  $Y$ , and let  $Y_0$  be the range of  $U$ . It follows that either  $X_0 = X$ , or else  $X_0$  is an initial segment of  $X$  and  $Y_0 = Y$ .

**EXERCISE.** Each subset of a well ordered set  $X$  is similar either to  $X$  or to an initial segment of  $X$ . If  $X$  and  $Y$  are well ordered sets and  $X \cong Y$  (i.e.,  $X$  is similar to  $Y$ ), then the similarity maps the least upper bound (if any) of each subset of  $X$  onto the least upper bound of the image of that subset.

## SECTION 19

---

### ORDINAL NUMBERS

---

The successor  $x^+$  of a set  $x$  was defined as  $x \cup \{x\}$ , and then  $\omega$  was constructed as the smallest set that contains 0 and that contains  $x^+$  whenever it contains  $x$ . What happens if we start with  $\omega$ , form its successor  $\omega^+$ , then form the successor of that, and proceed so on ad infinitum? In other words: is there something out beyond  $\omega, \omega^+, (\omega^+)^+, \dots$ , etc., in the same sense in which  $\omega$  is beyond 0, 1, 2,  $\dots$ , etc.?

The question calls for a set, say  $T$ , containing  $\omega$ , such that each element of  $T$  (other than  $\omega$  itself) can be obtained from  $\omega$  by the repeated formation of successors. To formulate this requirement more precisely we introduce some special and temporary terminology. Let us say that a function  $f$  whose domain is the set of strict predecessors of some natural number  $n$  (in other words,  $\text{dom } f = n$ ) is an  $\omega$ -successor function if  $f(0) = \omega$  (provided that  $n \neq 0$ , so that  $0 < n$ ), and  $f(m^+) = (f(m))^+$  whenever  $m^+ < n$ . An easy proof by mathematical induction shows that for each natural number  $n$  there exists a unique  $\omega$ -successor function with domain  $n$ . To say that something is either equal to  $\omega$  or can be obtained from  $\omega$  by the repeated formation of successors means that it belongs to the range of some  $\omega$ -successor function. Let  $S(n, x)$  be the sentence that says “ $n$  is a natural number and  $x$  belongs to the range of the  $\omega$ -successor function with domain  $n$ .” A set  $T$  such that  $x \in T$  if and only if  $S(n, x)$  is true for some  $n$  is what we are looking for; such a set is as far beyond  $\omega$  as  $\omega$  is beyond 0.

We know that for each natural number  $n$  we are permitted to form the set  $\{x: S(n, x)\}$ . In other words, for each natural number  $n$ , there exists a set  $F(n)$  such that  $x \in F(n)$  if and only if  $S(n, x)$  is true. The connection between  $n$  and  $F(n)$  looks very much like a function. It turns out, how-

ever, that none of the methods of set construction that we have seen so far is sufficiently strong to prove the existence of a set  $F$  of ordered pairs such that  $(n, x) \in F$  if and only if  $x \in F(n)$ . To achieve this obviously desirable state of affairs, we need one more set-theoretic principle (our last). The new principle says, roughly speaking, that anything intelligent that one can do to the elements of a set yields a set.

**Axiom of substitution.** *If  $S(a, b)$  is a sentence such that for each  $a$  in a set  $A$  the set  $\{b: S(a, b)\}$  can be formed, then there exists a function  $F$  with domain  $A$  such that  $F(a) = \{b: S(a, b)\}$  for each  $a$  in  $A$ .*

To say that  $\{b: S(a, b)\}$  can be formed means, of course, that there exists a set  $F(a)$  such that  $b \in F(a)$  if and only if  $S(a, b)$  is true. The axiom of extension implies that the function described in the axiom of substitution is uniquely determined by the given sentence and the given set. The reason for the name of the axiom is that it enables us to make a new set out of an old one by substituting something new for each element of the old.

The chief application of the axiom of substitution is in extending the process of counting far beyond the natural numbers. From the present point of view, the crucial property of a natural number is that it is a well ordered set such that the initial segment determined by each element is equal to that element. (Recall that if  $m$  and  $n$  are natural numbers, then  $m < n$  means  $m \in n$ ; this implies that  $\{m \in \omega: m < n\} = n$ .) This is the property on which the extended counting process is based; the fundamental definition in this circle of ideas is due to von Neumann. An *ordinal number* is defined as a well ordered set  $\alpha$  such that  $s(\xi) = \xi$  for all  $\xi$  in  $\alpha$ ; here  $s(\xi)$  is, as before, the initial segment  $\{\eta \in \alpha: \eta < \xi\}$ .

An example of an ordinal number that is not a natural number is the set  $\omega$  consisting of all the natural numbers. This means that we can already “count” farther than we could before; whereas before the only numbers at our disposal were the elements of  $\omega$ , now we have  $\omega$  itself. We have also the successor  $\omega^+$  of  $\omega$ ; this set is ordered in the obvious way, and, moreover, the obvious ordering is a well ordering that satisfies the condition imposed on ordinal numbers. Indeed, if  $\xi \in \omega^+$ , then, by the definition of successor, either  $\xi \in \omega$ , in which case we already know that  $s(\xi) = \xi$ , or else  $\xi = \omega$ , in which case  $s(\xi) = \omega$ , by the definition of order, so that again  $s(\xi) = \xi$ . The argument just presented is quite general; it proves that if  $\alpha$  is an ordinal number, then so is  $\alpha^+$ . It follows that our counting process extends now up to and including  $\omega$ , and  $\omega^+$ , and  $(\omega^+)^+$ , and so on ad infinitum.

At this point we make contact with our earlier discussion of what happens beyond  $\omega$ . The axiom of substitution implies easily that there exists a unique function  $F$  on  $\omega$  such that  $F(0) = \omega$  and  $F(n^+) = (F(n))^+$  for each natural number  $n$ . The range of this function is a set of interest for us; a set of even greater importance is the union of the set  $\omega$  with the range of the function  $F$ . For reasons that will become clear only after we have at least glanced at the arithmetic of ordinal numbers, that union is usually denoted by  $\omega_2$ . If, borrowing again from the notation of ordinal arithmetic, we write  $\omega + n$  for  $F(n)$ , then we can describe the set  $\omega_2$  as the set consisting of all  $n$  (with  $n$  in  $\omega$ ) and of all  $\omega + n$  (with  $n$  in  $\omega$ ).

It is now easy to verify that  $\omega_2$  is an ordinal number. The verification depends, of course, on the definition of order in  $\omega_2$ . At this point both that definition and the proof are left as exercises; our official attention turns to some general remarks that include the facts about  $\omega_2$  as easy special cases.

An order (partial or total) in a set  $X$  is uniquely determined by its initial segments. If, in other words,  $R$  and  $S$  are orders in  $X$ , and if, for each  $x$  in  $X$ , the set of all  $R$ -predecessors of  $x$  is the same as the set of all  $S$ -predecessors of  $x$ , then  $R$  and  $S$  are the same. This assertion is obvious whether predecessors are taken in the strict sense or not. The assertion applies, in particular, to well ordered sets. From this special case we infer that if it is possible at all to well order a set so as to make it an ordinal number, then there is only one way to do so. The set alone tells us what the relation that makes it an ordinal number must be; if that relation satisfies the requirements, then the set is an ordinal number, and otherwise it is not. To say that  $s(\xi) = \xi$  means that the predecessors of  $\xi$  must be just the elements of  $\xi$ . The relation in question is therefore simply the relation of belonging. If  $\eta < \xi$  is defined to mean  $\eta \in \xi$  whenever  $\xi$  and  $\eta$  are elements of a set  $\alpha$ , then the result either is or is not a well ordering of  $\alpha$  such that  $s(\xi) = \xi$  for each  $\xi$  in  $\alpha$ , and  $\alpha$  is an ordinal number in the one case and not in the other.

We conclude this preliminary discussion of ordinal numbers by mentioning the names of the first few of them. After  $0, 1, 2, \dots$  comes  $\omega$ , and after  $\omega, \omega + 1, \omega + 2, \dots$  comes  $\omega_2$ . After  $\omega_2 + 1$  (that is, the successor of  $\omega_2$ ) comes  $\omega_2 + 2$ , and then  $\omega_2 + 3$ ; next after all the terms of the sequence so begun comes  $\omega_3$ . (Another application of the axiom of substitution is needed at this point.) Next come  $\omega_3 + 1, \omega_3 + 2, \omega_3 + 3, \dots$ , and after them comes  $\omega_4$ . In this way we get successively  $\omega, \omega_2, \omega_3, \omega_4, \dots$ . An application of the axiom of substitution yields something that follows them all in the same sense in which  $\omega$  follows the natural numbers; that some-

thing is  $\omega^2$ . After that the whole thing starts over again:  $\omega^2 + 1, \omega^2 + 2, \dots, \omega^2 + \omega, \omega^2 + \omega + 1, \omega^2 + \omega + 2, \dots, \omega^2 + \omega 2, \omega^2 + \omega 2 + 1, \dots, \omega^2 + \omega 3, \dots, \omega^2 + \omega 4, \dots, \omega^2 2, \dots, \omega^2 3, \dots, \omega^3, \dots, \omega^4, \dots, \omega^\omega, \dots, \omega^{(\omega^\omega)}, \dots, \omega^{(\omega^{(\omega^\omega)})}, \dots$ . The next one after all that is  $\varepsilon_0$ ; then come  $\varepsilon_0 + 1, \varepsilon_0 + 2, \dots, \varepsilon_0 + \omega, \dots, \varepsilon_0 + \omega 2, \dots, \varepsilon_0 + \omega^2, \dots, \varepsilon_0 + \omega^\omega, \dots, \varepsilon_0 2, \dots, \varepsilon_0 \omega, \dots, \varepsilon_0 \omega^\omega, \dots, \varepsilon_0^2, \dots \dots \dots$ .

## SECTION 20

---

### SETS OF ORDINAL NUMBERS

---

An ordinal number is, by definition, a special kind of well ordered set; we proceed to examine its special properties.

The most elementary fact is that each element of an ordinal number  $\alpha$  is at the same time a subset of  $\alpha$ . (In other words, every ordinal number is a transitive set.) Indeed, if  $\xi \in \alpha$ , then the fact that  $s(\xi) = \xi$  implies that each element of  $\xi$  is a predecessor of  $\xi$  in  $\alpha$  and hence, in particular, an element of  $\alpha$ .

If  $\xi$  is an element of an ordinal number  $\alpha$ , then, as we have just seen,  $\xi$  is a subset of  $\alpha$ , and, consequently,  $\xi$  is a well ordered set (with respect to the ordering it inherits from  $\alpha$ ). Assertion:  $\xi$  is in fact an ordinal number. Indeed, if  $\eta \in \xi$ , then the initial segment determined by  $\eta$  in  $\xi$  is the same as the initial segment determined by  $\eta$  in  $\alpha$ ; since the latter is equal to  $\eta$ , so is the former. Another way of formulating the same result is to say that every initial segment of an ordinal number is an ordinal number.

The next thing to note is that if two ordinal numbers are similar, then they are equal. To prove this, suppose that  $\alpha$  and  $\beta$  are ordinal numbers and that  $f$  is a similarity from  $\alpha$  onto  $\beta$ ; we shall show that  $f(\xi) = \xi$  for each  $\xi$  in  $\alpha$ . The proof is a straightforward transfinite induction. Write  $S = \{\xi \in \alpha : f(\xi) = \xi\}$ . For each  $\xi$  in  $\alpha$ , the least element of  $\alpha$  that does not belong to  $s(\xi)$  is  $\xi$  itself. Since  $f$  is a similarity, it follows that the least element of  $\beta$  that does not belong to the image of  $s(\xi)$  under  $f$  is  $f(\xi)$ . These assertions imply that if  $s(\xi) \subset S$ , then  $f(\xi)$  and  $\xi$  are ordinal numbers with the same initial segments, and hence that  $f(\xi) = \xi$ . We have proved thus that  $\xi \in S$  whenever  $s(\xi) \subset S$ . The principle of transfinite induction implies that  $S = \alpha$ , and from this it follows that  $\alpha = \beta$ .

If  $\alpha$  and  $\beta$  are ordinal numbers, then, in particular, they are well ordered sets, and, consequently, either they are similar or else one of them is simi-

lar to an initial segment of the other. If, say,  $\beta$  is similar to an initial segment of  $\alpha$ , then  $\beta$  is similar to an element of  $\alpha$ . Since every element of  $\alpha$  is an ordinal number, it follows that  $\beta$  is an element of  $\alpha$ , or, in still other words, that  $\alpha$  is a continuation of  $\beta$ . We know by now that if  $\alpha$  and  $\beta$  are distinct ordinal numbers, then the statements

$$\beta \in \alpha,$$

$$\beta \subset \alpha,$$

$$\alpha \text{ is a continuation of } \beta,$$

are all equivalent to one another; if they hold, we may write

$$\beta < \alpha.$$

What we have just proved is that any two ordinal numbers are comparable; that is, if  $\alpha$  and  $\beta$  are ordinal numbers, then either  $\beta = \alpha$ , or  $\beta < \alpha$ , or  $\alpha < \beta$ .

The result of the preceding paragraph can be expressed by saying that every set of ordinal numbers is totally ordered. In fact more is true: every set of ordinal numbers is well ordered. Suppose indeed that  $E$  is a non-empty set of ordinal numbers, and let  $\alpha$  be an element of  $E$ . If  $\alpha \leq \beta$  for all  $\beta$  in  $E$ , then  $\alpha$  is the first element of  $E$  and all is well. If this is not the case, then there exists an element  $\beta$  in  $E$  such that  $\beta < \alpha$ , i.e.,  $\beta \in \alpha$ ; in other words, then  $\alpha \cap E$  is not empty. Since  $\alpha$  is a well ordered set,  $\alpha \cap E$  has a first element, say  $\alpha_0$ . If  $\beta \in E$ , then either  $\alpha \leq \beta$  (in which case  $\alpha_0 < \beta$ ), or  $\beta < \alpha$  (in which case  $\beta \in \alpha \cap E$  and therefore  $\alpha_0 \leq \beta$ ), and this proves that  $E$  has a first element, namely  $\alpha_0$ .

Some ordinal numbers are finite; they are just the natural numbers (i.e., the elements of  $\omega$ ). The others are called *transfinite*; the set  $\omega$  of all natural numbers is the smallest transfinite ordinal number. Each finite ordinal number (other than 0) has an immediate predecessor. If a transfinite ordinal number  $\alpha$  has an immediate predecessor  $\beta$ , then, just as for natural numbers,  $\alpha = \beta^+$ . Not every transfinite ordinal number does have an immediate predecessor; the ones that do not are called *limit numbers*.

Suppose now that  $C$  is a collection of ordinal numbers. Since, as we have just seen,  $C$  is a continuation chain, it follows that the union  $\alpha$  of the sets of  $C$  is a well ordered set such that for every  $\xi$  in  $C$ , distinct from  $\alpha$  itself,  $\alpha$  is a continuation of  $\xi$ . The initial segment determined by an element in  $\alpha$  is the same as the initial segment determined by that element whatever set of  $C$  it occurs in; this implies that  $\alpha$  is an ordinal number. If  $\xi \in C$ , then  $\xi \leq \alpha$ ; the number  $\alpha$  is an upper bound of the elements of

c. If  $\beta$  is another upper bound of  $C$ , then  $\xi \subset \beta$  whenever  $\xi \in C$ , and therefore, by the definition of unions,  $\alpha \subset \beta$ . This implies that  $\alpha$  is the least upper bound of  $C$ ; we have proved thus that every set of ordinal numbers has a supremum.

Is there a set that consists exactly of all the ordinal numbers? It is easy to see that the answer must be no. If there were such a set, then we could form the supremum of all ordinal numbers. That supremum would be an ordinal number greater than or equal to every ordinal number. Since, however, for each ordinal number there exists a strictly greater one (for example, its successor), this is impossible; it makes no sense to speak of the "set" of all ordinals. The contradiction, based on the assumption that there is such a set, is called the *Burali-Forti paradox*. (Burali-Forti was one man, not two.)

Our next purpose is to show that the concept of an ordinal number is not so special as it might appear, and that, in fact, each well ordered set resembles some ordinal number in all essential respects. "Resemblance" here is meant in the technical sense of similarity. An informal statement of the result is that each well ordered set can be counted.

**Counting theorem.** *Each well ordered set is similar to a unique ordinal number.*

**PROOF.** Since for ordinal numbers similarity is the same as equality, uniqueness is obvious. Suppose now that  $X$  is a well ordered set and suppose that an element  $a$  of  $X$  is such that the initial segment determined by each predecessor of  $a$  is similar to some (necessarily unique) ordinal number. If  $S(x, \alpha)$  is the sentence that says " $\alpha$  is an ordinal number and  $s(x) \cong \alpha$ ," then, for each  $x$  in  $s(a)$ , the set  $\{\alpha: S(x, \alpha)\}$  can be formed; in fact, that set is a singleton. The axiom of substitution implies the existence of a set consisting exactly of the ordinal numbers similar to the initial segments determined by the predecessors of  $a$ . It follows, whether  $a$  is the immediate successor of one of its predecessors or the supremum of them all, that  $s(a)$  is similar to an ordinal number. This argument prepares the way for an application of the principle of transfinite induction; the conclusion is that each initial segment in  $X$  is similar to some ordinal number. This fact, in turn, justifies another application of the axiom of substitution, just like the one made above; the final conclusion is, as desired, that  $X$  is similar to some ordinal number.

## SECTION 21

---

### ORDINAL ARITHMETIC

---

---

For natural numbers we used the recursion theorem to define the arithmetic operations, and, subsequently, we proved that those operations are related to the operations of set theory in various desirable ways. Thus, for instance, we know that the number of elements in the union of two disjoint finite sets  $E$  and  $F$  is equal to  $\#(E) + \#(F)$ . We observe now that this fact could have been used to define addition. If  $m$  and  $n$  are natural numbers, we could have defined their sum by finding disjoint sets  $E$  and  $F$ , with  $\#(E) = m$  and  $\#(F) = n$ , and writing  $m + n = \#(E \cup F)$ .

Corresponding to what was done and to what could have been done for natural numbers, there are two standard approaches to ordinal arithmetic. Partly for the sake of variety, and partly because in this context recursion seems less natural, we shall emphasize the set-theoretic approach instead of the recursive one.

We begin by pointing out that there is a more or less obvious way of putting two well ordered sets together to form a new well ordered set. Informally speaking, the idea is to write down one of them and then to follow it by the other. If we try to say this rigorously, we immediately encounter the difficulty that the two sets may not be disjoint. When are we supposed to write down an element that is common to the two sets? The way out of the difficulty is to make the sets disjoint. This can be done by painting their elements different colors. In more mathematical language, replace the elements of the sets by those same elements taken together with some distinguishing object, using two different objects for the two sets. In completely mathematical language: if  $E$  and  $F$  are arbitrary sets, let  $\hat{E}$  be the set of all ordered pairs  $(x, 0)$  with  $x$  in  $E$ , and let  $\hat{F}$  be the set of all ordered pairs  $(x, 1)$  with  $x$  in  $F$ . The sets  $\hat{E}$  and  $\hat{F}$  are clearly disjoint. There is an obvious one-to-one correspondence between  $E$  and  $\hat{E}$  ( $x \rightarrow (x, 0)$ ) and another one between  $F$  and  $\hat{F}$  ( $x \rightarrow (x, 1)$ ).

These correspondences can be used to carry over whatever structure  $E$  and  $F$  may possess (for example, order) to  $\hat{E}$  and  $\hat{F}$ . It follows that any time we are given two sets, with or without some additional structure, we may always replace them by disjoint sets with the same structure, and hence we may assume, with no loss of generality, that they were disjoint in the first place.

Before applying this construction to ordinal arithmetic, we observe that it can be generalized to arbitrary families of sets. If, indeed,  $\{E_i\}$  is a family, write  $\hat{E}_i$  for the set of all ordered pairs  $(x, i)$ , with  $x$  in  $E_i$ . (In other words,  $\hat{E}_i = E_i \times \{i\}$ .) The family  $\{\hat{E}_i\}$  is pairwise disjoint, and it can do anything the original family  $\{E_i\}$  could do.

Suppose now that  $E$  and  $F$  are disjoint well ordered sets. Define order in  $E \cup F$  so that pairs of elements in  $E$ , and also pairs of elements in  $F$ , retain the order they had, and so that each element of  $E$  precedes each element of  $F$ . (In ultraformal language: if  $R$  and  $S$  are the given order relations in  $E$  and  $F$  respectively, let  $E \cup F$  be ordered by  $R \cup S \cup (E \times F)$ .) The fact that  $E$  and  $F$  were well ordered implies that  $E \cup F$  is well ordered. The well ordered set  $E \cup F$  is called the *ordinal sum* of the well ordered sets  $E$  and  $F$ .

There is an easy and worth while way of extending the concept of ordinal sum to infinitely many summands. Suppose that  $\{E_i\}$  is a disjoint family of well ordered sets indexed by a well ordered set  $I$ . The ordinal sum of the family is the union  $\bigcup_i E_i$ , ordered as follows. If  $a$  and  $b$  are elements of the union, with  $a \in E_i$  and  $b \in E_j$ , then  $a < b$  means that either  $i < j$  or else  $i = j$  and  $a$  precedes  $b$  in the given order of  $E_i$ .

The definition of addition for ordinal numbers is now child's play. For each well ordered set  $X$ , let  $\text{ord } X$  be the unique ordinal number similar to  $X$ . (If  $X$  is finite, then  $\text{ord } X$  is the same as the natural number  $\#(X)$  defined earlier.) If  $\alpha$  and  $\beta$  are ordinal numbers, let  $A$  and  $B$  be disjoint well ordered sets with  $\text{ord } A = \alpha$  and  $\text{ord } B = \beta$ , and let  $C$  be the ordinal sum of  $A$  and  $B$ . The *sum*  $\alpha + \beta$  is, by definition, the ordinal number of  $C$ , so that  $\text{ord } A + \text{ord } B = \text{ord } C$ . It is important to note that the sum  $\alpha + \beta$  is independent of the particular choice of the sets  $A$  and  $B$ ; any other pair of disjoint sets, with the same ordinal numbers, would have given the same result.

These considerations extend without difficulty to the infinite case. If  $\{\alpha_i\}$  is a well ordered family of ordinal numbers indexed by a well ordered set  $I$ , let  $\{A_i\}$  be a disjoint family of well ordered sets with  $\text{ord } A_i = \alpha_i$  for each  $i$ , and let  $A$  be the ordinal sum of the family  $\{A_i\}$ . The sum  $\sum_{i \in I} \text{ord } A_i$  is, by definition, the ordinal number of  $A$ , so that  $\sum_{i \in I} \text{ord } A_i$

$= \text{ord } A$ . Here too the final result is independent of the arbitrary choice of the well ordered sets  $A_i$ ; any other choices (with the same ordinal numbers) would have given the same sum.

Some of the properties of addition for ordinal numbers are good and others are bad. On the good side of the ledger are the identities

$$\alpha + 0 = \alpha,$$

$$0 + \alpha = \alpha,$$

$$\alpha + 1 = \alpha^+,$$

and the associative law

$$\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma.$$

Equally laudable is the fact that  $\alpha < \beta$  if and only if there exists an ordinal number  $\gamma$  different from 0 such that  $\beta = \alpha + \gamma$ . The proofs of all these assertions are elementary.

Almost all the bad behavior of addition stems from the failure of the commutative law. Sample:  $1 + \omega = \omega$  (but, as we saw just above,  $\omega + 1 \neq \omega$ ). The misbehavior of addition expresses some intuitively clear facts about order. If, for instance, we tack a new element in front of an infinite sequence (of type  $\omega$ ), the result is clearly similar to what we started with, but if we tack it on at the end instead, then we have ruined similarity; the old set had no last element but the new set has one.

The main use of infinite sums is to motivate and facilitate the study of products. If  $A$  and  $B$  are well ordered sets, it is natural to define their product as the result of adding  $A$  to itself  $B$  times. To make sense out of this, we must first of all manufacture a disjoint family of well ordered sets, each of which is similar to  $A$ , indexed by the set  $B$ . The general prescription for doing this works well here; all we need to do is to write  $A_b = A \times \{b\}$  for each  $b$  in  $B$ . If now we examine the definition of ordinal sum as it applies to the family  $\{A_b\}$ , we are led to formulate the following definition. The *ordinal product* of two well ordered sets  $A$  and  $B$  is the Cartesian product  $A \times B$  with the reverse lexicographic order. In other words, if  $(a, b)$  and  $(c, d)$  are in  $A \times B$ , then  $(a, b) < (c, d)$  means that either  $b < d$  or else  $b = d$  and  $a < c$ .

If  $\alpha$  and  $\beta$  are ordinal numbers, let  $A$  and  $B$  be well ordered sets with  $\text{ord } A = \alpha$  and  $\text{ord } B = \beta$ , and let  $C$  be the ordinal product of  $A$  and  $B$ . The *product*  $\alpha\beta$  is, by definition, the ordinal number of  $C$ , so that  $(\text{ord } A)(\text{ord } B) = \text{ord } C$ . The product is unambiguously defined, independently of the arbitrary choice of the well ordered sets  $A$  and  $B$ . Alternatively, at this point we could have avoided any arbitrariness at all by

recalling that the most easily available well ordered set whose ordinal number is  $\alpha$  is the ordinal number  $\alpha$  itself (and similarly for  $\beta$ ).

Like addition, multiplication has its good and bad properties. Among the good ones are the identities

$$\alpha 0 = 0,$$

$$0\alpha = 0,$$

$$\alpha 1 = \alpha,$$

$$1\alpha = \alpha,$$

the associative law

$$\alpha(\beta\gamma) = (\alpha\beta)\gamma,$$

the left distributive law

$$\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma,$$

and the fact that if the product of two ordinal numbers is zero, then one of the factors must be zero. (Note that we use the standard convention about multiplication taking precedence over addition;  $\alpha\beta + \alpha\gamma$  denotes  $(\alpha\beta) + (\alpha\gamma)$ .)

The commutative law for multiplication fails, and so do many of its consequences. Thus, for instance,  $2\omega = \omega$  (think of an infinite sequence of ordered pairs), but  $\omega 2 \neq \omega$  (think of an ordered pair of infinite sequences). The right distributive law also fails; that is  $(\alpha + \beta)\gamma$  is in general different from  $\alpha\gamma + \beta\gamma$ . Example:  $(1 + 1)\omega = 2\omega = \omega$ , but  $1\omega + 1\omega = \omega + \omega = \omega 2$ .

Just as repeated addition led to the definition of ordinal products, repeated multiplication could be used to define ordinal exponents. Alternatively, exponentiation can be approached via transfinite recursion. The precise details are part of an extensive and highly specialized theory of ordinal numbers. At this point we shall be content with hinting at the definition and mentioning its easiest consequences. To define  $\alpha^\beta$  (where  $\alpha$  and  $\beta$  are ordinal numbers), use definition by transfinite induction (on  $\beta$ ). Begin by writing  $\alpha^0 = 1$  and  $\alpha^{\beta+1} = \alpha^\beta\alpha$ ; if  $\beta$  is a limit number, define  $\alpha^\beta$  as the supremum of the numbers of the form  $\alpha^\gamma$ , where  $\gamma < \beta$ . If this sketch of a definition is formulated with care, it follows that

$$0^\alpha = 0 \quad (\alpha \geq 1),$$

$$1^\gamma = 1,$$

$$\alpha^{\beta+\gamma} = \alpha^\beta\alpha^\gamma,$$

$$\alpha^{\beta\gamma} = (\alpha^\beta)^\gamma.$$

Not all the familiar laws of exponents hold; thus, for instance,  $(\alpha\beta)^\gamma$  is in general different from  $\alpha^\gamma\beta^\gamma$ . Example:  $(2 \cdot 2)^\omega = 4^\omega = \omega$ , but  $2^\omega \cdot 2^\omega = \omega \cdot \omega = \omega^2$ .

Warning: the exponent notation for ordinal numbers, here and below, is not consistent with our earlier use of it. The unordered set  $2^\omega$  of all functions from  $\omega$  to 2, and the well ordered set  $2^\omega$  that is the least upper bound of the sequence of ordinal numbers 2,  $2 \cdot 2$ ,  $2 \cdot 2 \cdot 2$ , etc., are not the same thing at all. There is no help for it; mathematical usage is firmly established in both camps. If, in a particular situation, the context does not reveal which of the two interpretations is to be used, then explicit verbal indication must be given.

## SECTION 22

---

### THE SCHRÖDER-BERNSTEIN THEOREM

---

The purpose of counting is to compare the size of one set with that of another; the most familiar method of counting the elements of a set is to arrange them in some appropriate order. The theory of ordinal numbers is an ingenious abstraction of the method, but it falls somewhat short of achieving the purpose. This is not to say that ordinal numbers are useless; it just turns out that their main use is elsewhere, in topology, for instance, as a source of illuminating examples and counterexamples. In what follows we shall continue to pay some attention to ordinal numbers, but they will cease to occupy the center of the stage. (It is of some importance to know that we could in fact dispense with them altogether. The theory of cardinal numbers can be constructed with the aid of ordinal numbers, or without it; both kinds of constructions have advantages.) With these prefatory remarks out of the way, we turn to the problem of comparing the sizes of sets.

The problem is to compare the sizes of sets when their elements do not appear to have anything to do with each other. It is easy enough to decide that there are more people in France than in Paris. It is not quite so easy, however, to compare the age of the universe in seconds with the population of Paris in electrons. For some mathematical examples, consider the following pairs of sets, defined in terms of an auxiliary set  $A$ : (i)  $X = A$ ,  $Y = A^+$ ; (ii)  $X = \mathcal{P}(A)$ ,  $Y = 2^A$ ; (iii)  $X$  is the set of all one-to-one mappings of  $A$  into itself,  $Y$  is the set of all finite subsets of  $A$ . In each case we may ask which of the two sets  $X$  and  $Y$  has more elements. The problem is first to find a rigorous interpretation of the question and then to answer it.

The well ordering theorem tells us that every set can be well ordered. For well ordered sets we have what seems to be a reasonable measure of

size, namely, their ordinal number. Do these two remarks solve the problem? To compare the sizes of  $X$  and  $Y$ , may we just well order each of them and then compare  $\text{ord } X$  and  $\text{ord } Y$ ? The answer is most emphatically no. The trouble is that one and the same set can be well ordered in many ways. The ordinal number of a well ordered set measures the well ordering more than it measures the set. For a concrete example consider the set  $\omega$  of all natural numbers. Introduce a new order by placing 0 after everything else. (In other words, if  $n$  and  $m$  are non-zero natural numbers, then arrange them in their usual order; if, however,  $n = 0$  and  $m \neq 0$ , let  $m$  precede  $n$ .) The result is a well ordering of  $\omega$ ; the ordinal number of this well ordering is  $\omega + 1$ .

If  $X$  and  $Y$  are well ordered sets, then a necessary and sufficient condition that  $\text{ord } X < \text{ord } Y$  is that  $X$  be similar to an initial segment of  $Y$ . It follows that we could compare the ordinal sizes of two well ordered sets even without knowing anything about ordinal numbers; all we would need to know is the concept of similarity. Similarity was defined for ordered sets; the central concept for arbitrary unordered sets is that of equivalence. (Recall that two sets  $X$  and  $Y$  are called equivalent,  $X \sim Y$ , in case there exists a one-to-one correspondence between them.) If we replace similarity by equivalence, then something like the suggestion of the preceding paragraph becomes usable. The point is that we do not have to know what size is if all we want is to compare sizes.

If  $X$  and  $Y$  are sets such that  $X$  is equivalent to a subset of  $Y$ , we shall write

$$X \precsim Y.$$

The notation is temporary and does not deserve a permanent name. As long as it lasts, however, it is convenient to have a way of referring to it; a reasonable possibility is to say that  $Y$  dominates  $X$ . The set of those ordered pairs  $(X, Y)$  of subsets of some set  $E$  for which  $X \precsim Y$  constitutes a relation in the power set of  $E$ . The symbolism correctly suggests some of the properties of the concept that it denotes. Since the symbolism is reminiscent of partial orders, and since a partial order is reflexive, antisymmetric, and transitive, we may expect that domination has similar properties.

Reflexivity and transitivity cause no trouble. Since each set  $X$  is equivalent to a subset (namely,  $X$ ) of itself, it follows that  $X \precsim X$  for all  $X$ . If  $f$  is a one-to-one correspondence between  $X$  and a subset of  $Y$ , and if  $g$  is a one-to-one correspondence between  $Y$  and a subset of  $Z$ , then we may restrict  $g$  to the range of  $f$  and compound the result with  $f$ ; the

conclusion is that  $X$  is equivalent to a subset of  $Z$ . In other words, if  $X \lesssim Y$  and  $Y \lesssim Z$ , then  $X \lesssim Z$ .

The interesting question is that of antisymmetry. If  $X \lesssim Y$  and  $Y \lesssim X$ , can we conclude that  $X = Y$ ? This is absurd; the assumptions are satisfied whenever  $X$  and  $Y$  are equivalent, and equivalent sets need not be identical. What then can we say about two sets if all we know is that each of them is equivalent to a subset of the other? The answer is contained in the following celebrated and important result.

**Schröder-Bernstein theorem.** *If  $X \lesssim Y$  and  $Y \lesssim X$ , then  $X \sim Y$ .*

**REMARK.** Observe that the converse, which is incidentally a considerable strengthening of the assertion of reflexivity, follows trivially from the definition of domination.

**PROOF.** Let  $f$  be a one-to-one mapping from  $X$  into  $Y$  and let  $g$  be a one-to-one mapping from  $Y$  into  $X$ ; the problem is to construct a one-to-one correspondence between  $X$  and  $Y$ . It is convenient to assume that the sets  $X$  and  $Y$  have no elements in common; if that is not true, we can so easily make it true that the added assumption involves no loss of generality.

We shall say that an element  $x$  in  $X$  is the *parent* of the element  $f(x)$  in  $Y$ , and, similarly, that an element  $y$  in  $Y$  is the parent of  $g(y)$  in  $X$ . Each element  $x$  of  $X$  has an infinite sequence of *descendants*, namely,  $f(x), g(f(x)), f(g(f(x))),$  etc., and similarly, the descendants of an element  $y$  of  $Y$  are  $g(y), f(g(y)), g(f(g(y))),$  etc. This definition implies that each term in the sequence is a descendant of all preceding terms; we shall also say that each term in the sequence is an *ancestor* of all following terms.

For each element (in either  $X$  or  $Y$ ) one of three things must happen. If we keep tracing the ancestry of the element back as far as possible, then either we ultimately come to an element of  $X$  that has no parent (these orphans are exactly the elements of  $X - g(Y)$ ), or we ultimately come to an element of  $Y$  that has no parent ( $Y - f(X)$ ), or the lineage regresses ad infinitum. Let  $X_X$  be the set of those elements of  $X$  that originate in  $X$  (i.e.,  $X_X$  consists of the elements of  $X - g(Y)$  together with all their descendants in  $X$ ), let  $X_Y$  be the set of those elements of  $X$  that originate in  $Y$  (i.e.,  $X_Y$  consists of all the descendants in  $X$  of the elements of  $Y - f(X)$ ), and let  $X_\infty$  be the set of those elements of  $X$  that have no parentless ancestor. Partition  $Y$  similarly into the three sets  $Y_X, Y_Y$ , and  $Y_\infty$ .

If  $x \in X_X$ , then  $f(x) \in Y_X$ , and, in fact, the restriction of  $f$  to  $X_X$  is a one-to-one correspondence between  $X_X$  and  $Y_X$ . If  $x \in X_Y$ , then  $x$  belongs to the domain of the inverse function  $g^{-1}$  and  $g^{-1}(x) \in Y_Y$ ; in fact the re-

striction of  $g^{-1}$  to  $X_Y$  is a one-to-one correspondence between  $X_Y$  and  $Y_Y$ . If, finally,  $x \in X_\infty$ , then  $f(x) \in Y_\infty$ , and the restriction of  $f$  to  $X_\infty$  is a one-to-one correspondence between  $X_\infty$  and  $Y_\infty$ ; alternatively, if  $x \in X_\infty$ , then  $g^{-1}(x) \in Y_\infty$ , and the restriction of  $g^{-1}$  to  $X_\infty$  is a one-to-one correspondence between  $X_\infty$  and  $Y_\infty$ . By combining these three one-to-one correspondences, we obtain a one-to-one correspondence between  $X$  and  $Y$ .

**EXERCISE.** Suppose that  $f$  is a mapping from  $X$  into  $Y$  and  $g$  is a mapping from  $Y$  into  $X$ . Prove that there exist subsets  $A$  and  $B$  of  $X$  and  $Y$  respectively, such that  $f(A) = B$  and  $g(Y - B) = X - A$ . This result can be used to give a proof of the Schröder-Bernstein theorem that looks quite different from the one above.

By now we know that domination has the essential properties of a partial order; we conclude this introductory discussion by observing that the order is in fact total. The assertion is known as the comparability theorem for sets: it says that if  $X$  and  $Y$  are sets, then either  $X \lesssim Y$  or  $Y \lesssim X$ . The proof is an immediate consequence of the well ordering theorem and of the comparability theorem for well ordered sets. Well order both  $X$  and  $Y$  and use the fact that either the well ordered sets so obtained are similar or one of them is similar to an initial segment of the other; in the former case  $X$  and  $Y$  are equivalent, and in the latter one of them is equivalent to a subset of the other.

## SECTION 23

---

### COUNTABLE SETS

---

If  $X$  and  $Y$  are sets such that  $Y$  dominates  $X$  and  $X$  dominates  $Y$ , then the Schröder-Bernstein theorem applies and says that  $X$  is equivalent to  $Y$ . If  $Y$  dominates  $X$  but  $X$  does not dominate  $Y$ , so that  $X$  is not equivalent to  $Y$ , we shall write

$$X < Y,$$

and we shall say that  $Y$  *strictly dominates*  $X$ .

Domination and strict domination can be used to express some of the facts about finite and infinite sets in a neat form. Recall that a set  $X$  is called finite in case it is equivalent to some natural number; otherwise it is infinite. We know that if  $X \lesssim Y$  and  $Y$  is finite, then  $X$  is finite, and we know that  $\omega$  is infinite (§ 13); we know also that if  $X$  is infinite, then  $\omega \lesssim X$  (§ 15). The converse of the last assertion is true and can be proved either directly (using the fact that a finite set cannot be equivalent to a proper subset of itself) or as an application of the Schröder-Bernstein theorem. (If  $\omega \lesssim X$ , then it is impossible that there exist a natural number  $n$  such that  $X \sim n$ , for then we should have  $\omega \lesssim n$ , and that contradicts the fact that  $\omega$  is infinite.)

We have just seen that a set  $X$  is infinite if and only if  $\omega \lesssim X$ ; next we shall prove that  $X$  is finite if and only if  $X < \omega$ . The proof depends on the transitivity of strict domination: if  $X \lesssim Y$  and  $Y \lesssim Z$ , and if at least one of these dominations is strict, then  $X < Z$ . Indeed, clearly,  $X \lesssim Z$ . If we had  $Z \lesssim X$ , then we should have  $Y \lesssim X$  and  $Z \lesssim Y$  and hence (by the Schröder-Bernstein theorem)  $X \sim Y$  and  $Y \sim Z$ , in contradiction to the assumption of strict domination. If now  $X$  is finite, then  $X \sim n$  for some natural number  $n$ , and, since  $\omega$  is infinite,  $n < \omega$ , so that  $X < \omega$ .

If, conversely,  $X \prec \omega$ , then  $X$  must be finite, for otherwise we should have  $\omega \lesssim X$ , and hence  $\omega \prec \omega$ , which is absurd.

A set  $X$  is called *countable* (or *denumerable*) in case  $X \lesssim \omega$  and *countably infinite* in case  $X \sim \omega$ . Clearly a countable set is either finite or countably infinite. Our main purpose in the immediate sequel is to show that many set-theoretic constructions when performed on countable sets lead again to countable sets.

We begin with the observation that every subset of  $\omega$  is countable, and we go on to deduce that every subset of each countable set is countable. These facts are trivial but useful.

If  $f$  is a function from  $\omega$  onto a set  $X$ , then  $X$  is countable. For the proof, observe that for each  $x$  in  $X$  the set  $f^{-1}(\{x\})$  is not empty (this is where the *onto* character of  $f$  is important), and consequently, for each  $x$  in  $X$ , we may find a natural number  $g(x)$  such that  $f(g(x)) = x$ . Since the function  $g$  is a one-to-one mapping from  $X$  into  $\omega$ , this proves that  $X \lesssim \omega$ . The reader who worries about such things might have noticed that this proof made use of the axiom of choice, and he may want to know that there is an alternative proof that does not depend on that axiom. (There is.) The same comment applies on a few other occasions in this section and its successors but we shall refrain from making it.

It follows from the preceding paragraph that a set  $X$  is countable if and only if there exists a function from some countable set onto  $X$ . A closely related result is this: if  $Y$  is any particular countably infinite set, then a necessary and sufficient condition that a non-empty set  $X$  be countable is that there exist a function from  $Y$  onto  $X$ .

The mapping  $n \rightarrow 2n$  is a one-to-one correspondence between  $\omega$  and the set  $A$  of all even numbers, so that  $A$  is countably infinite. This implies that if  $X$  is a countable set, then there exists a function  $f$  that maps  $A$  onto  $X$ . Since, similarly, the mapping  $n \rightarrow 2n + 1$  is a one-to-one correspondence between  $\omega$  and the set  $B$  of all odd numbers, it follows that if  $Y$  is a countable set, then there exists a function  $g$  that maps  $B$  onto  $Y$ . The function  $h$  that agrees with  $f$  on  $A$  and with  $g$  on  $B$  (i.e.,  $h(x) = f(x)$  when  $x \in A$  and  $h(x) = g(x)$  when  $x \in B$ ) maps  $\omega$  onto  $X \cup Y$ . Conclusion: the union of two countable sets is countable. From here on an easy argument by mathematical induction proves that the union of a finite set of countable sets is countable. The same result can be obtained by imitating the trick that worked for two sets; the basis of the method is the fact that for each non-zero natural number  $n$  there exists a pairwise disjoint family  $\{A_i\}$  ( $i < n$ ) of infinite subsets of  $\omega$  whose union is equal to  $\omega$ .

The same method can be used to prove still more. Assertion: there

exists a pairwise disjoint family  $\{A_n\}$  ( $n \in \omega$ ) of infinite subsets of  $\omega$  whose union is equal to  $\omega$ . One way to prove this is to write down the elements of  $\omega$  in an infinite array by counting down the diagonals, thus:

0	1	3	6	10	15	...
2	4	7	11	16	...	
5	8	12	17	...		
9	13	18	...			
14	19	...				
20	...					
...						

and then to consider the sequence of the rows of this array. Another way is to let  $A_0$  consist of 0 and the odd numbers, let  $A_1$  be the set obtained by doubling each non-zero element of  $A_0$ , and, inductively, let  $A_{n+1}$  be the set obtained by doubling each element of  $A_n$ ,  $n \geq 1$ . Either way (and there are many others still) the details are easy to fill in. Conclusion: the union of a countably infinite family of countable sets is countable. Proof: given the family  $\{X_n\}$  ( $n \in \omega$ ) of countable sets, find a family  $\{f_n\}$  of functions such that, for each  $n$ , the function  $f_n$  maps  $A_n$  onto  $X_n$ , and define a function  $f$  from  $\omega$  onto  $\bigcup_n X_n$  by writing  $f(k) = f_n(k)$  whenever  $k \in A_n$ . This result combined with the result of the preceding paragraph implies that the union of a countable set of countable sets is always countable.

An interesting and useful corollary is that the Cartesian product of two countable sets is also countable. Since

$$X \times Y = \bigcup_{y \in Y} (X \times \{y\}),$$

and since, if  $X$  is countable, then, for each fixed  $y$  in  $Y$ , the set  $X \times \{y\}$  is obviously countable (use the one-to-one correspondence  $x \rightarrow (x, y)$ ), the result follows from the preceding paragraph.

**EXERCISE.** Prove that the set of all finite subsets of a countable set is countable. Prove that if every countable subset of a totally ordered set  $X$  is well ordered, then  $X$  itself is well ordered.

On the basis of the preceding discussion it would not be unreasonable to guess that every set is countable. We proceed to show that that is not so; this negative result is what makes the theory of cardinal numbers interesting.

**Cantor's theorem.** *Every set is strictly dominated by its power set, or, in other words,*

$$X \prec \wp(X)$$

*for all  $X$ .*

**PROOF.** There is a natural one-to-one mapping from  $X$  into  $\wp(X)$ , namely, the mapping that associates with each element  $x$  of  $X$  the singleton  $\{x\}$ . The existence of this mapping proves that  $X \lesssim \wp(X)$ ; it remains to prove that  $X$  is not equivalent to  $\wp(X)$ .

Assume that  $f$  is a one-to-one mapping from  $X$  onto  $\wp(X)$ ; our purpose is to show that this assumption leads to a contradiction. Write  $A = \{x \in X : x \notin f(x)\}$ ; in words,  $A$  consists of those elements of  $X$  that are not contained in the corresponding set. Since  $A \in \wp(X)$  and since  $f$  maps  $X$  onto  $\wp(X)$ , there exists an element  $a$  in  $X$  such that  $f(a) = A$ . The element  $a$  either belongs to the set  $A$  or it does not. If  $a \in A$ , then, by the definition of  $A$ , we must have  $a \notin f(a)$ , and since  $f(a) = A$  this is impossible. If  $a \notin A$ , then, again by the definition of  $A$ , we must have  $a \in f(a)$ , and this too is impossible. The contradiction has arrived and the proof of Cantor's theorem is complete.

Since  $\wp(X)$  is always equivalent to  $2^X$  (where  $2^X$  is the set of all functions from  $X$  into 2), Cantor's theorem implies that  $X \prec 2^X$  for all  $X$ . If in particular we take  $\omega$  in the role of  $X$ , then we may conclude that the set of all sets of natural numbers is *uncountable* (i.e., not countable, non-denumerable), or, equivalently, that  $2^\omega$  is uncountable. Here  $2^\omega$  is the set of all infinite sequences of 0's and 1's (i.e., functions from  $\omega$  into 2). Note that if we interpret  $2^\omega$  in the sense of ordinal exponentiation, then  $2^\omega$  is countable (in fact  $2^\omega = \omega$ ).

## SECTION 24

---

### CARDINAL ARITHMETIC

---

One result of our study of the comparative sizes of sets will be to define a new concept, called *cardinal number*, and to associate with each set  $X$  a cardinal number, denoted by  $\text{card } X$ . The definitions are such that for each cardinal number  $a$  there exist sets  $A$  with  $\text{card } A = a$ . We shall also define an ordering for cardinal numbers, denoted as usual by  $\leq$ . The connection between these new concepts and the ones already at our disposal is easy to describe: it will turn out that  $\text{card } X = \text{card } Y$  if and only if  $X \sim Y$ , and  $\text{card } X < \text{card } Y$  if and only if  $X < Y$ . (If  $a$  and  $b$  are cardinal numbers,  $a < b$  means, of course, that  $a \leq b$  but  $a \neq b$ .)

The definition of cardinal numbers can be approached in several different ways, each of which has its strong advocates. To keep the peace as long as possible, and to demonstrate that the essential properties of the concept are independent of the approach, we shall postpone the basic construction. We proceed, instead, to study the arithmetic of cardinal numbers. In the course of that study we shall make use of the connection, described above, between cardinal inequality and set domination; that much of a loan from the future will be enough for the purpose.

If  $a$  and  $b$  are cardinal numbers, and if  $A$  and  $B$  are disjoint sets with  $\text{card } A = a$  and  $\text{card } B = b$ , we write, by definition,  $a + b = \text{card } (A \cup B)$ . If  $C$  and  $D$  are disjoint sets with  $\text{card } C = a$  and  $\text{card } D = b$ , then  $A \sim C$  and  $B \sim D$ ; it follows that  $A \cup B \sim C \cup D$ , and hence that  $a + b$  is unambiguously defined, independently of the arbitrary choice of  $A$  and  $B$ . Cardinal addition, thus defined, is commutative ( $a + b = b + a$ ), and associative ( $a + (b + c) = (a + b) + c$ ); these identities are immediate consequences of the corresponding facts about the formation of unions.

**EXERCISE.** Prove that if  $a, b, c$ , and  $d$  are cardinal numbers such that  $a \leq b$  and  $c \leq d$ , then  $a + c \leq b + d$ .

There is no difficulty about defining addition for infinitely many summands. If  $\{a_i\}$  is a family of cardinal numbers, and if  $\{A_i\}$  is a correspondingly indexed family of pairwise disjoint sets such that  $\text{card } A_i = a_i$  for each  $i$ , then we write, by definition,

$$\sum_i a_i = \text{card}(\bigcup_i A_i).$$

As before, the definition is unambiguous.

To define the product  $ab$  of two cardinal numbers  $a$  and  $b$ , we find sets  $A$  and  $B$  with  $\text{card } A = a$  and  $\text{card } B = b$ , and we write  $ab = \text{card}(A \times B)$ . The replacement of  $A$  and  $B$  by equivalent sets yields the same value of the product. Alternatively, we could have defined  $ab$  by “adding  $a$  to itself  $b$  times”; this refers to the formation of the infinite sum  $\sum_{i \in I} a_i$ , where the index set  $I$  has cardinal number  $b$ , and where  $a_i = a$  for each  $i$  in  $I$ . The reader should have no difficulty in verifying that this proposed alternative definition is indeed equivalent to the one that uses Cartesian products. Cardinal multiplication is commutative ( $ab = ba$ ) and associative ( $(a(bc)) = ((ab)c)$ ), and multiplication distributes over addition ( $a(b+c) = ab+ac$ ); the proofs are elementary.

**EXERCISE.** Prove that if  $a, b, c$ , and  $d$  are cardinal numbers such that  $a \leq b$  and  $c \leq d$ , then  $ac \leq bd$ .

There is no difficulty about defining multiplication for infinitely many factors. If  $\{a_i\}$  is a family of cardinal numbers, and if  $\{A_i\}$  is a correspondingly indexed family of sets such that  $\text{card } A_i = a_i$  for each  $i$ , then we write, by definition,

$$\prod_i a_i = \text{card}(\bigtimes_i A_i).$$

The definition is unambiguous.

**EXERCISE.** If  $\{a_i\}$  ( $i \in I$ ) and  $\{b_i\}$  ( $i \in I$ ) are families of cardinal numbers such that  $a_i < b_i$  for each  $i$  in  $I$ , then  $\sum_i a_i < \prod_i b_i$ .

We can go from products to exponents the same way as we went from sums to products. The definition of  $a^b$ , for cardinal numbers  $a$  and  $b$ , is most profitably given directly, but an alternative approach goes via repeated multiplication. For the direct definition, find sets  $A$  and  $B$  with  $\text{card } A = a$  and  $\text{card } B = b$ , and write  $a^b = \text{card } A^B$ . Alternatively, to define  $a^b$  “multiply  $a$  by itself  $b$  times.” More precisely: form  $\prod_{i \in I} a_i$ , where the index set  $I$  has cardinal number  $b$ , and where  $a_i = a$  for each  $i$

in I. The familiar laws of exponents hold. That is, if  $a$ ,  $b$ , and  $c$  are cardinal numbers, then

$$a^{b+c} = a^b a^c,$$

$$(ab)^c = a^c b^c,$$

$$a^{bc} = (a^b)^c.$$

**EXERCISE.** Prove that if  $a$ ,  $b$ , and  $c$  are cardinal numbers such that  $a \leq b$ , then  $a^c \leq b^c$ . Prove that if  $a$  and  $b$  are finite, greater than 1, and if  $c$  is infinite, then  $a^c = b^c$ .

The preceding definitions and their consequences are reasonably straightforward and not at all surprising. If they are restricted to finite sets only, the result is the familiar finite arithmetic. The novelty of the subject arises in the formation of sums, products, and powers in which at least one term is infinite. The words “finite” and “infinite” are used here in a very natural sense: a cardinal number is *finite* if it is the cardinal number of a finite set, and *infinite* otherwise.

If  $a$  and  $b$  are cardinal numbers such that  $a$  is finite and  $b$  is infinite, then

$$a + b = b.$$

For the proof, suppose that  $A$  and  $B$  are disjoint sets such that  $A$  is equivalent to some natural number  $k$  and  $B$  is infinite; we are to prove that  $A \cup B \sim B$ . Since  $\omega \lesssim B$ , we may and do assume that  $\omega \subset B$ . We define a mapping  $f$  from  $A \cup B$  to  $B$  as follows: the restriction of  $f$  to  $A$  is a one-to-one correspondence between  $A$  and  $k$ , the restriction of  $f$  to  $\omega$  is given by  $f(n) = n + k$  for all  $n$ , and the restriction of  $f$  to  $B - \omega$  is the identity mapping on  $B - \omega$ . Since the result is a one-to-one correspondence between  $A \cup B$  and  $B$ , the proof is complete.

Next: if  $a$  is an infinite cardinal number, then

$$a + a = a.$$

For the proof, let  $A$  be a set with  $\text{card } A = a$ . Since the set  $A \times 2$  is the union of two disjoint sets equivalent to  $A$  (namely,  $A \times \{0\}$  and  $A \times \{1\}$ ), it would be sufficient to prove that  $A \times 2$  is equivalent to  $A$ . The approach we shall use will not quite prove that much, but it will come close enough. The idea is to approximate the construction of the desired one-to-one correspondence by using larger and larger subsets of  $A$ .

Precisely speaking, let  $\mathfrak{F}$  be the collection of all functions  $f$  such that the domain of  $f$  is of the form  $X \times 2$ , for some subset  $X$  of  $A$ , and such that  $f$  is a one-to-one correspondence between  $X \times 2$  and  $X$ . If  $X$  is a count-

ably infinite subset of  $A$ , then  $X \times 2 \sim X$ . This implies that the collection  $\mathfrak{F}$  is not empty; at the very least it contains the one-to-one correspondences between  $X \times 2$  and  $X$  for the countably infinite subsets  $X$  of  $A$ . The collection  $\mathfrak{F}$  is partially ordered by extension. Since a straightforward verification shows that the hypotheses of Zorn's lemma are satisfied, it follows that  $\mathfrak{F}$  contains a maximal element  $f$  with  $\text{ran } f = X$ , say.

Assertion:  $A - X$  is finite. If  $A - X$  were infinite, then it would include a countably infinite set, say  $Y$ . By combining  $f$  with a one-to-one correspondence between  $Y \times 2$  and  $Y$  we could obtain a proper extension of  $f$ , in contradiction to the assumed maximality.

Since  $\text{card } X + \text{card } X = \text{card } X$ , and since  $\text{card } A = \text{card } X + \text{card } (A - X)$ , the fact that  $A - X$  is finite completes the proof that  $\text{card } A + \text{card } A = \text{card } A$ .

Here is one more result in additive cardinal arithmetic: if  $a$  and  $b$  are cardinal numbers at least one of which is infinite, and if  $c$  is equal to the larger one of  $a$  and  $b$ , then

$$a + b = c.$$

Suppose that  $b$  is infinite, and let  $A$  and  $B$  be disjoint sets with  $\text{card } A = a$  and  $\text{card } B = b$ . Since  $a \leq c$  and  $b \leq c$ , it follows that  $a + b \leq c + c$ , and since  $c \leq \text{card } (A \cup B)$ , it follows that  $c \leq a + b$ . The result follows from the antisymmetry of the ordering of cardinal numbers.

The principal result in multiplicative cardinal arithmetic is that if  $a$  is an infinite cardinal number, then

$$a \cdot a = a.$$

The proof resembles the proof of the corresponding additive fact. Let  $\mathfrak{F}$  be the collection of all functions  $f$  such that the domain of  $f$  is of the form  $X \times X$  for some subset  $X$  of  $A$ , and such that  $f$  is a one-to-one correspondence between  $X \times X$  and  $X$ . If  $X$  is a countably infinite subset of  $A$ , then  $X \times X \sim X$ . This implies that the collection  $\mathfrak{F}$  is not empty; at the very least it contains the one-to-one correspondences between  $X \times X$  and  $X$  for the countably infinite subsets  $X$  of  $A$ . The collection  $\mathfrak{F}$  is partially ordered by extension. The hypotheses of Zorn's lemma are easily verified, and it follows that  $\mathfrak{F}$  contains a maximal element  $f$  with  $\text{ran } f = X$ , say. Since  $(\text{card } X)(\text{card } X) = \text{card } X$ , the proof may be completed by showing that  $\text{card } X = \text{card } A$ .

Assume that  $\text{card } X < \text{card } A$ . Since  $\text{card } A$  is equal to the larger one of  $\text{card } X$  and  $\text{card } (A - X)$ , this implies that  $\text{card } A = \text{card } (A - X)$ , and hence that  $\text{card } X < \text{card } (A - X)$ . From this it follows that  $A - X$

has a subset  $Y$  equivalent to  $X$ . Since each of the disjoint sets  $X \times Y$ ,  $Y \times X$ , and  $Y \times Y$  is infinite and equivalent to  $X \times X$ , hence to  $X$ , and hence to  $Y$ , it follows that their union is equivalent to  $Y$ . By combining  $f$  with a one-to-one correspondence between that union and  $Y$ , we obtain a proper extension of  $f$ , in contradiction to the assumed maximality. This implies that our present hypothesis ( $\text{card } X < \text{card } A$ ) is untenable and hence completes the proof.

**EXERCISE.** Prove that if  $a$  and  $b$  are cardinal numbers at least one of which is infinite, then  $a + b = ab$ . Prove that if  $a$  and  $b$  are cardinal numbers such that  $a$  is infinite and  $b$  is finite, then  $a^b = a$ .

## SECTION 25

---

### CARDINAL NUMBERS

---

We know quite a bit about cardinal numbers by now, but we still do not know what they are. Speaking vaguely, we may say that the cardinal number of a set is the property that the set has in common with all sets equivalent to it. We may try to make this precise by saying that the cardinal number of  $X$  is equal to the set of all sets equivalent to  $X$ , but the attempt will fail; there is no set as large as that. The next thing to try, suggested by analogy with our approach to the definition of natural numbers, is to define the cardinal number of a set  $X$  as some particular carefully selected set equivalent to  $X$ . This is what we proceed to do.

For each set  $X$  there are too many other sets equivalent to  $X$ ; our first problem is to narrow the field. Since we know that every set is equivalent to some ordinal number, it is not unnatural to look for the typical sets, the representative sets, among ordinal numbers.

To be sure, a set can be equivalent to many ordinal numbers. A hopeful sign, however, is the fact that, for each set  $X$ , the ordinal numbers equivalent to  $X$  constitute a set. To prove this, observe first that it is easy to produce an ordinal number that is surely greater, strictly greater, than all the ordinal numbers equivalent to  $X$ . Suppose in fact that  $\gamma$  is an ordinal number equivalent to the power set  $\wp(X)$ . If  $\alpha$  is an ordinal number equivalent to  $X$ , then the set  $\alpha$  is strictly dominated by the set  $\gamma$  (i.e.,  $\text{card } \alpha < \text{card } \gamma$ ). It follows that we cannot have  $\gamma \leq \alpha$ , and, consequently, we must have  $\alpha < \gamma$ . Since, for ordinal numbers,  $\alpha < \gamma$  means the same thing as  $\alpha \in \gamma$ , we have found a set, namely  $\gamma$ , that contains every ordinal number equivalent to  $X$ , and this implies that the ordinal numbers equivalent to  $X$  do constitute a set.

Which one among the ordinal numbers equivalent to  $X$  deserves to be singled out and called the cardinal number of  $X$ ? The question has only one natural answer. Every set of ordinal numbers is well ordered; the

least element of a well ordered set is the only one that seems to clamor for special attention.

We are now prepared for the definition: a *cardinal number* is an ordinal number  $\alpha$  such that if  $\beta$  is an ordinal number equivalent to  $\alpha$  (i.e.,  $\text{card } \alpha = \text{card } \beta$ ), then  $\alpha \leq \beta$ . The ordinal numbers with this property have also been called *initial numbers*. If  $X$  is a set, then  $\text{card } X$ , the cardinal number of  $X$  (also known as the *power* of  $X$ ), is the least ordinal number equivalent to  $X$ .

**EXERCISE.** Prove that each infinite cardinal number is a limit number.

Since each set is equivalent to its cardinal number, it follows that if  $\text{card } X = \text{card } Y$ , then  $X \sim Y$ . If, conversely,  $X \sim Y$ , then  $\text{card } X \sim \text{card } Y$ . Since  $\text{card } X$  is the least ordinal number equivalent to  $X$ , it follows that  $\text{card } X \leq \text{card } Y$ , and, since the situation is symmetric in  $X$  and  $Y$ , we also have  $\text{card } Y \leq \text{card } X$ . In other words  $\text{card } X = \text{card } Y$  if and only if  $X \sim Y$ ; this was one of the conditions on cardinal numbers that we needed in the development of cardinal arithmetic.

A finite ordinal number (i.e., a natural number) is not equivalent to any finite ordinal number distinct from itself. It follows that if  $X$  is finite, then the set of ordinal numbers equivalent to  $X$  is a singleton, and, consequently, the cardinal number of  $X$  is the same as the ordinal number of  $X$ . Both cardinal numbers and ordinal numbers are generalizations of the natural numbers; in the familiar finite cases both the generalizations coincide with the special case that gave rise to them in the first place. As an almost trivial application of these remarks, we can now calculate the cardinal number of a power set  $\mathcal{P}(A)$ : if  $\text{card } A = a$ , then  $\text{card } \mathcal{P}(A) = 2^a$ . (Note that the result, though simple, could not have been stated before this; till now we did not know that 2 is a cardinal number.) The proof is immediate from the fact that  $\mathcal{P}(A)$  is equivalent to  $2^a$ .

If  $\alpha$  and  $\beta$  are ordinal numbers, we know what it means to say that  $\alpha < \beta$  or  $\alpha \leq \beta$ . It follows that cardinal numbers come to us automatically equipped with an order. The order satisfies the conditions we borrowed for our discussion of cardinal arithmetic. Indeed: if  $\text{card } X < \text{card } Y$ , then  $\text{card } X$  is a subset of  $\text{card } Y$ , and it follows that  $X \lesssim Y$ . If we had  $X \sim Y$ , then, as we have already seen, we should have  $\text{card } X = \text{card } Y$ ; it follows that we must have  $X < Y$ . If, finally,  $X < Y$ , then it is impossible that  $\text{card } Y \leq \text{card } X$  (for similarity implies equivalence), and hence  $\text{card } X < \text{card } Y$ .

As an application of these considerations we mention the inequality

$$\alpha < 2^\alpha,$$

valid for all cardinal numbers  $a$ . Proof: if  $A$  is a set with  $\text{card } A = a$ , then  $A \prec \wp(A)$ , hence  $\text{card } A < \text{card } \wp(A)$ , and therefore  $a < 2^a$ .

EXERCISE. If  $\text{card } A = a$ , what is the cardinal number of the set of all one-to-one mappings of  $A$  onto itself? What is the cardinal number of the set of all countably infinite subsets of  $A$ ?

The facts about the ordering of ordinal numbers are at the same time facts about the ordering of cardinal numbers. Thus, for instance, we know that any two cardinal numbers are comparable (always either  $a < b$ , or  $a = b$ , or  $b < a$ ), and that, in fact, every set of cardinal numbers is well ordered. We know also that every set of cardinal numbers has an upper bound (in fact, a supremum), and that, moreover, for every set of cardinal numbers, there is a cardinal number strictly greater than any of them. This implies of course that there is no largest cardinal number, or, equivalently, that there is no set that consists exactly of all the cardinal numbers. The contradiction, based on the assumption that there is such a set, is known as *Cantor's paradox*.

The fact that cardinal numbers are special ordinal numbers simplifies some aspects of the theory, but, at the same time, it introduces the possibility of some confusion that it is essential to avoid. One major source of difficulty is the notation for the arithmetic operations. If  $a$  and  $b$  are cardinal numbers, then they are also ordinal numbers, and, consequently, the sum  $a + b$  has two possible meanings. The cardinal sum of two cardinal numbers is in general not the same as their ordinal sum. All this sounds worse than it is; in practice it is easy to avoid confusion. The context, the use of special symbols for cardinal numbers, and an occasional explicit warning can make the discussion flow quite smoothly.

EXERCISE. Prove that if  $\alpha$  and  $\beta$  are ordinal numbers, then  $\text{card } (\alpha + \beta) = \text{card } \alpha + \text{card } \beta$  and  $\text{card } (\alpha\beta) = (\text{card } \alpha)(\text{card } \beta)$ . Use the ordinal interpretation of the operations on the left side and the cardinal interpretation on the right.

One of the special symbols for cardinal numbers that is used very frequently is the first letter ( $\aleph$ , aleph) of the Hebrew alphabet. Thus in particular the smallest transfinite ordinal number, i.e.,  $\omega$ , is a cardinal number, and, as such, it is always denoted by  $\aleph_0$ .

Every one of the ordinal numbers that we have explicitly named so far is countable. In many of the applications of set theory an important role is played by the smallest uncountable ordinal number, frequently denoted by  $\Omega$ . The most important property of  $\omega$  is that it is an infinite well or-

dered set each of whose initial segments is finite; correspondingly, the most important property of  $\Omega$  is that it is an uncountably infinite well ordered set each of whose initial segments is countable.

The least uncountable ordinal number  $\Omega$  clearly satisfies the defining condition of a cardinal number; in its cardinal role it is always denoted by  $\aleph_1$ . Equivalently,  $\aleph_1$  may be characterized as the least cardinal number strictly greater than  $\aleph_0$ , or, in other words, the immediate successor of  $\aleph_0$  in the ordering of cardinal numbers.

The arithmetic relation between  $\aleph_0$  and  $\aleph_1$  is the subject of a famous old problem about cardinal numbers. How do we get from  $\aleph_0$  to  $\aleph_1$  by arithmetic operations? We know by now that the most elementary steps, involving sums and products, just lead from  $\aleph_0$  back to  $\aleph_0$  again. The simplest thing we know to do that starts with  $\aleph_0$  and ends up with something larger is to form  $2^{\aleph_0}$ . We know therefore that  $\aleph_1 \leq 2^{\aleph_0}$ . Is the inequality strict? Is there an uncountable cardinal number strictly less than  $2^{\aleph_0}$ ? The celebrated *continuum hypothesis* asserts, as a guess, that the answer is no, or, in other words, that  $\aleph_1 = 2^{\aleph_0}$ . All that is known for sure is that the continuum hypothesis is consistent with the axioms of set theory.

For each infinite cardinal number  $a$ , consider the set  $c(a)$  of all infinite cardinal numbers that are strictly less than  $a$ . If  $a = \aleph_0$ , then  $c(a) = \emptyset$ ; if  $a = \aleph_1$ , then  $c(a) = \{\aleph_0\}$ . Since  $c(a)$  is a well ordered set, it has an ordinal number, say  $\alpha$ . The connection between  $a$  and  $\alpha$  is usually expressed by writing  $a = \aleph_\alpha$ . An equivalent definition of the cardinal numbers  $\aleph_\alpha$  proceeds by transfinite induction; according to that approach  $\aleph_\alpha$  (for  $\alpha > 0$ ) is the smallest cardinal number that is strictly greater than all the  $\aleph_\beta$ 's with  $\beta < \alpha$ . The *generalized continuum hypothesis* is the conjecture that  $\aleph_{\alpha+1} = 2^{\aleph_\alpha}$  for each ordinal number  $\alpha$ .

## INDEX

---

**all**, 5  
**ancestor**, 88  
**and**, 5  
**antisymmetric**, 3, 54  
**argument**, 30  
**associative**, 13  
**assume**, 30  
**atomic sentence**, 5  
**Aussonderungsaxiom**, 6  
**axiom of choice**, 59

**axiom of extension**, 2  
**axiom of infinity**, 44  
**axiom of pairing**, 9  
**axiom of powers**, 19  
**axiom of specification**, 6  
**axiom of substitution**, 75  
**axiom of unions**, 4  
  
**belonging**, 2  
**between**, 56

- binary, 26  
Boolean sum, 18  
Burali-Forti, 80
- canonical** map, 32  
Cantor, 93, 101  
cardinal number, 94, 100  
Cartesian product, 24  
chain, 54  
characteristic function, 33  
choice function, 60  
class, 1, 11  
cofinal, 68  
collection, 1  
commutative, 13  
comparability theorem, 73, 89  
comparable, 64  
complement, 17  
composite, 40  
condition, 6  
contain, 2  
continuation, 67  
continuum hypothesis, 102  
converse, 40  
coordinate, 23, 36  
correspondence, 30  
countable, 91  
counting theorem, 80
- Dedekind**, 61  
definition by induction, 49  
definition by transfinite induction, 71  
De Morgan, 17  
denumerable, 91  
descendant, 88  
difference, 17  
disjoint, 15  
distributive, 15  
domain, 27  
dominate, 87  
duality, 18
- element**, 1  
embedding, 31  
empty, 8  
equality, 2  
equivalence relation, 28  
equivalent, 52  
even, 72
- extension, 32  
**family**, 34  
finite, 45, 53  
first, 56  
first coordinate, 23  
from, 27, 30  
function, 30
- graph**, 30  
greater, 55  
greatest, 56  
greatest lower bound, 57
- idempotent**, 13  
identity map, 31  
if, 5  
image, 31  
imply, 5  
in, 27  
inclusion, 3  
inclusion map, 31  
index, 34  
induced relation, 28  
induction, 46  
infimum, 57  
infinite, 45, 53  
initial number, 100  
initial segment, 56  
injection, 31  
intersection, 14, 15  
into, 30  
inverse, 38, 40
- larger**, 55  
largest, 56  
last, 56  
least, 56  
least upper bound, 57  
less, 55  
lexicographical order, 58  
limit number, 79  
linear, 54  
logical operators, 5  
lower bound, 57
- mapping**, 30  
mathematical induction, 46  
maximal, 57

- member, 1  
 minimal, 56  
 modulo, 28
- natural** number, 44  
 non-empty family, 35  
 not, 5  
 number, 44  
 number of elements, 53
- on**, 30  
 one-to-one, 32  
 onto, 31  
 operator, 30  
 or, 5  
 order, 54  
 ordered pair, 23  
 ordered quadruple, 36  
 ordered triple, 36  
 order-preserving, 71  
 ordinal number, 75  
 ordinal product, 83  
 ordinal sum, 82
- pair**, 9  
 pairwise disjoint, 15  
 parent, 88  
 partial order, 54  
 partially ordered set, 55  
 partition, 28  
 Peano, 47  
 power, 51, 100  
 power set, 19  
 predecessor, 55  
 product, 50  
 projection, 24, 32, 36  
 proper, 3
- quadruple**, 14  
 quaternary, 26
- range**, 27  
 recursion, 48  
 reflexive, 3, 27  
 relation, 26  
 relative complement, 17  
 relative product, 41  
 restriction, 32  
 Russell, 7
- Schröder-Bernstein**, 88  
 second coordinate, 23  
 send, 30  
 sentence, 5  
 sequence, 45  
 sequence function, 70  
 set, 1  
 several variables, 37  
 similar, 71  
 simple, 54  
 singleton, 10  
 smaller, 55  
 smallest, 56  
 some, 5  
 strict, 56  
 subset, 3  
 successor, 43, 55  
 successor set, 44  
 sum, 50  
 supremum, 57  
 symmetric, 3, 27  
 symmetric difference, 18
- term**, 34  
 ternary, 26  
 to, 27, 30  
 total, 54  
 tower, 64  
 transfinite, 79  
 transfinite induction, 66  
 transfinite recursion, 70  
 transformation, 30  
 transitive, 3, 27  
 transitive set, 47  
 triple, 14
- union**, 12  
 universe, 7  
 unordered pair, 9  
 upper bound, 57
- value**, 30  
 variable, 37  
 von Neumann, 75
- weak**, 56  
 well ordering, 66
- Zorn**, 62

Undergraduate Texts in Mathematics

UTM

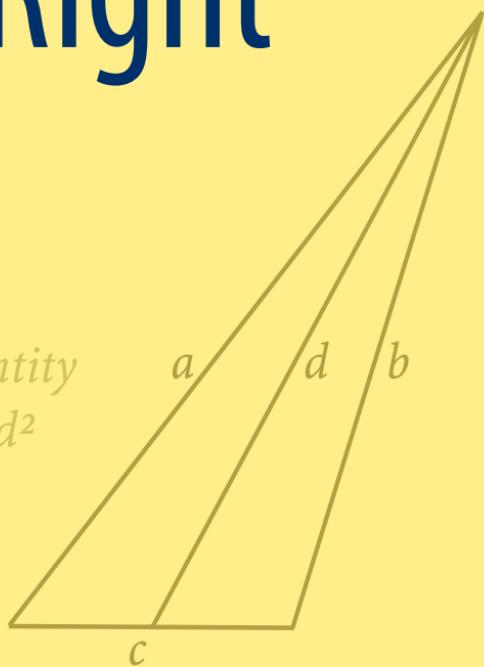
Sheldon Axler

# Linear Algebra Done Right

*Third Edition*

*Apollonius's Identity*

$$a^2 + b^2 = \frac{1}{2} c^2 + 2 d^2$$



Springer

# Undergraduate Texts in Mathematics

# Undergraduate Texts in Mathematics

---

## Series Editors:

Sheldon Axler

*San Francisco State University, San Francisco, CA, USA*

Kenneth Ribet

*University of California, Berkeley, CA, USA*

## Advisory Board:

Colin Adams, *Williams College, Williamstown, MA, USA*

Alejandro Adem, *University of British Columbia, Vancouver, BC, Canada*

Ruth Charney, *Brandeis University, Waltham, MA, USA*

Irene M. Gamba, *The University of Texas at Austin, Austin, TX, USA*

Roger E. Howe, *Yale University, New Haven, CT, USA*

David Jerison, *Massachusetts Institute of Technology, Cambridge, MA, USA*

Jeffrey C. Lagarias, *University of Michigan, Ann Arbor, MI, USA*

Jill Pipher, *Brown University, Providence, RI, USA*

Fadil Santosa, *University of Minnesota, Minneapolis, MN, USA*

Amie Wilkinson, *University of Chicago, Chicago, IL, USA*

**Undergraduate Texts in Mathematics** are generally aimed at third- and fourth-year undergraduate mathematics students at North American universities. These texts strive to provide students and teachers with new perspectives and novel approaches. The books include motivation that guides the reader to an appreciation of interrelations among different aspects of the subject. They feature examples that illustrate key concepts as well as exercises that strengthen understanding.

For further volumes:

<http://www.springer.com/series/666>

Sheldon Axler

# Linear Algebra Done Right

Third edition



Sheldon Axler  
Department of Mathematics  
San Francisco State University  
San Francisco, CA, USA

ISSN 0172-6056                   ISSN 2197-5604 (electronic)  
ISBN 978-3-319-11079-0       ISBN 978-3-319-11080-6 (eBook)  
DOI 10.1007/978-3-319-11080-6  
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014954079

Mathematics Subject Classification (2010): 15-01, 15A03, 15A04, 15A15, 15A18, 15A21

© Springer International Publishing 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Typeset by the author in LaTeX

*Cover figure:* For a statement of Apollonius's Identity and its connection to linear algebra, see the last exercise in Section 6.A.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# *Contents*

---

*Preface for the Instructor* xi

*Preface for the Student* xv

*Acknowledgments* xvii

## **1 Vector Spaces 1**

1.A  $\mathbf{R}^n$  and  $\mathbf{C}^n$  2

Complex Numbers 2

Lists 5

$\mathbf{F}^n$  6

Digression on Fields 10

Exercises 1.A 11

1.B Definition of Vector Space 12

Exercises 1.B 17

1.C Subspaces 18

Sums of Subspaces 20

Direct Sums 21

Exercises 1.C 24

## **2 Finite-Dimensional Vector Spaces 27**

2.A Span and Linear Independence 28

Linear Combinations and Span 28

Linear Independence 32

Exercises 2.A 37

2.B Bases **39**

Exercises 2.B **43**

2.C Dimension **44**

Exercises 2.C **48**

**3 Linear Maps 51**

3.A The Vector Space of Linear Maps **52**

Definition and Examples of Linear Maps **52**

Algebraic Operations on  $\mathcal{L}(V, W)$  **55**

Exercises 3.A **57**

3.B Null Spaces and Ranges **59**

Null Space and Injectivity **59**

Range and Surjectivity **61**

Fundamental Theorem of Linear Maps **63**

Exercises 3.B **67**

3.C Matrices **70**

Representing a Linear Map by a Matrix **70**

Addition and Scalar Multiplication of Matrices **72**

Matrix Multiplication **74**

Exercises 3.C **78**

3.D Invertibility and Isomorphic Vector Spaces **80**

Invertible Linear Maps **80**

Isomorphic Vector Spaces **82**

Linear Maps Thought of as Matrix Multiplication **84**

Operators **86**

Exercises 3.D **88**

3.E Products and Quotients of Vector Spaces **91**

Products of Vector Spaces **91**

Products and Direct Sums **93**

Quotients of Vector Spaces **94**

Exercises 3.E **98**

**3.F Duality 101**

- The Dual Space and the Dual Map **101**
- The Null Space and Range of the Dual of a Linear Map **104**
- The Matrix of the Dual of a Linear Map **109**
- The Rank of a Matrix **111**
- Exercises 3.F **113**

**4 Polynomials 117**

- Complex Conjugate and Absolute Value **118**
- Uniqueness of Coefficients for Polynomials **120**
- The Division Algorithm for Polynomials **121**
- Zeros of Polynomials **122**
- Factorization of Polynomials over **C** **123**
- Factorization of Polynomials over **R** **126**
- Exercises 4 **129**

**5 Eigenvalues, Eigenvectors, and Invariant Subspaces 131**

- 5.A Invariant Subspaces **132**
  - Eigenvalues and Eigenvectors **133**
  - Restriction and Quotient Operators **137**
  - Exercises 5.A **138**
- 5.B Eigenvectors and Upper-Triangular Matrices **143**
  - Polynomials Applied to Operators **143**
  - Existence of Eigenvalues **145**
  - Upper-Triangular Matrices **146**
  - Exercises 5.B **153**
- 5.C Eigenspaces and Diagonal Matrices **155**
  - Exercises 5.C **160**

**6 Inner Product Spaces 163**

- 6.A Inner Products and Norms **164**
  - Inner Products **164**
  - Norms **168**
  - Exercises 6.A **175**

**6.B Orthonormal Bases 180**

- Linear Functionals on Inner Product Spaces 187  
Exercises 6.B 189

**6.C Orthogonal Complements and Minimization Problems 193**

- Orthogonal Complements 193  
Minimization Problems 198  
Exercises 6.C 201

**7 Operators on Inner Product Spaces 203****7.A Self-Adjoint and Normal Operators 204**

- Adjoints 204  
Self-Adjoint Operators 209  
Normal Operators 212  
Exercises 7.A 214

**7.B The Spectral Theorem 217**

- The Complex Spectral Theorem 217  
The Real Spectral Theorem 219  
Exercises 7.B 223

**7.C Positive Operators and Isometries 225**

- Positive Operators 225  
Isometries 228  
Exercises 7.C 231

**7.D Polar Decomposition and Singular Value Decomposition 233**

- Polar Decomposition 233  
Singular Value Decomposition 236  
Exercises 7.D 238

**8 Operators on Complex Vector Spaces 241****8.A Generalized Eigenvectors and Nilpotent Operators 242**

- Null Spaces of Powers of an Operator 242  
Generalized Eigenvectors 244  
Nilpotent Operators 248  
Exercises 8.A 249

**8.B Decomposition of an Operator 252**

- Description of Operators on Complex Vector Spaces 252
- Multiplicity of an Eigenvalue 254
- Block Diagonal Matrices 255
- Square Roots 258
- Exercises 8.B 259

**8.C Characteristic and Minimal Polynomials 261**

- The Cayley–Hamilton Theorem 261
- The Minimal Polynomial 262
- Exercises 8.C 267

**8.D Jordan Form 270**

- Exercises 8.D 274

**9 *Operators on Real Vector Spaces* 275****9.A Complexification 276**

- Complexification of a Vector Space 276
- Complexification of an Operator 277
- The Minimal Polynomial of the Complexification 279
- Eigenvalues of the Complexification 280
- Characteristic Polynomial of the Complexification 283
- Exercises 9.A 285

**9.B Operators on Real Inner Product Spaces 287**

- Normal Operators on Real Inner Product Spaces 287
- Isometries on Real Inner Product Spaces 292
- Exercises 9.B 294

**10 *Trace and Determinant* 295****10.A Trace 296**

- Change of Basis 296
- Trace: A Connection Between Operators and Matrices 299
- Exercises 10.A 304

**10.B Determinant 307**

Determinant of an Operator **307**

Determinant of a Matrix **309**

The Sign of the Determinant **320**

Volume **323**

Exercises 10.B **330**

***Photo Credits 333***

***Symbol Index 335***

***Index 337***

# *Preface for the Instructor*

---

You are about to teach a course that will probably give students their second exposure to linear algebra. During their first brush with the subject, your students probably worked with Euclidean spaces and matrices. In contrast, this course will emphasize abstract vector spaces and linear maps.

The audacious title of this book deserves an explanation. Almost all linear algebra books use determinants to prove that every linear operator on a finite-dimensional complex vector space has an eigenvalue. Determinants are difficult, nonintuitive, and often defined without motivation. To prove the theorem about existence of eigenvalues on complex vector spaces, most books must define determinants, prove that a linear map is not invertible if and only if its determinant equals 0, and then define the characteristic polynomial. This tortuous (torturous?) path gives students little feeling for why eigenvalues exist.

In contrast, the simple determinant-free proofs presented here (for example, see 5.21) offer more insight. Once determinants have been banished to the end of the book, a new route opens to the main goal of linear algebra—understanding the structure of linear operators.

This book starts at the beginning of the subject, with no prerequisites other than the usual demand for suitable mathematical maturity. Even if your students have already seen some of the material in the first few chapters, they may be unaccustomed to working exercises of the type presented here, most of which require an understanding of proofs.

Here is a chapter-by-chapter summary of the highlights of the book:

- Chapter 1: Vector spaces are defined in this chapter, and their basic properties are developed.
- Chapter 2: Linear independence, span, basis, and dimension are defined in this chapter, which presents the basic theory of finite-dimensional vector spaces.

- Chapter 3: Linear maps are introduced in this chapter. The key result here is the Fundamental Theorem of Linear Maps (3.22): if  $T$  is a linear map on  $V$ , then  $\dim V = \dim \text{null } T + \dim \text{range } T$ . Quotient spaces and duality are topics in this chapter at a higher level of abstraction than other parts of the book; these topics can be skipped without running into problems elsewhere in the book.
- Chapter 4: The part of the theory of polynomials that will be needed to understand linear operators is presented in this chapter. This chapter contains no linear algebra. It can be covered quickly, especially if your students are already familiar with these results.
- Chapter 5: The idea of studying a linear operator by restricting it to small subspaces leads to eigenvectors in the early part of this chapter. The highlight of this chapter is a simple proof that on complex vector spaces, eigenvalues always exist. This result is then used to show that each linear operator on a complex vector space has an upper-triangular matrix with respect to some basis. All this is done without defining determinants or characteristic polynomials!
- Chapter 6: Inner product spaces are defined in this chapter, and their basic properties are developed along with standard tools such as orthonormal bases and the Gram–Schmidt Procedure. This chapter also shows how orthogonal projections can be used to solve certain minimization problems.
- Chapter 7: The Spectral Theorem, which characterizes the linear operators for which there exists an orthonormal basis consisting of eigenvectors, is the highlight of this chapter. The work in earlier chapters pays off here with especially simple proofs. This chapter also deals with positive operators, isometries, the Polar Decomposition, and the Singular Value Decomposition.
- Chapter 8: Minimal polynomials, characteristic polynomials, and generalized eigenvectors are introduced in this chapter. The main achievement of this chapter is the description of a linear operator on a complex vector space in terms of its generalized eigenvectors. This description enables one to prove many of the results usually proved using Jordan Form. For example, these tools are used to prove that every invertible linear operator on a complex vector space has a square root. The chapter concludes with a proof that every linear operator on a complex vector space can be put into Jordan Form.

- Chapter 9: Linear operators on real vector spaces occupy center stage in this chapter. Here the main technique is complexification, which is a natural extension of an operator on a real vector space to an operator on a complex vector space. Complexification allows our results about complex vector spaces to be transferred easily to real vector spaces. For example, this technique is used to show that every linear operator on a real vector space has an invariant subspace of dimension 1 or 2. As another example, we show that that every linear operator on an odd-dimensional real vector space has an eigenvalue.
- Chapter 10: The trace and determinant (on complex vector spaces) are defined in this chapter as the sum of the eigenvalues and the product of the eigenvalues, both counting multiplicity. These easy-to-remember definitions would not be possible with the traditional approach to eigenvalues, because the traditional method uses determinants to prove that sufficient eigenvalues exist. The standard theorems about determinants now become much clearer. The Polar Decomposition and the Real Spectral Theorem are used to derive the change of variables formula for multivariable integrals in a fashion that makes the appearance of the determinant there seem natural.

This book usually develops linear algebra simultaneously for real and complex vector spaces by letting  $\mathbf{F}$  denote either the real or the complex numbers. If you and your students prefer to think of  $\mathbf{F}$  as an arbitrary field, then see the comments at the end of Section 1.A. I prefer avoiding arbitrary fields at this level because they introduce extra abstraction without leading to any new linear algebra. Also, students are more comfortable thinking of polynomials as functions instead of the more formal objects needed for polynomials with coefficients in finite fields. Finally, even if the beginning part of the theory were developed with arbitrary fields, inner product spaces would push consideration back to just real and complex vector spaces.

You probably cannot cover everything in this book in one semester. Going through the first eight chapters is a good goal for a one-semester course. If you must reach Chapter 10, then consider covering Chapters 4 and 9 in fifteen minutes each, as well as skipping the material on quotient spaces and duality in Chapter 3.

A goal more important than teaching any particular theorem is to develop in students the ability to understand and manipulate the objects of linear algebra. Mathematics can be learned only by doing. Fortunately, linear algebra has many good homework exercises. When teaching this course, during each class I usually assign as homework several of the exercises, due the next class. Going over the homework might take up a third or even half of a typical class.

**Major changes from the previous edition:**

- This edition contains 561 exercises, including 337 new exercises that were not in the previous edition. Exercises now appear at the end of each section, rather than at the end of each chapter.
- Many new examples have been added to illustrate the key ideas of linear algebra.
- Beautiful new formatting, including the use of color, creates pages with an unusually pleasant appearance in both print and electronic versions. As a visual aid, definitions are in beige boxes and theorems are in blue boxes (in color versions of the book).
- Each theorem now has a descriptive name.
- New topics covered in the book include product spaces, quotient spaces, and duality.
- Chapter 9 (Operators on Real Vector Spaces) has been completely rewritten to take advantage of simplifications via complexification. This approach allows for more streamlined presentations in Chapters 5 and 7 because those chapters now focus mostly on complex vector spaces.
- Hundreds of improvements have been made throughout the book. For example, the proof of Jordan Form (Section 8.D) has been simplified.

Please check the website below for additional information about the book. I may occasionally write new sections on additional topics. These new sections will be posted on the website. Your suggestions, comments, and corrections are most welcome.

Best wishes for teaching a successful linear algebra class!

Sheldon Axler  
Mathematics Department  
San Francisco State University  
San Francisco, CA 94132, USA

website: [linear.axler.net](http://linear.axler.net)  
e-mail: [linear@axler.net](mailto:linear@axler.net)  
Twitter: @AxlerLinear

# *Preface for the Student*

---

You are probably about to begin your second exposure to linear algebra. Unlike your first brush with the subject, which probably emphasized Euclidean spaces and matrices, this encounter will focus on abstract vector spaces and linear maps. These terms will be defined later, so don't worry if you do not know what they mean. This book starts from the beginning of the subject, assuming no knowledge of linear algebra. The key point is that you are about to immerse yourself in serious mathematics, with an emphasis on attaining a deep understanding of the definitions, theorems, and proofs.

You cannot read mathematics the way you read a novel. If you zip through a page in less than an hour, you are probably going too fast. When you encounter the phrase "as you should verify", you should indeed do the verification, which will usually require some writing on your part. When steps are left out, you need to supply the missing pieces. You should ponder and internalize each definition. For each theorem, you should seek examples to show why each hypothesis is necessary. Discussions with other students should help.

As a visual aid, definitions are in beige boxes and theorems are in blue boxes (in color versions of the book). Each theorem has a descriptive name.

Please check the website below for additional information about the book. I may occasionally write new sections on additional topics. These new sections will be posted on the website. Your suggestions, comments, and corrections are most welcome.

Best wishes for success and enjoyment in learning linear algebra!

Sheldon Axler  
Mathematics Department  
San Francisco State University  
San Francisco, CA 94132, USA

website: [linear.axler.net](http://linear.axler.net)  
e-mail: [linear@axler.net](mailto:linear@axler.net)  
Twitter: @AxlerLinear

# *Acknowledgments*

---

I owe a huge intellectual debt to the many mathematicians who created linear algebra over the past two centuries. The results in this book belong to the common heritage of mathematics. A special case of a theorem may first have been proved in the nineteenth century, then slowly sharpened and improved by many mathematicians. Bestowing proper credit on all the contributors would be a difficult task that I have not undertaken. In no case should the reader assume that any theorem presented here represents my original contribution. However, in writing this book I tried to think about the best way to present linear algebra and to prove its theorems, without regard to the standard methods and proofs used in most textbooks.

Many people helped make this a better book. The two previous editions of this book were used as a textbook at about 300 universities and colleges. I received thousands of suggestions and comments from faculty and students who used the second edition. I looked carefully at all those suggestions as I was working on this edition. At first, I tried keeping track of whose suggestions I used so that those people could be thanked here. But as changes were made and then replaced with better suggestions, and as the list grew longer, keeping track of the sources of each suggestion became too complicated. And lists are boring to read anyway. Thus in lieu of a long list of people who contributed good ideas, I will just say how truly grateful I am to everyone who sent me suggestions and comments. Many many thanks!

Special thanks to Ken Ribet and his giant (220 students) linear algebra class at Berkeley that class-tested a preliminary version of this third edition and that sent me more suggestions and corrections than any other group.

Finally, I thank Springer for providing me with help when I needed it and for allowing me the freedom to make the final decisions about the content and appearance of this book. Special thanks to Elizabeth Loew for her wonderful work as editor and David Kramer for unusually skillful copyediting.



# CHAPTER

# 1

*René Descartes explaining his work to Queen Christina of Sweden. Vector spaces are a generalization of the description of a plane using two coordinates, as published by Descartes in 1637.*

## Vector Spaces

Linear algebra is the study of linear maps on finite-dimensional vector spaces. Eventually we will learn what all these terms mean. In this chapter we will define vector spaces and discuss their elementary properties.

In linear algebra, better theorems and more insight emerge if complex numbers are investigated along with real numbers. Thus we will begin by introducing the complex numbers and their basic properties.

We will generalize the examples of a plane and ordinary space to  $\mathbf{R}^n$  and  $\mathbf{C}^n$ , which we then will generalize to the notion of a vector space. The elementary properties of a vector space will already seem familiar to you.

Then our next topic will be subspaces, which play a role for vector spaces analogous to the role played by subsets for sets. Finally, we will look at sums of subspaces (analogous to unions of subsets) and direct sums of subspaces (analogous to unions of disjoint sets).

### LEARNING OBJECTIVES FOR THIS CHAPTER

- basic properties of the complex numbers
- $\mathbf{R}^n$  and  $\mathbf{C}^n$
- vector spaces
- subspaces
- sums and direct sums of subspaces

## 1.A $\mathbf{R}^n$ and $\mathbf{C}^n$

### Complex Numbers

You should already be familiar with basic properties of the set  $\mathbf{R}$  of real numbers. Complex numbers were invented so that we can take square roots of negative numbers. The idea is to assume we have a square root of  $-1$ , denoted  $i$ , that obeys the usual rules of arithmetic. Here are the formal definitions:

#### 1.1 Definition complex numbers

- A **complex number** is an ordered pair  $(a, b)$ , where  $a, b \in \mathbf{R}$ , but we will write this as  $a + bi$ .
- The set of all complex numbers is denoted by  $\mathbf{C}$ :

$$\mathbf{C} = \{a + bi : a, b \in \mathbf{R}\}.$$

- **Addition and multiplication** on  $\mathbf{C}$  are defined by

$$(a + bi) + (c + di) = (a + c) + (b + d)i, \\ (a + bi)(c + di) = (ac - bd) + (ad + bc)i;$$

here  $a, b, c, d \in \mathbf{R}$ .

If  $a \in \mathbf{R}$ , we identify  $a + 0i$  with the real number  $a$ . Thus we can think of  $\mathbf{R}$  as a subset of  $\mathbf{C}$ . We also usually write  $0 + bi$  as just  $bi$ , and we usually write  $0 + 1i$  as just  $i$ .

The symbol  $i$  was first used to denote  $\sqrt{-1}$  by Swiss mathematician Leonhard Euler in 1777.

Using multiplication as defined above, you should verify that  $i^2 = -1$ . Do not memorize the formula for the product of two complex numbers; you can always rederive it by recalling that  $i^2 = -1$  and then using the usual rules of arithmetic (as given by 1.3).

#### 1.2 Example Evaluate $(2 + 3i)(4 + 5i)$ .

**Solution**

$$\begin{aligned}
 (2 + 3i)(4 + 5i) &= 2 \cdot 4 + 2 \cdot (5i) + (3i) \cdot 4 + (3i)(5i) \\
 &= 8 + 10i + 12i - 15 \\
 &= -7 + 22i
 \end{aligned}$$

### 1.3 Properties of complex arithmetic

#### commutativity

$\alpha + \beta = \beta + \alpha$  and  $\alpha\beta = \beta\alpha$  for all  $\alpha, \beta \in \mathbf{C}$ ;

#### associativity

$(\alpha + \beta) + \lambda = \alpha + (\beta + \lambda)$  and  $(\alpha\beta)\lambda = \alpha(\beta\lambda)$  for all  $\alpha, \beta, \lambda \in \mathbf{C}$ ;

#### identities

$\lambda + 0 = \lambda$  and  $\lambda 1 = \lambda$  for all  $\lambda \in \mathbf{C}$ ;

#### additive inverse

for every  $\alpha \in \mathbf{C}$ , there exists a unique  $\beta \in \mathbf{C}$  such that  $\alpha + \beta = 0$ ;

#### multiplicative inverse

for every  $\alpha \in \mathbf{C}$  with  $\alpha \neq 0$ , there exists a unique  $\beta \in \mathbf{C}$  such that  $\alpha\beta = 1$ ;

#### distributive property

$\lambda(\alpha + \beta) = \lambda\alpha + \lambda\beta$  for all  $\lambda, \alpha, \beta \in \mathbf{C}$ .

The properties above are proved using the familiar properties of real numbers and the definitions of complex addition and multiplication. The next example shows how commutativity of complex multiplication is proved. Proofs of the other properties above are left as exercises.

---

#### 1.4 Example

Show that  $\alpha\beta = \beta\alpha$  for all  $\alpha, \beta, \lambda \in \mathbf{C}$ .

**Solution** Suppose  $\alpha = a + bi$  and  $\beta = c + di$ , where  $a, b, c, d \in \mathbf{R}$ . Then the definition of multiplication of complex numbers shows that

$$\begin{aligned}\alpha\beta &= (a + bi)(c + di) \\ &= (ac - bd) + (ad + bc)i\end{aligned}$$

and

$$\begin{aligned}\beta\alpha &= (c + di)(a + bi) \\ &= (ca - db) + (cb + da)i.\end{aligned}$$

The equations above and the commutativity of multiplication and addition of real numbers show that  $\alpha\beta = \beta\alpha$ .

---

### 1.5 Definition $-\alpha$ , subtraction, $1/\alpha$ , division

Let  $\alpha, \beta \in \mathbf{C}$ .

- Let  $-\alpha$  denote the additive inverse of  $\alpha$ . Thus  $-\alpha$  is the unique complex number such that

$$\alpha + (-\alpha) = 0.$$

- **Subtraction** on  $\mathbf{C}$  is defined by

$$\beta - \alpha = \beta + (-\alpha).$$

- For  $\alpha \neq 0$ , let  $1/\alpha$  denote the multiplicative inverse of  $\alpha$ . Thus  $1/\alpha$  is the unique complex number such that

$$\alpha(1/\alpha) = 1.$$

- **Division** on  $\mathbf{C}$  is defined by

$$\beta/\alpha = \beta(1/\alpha).$$

So that we can conveniently make definitions and prove theorems that apply to both real and complex numbers, we adopt the following notation:

### 1.6 Notation $\mathbf{F}$

Throughout this book,  $\mathbf{F}$  stands for either  $\mathbf{R}$  or  $\mathbf{C}$ .

*The letter  $\mathbf{F}$  is used because  $\mathbf{R}$  and  $\mathbf{C}$  are examples of what are called fields.*

Thus if we prove a theorem involving  $\mathbf{F}$ , we will know that it holds when  $\mathbf{F}$  is replaced with  $\mathbf{R}$  and when  $\mathbf{F}$  is replaced with  $\mathbf{C}$ .

Elements of  $\mathbf{F}$  are called **scalars**. The word “scalar”, a fancy word for “number”, is often used when we want to emphasize that an object is a number, as opposed to a vector (vectors will be defined soon).

For  $\alpha \in \mathbf{F}$  and  $m$  a positive integer, we define  $\alpha^m$  to denote the product of  $\alpha$  with itself  $m$  times:

$$\alpha^m = \underbrace{\alpha \cdots \alpha}_{m \text{ times}}.$$

Clearly  $(\alpha^m)^n = \alpha^{mn}$  and  $(\alpha\beta)^m = \alpha^m\beta^m$  for all  $\alpha, \beta \in \mathbf{F}$  and all positive integers  $m, n$ .

## Lists

Before defining  $\mathbf{R}^n$  and  $\mathbf{C}^n$ , we look at two important examples.

### 1.7 Example $\mathbf{R}^2$ and $\mathbf{R}^3$

- The set  $\mathbf{R}^2$ , which you can think of as a plane, is the set of all ordered pairs of real numbers:

$$\mathbf{R}^2 = \{(x, y) : x, y \in \mathbf{R}\}.$$

- The set  $\mathbf{R}^3$ , which you can think of as ordinary space, is the set of all ordered triples of real numbers:

$$\mathbf{R}^3 = \{(x, y, z) : x, y, z \in \mathbf{R}\}.$$

To generalize  $\mathbf{R}^2$  and  $\mathbf{R}^3$  to higher dimensions, we first need to discuss the concept of lists.

### 1.8 Definition *list, length*

Suppose  $n$  is a nonnegative integer. A *list* of *length*  $n$  is an ordered collection of  $n$  elements (which might be numbers, other lists, or more abstract entities) separated by commas and surrounded by parentheses. A list of length  $n$  looks like this:

$$(x_1, \dots, x_n).$$

Two lists are equal if and only if they have the same length and the same elements in the same order.

Thus a list of length 2 is an ordered pair, and a list of length 3 is an ordered triple.

*Many mathematicians call a list of length  $n$  an  $n$ -tuple.*

Sometimes we will use the word *list* without specifying its length. Remember, however, that by definition each list has a finite length that is a nonnegative integer. Thus an object that looks like

$$(x_1, x_2, \dots),$$

which might be said to have infinite length, is not a list.

A list of length 0 looks like this:  $( )$ . We consider such an object to be a list so that some of our theorems will not have trivial exceptions.

Lists differ from sets in two ways: in lists, order matters and repetitions have meaning; in sets, order and repetitions are irrelevant.

### 1.9 Example *lists versus sets*

- The lists  $(3, 5)$  and  $(5, 3)$  are not equal, but the sets  $\{3, 5\}$  and  $\{5, 3\}$  are equal.
- The lists  $(4, 4)$  and  $(4, 4, 4)$  are not equal (they do not have the same length), although the sets  $\{4, 4\}$  and  $\{4, 4, 4\}$  both equal the set  $\{4\}$ .

## $\mathbf{F}^n$

To define the higher-dimensional analogues of  $\mathbf{R}^2$  and  $\mathbf{R}^3$ , we will simply replace  $\mathbf{R}$  with  $\mathbf{F}$  (which equals  $\mathbf{R}$  or  $\mathbf{C}$ ) and replace the numbers 2 or 3 with an arbitrary positive integer. Specifically, fix a positive integer  $n$  for the rest of this section.

### 1.10 Definition $\mathbf{F}^n$

$\mathbf{F}^n$  is the set of all lists of length  $n$  of elements of  $\mathbf{F}$ :

$$\mathbf{F}^n = \{(x_1, \dots, x_n) : x_j \in \mathbf{F} \text{ for } j = 1, \dots, n\}.$$

For  $(x_1, \dots, x_n) \in \mathbf{F}^n$  and  $j \in \{1, \dots, n\}$ , we say that  $x_j$  is the  $j^{\text{th}}$  coordinate of  $(x_1, \dots, x_n)$ .

If  $\mathbf{F} = \mathbf{R}$  and  $n$  equals 2 or 3, then this definition of  $\mathbf{F}^n$  agrees with our previous notions of  $\mathbf{R}^2$  and  $\mathbf{R}^3$ .

### 1.11 Example $\mathbf{C}^4$ is the set of all lists of four complex numbers:

$$\mathbf{C}^4 = \{(z_1, z_2, z_3, z_4) : z_1, z_2, z_3, z_4 \in \mathbf{C}\}.$$

For an amusing account of how  $\mathbf{R}^3$  would be perceived by creatures living in  $\mathbf{R}^2$ , read *Flatland: A Romance of Many Dimensions*, by Edwin A. Abbott. This novel, published in 1884, may help you imagine a physical space of four or more dimensions.

If  $n \geq 4$ , we cannot visualize  $\mathbf{R}^n$  as a physical object. Similarly,  $\mathbf{C}^1$  can be thought of as a plane, but for  $n \geq 2$ , the human brain cannot provide a full image of  $\mathbf{C}^n$ . However, even if  $n$  is large, we can perform algebraic manipulations in  $\mathbf{F}^n$  as easily as in  $\mathbf{R}^2$  or  $\mathbf{R}^3$ . For example, addition in  $\mathbf{F}^n$  is defined as follows:

### 1.12 Definition addition in $\mathbf{F}^n$

**Addition** in  $\mathbf{F}^n$  is defined by adding corresponding coordinates:

$$(x_1, \dots, x_n) + (y_1, \dots, y_n) = (x_1 + y_1, \dots, x_n + y_n).$$

Often the mathematics of  $\mathbf{F}^n$  becomes cleaner if we use a single letter to denote a list of  $n$  numbers, without explicitly writing the coordinates. For example, the result below is stated with  $x$  and  $y$  in  $\mathbf{F}^n$  even though the proof requires the more cumbersome notation of  $(x_1, \dots, x_n)$  and  $(y_1, \dots, y_n)$ .

### 1.13 Commutativity of addition in $\mathbf{F}^n$

If  $x, y \in \mathbf{F}^n$ , then  $x + y = y + x$ .

**Proof** Suppose  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$ . Then

$$\begin{aligned} x + y &= (x_1, \dots, x_n) + (y_1, \dots, y_n) \\ &= (x_1 + y_1, \dots, x_n + y_n) \\ &= (y_1 + x_1, \dots, y_n + x_n) \\ &= (y_1, \dots, y_n) + (x_1, \dots, x_n) \\ &= y + x, \end{aligned}$$

where the second and fourth equalities above hold because of the definition of addition in  $\mathbf{F}^n$  and the third equality holds because of the usual commutativity of addition in  $\mathbf{F}$ . ■

If a single letter is used to denote an element of  $\mathbf{F}^n$ , then the same letter with appropriate subscripts is often used when coordinates must be displayed. For example, if  $x \in \mathbf{F}^n$ , then letting  $x$  equal  $(x_1, \dots, x_n)$  is good notation, as shown in the proof above. Even better, work with just  $x$  and avoid explicit coordinates when possible.

The symbol ■ means “end of the proof”.

### 1.14 Definition 0

Let 0 denote the list of length  $n$  whose coordinates are all 0:

$$0 = (0, \dots, 0).$$

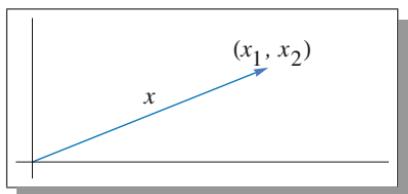
Here we are using the symbol 0 in two different ways—on the left side of the equation in 1.14, the symbol 0 denotes a list of length  $n$ , whereas on the right side, each 0 denotes a number. This potentially confusing practice actually causes no problems because the context always makes clear what is intended.

**1.15 Example** Consider the statement that 0 is an additive identity for  $\mathbf{F}^n$ :

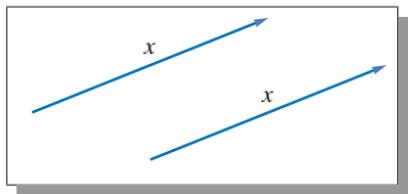
$$x + 0 = x \quad \text{for all } x \in \mathbf{F}^n.$$

Is the 0 above the number 0 or the list 0?

**Solution** Here 0 is a list, because we have not defined the sum of an element of  $\mathbf{F}^n$  (namely,  $x$ ) and the number 0.



*Elements of  $\mathbf{R}^2$  can be thought of as points or as vectors.*



*A vector.*

A picture can aid our intuition. We will draw pictures in  $\mathbf{R}^2$  because we can sketch this space on 2-dimensional surfaces such as paper and blackboards. A typical element of  $\mathbf{R}^2$  is a point  $x = (x_1, x_2)$ . Sometimes we think of  $x$  not as a point but as an arrow starting at the origin and ending at  $(x_1, x_2)$ , as shown here. When we think of  $x$  as an arrow, we refer to it as a *vector*.

When we think of vectors in  $\mathbf{R}^2$  as arrows, we can move an arrow parallel to itself (not changing its length or direction) and still think of it as the same vector. With that viewpoint, you will often gain better understanding by dispensing with the coordinate axes and the explicit coordinates and just thinking of the vector, as shown here.

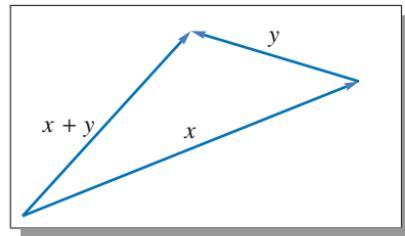
*Mathematical models of the economy can have thousands of variables, say  $x_1, \dots, x_{5000}$ , which means that we must operate in  $\mathbf{R}^{5000}$ . Such a space cannot be dealt with geometrically. However, the algebraic approach works well. Thus our subject is called linear algebra.*

Whenever we use pictures in  $\mathbf{R}^2$  or use the somewhat vague language of points and vectors, remember that these are just aids to our understanding, not substitutes for the actual mathematics that we will develop. Although we cannot draw good pictures in high-dimensional spaces, the elements of these spaces are as rigorously defined as elements of  $\mathbf{R}^2$ .

For example,  $(2, -3, 17, \pi, \sqrt{2})$  is an element of  $\mathbf{R}^5$ , and we may casually refer to it as a point in  $\mathbf{R}^5$  or a vector in  $\mathbf{R}^5$  without worrying about whether the geometry of  $\mathbf{R}^5$  has any physical meaning.

Recall that we defined the sum of two elements of  $\mathbf{F}^n$  to be the element of  $\mathbf{F}^n$  obtained by adding corresponding coordinates; see 1.12. As we will now see, addition has a simple geometric interpretation in the special case of  $\mathbf{R}^2$ .

Suppose we have two vectors  $x$  and  $y$  in  $\mathbf{R}^2$  that we want to add. Move the vector  $y$  parallel to itself so that its initial point coincides with the end point of the vector  $x$ , as shown here. The sum  $x + y$  then equals the vector whose initial point equals the initial point of  $x$  and whose end point equals the end point of the vector  $y$ , as shown here.



*The sum of two vectors.*

In the next definition, the 0 on the right side of the displayed equation below is the list  $0 \in \mathbf{F}^n$ .

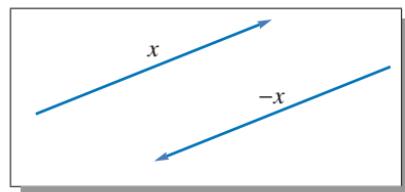
### 1.16 Definition *additive inverse in $\mathbf{F}^n$*

For  $x \in \mathbf{F}^n$ , the *additive inverse* of  $x$ , denoted  $-x$ , is the vector  $-x \in \mathbf{F}^n$  such that

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

In other words, if  $x = (x_1, \dots, x_n)$ , then  $-x = (-x_1, \dots, -x_n)$ .

For a vector  $x \in \mathbf{R}^2$ , the additive inverse  $-x$  is the vector parallel to  $x$  and with the same length as  $x$  but pointing in the opposite direction. The figure here illustrates this way of thinking about the additive inverse in  $\mathbf{R}^2$ .



*A vector and its additive inverse.*

Having dealt with addition in  $\mathbf{F}^n$ , we now turn to multiplication. We could define a multiplication in  $\mathbf{F}^n$  in a similar fashion, starting with two elements of  $\mathbf{F}^n$  and getting another element of  $\mathbf{F}^n$  by multiplying corresponding coordinates. Experience shows that this definition is not useful for our purposes. Another type of multiplication, called scalar multiplication, will be central to our subject. Specifically, we need to define what it means to multiply an element of  $\mathbf{F}^n$  by an element of  $\mathbf{F}$ .

### 1.17 Definition scalar multiplication in $\mathbf{F}^n$

The **product** of a number  $\lambda$  and a vector in  $\mathbf{F}^n$  is computed by multiplying each coordinate of the vector by  $\lambda$ :

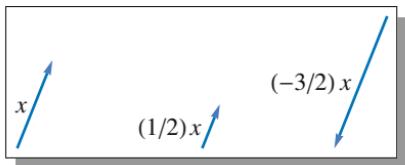
$$\lambda(x_1, \dots, x_n) = (\lambda x_1, \dots, \lambda x_n);$$

here  $\lambda \in \mathbf{F}$  and  $(x_1, \dots, x_n) \in \mathbf{F}^n$ .

*In scalar multiplication, we multiply together a scalar and a vector, getting a vector. You may be familiar with the dot product in  $\mathbf{R}^2$  or  $\mathbf{R}^3$ , in which we multiply together two vectors and get a scalar. Generalizations of the dot product will become important when we study inner products in Chapter 6.*

Scalar multiplication has a nice geometric interpretation in  $\mathbf{R}^2$ . If  $\lambda$  is a positive number and  $x$  is a vector in  $\mathbf{R}^2$ , then  $\lambda x$  is the vector that points in the same direction as  $x$  and whose length is  $\lambda$  times the length of  $x$ . In other words, to get  $\lambda x$ , we shrink or stretch  $x$  by a factor of  $\lambda$ , depending on whether  $\lambda < 1$  or  $\lambda > 1$ .

If  $\lambda$  is a negative number and  $x$  is a vector in  $\mathbf{R}^2$ , then  $\lambda x$  is the vector that points in the direction opposite to that of  $x$  and whose length is  $|\lambda|$  times the length of  $x$ , as shown here.



Scalar multiplication.

## Digression on Fields

A **field** is a set containing at least two distinct elements called 0 and 1, along with operations of addition and multiplication satisfying all the properties listed in 1.3. Thus **R** and **C** are fields, as is the set of rational numbers along with the usual operations of addition and multiplication. Another example of a field is the set  $\{0, 1\}$  with the usual operations of addition and multiplication except that  $1 + 1$  is defined to equal 0.

In this book we will not need to deal with fields other than **R** and **C**. However, many of the definitions, theorems, and proofs in linear algebra that work for both **R** and **C** also work without change for arbitrary fields. If you prefer to do so, throughout Chapters 1, 2, and 3 you can think of **F** as denoting an arbitrary field instead of **R** or **C**, except that some of the examples and exercises require that for each positive integer  $n$  we have  $\underbrace{1 + 1 + \cdots + 1}_{n \text{ times}} \neq 0$ .

## EXERCISES 1.A

---

- 1 Suppose  $a$  and  $b$  are real numbers, not both 0. Find real numbers  $c$  and  $d$  such that

$$1/(a + bi) = c + di.$$

- 2 Show that

$$\frac{-1 + \sqrt{3}i}{2}$$

is a cube root of 1 (meaning that its cube equals 1).

- 3 Find two distinct square roots of  $i$ .

- 4 Show that  $\alpha + \beta = \beta + \alpha$  for all  $\alpha, \beta \in \mathbf{C}$ .

- 5 Show that  $(\alpha + \beta) + \lambda = \alpha + (\beta + \lambda)$  for all  $\alpha, \beta, \lambda \in \mathbf{C}$ .

- 6 Show that  $(\alpha\beta)\lambda = \alpha(\beta\lambda)$  for all  $\alpha, \beta, \lambda \in \mathbf{C}$ .

- 7 Show that for every  $\alpha \in \mathbf{C}$ , there exists a unique  $\beta \in \mathbf{C}$  such that  $\alpha + \beta = 0$ .

- 8 Show that for every  $\alpha \in \mathbf{C}$  with  $\alpha \neq 0$ , there exists a unique  $\beta \in \mathbf{C}$  such that  $\alpha\beta = 1$ .

- 9 Show that  $\lambda(\alpha + \beta) = \lambda\alpha + \lambda\beta$  for all  $\lambda, \alpha, \beta \in \mathbf{C}$ .

- 10 Find  $x \in \mathbf{R}^4$  such that

$$(4, -3, 1, 7) + 2x = (5, 9, -6, 8).$$

- 11 Explain why there does not exist  $\lambda \in \mathbf{C}$  such that

$$\lambda(2 - 3i, 5 + 4i, -6 + 7i) = (12 - 5i, 7 + 22i, -32 - 9i).$$

- 12 Show that  $(x + y) + z = x + (y + z)$  for all  $x, y, z \in \mathbf{F}^n$ .

- 13 Show that  $(ab)x = a(bx)$  for all  $x \in \mathbf{F}^n$  and all  $a, b \in \mathbf{F}$ .

- 14 Show that  $1x = x$  for all  $x \in \mathbf{F}^n$ .

- 15 Show that  $\lambda(x + y) = \lambda x + \lambda y$  for all  $\lambda \in \mathbf{F}$  and all  $x, y \in \mathbf{F}^n$ .

- 16 Show that  $(a + b)x = ax + bx$  for all  $a, b \in \mathbf{F}$  and all  $x \in \mathbf{F}^n$ .

## 1.B Definition of Vector Space

The motivation for the definition of a vector space comes from properties of addition and scalar multiplication in  $\mathbf{F}^n$ : Addition is commutative, associative, and has an identity. Every element has an additive inverse. Scalar multiplication is associative. Scalar multiplication by 1 acts as expected. Addition and scalar multiplication are connected by distributive properties.

We will define a vector space to be a set  $V$  with an addition and a scalar multiplication on  $V$  that satisfy the properties in the paragraph above.

### 1.18 Definition *addition, scalar multiplication*

- An ***addition*** on a set  $V$  is a function that assigns an element  $u + v \in V$  to each pair of elements  $u, v \in V$ .
- A ***scalar multiplication*** on a set  $V$  is a function that assigns an element  $\lambda v \in V$  to each  $\lambda \in \mathbf{F}$  and each  $v \in V$ .

Now we are ready to give the formal definition of a vector space.

### 1.19 Definition *vector space*

A ***vector space*** is a set  $V$  along with an addition on  $V$  and a scalar multiplication on  $V$  such that the following properties hold:

#### **commutativity**

$$u + v = v + u \text{ for all } u, v \in V;$$

#### **associativity**

$$(u + v) + w = u + (v + w) \text{ and } (ab)v = a(bv) \text{ for all } u, v, w \in V \text{ and all } a, b \in \mathbf{F};$$

#### **additive identity**

$$\text{there exists an element } 0 \in V \text{ such that } v + 0 = v \text{ for all } v \in V;$$

#### **additive inverse**

$$\text{for every } v \in V, \text{ there exists } w \in V \text{ such that } v + w = 0;$$

#### **multiplicative identity**

$$1v = v \text{ for all } v \in V;$$

#### **distributive properties**

$$a(u + v) = au + av \text{ and } (a + b)v = av + bv \text{ for all } a, b \in \mathbf{F} \text{ and all } u, v \in V.$$

The following geometric language sometimes aids our intuition.

### 1.20 Definition *vector, point*

Elements of a vector space are called *vectors* or *points*.

The scalar multiplication in a vector space depends on  $\mathbf{F}$ . Thus when we need to be precise, we will say that  $V$  is a *vector space over  $\mathbf{F}$*  instead of saying simply that  $V$  is a vector space. For example,  $\mathbf{R}^n$  is a vector space over  $\mathbf{R}$ , and  $\mathbf{C}^n$  is a vector space over  $\mathbf{C}$ .

### 1.21 Definition *real vector space, complex vector space*

- A vector space over  $\mathbf{R}$  is called a *real vector space*.
- A vector space over  $\mathbf{C}$  is called a *complex vector space*.

Usually the choice of  $\mathbf{F}$  is either obvious from the context or irrelevant. Thus we often assume that  $\mathbf{F}$  is lurking in the background without specifically mentioning it.

With the usual operations of addition and scalar multiplication,  $\mathbf{F}^n$  is a vector space over  $\mathbf{F}$ , as you should verify. The example of  $\mathbf{F}^n$  motivated our definition of vector space.

*The simplest vector space contains only one point. In other words,  $\{0\}$  is a vector space.*

**1.22 Example**  $\mathbf{F}^\infty$  is defined to be the set of all sequences of elements of  $\mathbf{F}$ :

$$\mathbf{F}^\infty = \{(x_1, x_2, \dots) : x_j \in \mathbf{F} \text{ for } j = 1, 2, \dots\}.$$

Addition and scalar multiplication on  $\mathbf{F}^\infty$  are defined as expected:

$$(x_1, x_2, \dots) + (y_1, y_2, \dots) = (x_1 + y_1, x_2 + y_2, \dots),$$
$$\lambda(x_1, x_2, \dots) = (\lambda x_1, \lambda x_2, \dots).$$

With these definitions,  $\mathbf{F}^\infty$  becomes a vector space over  $\mathbf{F}$ , as you should verify. The additive identity in this vector space is the sequence of all 0's.

Our next example of a vector space involves a set of functions.

**1.23 Notation  $\mathbf{F}^S$** 

- If  $S$  is a set, then  $\mathbf{F}^S$  denotes the set of functions from  $S$  to  $\mathbf{F}$ .
- For  $f, g \in \mathbf{F}^S$ , the *sum*  $f + g \in \mathbf{F}^S$  is the function defined by

$$(f + g)(x) = f(x) + g(x)$$

for all  $x \in S$ .

- For  $\lambda \in \mathbf{F}$  and  $f \in \mathbf{F}^S$ , the *product*  $\lambda f \in \mathbf{F}^S$  is the function defined by

$$(\lambda f)(x) = \lambda f(x)$$

for all  $x \in S$ .

As an example of the notation above, if  $S$  is the interval  $[0, 1]$  and  $\mathbf{F} = \mathbf{R}$ , then  $\mathbf{R}^{[0,1]}$  is the set of real-valued functions on the interval  $[0, 1]$ .

You should verify all three bullet points in the next example.

**1.24 Example  $\mathbf{F}^S$  is a vector space**

- If  $S$  is a nonempty set, then  $\mathbf{F}^S$  (with the operations of addition and scalar multiplication as defined above) is a vector space over  $\mathbf{F}$ .
- The additive identity of  $\mathbf{F}^S$  is the function  $0 : S \rightarrow \mathbf{F}$  defined by

$$0(x) = 0$$

for all  $x \in S$ .

- For  $f \in \mathbf{F}^S$ , the additive inverse of  $f$  is the function  $-f : S \rightarrow \mathbf{F}$  defined by

$$(-f)(x) = -f(x)$$

for all  $x \in S$ .

*The elements of the vector space  $\mathbf{R}^{[0,1]}$  are real-valued functions on  $[0, 1]$ , not lists. In general, a vector space is an abstract entity whose elements might be lists, functions, or weird objects.*

Our previous examples of vector spaces,  $\mathbf{F}^n$  and  $\mathbf{F}^\infty$ , are special cases of the vector space  $\mathbf{F}^S$  because a list of length  $n$  of numbers in  $\mathbf{F}$  can be thought of as a function from  $\{1, 2, \dots, n\}$  to  $\mathbf{F}$  and a sequence of numbers in  $\mathbf{F}$  can be thought of as a function from the set of

positive integers to  $\mathbf{F}$ . In other words, we can think of  $\mathbf{F}^n$  as  $\mathbf{F}^{\{1, 2, \dots, n\}}$  and we can think of  $\mathbf{F}^\infty$  as  $\mathbf{F}^{\{1, 2, \dots\}}$ .

Soon we will see further examples of vector spaces, but first we need to develop some of the elementary properties of vector spaces.

The definition of a vector space requires that it have an additive identity. The result below states that this identity is unique.

### 1.25 Unique additive identity

A vector space has a unique additive identity.

**Proof** Suppose  $0$  and  $0'$  are both additive identities for some vector space  $V$ . Then

$$0' = 0' + 0 = 0 + 0' = 0,$$

where the first equality holds because  $0$  is an additive identity, the second equality comes from commutativity, and the third equality holds because  $0'$  is an additive identity. Thus  $0' = 0$ , proving that  $V$  has only one additive identity. ■

Each element  $v$  in a vector space has an additive inverse, an element  $w$  in the vector space such that  $v + w = 0$ . The next result shows that each element in a vector space has only one additive inverse.

### 1.26 Unique additive inverse

Every element in a vector space has a unique additive inverse.

**Proof** Suppose  $V$  is a vector space. Let  $v \in V$ . Suppose  $w$  and  $w'$  are additive inverses of  $v$ . Then

$$w = w + 0 = w + (v + w') = (w + v) + w' = 0 + w' = w'.$$

Thus  $w = w'$ , as desired. ■

Because additive inverses are unique, the following notation now makes sense.

### 1.27 Notation $-v, w - v$

Let  $v, w \in V$ . Then

- $-v$  denotes the additive inverse of  $v$ ;
- $w - v$  is defined to be  $w + (-v)$ .

Almost all the results in this book involve some vector space. To avoid having to restate frequently that  $V$  is a vector space, we now make the necessary declaration once and for all:

### 1.28 Notation $V$

For the rest of the book,  $V$  denotes a vector space over  $\mathbf{F}$ .

In the next result, 0 denotes a scalar (the number  $0 \in \mathbf{F}$ ) on the left side of the equation and a vector (the additive identity of  $V$ ) on the right side of the equation.

### 1.29 The number 0 times a vector

$0v = 0$  for every  $v \in V$ .

*Note that 1.29 asserts something about scalar multiplication and the additive identity of  $V$ . The only part of the definition of a vector space that connects scalar multiplication and vector addition is the distributive property. Thus the distributive property must be used in the proof of 1.29.*

**Proof** For  $v \in V$ , we have

$$0v = (0 + 0)v = 0v + 0v.$$

Adding the additive inverse of  $0v$  to both sides of the equation above gives  $0 = 0v$ , as desired. ■

In the next result, 0 denotes the additive identity of  $V$ . Although their proofs are similar, 1.29 and 1.30 are not identical. More precisely, 1.29 states that the product of the scalar 0 and any vector equals the vector 0, whereas 1.30 states that the product of any scalar and the vector 0 equals the vector 0.

### 1.30 A number times the vector 0

$a0 = 0$  for every  $a \in \mathbf{F}$ .

**Proof** For  $a \in \mathbf{F}$ , we have

$$a0 = a(0 + 0) = a0 + a0.$$

Adding the additive inverse of  $a0$  to both sides of the equation above gives  $0 = a0$ , as desired. ■

Now we show that if an element of  $V$  is multiplied by the scalar  $-1$ , then the result is the additive inverse of the element of  $V$ .

### 1.31 The number $-1$ times a vector

$(-1)v = -v$  for every  $v \in V$ .

**Proof** For  $v \in V$ , we have

$$v + (-1)v = 1v + (-1)v = (1 + (-1))v = 0v = 0.$$

This equation says that  $(-1)v$ , when added to  $v$ , gives 0. Thus  $(-1)v$  is the additive inverse of  $v$ , as desired. ■

## EXERCISES 1.B

---

- 1 Prove that  $-(-v) = v$  for every  $v \in V$ .
- 2 Suppose  $a \in \mathbf{F}$ ,  $v \in V$ , and  $av = 0$ . Prove that  $a = 0$  or  $v = 0$ .
- 3 Suppose  $v, w \in V$ . Explain why there exists a unique  $x \in V$  such that  $v + 3x = w$ .
- 4 The empty set is not a vector space. The empty set fails to satisfy only one of the requirements listed in 1.19. Which one?
- 5 Show that in the definition of a vector space (1.19), the additive inverse condition can be replaced with the condition that

$$0v = 0 \text{ for all } v \in V.$$

Here the 0 on the left side is the number 0, and the 0 on the right side is the additive identity of  $V$ . (The phrase “a condition can be replaced” in a definition means that the collection of objects satisfying the definition is unchanged if the original condition is replaced with the new condition.)

- 6 Let  $\infty$  and  $-\infty$  denote two distinct objects, neither of which is in  $\mathbf{R}$ . Define an addition and scalar multiplication on  $\mathbf{R} \cup \{\infty\} \cup \{-\infty\}$  as you could guess from the notation. Specifically, the sum and product of two real numbers is as usual, and for  $t \in \mathbf{R}$  define

$$t\infty = \begin{cases} -\infty & \text{if } t < 0, \\ 0 & \text{if } t = 0, \\ \infty & \text{if } t > 0, \end{cases} \quad t(-\infty) = \begin{cases} \infty & \text{if } t < 0, \\ 0 & \text{if } t = 0, \\ -\infty & \text{if } t > 0, \end{cases}$$

$$t + \infty = \infty + t = \infty, \quad t + (-\infty) = (-\infty) + t = -\infty,$$

$$\infty + \infty = \infty, \quad (-\infty) + (-\infty) = -\infty, \quad \infty + (-\infty) = 0.$$

Is  $\mathbf{R} \cup \{\infty\} \cup \{-\infty\}$  a vector space over  $\mathbf{R}$ ? Explain.

## 1.C Subspaces

By considering subspaces, we can greatly expand our examples of vector spaces.

### 1.32 Definition *subspace*

A subset  $U$  of  $V$  is called a *subspace* of  $V$  if  $U$  is also a vector space (using the same addition and scalar multiplication as on  $V$ ).

### 1.33 Example $\{(x_1, x_2, 0) : x_1, x_2 \in \mathbf{F}\}$ is a subspace of $\mathbf{F}^3$ .

*Some mathematicians use the term **linear subspace**, which means the same as subspace.*

The next result gives the easiest way to check whether a subset of a vector space is a subspace.

### 1.34 Conditions for a subspace

A subset  $U$  of  $V$  is a subspace of  $V$  if and only if  $U$  satisfies the following three conditions:

#### additive identity

$$0 \in U$$

#### closed under addition

$$u, w \in U \text{ implies } u + w \in U;$$

#### closed under scalar multiplication

$$a \in \mathbf{F} \text{ and } u \in U \text{ implies } au \in U.$$

*The additive identity condition above could be replaced with the condition that  $U$  is nonempty (then taking  $u \in U$ , multiplying it by 0, and using the condition that  $U$  is closed under scalar multiplication would imply that  $0 \in U$ ). However, if  $U$  is indeed a subspace of  $V$ , then the easiest way to show that  $U$  is nonempty is to show that  $0 \in U$ .*

**Proof** If  $U$  is a subspace of  $V$ , then  $U$  satisfies the three conditions above by the definition of vector space.

Conversely, suppose  $U$  satisfies the three conditions above. The first condition above ensures that the additive identity of  $V$  is in  $U$ .

The second condition above ensures that addition makes sense on  $U$ . The third condition ensures that scalar multiplication makes sense on  $U$ .

If  $u \in U$ , then  $-u$  [which equals  $(-1)u$  by 1.31] is also in  $U$  by the third condition above. Hence every element of  $U$  has an additive inverse in  $U$ .

The other parts of the definition of a vector space, such as associativity and commutativity, are automatically satisfied for  $U$  because they hold on the larger space  $V$ . Thus  $U$  is a vector space and hence is a subspace of  $V$ . ■

The three conditions in the result above usually enable us to determine quickly whether a given subset of  $V$  is a subspace of  $V$ . You should verify all the assertions in the next example.

### 1.35 Example subspaces

- (a) If  $b \in \mathbf{F}$ , then

$$\{(x_1, x_2, x_3, x_4) \in \mathbf{F}^4 : x_3 = 5x_4 + b\}$$

is a subspace of  $\mathbf{F}^4$  if and only if  $b = 0$ .

- (b) The set of continuous real-valued functions on the interval  $[0, 1]$  is a subspace of  $\mathbf{R}^{[0,1]}$ .
- (c) The set of differentiable real-valued functions on  $\mathbf{R}$  is a subspace of  $\mathbf{R}^\mathbf{R}$ .
- (d) The set of differentiable real-valued functions  $f$  on the interval  $(0, 3)$  such that  $f'(2) = b$  is a subspace of  $\mathbf{R}^{(0,3)}$  if and only if  $b = 0$ .
- (e) The set of all sequences of complex numbers with limit 0 is a subspace of  $\mathbf{C}^\infty$ .

Verifying some of the items above shows the linear structure underlying parts of calculus. For example, the second item above requires the result that the sum of two continuous functions is continuous. As another example, the fourth item above requires the result that for a constant  $c$ , the derivative of  $cf$  equals  $c$  times the derivative of  $f$ .

*Clearly  $\{0\}$  is the smallest subspace of  $V$  and  $V$  itself is the largest subspace of  $V$ . The empty set is not a subspace of  $V$  because a subspace must be a vector space and hence must contain at least one element, namely, an additive identity.*

The subspaces of  $\mathbf{R}^2$  are precisely  $\{0\}$ ,  $\mathbf{R}^2$ , and all lines in  $\mathbf{R}^2$  through the origin. The subspaces of  $\mathbf{R}^3$  are precisely  $\{0\}$ ,  $\mathbf{R}^3$ , all lines in  $\mathbf{R}^3$  through the origin, and all planes in  $\mathbf{R}^3$  through the origin. To prove that all these objects are indeed subspaces is easy—the hard part is to show that they are the only subspaces of  $\mathbf{R}^2$  and  $\mathbf{R}^3$ . That task will be easier after we introduce some additional tools in the next chapter.

## Sums of Subspaces

*The union of subspaces is rarely a subspace (see Exercise 12), which is why we usually work with sums rather than unions.*

When dealing with vector spaces, we are usually interested only in subspaces, as opposed to arbitrary subsets. The notion of the sum of subspaces will be useful.

### 1.36 Definition sum of subsets

Suppose  $U_1, \dots, U_m$  are subsets of  $V$ . The **sum** of  $U_1, \dots, U_m$ , denoted  $U_1 + \dots + U_m$ , is the set of all possible sums of elements of  $U_1, \dots, U_m$ . More precisely,

$$U_1 + \dots + U_m = \{u_1 + \dots + u_m : u_1 \in U_1, \dots, u_m \in U_m\}.$$

Let's look at some examples of sums of subspaces.

---

**1.37 Example** Suppose  $U$  is the set of all elements of  $\mathbf{F}^3$  whose second and third coordinates equal 0, and  $W$  is the set of all elements of  $\mathbf{F}^3$  whose first and third coordinates equal 0:

$$U = \{(x, 0, 0) \in \mathbf{F}^3 : x \in \mathbf{F}\} \quad \text{and} \quad W = \{(0, y, 0) \in \mathbf{F}^3 : y \in \mathbf{F}\}.$$

Then

$$U + W = \{(x, y, 0) : x, y \in \mathbf{F}\},$$

as you should verify.

---

**1.38 Example** Suppose that  $U = \{(x, x, y, y) \in \mathbf{F}^4 : x, y \in \mathbf{F}\}$  and  $W = \{(x, x, x, y) \in \mathbf{F}^4 : x, y \in \mathbf{F}\}$ . Then

$$U + W = \{(x, x, y, z) \in \mathbf{F}^4 : x, y, z \in \mathbf{F}\},$$

as you should verify.

---

The next result states that the sum of subspaces is a subspace, and is in fact the smallest subspace containing all the summands.

### 1.39 Sum of subspaces is the smallest containing subspace

Suppose  $U_1, \dots, U_m$  are subspaces of  $V$ . Then  $U_1 + \dots + U_m$  is the smallest subspace of  $V$  containing  $U_1, \dots, U_m$ .

**Proof** It is easy to see that  $0 \in U_1 + \cdots + U_m$  and that  $U_1 + \cdots + U_m$  is closed under addition and scalar multiplication. Thus 1.34 implies that  $U_1 + \cdots + U_m$  is a subspace of  $V$ .

Clearly  $U_1, \dots, U_m$  are all contained in  $U_1 + \cdots + U_m$  (to see this, consider sums  $u_1 + \cdots + u_m$  where all except one of the  $u$ 's are 0). Conversely, every subspace of  $V$  containing  $U_1, \dots, U_m$  contains  $U_1 + \cdots + U_m$  (because subspaces must contain all finite sums of their elements). Thus  $U_1 + \cdots + U_m$  is the smallest subspace of  $V$  containing  $U_1, \dots, U_m$ . ■

*Sums of subspaces in the theory of vector spaces are analogous to unions of subsets in set theory. Given two subspaces of a vector space, the smallest subspace containing them is their sum. Analogously, given two subsets of a set, the smallest subset containing them is their union.*

## Direct Sums

Suppose  $U_1, \dots, U_m$  are subspaces of  $V$ . Every element of  $U_1 + \cdots + U_m$  can be written in the form

$$u_1 + \cdots + u_m,$$

where each  $u_j$  is in  $U_j$ . We will be especially interested in cases where each vector in  $U_1 + \cdots + U_m$  can be represented in the form above in only one way. This situation is so important that we give it a special name: direct sum.

### 1.40 Definition direct sum

Suppose  $U_1, \dots, U_m$  are subspaces of  $V$ .

- The sum  $U_1 + \cdots + U_m$  is called a **direct sum** if each element of  $U_1 + \cdots + U_m$  can be written in only one way as a sum  $u_1 + \cdots + u_m$ , where each  $u_j$  is in  $U_j$ .
- If  $U_1 + \cdots + U_m$  is a direct sum, then  $U_1 \oplus \cdots \oplus U_m$  denotes  $U_1 + \cdots + U_m$ , with the  $\oplus$  notation serving as an indication that this is a direct sum.

**1.41 Example** Suppose  $U$  is the subspace of  $\mathbf{F}^3$  of those vectors whose last coordinate equals 0, and  $W$  is the subspace of  $\mathbf{F}^3$  of those vectors whose first two coordinates equal 0:

$U = \{(x, y, 0) \in \mathbf{F}^3 : x, y \in \mathbf{F}\}$  and  $W = \{(0, 0, z) \in \mathbf{F}^3 : z \in \mathbf{F}\}$ . Then  $\mathbf{F}^3 = U \oplus W$ , as you should verify.

**1.42 Example** Suppose  $U_j$  is the subspace of  $\mathbf{F}^n$  of those vectors whose coordinates are all 0, except possibly in the  $j^{\text{th}}$  slot (thus, for example,  $U_2 = \{(0, x, 0, \dots, 0) \in \mathbf{F}^n : x \in \mathbf{F}\}$ ). Then

$$\mathbf{F}^n = U_1 \oplus \cdots \oplus U_n,$$

as you should verify.

---

Sometimes nonexamples add to our understanding as much as examples.

---

**1.43 Example** Let

$$U_1 = \{(x, y, 0) \in \mathbf{F}^3 : x, y \in \mathbf{F}\},$$

$$U_2 = \{(0, 0, z) \in \mathbf{F}^3 : z \in \mathbf{F}\},$$

$$U_3 = \{(0, y, y) \in \mathbf{F}^3 : y \in \mathbf{F}\}.$$

Show that  $U_1 + U_2 + U_3$  is not a direct sum.

**Solution** Clearly  $\mathbf{F}^3 = U_1 + U_2 + U_3$ , because every vector  $(x, y, z) \in \mathbf{F}^3$  can be written as

$$(x, y, z) = (x, y, 0) + (0, 0, z) + (0, 0, 0),$$

where the first vector on the right side is in  $U_1$ , the second vector is in  $U_2$ , and the third vector is in  $U_3$ .

However,  $\mathbf{F}^3$  does not equal the direct sum of  $U_1, U_2, U_3$ , because the vector  $(0, 0, 0)$  can be written in two different ways as a sum  $u_1 + u_2 + u_3$ , with each  $u_j$  in  $U_j$ . Specifically, we have

$$(0, 0, 0) = (0, 1, 0) + (0, 0, 1) + (0, -1, -1)$$

and, of course,

$$(0, 0, 0) = (0, 0, 0) + (0, 0, 0) + (0, 0, 0),$$

where the first vector on the right side of each equation above is in  $U_1$ , the second vector is in  $U_2$ , and the third vector is in  $U_3$ .

---

The symbol  $\oplus$ , which is a plus sign inside a circle, serves as a reminder that we are dealing with a special type of sum of subspaces—each element in the direct sum can be represented only one way as a sum of elements from the specified subspaces.

The definition of direct sum requires that every vector in the sum have a unique representation as an appropriate sum. The next result shows that when deciding whether a sum of subspaces is a direct sum, we need only consider whether 0 can be uniquely written as an appropriate sum.

### 1.44 Condition for a direct sum

Suppose  $U_1, \dots, U_m$  are subspaces of  $V$ . Then  $U_1 + \dots + U_m$  is a direct sum if and only if the only way to write 0 as a sum  $u_1 + \dots + u_m$ , where each  $u_j$  is in  $U_j$ , is by taking each  $u_j$  equal to 0.

**Proof** First suppose  $U_1 + \dots + U_m$  is a direct sum. Then the definition of direct sum implies that the only way to write 0 as a sum  $u_1 + \dots + u_m$ , where each  $u_j$  is in  $U_j$ , is by taking each  $u_j$  equal to 0.

Now suppose that the only way to write 0 as a sum  $u_1 + \dots + u_m$ , where each  $u_j$  is in  $U_j$ , is by taking each  $u_j$  equal to 0. To show that  $U_1 + \dots + U_m$  is a direct sum, let  $v \in U_1 + \dots + U_m$ . We can write

$$v = u_1 + \dots + u_m$$

for some  $u_1 \in U_1, \dots, u_m \in U_m$ . To show that this representation is unique, suppose we also have

$$v = v_1 + \dots + v_m,$$

where  $v_1 \in U_1, \dots, v_m \in U_m$ . Subtracting these two equations, we have

$$0 = (u_1 - v_1) + \dots + (u_m - v_m).$$

Because  $u_1 - v_1 \in U_1, \dots, u_m - v_m \in U_m$ , the equation above implies that each  $u_j - v_j$  equals 0. Thus  $u_1 = v_1, \dots, u_m = v_m$ , as desired. ■

The next result gives a simple condition for testing which pairs of subspaces give a direct sum.

### 1.45 Direct sum of two subspaces

Suppose  $U$  and  $W$  are subspaces of  $V$ . Then  $U + W$  is a direct sum if and only if  $U \cap W = \{0\}$ .

**Proof** First suppose that  $U + W$  is a direct sum. If  $v \in U \cap W$ , then  $0 = v + (-v)$ , where  $v \in U$  and  $-v \in W$ . By the unique representation of 0 as the sum of a vector in  $U$  and a vector in  $W$ , we have  $v = 0$ . Thus  $U \cap W = \{0\}$ , completing the proof in one direction.

To prove the other direction, now suppose  $U \cap W = \{0\}$ . To prove that  $U + W$  is a direct sum, suppose  $u \in U, w \in W$ , and

$$0 = u + w.$$

To complete the proof, we need only show that  $u = w = 0$  (by 1.44). The equation above implies that  $u = -w \in W$ . Thus  $u \in U \cap W$ . Hence  $u = 0$ , which by the equation above implies that  $w = 0$ , completing the proof. ■

*Sums of subspaces are analogous to unions of subsets. Similarly, direct sums of subspaces are analogous to disjoint unions of subsets. No two subspaces of a vector space can be disjoint, because both contain 0. So disjointness is replaced, at least in the case of two subspaces, with the requirement that the intersection equals {0}.*

The result above deals only with the case of two subspaces. When asking about a possible direct sum with more than two subspaces, it is not enough to test that each pair of the subspaces intersect only at 0. To see this, consider Example 1.43. In that nonexample of a direct sum, we have  $U_1 \cap U_2 = U_1 \cap U_3 = U_2 \cap U_3 = \{0\}$ .

## EXERCISES 1.C

---

- 1 For each of the following subsets of  $\mathbf{F}^3$ , determine whether it is a subspace of  $\mathbf{F}^3$ :
  - (a)  $\{(x_1, x_2, x_3) \in \mathbf{F}^3 : x_1 + 2x_2 + 3x_3 = 0\}$ ;
  - (b)  $\{(x_1, x_2, x_3) \in \mathbf{F}^3 : x_1 + 2x_2 + 3x_3 = 4\}$ ;
  - (c)  $\{(x_1, x_2, x_3) \in \mathbf{F}^3 : x_1 x_2 x_3 = 0\}$ ;
  - (d)  $\{(x_1, x_2, x_3) \in \mathbf{F}^3 : x_1 = 5x_3\}$ .
- 2 Verify all the assertions in Example 1.35.
- 3 Show that the set of differentiable real-valued functions  $f$  on the interval  $(-4, 4)$  such that  $f'(-1) = 3f(2)$  is a subspace of  $\mathbf{R}^{(-4,4)}$ .
- 4 Suppose  $b \in \mathbf{R}$ . Show that the set of continuous real-valued functions  $f$  on the interval  $[0, 1]$  such that  $\int_0^1 f = b$  is a subspace of  $\mathbf{R}^{[0,1]}$  if and only if  $b = 0$ .
- 5 Is  $\mathbf{R}^2$  a subspace of the complex vector space  $\mathbf{C}^2$ ?
- 6 (a) Is  $\{(a, b, c) \in \mathbf{R}^3 : a^3 = b^3\}$  a subspace of  $\mathbf{R}^3$ ?  
 (b) Is  $\{(a, b, c) \in \mathbf{C}^3 : a^3 = b^3\}$  a subspace of  $\mathbf{C}^3$ ?
- 7 Give an example of a nonempty subset  $U$  of  $\mathbf{R}^2$  such that  $U$  is closed under addition and under taking additive inverses (meaning  $-u \in U$  whenever  $u \in U$ ), but  $U$  is not a subspace of  $\mathbf{R}^2$ .
- 8 Give an example of a nonempty subset  $U$  of  $\mathbf{R}^2$  such that  $U$  is closed under scalar multiplication, but  $U$  is not a subspace of  $\mathbf{R}^2$ .

- 9** A function  $f : \mathbf{R} \rightarrow \mathbf{R}$  is called *periodic* if there exists a positive number  $p$  such that  $f(x) = f(x + p)$  for all  $x \in \mathbf{R}$ . Is the set of periodic functions from  $\mathbf{R}$  to  $\mathbf{R}$  a subspace of  $\mathbf{R}^{\mathbf{R}}$ ? Explain.
- 10** Suppose  $U_1$  and  $U_2$  are subspaces of  $V$ . Prove that the intersection  $U_1 \cap U_2$  is a subspace of  $V$ .
- 11** Prove that the intersection of every collection of subspaces of  $V$  is a subspace of  $V$ .
- 12** Prove that the union of two subspaces of  $V$  is a subspace of  $V$  if and only if one of the subspaces is contained in the other.
- 13** Prove that the union of three subspaces of  $V$  is a subspace of  $V$  if and only if one of the subspaces contains the other two.  
*[This exercise is surprisingly harder than the previous exercise, possibly because this exercise is not true if we replace  $\mathbf{F}$  with a field containing only two elements.]*
- 14** Verify the assertion in Example 1.38.
- 15** Suppose  $U$  is a subspace of  $V$ . What is  $U + U$ ?
- 16** Is the operation of addition on the subspaces of  $V$  commutative? In other words, if  $U$  and  $W$  are subspaces of  $V$ , is  $U + W = W + U$ ?
- 17** Is the operation of addition on the subspaces of  $V$  associative? In other words, if  $U_1, U_2, U_3$  are subspaces of  $V$ , is

$$(U_1 + U_2) + U_3 = U_1 + (U_2 + U_3)?$$

- 18** Does the operation of addition on the subspaces of  $V$  have an additive identity? Which subspaces have additive inverses?
- 19** Prove or give a counterexample: if  $U_1, U_2, W$  are subspaces of  $V$  such that

$$U_1 + W = U_2 + W,$$

then  $U_1 = U_2$ .

- 20** Suppose

$$U = \{(x, x, y, y) \in \mathbf{F}^4 : x, y \in \mathbf{F}\}.$$

Find a subspace  $W$  of  $\mathbf{F}^4$  such that  $\mathbf{F}^4 = U \oplus W$ .

**21** Suppose

$$U = \{(x, y, x + y, x - y, 2x) \in \mathbf{F}^5 : x, y \in \mathbf{F}\}.$$

Find a subspace  $W$  of  $\mathbf{F}^5$  such that  $\mathbf{F}^5 = U \oplus W$ .

**22** Suppose

$$U = \{(x, y, x + y, x - y, 2x) \in \mathbf{F}^5 : x, y \in \mathbf{F}\}.$$

Find three subspaces  $W_1, W_2, W_3$  of  $\mathbf{F}^5$ , none of which equals  $\{0\}$ , such that  $\mathbf{F}^5 = U \oplus W_1 \oplus W_2 \oplus W_3$ .

**23** Prove or give a counterexample: if  $U_1, U_2, W$  are subspaces of  $V$  such that

$$V = U_1 \oplus W \quad \text{and} \quad V = U_2 \oplus W,$$

then  $U_1 = U_2$ .

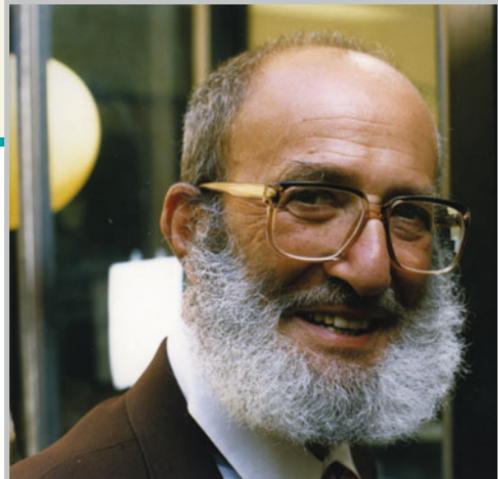
**24** A function  $f : \mathbf{R} \rightarrow \mathbf{R}$  is called *even* if

$$f(-x) = f(x)$$

for all  $x \in \mathbf{R}$ . A function  $f : \mathbf{R} \rightarrow \mathbf{R}$  is called *odd* if

$$f(-x) = -f(x)$$

for all  $x \in \mathbf{R}$ . Let  $U_e$  denote the set of real-valued even functions on  $\mathbf{R}$  and let  $U_o$  denote the set of real-valued odd functions on  $\mathbf{R}$ . Show that  $\mathbf{R}^\mathbf{R} = U_e \oplus U_o$ .



American mathematician Paul Halmos (1916–2006), who in 1942 published the first modern linear algebra book. The title of Halmos's book was the same as the title of this chapter.

# Finite-Dimensional Vector Spaces

Let's review our standing assumptions:

## 2.1 Notation $\mathbf{F}, V$

- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  denotes a vector space over  $\mathbf{F}$ .

In the last chapter we learned about vector spaces. Linear algebra focuses not on arbitrary vector spaces, but on finite-dimensional vector spaces, which we introduce in this chapter.

### LEARNING OBJECTIVES FOR THIS CHAPTER

- span
- linear independence
- bases
- dimension

## 2.A Span and Linear Independence

We have been writing lists of numbers surrounded by parentheses, and we will continue to do so for elements of  $\mathbf{F}^n$ ; for example,  $(2, -7, 8) \in \mathbf{F}^3$ . However, now we need to consider lists of vectors (which may be elements of  $\mathbf{F}^n$  or of other vector spaces). To avoid confusion, we will usually write lists of vectors without surrounding parentheses. For example,  $(4, 1, 6), (9, 5, 7)$  is a list of length 2 of vectors in  $\mathbf{R}^3$ .

### 2.2 Notation list of vectors

We will usually write lists of vectors without surrounding parentheses.

## Linear Combinations and Span

Adding up scalar multiples of vectors in a list gives what is called a linear combination of the list. Here is the formal definition:

### 2.3 Definition linear combination

A **linear combination** of a list  $v_1, \dots, v_m$  of vectors in  $V$  is a vector of the form

$$a_1v_1 + \cdots + a_mv_m,$$

where  $a_1, \dots, a_m \in \mathbf{F}$ .

---

### 2.4 Example In $\mathbf{F}^3$ ,

- $(17, -4, 2)$  is a linear combination of  $(2, 1, -3), (1, -2, 4)$  because  

$$(17, -4, 2) = 6(2, 1, -3) + 5(1, -2, 4).$$
- $(17, -4, 5)$  is not a linear combination of  $(2, 1, -3), (1, -2, 4)$  because there do not exist numbers  $a_1, a_2 \in \mathbf{F}$  such that  

$$(17, -4, 5) = a_1(2, 1, -3) + a_2(1, -2, 4).$$

In other words, the system of equations

$$\begin{aligned} 17 &= 2a_1 + a_2 \\ -4 &= a_1 - 2a_2 \\ 5 &= -3a_1 + 4a_2 \end{aligned}$$

---

has no solutions (as you should verify).

## 2.5 Definition *span*

The set of all linear combinations of a list of vectors  $v_1, \dots, v_m$  in  $V$  is called the *span* of  $v_1, \dots, v_m$ , denoted  $\text{span}(v_1, \dots, v_m)$ . In other words,

$$\text{span}(v_1, \dots, v_m) = \{a_1v_1 + \dots + a_mv_m : a_1, \dots, a_m \in \mathbf{F}\}.$$

The span of the empty list () is defined to be  $\{0\}$ .

## 2.6 Example

The previous example shows that in  $\mathbf{F}^3$ ,

- $(17, -4, 2) \in \text{span}((2, 1, -3), (1, -2, 4))$ ;
- $(17, -4, 5) \notin \text{span}((2, 1, -3), (1, -2, 4))$ .

Some mathematicians use the term *linear span*, which means the same as span.

## 2.7 Span is the smallest containing subspace

The span of a list of vectors in  $V$  is the smallest subspace of  $V$  containing all the vectors in the list.

**Proof** Suppose  $v_1, \dots, v_m$  is a list of vectors in  $V$ .

First we show that  $\text{span}(v_1, \dots, v_m)$  is a subspace of  $V$ . The additive identity is in  $\text{span}(v_1, \dots, v_m)$ , because

$$0 = 0v_1 + \dots + 0v_m.$$

Also,  $\text{span}(v_1, \dots, v_m)$  is closed under addition, because

$$(a_1v_1 + \dots + a_mv_m) + (c_1v_1 + \dots + c_mv_m) = (a_1 + c_1)v_1 + \dots + (a_m + c_m)v_m.$$

Furthermore,  $\text{span}(v_1, \dots, v_m)$  is closed under scalar multiplication, because

$$\lambda(a_1v_1 + \dots + a_mv_m) = \lambda a_1v_1 + \dots + \lambda a_mv_m.$$

Thus  $\text{span}(v_1, \dots, v_m)$  is a subspace of  $V$  (by 1.34).

Each  $v_j$  is a linear combination of  $v_1, \dots, v_m$  (to show this, set  $a_j = 1$  and let the other  $a$ 's in 2.3 equal 0). Thus  $\text{span}(v_1, \dots, v_m)$  contains each  $v_j$ . Conversely, because subspaces are closed under scalar multiplication and addition, every subspace of  $V$  containing each  $v_j$  contains  $\text{span}(v_1, \dots, v_m)$ . Thus  $\text{span}(v_1, \dots, v_m)$  is the smallest subspace of  $V$  containing all the vectors  $v_1, \dots, v_m$ . ■

## 2.8 Definition spans

If  $\text{span}(v_1, \dots, v_m)$  equals  $V$ , we say that  $v_1, \dots, v_m$  **spans**  $V$ .

### 2.9 Example Suppose $n$ is a positive integer. Show that

$$(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$$

spans  $\mathbf{F}^n$ . Here the  $j^{\text{th}}$  vector in the list above is the  $n$ -tuple with 1 in the  $j^{\text{th}}$  slot and 0 in all other slots.

**Solution** Suppose  $(x_1, \dots, x_n) \in \mathbf{F}^n$ . Then

$$(x_1, \dots, x_n) = x_1(1, 0, \dots, 0) + x_2(0, 1, 0, \dots, 0) + \dots + x_n(0, \dots, 0, 1).$$

Thus  $(x_1, \dots, x_n) \in \text{span}((1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1))$ , as desired.

Now we can make one of the key definitions in linear algebra.

## 2.10 Definition finite-dimensional vector space

A vector space is called **finite-dimensional** if some list of vectors in it spans the space.

Recall that by definition every list has finite length.

Example 2.9 above shows that  $\mathbf{F}^n$  is a finite-dimensional vector space for every positive integer  $n$ .

The definition of a polynomial is no doubt already familiar to you.

## 2.11 Definition polynomial, $\mathcal{P}(\mathbf{F})$

- A function  $p : \mathbf{F} \rightarrow \mathbf{F}$  is called a **polynomial** with coefficients in  $\mathbf{F}$  if there exist  $a_0, \dots, a_m \in \mathbf{F}$  such that

$$p(z) = a_0 + a_1 z + a_2 z^2 + \dots + a_m z^m$$

for all  $z \in \mathbf{F}$ .

- $\mathcal{P}(\mathbf{F})$  is the set of all polynomials with coefficients in  $\mathbf{F}$ .

With the usual operations of addition and scalar multiplication,  $\mathcal{P}(\mathbf{F})$  is a vector space over  $\mathbf{F}$ , as you should verify. In other words,  $\mathcal{P}(\mathbf{F})$  is a subspace of  $\mathbf{F}^{\mathbf{F}}$ , the vector space of functions from  $\mathbf{F}$  to  $\mathbf{F}$ .

If a polynomial (thought of as a function from  $\mathbf{F}$  to  $\mathbf{F}$ ) is represented by two sets of coefficients, then subtracting one representation of the polynomial from the other produces a polynomial that is identically zero as a function on  $\mathbf{F}$  and hence has all zero coefficients (if you are unfamiliar with this fact, just believe it for now; we will prove it later—see 4.7). **Conclusion:** the coefficients of a polynomial are uniquely determined by the polynomial. Thus the next definition uniquely defines the degree of a polynomial.

### 2.12 Definition *degree of a polynomial*, $\deg p$

- A polynomial  $p \in \mathcal{P}(\mathbf{F})$  is said to have *degree*  $m$  if there exist scalars  $a_0, a_1, \dots, a_m \in \mathbf{F}$  with  $a_m \neq 0$  such that

$$p(z) = a_0 + a_1 z + \cdots + a_m z^m$$

for all  $z \in \mathbf{F}$ . If  $p$  has degree  $m$ , we write  $\deg p = m$ .

- The polynomial that is identically 0 is said to have degree  $-\infty$ .

In the next definition, we use the convention that  $-\infty < m$ , which means that the polynomial 0 is in  $\mathcal{P}_m(\mathbf{F})$ .

### 2.13 Definition $\mathcal{P}_m(\mathbf{F})$

For  $m$  a nonnegative integer,  $\mathcal{P}_m(\mathbf{F})$  denotes the set of all polynomials with coefficients in  $\mathbf{F}$  and degree at most  $m$ .

To verify the next example, note that  $\mathcal{P}_m(\mathbf{F}) = \text{span}(1, z, \dots, z^m)$ ; here we are slightly abusing notation by letting  $z^k$  denote a function.

---

**2.14 Example**  $\mathcal{P}_m(\mathbf{F})$  is a finite-dimensional vector space for each non-negative integer  $m$ .

---

### 2.15 Definition *infinite-dimensional vector space*

A vector space is called *infinite-dimensional* if it is not finite-dimensional.

**2.16 Example** Show that  $\mathcal{P}(\mathbf{F})$  is infinite-dimensional.

**Solution** Consider any list of elements of  $\mathcal{P}(\mathbf{F})$ . Let  $m$  denote the highest degree of the polynomials in this list. Then every polynomial in the span of this list has degree at most  $m$ . Thus  $z^{m+1}$  is not in the span of our list. Hence no list spans  $\mathcal{P}(\mathbf{F})$ . Thus  $\mathcal{P}(\mathbf{F})$  is infinite-dimensional.

## Linear Independence

Suppose  $v_1, \dots, v_m \in V$  and  $v \in \text{span}(v_1, \dots, v_m)$ . By the definition of span, there exist  $a_1, \dots, a_m \in \mathbf{F}$  such that

$$v = a_1 v_1 + \cdots + a_m v_m.$$

Consider the question of whether the choice of scalars in the equation above is unique. Suppose  $c_1, \dots, c_m$  is another set of scalars such that

$$v = c_1 v_1 + \cdots + c_m v_m.$$

Subtracting the last two equations, we have

$$0 = (a_1 - c_1)v_1 + \cdots + (a_m - c_m)v_m.$$

Thus we have written 0 as a linear combination of  $(v_1, \dots, v_m)$ . If the only way to do this is the obvious way (using 0 for all scalars), then each  $a_j - c_j$  equals 0, which means that each  $a_j$  equals  $c_j$  (and thus the choice of scalars was indeed unique). This situation is so important that we give it a special name—linear independence—which we now define.

### 2.17 Definition *linearly independent*

- A list  $v_1, \dots, v_m$  of vectors in  $V$  is called ***linearly independent*** if the only choice of  $a_1, \dots, a_m \in \mathbf{F}$  that makes  $a_1 v_1 + \cdots + a_m v_m$  equal 0 is  $a_1 = \cdots = a_m = 0$ .
- The empty list () is also declared to be linearly independent.

The reasoning above shows that  $v_1, \dots, v_m$  is linearly independent if and only if each vector in  $\text{span}(v_1, \dots, v_m)$  has only one representation as a linear combination of  $v_1, \dots, v_m$ .

**2.18 Example** *linearly independent lists*

- (a) A list  $v$  of one vector  $v \in V$  is linearly independent if and only if  $v \neq 0$ .
- (b) A list of two vectors in  $V$  is linearly independent if and only if neither vector is a scalar multiple of the other.
- (c)  $(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0)$  is linearly independent in  $\mathbf{F}^4$ .
- (d) The list  $1, z, \dots, z^m$  is linearly independent in  $\mathcal{P}(\mathbf{F})$  for each nonnegative integer  $m$ .

If some vectors are removed from a linearly independent list, the remaining list is also linearly independent, as you should verify.

**2.19 Definition** *linearly dependent*

- A list of vectors in  $V$  is called ***linearly dependent*** if it is not linearly independent.
- In other words, a list  $v_1, \dots, v_m$  of vectors in  $V$  is linearly dependent if there exist  $a_1, \dots, a_m \in \mathbf{F}$ , not all 0, such that  $a_1v_1 + \dots + a_mv_m = 0$ .

**2.20 Example** *linearly dependent lists*

- $(2, 3, 1), (1, -1, 2), (7, 3, 8)$  is linearly dependent in  $\mathbf{F}^3$  because
$$2(2, 3, 1) + 3(1, -1, 2) + (-1)(7, 3, 8) = (0, 0, 0).$$
- The list  $(2, 3, 1), (1, -1, 2), (7, 3, c)$  is linearly dependent in  $\mathbf{F}^3$  if and only if  $c = 8$ , as you should verify.
- If some vector in a list of vectors in  $V$  is a linear combination of the other vectors, then the list is linearly dependent. (Proof: After writing one vector in the list as equal to a linear combination of the other vectors, move that vector to the other side of the equation, where it will be multiplied by  $-1$ .)
- Every list of vectors in  $V$  containing the 0 vector is linearly dependent. (This is a special case of the previous bullet point.)

The lemma below will often be useful. It states that given a linearly dependent list of vectors, one of the vectors is in the span of the previous ones and furthermore we can throw out that vector without changing the span of the original list.

### 2.21 Linear Dependence Lemma

Suppose  $v_1, \dots, v_m$  is a linearly dependent list in  $V$ . Then there exists  $j \in \{1, 2, \dots, m\}$  such that the following hold:

- (a)  $v_j \in \text{span}(v_1, \dots, v_{j-1})$ ;
- (b) if the  $j^{\text{th}}$  term is removed from  $v_1, \dots, v_m$ , the span of the remaining list equals  $\text{span}(v_1, \dots, v_m)$ .

**Proof** Because the list  $v_1, \dots, v_m$  is linearly dependent, there exist numbers  $a_1, \dots, a_m \in \mathbf{F}$ , not all 0, such that

$$a_1 v_1 + \cdots + a_m v_m = 0.$$

Let  $j$  be the largest element of  $\{1, \dots, m\}$  such that  $a_j \neq 0$ . Then

$$2.22 \quad v_j = -\frac{a_1}{a_j} v_1 - \cdots - \frac{a_{j-1}}{a_j} v_{j-1},$$

proving (a).

To prove (b), suppose  $u \in \text{span}(v_1, \dots, v_m)$ . Then there exist numbers  $c_1, \dots, c_m \in \mathbf{F}$  such that

$$u = c_1 v_1 + \cdots + c_m v_m.$$

In the equation above, we can replace  $v_j$  with the right side of 2.22, which shows that  $u$  is in the span of the list obtained by removing the  $j^{\text{th}}$  term from  $v_1, \dots, v_m$ . Thus (b) holds. ■

Choosing  $j = 1$  in the Linear Dependence Lemma above means that  $v_1 = 0$ , because if  $j = 1$  then condition (a) above is interpreted to mean that  $v_1 \in \text{span}()$ ; recall that  $\text{span}() = \{0\}$ . Note also that the proof of part (b) above needs to be modified in an obvious way if  $v_1 = 0$  and  $j = 1$ .

In general, the proofs in the rest of the book will not call attention to special cases that must be considered involving empty lists, lists of length 1, the subspace  $\{0\}$ , or other trivial cases for which the result is clearly true but needs a slightly different proof. Be sure to check these special cases yourself.

Now we come to a key result. It says that no linearly independent list in  $V$  is longer than a spanning list in  $V$ .

## 2.23 Length of linearly independent list $\leq$ length of spanning list

In a finite-dimensional vector space, the length of every linearly independent list of vectors is less than or equal to the length of every spanning list of vectors.

**Proof** Suppose  $u_1, \dots, u_m$  is linearly independent in  $V$ . Suppose also that  $w_1, \dots, w_n$  spans  $V$ . We need to prove that  $m \leq n$ . We do so through the multi-step process described below; note that in each step we add one of the  $u$ 's and remove one of the  $w$ 's.

### Step 1

Let  $B$  be the list  $w_1, \dots, w_n$ , which spans  $V$ . Thus adjoining any vector in  $V$  to this list produces a linearly dependent list (because the newly adjoined vector can be written as a linear combination of the other vectors). In particular, the list

$$u_1, w_1, \dots, w_n$$

is linearly dependent. Thus by the Linear Dependence Lemma (2.21), we can remove one of the  $w$ 's so that the new list  $B$  (of length  $n$ ) consisting of  $u_1$  and the remaining  $w$ 's spans  $V$ .

### Step $j$

The list  $B$  (of length  $n$ ) from step  $j - 1$  spans  $V$ . Thus adjoining any vector to this list produces a linearly dependent list. In particular, the list of length  $(n + 1)$  obtained by adjoining  $u_j$  to  $B$ , placing it just after  $u_1, \dots, u_{j-1}$ , is linearly dependent. By the Linear Dependence Lemma (2.21), one of the vectors in this list is in the span of the previous ones, and because  $u_1, \dots, u_j$  is linearly independent, this vector is one of the  $w$ 's, not one of the  $u$ 's. We can remove that  $w$  from  $B$  so that the new list  $B$  (of length  $n$ ) consisting of  $u_1, \dots, u_j$  and the remaining  $w$ 's spans  $V$ .

After step  $m$ , we have added all the  $u$ 's and the process stops. At each step as we add a  $u$  to  $B$ , the Linear Dependence Lemma implies that there is some  $w$  to remove. Thus there are at least as many  $w$ 's as  $u$ 's. ■

The next two examples show how the result above can be used to show, without any computations, that certain lists are not linearly independent and that certain lists do not span a given vector space.

**2.24 Example** Show that the list  $(1, 2, 3), (4, 5, 8), (9, 6, 7), (-3, 2, 8)$  is not linearly independent in  $\mathbf{R}^3$ .

**Solution** The list  $(1, 0, 0), (0, 1, 0), (0, 0, 1)$  spans  $\mathbf{R}^3$ . Thus no list of length larger than 3 is linearly independent in  $\mathbf{R}^3$ .

**2.25 Example** Show that the list  $(1, 2, 3, -5), (4, 5, 8, 3), (9, 6, 7, -1)$  does not span  $\mathbf{R}^4$ .

**Solution** The list  $(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)$  is linearly independent in  $\mathbf{R}^4$ . Thus no list of length less than 4 spans  $\mathbf{R}^4$ .

Our intuition suggests that every subspace of a finite-dimensional vector space should also be finite-dimensional. We now prove that this intuition is correct.

### 2.26 Finite-dimensional subspaces

Every subspace of a finite-dimensional vector space is finite-dimensional.

**Proof** Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . We need to prove that  $U$  is finite-dimensional. We do this through the following multi-step construction.

#### Step 1

If  $U = \{0\}$ , then  $U$  is finite-dimensional and we are done. If  $U \neq \{0\}$ , then choose a nonzero vector  $v_1 \in U$ .

#### Step j

If  $U = \text{span}(v_1, \dots, v_{j-1})$ , then  $U$  is finite-dimensional and we are done. If  $U \neq \text{span}(v_1, \dots, v_{j-1})$ , then choose a vector  $v_j \in U$  such that

$$v_j \notin \text{span}(v_1, \dots, v_{j-1}).$$

After each step, as long as the process continues, we have constructed a list of vectors such that no vector in this list is in the span of the previous vectors. Thus after each step we have constructed a linearly independent list, by the Linear Dependence Lemma (2.21). This linearly independent list cannot be longer than any spanning list of  $V$  (by 2.23). Thus the process eventually terminates, which means that  $U$  is finite-dimensional. ■

## EXERCISES 2.A

---

- 1** Suppose  $v_1, v_2, v_3, v_4$  spans  $V$ . Prove that the list

$$v_1 - v_2, v_2 - v_3, v_3 - v_4, v_4$$

also spans  $V$ .

- 2** Verify the assertions in Example 2.18.

- 3** Find a number  $t$  such that

$$(3, 1, 4), (2, -3, 5), (5, 9, t)$$

is not linearly independent in  $\mathbf{R}^3$ .

- 4** Verify the assertion in the second bullet point in Example 2.20.

- 5** (a) Show that if we think of  $\mathbf{C}$  as a vector space over  $\mathbf{R}$ , then the list  $(1+i, 1-i)$  is linearly independent.  
 (b) Show that if we think of  $\mathbf{C}$  as a vector space over  $\mathbf{C}$ , then the list  $(1+i, 1-i)$  is linearly dependent.

- 6** Suppose  $v_1, v_2, v_3, v_4$  is linearly independent in  $V$ . Prove that the list

$$v_1 - v_2, v_2 - v_3, v_3 - v_4, v_4$$

is also linearly independent.

- 7** Prove or give a counterexample: If  $v_1, v_2, \dots, v_m$  is a linearly independent list of vectors in  $V$ , then

$$5v_1 - 4v_2, v_2, v_3, \dots, v_m$$

is linearly independent.

- 8** Prove or give a counterexample: If  $v_1, v_2, \dots, v_m$  is a linearly independent list of vectors in  $V$  and  $\lambda \in \mathbf{F}$  with  $\lambda \neq 0$ , then  $\lambda v_1, \lambda v_2, \dots, \lambda v_m$  is linearly independent.

- 9** Prove or give a counterexample: If  $v_1, \dots, v_m$  and  $w_1, \dots, w_m$  are linearly independent lists of vectors in  $V$ , then  $v_1 + w_1, \dots, v_m + w_m$  is linearly independent.

- 10** Suppose  $v_1, \dots, v_m$  is linearly independent in  $V$  and  $w \in V$ . Prove that if  $v_1 + w, \dots, v_m + w$  is linearly dependent, then  $w \in \text{span}(v_1, \dots, v_m)$ .

- 11 Suppose  $v_1, \dots, v_m$  is linearly independent in  $V$  and  $w \in V$ . Show that  $v_1, \dots, v_m, w$  is linearly independent if and only if

$$w \notin \text{span}(v_1, \dots, v_m).$$

- 12 Explain why there does not exist a list of six polynomials that is linearly independent in  $\mathcal{P}_4(\mathbf{F})$ .
- 13 Explain why no list of four polynomials spans  $\mathcal{P}_4(\mathbf{F})$ .
- 14 Prove that  $V$  is infinite-dimensional if and only if there is a sequence  $v_1, v_2, \dots$  of vectors in  $V$  such that  $v_1, \dots, v_m$  is linearly independent for every positive integer  $m$ .
- 15 Prove that  $\mathbf{F}^\infty$  is infinite-dimensional.
- 16 Prove that the real vector space of all continuous real-valued functions on the interval  $[0, 1]$  is infinite-dimensional.
- 17 Suppose  $p_0, p_1, \dots, p_m$  are polynomials in  $\mathcal{P}_m(\mathbf{F})$  such that  $p_j(2) = 0$  for each  $j$ . Prove that  $p_0, p_1, \dots, p_m$  is not linearly independent in  $\mathcal{P}_m(\mathbf{F})$ .

## 2.B *Bases*

In the last section, we discussed linearly independent lists and spanning lists. Now we bring these concepts together.

### 2.27 Definition *basis*

A ***basis*** of  $V$  is a list of vectors in  $V$  that is linearly independent and spans  $V$ .

### 2.28 Example *bases*

- (a) The list  $(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$  is a basis of  $\mathbf{F}^n$ , called the ***standard basis*** of  $\mathbf{F}^n$ .
- (b) The list  $(1, 2), (3, 5)$  is a basis of  $\mathbf{F}^2$ .
- (c) The list  $(1, 2, -4), (7, -5, 6)$  is linearly independent in  $\mathbf{F}^3$  but is not a basis of  $\mathbf{F}^3$  because it does not span  $\mathbf{F}^3$ .
- (d) The list  $(1, 2), (3, 5), (4, 13)$  spans  $\mathbf{F}^2$  but is not a basis of  $\mathbf{F}^2$  because it is not linearly independent.
- (e) The list  $(1, 1, 0), (0, 0, 1)$  is a basis of  $\{(x, x, y) \in \mathbf{F}^3 : x, y \in \mathbf{F}\}$ .
- (f) The list  $(1, -1, 0), (1, 0, -1)$  is a basis of  

$$\{(x, y, z) \in \mathbf{F}^3 : x + y + z = 0\}.$$
- (g) The list  $1, z, \dots, z^m$  is a basis of  $\mathcal{P}_m(\mathbf{F})$ .

In addition to the standard basis,  $\mathbf{F}^n$  has many other bases. For example,  $(7, 5), (-4, 9)$  and  $(1, 2), (3, 5)$  are both bases of  $\mathbf{F}^2$ .

The next result helps explain why bases are useful. Recall that “uniquely” means “in only one way”.

### 2.29 Criterion for basis

A list  $v_1, \dots, v_n$  of vectors in  $V$  is a basis of  $V$  if and only if every  $v \in V$  can be written uniquely in the form

$$2.30 \quad v = a_1 v_1 + \cdots + a_n v_n,$$

where  $a_1, \dots, a_n \in \mathbf{F}$ .

**Proof** First suppose that  $v_1, \dots, v_n$  is a basis of  $V$ . Let  $v \in V$ . Because  $v_1, \dots, v_n$  spans  $V$ , there exist  $a_1, \dots, a_n \in \mathbf{F}$  such that 2.30 holds. To

*This proof is essentially a repetition of the ideas that led us to the definition of linear independence.*

show that the representation in 2.30 is unique, suppose  $c_1, \dots, c_n$  are scalars such that we also have

$$v = c_1 v_1 + \cdots + c_n v_n.$$

Subtracting the last equation from 2.30, we get

$$0 = (a_1 - c_1)v_1 + \cdots + (a_n - c_n)v_n.$$

This implies that each  $a_j - c_j$  equals 0 (because  $v_1, \dots, v_n$  is linearly independent). Hence  $a_1 = c_1, \dots, a_n = c_n$ . We have the desired uniqueness, completing the proof in one direction.

For the other direction, suppose every  $v \in V$  can be written uniquely in the form given by 2.30. Clearly this implies that  $v_1, \dots, v_n$  spans  $V$ . To show that  $v_1, \dots, v_n$  is linearly independent, suppose  $a_1, \dots, a_n \in \mathbf{F}$  are such that

$$0 = a_1 v_1 + \cdots + a_n v_n.$$

The uniqueness of the representation 2.30 (taking  $v = 0$ ) now implies that  $a_1 = \cdots = a_n = 0$ . Thus  $v_1, \dots, v_n$  is linearly independent and hence is a basis of  $V$ . ■

A spanning list in a vector space may not be a basis because it is not linearly independent. Our next result says that given any spanning list, some (possibly none) of the vectors in it can be discarded so that the remaining list is linearly independent and still spans the vector space.

As an example in the vector space  $\mathbf{F}^2$ , if the procedure in the proof below is applied to the list  $(1, 2), (3, 6), (4, 7), (5, 9)$ , then the second and fourth vectors will be removed. This leaves  $(1, 2), (4, 7)$ , which is a basis of  $\mathbf{F}^2$ .

### 2.31 Spanning list contains a basis

Every spanning list in a vector space can be reduced to a basis of the vector space.

**Proof** Suppose  $v_1, \dots, v_n$  spans  $V$ . We want to remove some of the vectors from  $v_1, \dots, v_n$  so that the remaining vectors form a basis of  $V$ . We do this through the multi-step process described below.

Start with  $B$  equal to the list  $v_1, \dots, v_n$ .

**Step 1**

If  $v_1 = 0$ , delete  $v_1$  from  $B$ . If  $v_1 \neq 0$ , leave  $B$  unchanged.

**Step j**

If  $v_j$  is in  $\text{span}(v_1, \dots, v_{j-1})$ , delete  $v_j$  from  $B$ . If  $v_j$  is not in  $\text{span}(v_1, \dots, v_{j-1})$ , leave  $B$  unchanged.

Stop the process after step  $n$ , getting a list  $B$ . This list  $B$  spans  $V$  because our original list spanned  $V$  and we have discarded only vectors that were already in the span of the previous vectors. The process ensures that no vector in  $B$  is in the span of the previous ones. Thus  $B$  is linearly independent, by the Linear Dependence Lemma (2.21). Hence  $B$  is a basis of  $V$ . ■

Our next result, an easy corollary of the previous result, tells us that every finite-dimensional vector space has a basis.

### 2.32 Basis of finite-dimensional vector space

Every finite-dimensional vector space has a basis.

**Proof** By definition, a finite-dimensional vector space has a spanning list. The previous result tells us that each spanning list can be reduced to a basis. ■

Our next result is in some sense a dual of 2.31, which said that every spanning list can be reduced to a basis. Now we show that given any linearly independent list, we can adjoin some additional vectors (this includes the possibility of adjoining no additional vectors) so that the extended list is still linearly independent but also spans the space.

### 2.33 Linearly independent list extends to a basis

Every linearly independent list of vectors in a finite-dimensional vector space can be extended to a basis of the vector space.

**Proof** Suppose  $u_1, \dots, u_m$  is linearly independent in a finite-dimensional vector space  $V$ . Let  $w_1, \dots, w_n$  be a basis of  $V$ . Thus the list

$$u_1, \dots, u_m, w_1, \dots, w_n$$

spans  $V$ . Applying the procedure of the proof of 2.31 to reduce this list to a basis of  $V$  produces a basis consisting of the vectors  $u_1, \dots, u_m$  (none of the  $u$ 's get deleted in this procedure because  $u_1, \dots, u_m$  is linearly independent) and some of the  $w$ 's. ■

As an example in  $\mathbf{F}^3$ , suppose we start with the linearly independent list  $(2, 3, 4), (9, 6, 8)$ . If we take  $w_1, w_2, w_3$  in the proof above to be the standard basis of  $\mathbf{F}^3$ , then the procedure in the proof above produces the list  $(2, 3, 4), (9, 6, 8), (0, 1, 0)$ , which is a basis of  $\mathbf{F}^3$ .

*Using the same basic ideas but considerably more advanced tools, the next result can be proved without the hypothesis that  $V$  is finite-dimensional.*

As an application of the result above, we now show that every subspace of a finite-dimensional vector space can be paired with another subspace to form a direct sum of the whole space.

### 2.34 Every subspace of $V$ is part of a direct sum equal to $V$

Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Then there is a subspace  $W$  of  $V$  such that  $V = U \oplus W$ .

**Proof** Because  $V$  is finite-dimensional, so is  $U$  (see 2.26). Thus there is a basis  $u_1, \dots, u_m$  of  $U$  (see 2.32). Of course  $u_1, \dots, u_m$  is a linearly independent list of vectors in  $V$ . Hence this list can be extended to a basis  $u_1, \dots, u_m, w_1, \dots, w_n$  of  $V$  (see 2.33). Let  $W = \text{span}(w_1, \dots, w_n)$ .

To prove that  $V = U \oplus W$ , by 1.45 we need only show that

$$V = U + W \quad \text{and} \quad U \cap W = \{0\}.$$

To prove the first equation above, suppose  $v \in V$ . Then, because the list  $u_1, \dots, u_m, w_1, \dots, w_n$  spans  $V$ , there exist  $a_1, \dots, a_m, b_1, \dots, b_n \in \mathbf{F}$  such that

$$v = \underbrace{a_1 u_1 + \cdots + a_m u_m}_u + \underbrace{b_1 w_1 + \cdots + b_n w_n}_w.$$

In other words, we have  $v = u + w$ , where  $u \in U$  and  $w \in W$  are defined as above. Thus  $v \in U + W$ , completing the proof that  $V = U + W$ .

To show that  $U \cap W = \{0\}$ , suppose  $v \in U \cap W$ . Then there exist scalars  $a_1, \dots, a_m, b_1, \dots, b_n \in \mathbf{F}$  such that

$$v = a_1 u_1 + \cdots + a_m u_m = b_1 w_1 + \cdots + b_n w_n.$$

Thus

$$a_1 u_1 + \cdots + a_m u_m - b_1 w_1 - \cdots - b_n w_n = 0.$$

Because  $u_1, \dots, u_m, w_1, \dots, w_n$  is linearly independent, this implies that  $a_1 = \cdots = a_m = b_1 = \cdots = b_n = 0$ . Thus  $v = 0$ , completing the proof that  $U \cap W = \{0\}$ . ■

## EXERCISES 2.B

---

**1** Find all vector spaces that have exactly one basis.

**2** Verify all the assertions in Example 2.28.

**3** (a) Let  $U$  be the subspace of  $\mathbf{R}^5$  defined by

$$U = \{(x_1, x_2, x_3, x_4, x_5) \in \mathbf{R}^5 : x_1 = 3x_2 \text{ and } x_3 = 7x_4\}.$$

Find a basis of  $U$ .

(b) Extend the basis in part (a) to a basis of  $\mathbf{R}^5$ .

(c) Find a subspace  $W$  of  $\mathbf{R}^5$  such that  $\mathbf{R}^5 = U \oplus W$ .

**4** (a) Let  $U$  be the subspace of  $\mathbf{C}^5$  defined by

$$U = \{(z_1, z_2, z_3, z_4, z_5) \in \mathbf{C}^5 : 6z_1 = z_2 \text{ and } z_3 + 2z_4 + 3z_5 = 0\}.$$

Find a basis of  $U$ .

(b) Extend the basis in part (a) to a basis of  $\mathbf{C}^5$ .

(c) Find a subspace  $W$  of  $\mathbf{C}^5$  such that  $\mathbf{C}^5 = U \oplus W$ .

**5** Prove or disprove: there exists a basis  $p_0, p_1, p_2, p_3$  of  $\mathcal{P}_3(\mathbf{F})$  such that none of the polynomials  $p_0, p_1, p_2, p_3$  has degree 2.

**6** Suppose  $v_1, v_2, v_3, v_4$  is a basis of  $V$ . Prove that

$$v_1 + v_2, v_2 + v_3, v_3 + v_4, v_4$$

is also a basis of  $V$ .

**7** Prove or give a counterexample: If  $v_1, v_2, v_3, v_4$  is a basis of  $V$  and  $U$  is a subspace of  $V$  such that  $v_1, v_2 \in U$  and  $v_3 \notin U$  and  $v_4 \notin U$ , then  $v_1, v_2$  is a basis of  $U$ .

**8** Suppose  $U$  and  $W$  are subspaces of  $V$  such that  $V = U \oplus W$ . Suppose also that  $u_1, \dots, u_m$  is a basis of  $U$  and  $w_1, \dots, w_n$  is a basis of  $W$ . Prove that

$$u_1, \dots, u_m, w_1, \dots, w_n$$

is a basis of  $V$ .

## 2.C Dimension

Although we have been discussing finite-dimensional vector spaces, we have not yet defined the dimension of such an object. How should dimension be defined? A reasonable definition should force the dimension of  $\mathbf{F}^n$  to equal  $n$ . Notice that the standard basis

$$(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$$

of  $\mathbf{F}^n$  has length  $n$ . Thus we are tempted to define the dimension as the length of a basis. However, a finite-dimensional vector space in general has many different bases, and our attempted definition makes sense only if all bases in a given vector space have the same length. Fortunately that turns out to be the case, as we now show.

### 2.35 Basis length does not depend on basis

Any two bases of a finite-dimensional vector space have the same length.

**Proof** Suppose  $V$  is finite-dimensional. Let  $B_1$  and  $B_2$  be two bases of  $V$ . Then  $B_1$  is linearly independent in  $V$  and  $B_2$  spans  $V$ , so the length of  $B_1$  is at most the length of  $B_2$  (by 2.23). Interchanging the roles of  $B_1$  and  $B_2$ , we also see that the length of  $B_2$  is at most the length of  $B_1$ . Thus the length of  $B_1$  equals the length of  $B_2$ , as desired. ■

Now that we know that any two bases of a finite-dimensional vector space have the same length, we can formally define the dimension of such spaces.

### 2.36 Definition *dimension*, $\dim V$

- The **dimension** of a finite-dimensional vector space is the length of any basis of the vector space.
- The dimension of  $V$  (if  $V$  is finite-dimensional) is denoted by  $\dim V$ .

### 2.37 Example *dimensions*

- $\dim \mathbf{F}^n = n$  because the standard basis of  $\mathbf{F}^n$  has length  $n$ .
- $\dim \mathcal{P}_m(\mathbf{F}) = m + 1$  because the basis  $1, z, \dots, z^m$  of  $\mathcal{P}_m(\mathbf{F})$  has length  $m + 1$ .

Every subspace of a finite-dimensional vector space is finite-dimensional (by 2.26) and so has a dimension. The next result gives the expected inequality about the dimension of a subspace.

### 2.38 Dimension of a subspace

If  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ , then  $\dim U \leq \dim V$ .

**Proof** Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Think of a basis of  $U$  as a linearly independent list in  $V$ , and think of a basis of  $V$  as a spanning list in  $V$ . Now use 2.23 to conclude that  $\dim U \leq \dim V$ . ■

To check that a list of vectors in  $V$  is a basis of  $V$ , we must, according to the definition, show that the list in question satisfies two properties: it must be linearly independent and it must span  $V$ . The next two results show that if the list in question has the right length, then we need only check that it satisfies one of the two required properties. First we prove that every linearly independent list with the right length is a basis.

The real vector space  $\mathbf{R}^2$  has dimension 2; the complex vector space  $\mathbf{C}$  has dimension 1. As sets,  $\mathbf{R}^2$  can be identified with  $\mathbf{C}$  (and addition is the same on both spaces, as is scalar multiplication by real numbers). Thus when we talk about the dimension of a vector space, the role played by the choice of  $\mathbf{F}$  cannot be neglected.

### 2.39 Linearly independent list of the right length is a basis

Suppose  $V$  is finite-dimensional. Then every linearly independent list of vectors in  $V$  with length  $\dim V$  is a basis of  $V$ .

**Proof** Suppose  $\dim V = n$  and  $v_1, \dots, v_n$  is linearly independent in  $V$ . The list  $v_1, \dots, v_n$  can be extended to a basis of  $V$  (by 2.33). However, every basis of  $V$  has length  $n$ , so in this case the extension is the trivial one, meaning that no elements are adjoined to  $v_1, \dots, v_n$ . In other words,  $v_1, \dots, v_n$  is a basis of  $V$ , as desired. ■

**2.40 Example** Show that the list  $(5, 7), (4, 3)$  is a basis of  $\mathbf{F}^2$ .

**Solution** This list of two vectors in  $\mathbf{F}^2$  is obviously linearly independent (because neither vector is a scalar multiple of the other). Note that  $\mathbf{F}^2$  has dimension 2. Thus 2.39 implies that the linearly independent list  $(5, 7), (4, 3)$  of length 2 is a basis of  $\mathbf{F}^2$  (we do not need to bother checking that it spans  $\mathbf{F}^2$ ).

**2.41 Example** Show that  $1, (x - 5)^2, (x - 5)^3$  is a basis of the subspace  $U$  of  $\mathcal{P}_3(\mathbf{R})$  defined by

$$U = \{p \in \mathcal{P}_3(\mathbf{R}) : p'(5) = 0\}.$$

**Solution** Clearly each of the polynomials  $1, (x - 5)^2$ , and  $(x - 5)^3$  is in  $U$ . Suppose  $a, b, c \in \mathbf{R}$  and

$$a + b(x - 5)^2 + c(x - 5)^3 = 0$$

for every  $x \in \mathbf{R}$ . Without explicitly expanding the left side of the equation above, we can see that the left side has a  $cx^3$  term. Because the right side has no  $x^3$  term, this implies that  $c = 0$ . Because  $c = 0$ , we see that the left side has a  $bx^2$  term, which implies that  $b = 0$ . Because  $b = c = 0$ , we can also conclude that  $a = 0$ .

Thus the equation above implies that  $a = b = c = 0$ . Hence the list  $1, (x - 5)^2, (x - 5)^3$  is linearly independent in  $U$ .

Thus  $\dim U \geq 3$ . Because  $U$  is a subspace of  $\mathcal{P}_3(\mathbf{R})$ , we know that  $\dim U \leq \dim \mathcal{P}_3(\mathbf{R}) = 4$  (by 2.38). However,  $\dim U$  cannot equal 4, because otherwise when we extend a basis of  $U$  to a basis of  $\mathcal{P}_3(\mathbf{R})$  we would get a list with length greater than 4. Hence  $\dim U = 3$ . Thus 2.39 implies that the linearly independent list  $1, (x - 5)^2, (x - 5)^3$  is a basis of  $U$ .

Now we prove that a spanning list with the right length is a basis.

#### 2.42 Spanning list of the right length is a basis

Suppose  $V$  is finite-dimensional. Then every spanning list of vectors in  $V$  with length  $\dim V$  is a basis of  $V$ .

**Proof** Suppose  $\dim V = n$  and  $v_1, \dots, v_n$  spans  $V$ . The list  $v_1, \dots, v_n$  can be reduced to a basis of  $V$  (by 2.31). However, every basis of  $V$  has length  $n$ , so in this case the reduction is the trivial one, meaning that no elements are deleted from  $v_1, \dots, v_n$ . In other words,  $v_1, \dots, v_n$  is a basis of  $V$ , as desired. ■

The next result gives a formula for the dimension of the sum of two subspaces of a finite-dimensional vector space. This formula is analogous to a familiar counting formula: the number of elements in the union of two finite sets equals the number of elements in the first set, plus the number of elements in the second set, minus the number of elements in the intersection of the two sets.

### 2.43 Dimension of a sum

If  $U_1$  and  $U_2$  are subspaces of a finite-dimensional vector space, then

$$\dim(U_1 + U_2) = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2).$$

**Proof** Let  $u_1, \dots, u_m$  be a basis of  $U_1 \cap U_2$ ; thus  $\dim(U_1 \cap U_2) = m$ . Because  $u_1, \dots, u_m$  is a basis of  $U_1 \cap U_2$ , it is linearly independent in  $U_1$ . Hence this list can be extended to a basis  $u_1, \dots, u_m, v_1, \dots, v_j$  of  $U_1$  (by 2.33). Thus  $\dim U_1 = m + j$ . Also extend  $u_1, \dots, u_m$  to a basis  $u_1, \dots, u_m, w_1, \dots, w_k$  of  $U_2$ ; thus  $\dim U_2 = m + k$ .

We will show that

$$u_1, \dots, u_m, v_1, \dots, v_j, w_1, \dots, w_k$$

is a basis of  $U_1 + U_2$ . This will complete the proof, because then we will have

$$\begin{aligned}\dim(U_1 + U_2) &= m + j + k \\ &= (m + j) + (m + k) - m \\ &= \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2).\end{aligned}$$

Clearly  $\text{span}(u_1, \dots, u_m, v_1, \dots, v_j, w_1, \dots, w_k)$  contains  $U_1$  and  $U_2$  and hence equals  $U_1 + U_2$ . So to show that this list is a basis of  $U_1 + U_2$  we need only show that it is linearly independent. To prove this, suppose

$$a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_jv_j + c_1w_1 + \cdots + c_kw_k = 0,$$

where all the  $a$ 's,  $b$ 's, and  $c$ 's are scalars. We need to prove that all the  $a$ 's,  $b$ 's, and  $c$ 's equal 0. The equation above can be rewritten as

$$c_1w_1 + \cdots + c_kw_k = -a_1u_1 - \cdots - a_mu_m - b_1v_1 - \cdots - b_jv_j,$$

which shows that  $c_1w_1 + \cdots + c_kw_k \in U_1$ . All the  $w$ 's are in  $U_2$ , so this implies that  $c_1w_1 + \cdots + c_kw_k \in U_1 \cap U_2$ . Because  $u_1, \dots, u_m$  is a basis of  $U_1 \cap U_2$ , we can write

$$c_1w_1 + \cdots + c_kw_k = d_1u_1 + \cdots + d_mu_m$$

for some choice of scalars  $d_1, \dots, d_m$ . But  $u_1, \dots, u_m, w_1, \dots, w_k$  is linearly independent, so the last equation implies that all the  $c$ 's (and  $d$ 's) equal 0. Thus our original equation involving the  $a$ 's,  $b$ 's, and  $c$ 's becomes

$$a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_jv_j = 0.$$

Because the list  $u_1, \dots, u_m, v_1, \dots, v_j$  is linearly independent, this equation implies that all the  $a$ 's and  $b$ 's are 0. We now know that all the  $a$ 's,  $b$ 's, and  $c$ 's equal 0, as desired. ■

## EXERCISES 2.C

---

- 1** Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$  such that  $\dim U = \dim V$ . Prove that  $U = V$ .
- 2** Show that the subspaces of  $\mathbf{R}^2$  are precisely  $\{0\}$ ,  $\mathbf{R}^2$ , and all lines in  $\mathbf{R}^2$  through the origin.
- 3** Show that the subspaces of  $\mathbf{R}^3$  are precisely  $\{0\}$ ,  $\mathbf{R}^3$ , all lines in  $\mathbf{R}^3$  through the origin, and all planes in  $\mathbf{R}^3$  through the origin.
- 4** (a) Let  $U = \{p \in \mathcal{P}_4(\mathbf{F}) : p(6) = 0\}$ . Find a basis of  $U$ .  
 (b) Extend the basis in part (a) to a basis of  $\mathcal{P}_4(\mathbf{F})$ .  
 (c) Find a subspace  $W$  of  $\mathcal{P}_4(\mathbf{F})$  such that  $\mathcal{P}_4(\mathbf{F}) = U \oplus W$ .
- 5** (a) Let  $U = \{p \in \mathcal{P}_4(\mathbf{R}) : p''(6) = 0\}$ . Find a basis of  $U$ .  
 (b) Extend the basis in part (a) to a basis of  $\mathcal{P}_4(\mathbf{R})$ .  
 (c) Find a subspace  $W$  of  $\mathcal{P}_4(\mathbf{R})$  such that  $\mathcal{P}_4(\mathbf{R}) = U \oplus W$ .
- 6** (a) Let  $U = \{p \in \mathcal{P}_4(\mathbf{F}) : p(2) = p(5)\}$ . Find a basis of  $U$ .  
 (b) Extend the basis in part (a) to a basis of  $\mathcal{P}_4(\mathbf{F})$ .  
 (c) Find a subspace  $W$  of  $\mathcal{P}_4(\mathbf{F})$  such that  $\mathcal{P}_4(\mathbf{F}) = U \oplus W$ .
- 7** (a) Let  $U = \{p \in \mathcal{P}_4(\mathbf{F}) : p(2) = p(5) = p(6)\}$ . Find a basis of  $U$ .  
 (b) Extend the basis in part (a) to a basis of  $\mathcal{P}_4(\mathbf{F})$ .  
 (c) Find a subspace  $W$  of  $\mathcal{P}_4(\mathbf{F})$  such that  $\mathcal{P}_4(\mathbf{F}) = U \oplus W$ .
- 8** (a) Let  $U = \{p \in \mathcal{P}_4(\mathbf{R}) : \int_{-1}^1 p = 0\}$ . Find a basis of  $U$ .  
 (b) Extend the basis in part (a) to a basis of  $\mathcal{P}_4(\mathbf{R})$ .  
 (c) Find a subspace  $W$  of  $\mathcal{P}_4(\mathbf{R})$  such that  $\mathcal{P}_4(\mathbf{R}) = U \oplus W$ .
- 9** Suppose  $v_1, \dots, v_m$  is linearly independent in  $V$  and  $w \in V$ . Prove that

$$\dim \text{span}(v_1 + w, \dots, v_m + w) \geq m - 1.$$

- 10** Suppose  $p_0, p_1, \dots, p_m \in \mathcal{P}(\mathbf{F})$  are such that each  $p_j$  has degree  $j$ . Prove that  $p_0, p_1, \dots, p_m$  is a basis of  $\mathcal{P}_m(\mathbf{F})$ .
- 11** Suppose that  $U$  and  $W$  are subspaces of  $\mathbf{R}^8$  such that  $\dim U = 3$ ,  $\dim W = 5$ , and  $U + W = \mathbf{R}^8$ . Prove that  $\mathbf{R}^8 = U \oplus W$ .

- 12** Suppose  $U$  and  $W$  are both five-dimensional subspaces of  $\mathbf{R}^9$ . Prove that  $U \cap W \neq \{0\}$ .
- 13** Suppose  $U$  and  $W$  are both 4-dimensional subspaces of  $\mathbf{C}^6$ . Prove that there exist two vectors in  $U \cap W$  such that neither of these vectors is a scalar multiple of the other.
- 14** Suppose  $U_1, \dots, U_m$  are finite-dimensional subspaces of  $V$ . Prove that  $U_1 + \dots + U_m$  is finite-dimensional and

$$\dim(U_1 + \dots + U_m) \leq \dim U_1 + \dots + \dim U_m.$$

- 15** Suppose  $V$  is finite-dimensional, with  $\dim V = n \geq 1$ . Prove that there exist 1-dimensional subspaces  $U_1, \dots, U_n$  of  $V$  such that

$$V = U_1 \oplus \dots \oplus U_n.$$

- 16** Suppose  $U_1, \dots, U_m$  are finite-dimensional subspaces of  $V$  such that  $U_1 + \dots + U_m$  is a direct sum. Prove that  $U_1 \oplus \dots \oplus U_m$  is finite-dimensional and

$$\dim U_1 \oplus \dots \oplus U_m = \dim U_1 + \dots + \dim U_m.$$

[The exercise above deepens the analogy between direct sums of subspaces and disjoint unions of subsets. Specifically, compare this exercise to the following obvious statement: if a set is written as a disjoint union of finite subsets, then the number of elements in the set equals the sum of the numbers of elements in the disjoint subsets.]

- 17** You might guess, by analogy with the formula for the number of elements in the union of three subsets of a finite set, that if  $U_1, U_2, U_3$  are subspaces of a finite-dimensional vector space, then

$$\begin{aligned} \dim(U_1 + U_2 + U_3) \\ = \dim U_1 + \dim U_2 + \dim U_3 \\ - \dim(U_1 \cap U_2) - \dim(U_1 \cap U_3) - \dim(U_2 \cap U_3) \\ + \dim(U_1 \cap U_2 \cap U_3). \end{aligned}$$

Prove this or give a counterexample.



*German mathematician Carl Friedrich Gauss (1777–1855), who in 1809 published a method for solving systems of linear equations. This method, now called Gaussian elimination, was also used in a Chinese book published over 1600 years earlier.*

## Linear Maps

So far our attention has focused on vector spaces. No one gets excited about vector spaces. The interesting part of linear algebra is the subject to which we now turn—linear maps.

In this chapter we will frequently need another vector space, which we will call  $W$ , in addition to  $V$ . Thus our standing assumptions are now as follows:

### 3.1 Notation $\mathbf{F}, V, W$

- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  and  $W$  denote vector spaces over  $\mathbf{F}$ .

#### LEARNING OBJECTIVES FOR THIS CHAPTER

- Fundamental Theorem of Linear Maps
- the matrix of a linear map with respect to given bases
- isomorphic vector spaces
- product spaces
- quotient spaces
- the dual space of a vector space and the dual of a linear map

## 3.A The Vector Space of Linear Maps

### Definition and Examples of Linear Maps

Now we are ready for one of the key definitions in linear algebra.

#### 3.2 Definition *linear map*

A *linear map* from  $V$  to  $W$  is a function  $T: V \rightarrow W$  with the following properties:

##### **additivity**

$$T(u + v) = Tu + Tv \text{ for all } u, v \in V;$$

##### **homogeneity**

$$T(\lambda v) = \lambda(Tv) \text{ for all } \lambda \in \mathbf{F} \text{ and all } v \in V.$$

Some mathematicians use the term **linear transformation**, which means the same as linear map.

Note that for linear maps we often use the notation  $Tv$  as well as the more standard functional notation  $T(v)$ .

#### 3.3 Notation $\mathcal{L}(V, W)$

The set of all linear maps from  $V$  to  $W$  is denoted  $\mathcal{L}(V, W)$ .

Let's look at some examples of linear maps. Make sure you verify that each of the functions defined below is indeed a linear map:

---

#### 3.4 Example *linear maps*

##### **zero**

In addition to its other uses, we let the symbol  $0$  denote the function that takes each element of some vector space to the additive identity of another vector space. To be specific,  $0 \in \mathcal{L}(V, W)$  is defined by

$$0v = 0.$$

The  $0$  on the left side of the equation above is a function from  $V$  to  $W$ , whereas the  $0$  on the right side is the additive identity in  $W$ . As usual, the context should allow you to distinguish between the many uses of the symbol  $0$ .

##### **identity**

The *identity map*, denoted  $I$ , is the function on some vector space that takes each element to itself. To be specific,  $I \in \mathcal{L}(V, V)$  is defined by

$$Iv = v.$$

**differentiation**

Define  $D \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  by

$$Dp = p'.$$

The assertion that this function is a linear map is another way of stating a basic result about differentiation:  $(f + g)' = f' + g'$  and  $(\lambda f)' = \lambda f'$  whenever  $f, g$  are differentiable and  $\lambda$  is a constant.

**integration**

Define  $T \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathbf{R})$  by

$$Tp = \int_0^1 p(x) dx.$$

The assertion that this function is linear is another way of stating a basic result about integration: the integral of the sum of two functions equals the sum of the integrals, and the integral of a constant times a function equals the constant times the integral of the function.

**multiplication by  $x^2$** 

Define  $T \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  by

$$(Tp)(x) = x^2 p(x)$$

for  $x \in \mathbf{R}$ .

**backward shift**

Recall that  $\mathbf{F}^\infty$  denotes the vector space of all sequences of elements of  $\mathbf{F}$ .

Define  $T \in \mathcal{L}(\mathbf{F}^\infty, \mathbf{F}^\infty)$  by

$$T(x_1, x_2, x_3, \dots) = (x_2, x_3, \dots).$$

**from  $\mathbf{R}^3$  to  $\mathbf{R}^2$** 

Define  $T \in \mathcal{L}(\mathbf{R}^3, \mathbf{R}^2)$  by

$$T(x, y, z) = (2x - y + 3z, 7x + 5y - 6z).$$

**from  $\mathbf{F}^n$  to  $\mathbf{F}^m$** 

Generalizing the previous example, let  $m$  and  $n$  be positive integers, let  $A_{j,k} \in \mathbf{F}$  for  $j = 1, \dots, m$  and  $k = 1, \dots, n$ , and define  $T \in \mathcal{L}(\mathbf{F}^n, \mathbf{F}^m)$  by

$$T(x_1, \dots, x_n) = (A_{1,1}x_1 + \dots + A_{1,n}x_n, \dots, A_{m,1}x_1 + \dots + A_{m,n}x_n).$$

Actually every linear map from  $\mathbf{F}^n$  to  $\mathbf{F}^m$  is of this form.

The existence part of the next result means that we can find a linear map that takes on whatever values we wish on the vectors in a basis. The uniqueness part of the next result means that a linear map is completely determined by its values on a basis.

### 3.5 Linear maps and basis of domain

Suppose  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_n \in W$ . Then there exists a unique linear map  $T: V \rightarrow W$  such that

$$Tv_j = w_j$$

for each  $j = 1, \dots, n$ .

**Proof** First we show the existence of a linear map  $T$  with the desired property. Define  $T: V \rightarrow W$  by

$$T(c_1v_1 + \dots + c_nv_n) = c_1w_1 + \dots + c_nw_n,$$

where  $c_1, \dots, c_n$  are arbitrary elements of  $\mathbf{F}$ . The list  $v_1, \dots, v_n$  is a basis of  $V$ , and thus the equation above does indeed define a function  $T$  from  $V$  to  $W$  (because each element of  $V$  can be uniquely written in the form  $c_1v_1 + \dots + c_nv_n$ ).

For each  $j$ , taking  $c_j = 1$  and the other  $c$ 's equal to 0 in the equation above shows that  $Tv_j = w_j$ .

If  $u, v \in V$  with  $u = a_1v_1 + \dots + a_nv_n$  and  $v = c_1v_1 + \dots + c_nv_n$ , then

$$\begin{aligned} T(u+v) &= T((a_1+c_1)v_1 + \dots + (a_n+c_n)v_n) \\ &= (a_1+c_1)w_1 + \dots + (a_n+c_n)w_n \\ &= (a_1w_1 + \dots + a_nw_n) + (c_1w_1 + \dots + c_nw_n) \\ &= Tu + Tv. \end{aligned}$$

Similarly, if  $\lambda \in \mathbf{F}$  and  $v = c_1v_1 + \dots + c_nv_n$ , then

$$\begin{aligned} T(\lambda v) &= T(\lambda c_1v_1 + \dots + \lambda c_nv_n) \\ &= \lambda c_1w_1 + \dots + \lambda c_nw_n \\ &= \lambda(c_1w_1 + \dots + c_nw_n) \\ &= \lambda Tv. \end{aligned}$$

Thus  $T$  is a linear map from  $V$  to  $W$ .

To prove uniqueness, now suppose that  $T \in \mathcal{L}(V, W)$  and that  $Tv_j = w_j$  for  $j = 1, \dots, n$ . Let  $c_1, \dots, c_n \in \mathbf{F}$ . The homogeneity of  $T$  implies that  $T(c_jv_j) = c_jw_j$  for  $j = 1, \dots, n$ . The additivity of  $T$  now implies that

$$T(c_1v_1 + \dots + c_nv_n) = c_1w_1 + \dots + c_nw_n.$$

Thus  $T$  is uniquely determined on  $\text{span}(v_1, \dots, v_n)$  by the equation above. Because  $v_1, \dots, v_n$  is a basis of  $V$ , this implies that  $T$  is uniquely determined on  $V$ . ■

## Algebraic Operations on $\mathcal{L}(V, W)$

We begin by defining addition and scalar multiplication on  $\mathcal{L}(V, W)$ .

### 3.6 Definition addition and scalar multiplication on $\mathcal{L}(V, W)$

Suppose  $S, T \in \mathcal{L}(V, W)$  and  $\lambda \in \mathbf{F}$ . The **sum**  $S + T$  and the **product**  $\lambda T$  are the linear maps from  $V$  to  $W$  defined by

$$(S + T)(v) = Sv + Tv \quad \text{and} \quad (\lambda T)(v) = \lambda(Tv)$$

for all  $v \in V$ .

You should verify that  $S + T$  and  $\lambda T$  as defined above are indeed linear maps. In other words, if  $S, T \in \mathcal{L}(V, W)$  and  $\lambda \in \mathbf{F}$ , then  $S + T \in \mathcal{L}(V, W)$  and  $\lambda T \in \mathcal{L}(V, W)$ .

Because we took the trouble to define addition and scalar multiplication on  $\mathcal{L}(V, W)$ , the next result should not be a surprise.

*Although linear maps are pervasive throughout mathematics, they are not as ubiquitous as imagined by some confused students who seem to think that cos is a linear map from  $\mathbf{R}$  to  $\mathbf{R}$  when they write that  $\cos 2x$  equals  $2 \cos x$  and that  $\cos(x + y)$  equals  $\cos x + \cos y$ .*

### 3.7 $\mathcal{L}(V, W)$ is a vector space

With the operations of addition and scalar multiplication as defined above,  $\mathcal{L}(V, W)$  is a vector space.

The routine proof of the result above is left to the reader. Note that the additive identity of  $\mathcal{L}(V, W)$  is the zero linear map defined earlier in this section.

Usually it makes no sense to multiply together two elements of a vector space, but for some pairs of linear maps a useful product exists. We will need a third vector space, so for the rest of this section suppose  $U$  is a vector space over  $\mathbf{F}$ .

### 3.8 Definition Product of Linear Maps

If  $T \in \mathcal{L}(U, V)$  and  $S \in \mathcal{L}(V, W)$ , then the **product**  $ST \in \mathcal{L}(U, W)$  is defined by

$$(ST)(u) = S(Tu)$$

for  $u \in U$ .

In other words,  $ST$  is just the usual composition  $S \circ T$  of two functions, but when both functions are linear, most mathematicians write  $ST$  instead of  $S \circ T$ . You should verify that  $ST$  is indeed a linear map from  $U$  to  $W$  whenever  $T \in \mathcal{L}(U, V)$  and  $S \in \mathcal{L}(V, W)$ .

Note that  $ST$  is defined only when  $T$  maps into the domain of  $S$ .

### 3.9 Algebraic properties of products of linear maps

#### associativity

$$(T_1 T_2) T_3 = T_1 (T_2 T_3)$$

whenever  $T_1$ ,  $T_2$ , and  $T_3$  are linear maps such that the products make sense (meaning that  $T_3$  maps into the domain of  $T_2$ , and  $T_2$  maps into the domain of  $T_1$ ).

#### identity

$$TI = IT = T$$

whenever  $T \in \mathcal{L}(V, W)$  (the first  $I$  is the identity map on  $V$ , and the second  $I$  is the identity map on  $W$ ).

#### distributive properties

$$(S_1 + S_2)T = S_1 T + S_2 T \quad \text{and} \quad S(T_1 + T_2) = ST_1 + ST_2$$

whenever  $T, T_1, T_2 \in \mathcal{L}(U, V)$  and  $S, S_1, S_2 \in \mathcal{L}(V, W)$ .

The routine proof of the result above is left to the reader.

Multiplication of linear maps is not commutative. In other words, it is not necessarily true that  $ST = TS$ , even if both sides of the equation make sense.

---

**3.10 Example** Suppose  $D \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  is the differentiation map defined in Example 3.4 and  $T \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  is the multiplication by  $x^2$  map defined earlier in this section. Show that  $TD \neq DT$ .

**Solution** We have

$$((TD)p)(x) = x^2 p'(x) \quad \text{but} \quad ((DT)p)(x) = x^2 p'(x) + 2xp(x).$$

In other words, differentiating and then multiplying by  $x^2$  is not the same as multiplying by  $x^2$  and then differentiating.

---

### 3.11 Linear maps take 0 to 0

Suppose  $T$  is a linear map from  $V$  to  $W$ . Then  $T(0) = 0$ .

**Proof** By additivity, we have

$$T(0) = T(0 + 0) = T(0) + T(0).$$

Add the additive inverse of  $T(0)$  to each side of the equation above to conclude that  $T(0) = 0$ . ■

## EXERCISES 3.A

---

- 1 Suppose  $b, c \in \mathbf{R}$ . Define  $T: \mathbf{R}^3 \rightarrow \mathbf{R}^2$  by

$$T(x, y, z) = (2x - 4y + 3z + b, 6x + cxyz).$$

Show that  $T$  is linear if and only if  $b = c = 0$ .

- 2 Suppose  $b, c \in \mathbf{R}$ . Define  $T: \mathcal{P}(\mathbf{R}) \rightarrow \mathbf{R}^2$  by

$$Tp = \left( 3p(4) + 5p'(6) + bp(1)p(2), \int_{-1}^2 x^3 p(x) dx + c \sin p(0) \right).$$

Show that  $T$  is linear if and only if  $b = c = 0$ .

- 3 Suppose  $T \in \mathcal{L}(\mathbf{F}^n, \mathbf{F}^m)$ . Show that there exist scalars  $A_{j,k} \in \mathbf{F}$  for  $j = 1, \dots, m$  and  $k = 1, \dots, n$  such that

$$T(x_1, \dots, x_n) = (A_{1,1}x_1 + \dots + A_{1,n}x_n, \dots, A_{m,1}x_1 + \dots + A_{m,n}x_n)$$

for every  $(x_1, \dots, x_n) \in \mathbf{F}^n$ .

[The exercise above shows that  $T$  has the form promised in the last item of Example 3.4.]

- 4 Suppose  $T \in \mathcal{L}(V, W)$  and  $v_1, \dots, v_m$  is a list of vectors in  $V$  such that  $Tv_1, \dots, Tv_m$  is a linearly independent list in  $W$ . Prove that  $v_1, \dots, v_m$  is linearly independent.
- 5 Prove the assertion in 3.7.
- 6 Prove the assertions in 3.9.

- 7 Show that every linear map from a 1-dimensional vector space to itself is multiplication by some scalar. More precisely, prove that if  $\dim V = 1$  and  $T \in \mathcal{L}(V, V)$ , then there exists  $\lambda \in \mathbf{F}$  such that  $Tv = \lambda v$  for all  $v \in V$ .
- 8 Give an example of a function  $\varphi: \mathbf{R}^2 \rightarrow \mathbf{R}$  such that

$$\varphi(av) = a\varphi(v)$$

for all  $a \in \mathbf{R}$  and all  $v \in \mathbf{R}^2$  but  $\varphi$  is not linear.

[The exercise above and the next exercise show that neither homogeneity nor additivity alone is enough to imply that a function is a linear map.]

- 9 Give an example of a function  $\varphi: \mathbf{C} \rightarrow \mathbf{C}$  such that

$$\varphi(w + z) = \varphi(w) + \varphi(z)$$

for all  $w, z \in \mathbf{C}$  but  $\varphi$  is not linear. (Here  $\mathbf{C}$  is thought of as a complex vector space.)

[There also exists a function  $\varphi: \mathbf{R} \rightarrow \mathbf{R}$  such that  $\varphi$  satisfies the additivity condition above but  $\varphi$  is not linear. However, showing the existence of such a function involves considerably more advanced tools.]

- 10 Suppose  $U$  is a subspace of  $V$  with  $U \neq V$ . Suppose  $S \in \mathcal{L}(U, W)$  and  $S \neq 0$  (which means that  $Su \neq 0$  for some  $u \in U$ ). Define  $T: V \rightarrow W$  by

$$Tv = \begin{cases} Sv & \text{if } v \in U, \\ 0 & \text{if } v \in V \text{ and } v \notin U. \end{cases}$$

Prove that  $T$  is not a linear map on  $V$ .

- 11 Suppose  $V$  is finite-dimensional. Prove that every linear map on a subspace of  $V$  can be extended to a linear map on  $V$ . In other words, show that if  $U$  is a subspace of  $V$  and  $S \in \mathcal{L}(U, W)$ , then there exists  $T \in \mathcal{L}(V, W)$  such that  $Tu = Su$  for all  $u \in U$ .
- 12 Suppose  $V$  is finite-dimensional with  $\dim V > 0$ , and suppose  $W$  is infinite-dimensional. Prove that  $\mathcal{L}(V, W)$  is infinite-dimensional.
- 13 Suppose  $v_1, \dots, v_m$  is a linearly dependent list of vectors in  $V$ . Suppose also that  $W \neq \{0\}$ . Prove that there exist  $w_1, \dots, w_m \in W$  such that no  $T \in \mathcal{L}(V, W)$  satisfies  $Tv_k = w_k$  for each  $k = 1, \dots, m$ .
- 14 Suppose  $V$  is finite-dimensional with  $\dim V \geq 2$ . Prove that there exist  $S, T \in \mathcal{L}(V, V)$  such that  $ST \neq TS$ .

## 3.B Null Spaces and Ranges

### Null Space and Injectivity

In this section we will learn about two subspaces that are intimately connected with each linear map. We begin with the set of vectors that get mapped to 0.

#### 3.12 Definition *null space*, $\text{null } T$

For  $T \in \mathcal{L}(V, W)$ , the ***null space*** of  $T$ , denoted  $\text{null } T$ , is the subset of  $V$  consisting of those vectors that  $T$  maps to 0:

$$\text{null } T = \{v \in V : Tv = 0\}.$$

#### 3.13 Example *null space*

- If  $T$  is the zero map from  $V$  to  $W$ , in other words if  $Tv = 0$  for every  $v \in V$ , then  $\text{null } T = V$ .
- Suppose  $\varphi \in \mathcal{L}(\mathbf{C}^3, \mathbf{F})$  is defined by  $\varphi(z_1, z_2, z_3) = z_1 + 2z_2 + 3z_3$ . Then  $\text{null } \varphi = \{(z_1, z_2, z_3) \in \mathbf{C}^3 : z_1 + 2z_2 + 3z_3 = 0\}$ . A basis of  $\text{null } \varphi$  is  $(-2, 1, 0), (-3, 0, 1)$ .
- Suppose  $D \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  is the differentiation map defined by  $Dp = p'$ . The only functions whose derivative equals the zero function are the constant functions. Thus the null space of  $D$  equals the set of constant functions.
- Suppose  $T \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  is the multiplication by  $x^2$  map defined by  $(Tp)(x) = x^2 p(x)$ . The only polynomial  $p$  such that  $x^2 p(x) = 0$  for all  $x \in \mathbf{R}$  is the 0 polynomial. Thus  $\text{null } T = \{0\}$ .
- Suppose  $T \in \mathcal{L}(\mathbf{F}^\infty, \mathbf{F}^\infty)$  is the backward shift defined by

$$T(x_1, x_2, x_3, \dots) = (x_2, x_3, \dots).$$

Clearly  $T(x_1, x_2, x_3, \dots)$  equals 0 if and only if  $x_2, x_3, \dots$  are all 0. Thus in this case we have  $\text{null } T = \{(a, 0, 0, \dots) : a \in \mathbf{F}\}$ .

The next result shows that the null space of each linear map is a subspace of the domain. In particular, 0 is in the null space of every linear map.

Some mathematicians use the term ***kernel*** instead of *null space*. The word “null” means zero. Thus the term “null space” should remind you of the connection to 0.

### 3.14 The null space is a subspace

Suppose  $T \in \mathcal{L}(V, W)$ . Then  $\text{null } T$  is a subspace of  $V$ .

**Proof** Because  $T$  is a linear map, we know that  $T(0) = 0$  (by 3.11). Thus  $0 \in \text{null } T$ .

Suppose  $u, v \in \text{null } T$ . Then

$$T(u + v) = Tu + Tv = 0 + 0 = 0.$$

Hence  $u + v \in \text{null } T$ . Thus  $\text{null } T$  is closed under addition.

Suppose  $u \in \text{null } T$  and  $\lambda \in \mathbf{F}$ . Then

$$T(\lambda u) = \lambda Tu = \lambda 0 = 0.$$

Hence  $\lambda u \in \text{null } T$ . Thus  $\text{null } T$  is closed under scalar multiplication.

*Take another look at the null spaces that were computed in Example 3.13 and note that all of them are subspaces.*

We have shown that  $\text{null } T$  contains 0 and is closed under addition and scalar multiplication. Thus  $\text{null } T$  is a subspace of  $V$  (by 1.34). ■

As we will soon see, for a linear map the next definition is closely connected to the null space.

### 3.15 Definition *injective*

A function  $T : V \rightarrow W$  is called *injective* if  $Tu = Tv$  implies  $u = v$ .

*Many mathematicians use the term **one-to-one**, which means the same as injective.*

The definition above could be rephrased to say that  $T$  is injective if  $u \neq v$  implies that  $Tu \neq Tv$ . In other words,  $T$  is injective if it maps distinct inputs to distinct outputs.

The next result says that we can check whether a linear map is injective by checking whether 0 is the only vector that gets mapped to 0. As a simple application of this result, we see that of the linear maps whose null spaces we computed in 3.13, only multiplication by  $x^2$  is injective (except that the zero map is injective in the special case  $V = \{0\}$ ).

### 3.16 Injectivity is equivalent to null space equals {0}

Let  $T \in \mathcal{L}(V, W)$ . Then  $T$  is injective if and only if  $\text{null } T = \{0\}$ .

**Proof** First suppose  $T$  is injective. We want to prove that  $\text{null } T = \{0\}$ . We already know that  $\{0\} \subset \text{null } T$  (by 3.11). To prove the inclusion in the other direction, suppose  $v \in \text{null } T$ . Then

$$T(v) = 0 = T(0).$$

Because  $T$  is injective, the equation above implies that  $v = 0$ . Thus we can conclude that  $\text{null } T = \{0\}$ , as desired.

To prove the implication in the other direction, now suppose  $\text{null } T = \{0\}$ . We want to prove that  $T$  is injective. To do this, suppose  $u, v \in V$  and  $Tu = Tv$ . Then

$$0 = Tu - Tv = T(u - v).$$

Thus  $u - v$  is in  $\text{null } T$ , which equals  $\{0\}$ . Hence  $u - v = 0$ , which implies that  $u = v$ . Hence  $T$  is injective, as desired. ■

### Range and Surjectivity

Now we give a name to the set of outputs of a function.

#### 3.17 Definition range

For  $T$  a function from  $V$  to  $W$ , the **range** of  $T$  is the subset of  $W$  consisting of those vectors that are of the form  $Tv$  for some  $v \in V$ :

$$\text{range } T = \{Tv : v \in V\}.$$

---

#### 3.18 Example range

- If  $T$  is the zero map from  $V$  to  $W$ , in other words if  $Tv = 0$  for every  $v \in V$ , then  $\text{range } T = \{0\}$ .
  - Suppose  $T \in \mathcal{L}(\mathbf{R}^2, \mathbf{R}^3)$  is defined by  $T(x, y) = (2x, 5y, x + y)$ , then  $\text{range } T = \{(2x, 5y, x + y) : x, y \in \mathbf{R}\}$ . A basis of  $\text{range } T$  is  $(2, 0, 1), (0, 5, 1)$ .
  - Suppose  $D \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  is the differentiation map defined by  $Dp = p'$ . Because for every polynomial  $q \in \mathcal{P}(\mathbf{R})$  there exists a polynomial  $p \in \mathcal{P}(\mathbf{R})$  such that  $p' = q$ , the range of  $D$  is  $\mathcal{P}(\mathbf{R})$ .
-

*Some mathematicians use the word **image**, which means the same as range.*

The next result shows that the range of each linear map is a subspace of the vector space into which it is being mapped.

### 3.19 The range is a subspace

If  $T \in \mathcal{L}(V, W)$ , then range  $T$  is a subspace of  $W$ .

**Proof** Suppose  $T \in \mathcal{L}(V, W)$ . Then  $T(0) = 0$  (by 3.11), which implies that  $0 \in \text{range } T$ .

If  $w_1, w_2 \in \text{range } T$ , then there exist  $v_1, v_2 \in V$  such that  $Tv_1 = w_1$  and  $Tv_2 = w_2$ . Thus

$$T(v_1 + v_2) = Tv_1 + Tv_2 = w_1 + w_2.$$

Hence  $w_1 + w_2 \in \text{range } T$ . Thus range  $T$  is closed under addition.

If  $w \in \text{range } T$  and  $\lambda \in \mathbf{F}$ , then there exists  $v \in V$  such that  $Tv = w$ . Thus

$$T(\lambda v) = \lambda Tv = \lambda w.$$

Hence  $\lambda w \in \text{range } T$ . Thus range  $T$  is closed under scalar multiplication.

We have shown that range  $T$  contains 0 and is closed under addition and scalar multiplication. Thus range  $T$  is a subspace of  $W$  (by 1.34). ■

### 3.20 Definition *surjective*

A function  $T: V \rightarrow W$  is called *surjective* if its range equals  $W$ .

To illustrate the definition above, note that of the ranges we computed in 3.18, only the differentiation map is surjective (except that the zero map is surjective in the special case  $W = \{0\}$ ).

*Many mathematicians use the term **onto**, which means the same as **surjective**.*

Whether a linear map is surjective depends on what we are thinking of as the vector space into which it maps.

**3.21 Example** The differentiation map  $D \in \mathcal{L}(\mathcal{P}_5(\mathbf{R}), \mathcal{P}_5(\mathbf{R}))$  defined by  $Dp = p'$  is not surjective, because the polynomial  $x^5$  is not in the range of  $D$ . However, the differentiation map  $S \in \mathcal{L}(\mathcal{P}_5(\mathbf{R}), \mathcal{P}_4(\mathbf{R}))$  defined by  $Sp = p'$  is surjective, because its range equals  $\mathcal{P}_4(\mathbf{R})$ , which is now the vector space into which  $S$  maps.

## Fundamental Theorem of Linear Maps

The next result is so important that it gets a dramatic name.

### 3.22 Fundamental Theorem of Linear Maps

Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Then  $\text{range } T$  is finite-dimensional and

$$\dim V = \dim \text{null } T + \dim \text{range } T.$$

**Proof** Let  $u_1, \dots, u_m$  be a basis of  $\text{null } T$ ; thus  $\dim \text{null } T = m$ . The linearly independent list  $u_1, \dots, u_m$  can be extended to a basis

$$u_1, \dots, u_m, v_1, \dots, v_n$$

of  $V$  (by 2.33). Thus  $\dim V = m + n$ . To complete the proof, we need only show that  $\text{range } T$  is finite-dimensional and  $\dim \text{range } T = n$ . We will do this by proving that  $Tv_1, \dots, Tv_n$  is a basis of  $\text{range } T$ .

Let  $v \in V$ . Because  $u_1, \dots, u_m, v_1, \dots, v_n$  spans  $V$ , we can write

$$v = a_1u_1 + \dots + a_mu_m + b_1v_1 + \dots + b_nv_n,$$

where the  $a$ 's and  $b$ 's are in  $\mathbf{F}$ . Applying  $T$  to both sides of this equation, we get

$$Tv = b_1Tv_1 + \dots + b_nTv_n,$$

where the terms of the form  $Tu_j$  disappeared because each  $u_j$  is in  $\text{null } T$ . The last equation implies that  $Tv_1, \dots, Tv_n$  spans  $\text{range } T$ . In particular,  $\text{range } T$  is finite-dimensional.

To show  $Tv_1, \dots, Tv_n$  is linearly independent, suppose  $c_1, \dots, c_n \in \mathbf{F}$  and

$$c_1Tv_1 + \dots + c_nTv_n = 0.$$

Then

$$T(c_1v_1 + \dots + c_nv_n) = 0.$$

Hence

$$c_1v_1 + \dots + c_nv_n \in \text{null } T.$$

Because  $u_1, \dots, u_m$  spans  $\text{null } T$ , we can write

$$c_1v_1 + \dots + c_nv_n = d_1u_1 + \dots + d_mu_m,$$

where the  $d$ 's are in  $\mathbf{F}$ . This equation implies that all the  $c$ 's (and  $d$ 's) are 0 (because  $u_1, \dots, u_m, v_1, \dots, v_n$  is linearly independent). Thus  $Tv_1, \dots, Tv_n$  is linearly independent and hence is a basis of  $\text{range } T$ , as desired. ■

Now we can show that no linear map from a finite-dimensional vector space to a “smaller” vector space can be injective, where “smaller” is measured by dimension.

### 3.23 A map to a smaller dimensional space is not injective

Suppose  $V$  and  $W$  are finite-dimensional vector spaces such that  $\dim V > \dim W$ . Then no linear map from  $V$  to  $W$  is injective.

**Proof** Let  $T \in \mathcal{L}(V, W)$ . Then

$$\begin{aligned}\dim \text{null } T &= \dim V - \dim \text{range } T \\ &\geq \dim V - \dim W \\ &> 0,\end{aligned}$$

where the equality above comes from the Fundamental Theorem of Linear Maps (3.22). The inequality above states that  $\dim \text{null } T > 0$ . This means that  $\text{null } T$  contains vectors other than 0. Thus  $T$  is not injective (by 3.16). ■

The next result shows that no linear map from a finite-dimensional vector space to a “bigger” vector space can be surjective, where “bigger” is measured by dimension.

### 3.24 A map to a larger dimensional space is not surjective

Suppose  $V$  and  $W$  are finite-dimensional vector spaces such that  $\dim V < \dim W$ . Then no linear map from  $V$  to  $W$  is surjective.

**Proof** Let  $T \in \mathcal{L}(V, W)$ . Then

$$\begin{aligned}\dim \text{range } T &= \dim V - \dim \text{null } T \\ &\leq \dim V \\ &< \dim W,\end{aligned}$$

where the equality above comes from the Fundamental Theorem of Linear Maps (3.22). The inequality above states that  $\dim \text{range } T < \dim W$ . This means that  $\text{range } T$  cannot equal  $W$ . Thus  $T$  is not surjective. ■

As we will now see, 3.23 and 3.24 have important consequences in the theory of linear equations. The idea here is to express questions about systems of linear equations in terms of linear maps.

**3.25 Example** Rephrase in terms of a linear map the question of whether a homogeneous system of linear equations has a nonzero solution.

### Solution

Fix positive integers  $m$  and  $n$ , and let  $A_{j,k} \in \mathbf{F}$  for  $j = 1, \dots, m$  and  $k = 1, \dots, n$ . Consider the homogeneous system of linear equations

*Homogeneous, in this context, means that the constant term on the right side of each equation below is 0.*

$$\begin{aligned} \sum_{k=1}^n A_{1,k}x_k &= 0 \\ &\vdots \\ \sum_{k=1}^n A_{m,k}x_k &= 0. \end{aligned}$$

Obviously  $x_1 = \dots = x_n = 0$  is a solution of the system of equations above; the question here is whether any other solutions exist.

Define  $T : \mathbf{F}^n \rightarrow \mathbf{F}^m$  by

$$T(x_1, \dots, x_n) = \left( \sum_{k=1}^n A_{1,k}x_k, \dots, \sum_{k=1}^n A_{m,k}x_k \right).$$

The equation  $T(x_1, \dots, x_n) = 0$  (the 0 here is the additive identity in  $\mathbf{F}^m$ , namely, the list of length  $m$  of all 0's) is the same as the homogeneous system of linear equations above.

Thus we want to know if  $\text{null } T$  is strictly bigger than  $\{0\}$ . In other words, we can rephrase our question about nonzero solutions as follows (by 3.16): What condition ensures that  $T$  is not injective?

### 3.26 Homogeneous system of linear equations

A homogeneous system of linear equations with more variables than equations has nonzero solutions.

**Proof** Use the notation and result from the example above. Thus  $T$  is a linear map from  $\mathbf{F}^n$  to  $\mathbf{F}^m$ , and we have a homogeneous system of  $m$  linear equations with  $n$  variables  $x_1, \dots, x_n$ . From 3.23 we see that  $T$  is not injective if  $n > m$ . ■

Example of the result above: a homogeneous system of four linear equations with five variables has nonzero solutions.

**3.27 Example** Consider the question of whether an inhomogeneous system of linear equations has no solutions for some choice of the constant terms. Rephrase this question in terms of a linear map.

**Solution** Fix positive integers  $m$  and  $n$ , and let  $A_{j,k} \in \mathbf{F}$  for  $j = 1, \dots, m$  and  $k = 1, \dots, n$ . For  $c_1, \dots, c_m \in \mathbf{F}$ , consider the system of linear equations

$$\sum_{k=1}^n A_{1,k}x_k = c_1$$

⋮

$$\sum_{k=1}^n A_{m,k}x_k = c_m.$$

**3.28**

The question here is whether there is some choice of  $c_1, \dots, c_m \in \mathbf{F}$  such that no solution exists to the system above.

Define  $T : \mathbf{F}^n \rightarrow \mathbf{F}^m$  by

$$T(x_1, \dots, x_n) = \left( \sum_{k=1}^n A_{1,k}x_k, \dots, \sum_{k=1}^n A_{m,k}x_k \right).$$

The equation  $T(x_1, \dots, x_n) = (c_1, \dots, c_m)$  is the same as the system of equations 3.28. Thus we want to know if  $\text{range } T \neq \mathbf{F}^m$ . Hence we can rephrase our question about not having a solution for some choice of  $c_1, \dots, c_m \in \mathbf{F}$  as follows: What condition ensures that  $T$  is not surjective?

### 3.29 Inhomogeneous system of linear equations

An inhomogeneous system of linear equations with more equations than variables has no solution for some choice of the constant terms.

*Our results about homogeneous systems with more variables than equations and inhomogeneous systems with more equations than variables (3.26 and 3.29) are often proved using Gaussian elimination. The abstract approach taken here leads to cleaner proofs.*

**Proof** Use the notation and result from the example above. Thus  $T$  is a linear map from  $\mathbf{F}^n$  to  $\mathbf{F}^m$ , and we have a system of  $m$  equations with  $n$  variables  $x_1, \dots, x_n$ . From 3.24 we see that  $T$  is not surjective if  $n < m$ . ■

Example of the result above: an inhomogeneous system of five linear equations with four variables has no solution for some choice of the constant terms.

## EXERCISES 3.B

- 1 Give an example of a linear map  $T$  such that  $\dim \text{null } T = 3$  and  $\dim \text{range } T = 2$ .

- 2 Suppose  $V$  is a vector space and  $S, T \in \mathcal{L}(V, V)$  are such that

$$\text{range } S \subset \text{null } T.$$

Prove that  $(ST)^2 = 0$ .

- 3 Suppose  $v_1, \dots, v_m$  is a list of vectors in  $V$ . Define  $T \in \mathcal{L}(\mathbf{F}^m, V)$  by

$$T(z_1, \dots, z_m) = z_1 v_1 + \dots + z_m v_m.$$

- (a) What property of  $T$  corresponds to  $v_1, \dots, v_m$  spanning  $V$ ?  
(b) What property of  $T$  corresponds to  $v_1, \dots, v_m$  being linearly independent?

- 4 Show that

$$\{T \in \mathcal{L}(\mathbf{R}^5, \mathbf{R}^4) : \dim \text{null } T > 2\}$$

is not a subspace of  $\mathcal{L}(\mathbf{R}^5, \mathbf{R}^4)$ .

- 5 Give an example of a linear map  $T : \mathbf{R}^4 \rightarrow \mathbf{R}^4$  such that

$$\text{range } T = \text{null } T.$$

- 6 Prove that there does not exist a linear map  $T : \mathbf{R}^5 \rightarrow \mathbf{R}^5$  such that

$$\text{range } T = \text{null } T.$$

- 7 Suppose  $V$  and  $W$  are finite-dimensional with  $2 \leq \dim V \leq \dim W$ . Show that  $\{T \in \mathcal{L}(V, W) : T \text{ is not injective}\}$  is not a subspace of  $\mathcal{L}(V, W)$ .

- 8 Suppose  $V$  and  $W$  are finite-dimensional with  $\dim V \geq \dim W \geq 2$ . Show that  $\{T \in \mathcal{L}(V, W) : T \text{ is not surjective}\}$  is not a subspace of  $\mathcal{L}(V, W)$ .

- 9 Suppose  $T \in \mathcal{L}(V, W)$  is injective and  $v_1, \dots, v_n$  is linearly independent in  $V$ . Prove that  $Tv_1, \dots, Tv_n$  is linearly independent in  $W$ .

- 10** Suppose  $v_1, \dots, v_n$  spans  $V$  and  $T \in \mathcal{L}(V, W)$ . Prove that the list  $Tv_1, \dots, Tv_n$  spans range  $T$ .
- 11** Suppose  $S_1, \dots, S_n$  are injective linear maps such that  $S_1S_2 \cdots S_n$  makes sense. Prove that  $S_1S_2 \cdots S_n$  is injective.
- 12** Suppose that  $V$  is finite-dimensional and that  $T \in \mathcal{L}(V, W)$ . Prove that there exists a subspace  $U$  of  $V$  such that  $U \cap \text{null } T = \{0\}$  and  $\text{range } T = \{Tu : u \in U\}$ .
- 13** Suppose  $T$  is a linear map from  $\mathbf{F}^4$  to  $\mathbf{F}^2$  such that
- $$\text{null } T = \{(x_1, x_2, x_3, x_4) \in \mathbf{F}^4 : x_1 = 5x_2 \text{ and } x_3 = 7x_4\}.$$
- Prove that  $T$  is surjective.
- 14** Suppose  $U$  is a 3-dimensional subspace of  $\mathbf{R}^8$  and that  $T$  is a linear map from  $\mathbf{R}^8$  to  $\mathbf{R}^5$  such that  $\text{null } T = U$ . Prove that  $T$  is surjective.
- 15** Prove that there does not exist a linear map from  $\mathbf{F}^5$  to  $\mathbf{F}^2$  whose null space equals
- $$\{(x_1, x_2, x_3, x_4, x_5) \in \mathbf{F}^5 : x_1 = 3x_2 \text{ and } x_3 = x_4 = x_5\}.$$
- 16** Suppose there exists a linear map on  $V$  whose null space and range are both finite-dimensional. Prove that  $V$  is finite-dimensional.
- 17** Suppose  $V$  and  $W$  are both finite-dimensional. Prove that there exists an injective linear map from  $V$  to  $W$  if and only if  $\dim V \leq \dim W$ .
- 18** Suppose  $V$  and  $W$  are both finite-dimensional. Prove that there exists a surjective linear map from  $V$  onto  $W$  if and only if  $\dim V \geq \dim W$ .
- 19** Suppose  $V$  and  $W$  are finite-dimensional and that  $U$  is a subspace of  $V$ . Prove that there exists  $T \in \mathcal{L}(V, W)$  such that  $\text{null } T = U$  if and only if  $\dim U \geq \dim V - \dim W$ .
- 20** Suppose  $W$  is finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Prove that  $T$  is injective if and only if there exists  $S \in \mathcal{L}(W, V)$  such that  $ST$  is the identity map on  $V$ .
- 21** Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Prove that  $T$  is surjective if and only if there exists  $S \in \mathcal{L}(W, V)$  such that  $TS$  is the identity map on  $W$ .

- 22** Suppose  $U$  and  $V$  are finite-dimensional vector spaces and  $S \in \mathcal{L}(V, W)$  and  $T \in \mathcal{L}(U, V)$ . Prove that

$$\dim \text{null } ST \leq \dim \text{null } S + \dim \text{null } T.$$

- 23** Suppose  $U$  and  $V$  are finite-dimensional vector spaces and  $S \in \mathcal{L}(V, W)$  and  $T \in \mathcal{L}(U, V)$ . Prove that

$$\dim \text{range } ST \leq \min\{\dim \text{range } S, \dim \text{range } T\}.$$

- 24** Suppose  $W$  is finite-dimensional and  $T_1, T_2 \in \mathcal{L}(V, W)$ . Prove that  $\text{null } T_1 \subset \text{null } T_2$  if and only if there exists  $S \in \mathcal{L}(W, W)$  such that  $T_2 = ST_1$ .

- 25** Suppose  $V$  is finite-dimensional and  $T_1, T_2 \in \mathcal{L}(V, W)$ . Prove that  $\text{range } T_1 \subset \text{range } T_2$  if and only if there exists  $S \in \mathcal{L}(V, V)$  such that  $T_1 = T_2S$ .

- 26** Suppose  $D \in \mathcal{L}(\mathcal{P}(\mathbf{R}), \mathcal{P}(\mathbf{R}))$  is such that  $\deg Dp = (\deg p) - 1$  for every nonconstant polynomial  $p \in \mathcal{P}(\mathbf{R})$ . Prove that  $D$  is surjective.

[The notation  $D$  is used above to remind you of the differentiation map that sends a polynomial  $p$  to  $p'$ . Without knowing the formula for the derivative of a polynomial (except that it reduces the degree by 1), you can use the exercise above to show that for every polynomial  $q \in \mathcal{P}(\mathbf{R})$ , there exists a polynomial  $p \in \mathcal{P}(\mathbf{R})$  such that  $p' = q$ .]

- 27** Suppose  $p \in \mathcal{P}(\mathbf{R})$ . Prove that there exists a polynomial  $q \in \mathcal{P}(\mathbf{R})$  such that  $5q'' + 3q' = p$ .

[This exercise can be done without linear algebra, but it's more fun to do it using linear algebra.]

- 28** Suppose  $T \in \mathcal{L}(V, W)$ , and  $w_1, \dots, w_m$  is a basis of  $\text{range } T$ . Prove that there exist  $\varphi_1, \dots, \varphi_m \in \mathcal{L}(V, \mathbf{F})$  such that

$$Tv = \varphi_1(v)w_1 + \cdots + \varphi_m(v)w_m$$

for every  $v \in V$ .

- 29** Suppose  $\varphi \in \mathcal{L}(V, \mathbf{F})$ . Suppose  $u \in V$  is not in  $\text{null } \varphi$ . Prove that

$$V = \text{null } \varphi \oplus \{au : a \in \mathbf{F}\}.$$

- 30** Suppose  $\varphi_1$  and  $\varphi_2$  are linear maps from  $V$  to  $\mathbf{F}$  that have the same null space. Show that there exists a constant  $c \in \mathbf{F}$  such that  $\varphi_1 = c\varphi_2$ .

- 31** Give an example of two linear maps  $T_1$  and  $T_2$  from  $\mathbf{R}^5$  to  $\mathbf{R}^2$  that have the same null space but are such that  $T_1$  is not a scalar multiple of  $T_2$ .

## 3.C Matrices

### Representing a Linear Map by a Matrix

We know that if  $v_1, \dots, v_n$  is a basis of  $V$  and  $T: V \rightarrow W$  is linear, then the values of  $Tv_1, \dots, T v_n$  determine the values of  $T$  on arbitrary vectors in  $V$  (see 3.5). As we will soon see, matrices are used as an efficient method of recording the values of the  $T v_j$ 's in terms of a basis of  $W$ .

#### 3.30 Definition **matrix**, $A_{j,k}$

Let  $m$  and  $n$  denote positive integers. An  $m$ -by- $n$  **matrix**  $A$  is a rectangular array of elements of  $\mathbb{F}$  with  $m$  rows and  $n$  columns:

$$A = \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{m,1} & \dots & A_{m,n} \end{pmatrix}.$$

The notation  $A_{j,k}$  denotes the entry in row  $j$ , column  $k$  of  $A$ . In other words, the first index refers to the row number and the second index refers to the column number.

Thus  $A_{2,3}$  refers to the entry in the second row, third column of a matrix  $A$ .

---

#### 3.31 Example If $A = \begin{pmatrix} 8 & 4 & 5 - 3i \\ 1 & 9 & 7 \end{pmatrix}$ , then $A_{2,3} = 7$ .

---

Now we come to the key definition in this section.

#### 3.32 Definition **matrix of a linear map**, $\mathcal{M}(T)$

Suppose  $T \in \mathcal{L}(V, W)$  and  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_m$  is a basis of  $W$ . The **matrix of**  $T$  with respect to these bases is the  $m$ -by- $n$  matrix  $\mathcal{M}(T)$  whose entries  $A_{j,k}$  are defined by

$$Tv_k = A_{1,k}w_1 + \cdots + A_{m,k}w_m.$$

If the bases are not clear from the context, then the notation  $\mathcal{M}(T, (v_1, \dots, v_n), (w_1, \dots, w_m))$  is used.

The matrix  $\mathcal{M}(T)$  of a linear map  $T \in \mathcal{L}(V, W)$  depends on the basis  $v_1, \dots, v_n$  of  $V$  and the basis  $w_1, \dots, w_m$  of  $W$ , as well as on  $T$ . However, the bases should be clear from the context, and thus they are often not included in the notation.

To remember how  $\mathcal{M}(T)$  is constructed from  $T$ , you might write across the top of the matrix the basis vectors  $v_1, \dots, v_n$  for the domain and along the left the basis vectors  $w_1, \dots, w_m$  for the vector space into which  $T$  maps, as follows:

$$\mathcal{M}(T) = \begin{array}{c} v_1 \quad \dots \quad v_k \quad \dots \quad v_n \\ \hline w_1 & \left( \begin{array}{c} A_{1,k} \\ \vdots \\ A_{m,k} \end{array} \right) \\ \vdots \\ w_m \end{array}.$$

In the matrix above only the  $k^{\text{th}}$  column is shown. Thus the second index of each displayed entry of the matrix above is  $k$ . The picture above should remind you that  $Tv_k$  can be computed from  $\mathcal{M}(T)$  by multiplying each entry in the  $k^{\text{th}}$  column by the corresponding  $w_j$  from the left column, and then adding up the resulting vectors.

The  $k^{\text{th}}$  column of  $\mathcal{M}(T)$  consists of the scalars needed to write  $Tv_k$  as a linear combination of  $(w_1, \dots, w_m)$ :

$$Tv_k = \sum_{j=1}^m A_{j,k} w_j.$$

If  $T$  is a linear map from  $\mathbf{F}^n$  to  $\mathbf{F}^m$ , then unless stated otherwise, assume the bases in question are the standard ones (where the  $k^{\text{th}}$  basis vector is 1 in the  $k^{\text{th}}$  slot and 0 in all the other slots). If you think of elements of  $\mathbf{F}^m$  as columns of  $m$  numbers, then you can think of the  $k^{\text{th}}$  column of  $\mathcal{M}(T)$  as  $T$  applied to the  $k^{\text{th}}$  standard basis vector.

If  $T$  maps an  $n$ -dimensional vector space to an  $m$ -dimensional vector space, then  $\mathcal{M}(T)$  is an  $m$ -by- $n$  matrix.

### 3.33 Example

Suppose  $T \in \mathcal{L}(\mathbf{F}^2, \mathbf{F}^3)$  is defined by

$$T(x, y) = (x + 3y, 2x + 5y, 7x + 9y).$$

Find the matrix of  $T$  with respect to the standard bases of  $\mathbf{F}^2$  and  $\mathbf{F}^3$ .

**Solution** Because  $T(1, 0) = (1, 2, 7)$  and  $T(0, 1) = (3, 5, 9)$ , the matrix of  $T$  with respect to the standard bases is the 3-by-2 matrix below:

$$\mathcal{M}(T) = \begin{pmatrix} 1 & 3 \\ 2 & 5 \\ 7 & 9 \end{pmatrix}.$$

When working with  $\mathcal{P}_m(\mathbf{F})$ , use the standard basis  $1, x, x^2, \dots, x^m$  unless the context indicates otherwise.

**3.34 Example** Suppose  $D \in \mathcal{L}(\mathcal{P}_3(\mathbf{R}), \mathcal{P}_2(\mathbf{R}))$  is the differentiation map defined by  $Dp = p'$ . Find the matrix of  $D$  with respect to the standard bases of  $\mathcal{P}_3(\mathbf{R})$  and  $\mathcal{P}_2(\mathbf{R})$ .

**Solution** Because  $(x^n)' = nx^{n-1}$ , the matrix of  $T$  with respect to the standard bases is the 3-by-4 matrix below:

$$\mathcal{M}(D) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

## Addition and Scalar Multiplication of Matrices

For the rest of this section, assume that  $V$  and  $W$  are finite-dimensional and that a basis has been chosen for each of these vector spaces. Thus for each linear map from  $V$  to  $W$ , we can talk about its matrix (with respect to the chosen bases, of course). Is the matrix of the sum of two linear maps equal to the sum of the matrices of the two maps?

Right now this question does not make sense, because although we have defined the sum of two linear maps, we have not defined the sum of two matrices. Fortunately, the obvious definition of the sum of two matrices has the right properties. Specifically, we make the following definition.

### 3.35 Definition matrix addition

The *sum of two matrices of the same size* is the matrix obtained by adding corresponding entries in the matrices:

$$\begin{aligned} & \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{m,1} & \dots & A_{m,n} \end{pmatrix} + \begin{pmatrix} C_{1,1} & \dots & C_{1,n} \\ \vdots & & \vdots \\ C_{m,1} & \dots & C_{m,n} \end{pmatrix} \\ &= \begin{pmatrix} A_{1,1} + C_{1,1} & \dots & A_{1,n} + C_{1,n} \\ \vdots & & \vdots \\ A_{m,1} + C_{m,1} & \dots & A_{m,n} + C_{m,n} \end{pmatrix}. \end{aligned}$$

In other words,  $(A + C)_{j,k} = A_{j,k} + C_{j,k}$ .

In the following result, the assumption is that the same bases are used for all three linear maps  $S + T$ ,  $S$ , and  $T$ .

### 3.36 The matrix of the sum of linear maps

Suppose  $S, T \in \mathcal{L}(V, W)$ . Then  $\mathcal{M}(S + T) = \mathcal{M}(S) + \mathcal{M}(T)$ .

The verification of the result above is left to the reader.

Still assuming that we have some bases in mind, is the matrix of a scalar times a linear map equal to the scalar times the matrix of the linear map? Again the question does not make sense, because we have not defined scalar multiplication on matrices. Fortunately, the obvious definition again has the right properties.

### 3.37 Definition scalar multiplication of a matrix

The product of a scalar and a matrix is the matrix obtained by multiplying each entry in the matrix by the scalar:

$$\lambda \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{m,1} & \dots & A_{m,n} \end{pmatrix} = \begin{pmatrix} \lambda A_{1,1} & \dots & \lambda A_{1,n} \\ \vdots & & \vdots \\ \lambda A_{m,1} & \dots & \lambda A_{m,n} \end{pmatrix}.$$

In other words,  $(\lambda A)_{j,k} = \lambda A_{j,k}$ .

In the following result, the assumption is that the same bases are used for both linear maps  $\lambda T$  and  $T$ .

### 3.38 The matrix of a scalar times a linear map

Suppose  $\lambda \in \mathbf{F}$  and  $T \in \mathcal{L}(V, W)$ . Then  $\mathcal{M}(\lambda T) = \lambda \mathcal{M}(T)$ .

The verification of the result above is also left to the reader.

Because addition and scalar multiplication have now been defined for matrices, you should not be surprised that a vector space is about to appear. We need only a bit of notation so that this new vector space has a name.

### 3.39 Notation $\mathbf{F}^{m,n}$

For  $m$  and  $n$  positive integers, the set of all  $m$ -by- $n$  matrices with entries in  $\mathbf{F}$  is denoted by  $\mathbf{F}^{m,n}$ .

### 3.40 $\dim \mathbf{F}^{m,n} = mn$

Suppose  $m$  and  $n$  are positive integers. With addition and scalar multiplication defined as above,  $\mathbf{F}^{m,n}$  is a vector space with dimension  $mn$ .

**Proof** The verification that  $\mathbf{F}^{m,n}$  is a vector space is left to the reader. Note that the additive identity of  $\mathbf{F}^{m,n}$  is the  $m$ -by- $n$  matrix whose entries all equal 0.

The reader should also verify that the list of  $m$ -by- $n$  matrices that have 0 in all entries except for a 1 in one entry is a basis of  $\mathbf{F}^{m,n}$ . There are  $mn$  such matrices, so the dimension of  $\mathbf{F}^{m,n}$  equals  $mn$ . ■

## Matrix Multiplication

Suppose, as previously, that  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_m$  is a basis of  $W$ . Suppose also that we have another vector space  $U$  and that  $u_1, \dots, u_p$  is a basis of  $U$ .

Consider linear maps  $T: U \rightarrow V$  and  $S: V \rightarrow W$ . The composition  $ST$  is a linear map from  $U$  to  $W$ . Does  $\mathcal{M}(ST)$  equal  $\mathcal{M}(S)\mathcal{M}(T)$ ? This question does not yet make sense, because we have not defined the product of two matrices. We will choose a definition of matrix multiplication that forces this question to have a positive answer. Let's see how to do this.

Suppose  $\mathcal{M}(S) = A$  and  $\mathcal{M}(T) = C$ . For  $1 \leq k \leq p$ , we have

$$\begin{aligned} (ST)u_k &= S\left(\sum_{r=1}^n C_{r,k}v_r\right) \\ &= \sum_{r=1}^n C_{r,k}Sv_r \\ &= \sum_{r=1}^n C_{r,k} \sum_{j=1}^m A_{j,r}w_j \\ &= \sum_{j=1}^m \left(\sum_{r=1}^n A_{j,r}C_{r,k}\right)w_j. \end{aligned}$$

Thus  $\mathcal{M}(ST)$  is the  $m$ -by- $p$  matrix whose entry in row  $j$ , column  $k$ , equals

$$\sum_{r=1}^n A_{j,r}C_{r,k}.$$

Now we see how to define matrix multiplication so that the desired equation  $\mathcal{M}(ST) = \mathcal{M}(S)\mathcal{M}(T)$  holds.

### 3.41 Definition *matrix multiplication*

Suppose  $A$  is an  $m$ -by- $n$  matrix and  $C$  is an  $n$ -by- $p$  matrix. Then  $AC$  is defined to be the  $m$ -by- $p$  matrix whose entry in row  $j$ , column  $k$ , is given by the following equation:

$$(AC)_{j,k} = \sum_{r=1}^n A_{j,r} C_{r,k}.$$

In other words, the entry in row  $j$ , column  $k$ , of  $AC$  is computed by taking row  $j$  of  $A$  and column  $k$  of  $C$ , multiplying together corresponding entries, and then summing.

Note that we define the product of two matrices only when the number of columns of the first matrix equals the number of rows of the second matrix.

*You may have learned this definition of matrix multiplication in an earlier course, although you may not have seen the motivation for it.*

**3.42 Example** Here we multiply together a 3-by-2 matrix and a 2-by-4 matrix, obtaining a 3-by-4 matrix:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 6 & 5 & 4 & 3 \\ 2 & 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 10 & 7 & 4 & 1 \\ 26 & 19 & 12 & 5 \\ 42 & 31 & 20 & 9 \end{pmatrix}.$$

Matrix multiplication is not commutative. In other words,  $AC$  is not necessarily equal to  $CA$  even if both products are defined (see Exercise 12). Matrix multiplication is distributive and associative (see Exercises 13 and 14).

In the following result, the assumption is that the same basis of  $V$  is used in considering  $T \in \mathcal{L}(U, V)$  and  $S \in \mathcal{L}(V, W)$ , the same basis of  $W$  is used in considering  $S \in \mathcal{L}(V, W)$  and  $ST \in \mathcal{L}(U, W)$ , and the same basis of  $U$  is used in considering  $T \in \mathcal{L}(U, V)$  and  $ST \in \mathcal{L}(U, W)$ .

### 3.43 The matrix of the product of linear maps

If  $T \in \mathcal{L}(U, V)$  and  $S \in \mathcal{L}(V, W)$ , then  $\mathcal{M}(ST) = \mathcal{M}(S)\mathcal{M}(T)$ .

The proof of the result above is the calculation that was done as motivation before the definition of matrix multiplication.

In the next piece of notation, note that as usual the first index refers to a row and the second index refers to a column, with a vertically centered dot used as a placeholder.

### 3.44 Notation $A_{j,\cdot}$ , $A_{\cdot,k}$

Suppose  $A$  is an  $m$ -by- $n$  matrix.

- If  $1 \leq j \leq m$ , then  $A_{j,\cdot}$  denotes the 1-by- $n$  matrix consisting of row  $j$  of  $A$ .
- If  $1 \leq k \leq n$ , then  $A_{\cdot,k}$  denotes the  $m$ -by-1 matrix consisting of column  $k$  of  $A$ .

**3.45 Example** If  $A = \begin{pmatrix} 8 & 4 & 5 \\ 1 & 9 & 7 \end{pmatrix}$ , then  $A_{2,\cdot}$  is row 2 of  $A$  and  $A_{\cdot,2}$  is column 2 of  $A$ . In other words,

$$A_{2,\cdot} = \begin{pmatrix} 1 & 9 & 7 \end{pmatrix} \quad \text{and} \quad A_{\cdot,2} = \begin{pmatrix} 4 \\ 9 \end{pmatrix}.$$

The product of a 1-by- $n$  matrix and an  $n$ -by-1 matrix is a 1-by-1 matrix. However, we will frequently identify a 1-by-1 matrix with its entry.

**3.46 Example**  $(3 \ 4) \begin{pmatrix} 6 \\ 2 \end{pmatrix} = (26)$  because  $3 \cdot 6 + 4 \cdot 2 = 26$ .

However, we can identify  $(26)$  with 26, writing  $(3 \ 4) \begin{pmatrix} 6 \\ 2 \end{pmatrix} = 26$ .

Our next result gives another way to think of matrix multiplication: the entry in row  $j$ , column  $k$ , of  $AC$  equals (row  $j$  of  $A$ ) times (column  $k$  of  $C$ ).

### 3.47 Entry of matrix product equals row times column

Suppose  $A$  is an  $m$ -by- $n$  matrix and  $C$  is an  $n$ -by- $p$  matrix. Then

$$(AC)_{j,k} = A_{j,\cdot} C_{\cdot,k}$$

for  $1 \leq j \leq m$  and  $1 \leq k \leq p$ .

The proof of the result above follows immediately from the definitions.

**3.48 Example** The result above and Example 3.46 show why the entry in row 2, column 1, of the product in Example 3.42 equals 26.

The next result gives yet another way to think of matrix multiplication. It states that column  $k$  of  $AC$  equals  $A$  times column  $k$  of  $C$ .

### 3.49 Column of matrix product equals matrix times column

Suppose  $A$  is an  $m$ -by- $n$  matrix and  $C$  is an  $n$ -by- $p$  matrix. Then

$$(AC)_{\cdot,k} = AC_{\cdot,k}$$

for  $1 \leq k \leq p$ .

Again, the proof of the result above follows immediately from the definitions and is left to the reader.

### 3.50 Example

From the result above and the equation

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 7 \\ 19 \\ 31 \end{pmatrix},$$

we see why column 2 in the matrix product in Example 3.42 is the right side of the equation above.

We give one more way of thinking about the product of an  $m$ -by- $n$  matrix and an  $n$ -by-1 matrix. The following example illustrates this approach.

### 3.51 Example

In the example above, the product of a 3-by-2 matrix and a 2-by-1 matrix is a linear combination of the columns of the 3-by-2 matrix, with the scalars that multiply the columns coming from the 2-by-1 matrix. Specifically,

$$\begin{pmatrix} 7 \\ 19 \\ 31 \end{pmatrix} = 5 \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} + 1 \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}.$$

The next result generalizes the example above. Again, the proof follows easily from the definitions and is left to the reader.

### 3.52 Linear combination of columns

Suppose  $A$  is an  $m$ -by- $n$  matrix and  $c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$  is an  $n$ -by-1 matrix.

Then

$$Ac = c_1 A_{\cdot,1} + \cdots + c_n A_{\cdot,n}.$$

In other words,  $Ac$  is a linear combination of the columns of  $A$ , with the scalars that multiply the columns coming from  $c$ .

Two more ways to think about matrix multiplication are given by Exercises 10 and 11.

## EXERCISES 3.C

---

- 1 Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Show that with respect to each choice of bases of  $V$  and  $W$ , the matrix of  $T$  has at least  $\dim \text{range } T$  nonzero entries.
- 2 Suppose  $D \in \mathcal{L}(\mathcal{P}_3(\mathbf{R}), \mathcal{P}_2(\mathbf{R}))$  is the differentiation map defined by  $Dp = p'$ . Find a basis of  $\mathcal{P}_3(\mathbf{R})$  and a basis of  $\mathcal{P}_2(\mathbf{R})$  such that the matrix of  $D$  with respect to these bases is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

[Compare the exercise above to Example 3.34.

The next exercise generalizes the exercise above.]

- 3 Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Prove that there exist a basis of  $V$  and a basis of  $W$  such that with respect to these bases, all entries of  $\mathcal{M}(T)$  are 0 except that the entries in row  $j$ , column  $j$ , equal 1 for  $1 \leq j \leq \dim \text{range } T$ .
- 4 Suppose  $v_1, \dots, v_m$  is a basis of  $V$  and  $W$  is finite-dimensional. Suppose  $T \in \mathcal{L}(V, W)$ . Prove that there exists a basis  $w_1, \dots, w_n$  of  $W$  such that all the entries in the first column of  $\mathcal{M}(T)$  (with respect to the bases  $v_1, \dots, v_m$  and  $w_1, \dots, w_n$ ) are 0 except for possibly a 1 in the first row, first column.

[In this exercise, unlike Exercise 3, you are given the basis of  $V$  instead of being able to choose a basis of  $V$ .]

- 5 Suppose  $w_1, \dots, w_n$  is a basis of  $W$  and  $V$  is finite-dimensional. Suppose  $T \in \mathcal{L}(V, W)$ . Prove that there exists a basis  $v_1, \dots, v_m$  of  $V$  such that all the entries in the first row of  $\mathcal{M}(T)$  (with respect to the bases  $v_1, \dots, v_m$  and  $w_1, \dots, w_n$ ) are 0 except for possibly a 1 in the first row, first column.

[In this exercise, unlike Exercise 3, you are given the basis of  $W$  instead of being able to choose a basis of  $W$ .]

- 6** Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Prove that  $\dim \text{range } T = 1$  if and only if there exist a basis of  $V$  and a basis of  $W$  such that with respect to these bases, all entries of  $\mathcal{M}(T)$  equal 1.
- 7** Verify 3.36.
- 8** Verify 3.38.
- 9** Prove 3.52.
- 10** Suppose  $A$  is an  $m$ -by- $n$  matrix and  $C$  is an  $n$ -by- $p$  matrix. Prove that

$$(AC)_{j,\cdot} = A_{j,\cdot} C$$

for  $1 \leq j \leq m$ . In other words, show that row  $j$  of  $AC$  equals (row  $j$  of  $A$ ) times  $C$ .

- 11** Suppose  $a = (a_1 \ \cdots \ a_n)$  is a 1-by- $n$  matrix and  $C$  is an  $n$ -by- $p$  matrix. Prove that

$$aC = a_1 C_{1,\cdot} + \cdots + a_n C_{n,\cdot}.$$

In other words, show that  $aC$  is a linear combination of the rows of  $C$ , with the scalars that multiply the rows coming from  $a$ .

- 12** Give an example with 2-by-2 matrices to show that matrix multiplication is not commutative. In other words, find 2-by-2 matrices  $A$  and  $C$  such that  $AC \neq CA$ .
- 13** Prove that the distributive property holds for matrix addition and matrix multiplication. In other words, suppose  $A, B, C, D, E$ , and  $F$  are matrices whose sizes are such that  $A(B + C)$  and  $(D + E)F$  make sense. Prove that  $AB + AC$  and  $DF + EF$  both make sense and that  $A(B + C) = AB + AC$  and  $(D + E)F = DF + EF$ .
- 14** Prove that matrix multiplication is associative. In other words, suppose  $A, B$ , and  $C$  are matrices whose sizes are such that  $(AB)C$  makes sense. Prove that  $A(BC)$  makes sense and that  $(AB)C = A(BC)$ .
- 15** Suppose  $A$  is an  $n$ -by- $n$  matrix and  $1 \leq j, k \leq n$ . Show that the entry in row  $j$ , column  $k$ , of  $A^3$  (which is defined to mean  $AAA$ ) is

$$\sum_{p=1}^n \sum_{r=1}^n A_{j,p} A_{p,r} A_{r,k}.$$

## 3.D Invertibility and Isomorphic Vector Spaces

### Invertible Linear Maps

We begin this section by defining the notions of invertible and inverse in the context of linear maps.

#### 3.53 Definition *invertible, inverse*

- A linear map  $T \in \mathcal{L}(V, W)$  is called ***invertible*** if there exists a linear map  $S \in \mathcal{L}(W, V)$  such that  $ST$  equals the identity map on  $V$  and  $TS$  equals the identity map on  $W$ .
- A linear map  $S \in \mathcal{L}(W, V)$  satisfying  $ST = I$  and  $TS = I$  is called an ***inverse*** of  $T$  (note that the first  $I$  is the identity map on  $V$  and the second  $I$  is the identity map on  $W$ ).

#### 3.54 Inverse is unique

An invertible linear map has a unique inverse.

**Proof** Suppose  $T \in \mathcal{L}(V, W)$  is invertible and  $S_1$  and  $S_2$  are inverses of  $T$ . Then

$$S_1 = S_1 I = S_1(TS_2) = (S_1 T)S_2 = IS_2 = S_2.$$

Thus  $S_1 = S_2$ . ■

Now that we know that the inverse is unique, we can give it a notation.

#### 3.55 Notation $T^{-1}$

If  $T$  is invertible, then its inverse is denoted by  $T^{-1}$ . In other words, if  $T \in \mathcal{L}(V, W)$  is invertible, then  $T^{-1}$  is the unique element of  $\mathcal{L}(W, V)$  such that  $T^{-1}T = I$  and  $TT^{-1} = I$ .

The following result characterizes the invertible linear maps.

#### 3.56 Invertibility is equivalent to injectivity and surjectivity

A linear map is invertible if and only if it is injective and surjective.

**Proof** Suppose  $T \in \mathcal{L}(V, W)$ . We need to show that  $T$  is invertible if and only if it is injective and surjective.

First suppose  $T$  is invertible. To show that  $T$  is injective, suppose  $u, v \in V$  and  $Tu = Tv$ . Then

$$u = T^{-1}(Tu) = T^{-1}(Tv) = v,$$

so  $u = v$ . Hence  $T$  is injective.

We are still assuming that  $T$  is invertible. Now we want to prove that  $T$  is surjective. To do this, let  $w \in W$ . Then  $w = T(T^{-1}w)$ , which shows that  $w$  is in the range of  $T$ . Thus  $\text{range } T = W$ . Hence  $T$  is surjective, completing this direction of the proof.

Now suppose  $T$  is injective and surjective. We want to prove that  $T$  is invertible. For each  $w \in W$ , define  $Sw$  to be the unique element of  $V$  such that  $T(Sw) = w$  (the existence and uniqueness of such an element follow from the surjectivity and injectivity of  $T$ ). Clearly  $T \circ S$  equals the identity map on  $W$ .

To prove that  $S \circ T$  equals the identity map on  $V$ , let  $v \in V$ . Then

$$T((S \circ T)v) = (T \circ S)(Tv) = I(Tv) = Tv.$$

This equation implies that  $(S \circ T)v = v$  (because  $T$  is injective). Thus  $S \circ T$  equals the identity map on  $V$ .

To complete the proof, we need to show that  $S$  is linear. To do this, suppose  $w_1, w_2 \in W$ . Then

$$T(Sw_1 + Sw_2) = T(Sw_1) + T(Sw_2) = w_1 + w_2.$$

Thus  $Sw_1 + Sw_2$  is the unique element of  $V$  that  $T$  maps to  $w_1 + w_2$ . By the definition of  $S$ , this implies that  $S(w_1 + w_2) = Sw_1 + Sw_2$ . Hence  $S$  satisfies the additive property required for linearity.

The proof of homogeneity is similar. Specifically, if  $w \in W$  and  $\lambda \in \mathbf{F}$ , then

$$T(\lambda Sw) = \lambda T(Sw) = \lambda w.$$

Thus  $\lambda Sw$  is the unique element of  $V$  that  $T$  maps to  $\lambda w$ . By the definition of  $S$ , this implies that  $S(\lambda w) = \lambda Sw$ . Hence  $S$  is linear, as desired. ■

### 3.57 Example linear maps that are not invertible

- The multiplication by  $x^2$  linear map from  $\mathcal{P}(\mathbf{R})$  to  $\mathcal{P}(\mathbf{R})$  (see 3.4) is not invertible because it is not surjective (1 is not in the range).
- The backward shift linear map from  $\mathbf{F}^\infty$  to  $\mathbf{F}^\infty$  (see 3.4) is not invertible because it is not injective  $[(1, 0, 0, 0, \dots)]$  is in the null space].

## Isomorphic Vector Spaces

The next definition captures the idea of two vector spaces that are essentially the same, except for the names of the elements of the vector spaces.

### 3.58 Definition *isomorphism, isomorphic*

- An ***isomorphism*** is an invertible linear map.
- Two vector spaces are called ***isomorphic*** if there is an isomorphism from one vector space onto the other one.

Think of an isomorphism  $T : V \rightarrow W$  as relabeling  $v \in V$  as  $Tv \in W$ . This viewpoint explains why two isomorphic vector spaces have the same vector space properties. The terms “isomorphism” and “invertible linear map” mean

*The Greek word **isos** means equal; the Greek word **morph** means shape. Thus **isomorphic** literally means equal shape.*

the same thing. Use “isomorphism” when you want to emphasize that the two spaces are essentially the same.

### 3.59 Dimension shows whether vector spaces are isomorphic

Two finite-dimensional vector spaces over  $\mathbf{F}$  are isomorphic if and only if they have the same dimension.

**Proof** First suppose  $V$  and  $W$  are isomorphic finite-dimensional vector spaces. Thus there exists an isomorphism  $T$  from  $V$  onto  $W$ . Because  $T$  is invertible, we have  $\text{null } T = \{0\}$  and  $\text{range } T = W$ . Thus  $\dim \text{null } T = 0$  and  $\dim \text{range } T = \dim W$ . The formula

$$\dim V = \dim \text{null } T + \dim \text{range } T$$

(the Fundamental Theorem of Linear Maps, which is 3.22) thus becomes the equation  $\dim V = \dim W$ , completing the proof in one direction.

To prove the other direction, suppose  $V$  and  $W$  are finite-dimensional vector spaces with the same dimension. Let  $v_1, \dots, v_n$  be a basis of  $V$  and  $w_1, \dots, w_n$  be a basis of  $W$ . Let  $T \in \mathcal{L}(V, W)$  be defined by

$$T(c_1v_1 + \cdots + c_nv_n) = c_1w_1 + \cdots + c_nw_n.$$

Then  $T$  is a well-defined linear map because  $v_1, \dots, v_n$  is a basis of  $V$  (see 3.5). Also,  $T$  is surjective because  $w_1, \dots, w_n$  spans  $W$ . Furthermore,  $\text{null } T = \{0\}$  because  $w_1, \dots, w_n$  is linearly independent; thus  $T$  is injective. Because  $T$  is injective and surjective, it is an isomorphism (see 3.56). Hence  $V$  and  $W$  are isomorphic, as desired. ■

The previous result implies that each finite-dimensional vector space  $V$  is isomorphic to  $\mathbf{F}^n$ , where  $n = \dim V$ .

If  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_m$  is a basis of  $W$ , then for each  $T \in \mathcal{L}(V, W)$ , we have a matrix  $\mathcal{M}(T) \in \mathbf{F}^{m,n}$ . In other words, once bases have been fixed for  $V$  and  $W$ ,  $\mathcal{M}$  becomes a function from  $\mathcal{L}(V, W)$  to  $\mathbf{F}^{m,n}$ . Notice that 3.36 and 3.38 show that  $\mathcal{M}$  is a linear map. This linear map is actually invertible, as we now show.

*Because every finite-dimensional vector space is isomorphic to some  $\mathbf{F}^n$ , why not just study  $\mathbf{F}^n$  instead of more general vector spaces? To answer this question, note that an investigation of  $\mathbf{F}^n$  would soon lead to other vector spaces. For example, we would encounter the null space and range of linear maps. Although each of these vector spaces is isomorphic to some  $\mathbf{F}^n$ , thinking of them that way often adds complexity but no new insight.*

### 3.60 $\mathcal{L}(V, W)$ and $\mathbf{F}^{m,n}$ are isomorphic

Suppose  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_m$  is a basis of  $W$ . Then  $\mathcal{M}$  is an isomorphism between  $\mathcal{L}(V, W)$  and  $\mathbf{F}^{m,n}$ .

**Proof** We already noted that  $\mathcal{M}$  is linear. We need to prove that  $\mathcal{M}$  is injective and surjective. Both are easy. We begin with injectivity. If  $T \in \mathcal{L}(V, W)$  and  $\mathcal{M}(T) = 0$ , then  $Tv_k = 0$  for  $k = 1, \dots, n$ . Because  $v_1, \dots, v_n$  is a basis of  $V$ , this implies  $T = 0$ . Thus  $\mathcal{M}$  is injective (by 3.16).

To prove that  $\mathcal{M}$  is surjective, suppose  $A \in \mathbf{F}^{m,n}$ . Let  $T$  be the linear map from  $V$  to  $W$  such that

$$Tv_k = \sum_{j=1}^m A_{j,k} w_j$$

for  $k = 1, \dots, n$  (see 3.5). Obviously  $\mathcal{M}(T)$  equals  $A$ , and thus the range of  $\mathcal{M}$  equals  $\mathbf{F}^{m,n}$ , as desired. ■

Now we can determine the dimension of the vector space of linear maps from one finite-dimensional vector space to another.

### 3.61 $\dim \mathcal{L}(V, W) = (\dim V)(\dim W)$

Suppose  $V$  and  $W$  are finite-dimensional. Then  $\mathcal{L}(V, W)$  is finite-dimensional and

$$\dim \mathcal{L}(V, W) = (\dim V)(\dim W).$$

**Proof** This follows from 3.60, 3.59, and 3.40. ■

## Linear Maps Thought of as Matrix Multiplication

Previously we defined the matrix of a linear map. Now we define the matrix of a vector.

### 3.62 Definition *matrix of a vector*, $\mathcal{M}(v)$

Suppose  $v \in V$  and  $v_1, \dots, v_n$  is a basis of  $V$ . The *matrix of  $v$*  with respect to this basis is the  $n$ -by-1 matrix

$$\mathcal{M}(v) = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix},$$

where  $c_1, \dots, c_n$  are the scalars such that

$$v = c_1 v_1 + \cdots + c_n v_n.$$

The matrix  $\mathcal{M}(v)$  of a vector  $v \in V$  depends on the basis  $v_1, \dots, v_n$  of  $V$ , as well as on  $v$ . However, the basis should be clear from the context and thus it is not included in the notation.

---

### 3.63 Example *matrix of a vector*

- The matrix of  $2 - 7x + 5x^3$  with respect to the standard basis of  $\mathcal{P}_3(\mathbf{R})$  is

$$\begin{pmatrix} 2 \\ -7 \\ 0 \\ 5 \end{pmatrix}.$$

- The matrix of a vector  $x \in \mathbf{F}^n$  with respect to the standard basis is obtained by writing the coordinates of  $x$  as the entries in an  $n$ -by-1 matrix. In other words, if  $x = (x_1, \dots, x_n) \in \mathbf{F}^n$ , then

$$\mathcal{M}(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$


---

Occasionally we want to think of elements of  $V$  as relabeled to be  $n$ -by-1 matrices. Once a basis  $v_1, \dots, v_n$  is chosen, the function  $\mathcal{M}$  that takes  $v \in V$  to  $\mathcal{M}(v)$  is an isomorphism of  $V$  onto  $\mathbf{F}^{n,1}$  that implements this relabeling.

Recall that if  $A$  is an  $m$ -by- $n$  matrix, then  $A_{\cdot,k}$  denotes the  $k^{\text{th}}$  column of  $A$ , thought of as an  $m$ -by-1 matrix. In the next result,  $\mathcal{M}(v_k)$  is computed with respect to the basis  $w_1, \dots, w_m$  of  $W$ .

### 3.64 $\mathcal{M}(T)_{\cdot,k} = \mathcal{M}(v_k)$ .

Suppose  $T \in \mathcal{L}(V, W)$  and  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_m$  is a basis of  $W$ . Let  $1 \leq k \leq n$ . Then the  $k^{\text{th}}$  column of  $\mathcal{M}(T)$ , which is denoted by  $\mathcal{M}(T)_{\cdot,k}$ , equals  $\mathcal{M}(v_k)$ .

**Proof** The desired result follows immediately from the definitions of  $\mathcal{M}(T)$  and  $\mathcal{M}(v_k)$ . ■

The next result shows how the notions of the matrix of a linear map, the matrix of a vector, and matrix multiplication fit together.

### 3.65 Linear maps act like matrix multiplication

Suppose  $T \in \mathcal{L}(V, W)$  and  $v \in V$ . Suppose  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_m$  is a basis of  $W$ . Then

$$\mathcal{M}(Tv) = \mathcal{M}(T)\mathcal{M}(v).$$

**Proof** Suppose  $v = c_1v_1 + \dots + c_nv_n$ , where  $c_1, \dots, c_n \in \mathbf{F}$ . Thus

$$3.66 \quad Tv = c_1Tv_1 + \dots + c_nTv_n.$$

Hence

$$\begin{aligned} \mathcal{M}(Tv) &= c_1\mathcal{M}(Tv_1) + \dots + c_n\mathcal{M}(Tv_n) \\ &= c_1\mathcal{M}(T)_{\cdot,1} + \dots + c_n\mathcal{M}(T)_{\cdot,n} \\ &= \mathcal{M}(T)\mathcal{M}(v), \end{aligned}$$

where the first equality follows from 3.66 and the linearity of  $\mathcal{M}$ , the second equality comes from 3.64, and the last equality comes from 3.52. ■

Each  $m$ -by- $n$  matrix  $A$  induces a linear map from  $\mathbf{F}^{n,1}$  to  $\mathbf{F}^{m,1}$ , namely the matrix multiplication function that takes  $x \in \mathbf{F}^{n,1}$  to  $Ax \in \mathbf{F}^{m,1}$ . The result above can be used to think of every linear map (from one finite-dimensional vector space to another finite-dimensional vector space) as a matrix multiplication map after suitable relabeling via the isomorphisms given by  $\mathcal{M}$ . Specifically, if  $T \in \mathcal{L}(V, W)$  and we identify  $v \in V$  with  $\mathcal{M}(v) \in \mathbf{F}^{n,1}$ , then the result above says that we can identify  $Tv$  with  $\mathcal{M}(T)\mathcal{M}(v)$ .

Because the result above allows us to think (via isomorphisms) of each linear map as multiplication on  $\mathbf{F}^{n,1}$  by some matrix  $A$ , keep in mind that the specific matrix  $A$  depends not only on the linear map but also on the choice of bases. One of the themes of many of the most important results in later chapters will be the choice of a basis that makes the matrix  $A$  as simple as possible.

In this book, we concentrate on linear maps rather than on matrices. However, sometimes thinking of linear maps as matrices (or thinking of matrices as linear maps) gives important insights that we will find useful.

## Operators

Linear maps from a vector space to itself are so important that they get a special name and special notation.

### 3.67 Definition *operator, $\mathcal{L}(V)$*

- A linear map from a vector space to itself is called an *operator*.
- The notation  $\mathcal{L}(V)$  denotes the set of all operators on  $V$ . In other words,  $\mathcal{L}(V) = \mathcal{L}(V, V)$ .

*The deepest and most important parts of linear algebra, as well as most of the rest of this book, deal with operators.*

A linear map is invertible if it is injective and surjective. For an operator, you might wonder whether injectivity alone, or surjectivity alone, is enough to imply invertibility. On

infinite-dimensional vector spaces, neither condition alone implies invertibility, as illustrated by the next example, which uses two familiar operators from Example 3.4.

---

### 3.68 Example *neither injectivity nor surjectivity implies invertibility*

- The multiplication by  $x^2$  operator on  $\mathcal{P}(\mathbf{R})$  is injective but not surjective.
  - The backward shift operator on  $\mathbf{F}^\infty$  is surjective but not injective.
- 

In view of the example above, the next result is remarkable—it states that for operators on a finite-dimensional vector space, either injectivity or surjectivity alone implies the other condition. Often it is easier to check that an operator on a finite-dimensional vector space is injective, and then we get surjectivity for free.

### 3.69 Injectivity is equivalent to surjectivity in finite dimensions

Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $T$  is invertible;
- (b)  $T$  is injective;
- (c)  $T$  is surjective.

**Proof** Clearly (a) implies (b).

Now suppose (b) holds, so that  $T$  is injective. Thus  $\text{null } T = \{0\}$  (by 3.16). From the Fundamental Theorem of Linear Maps (3.22) we have

$$\begin{aligned}\dim \text{range } T &= \dim V - \dim \text{null } T \\ &= \dim V.\end{aligned}$$

Thus  $\text{range } T$  equals  $V$ . Thus  $T$  is surjective. Hence (b) implies (c).

Now suppose (c) holds, so that  $T$  is surjective. Thus  $\text{range } T = V$ . From the Fundamental Theorem of Linear Maps (3.22) we have

$$\begin{aligned}\dim \text{null } T &= \dim V - \dim \text{range } T \\ &= 0.\end{aligned}$$

Thus  $\text{null } T$  equals  $\{0\}$ . Thus  $T$  is injective (by 3.16), and so  $T$  is invertible (we already knew that  $T$  was surjective). Hence (c) implies (a), completing the proof. ■

The next example illustrates the power of the previous result. Although it is possible to prove the result in the example below without using linear algebra, the proof using linear algebra is cleaner and easier.

---

**3.70 Example** Show that for each polynomial  $q \in \mathcal{P}(\mathbf{R})$ , there exists a polynomial  $p \in \mathcal{P}(\mathbf{R})$  with  $((x^2 + 5x + 7)p)'' = q$ .

**Solution** Example 3.68 shows that the magic of 3.69 does not apply to the infinite-dimensional vector space  $\mathcal{P}(\mathbf{R})$ . However, each nonzero polynomial  $q$  has some degree  $m$ . By restricting attention to  $\mathcal{P}_m(\mathbf{R})$ , we can work with a finite-dimensional vector space.

Suppose  $q \in \mathcal{P}_m(\mathbf{R})$ . Define  $T: \mathcal{P}_m(\mathbf{R}) \rightarrow \mathcal{P}_m(\mathbf{R})$  by

$$Tp = ((x^2 + 5x + 7)p)''.$$

Multiplying a nonzero polynomial by  $(x^2 + 5x + 7)$  increases the degree by 2, and then differentiating twice reduces the degree by 2. Thus  $T$  is indeed an operator on  $\mathcal{P}_m(\mathbf{R})$ .

Every polynomial whose second derivative equals 0 is of the form  $ax + b$ , where  $a, b \in \mathbf{R}$ . Thus  $\text{null } T = \{0\}$ . Hence  $T$  is injective.

Now 3.69 implies that  $T$  is surjective. Thus there exists a polynomial  $p \in \mathcal{P}_m(\mathbf{R})$  such that  $((x^2 + 5x + 7)p)'' = q$ , as desired.

---

Exercise 30 in Section 6.A gives a similar but more spectacular application of 3.69. The result in that exercise is quite difficult to prove without using linear algebra.

## EXERCISES 3.D

---

- 1 Suppose  $T \in \mathcal{L}(U, V)$  and  $S \in \mathcal{L}(V, W)$  are both invertible linear maps. Prove that  $ST \in \mathcal{L}(U, W)$  is invertible and that  $(ST)^{-1} = T^{-1}S^{-1}$ .
- 2 Suppose  $V$  is finite-dimensional and  $\dim V > 1$ . Prove that the set of noninvertible operators on  $V$  is not a subspace of  $\mathcal{L}(V)$ .
- 3 Suppose  $V$  is finite-dimensional,  $U$  is a subspace of  $V$ , and  $S \in \mathcal{L}(U, V)$ . Prove there exists an invertible operator  $T \in \mathcal{L}(V)$  such that  $Tu = Su$  for every  $u \in U$  if and only if  $S$  is injective.
- 4 Suppose  $W$  is finite-dimensional and  $T_1, T_2 \in \mathcal{L}(V, W)$ . Prove that  $\text{null } T_1 = \text{null } T_2$  if and only if there exists an invertible operator  $S \in \mathcal{L}(W)$  such that  $T_1 = ST_2$ .
- 5 Suppose  $V$  is finite-dimensional and  $T_1, T_2 \in \mathcal{L}(V, W)$ . Prove that  $\text{range } T_1 = \text{range } T_2$  if and only if there exists an invertible operator  $S \in \mathcal{L}(V)$  such that  $T_1 = T_2S$ .
- 6 Suppose  $V$  and  $W$  are finite-dimensional and  $T_1, T_2 \in \mathcal{L}(V, W)$ . Prove that there exist invertible operators  $R \in \mathcal{L}(V)$  and  $S \in \mathcal{L}(W)$  such that  $T_1 = ST_2R$  if and only if  $\dim \text{null } T_1 = \dim \text{null } T_2$ .
- 7 Suppose  $V$  and  $W$  are finite-dimensional. Let  $v \in V$ . Let

$$E = \{T \in \mathcal{L}(V, W) : Tv = 0\}.$$

- (a) Show that  $E$  is a subspace of  $\mathcal{L}(V, W)$ .
- (b) Suppose  $v \neq 0$ . What is  $\dim E$ ?

- 8 Suppose  $V$  is finite-dimensional and  $T: V \rightarrow W$  is a surjective linear map of  $V$  onto  $W$ . Prove that there is a subspace  $U$  of  $V$  such that  $T|_U$  is an isomorphism of  $U$  onto  $W$ . (Here  $T|_U$  means the function  $T$  restricted to  $U$ . In other words,  $T|_U$  is the function whose domain is  $U$ , with  $T|_U$  defined by  $T|_U(u) = Tu$  for every  $u \in U$ .)
- 9 Suppose  $V$  is finite-dimensional and  $S, T \in \mathcal{L}(V)$ . Prove that  $ST$  is invertible if and only if both  $S$  and  $T$  are invertible.
- 10 Suppose  $V$  is finite-dimensional and  $S, T \in \mathcal{L}(V)$ . Prove that  $ST = I$  if and only if  $TS = I$ .
- 11 Suppose  $V$  is finite-dimensional and  $S, T, U \in \mathcal{L}(V)$  and  $STU = I$ . Show that  $T$  is invertible and that  $T^{-1} = US$ .
- 12 Show that the result in the previous exercise can fail without the hypothesis that  $V$  is finite-dimensional.
- 13 Suppose  $V$  is a finite-dimensional vector space and  $R, S, T \in \mathcal{L}(V)$  are such that  $RST$  is surjective. Prove that  $S$  is injective.
- 14 Suppose  $v_1, \dots, v_n$  is a basis of  $V$ . Prove that the map  $T: V \rightarrow \mathbf{F}^{n,1}$  defined by

$$Tv = \mathcal{M}(v)$$

is an isomorphism of  $V$  onto  $\mathbf{F}^{n,1}$ ; here  $\mathcal{M}(v)$  is the matrix of  $v \in V$  with respect to the basis  $v_1, \dots, v_n$ .

- 15 Prove that every linear map from  $\mathbf{F}^{n,1}$  to  $\mathbf{F}^{m,1}$  is given by a matrix multiplication. In other words, prove that if  $T \in \mathcal{L}(\mathbf{F}^{n,1}, \mathbf{F}^{m,1})$ , then there exists an  $m$ -by- $n$  matrix  $A$  such that  $Tx = Ax$  for every  $x \in \mathbf{F}^{n,1}$ .
- 16 Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V)$ . Prove that  $T$  is a scalar multiple of the identity if and only if  $ST = TS$  for every  $S \in \mathcal{L}(V)$ .
- 17 Suppose  $V$  is finite-dimensional and  $\mathcal{E}$  is a subspace of  $\mathcal{L}(V)$  such that  $ST \in \mathcal{E}$  and  $TS \in \mathcal{E}$  for all  $S \in \mathcal{L}(V)$  and all  $T \in \mathcal{E}$ . Prove that  $\mathcal{E} = \{0\}$  or  $\mathcal{E} = \mathcal{L}(V)$ .
- 18 Show that  $V$  and  $\mathcal{L}(\mathbf{F}, V)$  are isomorphic vector spaces.
- 19 Suppose  $T \in \mathcal{L}(\mathcal{P}(\mathbf{R}))$  is such that  $T$  is injective and  $\deg Tp \leq \deg p$  for every nonzero polynomial  $p \in \mathcal{P}(\mathbf{R})$ .
- Prove that  $T$  is surjective.
  - Prove that  $\deg Tp = \deg p$  for every nonzero  $p \in \mathcal{P}(\mathbf{R})$ .

- 20** Suppose  $n$  is a positive integer and  $A_{i,j} \in \mathbf{F}$  for  $i, j = 1, \dots, n$ . Prove that the following are equivalent (note that in both parts below, the number of equations equals the number of variables):

- (a) The trivial solution  $x_1 = \dots = x_n = 0$  is the only solution to the homogeneous system of equations

$$\sum_{k=1}^n A_{1,k}x_k = 0$$

⋮

$$\sum_{k=1}^n A_{n,k}x_k = 0.$$

- (b) For every  $c_1, \dots, c_n \in \mathbf{F}$ , there exists a solution to the system of equations

$$\sum_{k=1}^n A_{1,k}x_k = c_1$$

⋮

$$\sum_{k=1}^n A_{n,k}x_k = c_n.$$

## 3.E Products and Quotients of Vector Spaces

### Products of Vector Spaces

As usual when dealing with more than one vector space, all the vector spaces in use should be over the same field.

#### 3.71 Definition *product of vector spaces*

Suppose  $V_1, \dots, V_m$  are vector spaces over  $\mathbf{F}$ .

- The **product**  $V_1 \times \dots \times V_m$  is defined by

$$V_1 \times \dots \times V_m = \{(v_1, \dots, v_m) : v_1 \in V_1, \dots, v_m \in V_m\}.$$

- Addition on  $V_1 \times \dots \times V_m$  is defined by

$$(u_1, \dots, u_m) + (v_1, \dots, v_m) = (u_1 + v_1, \dots, u_m + v_m).$$

- Scalar multiplication on  $V_1 \times \dots \times V_m$  is defined by

$$\lambda(v_1, \dots, v_m) = (\lambda v_1, \dots, \lambda v_m).$$

---

**3.72 Example** Elements of  $\mathcal{P}_2(\mathbf{R}) \times \mathbf{R}^3$  are lists of length 2, with the first item in the list an element of  $\mathcal{P}_2(\mathbf{R})$  and the second item in the list an element of  $\mathbf{R}^3$ .

For example,  $(5 - 6x + 4x^2, (3, 8, 7)) \in \mathcal{P}_2(\mathbf{R}) \times \mathbf{R}^3$ .

---

The next result should be interpreted to mean that the product of vector spaces is a vector space with the operations of addition and scalar multiplication as defined above.

#### 3.73 Product of vector spaces is a vector space

Suppose  $V_1, \dots, V_m$  are vector spaces over  $\mathbf{F}$ . Then  $V_1 \times \dots \times V_m$  is a vector space over  $\mathbf{F}$ .

The proof of the result above is left to the reader. Note that the additive identity of  $V_1 \times \dots \times V_m$  is  $(0, \dots, 0)$ , where the 0 in the  $j^{\text{th}}$  slot is the additive identity of  $V_j$ . The additive inverse of  $(v_1, \dots, v_m) \in V_1 \times \dots \times V_m$  is  $(-v_1, \dots, -v_m)$ .

**3.74 Example** Is  $\mathbf{R}^2 \times \mathbf{R}^3$  equal to  $\mathbf{R}^5$ ? Is  $\mathbf{R}^2 \times \mathbf{R}^3$  isomorphic to  $\mathbf{R}^5$ ?

**Solution** Elements of  $\mathbf{R}^2 \times \mathbf{R}^3$  are lists  $((x_1, x_2), (x_3, x_4, x_5))$ , where  $x_1, x_2, x_3, x_4, x_5 \in \mathbf{R}$ .

Elements of  $\mathbf{R}^5$  are lists  $(x_1, x_2, x_3, x_4, x_5)$ , where  $x_1, x_2, x_3, x_4, x_5 \in \mathbf{R}$ .

Although these look almost the same, they are not the same kind of object. Elements of  $\mathbf{R}^2 \times \mathbf{R}^3$  are lists of length 2 (with the first item itself a list of length 2 and the second item a list of length 3), and elements of  $\mathbf{R}^5$  are lists of length 5. Thus  $\mathbf{R}^2 \times \mathbf{R}^3$  does not equal  $\mathbf{R}^5$ .

The linear map that takes a vector  $((x_1, x_2), (x_3, x_4, x_5)) \in \mathbf{R}^2 \times \mathbf{R}^3$  to  $(x_1, x_2, x_3, x_4, x_5) \in \mathbf{R}^5$  is clearly an isomorphism of  $\mathbf{R}^2 \times \mathbf{R}^3$  onto  $\mathbf{R}^5$ . Thus these two vector spaces are isomorphic.

In this case, the isomorphism is so natural that we should think of it as a relabeling. Some people would even informally say that  $\mathbf{R}^2 \times \mathbf{R}^3$  equals  $\mathbf{R}^5$ , which is not technically correct but which captures the spirit of identification via relabeling.

The next example illustrates the idea of the proof of 3.76.

**3.75 Example** Find a basis of  $\mathcal{P}_2(\mathbf{R}) \times \mathbf{R}^2$ .

**Solution** Consider this list of length 5 of elements of  $\mathcal{P}_2(\mathbf{R}) \times \mathbf{R}^2$ :

$$(1, (0, 0)), (x, (0, 0)), (x^2, (0, 0)), (0, (1, 0)), (0, (0, 1)).$$

The list above is linearly independent and it spans  $\mathcal{P}_2(\mathbf{R}) \times \mathbf{R}^2$ . Thus it is a basis of  $\mathcal{P}_2(\mathbf{R}) \times \mathbf{R}^2$ .

### 3.76 Dimension of a product is the sum of dimensions

Suppose  $V_1, \dots, V_m$  are finite-dimensional vector spaces. Then  $V_1 \times \dots \times V_m$  is finite-dimensional and

$$\dim(V_1 \times \dots \times V_m) = \dim V_1 + \dots + \dim V_m.$$

**Proof** Choose a basis of each  $V_j$ . For each basis vector of each  $V_j$ , consider the element of  $V_1 \times \dots \times V_m$  that equals the basis vector in the  $j^{\text{th}}$  slot and 0 in the other slots. The list of all such vectors is linearly independent and spans  $V_1 \times \dots \times V_m$ . Thus it is a basis of  $V_1 \times \dots \times V_m$ . The length of this basis is  $\dim V_1 + \dots + \dim V_m$ , as desired. ■

## Products and Direct Sums

In the next result, the map  $\Gamma$  is surjective by the definition of  $U_1 + \cdots + U_m$ . Thus the last word in the result below could be changed from “injective” to “invertible”.

### 3.77 Products and direct sums

Suppose that  $U_1, \dots, U_m$  are subspaces of  $V$ . Define a linear map  $\Gamma : U_1 \times \cdots \times U_m \rightarrow U_1 + \cdots + U_m$  by

$$\Gamma(u_1, \dots, u_m) = u_1 + \cdots + u_m.$$

Then  $U_1 + \cdots + U_m$  is a direct sum if and only if  $\Gamma$  is injective.

**Proof** The linear map  $\Gamma$  is injective if and only if the only way to write 0 as a sum  $u_1 + \cdots + u_m$ , where each  $u_j$  is in  $U_j$ , is by taking each  $u_j$  equal to 0. Thus 1.44 shows that  $\Gamma$  is injective if and only if  $U_1 + \cdots + U_m$  is a direct sum, as desired. ■

### 3.78 A sum is a direct sum if and only if dimensions add up

Suppose  $V$  is finite-dimensional and  $U_1, \dots, U_m$  are subspaces of  $V$ . Then  $U_1 + \cdots + U_m$  is a direct sum if and only if

$$\dim(U_1 + \cdots + U_m) = \dim U_1 + \cdots + \dim U_m.$$

**Proof** The map  $\Gamma$  in 3.77 is surjective. Thus by the Fundamental Theorem of Linear Maps (3.22),  $\Gamma$  is injective if and only if

$$\dim(U_1 + \cdots + U_m) = \dim(U_1 \times \cdots \times U_m).$$

Combining 3.77 and 3.76 now shows that  $U_1 + \cdots + U_m$  is a direct sum if and only if

$$\dim(U_1 + \cdots + U_m) = \dim U_1 + \cdots + \dim U_m,$$

as desired. ■

In the special case  $m = 2$ , an alternative proof that  $U_1 + U_2$  is a direct sum if and only if  $\dim(U_1 + U_2) = \dim U_1 + \dim U_2$  can be obtained by combining 1.45 and 2.43.

## Quotients of Vector Spaces

We begin our approach to quotient spaces by defining the sum of a vector and a subspace.

### 3.79 Definition $v + U$

Suppose  $v \in V$  and  $U$  is a subspace of  $V$ . Then  $v + U$  is the subset of  $V$  defined by

$$v + U = \{v + u : u \in U\}.$$

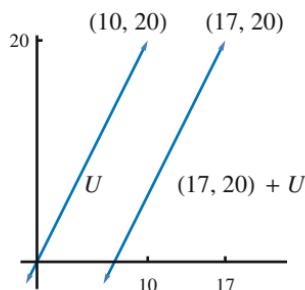
### 3.80 Example Suppose

$$U = \{(x, 2x) \in \mathbf{R}^2 : x \in \mathbf{R}\}.$$

Then  $U$  is the line in  $\mathbf{R}^2$  through the origin with slope 2. Thus

$$(17, 20) + U$$

is the line in  $\mathbf{R}^2$  that contains the point  $(17, 20)$  and has slope 2.



### 3.81 Definition *affine subset, parallel*

- An *affine subset* of  $V$  is a subset of  $V$  of the form  $v + U$  for some  $v \in V$  and some subspace  $U$  of  $V$ .
- For  $v \in V$  and  $U$  a subspace of  $V$ , the affine subset  $v + U$  is said to be *parallel* to  $U$ .

### 3.82 Example *parallel affine subsets*

- In Example 3.80 above, all the lines in  $\mathbf{R}^2$  with slope 2 are parallel to  $U$ .
- If  $U = \{(x, y, 0) \in \mathbf{R}^3 : x, y \in \mathbf{R}\}$ , then the affine subsets of  $\mathbf{R}^3$  parallel to  $U$  are the planes in  $\mathbf{R}^3$  that are parallel to the  $xy$ -plane  $U$  in the usual sense.

**Important:** With the definition of *parallel* given in 3.81, no line in  $\mathbf{R}^3$  is considered to be an affine subset that is parallel to the plane  $U$ .

### 3.83 Definition *quotient space*, $V/U$

Suppose  $U$  is a subspace of  $V$ . Then the *quotient space*  $V/U$  is the set of all affine subsets of  $V$  parallel to  $U$ . In other words,

$$V/U = \{v + U : v \in V\}.$$

### 3.84 Example *quotient spaces*

- If  $U = \{(x, 2x) \in \mathbf{R}^2 : x \in \mathbf{R}\}$ , then  $\mathbf{R}^2/U$  is the set of all lines in  $\mathbf{R}^2$  that have slope 2.
- If  $U$  is a line in  $\mathbf{R}^3$  containing the origin, then  $\mathbf{R}^3/U$  is the set of all lines in  $\mathbf{R}^3$  parallel to  $U$ .
- If  $U$  is a plane in  $\mathbf{R}^3$  containing the origin, then  $\mathbf{R}^3/U$  is the set of all planes in  $\mathbf{R}^3$  parallel to  $U$ .

Our next goal is to make  $V/U$  into a vector space. To do this, we will need the following result.

### 3.85 Two affine subsets parallel to $U$ are equal or disjoint

Suppose  $U$  is a subspace of  $V$  and  $v, w \in V$ . Then the following are equivalent:

- $v - w \in U$ ;
- $v + U = w + U$ ;
- $(v + U) \cap (w + U) \neq \emptyset$ .

**Proof** First suppose (a) holds, so  $v - w \in U$ . If  $u \in U$ , then

$$v + u = w + ((v - w) + u) \in w + U.$$

Thus  $v + U \subset w + U$ . Similarly,  $w + U \subset v + U$ . Thus  $v + U = w + U$ , completing the proof that (a) implies (b).

Obviously (b) implies (c).

Now suppose (c) holds, so  $(v + U) \cap (w + U) \neq \emptyset$ . Thus there exist  $u_1, u_2 \in U$  such that

$$v + u_1 = w + u_2.$$

Thus  $v - w = u_2 - u_1$ . Hence  $v - w \in U$ , showing that (c) implies (a) and completing the proof. ■

Now we can define addition and scalar multiplication on  $V/U$ .

### 3.86 Definition *addition and scalar multiplication on $V/U$*

Suppose  $U$  is a subspace of  $V$ . Then **addition** and **scalar multiplication** are defined on  $V/U$  by

$$(v + U) + (w + U) = (v + w) + U$$

$$\lambda(v + U) = (\lambda v) + U$$

for  $v, w \in V$  and  $\lambda \in \mathbf{F}$ .

As part of the proof of the next result, we will show that the definitions above make sense.

### 3.87 Quotient space is a vector space

Suppose  $U$  is a subspace of  $V$ . Then  $V/U$ , with the operations of addition and scalar multiplication as defined above, is a vector space.

**Proof** The potential problem with the definitions above of addition and scalar multiplication on  $V/U$  is that the representation of an affine subset parallel to  $U$  is not unique. Specifically, suppose  $v, w \in V$ . Suppose also that  $\hat{v}, \hat{w} \in V$  are such that  $v + U = \hat{v} + U$  and  $w + U = \hat{w} + U$ . To show that the definition of addition on  $V/U$  given above makes sense, we must show that  $(v + w) + U = (\hat{v} + \hat{w}) + U$ .

By 3.85, we have

$$v - \hat{v} \in U \quad \text{and} \quad w - \hat{w} \in U.$$

Because  $U$  is a subspace of  $V$  and thus is closed under addition, this implies that  $(v - \hat{v}) + (w - \hat{w}) \in U$ . Thus  $(v + w) - (\hat{v} + \hat{w}) \in U$ . Using 3.85 again, we see that

$$(v + w) + U = (\hat{v} + \hat{w}) + U,$$

as desired. Thus the definition of addition on  $V/U$  makes sense.

Similarly, suppose  $\lambda \in \mathbf{F}$ . Because  $U$  is a subspace of  $V$  and thus is closed under scalar multiplication, we have  $\lambda(v - \hat{v}) \in U$ . Thus  $\lambda v - \lambda \hat{v} \in U$ . Hence 3.85 implies that  $(\lambda v) + U = (\lambda \hat{v}) + U$ . Thus the definition of scalar multiplication on  $V/U$  makes sense.

Now that addition and scalar multiplication have been defined on  $V/U$ , the verification that these operations make  $V/U$  into a vector space is straightforward and is left to the reader. Note that the additive identity of  $V/U$  is  $0 + U$  (which equals  $U$ ) and that the additive inverse of  $v + U$  is  $(-v) + U$ . ■

The next concept will give us an easy way to compute the dimension of  $V/U$ .

### 3.88 Definition *quotient map, $\pi$*

Suppose  $U$  is a subspace of  $V$ . The *quotient map*  $\pi$  is the linear map  $\pi: V \rightarrow V/U$  defined by

$$\pi(v) = v + U$$

for  $v \in V$ .

The reader should verify that  $\pi$  is indeed a linear map. Although  $\pi$  depends on  $U$  as well as  $V$ , these spaces are left out of the notation because they should be clear from the context.

### 3.89 Dimension of a quotient space

Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Then

$$\dim V/U = \dim V - \dim U.$$

**Proof** Let  $\pi$  be the quotient map from  $V$  to  $V/U$ . From 3.85, we see that  $\text{null } \pi = U$ . Clearly  $\text{range } \pi = V/U$ . The Fundamental Theorem of Linear Maps (3.22) thus tells us that

$$\dim V = \dim U + \dim V/U,$$

which gives the desired result. ■

Each linear map  $T$  on  $V$  induces a linear map  $\tilde{T}$  on  $V/(\text{null } T)$ , which we now define.

### 3.90 Definition $\tilde{T}$

Suppose  $T \in \mathcal{L}(V, W)$ . Define  $\tilde{T}: V/(\text{null } T) \rightarrow W$  by

$$\tilde{T}(v + \text{null } T) = Tv.$$

To show that the definition of  $\tilde{T}$  makes sense, suppose  $u, v \in V$  are such that  $u + \text{null } T = v + \text{null } T$ . By 3.85, we have  $u - v \in \text{null } T$ . Thus  $T(u - v) = 0$ . Hence  $Tu = Tv$ . Thus the definition of  $\tilde{T}$  indeed makes sense.

### 3.91 Null space and range of $\tilde{T}$

Suppose  $T \in \mathcal{L}(V, W)$ . Then

- (a)  $\tilde{T}$  is a linear map from  $V/\text{null } T$  to  $W$ ;
- (b)  $\tilde{T}$  is injective;
- (c)  $\text{range } \tilde{T} = \text{range } T$ ;
- (d)  $V/\text{null } T$  is isomorphic to  $\text{range } T$ .

#### Proof

- (a) The routine verification that  $\tilde{T}$  is linear is left to the reader.
- (b) Suppose  $v \in V$  and  $\tilde{T}(v + \text{null } T) = 0$ . Then  $Tv = 0$ . Thus  $v \in \text{null } T$ . Hence 3.85 implies that  $v + \text{null } T = 0 + \text{null } T$ . This implies that  $\text{null } \tilde{T} = 0$ , and hence  $\tilde{T}$  is injective, as desired.
- (c) The definition of  $\tilde{T}$  shows that  $\text{range } \tilde{T} = \text{range } T$ .
- (d) Parts (b) and (c) imply that if we think of  $\tilde{T}$  as mapping into  $\text{range } T$ , then  $\tilde{T}$  is an isomorphism from  $V/\text{null } T$  onto  $\text{range } T$ . ■

## EXERCISES 3.E

---

- 1** Suppose  $T$  is a function from  $V$  to  $W$ . The **graph** of  $T$  is the subset of  $V \times W$  defined by

$$\text{graph of } T = \{(v, Tv) \in V \times W : v \in V\}.$$

Prove that  $T$  is a linear map if and only if the graph of  $T$  is a subspace of  $V \times W$ .

[Formally, a function  $T$  from  $V$  to  $W$  is a subset  $T$  of  $V \times W$  such that for each  $v \in V$ , there exists exactly one element  $(v, w) \in T$ . In other words, formally a function is what is called above its graph. We do not usually think of functions in this formal manner. However, if we do become formal, then the exercise above could be rephrased as follows: Prove that a function  $T$  from  $V$  to  $W$  is a linear map if and only if  $T$  is a subspace of  $V \times W$ .]

- 2 Suppose  $V_1, \dots, V_m$  are vector spaces such that  $V_1 \times \cdots \times V_m$  is finite-dimensional. Prove that  $V_j$  is finite-dimensional for each  $j = 1, \dots, m$ .

- 3 Give an example of a vector space  $V$  and subspaces  $U_1, U_2$  of  $V$  such that  $U_1 \times U_2$  is isomorphic to  $U_1 + U_2$  but  $U_1 + U_2$  is not a direct sum.

*Hint:* The vector space  $V$  must be infinite-dimensional.

- 4 Suppose  $V_1, \dots, V_m$  are vector spaces. Prove that  $\mathcal{L}(V_1 \times \cdots \times V_m, W)$  and  $\mathcal{L}(V_1, W) \times \cdots \times \mathcal{L}(V_m, W)$  are isomorphic vector spaces.

- 5 Suppose  $W_1, \dots, W_m$  are vector spaces. Prove that  $\mathcal{L}(V, W_1 \times \cdots \times W_m)$  and  $\mathcal{L}(V, W_1) \times \cdots \times \mathcal{L}(V, W_m)$  are isomorphic vector spaces.

- 6 For  $n$  a positive integer, define  $V^n$  by

$$V^n = \underbrace{V \times \cdots \times V}_{n \text{ times}}.$$

Prove that  $V^n$  and  $\mathcal{L}(\mathbf{F}^n, V)$  are isomorphic vector spaces.

- 7 Suppose  $v, x$  are vectors in  $V$  and  $U, W$  are subspaces of  $V$  such that  $v + U = x + W$ . Prove that  $U = W$ .

- 8 Prove that a nonempty subset  $A$  of  $V$  is an affine subset of  $V$  if and only if  $\lambda v + (1 - \lambda)w \in A$  for all  $v, w \in A$  and all  $\lambda \in \mathbf{F}$ .

- 9 Suppose  $A_1$  and  $A_2$  are affine subsets of  $V$ . Prove that the intersection  $A_1 \cap A_2$  is either an affine subset of  $V$  or the empty set.

- 10 Prove that the intersection of every collection of affine subsets of  $V$  is either an affine subset of  $V$  or the empty set.

- 11 Suppose  $v_1, \dots, v_m \in V$ . Let

$$A = \{\lambda_1 v_1 + \cdots + \lambda_m v_m : \lambda_1, \dots, \lambda_m \in \mathbf{F} \text{ and } \lambda_1 + \cdots + \lambda_m = 1\}.$$

- (a) Prove that  $A$  is an affine subset of  $V$ .

- (b) Prove that every affine subset of  $V$  that contains  $v_1, \dots, v_m$  also contains  $A$ .

- (c) Prove that  $A = v + U$  for some  $v \in V$  and some subspace  $U$  of  $V$  with  $\dim U \leq m - 1$ .

- 12 Suppose  $U$  is a subspace of  $V$  such that  $V/U$  is finite-dimensional. Prove that  $V$  is isomorphic to  $U \times (V/U)$ .

- 13** Suppose  $U$  is a subspace of  $V$  and  $v_1 + U, \dots, v_m + U$  is a basis of  $V/U$  and  $u_1, \dots, u_n$  is a basis of  $U$ . Prove that  $v_1, \dots, v_m, u_1, \dots, u_n$  is a basis of  $V$ .
- 14** Suppose  $U = \{(x_1, x_2, \dots) \in \mathbf{F}^\infty : x_j \neq 0 \text{ for only finitely many } j\}$ .
- Show that  $U$  is a subspace of  $\mathbf{F}^\infty$ .
  - Prove that  $\mathbf{F}^\infty/U$  is infinite-dimensional.
- 15** Suppose  $\varphi \in \mathcal{L}(V, \mathbf{F})$  and  $\varphi \neq 0$ . Prove that  $\dim V/(\text{null } \varphi) = 1$ .
- 16** Suppose  $U$  is a subspace of  $V$  such that  $\dim V/U = 1$ . Prove that there exists  $\varphi \in \mathcal{L}(V, \mathbf{F})$  such that  $\text{null } \varphi = U$ .
- 17** Suppose  $U$  is a subspace of  $V$  such that  $V/U$  is finite-dimensional. Prove that there exists a subspace  $W$  of  $V$  such that  $\dim W = \dim V/U$  and  $V = U \oplus W$ .
- 18** Suppose  $T \in \mathcal{L}(V, W)$  and  $U$  is a subspace of  $V$ . Let  $\pi$  denote the quotient map from  $V$  onto  $V/U$ . Prove that there exists  $S \in \mathcal{L}(V/U, W)$  such that  $T = S \circ \pi$  if and only if  $U \subset \text{null } T$ .
- 19** Find a correct statement analogous to 3.78 that is applicable to finite sets, with unions analogous to sums of subspaces and disjoint unions analogous to direct sums.
- 20** Suppose  $U$  is a subspace of  $V$ . Define  $\Gamma: \mathcal{L}(V/U, W) \rightarrow \mathcal{L}(V, W)$  by

$$\Gamma(S) = S \circ \pi.$$

- Show that  $\Gamma$  is a linear map.
- Show that  $\Gamma$  is injective.
- Show that  $\text{range } \Gamma = \{T \in \mathcal{L}(V, W) : Tu = 0 \text{ for every } u \in U\}$ .

## 3.F Duality

### The Dual Space and the Dual Map

Linear maps into the scalar field  $\mathbf{F}$  play a special role in linear algebra, and thus they get a special name:

#### 3.92 Definition *linear functional*

A *linear functional* on  $V$  is a linear map from  $V$  to  $\mathbf{F}$ . In other words, a linear functional is an element of  $\mathcal{L}(V, \mathbf{F})$ .

#### 3.93 Example *linear functionals*

- Define  $\varphi: \mathbf{R}^3 \rightarrow \mathbf{R}$  by  $\varphi(x, y, z) = 4x - 5y + 2z$ . Then  $\varphi$  is a linear functional on  $\mathbf{R}^3$ .
- Fix  $(c_1, \dots, c_n) \in \mathbf{F}^n$ . Define  $\varphi: \mathbf{F}^n \rightarrow \mathbf{F}$  by

$$\varphi(x_1, \dots, x_n) = c_1 x_1 + \dots + c_n x_n.$$

Then  $\varphi$  is a linear functional on  $\mathbf{F}^n$ .

- Define  $\varphi: \mathcal{P}(\mathbf{R}) \rightarrow \mathbf{R}$  by  $\varphi(p) = 3p''(5) + 7p(4)$ . Then  $\varphi$  is a linear functional on  $\mathcal{P}(\mathbf{R})$ .
- Define  $\varphi: \mathcal{P}(\mathbf{R}) \rightarrow \mathbf{R}$  by  $\varphi(p) = \int_0^1 p(x) dx$ . Then  $\varphi$  is a linear functional on  $\mathcal{P}(\mathbf{R})$ .

The vector space  $\mathcal{L}(V, \mathbf{F})$  also gets a special name and special notation:

#### 3.94 Definition *dual space, $V'$*

The *dual space* of  $V$ , denoted  $V'$ , is the vector space of all linear functionals on  $V$ . In other words,  $V' = \mathcal{L}(V, \mathbf{F})$ .

#### 3.95 $\dim V' = \dim V$

Suppose  $V$  is finite-dimensional. Then  $V'$  is also finite-dimensional and  $\dim V' = \dim V$ .

**Proof** This result follows from 3.61. ■

In the following definition, 3.5 implies that each  $\varphi_j$  is well defined.

### 3.96 Definition *dual basis*

If  $v_1, \dots, v_n$  is a basis of  $V$ , then the ***dual basis*** of  $v_1, \dots, v_n$  is the list  $\varphi_1, \dots, \varphi_n$  of elements of  $V'$ , where each  $\varphi_j$  is the linear functional on  $V$  such that

$$\varphi_j(v_k) = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } k \neq j. \end{cases}$$

**3.97 Example** What is the dual basis of the standard basis  $e_1, \dots, e_n$  of  $\mathbf{F}^n$ ?

**Solution** For  $1 \leq j \leq n$ , define  $\varphi_j$  to be the linear functional on  $\mathbf{F}^n$  that selects the  $j^{\text{th}}$  coordinate of a vector in  $\mathbf{F}^n$ . In other words,

$$\varphi_j(x_1, \dots, x_n) = x_j$$

for  $(x_1, \dots, x_n) \in \mathbf{F}^n$ . Clearly

$$\varphi_j(e_k) = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } k \neq j. \end{cases}$$

Thus  $\varphi_1, \dots, \varphi_n$  is the dual basis of the standard basis  $e_1, \dots, e_n$  of  $\mathbf{F}^n$ .

The next result shows that the dual basis is indeed a basis. Thus the terminology “dual basis” is justified.

### 3.98 Dual basis is a basis of the dual space

Suppose  $V$  is finite-dimensional. Then the dual basis of a basis of  $V$  is a basis of  $V'$ .

**Proof** Suppose  $v_1, \dots, v_n$  is a basis of  $V$ . Let  $\varphi_1, \dots, \varphi_n$  denote the dual basis.

To show that  $\varphi_1, \dots, \varphi_n$  is a linearly independent list of elements of  $V'$ , suppose  $a_1, \dots, a_n \in F$  are such that

$$a_1\varphi_1 + \cdots + a_n\varphi_n = 0.$$

Now  $(a_1\varphi_1 + \cdots + a_n\varphi_n)(v_j) = a_j$  for  $j = 1, \dots, n$ . The equation above thus shows that  $a_1 = \cdots = a_n = 0$ . Hence  $\varphi_1, \dots, \varphi_n$  is linearly independent.

Now 2.39 and 3.95 imply that  $\varphi_1, \dots, \varphi_n$  is a basis of  $V'$ . ■

In the definition below, note that if  $T$  is a linear map from  $V$  to  $W$  then  $T'$  is a linear map from  $W'$  to  $V'$ .

### 3.99 Definition *dual map*, $T'$

If  $T \in \mathcal{L}(V, W)$ , then the **dual map** of  $T$  is the linear map  $T' \in \mathcal{L}(W', V')$  defined by  $T'(\varphi) = \varphi \circ T$  for  $\varphi \in W'$ .

If  $T \in \mathcal{L}(V, W)$  and  $\varphi \in W'$ , then  $T'(\varphi)$  is defined above to be the composition of the linear maps  $\varphi$  and  $T$ . Thus  $T'(\varphi)$  is indeed a linear map from  $V$  to  $\mathbf{F}$ ; in other words,  $T'(\varphi) \in V'$ .

The verification that  $T'$  is a linear map from  $W'$  to  $V'$  is easy:

- If  $\varphi, \psi \in W'$ , then

$$T'(\varphi + \psi) = (\varphi + \psi) \circ T = \varphi \circ T + \psi \circ T = T'(\varphi) + T'(\psi).$$

- If  $\lambda \in \mathbf{F}$  and  $\varphi \in W'$ , then

$$T'(\lambda\varphi) = (\lambda\varphi) \circ T = \lambda(\varphi \circ T) = \lambda T'(\varphi).$$

In the next example, the prime notation is used with two unrelated meanings:  $D'$  denotes the dual of a linear map  $D$ , and  $p'$  denotes the derivative of a polynomial  $p$ .

---

### 3.100 Example Define $D : \mathcal{P}(\mathbf{R}) \rightarrow \mathcal{P}(\mathbf{R})$ by $Dp = p'$ .

- Suppose  $\varphi$  is the linear functional on  $\mathcal{P}(\mathbf{R})$  defined by  $\varphi(p) = p(3)$ . Then  $D'(\varphi)$  is the linear functional on  $\mathcal{P}(\mathbf{R})$  given by

$$(D'(\varphi))(p) = (\varphi \circ D)(p) = \varphi(Dp) = \varphi(p') = p'(3).$$

In other words,  $D'(\varphi)$  is the linear functional on  $\mathcal{P}(\mathbf{R})$  that takes  $p$  to  $p'(3)$ .

- Suppose  $\varphi$  is the linear functional on  $\mathcal{P}(\mathbf{R})$  defined by  $\varphi(p) = \int_0^1 p$ . Then  $D'(\varphi)$  is the linear functional on  $\mathcal{P}(\mathbf{R})$  given by

$$(D'(\varphi))(p) = (\varphi \circ D)(p) = \varphi(Dp) = \varphi(p') = \int_0^1 p' = p(1) - p(0).$$

In other words,  $D'(\varphi)$  is the linear functional on  $\mathcal{P}(\mathbf{R})$  that takes  $p$  to  $p(1) - p(0)$ .

---

The first two bullet points in the result below imply that the function that takes  $T$  to  $T'$  is a linear map from  $\mathcal{L}(V, W)$  to  $\mathcal{L}(W', V')$ .

In the third bullet point below, note the reversal of order from  $ST$  on the left to  $T'S'$  on the right (here we assume that  $U$  is a vector space over  $\mathbf{F}$ ).

### 3.101 Algebraic properties of dual maps

- $(S + T)' = S' + T'$  for all  $S, T \in \mathcal{L}(V, W)$ .
- $(\lambda T)' = \lambda T'$  for all  $\lambda \in \mathbf{F}$  and all  $T \in \mathcal{L}(V, W)$ .
- $(ST)' = T'S'$  for all  $T \in \mathcal{L}(U, V)$  and all  $S \in \mathcal{L}(V, W)$ .

**Proof** The proofs of the first two bullet points above are left to the reader.

To prove the third bullet point, suppose  $\varphi \in W'$ . Then

$$(ST)'(\varphi) = \varphi \circ (ST) = (\varphi \circ S) \circ T = T'(\varphi \circ S) = T'(S'(\varphi)) = (T'S')(\varphi),$$

Some books use the notation  $V^*$  and  $T^*$  for duality instead of  $V'$  and  $T'$ . However, here we reserve the notation  $T^*$  for the adjoint, which will be introduced when we study linear maps on inner product spaces in Chapter 7.

where the first, third, and fourth equalities above hold because of the definition of the dual map, the second equality holds because composition of functions is associative, and the last equality follows from the definition of composition.

The equality of the first and last terms above for all  $\varphi \in W'$  means that  $(ST)' = T'S'$ . ■

## The Null Space and Range of the Dual of a Linear Map

Our goal in this subsection is to describe null  $T'$  and range  $T'$  in terms of range  $T$  and null  $T$ . To do this, we will need the following definition.

### 3.102 Definition *annihilator*, $U^0$

For  $U \subset V$ , the **annihilator** of  $U$ , denoted  $U^0$ , is defined by

$$U^0 = \{\varphi \in V' : \varphi(u) = 0 \text{ for all } u \in U\}.$$

**3.103 Example** Suppose  $U$  is the subspace of  $\mathcal{P}(\mathbf{R})$  consisting of all polynomial multiples of  $x^2$ . If  $\varphi$  is the linear functional on  $\mathcal{P}(\mathbf{R})$  defined by  $\varphi(p) = p'(0)$ , then  $\varphi \in U^0$ .

For  $U \subset V$ , the annihilator  $U^0$  is a subset of the dual space  $V'$ . Thus  $U^0$  depends on the vector space containing  $U$ , so a notation such as  $U_V^0$  would be more precise. However, the containing vector space will always be clear from the context, so we will use the simpler notation  $U^0$ .

**3.104 Example** Let  $e_1, e_2, e_3, e_4, e_5$  denote the standard basis of  $\mathbf{R}^5$ , and let  $\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5$  denote the dual basis of  $(\mathbf{R}^5)'$ . Suppose

$$U = \text{span}(e_1, e_2) = \{(x_1, x_2, 0, 0, 0) \in \mathbf{R}^5 : x_1, x_2 \in \mathbf{R}\}.$$

Show that  $U^0 = \text{span}(\varphi_3, \varphi_4, \varphi_5)$ .

**Solution** Recall (see 3.97) that  $\varphi_j$  is the linear functional on  $\mathbf{R}^5$  that selects that  $j^{\text{th}}$  coordinate:  $\varphi_j(x_1, x_2, x_3, x_4, x_5) = x_j$ .

First suppose  $\varphi \in \text{span}(\varphi_3, \varphi_4, \varphi_5)$ . Then there exist  $c_3, c_4, c_5 \in \mathbf{R}$  such that  $\varphi = c_3\varphi_3 + c_4\varphi_4 + c_5\varphi_5$ . If  $(x_1, x_2, 0, 0, 0) \in U$ , then

$$\varphi(x_1, x_2, 0, 0, 0) = (c_3\varphi_3 + c_4\varphi_4 + c_5\varphi_5)(x_1, x_2, 0, 0, 0) = 0.$$

Thus  $\varphi \in U^0$ . In other words, we have shown that  $\text{span}(\varphi_3, \varphi_4, \varphi_5) \subset U^0$ .

To show the inclusion in the other direction, suppose  $\varphi \in U^0$ . Because the dual basis is a basis of  $(\mathbf{R}^5)'$ , there exist  $c_1, c_2, c_3, c_4, c_5 \in \mathbf{R}$  such that  $\varphi = c_1\varphi_1 + c_2\varphi_2 + c_3\varphi_3 + c_4\varphi_4 + c_5\varphi_5$ . Because  $e_1 \in U$  and  $\varphi \in U^0$ , we have

$$0 = \varphi(e_1) = (c_1\varphi_1 + c_2\varphi_2 + c_3\varphi_3 + c_4\varphi_4 + c_5\varphi_5)(e_1) = c_1.$$

Similarly,  $e_2 \in U$  and thus  $c_2 = 0$ . Hence  $\varphi = c_3\varphi_3 + c_4\varphi_4 + c_5\varphi_5$ . Thus  $\varphi \in \text{span}(\varphi_3, \varphi_4, \varphi_5)$ , which shows that  $U^0 \subset \text{span}(\varphi_3, \varphi_4, \varphi_5)$ .

### 3.105 The annihilator is a subspace

Suppose  $U \subset V$ . Then  $U^0$  is a subspace of  $V'$ .

**Proof** Clearly  $0 \in U^0$  (here  $0$  is the zero linear functional on  $V$ ), because the zero linear functional applied to every vector in  $U$  is  $0$ .

Suppose  $\varphi, \psi \in U^0$ . Thus  $\varphi, \psi \in V'$  and  $\varphi(u) = \psi(u) = 0$  for every  $u \in U$ . If  $u \in U$ , then  $(\varphi + \psi)(u) = \varphi(u) + \psi(u) = 0 + 0 = 0$ . Thus  $\varphi + \psi \in U^0$ .

Similarly,  $U^0$  is closed under scalar multiplication. Thus 1.34 implies that  $U^0$  is a subspace of  $V'$ . ■

The next result shows that  $\dim U^0$  is the difference of  $\dim V$  and  $\dim U$ . For example, this shows that if  $U$  is a 2-dimensional subspace of  $\mathbf{R}^5$ , then  $U^0$  is a 3-dimensional subspace of  $(\mathbf{R}^5)'$ , as in Example 3.104.

The next result can be proved following the pattern of Example 3.104: choose a basis  $u_1, \dots, u_m$  of  $U$ , extend to a basis  $u_1, \dots, u_m, \dots, u_n$  of  $V$ , let  $\varphi_1, \dots, \varphi_m, \dots, \varphi_n$  be the dual basis of  $V'$ , and then show  $\varphi_{m+1}, \dots, \varphi_n$  is a basis of  $U^0$ , which implies the desired result.

You should construct the proof outlined in the paragraph above, even though a slicker proof is presented here.

### 3.106 Dimension of the annihilator

Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Then

$$\dim U + \dim U^0 = \dim V.$$

**Proof** Let  $i \in \mathcal{L}(U, V)$  be the inclusion map defined by  $i(u) = u$  for  $u \in U$ . Thus  $i'$  is a linear map from  $V'$  to  $U'$ . The Fundamental Theorem of Linear Maps (3.22) applied to  $i'$  shows that

$$\dim \text{range } i' + \dim \text{null } i' = \dim V'.$$

However,  $\text{null } i' = U^0$  (as can be seen by thinking about the definitions) and  $\dim V' = \dim V$  (by 3.95), so we can rewrite the equation above as

$$\dim \text{range } i' + \dim U^0 = \dim V.$$

If  $\varphi \in U'$ , then  $\varphi$  can be extended to a linear functional  $\psi$  on  $V$  (see, for example, Exercise 11 in Section 3.A). The definition of  $i'$  shows that  $i'(\psi) = \varphi$ . Thus  $\varphi \in \text{range } i'$ , which implies that  $\text{range } i' = U'$ . Hence  $\dim \text{range } i' = \dim U' = \dim U$ , and the displayed equation above becomes the desired result. ■

The proof of part (a) of the result below does not use the hypothesis that  $V$  and  $W$  are finite-dimensional.

### 3.107 The null space of $T'$

Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Then

- (a)  $\text{null } T' = (\text{range } T)^0$ ;
- (b)  $\dim \text{null } T' = \dim \text{null } T + \dim W - \dim V$ .

**Proof**

- (a) First suppose  $\varphi \in \text{null } T'$ . Thus  $0 = T'(\varphi) = \varphi \circ T$ . Hence

$$0 = (\varphi \circ T)(v) = \varphi(Tv) \quad \text{for every } v \in V.$$

Thus  $\varphi \in (\text{range } T)^0$ . This implies that  $\text{null } T' \subset (\text{range } T)^0$ .

To prove the inclusion in the opposite direction, now suppose that  $\varphi \in (\text{range } T)^0$ . Thus  $\varphi(Tv) = 0$  for every vector  $v \in V$ . Hence  $0 = \varphi \circ T = T'(\varphi)$ . In other words,  $\varphi \in \text{null } T'$ , which shows that  $(\text{range } T)^0 \subset \text{null } T'$ , completing the proof of (a).

- (b) We have

$$\begin{aligned} \dim \text{null } T' &= \dim(\text{range } T)^0 \\ &= \dim W - \dim \text{range } T \\ &= \dim W - (\dim V - \dim \text{null } T) \\ &= \dim \text{null } T + \dim W - \dim V, \end{aligned}$$

where the first equality comes from (a), the second equality comes from 3.106, and the third equality comes from the Fundamental Theorem of Linear Maps (3.22). ■

The next result can be useful because sometimes it is easier to verify that  $T'$  is injective than to show directly that  $T$  is surjective.

### 3.108 $T$ surjective is equivalent to $T'$ injective

Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Then  $T$  is surjective if and only if  $T'$  is injective.

**Proof** The map  $T \in \mathcal{L}(V, W)$  is surjective if and only if  $\text{range } T = W$ , which happens if and only if  $(\text{range } T)^0 = \{0\}$ , which happens if and only if  $\text{null } T' = \{0\}$  [by 3.107(a)], which happens if and only if  $T'$  is injective. ■

### 3.109 The range of $T'$

Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Then

- (a)  $\dim \text{range } T' = \dim \text{range } T$ ;
- (b)  $\text{range } T' = (\text{null } T)^0$ .

**Proof**

(a) We have

$$\begin{aligned}\dim \text{range } T' &= \dim W' - \dim \text{null } T' \\ &= \dim W - \dim(\text{range } T)^0 \\ &= \dim \text{range } T,\end{aligned}$$

where the first equality comes from the Fundamental Theorem of Linear Maps (3.22), the second equality comes from 3.95 and 3.107(a), and the third equality comes from 3.106.

(b) First suppose  $\varphi \in \text{range } T'$ . Thus there exists  $\psi \in W'$  such that  $\varphi = T'(\psi)$ . If  $v \in \text{null } T$ , then

$$\varphi(v) = (T'(\psi))v = (\psi \circ T)(v) = \psi(Tv) = \psi(0) = 0.$$

Hence  $\varphi \in (\text{null } T)^0$ . This implies that  $\text{range } T' \subset (\text{null } T)^0$ .

We will complete the proof by showing that  $\text{range } T'$  and  $(\text{null } T)^0$  have the same dimension. To do this, note that

$$\begin{aligned}\dim \text{range } T' &= \dim \text{range } T \\ &= \dim V - \dim \text{null } T \\ &= \dim(\text{null } T)^0,\end{aligned}$$

where the first equality comes from (a), the second equality comes from the Fundamental Theorem of Linear Maps (3.22), and the third equality comes from 3.106. ■

The next result should be compared to 3.108.

### 3.110 $T$ injective is equivalent to $T'$ surjective

Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Then  $T$  is injective if and only if  $T'$  is surjective.

**Proof** The map  $T \in \mathcal{L}(V, W)$  is injective if and only if  $\text{null } T = \{0\}$ , which happens if and only if  $(\text{null } T)^0 = V'$ , which happens if and only if  $\text{range } T' = V'$  [by 3.109(b)], which happens if and only if  $T'$  is surjective. ■

## The Matrix of the Dual of a Linear Map

We now define the transpose of a matrix.

### 3.111 Definition transpose, $A^t$

The **transpose** of a matrix  $A$ , denoted  $A^t$ , is the matrix obtained from  $A$  by interchanging the rows and columns. More specifically, if  $A$  is an  $m$ -by- $n$  matrix, then  $A^t$  is the  $n$ -by- $m$  matrix whose entries are given by the equation

$$(A^t)_{k,j} = A_{j,k}.$$

---

### 3.112 Example

If  $A = \begin{pmatrix} 5 & -7 \\ 3 & 8 \\ -4 & 2 \end{pmatrix}$ , then  $A^t = \begin{pmatrix} 5 & 3 & -4 \\ -7 & 8 & 2 \end{pmatrix}$ .

Note that here  $A$  is a 3-by-2 matrix and  $A^t$  is a 2-by-3 matrix.

---

The transpose has nice algebraic properties:  $(A + C)^t = A^t + C^t$  and  $(\lambda A)^t = \lambda A^t$  for all  $m$ -by- $n$  matrices  $A, C$  and all  $\lambda \in \mathbf{F}$  (see Exercise 33).

The next result shows that the transpose of the product of two matrices is the product of the transposes in the opposite order.

### 3.113 The transpose of the product of matrices

If  $A$  is an  $m$ -by- $n$  matrix and  $C$  is an  $n$ -by- $p$  matrix, then

$$(AC)^t = C^t A^t.$$

**Proof** Suppose  $1 \leq k \leq p$  and  $1 \leq j \leq m$ . Then

$$\begin{aligned} ((AC)^t)_{k,j} &= (AC)_{j,k} \\ &= \sum_{r=1}^n A_{j,r} C_{r,k} \\ &= \sum_{r=1}^n (C^t)_{k,r} (A^t)_{r,j} \\ &= (C^t A^t)_{k,j}. \end{aligned}$$

Thus  $(AC)^t = C^t A^t$ , as desired. ■

The setting for the next result is the assumption that we have a basis  $v_1, \dots, v_n$  of  $V$ , along with its dual basis  $\varphi_1, \dots, \varphi_n$  of  $V'$ . We also have a basis  $w_1, \dots, w_m$  of  $W$ , along with its dual basis  $\psi_1, \dots, \psi_m$  of  $W'$ . Thus  $\mathcal{M}(T)$  is computed with respect to the bases just mentioned of  $V$  and  $W$ , and  $\mathcal{M}(T')$  is computed with respect to the dual bases just mentioned of  $W'$  and  $V'$ .

### 3.114 The matrix of $T'$ is the transpose of the matrix of $T$

Suppose  $T \in \mathcal{L}(V, W)$ . Then  $\mathcal{M}(T') = (\mathcal{M}(T))^t$ .

**Proof** Let  $A = \mathcal{M}(T)$  and  $C = \mathcal{M}(T')$ . Suppose  $1 \leq j \leq m$  and  $1 \leq k \leq n$ .

From the definition of  $\mathcal{M}(T')$  we have

$$T'(\psi_j) = \sum_{r=1}^n C_{r,j} \varphi_r.$$

The left side of the equation above equals  $\psi_j \circ T$ . Thus applying both sides of the equation above to  $v_k$  gives

$$\begin{aligned} (\psi_j \circ T)(v_k) &= \sum_{r=1}^n C_{r,j} \varphi_r(v_k) \\ &= C_{k,j}. \end{aligned}$$

We also have

$$\begin{aligned} (\psi_j \circ T)(v_k) &= \psi_j(Tv_k) \\ &= \psi_j\left(\sum_{r=1}^m A_{r,k} w_r\right) \\ &= \sum_{r=1}^m A_{r,k} \psi_j(w_r) \\ &= A_{j,k}. \end{aligned}$$

Comparing the last line of the last two sets of equations, we have  $C_{k,j} = A_{j,k}$ . Thus  $C = A^t$ . In other words,  $\mathcal{M}(T') = (\mathcal{M}(T))^t$ , as desired. ■

## The Rank of a Matrix

We begin by defining two nonnegative integers that are associated with each matrix.

### 3.115 Definition *row rank, column rank*

Suppose  $A$  is an  $m$ -by- $n$  matrix with entries in  $\mathbf{F}$ .

- The ***row rank*** of  $A$  is the dimension of the span of the rows of  $A$  in  $\mathbf{F}^{1,n}$ .
- The ***column rank*** of  $A$  is the dimension of the span of the columns of  $A$  in  $\mathbf{F}^{m,1}$ .

**3.116 Example** Suppose  $A = \begin{pmatrix} 4 & 7 & 1 & 8 \\ 3 & 5 & 2 & 9 \end{pmatrix}$ . Find the row rank of  $A$  and the column rank of  $A$ .

**Solution** The row rank of  $A$  is the dimension of

$$\text{span}\left(\left(\begin{array}{cccc} 4 & 7 & 1 & 8 \end{array}\right), \left(\begin{array}{cccc} 3 & 5 & 2 & 9 \end{array}\right)\right)$$

in  $\mathbf{F}^{1,4}$ . Neither of the two vectors listed above in  $\mathbf{F}^{1,4}$  is a scalar multiple of the other. Thus the span of this list of length 2 has dimension 2. In other words, the row rank of  $A$  is 2.

The column rank of  $A$  is the dimension of

$$\text{span}\left(\left(\begin{array}{c} 4 \\ 3 \end{array}\right), \left(\begin{array}{c} 7 \\ 5 \end{array}\right), \left(\begin{array}{c} 1 \\ 2 \end{array}\right), \left(\begin{array}{c} 8 \\ 9 \end{array}\right)\right)$$

in  $\mathbf{F}^{2,1}$ . Neither of the first two vectors listed above in  $\mathbf{F}^{2,1}$  is a scalar multiple of the other. Thus the span of this list of length 4 has dimension at least 2. The span of this list of vectors in  $\mathbf{F}^{2,1}$  cannot have dimension larger than 2 because  $\dim \mathbf{F}^{2,1} = 2$ . Thus the span of this list has dimension 2. In other words, the column rank of  $A$  is 2.

Notice that no bases are in sight in the statement of the next result. Although  $\mathcal{M}(T)$  in the next result depends on a choice of bases of  $V$  and  $W$ , the next result shows that the column rank of  $\mathcal{M}(T)$  is the same for all such choices (because range  $T$  does not depend on a choice of basis).

**3.117 Dimension of range  $T$  equals column rank of  $\mathcal{M}(T)$** 

Suppose  $V$  and  $W$  are finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Then  $\dim \text{range } T$  equals the column rank of  $\mathcal{M}(T)$ .

**Proof** Suppose  $v_1, \dots, v_n$  is a basis of  $V$  and  $w_1, \dots, w_m$  is a basis of  $W$ . The function that takes  $w \in \text{span}(Tv_1, \dots, Tv_n)$  to  $\mathcal{M}(w)$  is easily seen to be an isomorphism from  $\text{span}(Tv_1, \dots, Tv_n)$  onto  $\text{span}(\mathcal{M}(Tv_1), \dots, \mathcal{M}(Tv_n))$ . Thus  $\dim \text{span}(Tv_1, \dots, Tv_n) = \dim \text{span}(\mathcal{M}(Tv_1), \dots, \mathcal{M}(Tv_n))$ , where the last dimension equals the column rank of  $\mathcal{M}(T)$ .

It is easy to see that  $\text{range } T = \text{span}(Tv_1, \dots, Tv_n)$ . Thus we have  $\dim \text{range } T = \dim \text{span}(Tv_1, \dots, Tv_n) =$  the column rank of  $\mathcal{M}(T)$ , as desired. ■

In Example 3.116, the row rank and column rank turned out to equal each other. The next result shows that this always happens.

**3.118 Row rank equals column rank**

Suppose  $A \in \mathbf{F}^{m,n}$ . Then the row rank of  $A$  equals the column rank of  $A$ .

**Proof** Define  $T : \mathbf{F}^{n,1} \rightarrow \mathbf{F}^{m,1}$  by  $Tx = Ax$ . Thus  $\mathcal{M}(T) = A$ , where  $\mathcal{M}(T)$  is computed with respect to the standard bases of  $\mathbf{F}^{n,1}$  and  $\mathbf{F}^{m,1}$ . Now

$$\begin{aligned} \text{column rank of } A &= \text{column rank of } \mathcal{M}(T) \\ &= \dim \text{range } T \\ &= \dim \text{range } T' \\ &= \text{column rank of } \mathcal{M}(T') \\ &= \text{column rank of } A^t \\ &= \text{row rank of } A, \end{aligned}$$

where the second equality above comes from 3.117, the third equality comes from 3.109(a), the fourth equality comes from 3.117 (where  $\mathcal{M}(T')$  is computed with respect to the dual bases of the standard bases), the fifth equality comes from 3.114, and the last equality follows easily from the definitions. ■

The last result allows us to dispense with the terms “row rank” and “column rank” and just use the simpler term “rank”.

**3.119 Definition rank**

The **rank** of a matrix  $A \in \mathbf{F}^{m,n}$  is the column rank of  $A$ .

## EXERCISES 3.F

---

- 1** Explain why every linear functional is either surjective or the zero map.
- 2** Give three distinct examples of linear functionals on  $\mathbf{R}^{[0,1]}$ .
- 3** Suppose  $V$  is finite-dimensional and  $v \in V$  with  $v \neq 0$ . Prove that there exists  $\varphi \in V'$  such that  $\varphi(v) = 1$ .
- 4** Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$  such that  $U \neq V$ . Prove that there exists  $\varphi \in V'$  such that  $\varphi(u) = 0$  for every  $u \in U$  but  $\varphi \neq 0$ .
- 5** Suppose  $V_1, \dots, V_m$  are vector spaces. Prove that  $(V_1 \times \dots \times V_m)'$  and  $V_1' \times \dots \times V_m'$  are isomorphic vector spaces.
- 6** Suppose  $V$  is finite-dimensional and  $v_1, \dots, v_m \in V$ . Define a linear map  $\Gamma: V' \rightarrow \mathbf{F}^m$  by
 
$$\Gamma(\varphi) = (\varphi(v_1), \dots, \varphi(v_m)).$$
  - (a) Prove that  $v_1, \dots, v_m$  spans  $V$  if and only if  $\Gamma$  is injective.
  - (b) Prove that  $v_1, \dots, v_m$  is linearly independent if and only if  $\Gamma$  is surjective.
- 7** Suppose  $m$  is a positive integer. Show that the dual basis of the basis  $1, x, \dots, x^m$  of  $\mathcal{P}_m(\mathbf{R})$  is  $\varphi_0, \varphi_1, \dots, \varphi_m$ , where  $\varphi_j(p) = \frac{p^{(j)}(0)}{j!}$ . Here  $p^{(j)}$  denotes the  $j^{\text{th}}$  derivative of  $p$ , with the understanding that the  $0^{\text{th}}$  derivative of  $p$  is  $p$ .
- 8** Suppose  $m$  is a positive integer.
  - (a) Show that  $1, x - 5, \dots, (x - 5)^m$  is a basis of  $\mathcal{P}_m(\mathbf{R})$ .
  - (b) What is the dual basis of the basis in part (a)?
- 9** Suppose  $v_1, \dots, v_n$  is a basis of  $V$  and  $\varphi_1, \dots, \varphi_n$  is the corresponding dual basis of  $V'$ . Suppose  $\psi \in V'$ . Prove that
 
$$\psi = \psi(v_1)\varphi_1 + \dots + \psi(v_n)\varphi_n.$$
- 10** Prove the first two bullet points in 3.101.

- 11** Suppose  $A$  is an  $m$ -by- $n$  matrix with  $A \neq 0$ . Prove that the rank of  $A$  is 1 if and only if there exist  $(c_1, \dots, c_m) \in \mathbf{F}^m$  and  $(d_1, \dots, d_n) \in \mathbf{F}^n$  such that  $A_{j,k} = c_j d_k$  for every  $j = 1, \dots, m$  and every  $k = 1, \dots, n$ .
- 12** Show that the dual map of the identity map on  $V$  is the identity map on  $V'$ .
- 13** Define  $T: \mathbf{R}^3 \rightarrow \mathbf{R}^2$  by  $T(x, y, z) = (4x + 5y + 6z, 7x + 8y + 9z)$ . Suppose  $\varphi_1, \varphi_2$  denotes the dual basis of the standard basis of  $\mathbf{R}^2$  and  $\psi_1, \psi_2, \psi_3$  denotes the dual basis of the standard basis of  $\mathbf{R}^3$ .
- Describe the linear functionals  $T'(\varphi_1)$  and  $T'(\varphi_2)$ .
  - Write  $T'(\varphi_1)$  and  $T'(\varphi_2)$  as linear combinations of  $\psi_1, \psi_2, \psi_3$ .
- 14** Define  $T: \mathcal{P}(\mathbf{R}) \rightarrow \mathcal{P}(\mathbf{R})$  by  $(Tp)(x) = x^2 p(x) + p''(x)$  for  $x \in \mathbf{R}$ .
- Suppose  $\varphi \in \mathcal{P}(\mathbf{R})'$  is defined by  $\varphi(p) = p'(4)$ . Describe the linear functional  $T'(\varphi)$  on  $\mathcal{P}(\mathbf{R})$ .
  - Suppose  $\varphi \in \mathcal{P}(\mathbf{R})'$  is defined by  $\varphi(p) = \int_0^1 p(x) dx$ . Evaluate  $(T'(\varphi))(x^3)$ .
- 15** Suppose  $W$  is finite-dimensional and  $T \in \mathcal{L}(V, W)$ . Prove that  $T' = 0$  if and only if  $T = 0$ .
- 16** Suppose  $V$  and  $W$  are finite-dimensional. Prove that the map that takes  $T \in \mathcal{L}(V, W)$  to  $T' \in \mathcal{L}(W', V')$  is an isomorphism of  $\mathcal{L}(V, W)$  onto  $\mathcal{L}(W', V')$ .
- 17** Suppose  $U \subset V$ . Explain why  $U^0 = \{\varphi \in V': U \subset \text{null } \varphi\}$ .
- 18** Suppose  $V$  is finite-dimensional and  $U \subset V$ . Show that  $U = \{0\}$  if and only if  $U^0 = V'$ .
- 19** Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Show that  $U = V$  if and only if  $U^0 = \{0\}$ .
- 20** Suppose  $U$  and  $W$  are subsets of  $V$  with  $U \subset W$ . Prove that  $W^0 \subset U^0$ .
- 21** Suppose  $V$  is finite-dimensional and  $U$  and  $W$  are subspaces of  $V$  with  $W^0 \subset U^0$ . Prove that  $U \subset W$ .
- 22** Suppose  $U, W$  are subspaces of  $V$ . Show that  $(U + W)^0 = U^0 \cap W^0$ .

- 23** Suppose  $V$  is finite-dimensional and  $U$  and  $W$  are subspaces of  $V$ . Prove that  $(U \cap W)^0 = U^0 + W^0$ .

- 24** Prove 3.106 using the ideas sketched in the discussion before the statement of 3.106.

- 25** Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Show that

$$U = \{v \in V : \varphi(v) = 0 \text{ for every } \varphi \in U^0\}.$$

- 26** Suppose  $V$  is finite-dimensional and  $\Gamma$  is a subspace of  $V'$ . Show that

$$\Gamma = \{v \in V : \varphi(v) = 0 \text{ for every } \varphi \in \Gamma^0\}.$$

- 27** Suppose  $T \in \mathcal{L}(\mathcal{P}_5(\mathbf{R}), \mathcal{P}_5(\mathbf{R}))$  and  $\text{null } T' = \text{span}(\varphi)$ , where  $\varphi$  is the linear functional on  $\mathcal{P}_5(\mathbf{R})$  defined by  $\varphi(p) = p(8)$ . Prove that  $\text{range } T = \{p \in \mathcal{P}_5(\mathbf{R}) : p(8) = 0\}$ .

- 28** Suppose  $V$  and  $W$  are finite-dimensional,  $T \in \mathcal{L}(V, W)$ , and there exists  $\varphi \in W'$  such that  $\text{null } T' = \text{span}(\varphi)$ . Prove that  $\text{range } T = \text{null } \varphi$ .

- 29** Suppose  $V$  and  $W$  are finite-dimensional,  $T \in \mathcal{L}(V, W)$ , and there exists  $\varphi \in V'$  such that  $\text{range } T' = \text{span}(\varphi)$ . Prove that  $\text{null } T = \text{null } \varphi$ .

- 30** Suppose  $V$  is finite-dimensional and  $\varphi_1, \dots, \varphi_m$  is a linearly independent list in  $V'$ . Prove that

$$\dim((\text{null } \varphi_1) \cap \dots \cap (\text{null } \varphi_m)) = (\dim V) - m.$$

- 31** Suppose  $V$  is finite-dimensional and  $\varphi_1, \dots, \varphi_n$  is a basis of  $V'$ . Show that there exists a basis of  $V$  whose dual basis is  $\varphi_1, \dots, \varphi_n$ .

- 32** Suppose  $T \in \mathcal{L}(V)$ , and  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  are bases of  $V$ . Prove that the following are equivalent:

- (a)  $T$  is invertible.
- (b) The columns of  $\mathcal{M}(T)$  are linearly independent in  $\mathbf{F}^{n,1}$ .
- (c) The columns of  $\mathcal{M}(T)$  span  $\mathbf{F}^{n,1}$ .
- (d) The rows of  $\mathcal{M}(T)$  are linearly independent in  $\mathbf{F}^{1,n}$ .
- (e) The rows of  $\mathcal{M}(T)$  span  $\mathbf{F}^{1,n}$ .

Here  $\mathcal{M}(T)$  means  $\mathcal{M}(T, (u_1, \dots, u_n), (v_1, \dots, v_n))$ .

**33** Suppose  $m$  and  $n$  are positive integers. Prove that the function that takes  $A$  to  $A^t$  is a linear map from  $\mathbf{F}^{m,n}$  to  $\mathbf{F}^{n,m}$ . Furthermore, prove that this linear map is invertible.

**34** The **double dual space** of  $V$ , denoted  $V''$ , is defined to be the dual space of  $V'$ . In other words,  $V'' = (V')'$ . Define  $\Lambda : V \rightarrow V''$  by

$$(\Lambda v)(\varphi) = \varphi(v)$$

for  $v \in V$  and  $\varphi \in V'$ .

- (a) Show that  $\Lambda$  is a linear map from  $V$  to  $V''$ .
- (b) Show that if  $T \in \mathcal{L}(V)$ , then  $T'' \circ \Lambda = \Lambda \circ T$ , where  $T'' = (T')'$ .
- (c) Show that if  $V$  is finite-dimensional, then  $\Lambda$  is an isomorphism from  $V$  onto  $V''$ .

[Suppose  $V$  is finite-dimensional. Then  $V$  and  $V'$  are isomorphic, but finding an isomorphism from  $V$  onto  $V'$  generally requires choosing a basis of  $V$ . In contrast, the isomorphism  $\Lambda$  from  $V$  onto  $V''$  does not require a choice of basis and thus is considered more natural.]

**35** Show that  $(\mathcal{P}(\mathbf{R}))'$  and  $\mathbf{R}^\infty$  are isomorphic.

**36** Suppose  $U$  is a subspace of  $V$ . Let  $i : U \rightarrow V$  be the inclusion map defined by  $i(u) = u$ . Thus  $i' \in \mathcal{L}(V', U')$ .

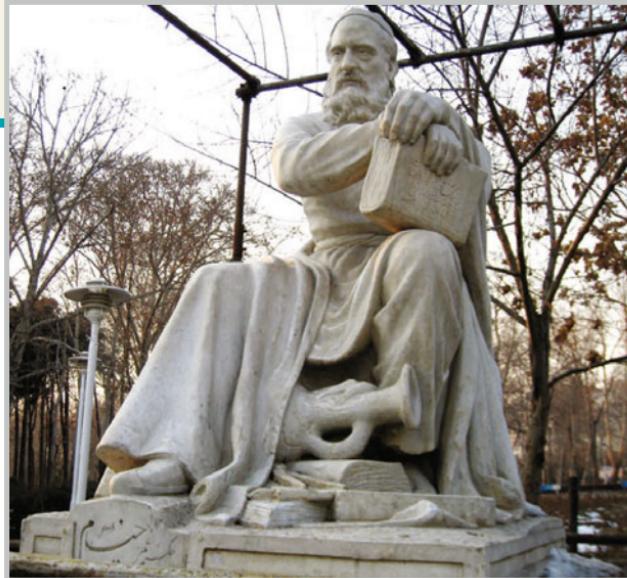
- (a) Show that  $\text{null } i' = U^0$ .
- (b) Prove that if  $V$  is finite-dimensional, then  $\text{range } i' = U'$ .
- (c) Prove that if  $V$  is finite-dimensional, then  $\tilde{i}'$  is an isomorphism from  $V'/U^0$  onto  $U'$ .

[The isomorphism in part (c) is natural in that it does not depend on a choice of basis in either vector space.]

**37** Suppose  $U$  is a subspace of  $V$ . Let  $\pi : V \rightarrow V/U$  be the usual quotient map. Thus  $\pi' \in \mathcal{L}((V/U)', V')$ .

- (a) Show that  $\pi'$  is injective.
- (b) Show that  $\text{range } \pi' = U^0$ .
- (c) Conclude that  $\pi'$  is an isomorphism from  $(V/U)'$  onto  $U^0$ .

[The isomorphism in part (c) is natural in that it does not depend on a choice of basis in either vector space. In fact, there is no assumption here that any of these vector spaces are finite-dimensional.]



*Statue of Persian mathematician and poet Omar Khayyám (1048–1131), whose algebra book written in 1070 contained the first serious study of cubic polynomials.*

# Polynomials

This short chapter contains material on polynomials that we will need to understand operators. Many of the results in this chapter will already be familiar to you from other courses; they are included here for completeness.

Because this chapter is not about linear algebra, your instructor may go through it rapidly. You may not be asked to scrutinize all the proofs. Make sure, however, that you at least read and understand the statements of all the results in this chapter—they will be used in later chapters.

The standing assumption we need for this chapter is as follows:

## 4.1 Notation F

F denotes **R** or **C**.

### LEARNING OBJECTIVES FOR THIS CHAPTER

- Division Algorithm for Polynomials
- factorization of polynomials over **C**
- factorization of polynomials over **R**

## Complex Conjugate and Absolute Value

Before discussing polynomials with complex or real coefficients, we need to learn a bit more about the complex numbers.

### 4.2 Definition $\operatorname{Re} z, \operatorname{Im} z$

Suppose  $z = a + bi$ , where  $a$  and  $b$  are real numbers.

- The ***real part*** of  $z$ , denoted  $\operatorname{Re} z$ , is defined by  $\operatorname{Re} z = a$ .
- The ***imaginary part*** of  $z$ , denoted  $\operatorname{Im} z$ , is defined by  $\operatorname{Im} z = b$ .

Thus for every complex number  $z$ , we have

$$z = \operatorname{Re} z + (\operatorname{Im} z)i.$$

### 4.3 Definition *complex conjugate, $\bar{z}$ , absolute value, $|z|$*

Suppose  $z \in \mathbf{C}$ .

- The ***complex conjugate*** of  $z \in \mathbf{C}$ , denoted  $\bar{z}$ , is defined by

$$\bar{z} = \operatorname{Re} z - (\operatorname{Im} z)i.$$

- The ***absolute value*** of a complex number  $z$ , denoted  $|z|$ , is defined by

$$|z| = \sqrt{(\operatorname{Re} z)^2 + (\operatorname{Im} z)^2}.$$

---

### 4.4 Example Suppose $z = 3 + 2i$ . Then

- $\operatorname{Re} z = 3$  and  $\operatorname{Im} z = 2$ ;
  - $\bar{z} = 3 - 2i$ ;
  - $|z| = \sqrt{3^2 + 2^2} = \sqrt{13}$ .
- 

Note that  $|z|$  is a nonnegative number for every  $z \in \mathbf{C}$ .

You should verify that  $z = \bar{z}$  if and only if  $z$  is a real number.

The real and imaginary parts, complex conjugate, and absolute value have the following properties:

## 4.5 Properties of complex numbers

Suppose  $w, z \in \mathbf{C}$ . Then

**sum of  $z$  and  $\bar{z}$**

$$z + \bar{z} = 2 \operatorname{Re} z;$$

**difference of  $z$  and  $\bar{z}$**

$$z - \bar{z} = 2(\operatorname{Im} z)i;$$

**product of  $z$  and  $\bar{z}$**

$$z\bar{z} = |z|^2;$$

**additivity and multiplicativity of complex conjugate**

$$\overline{w+z} = \bar{w} + \bar{z} \text{ and } \overline{wz} = \bar{w}\bar{z};$$

**conjugate of conjugate**

$$\bar{\bar{z}} = z;$$

**real and imaginary parts are bounded by  $|z|$**

$$|\operatorname{Re} z| \leq |z| \text{ and } |\operatorname{Im} z| \leq |z|$$

**absolute value of the complex conjugate**

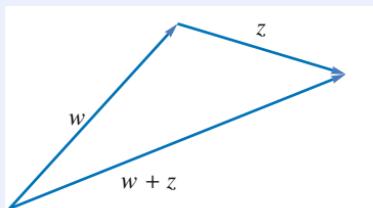
$$|\bar{z}| = |z|;$$

**multiplicativity of absolute value**

$$|wz| = |w||z|;$$

**Triangle Inequality**

$$|w+z| \leq |w| + |z|.$$



**Proof** Except for the last item, the routine verifications of the assertions above are left to the reader. To verify the last item, we have

$$\begin{aligned} |w+z|^2 &= (w+z)(\bar{w}+\bar{z}) \\ &= w\bar{w} + z\bar{z} + w\bar{z} + z\bar{w} \\ &= |w|^2 + |z|^2 + w\bar{z} + \bar{w}z \\ &= |w|^2 + |z|^2 + 2 \operatorname{Re}(w\bar{z}) \\ &\leq |w|^2 + |z|^2 + 2|w\bar{z}| \\ &= |w|^2 + |z|^2 + 2|w||z| \\ &= (|w| + |z|)^2. \end{aligned}$$

Taking the square root of both sides of the inequality  $|w+z|^2 \leq (|w| + |z|)^2$  now gives the desired inequality. ■

## Uniqueness of Coefficients for Polynomials

Recall that a function  $p : \mathbf{F} \rightarrow \mathbf{F}$  is called a polynomial with coefficients in  $\mathbf{F}$  if there exist  $a_0, \dots, a_m \in \mathbf{F}$  such that

$$4.6 \quad p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m$$

for all  $z \in \mathbf{F}$ .

### 4.7 If a polynomial is the zero function, then all coefficients are 0

Suppose  $a_0, \dots, a_m \in \mathbf{F}$ . If

$$a_0 + a_1 z + \cdots + a_m z^m = 0$$

for every  $z \in \mathbf{F}$ , then  $a_0 = \cdots = a_m = 0$ .

**Proof** We will prove the contrapositive. If not all the coefficients are 0, then by changing  $m$  we can assume  $a_m \neq 0$ . Let

$$z = \frac{|a_0| + |a_1| + \cdots + |a_{m-1}|}{|a_m|} + 1.$$

Note that  $z \geq 1$ , and thus  $z^j \leq z^{m-1}$  for  $j = 0, 1, \dots, m-1$ . Using the Triangle Inequality, we have

$$\begin{aligned} |a_0 + a_1 z + \cdots + a_{m-1} z^{m-1}| &\leq (|a_0| + |a_1| + \cdots + |a_{m-1}|)z^{m-1} \\ &< |a_m z^m|. \end{aligned}$$

Thus  $a_0 + a_1 z + \cdots + a_{m-1} z^{m-1} \neq -a_m z^m$ . Hence we conclude that  $a_0 + a_1 z + \cdots + a_{m-1} z^{m-1} + a_m z^m \neq 0$ . ■

The result above implies that the coefficients of a polynomial are uniquely determined (because if a polynomial had two different sets of coefficients, then subtracting the two representations of the polynomial would give a contradiction to the result above).

Recall that if a polynomial  $p$  can be written in the form 4.6 with  $a_m \neq 0$ , then we say that  $p$  has degree  $m$  and we write  $\deg p = m$ .

*The 0 polynomial is declared to have degree  $-\infty$  so that exceptions are not needed for various reasonable results. For example,  $\deg(pq) = \deg p + \deg q$  even if  $p = 0$ .*

The degree of the 0 polynomial is defined to be  $-\infty$ . When necessary, use the obvious arithmetic with  $-\infty$ . For example,  $-\infty < m$  and  $-\infty + m = -\infty$  for every integer  $m$ .

## The Division Algorithm for Polynomials

If  $p$  and  $s$  are nonnegative integers, with  $s \neq 0$ , then there exist nonnegative integers  $q$  and  $r$  such that

$$p = sq + r$$

and  $r < s$ . Think of dividing  $p$  by  $s$ , getting quotient  $q$  with remainder  $r$ . Our next task is to prove an analogous result for polynomials.

The result below is often called the Division Algorithm for Polynomials, although as stated here it is not really an algorithm, just a useful result.

*Think of the Division Algorithm for Polynomials as giving the remainder  $r$  when  $p$  is divided by  $s$ .*

Recall that  $\mathcal{P}(\mathbf{F})$  denotes the vector space of all polynomials with coefficients in  $\mathbf{F}$  and that  $\mathcal{P}_m(\mathbf{F})$  is the subspace of  $\mathcal{P}(\mathbf{F})$  consisting of the polynomials with coefficients in  $\mathbf{F}$  and degree at most  $m$ .

The next result can be proved without linear algebra, but the proof given here using linear algebra is appropriate for a linear algebra textbook.

### 4.8 Division Algorithm for Polynomials

Suppose that  $p, s \in \mathcal{P}(\mathbf{F})$ , with  $s \neq 0$ . Then there exist unique polynomials  $q, r \in \mathcal{P}(\mathbf{F})$  such that

$$p = sq + r$$

and  $\deg r < \deg s$ .

**Proof** Let  $n = \deg p$  and  $m = \deg s$ . If  $n < m$ , then take  $q = 0$  and  $r = p$  to get the desired result. Thus we can assume that  $n \geq m$ .

Define  $T: \mathcal{P}_{n-m}(\mathbf{F}) \times \mathcal{P}_{m-1}(\mathbf{F}) \rightarrow \mathcal{P}_n(\mathbf{F})$  by

$$T(q, r) = sq + r.$$

The reader can easily verify that  $T$  is a linear map. If  $(q, r) \in \text{null } T$ , then  $sq + r = 0$ , which implies that  $q = 0$  and  $r = 0$  [because otherwise  $\deg sq \geq m$  and thus  $sq$  cannot equal  $-r$ ]. Thus  $\dim \text{null } T = 0$  (proving the “unique” part of the result).

From 3.76 we have

$$\dim(\mathcal{P}_{n-m}(\mathbf{F}) \times \mathcal{P}_{m-1}(\mathbf{F})) = (n - m + 1) + (m - 1 + 1) = n + 1.$$

The Fundamental Theorem of Linear Maps (3.22) and the equation displayed above now imply that  $\dim \text{range } T = n + 1$ , which equals  $\dim \mathcal{P}_n(\mathbf{F})$ . Thus  $\text{range } T = \mathcal{P}_n(\mathbf{F})$ , and hence there exist  $q \in \mathcal{P}_{n-m}(\mathbf{F})$  and  $r \in \mathcal{P}_{m-1}(\mathbf{F})$  such that  $p = T(q, r) = sq + r$ . ■

## Zeros of Polynomials

The solutions to the equation  $p(z) = 0$  play a crucial role in the study of a polynomial  $p \in \mathcal{P}(\mathbf{F})$ . Thus these solutions have a special name.

### 4.9 Definition zero of a polynomial

A number  $\lambda \in \mathbf{F}$  is called a *zero* (or *root*) of a polynomial  $p \in \mathcal{P}(\mathbf{F})$  if

$$p(\lambda) = 0.$$

### 4.10 Definition factor

A polynomial  $s \in \mathcal{P}(\mathbf{F})$  is called a *factor* of  $p \in \mathcal{P}(\mathbf{F})$  if there exists a polynomial  $q \in \mathcal{P}(\mathbf{F})$  such that  $p = sq$ .

We begin by showing that  $\lambda$  is a zero of a polynomial  $p \in \mathcal{P}(\mathbf{F})$  if and only if  $z - \lambda$  is a factor of  $p$ .

### 4.11 Each zero of a polynomial corresponds to a degree-1 factor

Suppose  $p \in \mathcal{P}(\mathbf{F})$  and  $\lambda \in \mathbf{F}$ . Then  $p(\lambda) = 0$  if and only if there is a polynomial  $q \in \mathcal{P}(\mathbf{F})$  such that

$$p(z) = (z - \lambda)q(z)$$

for every  $z \in \mathbf{F}$ .

**Proof** One direction is obvious. Namely, suppose there is a polynomial  $q \in \mathcal{P}(\mathbf{F})$  such that  $p(z) = (z - \lambda)q(z)$  for all  $z \in \mathbf{F}$ . Then

$$p(\lambda) = (\lambda - \lambda)q(\lambda) = 0,$$

as desired.

To prove the other direction, suppose  $p(\lambda) = 0$ . The polynomial  $z - \lambda$  has degree 1. Because a polynomial with degree less than 1 is a constant function, the Division Algorithm for Polynomials (4.8) implies that there exist a polynomial  $q \in \mathcal{P}(\mathbf{F})$  and a number  $r \in \mathbf{F}$  such that

$$p(z) = (z - \lambda)q(z) + r$$

for every  $z \in \mathbf{F}$ . The equation above and the equation  $p(\lambda) = 0$  imply that  $r = 0$ . Thus  $p(z) = (z - \lambda)q(z)$  for every  $z \in \mathbf{F}$ . ■

Now we can prove that polynomials do not have too many zeros.

### 4.12 A polynomial has at most as many zeros as its degree

Suppose  $p \in \mathcal{P}(\mathbf{F})$  is a polynomial with degree  $m \geq 0$ . Then  $p$  has at most  $m$  distinct zeros in  $\mathbf{F}$ .

**Proof** If  $m = 0$ , then  $p(z) = a_0 \neq 0$  and so  $p$  has no zeros.

If  $m = 1$ , then  $p(z) = a_0 + a_1 z$ , with  $a_1 \neq 0$ , and thus  $p$  has exactly one zero, namely,  $-a_0/a_1$ .

Now suppose  $m > 1$ . We use induction on  $m$ , assuming that every polynomial with degree  $m - 1$  has at most  $m - 1$  distinct zeros. If  $p$  has no zeros in  $\mathbf{F}$ , then we are done. If  $p$  has a zero  $\lambda \in \mathbf{F}$ , then by 4.11 there is a polynomial  $q$  such that

$$p(z) = (z - \lambda)q(z)$$

for all  $z \in \mathbf{F}$ . Clearly  $\deg q = m - 1$ . The equation above shows that if  $p(z) = 0$ , then either  $z = \lambda$  or  $q(z) = 0$ . In other words, the zeros of  $p$  consist of  $\lambda$  and the zeros of  $q$ . By our induction hypothesis,  $q$  has at most  $m - 1$  distinct zeros in  $\mathbf{F}$ . Thus  $p$  has at most  $m$  distinct zeros in  $\mathbf{F}$ . ■

## Factorization of Polynomials over $\mathbf{C}$

So far we have been handling polynomials with complex coefficients and polynomials with real coefficients simultaneously through our convention that  $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ . Now we will see some differences between these two cases. First we treat polynomials with complex coefficients. Then we will use our results about polynomials with complex coefficients to prove corresponding results for polynomials with real coefficients.

The next result, although called the Fundamental Theorem of Algebra, uses analysis in its proof. The short proof presented here uses tools from complex analysis. If you have not had a course in complex analysis, this proof will almost certainly be meaningless to you. In that case, just accept the Fundamental Theorem of Algebra as something that we need to use but whose proof requires more advanced tools that you may learn in later courses.

*The Fundamental Theorem of Algebra is an existence theorem. Its proof does not lead to a method for finding zeros. The quadratic formula gives the zeros explicitly for polynomials of degree 2. Similar but more complicated formulas exist for polynomials of degree 3 and 4. No such formulas exist for polynomials of degree 5 and above.*

### 4.13 Fundamental Theorem of Algebra

Every nonconstant polynomial with complex coefficients has a zero.

**Proof** Let  $p$  be a nonconstant polynomial with complex coefficients. Suppose  $p$  has no zeros. Then  $1/p$  is an analytic function on  $\mathbf{C}$ . Furthermore,  $|p(z)| \rightarrow \infty$  as  $|z| \rightarrow \infty$ , which implies that  $1/p \rightarrow 0$  as  $|z| \rightarrow \infty$ . Thus  $1/p$  is a bounded analytic function on  $\mathbf{C}$ . By Liouville's theorem, every such function is constant. But if  $1/p$  is constant, then  $p$  is constant, contradicting our assumption that  $p$  is nonconstant. ■

Although the proof given above is probably the shortest proof of the Fundamental Theorem of Algebra, a web search can lead you to several other proofs that use different techniques. All proofs of the Fundamental Theorem of Algebra need to use some analysis, because the result is not true if  $\mathbf{C}$  is replaced, for example, with the set of numbers of the form  $c + di$  where  $c, d$  are rational numbers.

*The cubic formula, which was discovered in the 16<sup>th</sup> century, is presented below for your amusement only. Do not memorize it.*

*Suppose*

$$p(x) = ax^3 + bx^2 + cx + d,$$

*where  $a \neq 0$ . Set*

$$u = \frac{9abc - 2b^3 - 27a^2d}{54a^3}$$

*and then set*

$$v = u^2 + \left( \frac{3ac - b^2}{9a^2} \right)^3.$$

*Suppose  $v \geq 0$ . Then*

$$-\frac{b}{3a} + \sqrt[3]{u + \sqrt{v}} + \sqrt[3]{u - \sqrt{v}}$$

*is a zero of  $p$ .*

Remarkably, mathematicians have proved that no formula exists for the zeros of polynomials of degree 5 or higher. But computers and calculators can use clever numerical methods to find good approximations to the zeros of any polynomial, even when exact zeros cannot be found.

For example, no one will ever be able to give an exact formula for a zero of the polynomial  $p$  defined by

$$p(x) = x^5 - 5x^4 - 6x^3 + 17x^2 + 4x - 7.$$

However, a computer or symbolic calculator can find approximate zeros of this polynomial.

The Fundamental Theorem of Algebra leads to the following factorization result for polynomials with complex coefficients. Note that in this factorization, the numbers  $\lambda_1, \dots, \lambda_m$  are precisely the zeros of  $p$ , for these are the only values of  $z$  for which the right side of the equation in the next result equals 0.

#### 4.14 Factorization of a polynomial over $\mathbf{C}$

If  $p \in \mathcal{P}(\mathbf{C})$  is a nonconstant polynomial, then  $p$  has a unique factorization (except for the order of the factors) of the form

$$p(z) = c(z - \lambda_1) \cdots (z - \lambda_m),$$

where  $c, \lambda_1, \dots, \lambda_m \in \mathbf{C}$ .

**Proof** Let  $p \in \mathcal{P}(\mathbf{C})$  and let  $m = \deg p$ . We will use induction on  $m$ . If  $m = 1$ , then clearly the desired factorization exists and is unique. So assume that  $m > 1$  and that the desired factorization exists and is unique for all polynomials of degree  $m - 1$ .

First we will show that the desired factorization of  $p$  exists. By the Fundamental Theorem of Algebra (4.13),  $p$  has a zero  $\lambda$ . By 4.11, there is a polynomial  $q$  such that

$$p(z) = (z - \lambda)q(z)$$

for all  $z \in \mathbf{C}$ . Because  $\deg q = m - 1$ , our induction hypothesis implies that  $q$  has the desired factorization, which when plugged into the equation above gives the desired factorization of  $p$ .

Now we turn to the question of uniqueness. Clearly  $c$  is uniquely determined as the coefficient of  $z^m$  in  $p$ . So we need only show that except for the order, there is only one way to choose  $\lambda_1, \dots, \lambda_m$ . If

$$(z - \lambda_1) \cdots (z - \lambda_m) = (z - \tau_1) \cdots (z - \tau_m)$$

for all  $z \in \mathbf{C}$ , then because the left side of the equation above equals 0 when  $z = \lambda_1$ , one of the  $\tau$ 's on the right side equals  $\lambda_1$ . Relabeling, we can assume that  $\tau_1 = \lambda_1$ . Now for  $z \neq \lambda_1$ , we can divide both sides of the equation above by  $z - \lambda_1$ , getting

$$(z - \lambda_2) \cdots (z - \lambda_m) = (z - \tau_2) \cdots (z - \tau_m)$$

for all  $z \in \mathbf{C}$  except possibly  $z = \lambda_1$ . Actually the equation above holds for all  $z \in \mathbf{C}$ , because otherwise by subtracting the right side from the left side we would get a nonzero polynomial that has infinitely many zeros. The equation above and our induction hypothesis imply that except for the order, the  $\lambda$ 's are the same as the  $\tau$ 's, completing the proof of uniqueness. ■

## Factorization of Polynomials over $\mathbf{R}$

*The failure of the Fundamental Theorem of Algebra for  $\mathbf{R}$  accounts for the differences between operators on real and complex vector spaces, as we will see in later chapters.*

A polynomial with real coefficients may have no real zeros. For example, the polynomial  $1 + x^2$  has no real zeros.

To obtain a factorization theorem over  $\mathbf{R}$ , we will use our factorization theorem over  $\mathbf{C}$ . We begin with the following result.

### 4.15 Polynomials with real coefficients have zeros in pairs

Suppose  $p \in \mathcal{P}(\mathbf{C})$  is a polynomial with real coefficients. If  $\lambda \in \mathbf{C}$  is a zero of  $p$ , then so is  $\bar{\lambda}$ .

**Proof** Let

$$p(z) = a_0 + a_1 z + \cdots + a_m z^m,$$

where  $a_0, \dots, a_m$  are real numbers. Suppose  $\lambda \in \mathbf{C}$  is a zero of  $p$ . Then

$$a_0 + a_1 \lambda + \cdots + a_m \lambda^m = 0.$$

Take the complex conjugate of both sides of this equation, obtaining

$$a_0 + a_1 \bar{\lambda} + \cdots + a_m \bar{\lambda}^m = 0,$$

where we have used basic properties of complex conjugation (see 4.5). The equation above shows that  $\bar{\lambda}$  is a zero of  $p$ . ■

*Think about the connection between the quadratic formula and 4.16.*

We want a factorization theorem for polynomials with real coefficients. First we need to characterize the polynomials of degree 2 with real coefficients that can be written as the product of two polynomials of degree 1 with real coefficients.

### 4.16 Factorization of a quadratic polynomial

Suppose  $b, c \in \mathbf{R}$ . Then there is a polynomial factorization of the form

$$x^2 + bx + c = (x - \lambda_1)(x - \lambda_2)$$

with  $\lambda_1, \lambda_2 \in \mathbf{R}$  if and only if  $b^2 \geq 4c$ .

**Proof** Notice that

$$x^2 + bx + c = \left(x + \frac{b}{2}\right)^2 + \left(c - \frac{b^2}{4}\right).$$

First suppose  $b^2 < 4c$ . Then clearly the right side of the equation above is positive for every  $x \in \mathbf{R}$ . Hence the polynomial  $x^2 + bx + c$  has no real zeros and thus cannot be factored in the form  $(x - \lambda_1)(x - \lambda_2)$  with  $\lambda_1, \lambda_2 \in \mathbf{R}$ .

The equation above is the basis of the technique called **completing the square**.

Conversely, now suppose  $b^2 \geq 4c$ . Then there is a real number  $d$  such that  $d^2 = \frac{b^2}{4} - c$ . From the displayed equation above, we have

$$\begin{aligned} x^2 + bx + c &= \left(x + \frac{b}{2}\right)^2 - d^2 \\ &= \left(x + \frac{b}{2} + d\right)\left(x + \frac{b}{2} - d\right), \end{aligned}$$

which gives the desired factorization. ■

The next result gives a factorization of a polynomial over  $\mathbf{R}$ . The idea of the proof is to use the factorization 4.14 of  $p$  as a polynomial with complex coefficients. Complex but nonreal zeros of  $p$  come in pairs; see 4.15. Thus if the factorization of  $p$  as an element of  $\mathcal{P}(\mathbf{C})$  includes terms of the form  $(x - \lambda)$  with  $\lambda$  a nonreal complex number, then  $(x - \bar{\lambda})$  is also a term in the factorization. Multiplying together these two terms, we get

$$(x^2 - 2(\operatorname{Re} \lambda)x + |\lambda|^2),$$

which is a quadratic term of the required form.

The idea sketched in the paragraph above almost provides a proof of the existence of our desired factorization. However, we need to be careful about one point. Suppose  $\lambda$  is a nonreal complex number and  $(x - \lambda)$  is a term in the factorization of  $p$  as an element of  $\mathcal{P}(\mathbf{C})$ . We are guaranteed by 4.15 that  $(x - \bar{\lambda})$  also appears as a term in the factorization, but 4.15 does not state that these two factors appear the same number of times, as needed to make the idea above work. However, the proof works around this point.

In the next result, either  $m$  or  $M$  may equal 0. The numbers  $\lambda_1, \dots, \lambda_m$  are precisely the real zeros of  $p$ , for these are the only real values of  $x$  for which the right side of the equation in the next result equals 0.

### 4.17 Factorization of a polynomial over $\mathbf{R}$

Suppose  $p \in \mathcal{P}(\mathbf{R})$  is a nonconstant polynomial. Then  $p$  has a unique factorization (except for the order of the factors) of the form

$$p(x) = c(x - \lambda_1) \cdots (x - \lambda_m)(x^2 + b_1x + c_1) \cdots (x^2 + b_Mx + c_M),$$

where  $c, \lambda_1, \dots, \lambda_m, b_1, \dots, b_M, c_1, \dots, c_M \in \mathbf{R}$ , with  $b_j^2 < 4c_j$  for each  $j$ .

**Proof** Think of  $p$  as an element of  $\mathcal{P}(\mathbf{C})$ . If all the (complex) zeros of  $p$  are real, then we are done by 4.14. Thus suppose  $p$  has a zero  $\lambda \in \mathbf{C}$  with  $\lambda \notin \mathbf{R}$ . By 4.15,  $\bar{\lambda}$  is a zero of  $p$ . Thus we can write

$$\begin{aligned} p(x) &= (x - \lambda)(x - \bar{\lambda})q(x) \\ &= (x^2 - 2(\operatorname{Re} \lambda)x + |\lambda|^2)q(x) \end{aligned}$$

for some polynomial  $q \in \mathcal{P}(\mathbf{C})$  with degree two less than the degree of  $p$ . If we can prove that  $q$  has real coefficients, then by using induction on the degree of  $p$ , we can conclude that  $(x - \lambda)$  appears in the factorization of  $p$  exactly as many times as  $(x - \bar{\lambda})$ .

To prove that  $q$  has real coefficients, we solve the equation above for  $q$ , getting

$$q(x) = \frac{p(x)}{x^2 - 2(\operatorname{Re} \lambda)x + |\lambda|^2}$$

for all  $x \in \mathbf{R}$ . The equation above implies that  $q(x) \in \mathbf{R}$  for all  $x \in \mathbf{R}$ . Writing

$$q(x) = a_0 + a_1x + \cdots + a_{n-2}x^{n-2},$$

where  $n = \deg p$  and  $a_0, \dots, a_{n-2} \in \mathbf{C}$ , we thus have

$$0 = \operatorname{Im} q(x) = (\operatorname{Im} a_0) + (\operatorname{Im} a_1)x + \cdots + (\operatorname{Im} a_{n-2})x^{n-2}$$

for all  $x \in \mathbf{R}$ . This implies that  $\operatorname{Im} a_0, \dots, \operatorname{Im} a_{n-2}$  all equal 0 (by 4.7). Thus all the coefficients of  $q$  are real, as desired. Hence the desired factorization exists.

Now we turn to the question of uniqueness of our factorization. A factor of  $p$  of the form  $x^2 + b_jx + c_j$  with  $b_j^2 < 4c_j$  can be uniquely written as  $(x - \lambda_j)(x - \bar{\lambda}_j)$  with  $\lambda_j \in \mathbf{C}$ . A moment's thought shows that two different factorizations of  $p$  as an element of  $\mathcal{P}(\mathbf{R})$  would lead to two different factorizations of  $p$  as an element of  $\mathcal{P}(\mathbf{C})$ , contradicting 4.14. ■

## EXERCISES 4

1 Verify all the assertions in 4.5 except the last one.

2 Suppose  $m$  is a positive integer. Is the set

$$\{0\} \cup \{p \in \mathcal{P}(\mathbf{F}) : \deg p = m\}$$

a subspace of  $\mathcal{P}(\mathbf{F})$ ?

3 Is the set

$$\{0\} \cup \{p \in \mathcal{P}(\mathbf{F}) : \deg p \text{ is even}\}$$

a subspace of  $\mathcal{P}(\mathbf{F})$ ?

4 Suppose  $m$  and  $n$  are positive integers with  $m \leq n$ , and suppose  $\lambda_1, \dots, \lambda_m \in \mathbf{F}$ . Prove that there exists a polynomial  $p \in \mathcal{P}(\mathbf{F})$  with  $\deg p = n$  such that  $0 = p(\lambda_1) = \dots = p(\lambda_m)$  and such that  $p$  has no other zeros.

5 Suppose  $m$  is a nonnegative integer,  $z_1, \dots, z_{m+1}$  are distinct elements of  $\mathbf{F}$ , and  $w_1, \dots, w_{m+1} \in \mathbf{F}$ . Prove that there exists a unique polynomial  $p \in \mathcal{P}_m(\mathbf{F})$  such that

$$p(z_j) = w_j$$

for  $j = 1, \dots, m + 1$ .

[This result can be proved without using linear algebra. However, try to find the clearer, shorter proof that uses some linear algebra.]

6 Suppose  $p \in \mathcal{P}(\mathbf{C})$  has degree  $m$ . Prove that  $p$  has  $m$  distinct zeros if and only if  $p$  and its derivative  $p'$  have no zeros in common.

7 Prove that every polynomial of odd degree with real coefficients has a real zero.

8 Define  $T: \mathcal{P}(\mathbf{R}) \rightarrow \mathbf{R}^{\mathbf{R}}$  by

$$Tp = \begin{cases} \frac{p - p(3)}{x - 3} & \text{if } x \neq 3, \\ p'(3) & \text{if } x = 3. \end{cases}$$

Show that  $Tp \in \mathcal{P}(\mathbf{R})$  for every polynomial  $p \in \mathcal{P}(\mathbf{R})$  and that  $T$  is a linear map.

- 9 Suppose  $p \in \mathcal{P}(\mathbf{C})$ . Define  $q: \mathbf{C} \rightarrow \mathbf{C}$  by

$$q(z) = p(z)\overline{p(\bar{z})}.$$

Prove that  $q$  is a polynomial with real coefficients.

- 10 Suppose  $m$  is a nonnegative integer and  $p \in \mathcal{P}_m(\mathbf{C})$  is such that there exist distinct real numbers  $x_0, x_1, \dots, x_m$  such that  $p(x_j) \in \mathbf{R}$  for  $j = 0, 1, \dots, m$ . Prove that all the coefficients of  $p$  are real.
- 11 Suppose  $p \in \mathcal{P}(\mathbf{F})$  with  $p \neq 0$ . Let  $U = \{pq : q \in \mathcal{P}(\mathbf{F})\}$ .
- Show that  $\dim \mathcal{P}(\mathbf{F})/U = \deg p$ .
  - Find a basis of  $\dim \mathcal{P}(\mathbf{F})/U$ .



# CHAPTER

# 5

*Statue of Italian mathematician Leonardo of Pisa (1170–1250, approximate dates), also known as Fibonacci. Exercise 16 in Section 5.C shows how linear algebra can be used to find an explicit formula for the Fibonacci sequence.*

# Eigenvalues, Eigenvectors, and Invariant Subspaces

Linear maps from one vector space to another vector space were the objects of study in Chapter 3. Now we begin our investigation of linear maps from a finite-dimensional vector space to itself. Their study constitutes the most important part of linear algebra.

Our standing assumptions are as follows:

## 5.1 Notation $\mathbf{F}, V$

- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  denotes a vector space over  $\mathbf{F}$ .

## LEARNING OBJECTIVES FOR THIS CHAPTER

- invariant subspaces
- eigenvalues, eigenvectors, and eigenspaces
- each operator on a finite-dimensional complex vector space has an eigenvalue and an upper-triangular matrix with respect to some basis

## 5.A Invariant Subspaces

In this chapter we develop the tools that will help us understand the structure of operators. Recall that an operator is a linear map from a vector space to itself. Recall also that we denote the set of operators on  $V$  by  $\mathcal{L}(V)$ ; in other words,  $\mathcal{L}(V) = \mathcal{L}(V, V)$ .

Let's see how we might better understand what an operator looks like. Suppose  $T \in \mathcal{L}(V)$ . If we have a direct sum decomposition

$$V = U_1 \oplus \cdots \oplus U_m,$$

where each  $U_j$  is a proper subspace of  $V$ , then to understand the behavior of  $T$ , we need only understand the behavior of each  $T|_{U_j}$ ; here  $T|_{U_j}$  denotes the restriction of  $T$  to the smaller domain  $U_j$ . Dealing with  $T|_{U_j}$  should be easier than dealing with  $T$  because  $U_j$  is a smaller vector space than  $V$ .

However, if we intend to apply tools useful in the study of operators (such as taking powers), then we have a problem:  $T|_{U_j}$  may not map  $U_j$  into itself; in other words,  $T|_{U_j}$  may not be an operator on  $U_j$ . Thus we are led to consider only decompositions of  $V$  of the form above where  $T$  maps each  $U_j$  into itself.

The notion of a subspace that gets mapped into itself is sufficiently important to deserve a name.

### 5.2 Definition invariant subspace

Suppose  $T \in \mathcal{L}(V)$ . A subspace  $U$  of  $V$  is called **invariant** under  $T$  if  $u \in U$  implies  $Tu \in U$ .

In other words,  $U$  is invariant under  $T$  if  $T|_U$  is an operator on  $U$ .

**5.3 Example** Suppose  $T \in \mathcal{L}(V)$ . Show that each of the following subspaces of  $V$  is invariant under  $T$ :

- (a)  $\{0\}$ ;
- (b)  $V$ ;
- (c)  $\text{null } T$ ;
- (d)  $\text{range } T$ .

The most famous unsolved problem in functional analysis is called the **invariant subspace problem**. It deals with invariant subspaces of operators on infinite-dimensional vector spaces.

**Solution**

- (a) If  $u \in \{0\}$ , then  $u = 0$  and hence  $Tu = 0 \in \{0\}$ . Thus  $\{0\}$  is invariant under  $T$ .
  - (b) If  $u \in V$ , then  $Tu \in V$ . Thus  $V$  is invariant under  $T$ .
  - (c) If  $u \in \text{null } T$ , then  $Tu = 0$ , and hence  $Tu \in \text{null } T$ . Thus  $\text{null } T$  is invariant under  $T$ .
  - (d) If  $u \in \text{range } T$ , then  $Tu \in \text{range } T$ . Thus  $\text{range } T$  is invariant under  $T$ .
- 

Must an operator  $T \in \mathcal{L}(V)$  have any invariant subspaces other than  $\{0\}$  and  $V$ ? Later we will see that this question has an affirmative answer if  $V$  is finite-dimensional and  $\dim V > 1$  (for  $\mathbf{F} = \mathbf{C}$ ) or  $\dim V > 2$  (for  $\mathbf{F} = \mathbf{R}$ ); see 5.21 and 9.8.

Although  $\text{null } T$  and  $\text{range } T$  are invariant under  $T$ , they do not necessarily provide easy answers to the question about the existence of invariant subspaces other than  $\{0\}$  and  $V$ , because  $\text{null } T$  may equal  $\{0\}$  and  $\text{range } T$  may equal  $V$  (this happens when  $T$  is invertible).

---

**5.4 Example** Suppose that  $T \in \mathcal{L}(\mathcal{P}(\mathbf{R}))$  is defined by  $Tp = p'$ . Then  $\mathcal{P}_4(\mathbf{R})$ , which is a subspace of  $\mathcal{P}(\mathbf{R})$ , is invariant under  $T$  because if  $p \in \mathcal{P}(\mathbf{R})$  has degree at most 4, then  $p'$  also has degree at most 4.

---

## Eigenvalues and Eigenvectors

We will return later to a deeper study of invariant subspaces. Now we turn to an investigation of the simplest possible nontrivial invariant subspaces—*invariant subspaces with dimension 1*.

Take any  $v \in V$  with  $v \neq 0$  and let  $U$  equal the set of all scalar multiples of  $v$ :

$$U = \{\lambda v : \lambda \in \mathbf{F}\} = \text{span}(v).$$

Then  $U$  is a 1-dimensional subspace of  $V$  (and every 1-dimensional subspace of  $V$  is of this form for an appropriate choice of  $v$ ). If  $U$  is invariant under an operator  $T \in \mathcal{L}(V)$ , then  $Tv \in U$ , and hence there is a scalar  $\lambda \in \mathbf{F}$  such that

$$Tv = \lambda v.$$

Conversely, if  $Tv = \lambda v$  for some  $\lambda \in \mathbf{F}$ , then  $\text{span}(v)$  is a 1-dimensional subspace of  $V$  invariant under  $T$ .

The equation

$$Tv = \lambda v,$$

which we have just seen is intimately connected with 1-dimensional invariant subspaces, is important enough that the vectors  $v$  and scalars  $\lambda$  satisfying it are given special names.

### 5.5 Definition eigenvalue

Suppose  $T \in \mathcal{L}(V)$ . A number  $\lambda \in \mathbf{F}$  is called an *eigenvalue* of  $T$  if there exists  $v \in V$  such that  $v \neq 0$  and  $Tv = \lambda v$ .

*The word eigenvalue is half-German, half-English. The German adjective **eigen** means “own” in the sense of characterizing an intrinsic property. Some mathematicians use the term **characteristic value** instead of eigenvalue.*

The comments above show that  $T$  has a 1-dimensional invariant subspace if and only if  $T$  has an eigenvalue.

In the definition above, we require that  $v \neq 0$  because every scalar  $\lambda \in \mathbf{F}$  satisfies  $T0 = \lambda 0$ .

### 5.6 Equivalent conditions to be an eigenvalue

Suppose  $V$  is finite-dimensional,  $T \in \mathcal{L}(V)$ , and  $\lambda \in F$ . Then the following are equivalent:

- (a)  $\lambda$  is an eigenvalue of  $T$ ;
- (b)  $T - \lambda I$  is not injective;
- (c)  $T - \lambda I$  is not surjective;
- (d)  $T - \lambda I$  is not invertible.

*Recall that  $I \in \mathcal{L}(V)$  is the identity operator defined by  $Iv = v$  for all  $v \in V$ .*

**Proof** Conditions (a) and (b) are equivalent because the equation  $Tv = \lambda v$  is equivalent to the equation  $(T - \lambda I)v = 0$ . Conditions (b), (c), and (d) are equivalent by 3.69. ■

### 5.7 Definition eigenvector

Suppose  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$  is an eigenvalue of  $T$ . A vector  $v \in V$  is called an *eigenvector* of  $T$  corresponding to  $\lambda$  if  $v \neq 0$  and  $Tv = \lambda v$ .

Because  $Tv = \lambda v$  if and only if  $(T - \lambda I)v = 0$ , a vector  $v \in V$  with  $v \neq 0$  is an eigenvector of  $T$  corresponding to  $\lambda$  if and only if  $v \in \text{null}(T - \lambda I)$ .

---

**5.8 Example** Suppose  $T \in \mathcal{L}(\mathbf{F}^2)$  is defined by

$$T(w, z) = (-z, w).$$

- (a) Find the eigenvalues and eigenvectors of  $T$  if  $\mathbf{F} = \mathbf{R}$ .
- (b) Find the eigenvalues and eigenvectors of  $T$  if  $\mathbf{F} = \mathbf{C}$ .

**Solution**

- (a) If  $\mathbf{F} = \mathbf{R}$ , then  $T$  is a counterclockwise rotation by  $90^\circ$  about the origin in  $\mathbf{R}^2$ . An operator has an eigenvalue if and only if there exists a nonzero vector in its domain that gets sent by the operator to a scalar multiple of itself. A  $90^\circ$  counterclockwise rotation of a nonzero vector in  $\mathbf{R}^2$  obviously never equals a scalar multiple of itself. Conclusion: if  $\mathbf{F} = \mathbf{R}$ , then  $T$  has no eigenvalues (and thus has no eigenvectors).

- (b) To find eigenvalues of  $T$ , we must find the scalars  $\lambda$  such that

$$T(w, z) = \lambda(w, z)$$

has some solution other than  $w = z = 0$ . The equation above is equivalent to the simultaneous equations

$$5.9 \quad -z = \lambda w, \quad w = \lambda z.$$

Substituting the value for  $w$  given by the second equation into the first equation gives

$$-z = \lambda^2 z.$$

Now  $z$  cannot equal 0 [otherwise 5.9 implies that  $w = 0$ ; we are looking for solutions to 5.9 where  $(w, z)$  is not the 0 vector], so the equation above leads to the equation

$$-1 = \lambda^2.$$

The solutions to this equation are  $\lambda = i$  and  $\lambda = -i$ . You should be able to verify easily that  $i$  and  $-i$  are eigenvalues of  $T$ . Indeed, the eigenvectors corresponding to the eigenvalue  $i$  are the vectors of the form  $(w, -wi)$ , with  $w \in \mathbf{C}$  and  $w \neq 0$ , and the eigenvectors corresponding to the eigenvalue  $-i$  are the vectors of the form  $(w, wi)$ , with  $w \in \mathbf{C}$  and  $w \neq 0$ .

---

Now we show that eigenvectors corresponding to distinct eigenvalues are linearly independent.

### 5.10 Linearly independent eigenvectors

Let  $T \in \mathcal{L}(V)$ . Suppose  $\lambda_1, \dots, \lambda_m$  are distinct eigenvalues of  $T$  and  $v_1, \dots, v_m$  are corresponding eigenvectors. Then  $v_1, \dots, v_m$  is linearly independent.

**Proof** Suppose  $v_1, \dots, v_m$  is linearly dependent. Let  $k$  be the smallest positive integer such that

$$5.11 \quad v_k \in \text{span}(v_1, \dots, v_{k-1});$$

the existence of  $k$  with this property follows from the Linear Dependence Lemma (2.21). Thus there exist  $a_1, \dots, a_{k-1} \in \mathbf{F}$  such that

$$5.12 \quad v_k = a_1 v_1 + \cdots + a_{k-1} v_{k-1}.$$

Apply  $T$  to both sides of this equation, getting

$$\lambda_k v_k = a_1 \lambda_1 v_1 + \cdots + a_{k-1} \lambda_{k-1} v_{k-1}.$$

Multiply both sides of 5.12 by  $\lambda_k$  and then subtract the equation above, getting

$$0 = a_1(\lambda_k - \lambda_1)v_1 + \cdots + a_{k-1}(\lambda_k - \lambda_{k-1})v_{k-1}.$$

Because we chose  $k$  to be the smallest positive integer satisfying 5.11,  $v_1, \dots, v_{k-1}$  is linearly independent. Thus the equation above implies that all the  $a$ 's are 0 (recall that  $\lambda_k$  is not equal to any of  $\lambda_1, \dots, \lambda_{k-1}$ ). However, this means that  $v_k$  equals 0 (see 5.12), contradicting our hypothesis that  $v_k$  is an eigenvector. Therefore our assumption that  $v_1, \dots, v_m$  is linearly dependent was false. ■

The corollary below states that an operator cannot have more distinct eigenvalues than the dimension of the vector space on which it acts.

### 5.13 Number of eigenvalues

Suppose  $V$  is finite-dimensional. Then each operator on  $V$  has at most  $\dim V$  distinct eigenvalues.

**Proof** Let  $T \in \mathcal{L}(V)$ . Suppose  $\lambda_1, \dots, \lambda_m$  are distinct eigenvalues of  $T$ . Let  $v_1, \dots, v_m$  be corresponding eigenvectors. Then 5.10 implies that the list  $v_1, \dots, v_m$  is linearly independent. Thus  $m \leq \dim V$  (see 2.23), as desired. ■

## Restriction and Quotient Operators

If  $T \in \mathcal{L}(V)$  and  $U$  is a subspace of  $V$  invariant under  $T$ , then  $U$  determines two other operators  $T|_U \in \mathcal{L}(U)$  and  $T/U \in \mathcal{L}(V/U)$  in a natural way, as defined below.

### 5.14 Definition $T|_U$ and $T/U$

Suppose  $T \in \mathcal{L}(V)$  and  $U$  is a subspace of  $V$  invariant under  $T$ .

- The **restriction operator**  $T|_U \in \mathcal{L}(U)$  is defined by

$$T|_U(u) = Tu$$

for  $u \in U$ .

- The **quotient operator**  $T/U \in \mathcal{L}(V/U)$  is defined by

$$(T/U)(v + U) = Tv + U$$

for  $v \in V$ .

For both the operators defined above, it is worthwhile to pay attention to their domains and to spend a moment thinking about why they are well defined as operators on their domains. First consider the restriction operator  $T|_U \in \mathcal{L}(U)$ , which is  $T$  with its domain restricted to  $U$ , thought of as mapping into  $U$  instead of into  $V$ . The condition that  $U$  is invariant under  $T$  is what allows us to think of  $T|_U$  as an operator on  $U$ , meaning a linear map into the same space as the domain, rather than as simply a linear map from one vector space to another vector space.

To show that the definition above of the quotient operator makes sense, we need to verify that if  $v + U = w + U$ , then  $Tv + U = Tw + U$ . Hence suppose  $v + U = w + U$ . Thus  $v - w \in U$  (see 3.85). Because  $U$  is invariant under  $T$ , we also have  $T(v - w) \in U$ , which implies that  $Tv - Tw \in U$ , which implies that  $Tv + U = Tw + U$ , as desired.

Suppose  $T$  is an operator on a finite-dimensional vector space  $V$  and  $U$  is a subspace of  $V$  invariant under  $T$ , with  $U \neq \{0\}$  and  $U \neq V$ . In some sense, we can learn about  $T$  by studying the operators  $T|_U$  and  $T/U$ , each of which is an operator on a vector space with smaller dimension than  $V$ . For example, proof 2 of 5.27 makes nice use of  $T/U$ .

However, sometimes  $T|_U$  and  $T/U$  do not provide enough information about  $T$ . In the next example, both  $T|_U$  and  $T/U$  are 0 even though  $T$  is not the 0 operator.

**5.15 Example** Define an operator  $T \in \mathcal{L}(\mathbf{F}^2)$  by  $T(x, y) = (y, 0)$ . Let  $U = \{(x, 0) : x \in \mathbf{F}\}$ . Show that

- (a)  $U$  is invariant under  $T$  and  $T|_U$  is the 0 operator on  $U$ ;
- (b) there does not exist a subspace  $W$  of  $\mathbf{F}^2$  that is invariant under  $T$  and such that  $\mathbf{F}^2 = U \oplus W$ ;
- (c)  $T/U$  is the 0 operator on  $\mathbf{F}^2/U$ .

### Solution

- (a) For  $(x, 0) \in U$ , we have  $T(x, 0) = (0, 0) \in U$ . Thus  $U$  is invariant under  $T$  and  $T|_U$  is the 0 operator on  $U$ .
- (b) Suppose  $W$  is a subspace of  $V$  such that  $\mathbf{F}^2 = U \oplus W$ . Because  $\dim \mathbf{F}^2 = 2$  and  $\dim U = 1$ , we have  $\dim W = 1$ . If  $W$  were invariant under  $T$ , then each nonzero vector in  $W$  would be an eigenvector of  $T$ . However, it is easy to see that 0 is the only eigenvalue of  $T$  and that all eigenvectors of  $T$  are in  $U$ . Thus  $W$  is not invariant under  $T$ .
- (c) For  $(x, y) \in \mathbf{F}^2$ , we have

$$\begin{aligned}(T/U)((x, y) + U) &= T(x, y) + U \\ &= (y, 0) + U \\ &= 0 + U,\end{aligned}$$

where the last equality holds because  $(y, 0) \in U$ . The equation above shows that  $T/U$  is the 0 operator.

## EXERCISES 5.A

- 1 Suppose  $T \in \mathcal{L}(V)$  and  $U$  is a subspace of  $V$ .
  - (a) Prove that if  $U \subset \text{null } T$ , then  $U$  is invariant under  $T$ .
  - (b) Prove that if  $\text{range } T \subset U$ , then  $U$  is invariant under  $T$ .
- 2 Suppose  $S, T \in \mathcal{L}(V)$  are such that  $ST = TS$ . Prove that  $\text{null } S$  is invariant under  $T$ .

- 3** Suppose  $S, T \in \mathcal{L}(V)$  are such that  $ST = TS$ . Prove that  $\text{range } S$  is invariant under  $T$ .
- 4** Suppose that  $T \in \mathcal{L}(V)$  and  $U_1, \dots, U_m$  are subspaces of  $V$  invariant under  $T$ . Prove that  $U_1 + \dots + U_m$  is invariant under  $T$ .
- 5** Suppose  $T \in \mathcal{L}(V)$ . Prove that the intersection of every collection of subspaces of  $V$  invariant under  $T$  is invariant under  $T$ .
- 6** Prove or give a counterexample: if  $V$  is finite-dimensional and  $U$  is a subspace of  $V$  that is invariant under every operator on  $V$ , then  $U = \{0\}$  or  $U = V$ .
- 7** Suppose  $T \in \mathcal{L}(\mathbf{R}^2)$  is defined by  $T(x, y) = (-3y, x)$ . Find the eigenvalues of  $T$ .

- 8** Define  $T \in \mathcal{L}(\mathbf{F}^2)$  by

$$T(w, z) = (z, w).$$

Find all eigenvalues and eigenvectors of  $T$ .

- 9** Define  $T \in \mathcal{L}(\mathbf{F}^3)$  by

$$T(z_1, z_2, z_3) = (2z_2, 0, 5z_3).$$

Find all eigenvalues and eigenvectors of  $T$ .

- 10** Define  $T \in \mathcal{L}(\mathbf{F}^n)$  by

$$T(x_1, x_2, x_3, \dots, x_n) = (x_1, 2x_2, 3x_3, \dots, nx_n).$$

- (a) Find all eigenvalues and eigenvectors of  $T$ .  
 (b) Find all invariant subspaces of  $T$ .

- 11** Define  $T: \mathcal{P}(\mathbf{R}) \rightarrow \mathcal{P}(\mathbf{R})$  by  $Tp = p'$ . Find all eigenvalues and eigenvectors of  $T$ .

- 12** Define  $T \in \mathcal{L}(\mathcal{P}_4(\mathbf{R}))$  by

$$(Tp)(x) = xp'(x)$$

for all  $x \in \mathbf{R}$ . Find all eigenvalues and eigenvectors of  $T$ .

- 13** Suppose  $V$  is finite-dimensional,  $T \in \mathcal{L}(V)$ , and  $\lambda \in \mathbf{F}$ . Prove that there exists  $\alpha \in \mathbf{F}$  such that  $|\alpha - \lambda| < \frac{1}{1000}$  and  $T - \alpha I$  is invertible.

- 14** Suppose  $V = U \oplus W$ , where  $U$  and  $W$  are nonzero subspaces of  $V$ . Define  $P \in \mathcal{L}(V)$  by  $P(u + w) = u$  for  $u \in U$  and  $w \in W$ . Find all eigenvalues and eigenvectors of  $P$ .
- 15** Suppose  $T \in \mathcal{L}(V)$ . Suppose  $S \in \mathcal{L}(V)$  is invertible.
- Prove that  $T$  and  $S^{-1}TS$  have the same eigenvalues.
  - What is the relationship between the eigenvectors of  $T$  and the eigenvectors of  $S^{-1}TS$ ?
- 16** Suppose  $V$  is a complex vector space,  $T \in \mathcal{L}(V)$ , and the matrix of  $T$  with respect to some basis of  $V$  contains only real entries. Show that if  $\lambda$  is an eigenvalue of  $T$ , then so is  $\bar{\lambda}$ .
- 17** Give an example of an operator  $T \in \mathcal{L}(\mathbf{R}^4)$  such that  $T$  has no (real) eigenvalues.
- 18** Show that the operator  $T \in \mathcal{L}(\mathbf{C}^\infty)$  defined by

$$T(z_1, z_2, \dots) = (0, z_1, z_2, \dots)$$

has no eigenvalues.

- 19** Suppose  $n$  is a positive integer and  $T \in \mathcal{L}(\mathbf{F}^n)$  is defined by

$$T(x_1, \dots, x_n) = (x_1 + \dots + x_n, \dots, x_1 + \dots + x_n);$$

in other words,  $T$  is the operator whose matrix (with respect to the standard basis) consists of all 1's. Find all eigenvalues and eigenvectors of  $T$ .

- 20** Find all eigenvalues and eigenvectors of the backward shift operator  $T \in \mathcal{L}(\mathbf{F}^\infty)$  defined by

$$T(z_1, z_2, z_3, \dots) = (z_2, z_3, \dots).$$

- 21** Suppose  $T \in \mathcal{L}(V)$  is invertible.

- Suppose  $\lambda \in \mathbf{F}$  with  $\lambda \neq 0$ . Prove that  $\lambda$  is an eigenvalue of  $T$  if and only if  $\frac{1}{\lambda}$  is an eigenvalue of  $T^{-1}$ .
- Prove that  $T$  and  $T^{-1}$  have the same eigenvectors.

- 22** Suppose  $T \in \mathcal{L}(V)$  and there exist nonzero vectors  $v$  and  $w$  in  $V$  such that

$$Tv = 3w \quad \text{and} \quad Tw = 3v.$$

Prove that 3 or  $-3$  is an eigenvalue of  $T$ .

- 23** Suppose  $V$  is finite-dimensional and  $S, T \in \mathcal{L}(V)$ . Prove that  $ST$  and  $TS$  have the same eigenvalues.

- 24** Suppose  $A$  is an  $n$ -by- $n$  matrix with entries in  $\mathbf{F}$ . Define  $T \in \mathcal{L}(\mathbf{F}^n)$  by  $Tx = Ax$ , where elements of  $\mathbf{F}^n$  are thought of as  $n$ -by-1 column vectors.

- (a) Suppose the sum of the entries in each row of  $A$  equals 1. Prove that 1 is an eigenvalue of  $T$ .
- (b) Suppose the sum of the entries in each column of  $A$  equals 1. Prove that 1 is an eigenvalue of  $T$ .

- 25** Suppose  $T \in \mathcal{L}(V)$  and  $u, v$  are eigenvectors of  $T$  such that  $u + v$  is also an eigenvector of  $T$ . Prove that  $u$  and  $v$  are eigenvectors of  $T$  corresponding to the same eigenvalue.

- 26** Suppose  $T \in \mathcal{L}(V)$  is such that every nonzero vector in  $V$  is an eigenvector of  $T$ . Prove that  $T$  is a scalar multiple of the identity operator.

- 27** Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V)$  is such that every subspace of  $V$  with dimension  $\dim V - 1$  is invariant under  $T$ . Prove that  $T$  is a scalar multiple of the identity operator.

- 28** Suppose  $V$  is finite-dimensional with  $\dim V \geq 3$  and  $T \in \mathcal{L}(V)$  is such that every 2-dimensional subspace of  $V$  is invariant under  $T$ . Prove that  $T$  is a scalar multiple of the identity operator.

- 29** Suppose  $T \in \mathcal{L}(V)$  and  $\dim \text{range } T = k$ . Prove that  $T$  has at most  $k + 1$  distinct eigenvalues.

- 30** Suppose  $T \in \mathcal{L}(\mathbf{R}^3)$  and  $-4, 5$ , and  $\sqrt{7}$  are eigenvalues of  $T$ . Prove that there exists  $x \in \mathbf{R}^3$  such that  $Tx - 9x = (-4, 5, \sqrt{7})$ .

- 31** Suppose  $V$  is finite-dimensional and  $v_1, \dots, v_m$  is a list of vectors in  $V$ . Prove that  $v_1, \dots, v_m$  is linearly independent if and only if there exists  $T \in \mathcal{L}(V)$  such that  $v_1, \dots, v_m$  are eigenvectors of  $T$  corresponding to distinct eigenvalues.

- 32** Suppose  $\lambda_1, \dots, \lambda_n$  is a list of distinct real numbers. Prove that the list  $e^{\lambda_1 x}, \dots, e^{\lambda_n x}$  is linearly independent in the vector space of real-valued functions on  $\mathbf{R}$ .

*Hint:* Let  $V = \text{span}(e^{\lambda_1 x}, \dots, e^{\lambda_n x})$ , and define an operator  $T \in \mathcal{L}(V)$  by  $Tf = f'$ . Find eigenvalues and eigenvectors of  $T$ .

- 33** Suppose  $T \in \mathcal{L}(V)$ . Prove that  $T/(\text{range } T) = 0$ .
- 34** Suppose  $T \in \mathcal{L}(V)$ . Prove that  $T/(\text{null } T)$  is injective if and only if  $(\text{null } T) \cap (\text{range } T) = \{0\}$ .
- 35** Suppose  $V$  is finite-dimensional,  $T \in \mathcal{L}(V)$ , and  $U$  is invariant under  $T$ . Prove that each eigenvalue of  $T/U$  is an eigenvalue of  $T$ .  
[The exercise below asks you to verify that the hypothesis that  $V$  is finite-dimensional is needed for the exercise above.]
- 36** Give an example of a vector space  $V$ , an operator  $T \in \mathcal{L}(V)$ , and a subspace  $U$  of  $V$  that is invariant under  $T$  such that  $T/U$  has an eigenvalue that is not an eigenvalue of  $T$ .

## 5.B Eigenvectors and Upper-Triangular Matrices

### Polynomials Applied to Operators

The main reason that a richer theory exists for operators (which map a vector space into itself) than for more general linear maps is that operators can be raised to powers. We begin this section by defining that notion and the key concept of applying a polynomial to an operator.

If  $T \in \mathcal{L}(V)$ , then  $TT$  makes sense and is also in  $\mathcal{L}(V)$ . We usually write  $T^2$  instead of  $TT$ . More generally, we have the following definition.

#### 5.16 Definition $T^m$

Suppose  $T \in \mathcal{L}(V)$  and  $m$  is a positive integer.

- $T^m$  is defined by

$$T^m = \underbrace{T \cdots T}_{m \text{ times}}.$$

- $T^0$  is defined to be the identity operator  $I$  on  $V$ .
- If  $T$  is invertible with inverse  $T^{-1}$ , then  $T^{-m}$  is defined by

$$T^{-m} = (T^{-1})^m.$$

You should verify that if  $T$  is an operator, then

$$T^m T^n = T^{m+n} \quad \text{and} \quad (T^m)^n = T^{mn},$$

where  $m$  and  $n$  are allowed to be arbitrary integers if  $T$  is invertible and nonnegative integers if  $T$  is not invertible.

#### 5.17 Definition $p(T)$

Suppose  $T \in \mathcal{L}(V)$  and  $p \in \mathcal{P}(\mathbf{F})$  is a polynomial given by

$$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_m z^m$$

for  $z \in \mathbf{F}$ . Then  $p(T)$  is the operator defined by

$$p(T) = a_0 I + a_1 T + a_2 T^2 + \cdots + a_m T^m.$$

This is a new use of the symbol  $p$  because we are applying it to operators, not just elements of  $\mathbf{F}$ .

**5.18 Example** Suppose  $D \in \mathcal{L}(\mathcal{P}(\mathbf{R}))$  is the differentiation operator defined by  $Dq = q'$  and  $p$  is the polynomial defined by  $p(x) = 7 - 3x + 5x^2$ . Then  $p(D) = 7I - 3D + 5D^2$ ; thus

$$(p(D))q = 7q - 3q' + 5q''$$

for every  $q \in \mathcal{P}(\mathbf{R})$ .

If we fix an operator  $T \in \mathcal{L}(V)$ , then the function from  $\mathcal{P}(\mathbf{F})$  to  $\mathcal{L}(V)$  given by  $p \mapsto p(T)$  is linear, as you should verify.

### 5.19 Definition *product of polynomials*

If  $p, q \in \mathcal{P}(\mathbf{F})$ , then  $pq \in \mathcal{P}(\mathbf{F})$  is the polynomial defined by

$$(pq)(z) = p(z)q(z)$$

for  $z \in \mathbf{F}$ .

Any two polynomials of an operator commute, as shown below.

### 5.20 Multiplicative properties

Suppose  $p, q \in \mathcal{P}(\mathbf{F})$  and  $T \in \mathcal{L}(V)$ . Then

- (a)  $(pq)(T) = p(T)q(T)$ ;
- (b)  $p(T)q(T) = q(T)p(T)$ .

Part (a) holds because when expanding a product of polynomials using the distributive property, it does not matter whether the symbol is  $z$  or  $T$ .

#### Proof

- (a) Suppose  $p(z) = \sum_{j=0}^m a_j z^j$  and  $q(z) = \sum_{k=0}^n b_k z^k$  for  $z \in \mathbf{F}$ . Then

$$(pq)(z) = \sum_{j=0}^m \sum_{k=0}^n a_j b_k z^{j+k}.$$

Thus

$$\begin{aligned} (pq)(T) &= \sum_{j=0}^m \sum_{k=0}^n a_j b_k T^{j+k} \\ &= \left( \sum_{j=0}^m a_j T^j \right) \left( \sum_{k=0}^n b_k T^k \right) \\ &= p(T)q(T). \end{aligned}$$

- (b) Part (a) implies  $p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T)$ . ■

## Existence of Eigenvalues

Now we come to one of the central results about operators on complex vector spaces.

### 5.21 Operators on complex vector spaces have an eigenvalue

Every operator on a finite-dimensional, nonzero, complex vector space has an eigenvalue.

**Proof** Suppose  $V$  is a complex vector space with dimension  $n > 0$  and  $T \in \mathcal{L}(V)$ . Choose  $v \in V$  with  $v \neq 0$ . Then

$$v, T v, T^2 v, \dots, T^n v$$

is not linearly independent, because  $V$  has dimension  $n$  and we have  $n + 1$  vectors. Thus there exist complex numbers  $a_0, \dots, a_n$ , not all 0, such that

$$0 = a_0 v + a_1 T v + \cdots + a_n T^n v.$$

Note that  $a_1, \dots, a_n$  cannot all be 0, because otherwise the equation above would become  $0 = a_0 v$ , which would force  $a_0$  also to be 0.

Make the  $a$ 's the coefficients of a polynomial, which by the Fundamental Theorem of Algebra (4.14) has a factorization

$$a_0 + a_1 z + \cdots + a_n z^n = c(z - \lambda_1) \cdots (z - \lambda_m),$$

where  $c$  is a nonzero complex number, each  $\lambda_j$  is in  $\mathbf{C}$ , and the equation holds for all  $z \in \mathbf{C}$  (here  $m$  is not necessarily equal to  $n$ , because  $a_n$  may equal 0). We then have

$$\begin{aligned} 0 &= a_0 v + a_1 T v + \cdots + a_n T^n v \\ &= (a_0 I + a_1 T + \cdots + a_n T^n) v \\ &= c(T - \lambda_1 I) \cdots (T - \lambda_m I) v. \end{aligned}$$

Thus  $T - \lambda_j I$  is not injective for at least one  $j$ . In other words,  $T$  has an eigenvalue. ■

The proof above depends on the Fundamental Theorem of Algebra, which is typical of proofs of this result. See Exercises 16 and 17 for possible ways to rewrite the proof above using the idea of the proof in a slightly different form.

## Upper-Triangular Matrices

In Chapter 3 we discussed the matrix of a linear map from one vector space to another vector space. That matrix depended on a choice of a basis of each of the two vector spaces. Now that we are studying operators, which map a vector space to itself, the emphasis is on using only one basis.

### 5.22 Definition *matrix of an operator*, $\mathcal{M}(T)$

Suppose  $T \in \mathcal{L}(V)$  and  $v_1, \dots, v_n$  is a basis of  $V$ . The *matrix of  $T$*  with respect to this basis is the  $n$ -by- $n$  matrix

$$\mathcal{M}(T) = \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{n,1} & \dots & A_{n,n} \end{pmatrix}$$

whose entries  $A_{j,k}$  are defined by

$$Tv_k = A_{1,k}v_1 + \dots + A_{n,k}v_n.$$

If the basis is not clear from the context, then the notation  $\mathcal{M}(T, (v_1, \dots, v_n))$  is used.

Note that the matrices of operators are square arrays, rather than the more general rectangular arrays that we considered earlier for linear maps.

*The  $k^{\text{th}}$  column of the matrix  $\mathcal{M}(T)$  is formed from the coefficients used to write  $Tv_k$  as a linear combination of  $v_1, \dots, v_n$ .*

If  $T$  is an operator on  $\mathbf{F}^n$  and no basis is specified, assume that the basis in question is the standard one (where the  $j^{\text{th}}$  basis vector is 1 in the  $j^{\text{th}}$  slot and 0 in all the other slots). You can

then think of the  $j^{\text{th}}$  column of  $\mathcal{M}(T)$  as  $T$  applied to the  $j^{\text{th}}$  basis vector.

**5.23 Example** Define  $T \in \mathcal{L}(\mathbf{F}^3)$  by  $T(x, y, z) = (2x+y, 5y+3z, 8z)$ . Then

$$\mathcal{M}(T) = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 5 & 3 \\ 0 & 0 & 8 \end{pmatrix}.$$

A central goal of linear algebra is to show that given an operator  $T \in \mathcal{L}(V)$ , there exists a basis of  $V$  with respect to which  $T$  has a reasonably simple matrix. To make this vague formulation a bit more precise, we might try to choose a basis of  $V$  such that  $\mathcal{M}(T)$  has many 0's.

If  $V$  is a finite-dimensional complex vector space, then we already know enough to show that there is a basis of  $V$  with respect to which the matrix of  $T$  has 0's everywhere in the first column, except possibly the first entry. In other words, there is a basis of  $V$  with respect to which the matrix of  $T$  looks like

$$\begin{pmatrix} \lambda & & \\ 0 & * & \\ \vdots & & \\ 0 & & \end{pmatrix};$$

here the  $*$  denotes the entries in all the columns other than the first column. To prove this, let  $\lambda$  be an eigenvalue of  $T$  (one exists by 5.21) and let  $v$  be a corresponding eigenvector. Extend  $v$  to a basis of  $V$ . Then the matrix of  $T$  with respect to this basis has the form above.

Soon we will see that we can choose a basis of  $V$  with respect to which the matrix of  $T$  has even more 0's.

### 5.24 Definition *diagonal of a matrix*

The ***diagonal*** of a square matrix consists of the entries along the line from the upper left corner to the bottom right corner.

For example, the diagonal of the matrix in 5.23 consists of the entries 2, 5, 8.

### 5.25 Definition *upper-triangular matrix*

A matrix is called ***upper triangular*** if all the entries below the diagonal equal 0.

For example, the matrix in 5.23 is upper triangular.

Typically we represent an upper-triangular matrix in the form

$$\begin{pmatrix} \lambda_1 & * & \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix};$$

the 0 in the matrix above indicates that all entries below the diagonal in this  $n$ -by- $n$  matrix equal 0. Upper-triangular matrices can be considered reasonably simple—for  $n$  large, almost half its entries in an  $n$ -by- $n$  upper-triangular matrix are 0.

We often use  $*$  to denote matrix entries that we do not know about or that are irrelevant to the questions being discussed.

The following proposition demonstrates a useful connection between upper-triangular matrices and invariant subspaces.

### 5.26 Conditions for upper-triangular matrix

Suppose  $T \in \mathcal{L}(V)$  and  $v_1, \dots, v_n$  is a basis of  $V$ . Then the following are equivalent:

- (a) the matrix of  $T$  with respect to  $v_1, \dots, v_n$  is upper triangular;
- (b)  $Tv_j \in \text{span}(v_1, \dots, v_j)$  for each  $j = 1, \dots, n$ ;
- (c)  $\text{span}(v_1, \dots, v_j)$  is invariant under  $T$  for each  $j = 1, \dots, n$ .

**Proof** The equivalence of (a) and (b) follows easily from the definitions and a moment's thought. Obviously (c) implies (b). Hence to complete the proof, we need only prove that (b) implies (c).

Thus suppose (b) holds. Fix  $j \in \{1, \dots, n\}$ . From (b), we know that

$$\begin{aligned}Tv_1 &\in \text{span}(v_1) \subset \text{span}(v_1, \dots, v_j); \\Tv_2 &\in \text{span}(v_1, v_2) \subset \text{span}(v_1, \dots, v_j); \\&\vdots \\Tv_j &\in \text{span}(v_1, \dots, v_j).\end{aligned}$$

Thus if  $v$  is a linear combination of  $v_1, \dots, v_j$ , then

$$Tv \in \text{span}(v_1, \dots, v_j).$$

In other words,  $\text{span}(v_1, \dots, v_j)$  is invariant under  $T$ , completing the proof. ■

*The next result does not hold on real vector spaces, because the first vector in a basis with respect to which an operator has an upper-triangular matrix is an eigenvector of the operator. Thus if an operator on a real vector space has no eigenvalues [see 5.8(a) for an example], then there is no basis with respect to which the operator has an upper-triangular matrix.*

Now we can prove that for each operator on a finite-dimensional complex vector space, there is a basis of the vector space with respect to which the matrix of the operator has only 0's below the diagonal. In Chapter 8 we will improve even this result.

Sometimes more insight comes from seeing more than one proof of a theorem. Thus two proofs are presented of the next result. Use whichever appeals more to you.

### 5.27 Over $\mathbb{C}$ , every operator has an upper-triangular matrix

Suppose  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ . Then  $T$  has an upper-triangular matrix with respect to some basis of  $V$ .

**Proof 1** We will use induction on the dimension of  $V$ . Clearly the desired result holds if  $\dim V = 1$ .

Suppose now that  $\dim V > 1$  and the desired result holds for all complex vector spaces whose dimension is less than the dimension of  $V$ . Let  $\lambda$  be any eigenvalue of  $T$  (5.21 guarantees that  $T$  has an eigenvalue). Let

$$U = \text{range}(T - \lambda I).$$

Because  $T - \lambda I$  is not surjective (see 3.69),  $\dim U < \dim V$ . Furthermore,  $U$  is invariant under  $T$ . To prove this, suppose  $u \in U$ . Then

$$Tu = (T - \lambda I)u + \lambda u.$$

Obviously  $(T - \lambda I)u \in U$  (because  $U$  equals the range of  $T - \lambda I$ ) and  $\lambda u \in U$ . Thus the equation above shows that  $Tu \in U$ . Hence  $U$  is invariant under  $T$ , as claimed.

Thus  $T|_U$  is an operator on  $U$ . By our induction hypothesis, there is a basis  $u_1, \dots, u_m$  of  $U$  with respect to which  $T|_U$  has an upper-triangular matrix. Thus for each  $j$  we have (using 5.26)

$$\mathbf{5.28} \quad Tu_j = (T|_U)(u_j) \in \text{span}(u_1, \dots, u_j).$$

Extend  $u_1, \dots, u_m$  to a basis  $u_1, \dots, u_m, v_1, \dots, v_n$  of  $V$ . For each  $k$ , we have

$$Tv_k = (T - \lambda I)v_k + \lambda v_k.$$

The definition of  $U$  shows that  $(T - \lambda I)v_k \in U = \text{span}(u_1, \dots, u_m)$ . Thus the equation above shows that

$$\mathbf{5.29} \quad Tv_k \in \text{span}(u_1, \dots, u_m, v_1, \dots, v_k).$$

From 5.28 and 5.29, we conclude (using 5.26) that  $T$  has an upper-triangular matrix with respect to the basis  $u_1, \dots, u_m, v_1, \dots, v_n$  of  $V$ , as desired. ■

**Proof 2** We will use induction on the dimension of  $V$ . Clearly the desired result holds if  $\dim V = 1$ .

Suppose now that  $\dim V = n > 1$  and the desired result holds for all complex vector spaces whose dimension is  $n - 1$ . Let  $v_1$  be any eigenvector of  $T$  (5.21 guarantees that  $T$  has an eigenvector). Let  $U = \text{span}(v_1)$ . Then  $U$  is an invariant subspace of  $T$  and  $\dim U = 1$ .

Because  $\dim V/U = n - 1$  (see 3.89), we can apply our induction hypothesis to  $T/U \in \mathcal{L}(V/U)$ . Thus there is a basis  $v_2 + U, \dots, v_n + U$  of  $V/U$  such that  $T/U$  has an upper-triangular matrix with respect to this basis. Hence by 5.26,

$$(T/U)(v_j + U) \in \text{span}(v_2 + U, \dots, v_j + U)$$

for each  $j = 2, \dots, n$ . Unraveling the meaning of the inclusion above, we see that

$$Tv_j \in \text{span}(v_1, \dots, v_j)$$

for each  $j = 1, \dots, n$ . Thus by 5.26,  $T$  has an upper-triangular matrix with respect to the basis  $v_1, \dots, v_n$  of  $V$ , as desired (it is easy to verify that  $v_1, \dots, v_n$  is a basis of  $V$ ; see Exercise 13 in Section 3.E for a more general result). ■

How does one determine from looking at the matrix of an operator whether the operator is invertible? If we are fortunate enough to have a basis with respect to which the matrix of the operator is upper triangular, then this problem becomes easy, as the following proposition shows.

### 5.30 Determination of invertibility from upper-triangular matrix

Suppose  $T \in \mathcal{L}(V)$  has an upper-triangular matrix with respect to some basis of  $V$ . Then  $T$  is invertible if and only if all the entries on the diagonal of that upper-triangular matrix are nonzero.

**Proof** Suppose  $v_1, \dots, v_n$  is a basis of  $V$  with respect to which  $T$  has an upper-triangular matrix

$$5.31 \quad \mathcal{M}(T) = \begin{pmatrix} \lambda_1 & & * \\ & \lambda_2 & \\ 0 & & \ddots & \lambda_n \end{pmatrix}.$$

We need to prove that  $T$  is invertible if and only if all the  $\lambda_j$ 's are nonzero.

First suppose the diagonal entries  $\lambda_1, \dots, \lambda_n$  are all nonzero. The upper-triangular matrix in 5.31 implies that  $Tv_1 = \lambda_1 v_1$ . Because  $\lambda_1 \neq 0$ , we have  $T(v_1/\lambda_1) = v_1$ ; thus  $v_1 \in \text{range } T$ .

Now

$$T(v_2/\lambda_2) = av_1 + v_2$$

for some  $a \in \mathbf{F}$ . The left side of the equation above and  $av_1$  are both in range  $T$ ; thus  $v_2 \in \text{range } T$ .

Similarly, we see that

$$T(v_3/\lambda_3) = bv_1 + cv_2 + v_3$$

for some  $b, c \in \mathbf{F}$ . The left side of the equation above and  $bv_1, cv_2$  are all in range  $T$ ; thus  $v_3 \in \text{range } T$ .

Continuing in this fashion, we conclude that  $v_1, \dots, v_n \in \text{range } T$ . Because  $v_1, \dots, v_n$  is a basis of  $V$ , this implies that  $\text{range } T = V$ . In other words,  $T$  is surjective. Hence  $T$  is invertible (by 3.69), as desired.

To prove the other direction, now suppose that  $T$  is invertible. This implies that  $\lambda_1 \neq 0$ , because otherwise we would have  $Tv_1 = 0$ .

Let  $1 < j \leq n$ , and suppose  $\lambda_j = 0$ . Then 5.31 implies that  $T$  maps  $\text{span}(v_1, \dots, v_j)$  into  $\text{span}(v_1, \dots, v_{j-1})$ . Because

$$\dim \text{span}(v_1, \dots, v_j) = j \quad \text{and} \quad \dim \text{span}(v_1, \dots, v_{j-1}) = j - 1,$$

this implies that  $T$  restricted to  $\dim \text{span}(v_1, \dots, v_j)$  is not injective (by 3.23). Thus there exists  $v \in \text{span}(v_1, \dots, v_j)$  such that  $v \neq 0$  and  $Tv = 0$ . Thus  $T$  is not injective, which contradicts our hypothesis (for this direction) that  $T$  is invertible. This contradiction means that our assumption that  $\lambda_j = 0$  must be false. Hence  $\lambda_j \neq 0$ , as desired. ■

As an example of the result above, we see that the operator in Example 5.23 is invertible.

Unfortunately no method exists for exactly computing the eigenvalues of an operator from its matrix. However, if we are fortunate enough to find a basis with respect to which the matrix of the operator is upper triangular, then the problem of computing the eigenvalues becomes trivial, as the following proposition shows.

*Powerful numeric techniques exist for finding good approximations to the eigenvalues of an operator from its matrix.*

### 5.32 Determination of eigenvalues from upper-triangular matrix

Suppose  $T \in \mathcal{L}(V)$  has an upper-triangular matrix with respect to some basis of  $V$ . Then the eigenvalues of  $T$  are precisely the entries on the diagonal of that upper-triangular matrix.

**Proof** Suppose  $v_1, \dots, v_n$  is a basis of  $V$  with respect to which  $T$  has an upper-triangular matrix

$$\mathcal{M}(T) = \begin{pmatrix} \lambda_1 & & * \\ & \lambda_2 & \\ 0 & & \ddots & \lambda_n \end{pmatrix}.$$

Let  $\lambda \in \mathbf{F}$ . Then

$$\mathcal{M}(T - \lambda I) = \begin{pmatrix} \lambda_1 - \lambda & & * \\ & \lambda_2 - \lambda & \\ 0 & & \ddots & \lambda_n - \lambda \end{pmatrix}.$$

Hence  $T - \lambda I$  is not invertible if and only if  $\lambda$  equals one of the numbers  $\lambda_1, \dots, \lambda_n$  (by 5.30). Thus  $\lambda$  is an eigenvalue of  $T$  if and only if  $\lambda$  equals one of the numbers  $\lambda_1, \dots, \lambda_n$ . ■

**5.33 Example** Define  $T \in \mathcal{L}(\mathbf{F}^3)$  by  $T(x, y, z) = (2x + y, 5y + 3z, 8z)$ . What are the eigenvalues of  $T$ ?

**Solution** The matrix of  $T$  with respect to the standard basis is

$$\mathcal{M}(T) = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 5 & 3 \\ 0 & 0 & 8 \end{pmatrix}.$$

Thus  $\mathcal{M}(T)$  is an upper-triangular matrix. Now 5.32 implies that the eigenvalues of  $T$  are 2, 5, and 8.

Once the eigenvalues of an operator on  $\mathbf{F}^n$  are known, the eigenvectors can be found easily using Gaussian elimination.

## EXERCISES 5.B

---

- 1 Suppose  $T \in \mathcal{L}(V)$  and there exists a positive integer  $n$  such that  $T^n = 0$ .
- Prove that  $I - T$  is invertible and that
$$(I - T)^{-1} = I + T + \cdots + T^{n-1}.$$
  - Explain how you would guess the formula above.
- 2 Suppose  $T \in \mathcal{L}(V)$  and  $(T - 2I)(T - 3I)(T - 4I) = 0$ . Suppose  $\lambda$  is an eigenvalue of  $T$ . Prove that  $\lambda = 2$  or  $\lambda = 3$  or  $\lambda = 4$ .
- 3 Suppose  $T \in \mathcal{L}(V)$  and  $T^2 = I$  and  $-1$  is not an eigenvalue of  $T$ . Prove that  $T = I$ .
- 4 Suppose  $P \in \mathcal{L}(V)$  and  $P^2 = P$ . Prove that  $V = \text{null } P \oplus \text{range } P$ .
- 5 Suppose  $S, T \in \mathcal{L}(V)$  and  $S$  is invertible. Suppose  $p \in \mathcal{P}(\mathbf{F})$  is a polynomial. Prove that
$$p(STS^{-1}) = Sp(T)S^{-1}.$$
- 6 Suppose  $T \in \mathcal{L}(V)$  and  $U$  is a subspace of  $V$  invariant under  $T$ . Prove that  $U$  is invariant under  $p(T)$  for every polynomial  $p \in \mathcal{P}(\mathbf{F})$ .
- 7 Suppose  $T \in \mathcal{L}(V)$ . Prove that  $9$  is an eigenvalue of  $T^2$  if and only if  $3$  or  $-3$  is an eigenvalue of  $T$ .
- 8 Give an example of  $T \in \mathcal{L}(\mathbf{R}^2)$  such that  $T^4 = -1$ .
- 9 Suppose  $V$  is finite-dimensional,  $T \in \mathcal{L}(V)$ , and  $v \in V$  with  $v \neq 0$ . Let  $p$  be a nonzero polynomial of smallest degree such that  $p(T)v = 0$ . Prove that every zero of  $p$  is an eigenvalue of  $T$ .
- 10 Suppose  $T \in \mathcal{L}(V)$  and  $v$  is an eigenvector of  $T$  with eigenvalue  $\lambda$ . Suppose  $p \in \mathcal{P}(\mathbf{F})$ . Prove that  $p(T)v = p(\lambda)v$ .
- 11 Suppose  $\mathbf{F} = \mathbf{C}$ ,  $T \in \mathcal{L}(V)$ ,  $p \in \mathcal{P}(\mathbf{C})$  is a polynomial, and  $\alpha \in \mathbf{C}$ . Prove that  $\alpha$  is an eigenvalue of  $p(T)$  if and only if  $\alpha = p(\lambda)$  for some eigenvalue  $\lambda$  of  $T$ .
- 12 Show that the result in the previous exercise does not hold if  $\mathbf{C}$  is replaced with  $\mathbf{R}$ .

- 13** Suppose  $W$  is a complex vector space and  $T \in \mathcal{L}(W)$  has no eigenvalues. Prove that every subspace of  $W$  invariant under  $T$  is either  $\{0\}$  or infinite-dimensional.
- 14** Give an example of an operator whose matrix with respect to some basis contains only 0's on the diagonal, but the operator is invertible.  
*[The exercise above and the exercise below show that 5.30 fails without the hypothesis that an upper-triangular matrix is under consideration.]*
- 15** Give an example of an operator whose matrix with respect to some basis contains only nonzero numbers on the diagonal, but the operator is not invertible.
- 16** Rewrite the proof of 5.21 using the linear map that sends  $p \in \mathcal{P}_n(\mathbf{C})$  to  $(p(T))v \in V$  (and use 3.23).
- 17** Rewrite the proof of 5.21 using the linear map that sends  $p \in \mathcal{P}_{n^2}(\mathbf{C})$  to  $p(T) \in \mathcal{L}(V)$  (and use 3.23).
- 18** Suppose  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ . Define a function  $f : \mathbf{C} \rightarrow \mathbf{R}$  by

$$f(\lambda) = \dim \text{range}(T - \lambda I).$$

Prove that  $f$  is not a continuous function.

- 19** Suppose  $V$  is finite-dimensional with  $\dim V > 1$  and  $T \in \mathcal{L}(V)$ . Prove that

$$\{p(T) : p \in \mathcal{P}(\mathbf{F})\} \neq \mathcal{L}(V).$$

- 20** Suppose  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ . Prove that  $T$  has an invariant subspace of dimension  $k$  for each  $k = 1, \dots, \dim V$ .

## 5.C Eigenspaces and Diagonal Matrices

### 5.34 Definition *diagonal matrix*

A ***diagonal matrix*** is a square matrix that is 0 everywhere except possibly along the diagonal.

### 5.35 Example

$$\begin{pmatrix} 8 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

is a diagonal matrix.

Obviously every diagonal matrix is upper triangular. In general, a diagonal matrix has many more 0's than an upper-triangular matrix.

If an operator has a diagonal matrix with respect to some basis, then the entries along the diagonal are precisely the eigenvalues of the operator; this follows from 5.32 (or find an easier proof for diagonal matrices).

### 5.36 Definition *eigenspace*, $E(\lambda, T)$

Suppose  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ . The ***eigenspace*** of  $T$  corresponding to  $\lambda$ , denoted  $E(\lambda, T)$ , is defined by

$$E(\lambda, T) = \text{null}(T - \lambda I).$$

In other words,  $E(\lambda, T)$  is the set of all eigenvectors of  $T$  corresponding to  $\lambda$ , along with the 0 vector.

For  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ , the eigenspace  $E(\lambda, T)$  is a subspace of  $V$  (because the null space of each linear map on  $V$  is a subspace of  $V$ ). The definitions imply that  $\lambda$  is an eigenvalue of  $T$  if and only if  $E(\lambda, T) \neq \{0\}$ .

**5.37 Example** Suppose the matrix of an operator  $T \in \mathcal{L}(V)$  with respect to a basis  $v_1, v_2, v_3$  of  $V$  is the matrix in Example 5.35 above. Then

$$E(8, T) = \text{span}(v_1), \quad E(5, T) = \text{span}(v_2, v_3).$$

If  $\lambda$  is an eigenvalue of an operator  $T \in \mathcal{L}(V)$ , then  $T$  restricted to  $E(\lambda, T)$  is just the operator of multiplication by  $\lambda$ .

### 5.38 Sum of eigenspaces is a direct sum

Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V)$ . Suppose also that  $\lambda_1, \dots, \lambda_m$  are distinct eigenvalues of  $T$ . Then

$$E(\lambda_1, T) + \cdots + E(\lambda_m, T)$$

is a direct sum. Furthermore,

$$\dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T) \leq \dim V.$$

**Proof** To show that  $E(\lambda_1, T) + \cdots + E(\lambda_m, T)$  is a direct sum, suppose

$$u_1 + \cdots + u_m = 0,$$

where each  $u_j$  is in  $E(\lambda_j, T)$ . Because eigenvectors corresponding to distinct eigenvalues are linearly independent (see 5.10), this implies that each  $u_j$  equals 0. This implies (using 1.44) that  $E(\lambda_1, T) + \cdots + E(\lambda_m, T)$  is a direct sum, as desired.

Now

$$\begin{aligned} \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T) &= \dim(E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T)) \\ &\leq \dim V, \end{aligned}$$

where the equality above follows from Exercise 16 in Section 2.C. ■

### 5.39 Definition *diagonalizable*

An operator  $T \in \mathcal{L}(V)$  is called **diagonalizable** if the operator has a diagonal matrix with respect to some basis of  $V$ .

### 5.40 Example

Define  $T \in \mathcal{L}(\mathbf{R}^2)$  by

$$T(x, y) = (41x + 7y, -20x + 74y).$$

The matrix of  $T$  with respect to the standard basis of  $\mathbf{R}^2$  is

$$\begin{pmatrix} 41 & 7 \\ -20 & 74 \end{pmatrix},$$

which is not a diagonal matrix. However,  $T$  is diagonalizable, because the matrix of  $T$  with respect to the basis  $(1, 4), (7, 5)$  is

$$\begin{pmatrix} 69 & 0 \\ 0 & 46 \end{pmatrix},$$

as you should verify.

### 5.41 Conditions equivalent to diagonalizability

Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V)$ . Let  $\lambda_1, \dots, \lambda_m$  denote the distinct eigenvalues of  $T$ . Then the following are equivalent:

- (a)  $T$  is diagonalizable;
- (b)  $V$  has a basis consisting of eigenvectors of  $T$ ;
- (c) there exist 1-dimensional subspaces  $U_1, \dots, U_n$  of  $V$ , each invariant under  $T$ , such that

$$V = U_1 \oplus \cdots \oplus U_n;$$

- (d)  $V = E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T)$ ;
- (e)  $\dim V = \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T)$ .

**Proof** An operator  $T \in \mathcal{L}(V)$  has a diagonal matrix

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

with respect to a basis  $v_1, \dots, v_n$  of  $V$  if and only if  $Tv_j = \lambda_j v_j$  for each  $j$ . Thus (a) and (b) are equivalent.

Suppose (b) holds; thus  $V$  has a basis  $v_1, \dots, v_n$  consisting of eigenvectors of  $T$ . For each  $j$ , let  $U_j = \text{span}(v_j)$ . Obviously each  $U_j$  is a 1-dimensional subspace of  $V$  that is invariant under  $T$ . Because  $v_1, \dots, v_n$  is a basis of  $V$ , each vector in  $V$  can be written uniquely as a linear combination of  $v_1, \dots, v_n$ . In other words, each vector in  $V$  can be written uniquely as a sum  $u_1 + \cdots + u_n$ , where each  $u_j$  is in  $U_j$ . Thus  $V = U_1 \oplus \cdots \oplus U_n$ . Hence (b) implies (c).

Suppose now that (c) holds; thus there are 1-dimensional subspaces  $U_1, \dots, U_n$  of  $V$ , each invariant under  $T$ , such that  $V = U_1 \oplus \cdots \oplus U_n$ . For each  $j$ , let  $v_j$  be a nonzero vector in  $U_j$ . Then each  $v_j$  is an eigenvector of  $T$ . Because each vector in  $V$  can be written uniquely as a sum  $u_1 + \cdots + u_n$ , where each  $u_j$  is in  $U_j$  (so each  $u_j$  is a scalar multiple of  $v_j$ ), we see that  $v_1, \dots, v_n$  is a basis of  $V$ . Thus (c) implies (b).

At this stage of the proof we know that (a), (b), and (c) are all equivalent. We will finish the proof by showing that (b) implies (d), that (d) implies (e), and that (e) implies (b).

Suppose (b) holds; thus  $V$  has a basis consisting of eigenvectors of  $T$ . Hence every vector in  $V$  is a linear combination of eigenvectors of  $T$ , which implies that

$$V = E(\lambda_1, T) + \cdots + E(\lambda_m, T).$$

Now 5.38 shows that (d) holds.

That (d) implies (e) follows immediately from Exercise 16 in Section 2.C.

Finally, suppose (e) holds; thus

**5.42**  $\dim V = \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T).$

Choose a basis of each  $E(\lambda_j, T)$ ; put all these bases together to form a list  $v_1, \dots, v_n$  of eigenvectors of  $T$ , where  $n = \dim V$  (by 5.42). To show that this list is linearly independent, suppose

$$a_1v_1 + \cdots + a_nv_n = 0,$$

where  $a_1, \dots, a_n \in \mathbf{F}$ . For each  $j = 1, \dots, m$ , let  $u_j$  denote the sum of all the terms  $a_k v_k$  such that  $v_k \in E(\lambda_j, T)$ . Thus each  $u_j$  is in  $E(\lambda_j, T)$ , and

$$u_1 + \cdots + u_m = 0.$$

Because eigenvectors corresponding to distinct eigenvalues are linearly independent (see 5.10), this implies that each  $u_j$  equals 0. Because each  $u_j$  is a sum of terms  $a_k v_k$ , where the  $v_k$ 's were chosen to be a basis of  $E(\lambda_j, T)$ , this implies that all the  $a_k$ 's equal 0. Thus  $v_1, \dots, v_n$  is linearly independent and hence is a basis of  $V$  (by 2.39). Thus (e) implies (b), completing the proof. ■

Unfortunately not every operator is diagonalizable. This sad state of affairs can arise even on complex vector spaces, as shown by the next example.

**5.43 Example** Show that the operator  $T \in \mathcal{L}(\mathbf{C}^2)$  defined by

$$T(w, z) = (z, 0)$$

is not diagonalizable.

**Solution** As you should verify, 0 is the only eigenvalue of  $T$  and furthermore  $E(0, T) = \{(w, 0) \in \mathbf{C}^2 : w \in \mathbf{C}\}$ .

Thus conditions (b), (c), (d), and (e) of 5.41 are easily seen to fail (of course, because these conditions are equivalent, it is only necessary to check that one of them fails). Thus condition (a) of 5.41 also fails, and hence  $T$  is not diagonalizable.

The next result shows that if an operator has as many distinct eigenvalues as the dimension of its domain, then the operator is diagonalizable.

### 5.44 Enough eigenvalues implies diagonalizability

If  $T \in \mathcal{L}(V)$  has  $\dim V$  distinct eigenvalues, then  $T$  is diagonalizable.

**Proof** Suppose  $T \in \mathcal{L}(V)$  has  $\dim V$  distinct eigenvalues  $\lambda_1, \dots, \lambda_{\dim V}$ . For each  $j$ , let  $v_j \in V$  be an eigenvector corresponding to the eigenvalue  $\lambda_j$ . Because eigenvectors corresponding to distinct eigenvalues are linearly independent (see 5.10),  $v_1, \dots, v_{\dim V}$  is linearly independent. A linearly independent list of  $\dim V$  vectors in  $V$  is a basis of  $V$  (see 2.39); thus  $v_1, \dots, v_{\dim V}$  is a basis of  $V$ . With respect to this basis consisting of eigenvectors,  $T$  has a diagonal matrix. ■

**5.45 Example** Define  $T \in \mathcal{L}(\mathbf{F}^3)$  by  $T(x, y, z) = (2x + y, 5y + 3z, 8z)$ . Find a basis of  $\mathbf{F}^3$  with respect to which  $T$  has a diagonal matrix.

**Solution** With respect to the standard basis, the matrix of  $T$  is

$$\begin{pmatrix} 2 & 1 & 0 \\ 0 & 5 & 3 \\ 0 & 0 & 8 \end{pmatrix}.$$

The matrix above is upper triangular. Thus by 5.32, the eigenvalues of  $T$  are 2, 5, and 8. Because  $T$  is an operator on a vector space with dimension 3 and  $T$  has three distinct eigenvalues, 5.44 assures us that there exists a basis of  $\mathbf{F}^3$  with respect to which  $T$  has a diagonal matrix.

To find this basis, we only have to find an eigenvector for each eigenvalue. In other words, we have to find a nonzero solution to the equation

$$T(x, y, z) = \lambda(x, y, z)$$

for  $\lambda = 2$ , then for  $\lambda = 5$ , and then for  $\lambda = 8$ . These simple equations are easy to solve: for  $\lambda = 2$  we have the eigenvector  $(1, 0, 0)$ ; for  $\lambda = 5$  we have the eigenvector  $(1, 3, 0)$ ; for  $\lambda = 8$  we have the eigenvector  $(1, 6, 6)$ .

Thus  $(1, 0, 0), (1, 3, 0), (1, 6, 6)$  is a basis of  $\mathbf{F}^3$ , and with respect to this basis the matrix of  $T$  is

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 8 \end{pmatrix}.$$

The converse of 5.44 is not true. For example, the operator  $T$  defined on the three-dimensional space  $\mathbf{F}^3$  by

$$T(z_1, z_2, z_3) = (4z_1, 4z_2, 5z_3)$$

has only two eigenvalues (4 and 5), but this operator has a diagonal matrix with respect to the standard basis.

In later chapters we will find additional conditions that imply that certain operators are diagonalizable.

## EXERCISES 5.C

---

- 1 Suppose  $T \in \mathcal{L}(V)$  is diagonalizable. Prove that  $V = \text{null } T \oplus \text{range } T$ .
- 2 Prove the converse of the statement in the exercise above or give a counterexample to the converse.
- 3 Suppose  $V$  is finite-dimensional and  $T \in \mathcal{L}(V)$ . Prove that the following are equivalent:
  - (a)  $V = \text{null } T \oplus \text{range } T$ .
  - (b)  $V = \text{null } T + \text{range } T$ .
  - (c)  $\text{null } T \cap \text{range } T = \{0\}$ .
- 4 Give an example to show that the exercise above is false without the hypothesis that  $V$  is finite-dimensional.
- 5 Suppose  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ . Prove that  $T$  is diagonalizable if and only if

$$V = \text{null}(T - \lambda I) \oplus \text{range}(T - \lambda I)$$

for every  $\lambda \in \mathbf{C}$ .

- 6 Suppose  $V$  is finite-dimensional,  $T \in \mathcal{L}(V)$  has  $\dim V$  distinct eigenvalues, and  $S \in \mathcal{L}(V)$  has the same eigenvectors as  $T$  (not necessarily with the same eigenvalues). Prove that  $ST = TS$ .
- 7 Suppose  $T \in \mathcal{L}(V)$  has a diagonal matrix  $A$  with respect to some basis of  $V$  and that  $\lambda \in \mathbf{F}$ . Prove that  $\lambda$  appears on the diagonal of  $A$  precisely  $\dim E(\lambda, T)$  times.
- 8 Suppose  $T \in \mathcal{L}(\mathbf{F}^5)$  and  $\dim E(8, T) = 4$ . Prove that  $T - 2I$  or  $T - 6I$  is invertible.

**9** Suppose  $T \in \mathcal{L}(V)$  is invertible. Prove that  $E(\lambda, T) = E(\frac{1}{\lambda}, T^{-1})$  for every  $\lambda \in \mathbf{F}$  with  $\lambda \neq 0$ .

**10** Suppose that  $V$  is finite-dimensional and  $T \in \mathcal{L}(V)$ . Let  $\lambda_1, \dots, \lambda_m$  denote the distinct nonzero eigenvalues of  $T$ . Prove that

$$\dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T) \leq \dim \text{range } T.$$

**11** Verify the assertion in Example 5.40.

**12** Suppose  $R, T \in \mathcal{L}(\mathbf{F}^3)$  each have 2, 6, 7 as eigenvalues. Prove that there exists an invertible operator  $S \in \mathcal{L}(\mathbf{F}^3)$  such that  $R = S^{-1}TS$ .

**13** Find  $R, T \in \mathcal{L}(\mathbf{F}^4)$  such that  $R$  and  $T$  each have 2, 6, 7 as eigenvalues,  $R$  and  $T$  have no other eigenvalues, and there does not exist an invertible operator  $S \in \mathcal{L}(\mathbf{F}^4)$  such that  $R = S^{-1}TS$ .

**14** Find  $T \in \mathcal{L}(\mathbf{C}^3)$  such that 6 and 7 are eigenvalues of  $T$  and such that  $T$  does not have a diagonal matrix with respect to any basis of  $\mathbf{C}^3$ .

**15** Suppose  $T \in \mathcal{L}(\mathbf{C}^3)$  is such that 6 and 7 are eigenvalues of  $T$ . Furthermore, suppose  $T$  does not have a diagonal matrix with respect to any basis of  $\mathbf{C}^3$ . Prove that there exists  $(x, y, z) \in \mathbf{F}^3$  such that  $T(x, y, z) = (17 + 8x, \sqrt{5} + 8y, 2\pi + 8z)$ .

**16** The **Fibonacci sequence**  $F_1, F_2, \dots$  is defined by

$$F_1 = 1, \quad F_2 = 1, \quad \text{and} \quad F_n = F_{n-2} + F_{n-1} \text{ for } n \geq 3.$$

Define  $T \in \mathcal{L}(\mathbf{R}^2)$  by  $T(x, y) = (y, x + y)$ .

- (a) Show that  $T^n(0, 1) = (F_n, F_{n+1})$  for each positive integer  $n$ .
- (b) Find the eigenvalues of  $T$ .
- (c) Find a basis of  $\mathbf{R}^2$  consisting of eigenvectors of  $T$ .
- (d) Use the solution to part (c) to compute  $T^n(0, 1)$ . Conclude that

$$F_n = \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^n - \left( \frac{1 - \sqrt{5}}{2} \right)^n \right]$$

for each positive integer  $n$ .

- (e) Use part (d) to conclude that for each positive integer  $n$ , the Fibonacci number  $F_n$  is the integer that is closest to

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



*Woman teaching geometry, from a fourteenth-century edition of Euclid's geometry book.*

## Inner Product Spaces

In making the definition of a vector space, we generalized the linear structure (addition and scalar multiplication) of  $\mathbf{R}^2$  and  $\mathbf{R}^3$ . We ignored other important features, such as the notions of length and angle. These ideas are embedded in the concept we now investigate, inner products.

Our standing assumptions are as follows:

### 6.1 Notation $\mathbf{F}, V$

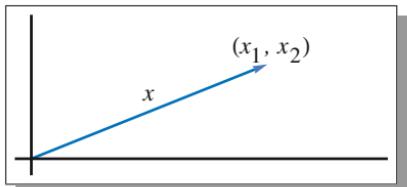
- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  denotes a vector space over  $\mathbf{F}$ .

### LEARNING OBJECTIVES FOR THIS CHAPTER

- Cauchy–Schwarz Inequality
- Gram–Schmidt Procedure
- linear functionals on inner product spaces
- calculating minimum distance to a subspace

## 6.A Inner Products and Norms

### Inner Products



The length of this vector  $x$  is  $\sqrt{x_1^2 + x_2^2}$ .

To motivate the concept of inner product, think of vectors in  $\mathbf{R}^2$  and  $\mathbf{R}^3$  as arrows with initial point at the origin. The length of a vector  $x$  in  $\mathbf{R}^2$  or  $\mathbf{R}^3$  is called the **norm** of  $x$ , denoted  $\|x\|$ . Thus for  $x = (x_1, x_2) \in \mathbf{R}^2$ , we have  $\|x\| = \sqrt{x_1^2 + x_2^2}$ .

Similarly, if  $x = (x_1, x_2, x_3) \in \mathbf{R}^3$ , then  $\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$ .

Even though we cannot draw pictures in higher dimensions, the generalization to  $\mathbf{R}^n$  is obvious: we define the norm of  $x = (x_1, \dots, x_n) \in \mathbf{R}^n$  by

$$\|x\| = \sqrt{x_1^2 + \dots + x_n^2}.$$

The norm is not linear on  $\mathbf{R}^n$ . To inject linearity into the discussion, we introduce the dot product.

### 6.2 Definition dot product

For  $x, y \in \mathbf{R}^n$ , the **dot product** of  $x$  and  $y$ , denoted  $x \cdot y$ , is defined by

$$x \cdot y = x_1 y_1 + \dots + x_n y_n,$$

where  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$ .

If we think of vectors as points instead of arrows, then  $\|x\|$  should be interpreted as the distance from the origin to the point  $x$ .

Note that the dot product of two vectors in  $\mathbf{R}^n$  is a number, not a vector. Obviously  $x \cdot x = \|x\|^2$  for all  $x \in \mathbf{R}^n$ . The dot product on  $\mathbf{R}^n$  has the following properties:

- $x \cdot x \geq 0$  for all  $x \in \mathbf{R}^n$ ;
- $x \cdot x = 0$  if and only if  $x = 0$ ;
- for  $y \in \mathbf{R}^n$  fixed, the map from  $\mathbf{R}^n$  to  $\mathbf{R}$  that sends  $x \in \mathbf{R}^n$  to  $x \cdot y$  is linear;
- $x \cdot y = y \cdot x$  for all  $x, y \in \mathbf{R}^n$ .

An inner product is a generalization of the dot product. At this point you may be tempted to guess that an inner product is defined by abstracting the properties of the dot product discussed in the last paragraph. For real vector spaces, that guess is correct. However, so that we can make a definition that will be useful for both real and complex vector spaces, we need to examine the complex case before making the definition.

Recall that if  $\lambda = a + bi$ , where  $a, b \in \mathbf{R}$ , then

- the absolute value of  $\lambda$ , denoted  $|\lambda|$ , is defined by  $|\lambda| = \sqrt{a^2 + b^2}$ ;
- the complex conjugate of  $\lambda$ , denoted  $\bar{\lambda}$ , is defined by  $\bar{\lambda} = a - bi$ ;
- $|\lambda|^2 = \lambda\bar{\lambda}$ .

See Chapter 4 for the definitions and the basic properties of the absolute value and complex conjugate.

For  $z = (z_1, \dots, z_n) \in \mathbf{C}^n$ , we define the norm of  $z$  by

$$\|z\| = \sqrt{|z_1|^2 + \cdots + |z_n|^2}.$$

The absolute values are needed because we want  $\|z\|$  to be a nonnegative number. Note that

$$\|z\|^2 = z_1\bar{z}_1 + \cdots + z_n\bar{z}_n.$$

We want to think of  $\|z\|^2$  as the inner product of  $z$  with itself, as we did in  $\mathbf{R}^n$ . The equation above thus suggests that the inner product of  $w = (w_1, \dots, w_n) \in \mathbf{C}^n$  with  $z$  should equal

$$w_1\bar{z}_1 + \cdots + w_n\bar{z}_n.$$

If the roles of the  $w$  and  $z$  were interchanged, the expression above would be replaced with its complex conjugate. In other words, we should expect that the inner product of  $w$  with  $z$  equals the complex conjugate of the inner product of  $z$  with  $w$ . With that motivation, we are now ready to define an inner product on  $V$ , which may be a real or a complex vector space.

Two comments about the notation used in the next definition:

- If  $\lambda$  is a complex number, then the notation  $\lambda \geq 0$  means that  $\lambda$  is real and nonnegative.
- We use the common notation  $\langle u, v \rangle$ , with angle brackets denoting an inner product. Some people use parentheses instead, but then  $(u, v)$  becomes ambiguous because it could denote either an ordered pair or an inner product.

### 6.3 Definition inner product

An *inner product* on  $V$  is a function that takes each ordered pair  $(u, v)$  of elements of  $V$  to a number  $\langle u, v \rangle \in \mathbf{F}$  and has the following properties:

#### positivity

$\langle v, v \rangle \geq 0$  for all  $v \in V$ ;

#### definiteness

$\langle v, v \rangle = 0$  if and only if  $v = 0$ ;

#### additivity in first slot

$\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$  for all  $u, v, w \in V$ ;

#### homogeneity in first slot

$\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$  for all  $\lambda \in \mathbf{F}$  and all  $u, v \in V$ ;

#### conjugate symmetry

$\langle u, v \rangle = \langle v, u \rangle$  for all  $u, v \in V$ .

Although most mathematicians define an inner product as above, many physicists use a definition that requires homogeneity in the second slot instead of the first slot.

Every real number equals its complex conjugate. Thus if we are dealing with a real vector space, then in the last condition above we can dispense with the complex conjugate and simply state that  $\langle u, v \rangle = \langle v, u \rangle$  for all  $v, w \in V$ .

### 6.4 Example inner products

- (a) The *Euclidean inner product* on  $\mathbf{F}^n$  is defined by

$$\langle (w_1, \dots, w_n), (z_1, \dots, z_n) \rangle = w_1\bar{z}_1 + \dots + w_n\bar{z}_n.$$

- (b) If  $c_1, \dots, c_n$  are positive numbers, then an inner product can be defined on  $\mathbf{F}^n$  by

$$\langle (w_1, \dots, w_n), (z_1, \dots, z_n) \rangle = c_1 w_1 \bar{z}_1 + \dots + c_n w_n \bar{z}_n.$$

- (c) An inner product can be defined on the vector space of continuous real-valued functions on the interval  $[-1, 1]$  by

$$\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx.$$

- (d) An inner product can be defined on  $\mathcal{P}(\mathbf{R})$  by

$$\langle p, q \rangle = \int_0^\infty p(x)q(x)e^{-x} dx.$$

### 6.5 Definition inner product space

An *inner product space* is a vector space  $V$  along with an inner product on  $V$ .

The most important example of an inner product space is  $\mathbf{F}^n$  with the Euclidean inner product given by part (a) of the last example. When  $\mathbf{F}^n$  is referred to as an inner product space, you should assume that the inner product is the Euclidean inner product unless explicitly told otherwise.

So that we do not have to keep repeating the hypothesis that  $V$  is an inner product space, for the rest of this chapter we make the following assumption:

### 6.6 Notation $V$

For the rest of this chapter,  $V$  denotes an inner product space over  $\mathbf{F}$ .

Note the slight abuse of language here. An inner product space is a vector space along with an inner product on that vector space. When we say that a vector space  $V$  is an inner product space, we are also thinking that an inner product on  $V$  is lurking nearby or is obvious from the context (or is the Euclidean inner product if the vector space is  $\mathbf{F}^n$ ).

### 6.7 Basic properties of an inner product

- (a) For each fixed  $u \in V$ , the function that takes  $v$  to  $\langle v, u \rangle$  is a linear map from  $V$  to  $\mathbf{F}$ .
- (b)  $\langle 0, u \rangle = 0$  for every  $u \in V$ .
- (c)  $\langle u, 0 \rangle = 0$  for every  $u \in V$ .
- (d)  $\langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$  for all  $u, v, w \in V$ .
- (e)  $\langle u, \lambda v \rangle = \bar{\lambda} \langle u, v \rangle$  for all  $\lambda \in \mathbf{F}$  and  $u, v \in V$ .

#### Proof

- (a) Part (a) follows from the conditions of additivity in the first slot and homogeneity in the first slot in the definition of an inner product.
- (b) Part (b) follows from part (a) and the result that every linear map takes 0 to 0.

- (c) Part (c) follows from part (a) and the conjugate symmetry property in the definition of an inner product.

- (d) Suppose  $u, v, w \in V$ . Then

$$\begin{aligned}\langle u, v + w \rangle &= \overline{\langle v + w, u \rangle} \\ &= \overline{\langle v, u \rangle + \langle w, u \rangle} \\ &= \overline{\langle v, u \rangle} + \overline{\langle w, u \rangle} \\ &= \langle u, v \rangle + \langle u, w \rangle.\end{aligned}$$

- (e) Suppose  $\lambda \in \mathbf{F}$  and  $u, v \in V$ . Then

$$\begin{aligned}\langle u, \lambda v \rangle &= \overline{\langle \lambda v, u \rangle} \\ &= \overline{\lambda \langle v, u \rangle} \\ &= \bar{\lambda} \overline{\langle v, u \rangle} \\ &= \bar{\lambda} \langle u, v \rangle,\end{aligned}$$

as desired. ■

## Norms

Our motivation for defining inner products came initially from the norms of vectors on  $\mathbf{R}^2$  and  $\mathbf{R}^3$ . Now we see that each inner product determines a norm.

### 6.8 Definition norm, $\|v\|$

For  $v \in V$ , the **norm** of  $v$ , denoted  $\|v\|$ , is defined by

$$\|v\| = \sqrt{\langle v, v \rangle}.$$

---

### 6.9 Example norms

- (a) If  $(z_1, \dots, z_n) \in \mathbf{F}^n$  (with the Euclidean inner product), then

$$\|(z_1, \dots, z_n)\| = \sqrt{|z_1|^2 + \cdots + |z_n|^2}.$$

- (b) In the vector space of continuous real-valued functions on  $[-1, 1]$  [with inner product given as in part (c) of 6.4], we have

$$\|f\| = \sqrt{\int_{-1}^1 (f(x))^2 dx}.$$


---

### 6.10 Basic properties of the norm

Suppose  $v \in V$ .

- (a)  $\|v\| = 0$  if and only if  $v = 0$ .
- (b)  $\|\lambda v\| = |\lambda| \|v\|$  for all  $\lambda \in \mathbf{F}$ .

#### Proof

- (a) The desired result holds because  $\langle v, v \rangle = 0$  if and only if  $v = 0$ .
- (b) Suppose  $\lambda \in \mathbf{F}$ . Then

$$\begin{aligned}\|\lambda v\|^2 &= \langle \lambda v, \lambda v \rangle \\ &= \lambda \langle v, \lambda v \rangle \\ &= \lambda \bar{\lambda} \langle v, v \rangle \\ &= |\lambda|^2 \|v\|^2.\end{aligned}$$

Taking square roots now gives the desired equality. ■

The proof above of part (b) illustrates a general principle: working with norms squared is usually easier than working directly with norms.

Now we come to a crucial definition.

### 6.11 Definition *orthogonal*

Two vectors  $u, v \in V$  are called ***orthogonal*** if  $\langle u, v \rangle = 0$ .

In the definition above, the order of the vectors does not matter, because  $\langle u, v \rangle = 0$  if and only if  $\langle v, u \rangle = 0$ . Instead of saying that  $u$  and  $v$  are orthogonal, sometimes we say that  $u$  is orthogonal to  $v$ .

Exercise 13 asks you to prove that if  $u, v$  are nonzero vectors in  $\mathbf{R}^2$ , then

$$\langle u, v \rangle = \|u\| \|v\| \cos \theta,$$

where  $\theta$  is the angle between  $u$  and  $v$  (thinking of  $u$  and  $v$  as arrows with initial point at the origin). Thus two vectors in  $\mathbf{R}^2$  are orthogonal (with respect to the usual Euclidean inner product) if and only if the cosine of the angle between them is 0, which happens if and only if the vectors are perpendicular in the usual sense of plane geometry. Thus you can think of the word *orthogonal* as a fancy word meaning *perpendicular*.

We begin our study of orthogonality with an easy result.

### 6.12 Orthogonality and 0

- (a) 0 is orthogonal to every vector in  $V$ .
- (b) 0 is the only vector in  $V$  that is orthogonal to itself.

#### Proof

- (a) Part (b) of 6.7 states that  $\langle 0, u \rangle = 0$  for every  $u \in V$ .
- (b) If  $v \in V$  and  $\langle v, v \rangle = 0$ , then  $v = 0$  (by definition of inner product). ■

The word **orthogonal** comes from the Greek word **orthogonios**, which means right-angled.

For the special case  $V = \mathbf{R}^2$ , the next theorem is over 2,500 years old. Of course, the proof below is not the original proof.

### 6.13 Pythagorean Theorem

Suppose  $u$  and  $v$  are orthogonal vectors in  $V$ . Then

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2.$$

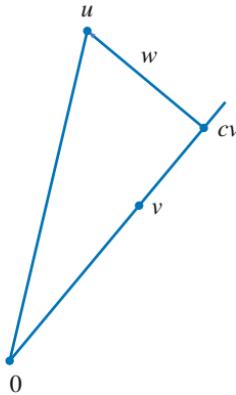
**Proof** We have

$$\begin{aligned}\|u + v\|^2 &= \langle u + v, u + v \rangle \\ &= \langle u, u \rangle + \langle u, v \rangle + \langle v, u \rangle + \langle v, v \rangle \\ &= \|u\|^2 + \|v\|^2,\end{aligned}$$

as desired. ■

The proof given above of the Pythagorean Theorem shows that the conclusion holds if and only if  $\langle u, v \rangle + \langle v, u \rangle$ , which equals  $2\operatorname{Re}\langle u, v \rangle$ , is 0. Thus the converse of the Pythagorean Theorem holds in real inner product spaces.

Suppose  $u, v \in V$ , with  $v \neq 0$ . We would like to write  $u$  as a scalar multiple of  $v$  plus a vector  $w$  orthogonal to  $v$ , as suggested in the next picture.



*An orthogonal decomposition.*

To discover how to write  $u$  as a scalar multiple of  $v$  plus a vector orthogonal to  $v$ , let  $c \in \mathbf{F}$  denote a scalar. Then

$$u = cv + (u - cv).$$

Thus we need to choose  $c$  so that  $v$  is orthogonal to  $(u - cv)$ . In other words, we want

$$0 = \langle u - cv, v \rangle = \langle u, v \rangle - c\|v\|^2.$$

The equation above shows that we should choose  $c$  to be  $\langle u, v \rangle / \|v\|^2$ . Making this choice of  $c$ , we can write

$$u = \frac{\langle u, v \rangle}{\|v\|^2} v + \left( u - \frac{\langle u, v \rangle}{\|v\|^2} v \right).$$

As you should verify, the equation above writes  $u$  as a scalar multiple of  $v$  plus a vector orthogonal to  $v$ . In other words, we have proved the following result.

#### 6.14 An orthogonal decomposition

Suppose  $u, v \in V$ , with  $v \neq 0$ . Set  $c = \frac{\langle u, v \rangle}{\|v\|^2}$  and  $w = u - \frac{\langle u, v \rangle}{\|v\|^2} v$ . Then

$$\langle w, v \rangle = 0 \quad \text{and} \quad u = cv + w.$$

The orthogonal decomposition 6.14 will be used in the proof of the Cauchy–Schwarz Inequality, which is our next result and is one of the most important inequalities in mathematics.

French mathematician Augustin-Louis Cauchy (1789–1857) proved 6.17(a) in 1821. German mathematician Hermann Schwarz (1843–1921) proved 6.17(b) in 1886.

### 6.15 Cauchy–Schwarz Inequality

Suppose  $u, v \in V$ . Then

$$|\langle u, v \rangle| \leq \|u\| \|v\|.$$

This inequality is an equality if and only if one of  $u, v$  is a scalar multiple of the other.

**Proof** If  $v = 0$ , then both sides of the desired inequality equal 0. Thus we can assume that  $v \neq 0$ . Consider the orthogonal decomposition

$$u = \frac{\langle u, v \rangle}{\|v\|^2} v + w$$

given by 6.14, where  $w$  is orthogonal to  $v$ . By the Pythagorean Theorem,

$$\begin{aligned} \|u\|^2 &= \left\| \frac{\langle u, v \rangle}{\|v\|^2} v \right\|^2 + \|w\|^2 \\ &= \frac{|\langle u, v \rangle|^2}{\|v\|^2} + \|w\|^2 \\ \mathbf{6.16} \quad &\geq \frac{|\langle u, v \rangle|^2}{\|v\|^2}. \end{aligned}$$

Multiplying both sides of this inequality by  $\|v\|^2$  and then taking square roots gives the desired inequality.

Looking at the proof in the paragraph above, note that the Cauchy–Schwarz Inequality is an equality if and only if 6.16 is an equality. Obviously this happens if and only if  $w = 0$ . But  $w = 0$  if and only if  $u$  is a multiple of  $v$  (see 6.14). Thus the Cauchy–Schwarz Inequality is an equality if and only if  $u$  is a scalar multiple of  $v$  or  $v$  is a scalar multiple of  $u$  (or both; the phrasing has been chosen to cover cases in which either  $u$  or  $v$  equals 0). ■

---

### 6.17 Example *examples of the Cauchy–Schwarz Inequality*

(a) If  $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbf{R}$ , then

$$|x_1 y_1 + \cdots + x_n y_n|^2 \leq (x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2).$$

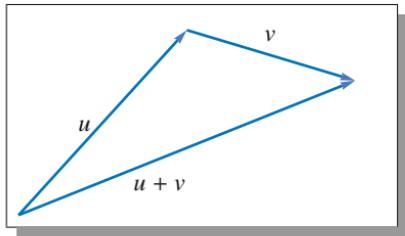
(b) If  $f, g$  are continuous real-valued functions on  $[-1, 1]$ , then

$$\left| \int_{-1}^1 f(x)g(x) dx \right|^2 \leq \left( \int_{-1}^1 (f(x))^2 dx \right) \left( \int_{-1}^1 (g(x))^2 dx \right).$$


---

The next result, called the Triangle Inequality, has the geometric interpretation that the length of each side of a triangle is less than the sum of the lengths of the other two sides.

Note that the Triangle Inequality implies that the shortest path between two points is a line segment.



### 6.18 Triangle Inequality

Suppose  $u, v \in V$ . Then

$$\|u + v\| \leq \|u\| + \|v\|.$$

This inequality is an equality if and only if one of  $u, v$  is a nonnegative multiple of the other.

**Proof** We have

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v \rangle \\ &= \langle u, u \rangle + \langle v, v \rangle + \langle u, v \rangle + \langle v, u \rangle \\ &= \langle u, u \rangle + \langle v, v \rangle + \langle u, v \rangle + \overline{\langle u, v \rangle} \\ &= \|u\|^2 + \|v\|^2 + 2 \operatorname{Re}\langle u, v \rangle \\ &\stackrel{6.19}{\leq} \|u\|^2 + \|v\|^2 + 2|\langle u, v \rangle| \\ &\stackrel{6.20}{\leq} \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| \\ &= (\|u\| + \|v\|)^2, \end{aligned}$$

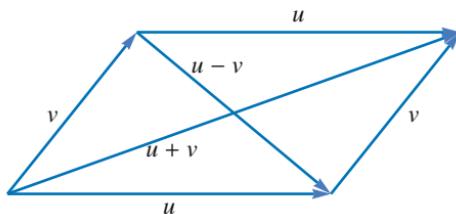
where 6.20 follows from the Cauchy–Schwarz Inequality (6.15). Taking square roots of both sides of the inequality above gives the desired inequality.

The proof above shows that the Triangle Inequality is an equality if and only if we have equality in 6.19 and 6.20. Thus we have equality in the Triangle Inequality if and only if

$$6.21 \quad \langle u, v \rangle = \|u\|\|v\|.$$

If one of  $u, v$  is a nonnegative multiple of the other, then 6.21 holds, as you should verify. Conversely, suppose 6.21 holds. Then the condition for equality in the Cauchy–Schwarz Inequality (6.15) implies that one of  $u, v$  is a scalar multiple of the other. Clearly 6.21 forces the scalar in question to be nonnegative, as desired. ■

The next result is called the parallelogram equality because of its geometric interpretation: in every parallelogram, the sum of the squares of the lengths of the diagonals equals the sum of the squares of the lengths of the four sides.



*The parallelogram equality.*

## 6.22 Parallelogram Equality

Suppose  $u, v \in V$ . Then

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

**Proof** We have

$$\begin{aligned} \|u + v\|^2 + \|u - v\|^2 &= \langle u + v, u + v \rangle + \langle u - v, u - v \rangle \\ &= \|u\|^2 + \|v\|^2 + \langle u, v \rangle + \langle v, u \rangle \\ &\quad + \|u\|^2 + \|v\|^2 - \langle u, v \rangle - \langle v, u \rangle \\ &= 2(\|u\|^2 + \|v\|^2), \end{aligned}$$

as desired. ■

Law professor Richard Friedman presenting a case before the U.S. Supreme Court in 2010:

*Mr. Friedman:* I think that issue is entirely orthogonal to the issue here because the Commonwealth is acknowledging—

*Chief Justice Roberts:* I'm sorry. Entirely what?

*Mr. Friedman:* Orthogonal. Right angle. Unrelated. Irrelevant.

*Chief Justice Roberts:* Oh.

*Justice Scalia:* What was that adjective? I liked that.

*Mr. Friedman:* Orthogonal.

*Chief Justice Roberts:* Orthogonal.

*Mr. Friedman:* Right, right.

*Justice Scalia:* Orthogonal, ooh. (Laughter.)

*Justice Kennedy:* I knew this case presented us a problem. (Laughter.)

## EXERCISES 6.A

- 1 Show that the function that takes  $((x_1, x_2), (y_1, y_2)) \in \mathbf{R}^2 \times \mathbf{R}^2$  to  $|x_1 y_1| + |x_2 y_2|$  is not an inner product on  $\mathbf{R}^2$ .
- 2 Show that the function that takes  $((x_1, x_2, x_3), (y_1, y_2, y_3)) \in \mathbf{R}^3 \times \mathbf{R}^3$  to  $x_1 y_1 + x_3 y_3$  is not an inner product on  $\mathbf{R}^3$ .
- 3 Suppose  $\mathbf{F} = \mathbf{R}$  and  $V \neq \{0\}$ . Replace the positivity condition (which states that  $\langle v, v \rangle \geq 0$  for all  $v \in V$ ) in the definition of an inner product (6.3) with the condition that  $\langle v, v \rangle > 0$  for some  $v \in V$ . Show that this change in the definition does not change the set of functions from  $V \times V$  to  $\mathbf{R}$  that are inner products on  $V$ .
- 4 Suppose  $V$  is a real inner product space.
  - (a) Show that  $\langle u + v, u - v \rangle = \|u\|^2 - \|v\|^2$  for every  $u, v \in V$ .
  - (b) Show that if  $u, v \in V$  have the same norm, then  $u + v$  is orthogonal to  $u - v$ .
  - (c) Use part (b) to show that the diagonals of a rhombus are perpendicular to each other.
- 5 Suppose  $T \in \mathcal{L}(V)$  is such that  $\|Tv\| \leq \|v\|$  for every  $v \in V$ . Prove that  $T - \sqrt{2}I$  is invertible.
- 6 Suppose  $u, v \in V$ . Prove that  $\langle u, v \rangle = 0$  if and only if

$$\|u\| \leq \|u + av\|$$

for all  $a \in \mathbf{F}$ .

- 7 Suppose  $u, v \in V$ . Prove that  $\|au + bv\| = \|bu + av\|$  for all  $a, b \in \mathbf{R}$  if and only if  $\|u\| = \|v\|$ .
- 8 Suppose  $u, v \in V$  and  $\|u\| = \|v\| = 1$  and  $\langle u, v \rangle = 1$ . Prove that  $u = v$ .
- 9 Suppose  $u, v \in V$  and  $\|u\| \leq 1$  and  $\|v\| \leq 1$ . Prove that

$$\sqrt{1 - \|u\|^2} \sqrt{1 - \|v\|^2} \leq 1 - |\langle u, v \rangle|.$$

- 10 Find vectors  $u, v \in \mathbf{R}^2$  such that  $u$  is a scalar multiple of  $(1, 3)$ ,  $v$  is orthogonal to  $(1, 3)$ , and  $(1, 2) = u + v$ .

**11** Prove that

$$16 \leq (a + b + c + d) \left( \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)$$

for all positive numbers  $a, b, c, d$ .

**12** Prove that

$$(x_1 + \cdots + x_n)^2 \leq n(x_1^2 + \cdots + x_n^2)$$

for all positive integers  $n$  and all real numbers  $x_1, \dots, x_n$ .

**13** Suppose  $u, v$  are nonzero vectors in  $\mathbf{R}^2$ . Prove that

$$\langle u, v \rangle = \|u\| \|v\| \cos \theta,$$

where  $\theta$  is the angle between  $u$  and  $v$  (thinking of  $u$  and  $v$  as arrows with initial point at the origin).

*Hint:* Draw the triangle formed by  $u, v$ , and  $u - v$ ; then use the law of cosines.

**14** The angle between two vectors (thought of as arrows with initial point at the origin) in  $\mathbf{R}^2$  or  $\mathbf{R}^3$  can be defined geometrically. However, geometry is not as clear in  $\mathbf{R}^n$  for  $n > 3$ . Thus the angle between two nonzero vectors  $x, y \in \mathbf{R}^n$  is defined to be

$$\arccos \frac{\langle x, y \rangle}{\|x\| \|y\|},$$

where the motivation for this definition comes from the previous exercise. Explain why the Cauchy–Schwarz Inequality is needed to show that this definition makes sense.

**15** Prove that

$$\left( \sum_{j=1}^n a_j b_j \right)^2 \leq \left( \sum_{j=1}^n j a_j^2 \right) \left( \sum_{j=1}^n \frac{b_j^2}{j} \right)$$

for all real numbers  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$ .

**16** Suppose  $u, v \in V$  are such that

$$\|u\| = 3, \quad \|u + v\| = 4, \quad \|u - v\| = 6.$$

What number does  $\|v\|$  equal?

- 17** Prove or disprove: there is an inner product on  $\mathbf{R}^2$  such that the associated norm is given by

$$\|(x, y)\| = \max\{x, y\}$$

for all  $(x, y) \in \mathbf{R}^2$ .

- 18** Suppose  $p > 0$ . Prove that there is an inner product on  $\mathbf{R}^2$  such that the associated norm is given by

$$\|(x, y)\| = (x^p + y^p)^{1/p}$$

for all  $(x, y) \in \mathbf{R}^2$  if and only if  $p = 2$ .

- 19** Suppose  $V$  is a real inner product space. Prove that

$$\langle u, v \rangle = \frac{\|u + v\|^2 - \|u - v\|^2}{4}$$

for all  $u, v \in V$ .

- 20** Suppose  $V$  is a complex inner product space. Prove that

$$\langle u, v \rangle = \frac{\|u + v\|^2 - \|u - v\|^2 + \|u + iv\|^2 i - \|u - iv\|^2 i}{4}$$

for all  $u, v \in V$ .

- 21** A norm on a vector space  $U$  is a function  $\| \cdot \|: U \rightarrow [0, \infty)$  such that  $\|u\| = 0$  if and only if  $u = 0$ ,  $\|\alpha u\| = |\alpha| \|u\|$  for all  $\alpha \in \mathbf{F}$  and all  $u \in U$ , and  $\|u + v\| \leq \|u\| + \|v\|$  for all  $u, v \in U$ . Prove that a norm satisfying the parallelogram equality comes from an inner product (in other words, show that if  $\| \cdot \|$  is a norm on  $U$  satisfying the parallelogram equality, then there is an inner product  $\langle \cdot, \cdot \rangle$  on  $U$  such that  $\|u\| = \langle u, u \rangle^{1/2}$  for all  $u \in U$ ).

- 22** Show that the square of an average is less than or equal to the average of the squares. More precisely, show that if  $a_1, \dots, a_n \in \mathbf{R}$ , then the square of the average of  $a_1, \dots, a_n$  is less than or equal to the average of  $a_1^2, \dots, a_n^2$ .

- 23** Suppose  $V_1, \dots, V_m$  are inner product spaces. Show that the equation

$$\langle (u_1, \dots, u_m), (v_1, \dots, v_m) \rangle = \langle u_1, v_1 \rangle + \cdots + \langle u_m, v_m \rangle$$

defines an inner product on  $V_1 \times \cdots \times V_m$ .

[In the expression above on the right,  $\langle u_1, v_1 \rangle$  denotes the inner product on  $V_1, \dots, \langle u_m, v_m \rangle$  denotes the inner product on  $V_m$ . Each of the spaces  $V_1, \dots, V_m$  may have a different inner product, even though the same notation is used here.]

- 24** Suppose  $S \in \mathcal{L}(V)$  is an injective operator on  $V$ . Define  $\langle \cdot, \cdot \rangle_1$  by

$$\langle u, v \rangle_1 = \langle Su, Sv \rangle$$

for  $u, v \in V$ . Show that  $\langle \cdot, \cdot \rangle_1$  is an inner product on  $V$ .

- 25** Suppose  $S \in \mathcal{L}(V)$  is not injective. Define  $\langle \cdot, \cdot \rangle_1$  as in the exercise above. Explain why  $\langle \cdot, \cdot \rangle_1$  is not an inner product on  $V$ .

- 26** Suppose  $f, g$  are differentiable functions from  $\mathbf{R}$  to  $\mathbf{R}^n$ .

- (a) Show that

$$\langle f(t), g(t) \rangle' = \langle f'(t), g(t) \rangle + \langle f(t), g'(t) \rangle.$$

- (b) Suppose  $c > 0$  and  $\|f(t)\| = c$  for every  $t \in \mathbf{R}$ . Show that  $\langle f'(t), f(t) \rangle = 0$  for every  $t \in \mathbf{R}$ .  
(c) Interpret the result in part (b) geometrically in terms of the tangent vector to a curve lying on a sphere in  $\mathbf{R}^n$  centered at the origin.

[For the exercise above, a function  $f : \mathbf{R} \rightarrow \mathbf{R}^n$  is called differentiable if there exist differentiable functions  $f_1, \dots, f_n$  from  $\mathbf{R}$  to  $\mathbf{R}$  such that  $f(t) = (f_1(t), \dots, f_n(t))$  for each  $t \in \mathbf{R}$ . Furthermore, for each  $t \in \mathbf{R}$ , the derivative  $f'(t) \in \mathbf{R}^n$  is defined by  $f'(t) = (f'_1(t), \dots, f'_n(t))$ .]

- 27** Suppose  $u, v, w \in V$ . Prove that

$$\|w - \frac{1}{2}(u + v)\|^2 = \frac{\|w - u\|^2 + \|w - v\|^2}{2} - \frac{\|u - v\|^2}{4}.$$

- 28** Suppose  $C$  is a subset of  $V$  with the property that  $u, v \in C$  implies  $\frac{1}{2}(u + v) \in C$ . Let  $w \in V$ . Show that there is at most one point in  $C$  that is closest to  $w$ . In other words, show that there is at most one  $u \in C$  such that

$$\|w - u\| \leq \|w - v\| \quad \text{for all } v \in C.$$

*Hint:* Use the previous exercise.

- 29** For  $u, v \in V$ , define  $d(u, v) = \|u - v\|$ .

- (a) Show that  $d$  is a metric on  $V$ .  
(b) Show that if  $V$  is finite-dimensional, then  $d$  is a complete metric on  $V$  (meaning that every Cauchy sequence converges).  
(c) Show that every finite-dimensional subspace of  $V$  is a closed subset of  $V$  (with respect to the metric  $d$ ).

- 30** Fix a positive integer  $n$ . The **Laplacian**  $\Delta p$  of a twice differentiable function  $p$  on  $\mathbf{R}^n$  is the function on  $\mathbf{R}^n$  defined by

$$\Delta p = \frac{\partial^2 p}{\partial x_1^2} + \cdots + \frac{\partial^2 p}{\partial x_n^2}.$$

The function  $p$  is called **harmonic** if  $\Delta p = 0$ .

A **polynomial** on  $\mathbf{R}^n$  is a linear combination of functions of the form  $x_1^{m_1} \cdots x_n^{m_n}$ , where  $m_1, \dots, m_n$  are nonnegative integers.

Suppose  $q$  is a polynomial on  $\mathbf{R}^n$ . Prove that there exists a harmonic polynomial  $p$  on  $\mathbf{R}^n$  such that  $p(x) = q(x)$  for every  $x \in \mathbf{R}^n$  with  $\|x\| = 1$ .

[The only fact about harmonic functions that you need for this exercise is that if  $p$  is a harmonic function on  $\mathbf{R}^n$  and  $p(x) = 0$  for all  $x \in \mathbf{R}^n$  with  $\|x\| = 1$ , then  $p = 0$ .]

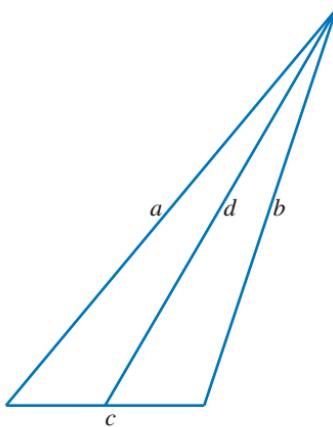
*Hint:* A reasonable guess is that the desired harmonic polynomial  $p$  is of the form  $q + (1 - \|x\|^2)r$  for some polynomial  $r$ . Prove that there is a polynomial  $r$  on  $\mathbf{R}^n$  such that  $q + (1 - \|x\|^2)r$  is harmonic by defining an operator  $T$  on a suitable vector space by

$$Tr = \Delta((1 - \|x\|^2)r)$$

and then showing that  $T$  is injective and hence surjective.

- 31** Use inner products to prove Apollonius's Identity: In a triangle with sides of length  $a$ ,  $b$ , and  $c$ , let  $d$  be the length of the line segment from the midpoint of the side of length  $c$  to the opposite vertex. Then

$$a^2 + b^2 = \frac{1}{2}c^2 + 2d^2.$$



## 6.B Orthonormal Bases

### 6.23 Definition orthonormal

- A list of vectors is called **orthonormal** if each vector in the list has norm 1 and is orthogonal to all the other vectors in the list.
- In other words, a list  $e_1, \dots, e_m$  of vectors in  $V$  is orthonormal if

$$\langle e_j, e_k \rangle = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}$$

### 6.24 Example orthonormal lists

- (a) The standard basis in  $\mathbf{F}^n$  is an orthonormal list.
- (b)  $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}), (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$  is an orthonormal list in  $\mathbf{F}^3$ .
- (c)  $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}), (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0), (\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}})$  is an orthonormal list in  $\mathbf{F}^3$ .

Orthonormal lists are particularly easy to work with, as illustrated by the next result.

### 6.25 The norm of an orthonormal linear combination

If  $e_1, \dots, e_m$  is an orthonormal list of vectors in  $V$ , then

$$\|a_1e_1 + \cdots + a_m e_m\|^2 = |a_1|^2 + \cdots + |a_m|^2$$

for all  $a_1, \dots, a_m \in \mathbf{F}$ .

**Proof** Because each  $e_j$  has norm 1, this follows easily from repeated applications of the Pythagorean Theorem (6.13). ■

The result above has the following important corollary.

### 6.26 An orthonormal list is linearly independent

Every orthonormal list of vectors is linearly independent.

**Proof** Suppose  $e_1, \dots, e_m$  is an orthonormal list of vectors in  $V$  and  $a_1, \dots, a_m \in \mathbf{F}$  are such that

$$a_1e_1 + \cdots + a_m e_m = 0.$$

Then  $|a_1|^2 + \cdots + |a_m|^2 = 0$  (by 6.25), which means that all the  $a_j$ 's are 0. Thus  $e_1, \dots, e_m$  is linearly independent. ■

### 6.27 Definition orthonormal basis

An **orthonormal basis** of  $V$  is an orthonormal list of vectors in  $V$  that is also a basis of  $V$ .

For example, the standard basis is an orthonormal basis of  $\mathbf{F}^n$ .

### 6.28 An orthonormal list of the right length is an orthonormal basis

Every orthonormal list of vectors in  $V$  with length  $\dim V$  is an orthonormal basis of  $V$ .

**Proof** By 6.26, any such list must be linearly independent; because it has the right length, it is a basis—see 2.39. ■

### 6.29 Example Show that

$$\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}\right), \left(\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}\right), \left(-\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, \frac{1}{2}\right)$$

is an orthonormal basis of  $\mathbf{F}^4$ .

**Solution** We have

$$\left\| \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) \right\| = \sqrt{\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2} = 1.$$

Similarly, the other three vectors in the list above also have norm 1.

We have

$$\langle \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}\right) \rangle = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \left(-\frac{1}{2}\right) + \frac{1}{2} \cdot \left(-\frac{1}{2}\right) = 0.$$

Similarly, the inner product of any two distinct vectors in the list above also equals 0.

Thus the list above is orthonormal. Because we have an orthonormal list of length four in the four-dimensional vector space  $\mathbf{F}^4$ , this list is an orthonormal basis of  $\mathbf{F}^4$  (by 6.28).

In general, given a basis  $e_1, \dots, e_n$  of  $V$  and a vector  $v \in V$ , we know that there is some choice of scalars  $a_1, \dots, a_n \in \mathbf{F}$  such that

$$v = a_1e_1 + \cdots + a_ne_n.$$

*The importance of orthonormal bases stems mainly from the next result.*

Computing the numbers  $a_1, \dots, a_n$  that satisfy the equation above can be difficult for an arbitrary basis of  $V$ . The next result shows, however, that this is easy for an orthonormal basis—just take  $a_j = \langle v, e_j \rangle$ .

### 6.30 Writing a vector as linear combination of orthonormal basis

Suppose  $e_1, \dots, e_n$  is an orthonormal basis of  $V$  and  $v \in V$ . Then

$$v = \langle v, e_1 \rangle e_1 + \cdots + \langle v, e_n \rangle e_n$$

and

$$\|v\|^2 = |\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_n \rangle|^2.$$

**Proof** Because  $e_1, \dots, e_n$  is a basis of  $V$ , there exist scalars  $a_1, \dots, a_n$  such that

$$v = a_1e_1 + \cdots + a_ne_n.$$

Because  $e_1, \dots, e_n$  is orthonormal, taking the inner product of both sides of this equation with  $e_j$  gives  $\langle v, e_j \rangle = a_j$ . Thus the first equation in 6.30 holds.

The second equation in 6.30 follows immediately from the first equation and 6.25. ■

Now that we understand the usefulness of orthonormal bases, how do we go about finding them? For example, does  $\mathcal{P}_m(\mathbf{R})$ , with inner product given by integration on  $[-1, 1]$  [see 6.4(c)], have an orthonormal basis? The next result will lead to answers to these questions.

*Danish mathematician Jørgen Gram (1850–1916) and German mathematician Erhard Schmidt (1876–1959) popularized this algorithm that constructs orthonormal lists.*

The algorithm used in the next proof is called the **Gram–Schmidt Procedure**. It gives a method for turning a linearly independent list into an orthonormal list with the same span as the original list.

### 6.31 Gram–Schmidt Procedure

Suppose  $v_1, \dots, v_m$  is a linearly independent list of vectors in  $V$ . Let  $e_1 = v_1/\|v_1\|$ . For  $j = 2, \dots, m$ , define  $e_j$  inductively by

$$e_j = \frac{v_j - \langle v_j, e_1 \rangle e_1 - \cdots - \langle v_j, e_{j-1} \rangle e_{j-1}}{\|v_j - \langle v_j, e_1 \rangle e_1 - \cdots - \langle v_j, e_{j-1} \rangle e_{j-1}\|}.$$

Then  $e_1, \dots, e_m$  is an orthonormal list of vectors in  $V$  such that

$$\text{span}(v_1, \dots, v_j) = \text{span}(e_1, \dots, e_j)$$

for  $j = 1, \dots, m$ .

**Proof** We will show by induction on  $j$  that the desired conclusion holds. To get started with  $j = 1$ , note that  $\text{span}(v_1) = \text{span}(e_1)$  because  $v_1$  is a positive multiple of  $e_1$ .

Suppose  $1 < j < m$  and we have verified that

$$6.32 \quad \text{span}(v_1, \dots, v_{j-1}) = \text{span}(e_1, \dots, e_{j-1}).$$

Note that  $v_j \notin \text{span}(v_1, \dots, v_{j-1})$  (because  $v_1, \dots, v_m$  is linearly independent). Thus  $v_j \notin \text{span}(e_1, \dots, e_{j-1})$ . Hence we are not dividing by 0 in the definition of  $e_j$  given in 6.31. Dividing a vector by its norm produces a new vector with norm 1; thus  $\|e_j\| = 1$ .

Let  $1 \leq k < j$ . Then

$$\begin{aligned} \langle e_j, e_k \rangle &= \left\langle \frac{v_j - \langle v_j, e_1 \rangle e_1 - \cdots - \langle v_j, e_{j-1} \rangle e_{j-1}}{\|v_j - \langle v_j, e_1 \rangle e_1 - \cdots - \langle v_j, e_{j-1} \rangle e_{j-1}\|}, e_k \right\rangle \\ &= \frac{\langle v_j, e_k \rangle - \langle v_j, e_k \rangle}{\|v_j - \langle v_j, e_1 \rangle e_1 - \cdots - \langle v_j, e_{j-1} \rangle e_{j-1}\|} \\ &= 0. \end{aligned}$$

Thus  $e_1, \dots, e_j$  is an orthonormal list.

From the definition of  $e_j$  given in 6.31, we see that  $v_j \in \text{span}(e_1, \dots, e_j)$ . Combining this information with 6.32 shows that

$$\text{span}(v_1, \dots, v_j) \subset \text{span}(e_1, \dots, e_j).$$

Both lists above are linearly independent (the  $v$ 's by hypothesis, the  $e$ 's by orthonormality and 6.26). Thus both subspaces above have dimension  $j$ , and hence they are equal, completing the proof. ■

**6.33 Example** Find an orthonormal basis of  $\mathcal{P}_2(\mathbf{R})$ , where the inner product is given by  $\langle p, q \rangle = \int_{-1}^1 p(x)q(x) dx$ .

**Solution** We will apply the Gram–Schmidt Procedure (6.31) to the basis  $1, x, x^2$ .

To get started, with this inner product we have

$$\|1\|^2 = \int_{-1}^1 1^2 dx = 2.$$

Thus  $\|1\| = \sqrt{2}$ , and hence  $e_1 = \sqrt{\frac{1}{2}}$ .

Now the numerator in the expression for  $e_2$  is

$$x - \langle x, e_1 \rangle e_1 = x - \left( \int_{-1}^1 x \sqrt{\frac{1}{2}} dx \right) \sqrt{\frac{1}{2}} = x.$$

We have

$$\|x\|^2 = \int_{-1}^1 x^2 dx = \frac{2}{3}.$$

Thus  $\|x\| = \sqrt{\frac{2}{3}}$ , and hence  $e_2 = \sqrt{\frac{3}{2}}x$ .

Now the numerator in the expression for  $e_3$  is

$$\begin{aligned} & x^2 - \langle x^2, e_1 \rangle e_1 - \langle x^2, e_2 \rangle e_2 \\ &= x^2 - \left( \int_{-1}^1 x^2 \sqrt{\frac{1}{2}} dx \right) \sqrt{\frac{1}{2}} - \left( \int_{-1}^1 x^2 \sqrt{\frac{3}{2}}x dx \right) \sqrt{\frac{3}{2}}x \\ &= x^2 - \frac{1}{3}. \end{aligned}$$

We have

$$\|x^2 - \frac{1}{3}\|^2 = \int_{-1}^1 \left( x^4 - \frac{2}{3}x^2 + \frac{1}{9} \right) dx = \frac{8}{45}.$$

Thus  $\|x^2 - \frac{1}{3}\| = \sqrt{\frac{8}{45}}$ , and hence  $e_3 = \sqrt{\frac{45}{8}}(x^2 - \frac{1}{3})$ .

Thus

$$\sqrt{\frac{1}{2}}, \sqrt{\frac{3}{2}}x, \sqrt{\frac{45}{8}}(x^2 - \frac{1}{3})$$

is an orthonormal list of length 3 in  $\mathcal{P}_2(\mathbf{R})$ . Hence this orthonormal list is an orthonormal basis of  $\mathcal{P}_2(\mathbf{R})$  by 6.28.

Now we can answer the question about the existence of orthonormal bases.

### 6.34 Existence of orthonormal basis

Every finite-dimensional inner product space has an orthonormal basis.

**Proof** Suppose  $V$  is finite-dimensional. Choose a basis of  $V$ . Apply the Gram–Schmidt Procedure (6.31) to it, producing an orthonormal list with length  $\dim V$ . By 6.28, this orthonormal list is an orthonormal basis of  $V$ . ■

Sometimes we need to know not only that an orthonormal basis exists, but also that every orthonormal list can be extended to an orthonormal basis. In the next corollary, the Gram–Schmidt Procedure shows that such an extension is always possible.

### 6.35 Orthonormal list extends to orthonormal basis

Suppose  $V$  is finite-dimensional. Then every orthonormal list of vectors in  $V$  can be extended to an orthonormal basis of  $V$ .

**Proof** Suppose  $e_1, \dots, e_m$  is an orthonormal list of vectors in  $V$ . Then  $e_1, \dots, e_m$  is linearly independent (by 6.26). Hence this list can be extended to a basis  $e_1, \dots, e_m, v_1, \dots, v_n$  of  $V$  (see 2.33). Now apply the Gram–Schmidt Procedure (6.31) to  $e_1, \dots, e_m, v_1, \dots, v_n$ , producing an orthonormal list

$$6.36 \quad e_1, \dots, e_m, f_1, \dots, f_n;$$

here the formula given by the Gram–Schmidt Procedure leaves the first  $m$  vectors unchanged because they are already orthonormal. The list above is an orthonormal basis of  $V$  by 6.28. ■

Recall that a matrix is called upper triangular if all entries below the diagonal equal 0. In other words, an upper-triangular matrix looks like this:

$$\begin{pmatrix} * & & * \\ & \ddots & \\ 0 & & * \end{pmatrix},$$

where the 0 in the matrix above indicates that all entries below the diagonal equal 0, and asterisks are used to denote entries on and above the diagonal.

In the last chapter we showed that if  $V$  is a finite-dimensional complex vector space, then for each operator on  $V$  there is a basis with respect to which the matrix of the operator is upper triangular (see 5.27). Now that we are dealing with inner product spaces, we would like to know whether there exists an *orthonormal* basis with respect to which we have an upper-triangular matrix.

The next result shows that the existence of a basis with respect to which  $T$  has an upper-triangular matrix implies the existence of an orthonormal basis with this property. This result is true on both real and complex vector spaces (although on a real vector space, the hypothesis holds only for some operators).

### 6.37 Upper-triangular matrix with respect to orthonormal basis

Suppose  $T \in \mathcal{L}(V)$ . If  $T$  has an upper-triangular matrix with respect to some basis of  $V$ , then  $T$  has an upper-triangular matrix with respect to some orthonormal basis of  $V$ .

**Proof** Suppose  $T$  has an upper-triangular matrix with respect to some basis  $v_1, \dots, v_n$  of  $V$ . Thus  $\text{span}(v_1, \dots, v_j)$  is invariant under  $T$  for each  $j = 1, \dots, n$  (see 5.26).

Apply the Gram–Schmidt Procedure to  $v_1, \dots, v_n$ , producing an orthonormal basis  $e_1, \dots, e_n$  of  $V$ . Because

$$\text{span}(e_1, \dots, e_j) = \text{span}(v_1, \dots, v_j)$$

for each  $j$  (see 6.31), we conclude that  $\text{span}(e_1, \dots, e_j)$  is invariant under  $T$  for each  $j = 1, \dots, n$ . Thus, by 5.26,  $T$  has an upper-triangular matrix with respect to the orthonormal basis  $e_1, \dots, e_n$ . ■

*German mathematician Issai Schur (1875–1941) published the first proof of the next result in 1909.*

The next result is an important application of the result above.

### 6.38 Schur's Theorem

Suppose  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ . Then  $T$  has an upper-triangular matrix with respect to some orthonormal basis of  $V$ .

**Proof** Recall that  $T$  has an upper-triangular matrix with respect to some basis of  $V$  (see 5.27). Now apply 6.37. ■

## Linear Functionals on Inner Product Spaces

Because linear maps into the scalar field  $\mathbf{F}$  play a special role, we defined a special name for them in Section 3.F. That definition is repeated below in case you skipped Section 3.F.

### 6.39 Definition *linear functional*

A *linear functional* on  $V$  is a linear map from  $V$  to  $\mathbf{F}$ . In other words, a linear functional is an element of  $\mathcal{L}(V, \mathbf{F})$ .

### 6.40 Example The function $\varphi: \mathbf{F}^3 \rightarrow \mathbf{F}$ defined by

$$\varphi(z_1, z_2, z_3) = 2z_1 - 5z_2 + z_3$$

is a linear functional on  $\mathbf{F}^3$ . We could write this linear functional in the form

$$\varphi(z) = \langle z, u \rangle$$

for every  $z \in \mathbf{F}^3$ , where  $u = (2, -5, 1)$ .

### 6.41 Example The function $\varphi: \mathcal{P}_2(\mathbf{R}) \rightarrow \mathbf{R}$ defined by

$$\varphi(p) = \int_{-1}^1 p(t)(\cos(\pi t)) dt$$

is a linear functional on  $\mathcal{P}_2(\mathbf{R})$  (here the inner product on  $\mathcal{P}_2(\mathbf{R})$  is multiplication followed by integration on  $[-1, 1]$ ; see 6.33). It is not obvious that there exists  $u \in \mathcal{P}_2(\mathbf{R})$  such that

$$\varphi(p) = \langle p, u \rangle$$

for every  $p \in \mathcal{P}_2(\mathbf{R})$  [we cannot take  $u(t) = \cos(\pi t)$  because that function is not an element of  $\mathcal{P}_2(\mathbf{R})$ ].

If  $u \in V$ , then the map that sends  $v$  to  $\langle v, u \rangle$  is a linear functional on  $V$ . The next result shows that every linear functional on  $V$  is of this form. Example 6.41 above illustrates the power of the next result because for the linear functional in that example, there is no obvious candidate for  $u$ .

*The next result is named in honor of Hungarian mathematician Frigyes Riesz (1880–1956), who proved several results early in the twentieth century that look very much like the result below.*

### 6.42 Riesz Representation Theorem

Suppose  $V$  is finite-dimensional and  $\varphi$  is a linear functional on  $V$ . Then there is a unique vector  $u \in V$  such that

$$\varphi(v) = \langle v, u \rangle$$

for every  $v \in V$ .

**Proof** First we show there exists a vector  $u \in V$  such that  $\varphi(v) = \langle v, u \rangle$  for every  $v \in V$ . Let  $e_1, \dots, e_n$  be an orthonormal basis of  $V$ . Then

$$\begin{aligned}\varphi(v) &= \varphi(\langle v, e_1 \rangle e_1 + \dots + \langle v, e_n \rangle e_n) \\ &= \langle v, e_1 \rangle \varphi(e_1) + \dots + \langle v, e_n \rangle \varphi(e_n) \\ &= \langle v, \overline{\varphi(e_1)} e_1 + \dots + \overline{\varphi(e_n)} e_n \rangle\end{aligned}$$

for every  $v \in V$ , where the first equality comes from 6.30. Thus setting

$$6.43 \quad u = \overline{\varphi(e_1)} e_1 + \dots + \overline{\varphi(e_n)} e_n,$$

we have  $\varphi(v) = \langle v, u \rangle$  for every  $v \in V$ , as desired.

Now we prove that only one vector  $u \in V$  has the desired behavior. Suppose  $u_1, u_2 \in V$  are such that

$$\varphi(v) = \langle v, u_1 \rangle = \langle v, u_2 \rangle$$

for every  $v \in V$ . Then

$$0 = \langle v, u_1 \rangle - \langle v, u_2 \rangle = \langle v, u_1 - u_2 \rangle$$

for every  $v \in V$ . Taking  $v = u_1 - u_2$  shows that  $u_1 - u_2 = 0$ . In other words,  $u_1 = u_2$ , completing the proof of the uniqueness part of the result. ■

6.44 **Example** Find  $u \in \mathcal{P}_2(\mathbf{R})$  such that

$$\int_{-1}^1 p(t)(\cos(\pi t)) dt = \int_{-1}^1 p(t)u(t) dt$$

for every  $p \in \mathcal{P}_2(\mathbf{R})$ .

**Solution** Let  $\varphi(p) = \int_{-1}^1 p(t)(\cos(\pi t)) dt$ . Applying formula 6.43 from the proof above, and using the orthonormal basis from Example 6.33, we have

$$\begin{aligned} u(x) &= \left( \int_{-1}^1 \sqrt{\frac{1}{2}}(\cos(\pi t)) dt \right) \sqrt{\frac{1}{2}} + \left( \int_{-1}^1 \sqrt{\frac{3}{2}}t(\cos(\pi t)) dt \right) \sqrt{\frac{3}{2}}x \\ &\quad + \left( \int_{-1}^1 \sqrt{\frac{45}{8}}(t^2 - \frac{1}{3})(\cos(\pi t)) dt \right) \sqrt{\frac{45}{8}}(x^2 - \frac{1}{3}). \end{aligned}$$

A bit of calculus shows that

$$u(x) = -\frac{45}{2\pi^2}(x^2 - \frac{1}{3}).$$


---

Suppose  $V$  is finite-dimensional and  $\varphi$  a linear functional on  $V$ . Then 6.43 gives a formula for the vector  $u$  that satisfies  $\varphi(v) = \langle v, u \rangle$  for all  $v \in V$ . Specifically, we have

$$u = \overline{\varphi(e_1)}e_1 + \cdots + \overline{\varphi(e_n)}e_n.$$

The right side of the equation above seems to depend on the orthonormal basis  $e_1, \dots, e_n$  as well as on  $\varphi$ . However, 6.42 tells us that  $u$  is uniquely determined by  $\varphi$ . Thus the right side of the equation above is the same regardless of which orthonormal basis  $e_1, \dots, e_n$  of  $V$  is chosen.

## EXERCISES 6.B

---

- 1 (a) Suppose  $\theta \in \mathbf{R}$ . Show that  $(\cos \theta, \sin \theta), (-\sin \theta, \cos \theta)$  and  $(\cos \theta, \sin \theta), (\sin \theta, -\cos \theta)$  are orthonormal bases of  $\mathbf{R}^2$ .  
(b) Show that each orthonormal basis of  $\mathbf{R}^2$  is of the form given by one of the two possibilities of part (a).
- 2 Suppose  $e_1, \dots, e_m$  is an orthonormal list of vectors in  $V$ . Let  $v \in V$ . Prove that
$$\|v\|^2 = |\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_m \rangle|^2$$
if and only if  $v \in \text{span}(e_1, \dots, e_m)$ .
- 3 Suppose  $T \in \mathcal{L}(\mathbf{R}^3)$  has an upper-triangular matrix with respect to the basis  $(1, 0, 0), (1, 1, 1), (1, 1, 2)$ . Find an orthonormal basis of  $\mathbf{R}^3$  (use the usual inner product on  $\mathbf{R}^3$ ) with respect to which  $T$  has an upper-triangular matrix.

- 4 Suppose  $n$  is a positive integer. Prove that

$$\frac{1}{\sqrt{2\pi}}, \frac{\cos x}{\sqrt{\pi}}, \frac{\cos 2x}{\sqrt{\pi}}, \dots, \frac{\cos nx}{\sqrt{\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\sin 2x}{\sqrt{\pi}}, \dots, \frac{\sin nx}{\sqrt{\pi}}$$

is an orthonormal list of vectors in  $C[-\pi, \pi]$ , the vector space of continuous real-valued functions on  $[-\pi, \pi]$  with inner product

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)g(x) dx.$$

[The orthonormal list above is often used for modeling periodic phenomena such as tides.]

- 5 On  $\mathcal{P}_2(\mathbf{R})$ , consider the inner product given by

$$\langle p, q \rangle = \int_0^1 p(x)q(x) dx.$$

Apply the Gram–Schmidt Procedure to the basis  $1, x, x^2$  to produce an orthonormal basis of  $\mathcal{P}_2(\mathbf{R})$ .

- 6 Find an orthonormal basis of  $\mathcal{P}_2(\mathbf{R})$  (with inner product as in Exercise 5) such that the differentiation operator (the operator that takes  $p$  to  $p'$ ) on  $\mathcal{P}_2(\mathbf{R})$  has an upper-triangular matrix with respect to this basis.

- 7 Find a polynomial  $q \in \mathcal{P}_2(\mathbf{R})$  such that

$$p\left(\frac{1}{2}\right) = \int_0^1 p(x)q(x) dx$$

for every  $p \in \mathcal{P}_2(\mathbf{R})$ .

- 8 Find a polynomial  $q \in \mathcal{P}_2(\mathbf{R})$  such that

$$\int_0^1 p(x)(\cos \pi x) dx = \int_0^1 p(x)q(x) dx$$

for every  $p \in \mathcal{P}_2(\mathbf{R})$ .

- 9 What happens if the Gram–Schmidt Procedure is applied to a list of vectors that is not linearly independent?

- 10** Suppose  $V$  is a real inner product space and  $v_1, \dots, v_m$  is a linearly independent list of vectors in  $V$ . Prove that there exist exactly  $2^m$  orthonormal lists  $e_1, \dots, e_m$  of vectors in  $V$  such that

$$\text{span}(v_1, \dots, v_j) = \text{span}(e_1, \dots, e_j)$$

for all  $j \in \{1, \dots, m\}$ .

- 11** Suppose  $\langle \cdot, \cdot \rangle_1$  and  $\langle \cdot, \cdot \rangle_2$  are inner products on  $V$  such that  $\langle v, w \rangle_1 = 0$  if and only if  $\langle v, w \rangle_2 = 0$ . Prove that there is a positive number  $c$  such that  $\langle v, w \rangle_1 = c \langle v, w \rangle_2$  for every  $v, w \in V$ .
- 12** Suppose  $V$  is finite-dimensional and  $\langle \cdot, \cdot \rangle_1, \langle \cdot, \cdot \rangle_2$  are inner products on  $V$  with corresponding norms  $\|\cdot\|_1$  and  $\|\cdot\|_2$ . Prove that there exists a positive number  $c$  such that

$$\|v\|_1 \leq c \|v\|_2$$

for every  $v \in V$ .

- 13** Suppose  $v_1, \dots, v_m$  is a linearly independent list in  $V$ . Show that there exists  $w \in V$  such that  $\langle w, v_j \rangle > 0$  for all  $j \in \{1, \dots, m\}$ .
- 14** Suppose  $e_1, \dots, e_n$  is an orthonormal basis of  $V$  and  $v_1, \dots, v_n$  are vectors in  $V$  such that

$$\|e_j - v_j\| < \frac{1}{\sqrt{n}}$$

for each  $j$ . Prove that  $v_1, \dots, v_n$  is a basis of  $V$ .

- 15** Suppose  $C_{\mathbf{R}}([-1, 1])$  is the vector space of continuous real-valued functions on the interval  $[-1, 1]$  with inner product given by

$$\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx$$

for  $f, g \in C_{\mathbf{R}}([-1, 1])$ . Let  $\varphi$  be the linear functional on  $C_{\mathbf{R}}([-1, 1])$  defined by  $\varphi(f) = f(0)$ . Show that there does not exist  $g \in C_{\mathbf{R}}([-1, 1])$  such that

$$\varphi(f) = \langle f, g \rangle$$

for every  $f \in C_{\mathbf{R}}([-1, 1])$ .

[The exercise above shows that the Riesz Representation Theorem (6.42) does not hold on infinite-dimensional vector spaces without additional hypotheses on  $V$  and  $\varphi$ .]

- 16** Suppose  $\mathbf{F} = \mathbf{C}$ ,  $V$  is finite-dimensional,  $T \in \mathcal{L}(V)$ , all the eigenvalues of  $T$  have absolute value less than 1, and  $\epsilon > 0$ . Prove that there exists a positive integer  $m$  such that  $\|T^m v\| \leq \epsilon \|v\|$  for every  $v \in V$ .
- 17** For  $u \in V$ , let  $\Phi u$  denote the linear functional on  $V$  defined by

$$(\Phi u)(v) = \langle v, u \rangle$$

for  $v \in V$ .

- (a) Show that if  $\mathbf{F} = \mathbf{R}$ , then  $\Phi$  is a linear map from  $V$  to  $V'$ . (Recall from Section 3.F that  $V' = \mathcal{L}(V, \mathbf{F})$  and that  $V'$  is called the dual space of  $V$ .)
- (b) Show that if  $\mathbf{F} = \mathbf{C}$  and  $V \neq \{0\}$ , then  $\Phi$  is not a linear map.
- (c) Show that  $\Phi$  is injective.
- (d) Suppose  $\mathbf{F} = \mathbf{R}$  and  $V$  is finite-dimensional. Use parts (a) and (c) and a dimension-counting argument (but without using 6.42) to show that  $\Phi$  is an isomorphism from  $V$  onto  $V'$ .

[Part (d) gives an alternative proof of the Riesz Representation Theorem (6.42) when  $\mathbf{F} = \mathbf{R}$ . Part (d) also gives a natural isomorphism (meaning that it does not depend on a choice of basis) from a finite-dimensional real inner product space onto its dual space.]

## 6.C Orthogonal Complements and Minimization Problems

### Orthogonal Complements

#### 6.45 Definition *orthogonal complement*, $U^\perp$

If  $U$  is a subset of  $V$ , then the *orthogonal complement* of  $U$ , denoted  $U^\perp$ , is the set of all vectors in  $V$  that are orthogonal to every vector in  $U$ :

$$U^\perp = \{v \in V : \langle v, u \rangle = 0 \text{ for every } u \in U\}.$$

For example, if  $U$  is a line in  $\mathbf{R}^3$ , then  $U^\perp$  is the plane containing the origin that is perpendicular to  $U$ . If  $U$  is a plane in  $\mathbf{R}^3$ , then  $U^\perp$  is the line containing the origin that is perpendicular to  $U$ .

#### 6.46 Basic properties of orthogonal complement

- (a) If  $U$  is a subset of  $V$ , then  $U^\perp$  is a subspace of  $V$ .
- (b)  $\{0\}^\perp = V$ .
- (c)  $V^\perp = \{0\}$ .
- (d) If  $U$  is a subset of  $V$ , then  $U \cap U^\perp \subset \{0\}$ .
- (e) If  $U$  and  $W$  are subsets of  $V$  and  $U \subset W$ , then  $W^\perp \subset U^\perp$ .

#### Proof

- (a) Suppose  $U$  is a subset of  $V$ . Then  $\langle 0, u \rangle = 0$  for every  $u \in U$ ; thus  $0 \in U^\perp$ .

Suppose  $v, w \in U^\perp$ . If  $u \in U$ , then

$$\langle v + w, u \rangle = \langle v, u \rangle + \langle w, u \rangle = 0 + 0 = 0.$$

Thus  $v + w \in U^\perp$ . In other words,  $U^\perp$  is closed under addition.

Similarly, suppose  $\lambda \in \mathbf{F}$  and  $v \in U^\perp$ . If  $u \in U$ , then

$$\langle \lambda v, u \rangle = \lambda \langle v, u \rangle = \lambda \cdot 0 = 0.$$

Thus  $\lambda v \in U^\perp$ . In other words,  $U^\perp$  is closed under scalar multiplication. Thus  $U^\perp$  is a subspace of  $V$ .

- (b) Suppose  $v \in V$ . Then  $\langle v, 0 \rangle = 0$ , which implies that  $v \in \{0\}^\perp$ . Thus  $\{0\}^\perp = V$ .
- (c) Suppose  $v \in V^\perp$ . Then  $\langle v, v \rangle = 0$ , which implies that  $v = 0$ . Thus  $V^\perp = \{0\}$ .
- (d) Suppose  $U$  is a subset of  $V$  and  $v \in U \cap U^\perp$ . Then  $\langle v, v \rangle = 0$ , which implies that  $v = 0$ . Thus  $U \cap U^\perp \subset \{0\}$ .
- (e) Suppose  $U$  and  $W$  are subsets of  $V$  and  $U \subset W$ . Suppose  $v \in W^\perp$ . Then  $\langle v, u \rangle = 0$  for every  $u \in W$ , which implies that  $\langle v, u \rangle = 0$  for every  $u \in U$ . Hence  $v \in U^\perp$ . Thus  $W^\perp \subset U^\perp$ . ■

Recall that if  $U, W$  are subspaces of  $V$ , then  $V$  is the direct sum of  $U$  and  $W$  (written  $V = U \oplus W$ ) if each element of  $V$  can be written in exactly one way as a vector in  $U$  plus a vector in  $W$  (see 1.40).

The next result shows that every finite-dimensional subspace of  $V$  leads to a natural direct sum decomposition of  $V$ .

#### 6.47 Direct sum of a subspace and its orthogonal complement

Suppose  $U$  is a finite-dimensional subspace of  $V$ . Then

$$V = U \oplus U^\perp.$$

**Proof** First we will show that

$$\mathbf{6.48} \quad V = U + U^\perp.$$

To do this, suppose  $v \in V$ . Let  $e_1, \dots, e_m$  be an orthonormal basis of  $U$ . Obviously

$$\mathbf{6.49} \quad v = \underbrace{\langle v, e_1 \rangle e_1 + \cdots + \langle v, e_m \rangle e_m}_u + \underbrace{v - \langle v, e_1 \rangle e_1 - \cdots - \langle v, e_m \rangle e_m}_w.$$

Let  $u$  and  $w$  be defined as in the equation above. Clearly  $u \in U$ . Because  $e_1, \dots, e_m$  is an orthonormal list, for each  $j = 1, \dots, m$  we have

$$\begin{aligned} \langle w, e_j \rangle &= \langle v, e_j \rangle - \langle v, e_j \rangle \\ &= 0. \end{aligned}$$

Thus  $w$  is orthogonal to every vector in  $\text{span}(e_1, \dots, e_m)$ . In other words,  $w \in U^\perp$ . Thus we have written  $v = u + w$ , where  $u \in U$  and  $w \in U^\perp$ , completing the proof of 6.48.

From 6.46(d), we know that  $U \cap U^\perp = \{0\}$ . Along with 6.48, this implies that  $V = U \oplus U^\perp$  (see 1.45). ■

Now we can see how to compute  $\dim U^\perp$  from  $\dim U$ .

### 6.50 Dimension of the orthogonal complement

Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Then

$$\dim U^\perp = \dim V - \dim U.$$

**Proof** The formula for  $\dim U^\perp$  follows immediately from 6.47 and 3.78. ■

The next result is an important consequence of 6.47.

### 6.51 The orthogonal complement of the orthogonal complement

Suppose  $U$  is a finite-dimensional subspace of  $V$ . Then

$$U = (U^\perp)^\perp.$$

**Proof** First we will show that

$$6.52 \quad U \subset (U^\perp)^\perp.$$

To do this, suppose  $u \in U$ . Then  $\langle u, v \rangle = 0$  for every  $v \in U^\perp$  (by the definition of  $U^\perp$ ). Because  $u$  is orthogonal to every vector in  $U^\perp$ , we have  $u \in (U^\perp)^\perp$ , completing the proof of 6.52.

To prove the inclusion in the other direction, suppose  $v \in (U^\perp)^\perp$ . By 6.47, we can write  $v = u + w$ , where  $u \in U$  and  $w \in U^\perp$ . We have  $v - u = w \in U^\perp$ . Because  $v \in (U^\perp)^\perp$  and  $u \in (U^\perp)^\perp$  (from 6.52), we have  $v - u \in (U^\perp)^\perp$ . Thus  $v - u \in U^\perp \cap (U^\perp)^\perp$ , which implies that  $v - u$  is orthogonal to itself, which implies that  $v - u = 0$ , which implies that  $v = u$ , which implies that  $v \in U$ . Thus  $(U^\perp)^\perp \subset U$ , which along with 6.52 completes the proof. ■

We now define an operator  $P_U$  for each finite-dimensional subspace of  $V$ .

### 6.53 Definition *orthogonal projection*, $P_U$

Suppose  $U$  is a finite-dimensional subspace of  $V$ . The *orthogonal projection* of  $V$  onto  $U$  is the operator  $P_U \in \mathcal{L}(V)$  defined as follows: For  $v \in V$ , write  $v = u + w$ , where  $u \in U$  and  $w \in U^\perp$ . Then  $P_U v = u$ .

The direct sum decomposition  $V = U \oplus U^\perp$  given by 6.47 shows that each  $v \in V$  can be uniquely written in the form  $v = u + w$  with  $u \in U$  and  $w \in U^\perp$ . Thus  $P_U v$  is well defined.

**6.54 Example** Suppose  $x \in V$  with  $x \neq 0$  and  $U = \text{span}(x)$ . Show that

$$P_U v = \frac{\langle v, x \rangle}{\|x\|^2} x$$

for every  $v \in V$ .

**Solution** Suppose  $v \in V$ . Then

$$v = \frac{\langle v, x \rangle}{\|x\|^2} x + \left( v - \frac{\langle v, x \rangle}{\|x\|^2} x \right),$$

where the first term on the right is in  $\text{span}(x)$  (and thus in  $U$ ) and the second term on the right is orthogonal to  $x$  (and thus is in  $U^\perp$ ). Thus  $P_U v$  equals the first term on the right, as desired.

### 6.55 Properties of the orthogonal projection $P_U$

Suppose  $U$  is a finite-dimensional subspace of  $V$  and  $v \in V$ . Then

- (a)  $P_U \in \mathcal{L}(V)$ ;
- (b)  $P_U u = u$  for every  $u \in U$ ;
- (c)  $P_U w = 0$  for every  $w \in U^\perp$ ;
- (d)  $\text{range } P_U = U$ ;
- (e)  $\text{null } P_U = U^\perp$ ;
- (f)  $v - P_U v \in U^\perp$ ;
- (g)  $P_U^2 = P_U$ ;
- (h)  $\|P_U v\| \leq \|v\|$ ;
- (i) for every orthonormal basis  $e_1, \dots, e_m$  of  $U$ ,

$$P_U v = \langle v, e_1 \rangle e_1 + \cdots + \langle v, e_m \rangle e_m.$$

**Proof**

- (a) To show that  $P_U$  is a linear map on  $V$ , suppose  $v_1, v_2 \in V$ . Write

$$v_1 = u_1 + w_1 \quad \text{and} \quad v_2 = u_2 + w_2$$

with  $u_1, u_2 \in U$  and  $w_1, w_2 \in U^\perp$ . Thus  $P_U v_1 = u_1$  and  $P_U v_2 = u_2$ . Now

$$v_1 + v_2 = (u_1 + u_2) + (w_1 + w_2),$$

where  $u_1 + u_2 \in U$  and  $w_1 + w_2 \in U^\perp$ . Thus

$$P_U(v_1 + v_2) = u_1 + u_2 = P_U v_1 + P_U v_2.$$

Similarly, suppose  $\lambda \in \mathbf{F}$ . The equation  $v = u + w$  with  $u \in U$  and  $w \in U^\perp$  implies that  $\lambda v = \lambda u + \lambda w$  with  $\lambda u \in U$  and  $\lambda w \in U^\perp$ . Thus  $P_U(\lambda v) = \lambda u = \lambda P_U v$ .

Hence  $P_U$  is a linear map from  $V$  to  $V$ .

- (b) Suppose  $u \in U$ . We can write  $u = u + 0$ , where  $u \in U$  and  $0 \in U^\perp$ . Thus  $P_U u = u$ .
- (c) Suppose  $w \in U^\perp$ . We can write  $w = 0 + w$ , where  $0 \in U$  and  $w \in U^\perp$ . Thus  $P_U w = 0$ .
- (d) The definition of  $P_U$  implies that  $\text{range } P_U \subset U$ . Part (b) implies that  $U \subset \text{range } P_U$ . Thus  $\text{range } P_U = U$ .
- (e) Part (c) implies that  $U^\perp \subset \text{null } P_U$ . To prove the inclusion in the other direction, note that if  $v \in \text{null } P_U$  then the decomposition given by 6.47 must be  $v = 0 + v$ , where  $0 \in U$  and  $v \in U^\perp$ . Thus  $\text{null } P_U \subset U^\perp$ .
- (f) If  $v = u + w$  with  $u \in U$  and  $w \in U^\perp$ , then

$$v - P_U v = v - u = w \in U^\perp.$$

- (g) If  $v = u + w$  with  $u \in U$  and  $w \in U^\perp$ , then

$$(P_U^2)v = P_U(P_U v) = P_U u = u = P_U v.$$

- (h) If  $v = u + w$  with  $u \in U$  and  $w \in U^\perp$ , then

$$\|P_U v\|^2 = \|u\|^2 \leq \|u\|^2 + \|w\|^2 = \|v\|^2,$$

where the last equality comes from the Pythagorean Theorem.

- (i) The formula for  $P_U v$  follows from equation 6.49 in the proof of 6.47. ■

## Minimization Problems

*The remarkable simplicity of the solution to this minimization problem has led to many important applications of inner product spaces outside of pure mathematics.*

The following problem often arises: given a subspace  $U$  of  $V$  and a point  $v \in V$ , find a point  $u \in U$  such that  $\|v - u\|$  is as small as possible. The next proposition shows that this minimization problem is solved by taking  $u = P_U v$ .

### 6.56 Minimizing the distance to a subspace

Suppose  $U$  is a finite-dimensional subspace of  $V$ ,  $v \in V$ , and  $u \in U$ . Then

$$\|v - P_U v\| \leq \|v - u\|.$$

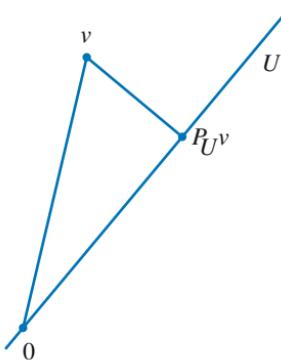
Furthermore, the inequality above is an equality if and only if  $u = P_U v$ .

**Proof** We have

$$\begin{aligned} 6.57 \quad \|v - P_U v\|^2 &\leq \|v - P_U v\|^2 + \|P_U v - u\|^2 \\ &= \|(v - P_U v) + (P_U v - u)\|^2 \\ &= \|v - u\|^2, \end{aligned}$$

where the first line above holds because  $0 \leq \|P_U v - u\|^2$ , the second line above comes from the Pythagorean Theorem [which applies because  $v - P_U v \in U^\perp$  by 6.55(f), and  $P_U v - u \in U$ ], and the third line above holds by simple algebra. Taking square roots gives the desired inequality.

Our inequality above is an equality if and only if 6.57 is an equality, which happens if and only if  $\|P_U v - u\| = 0$ , which happens if and only if  $u = P_U v$ . ■



$P_U v$  is the closest point in  $U$  to  $v$ .

The last result is often combined with the formula 6.55(i) to compute explicit solutions to minimization problems.

**6.58 Example** Find a polynomial  $u$  with real coefficients and degree at most 5 that approximates  $\sin x$  as well as possible on the interval  $[-\pi, \pi]$ , in the sense that

$$\int_{-\pi}^{\pi} |\sin x - u(x)|^2 dx$$

is as small as possible. Compare this result to the Taylor series approximation.

**Solution** Let  $C_{\mathbf{R}}[-\pi, \pi]$  denote the real inner product space of continuous real-valued functions on  $[-\pi, \pi]$  with inner product

**6.59** 
$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)g(x) dx.$$

Let  $v \in C_{\mathbf{R}}[-\pi, \pi]$  be the function defined by  $v(x) = \sin x$ . Let  $U$  denote the subspace of  $C_{\mathbf{R}}[-\pi, \pi]$  consisting of the polynomials with real coefficients and degree at most 5. Our problem can now be reformulated as follows:

Find  $u \in U$  such that  $\|v - u\|$  is as small as possible.

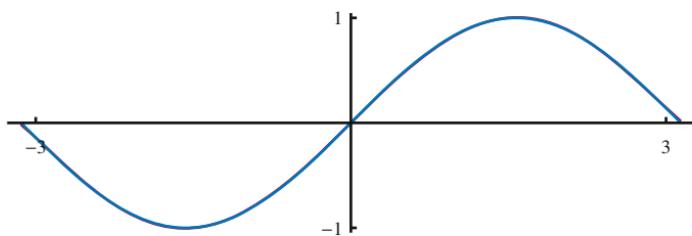
To compute the solution to our approximation problem, first apply the Gram–Schmidt Procedure (using the inner product given by 6.59) to the basis  $1, x, x^2, x^3, x^4, x^5$  of  $U$ , producing an orthonormal basis  $e_1, e_2, e_3, e_4, e_5, e_6$  of  $U$ . Then, again using the inner product given by 6.59, compute  $P_U v$  using 6.55(i) (with  $m = 6$ ). Doing this computation shows that  $P_U v$  is the function  $u$  defined by

A computer that can perform integrations is useful here.

**6.60** 
$$u(x) = 0.987862x - 0.155271x^3 + 0.00564312x^5,$$

where the  $\pi$ 's that appear in the exact answer have been replaced with a good decimal approximation.

By 6.56, the polynomial  $u$  above is the best approximation to  $\sin x$  on  $[-\pi, \pi]$  using polynomials of degree at most 5 (here “best approximation” means in the sense of minimizing  $\int_{-\pi}^{\pi} |\sin x - u(x)|^2 dx$ ). To see how good this approximation is, the next figure shows the graphs of both  $\sin x$  and our approximation  $u(x)$  given by 6.60 over the interval  $[-\pi, \pi]$ .



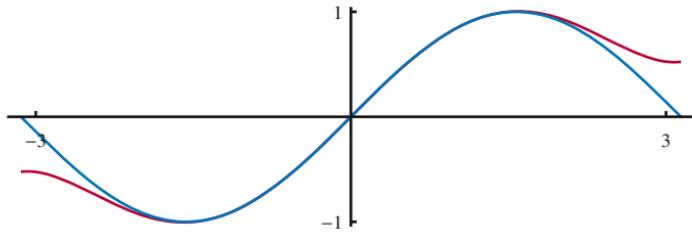
*Graphs on  $[-\pi, \pi]$  of  $\sin x$  (blue) and its approximation  $u(x)$  (red) given by 6.60.*

Our approximation 6.60 is so accurate that the two graphs are almost identical—our eyes may see only one graph! Here the blue graph is placed almost exactly over the red graph. If you are viewing this on an electronic device, try enlarging the picture above, especially near 3 or  $-3$ , to see a small gap between the two graphs.

Another well-known approximation to  $\sin x$  by a polynomial of degree 5 is given by the Taylor polynomial

$$\mathbf{6.61} \quad x - \frac{x^3}{3!} + \frac{x^5}{5!}.$$

To see how good this approximation is, the next picture shows the graphs of both  $\sin x$  and the Taylor polynomial 6.61 over the interval  $[-\pi, \pi]$ .



*Graphs on  $[-\pi, \pi]$  of  $\sin x$  (blue) and the Taylor polynomial 6.61 (red).*

The Taylor polynomial is an excellent approximation to  $\sin x$  for  $x$  near 0. But the picture above shows that for  $|x| > 2$ , the Taylor polynomial is not so accurate, especially compared to 6.60. For example, taking  $x = 3$ , our approximation 6.60 estimates  $\sin 3$  with an error of about 0.001, but the Taylor series 6.61 estimates  $\sin 3$  with an error of about 0.4. Thus at  $x = 3$ , the error in the Taylor series is hundreds of times larger than the error given by 6.60. Linear algebra has helped us discover an approximation to  $\sin x$  that improves upon what we learned in calculus!

## EXERCISES 6.C

- 1 Suppose  $v_1, \dots, v_m \in V$ . Prove that

$$\{v_1, \dots, v_m\}^\perp = (\text{span}(v_1, \dots, v_m))^\perp.$$

- 2 Suppose  $U$  is a finite-dimensional subspace of  $V$ . Prove that  $U^\perp = \{0\}$  if and only if  $U = V$ .

[Exercise 14(a) shows that the result above is not true without the hypothesis that  $U$  is finite-dimensional.]

- 3 Suppose  $U$  is a subspace of  $V$  with basis  $u_1, \dots, u_m$  and

$$u_1, \dots, u_m, w_1, \dots, w_n$$

is a basis of  $V$ . Prove that if the Gram–Schmidt Procedure is applied to the basis of  $V$  above, producing a list  $e_1, \dots, e_m, f_1, \dots, f_n$ , then  $e_1, \dots, e_m$  is an orthonormal basis of  $U$  and  $f_1, \dots, f_n$  is an orthonormal basis of  $U^\perp$ .

- 4 Suppose  $U$  is the subspace of  $\mathbf{R}^4$  defined by

$$U = \text{span}((1, 2, 3, -4), (-5, 4, 3, 2)).$$

Find an orthonormal basis of  $U$  and an orthonormal basis of  $U^\perp$ .

- 5 Suppose  $V$  is finite-dimensional and  $U$  is a subspace of  $V$ . Show that  $P_{U^\perp} = I - P_U$ , where  $I$  is the identity operator on  $V$ .

- 6 Suppose  $U$  and  $W$  are finite-dimensional subspaces of  $V$ . Prove that  $P_U P_W = 0$  if and only if  $\langle u, w \rangle = 0$  for all  $u \in U$  and all  $w \in W$ .

- 7 Suppose  $V$  is finite-dimensional and  $P \in \mathcal{L}(V)$  is such that  $P^2 = P$  and every vector in null  $P$  is orthogonal to every vector in range  $P$ . Prove that there exists a subspace  $U$  of  $V$  such that  $P = P_U$ .

- 8 Suppose  $V$  is finite-dimensional and  $P \in \mathcal{L}(V)$  is such that  $P^2 = P$  and

$$\|Pv\| \leq \|v\|$$

for every  $v \in V$ . Prove that there exists a subspace  $U$  of  $V$  such that  $P = P_U$ .

- 9 Suppose  $T \in \mathcal{L}(V)$  and  $U$  is a finite-dimensional subspace of  $V$ . Prove that  $U$  is invariant under  $T$  if and only if  $P_U T P_U = T P_U$ .

- 10** Suppose  $V$  is finite-dimensional,  $T \in \mathcal{L}(V)$ , and  $U$  is a subspace of  $V$ . Prove that  $U$  and  $U^\perp$  are both invariant under  $T$  if and only if  $P_U T = T P_U$ .

- 11** In  $\mathbf{R}^4$ , let

$$U = \text{span}((1, 1, 0, 0), (1, 1, 1, 2)).$$

Find  $u \in U$  such that  $\|u - (1, 2, 3, 4)\|$  is as small as possible.

- 12** Find  $p \in \mathcal{P}_3(\mathbf{R})$  such that  $p(0) = 0$ ,  $p'(0) = 0$ , and

$$\int_0^1 |2 + 3x - p(x)|^2 dx$$

is as small as possible.

- 13** Find  $p \in \mathcal{P}_5(\mathbf{R})$  that makes

$$\int_{-\pi}^{\pi} |\sin x - p(x)|^2 dx$$

as small as possible.

[The polynomial 6.60 is an excellent approximation to the answer to this exercise, but here you are asked to find the exact solution, which involves powers of  $\pi$ . A computer that can perform symbolic integration will be useful.]

- 14** Suppose  $C_{\mathbf{R}}([-1, 1])$  is the vector space of continuous real-valued functions on the interval  $[-1, 1]$  with inner product given by

$$\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx$$

for  $f, g \in C_{\mathbf{R}}([-1, 1])$ . Let  $U$  be the subspace of  $C_{\mathbf{R}}([-1, 1])$  defined by

$$U = \{f \in C_{\mathbf{R}}([-1, 1]) : f(0) = 0\}.$$

- (a) Show that  $U^\perp = \{0\}$ .
- (b) Show that 6.47 and 6.51 do not hold without the finite-dimensional hypothesis.



Isaac Newton (1642–1727), as envisioned by British poet and artist William Blake in this 1795 painting.

# Operators on Inner Product Spaces

The deepest results related to inner product spaces deal with the subject to which we now turn—operators on inner product spaces. By exploiting properties of the adjoint, we will develop a detailed description of several important classes of operators on inner product spaces.

A new assumption for this chapter is listed in the second bullet point below:

## 7.1 Notation $\mathbf{F}, V$

- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  and  $W$  denote finite-dimensional inner product spaces over  $\mathbf{F}$ .

### LEARNING OBJECTIVES FOR THIS CHAPTER

- adjoint
- Spectral Theorem
- positive operators
- isometries
- Polar Decomposition
- Singular Value Decomposition

## 7.A Self-Adjoint and Normal Operators

### Adjoints

#### 7.2 Definition adjoint, $T^*$

Suppose  $T \in \mathcal{L}(V, W)$ . The **adjoint** of  $T$  is the function  $T^*: W \rightarrow V$  such that

$$\langle Tv, w \rangle = \langle v, T^*w \rangle$$

for every  $v \in V$  and every  $w \in W$ .

The word **adjoint** has another meaning in linear algebra. We do not need the second meaning in this book. In case you encounter the second meaning for adjoint elsewhere, be warned that the two meanings for adjoint are unrelated to each other.

To see why the definition above makes sense, suppose  $T \in \mathcal{L}(V, W)$ . Fix  $w \in W$ . Consider the linear functional on  $V$  that maps  $v \in V$  to  $\langle Tv, w \rangle$ ; this linear functional depends on  $T$  and  $w$ . By the Riesz Representation Theorem (6.42), there exists a unique vector in  $V$  such that this linear functional is

given by taking the inner product with it. We call this unique vector  $T^*w$ . In other words,  $T^*w$  is the unique vector in  $V$  such that  $\langle Tv, w \rangle = \langle v, T^*w \rangle$  for every  $v \in V$ .

#### 7.3 Example Define $T: \mathbf{R}^3 \rightarrow \mathbf{R}^2$ by

$$T(x_1, x_2, x_3) = (x_2 + 3x_3, 2x_1).$$

Find a formula for  $T^*$ .

**Solution** Here  $T^*$  will be a function from  $\mathbf{R}^2$  to  $\mathbf{R}^3$ . To compute  $T^*$ , fix a point  $(y_1, y_2) \in \mathbf{R}^2$ . Then for every  $(x_1, x_2, x_3) \in \mathbf{R}^3$  we have

$$\begin{aligned} \langle (x_1, x_2, x_3), T^*(y_1, y_2) \rangle &= \langle T(x_1, x_2, x_3), (y_1, y_2) \rangle \\ &= \langle (x_2 + 3x_3, 2x_1), (y_1, y_2) \rangle \\ &= x_2 y_1 + 3x_3 y_1 + 2x_1 y_2 \\ &= \langle (x_1, x_2, x_3), (2y_2, y_1, 3y_1) \rangle. \end{aligned}$$

Thus

$$T^*(y_1, y_2) = (2y_2, y_1, 3y_1).$$

**7.4 Example** Fix  $u \in V$  and  $x \in W$ . Define  $T \in \mathcal{L}(V, W)$  by

$$Tv = \langle v, u \rangle x$$

for every  $v \in V$ . Find a formula for  $T^*$ .

**Solution** Fix  $w \in W$ . Then for every  $v \in V$  we have

$$\begin{aligned} \langle v, T^*w \rangle &= \langle Tv, w \rangle \\ &= \langle \langle v, u \rangle x, w \rangle \\ &= \langle v, u \rangle \langle x, w \rangle \\ &= \langle v, \langle w, x \rangle u \rangle. \end{aligned}$$

Thus

$$T^*w = \langle w, x \rangle u.$$

In the two examples above,  $T^*$  turned out to be not just a function but a linear map. This is true in general, as shown by the next result.

The proofs of the next two results use a common technique: flip  $T^*$  from one side of an inner product to become  $T$  on the other side.

## 7.5 The adjoint is a linear map

If  $T \in \mathcal{L}(V, W)$ , then  $T^* \in \mathcal{L}(W, V)$ .

**Proof** Suppose  $T \in \mathcal{L}(V, W)$ . Fix  $w_1, w_2 \in W$ . If  $v \in V$ , then

$$\begin{aligned} \langle v, T^*(w_1 + w_2) \rangle &= \langle Tv, w_1 + w_2 \rangle \\ &= \langle Tv, w_1 \rangle + \langle Tv, w_2 \rangle \\ &= \langle v, T^*w_1 \rangle + \langle v, T^*w_2 \rangle \\ &= \langle v, T^*w_1 + T^*w_2 \rangle, \end{aligned}$$

which shows that  $T^*(w_1 + w_2) = T^*w_1 + T^*w_2$ .

Fix  $w \in W$  and  $\lambda \in \mathbf{F}$ . If  $v \in V$ , then

$$\begin{aligned} \langle v, T^*(\lambda w) \rangle &= \langle Tv, \lambda w \rangle \\ &= \bar{\lambda} \langle Tv, w \rangle \\ &= \bar{\lambda} \langle v, T^*w \rangle \\ &= \langle v, \lambda T^*w \rangle, \end{aligned}$$

which shows that  $T^*(\lambda w) = \lambda T^*w$ .

Thus  $T^*$  is a linear map, as desired. ■

## 7.6 Properties of the adjoint

- (a)  $(S + T)^* = S^* + T^*$  for all  $S, T \in \mathcal{L}(V, W)$ ;
- (b)  $(\lambda T)^* = \bar{\lambda}T^*$  for all  $\lambda \in \mathbf{F}$  and  $T \in \mathcal{L}(V, W)$ ;
- (c)  $(T^*)^* = T$  for all  $T \in \mathcal{L}(V, W)$ ;
- (d)  $I^* = I$ , where  $I$  is the identity operator on  $V$ ;
- (e)  $(ST)^* = T^*S^*$  for all  $T \in \mathcal{L}(V, W)$  and  $S \in \mathcal{L}(W, U)$  (here  $U$  is an inner product space over  $\mathbf{F}$ ).

### Proof

- (a) Suppose  $S, T \in \mathcal{L}(V, W)$ . If  $v \in V$  and  $w \in W$ , then

$$\begin{aligned}\langle v, (S + T)^*w \rangle &= \langle (S + T)v, w \rangle \\ &= \langle Sv, w \rangle + \langle Tv, w \rangle \\ &= \langle v, S^*w \rangle + \langle v, T^*w \rangle \\ &= \langle v, S^*w + T^*w \rangle.\end{aligned}$$

Thus  $(S + T)^*w = S^*w + T^*w$ , as desired.

- (b) Suppose  $\lambda \in \mathbf{F}$  and  $T \in \mathcal{L}(V, W)$ . If  $v \in V$  and  $w \in W$ , then

$$\langle v, (\lambda T)^*w \rangle = \langle \lambda T v, w \rangle = \lambda \langle T v, w \rangle = \lambda \langle v, T^*w \rangle = \langle v, \bar{\lambda}T^*w \rangle.$$

Thus  $(\lambda T)^*w = \bar{\lambda}T^*w$ , as desired.

- (c) Suppose  $T \in \mathcal{L}(V, W)$ . If  $v \in V$  and  $w \in W$ , then

$$\langle w, (T^*)^*v \rangle = \langle T^*w, v \rangle = \overline{\langle v, T^*w \rangle} = \overline{\langle T v, w \rangle} = \langle w, T v \rangle.$$

Thus  $(T^*)^*v = T v$ , as desired.

- (d) If  $v, u \in V$ , then

$$\langle v, I^*u \rangle = \langle Iv, u \rangle = \langle v, u \rangle.$$

Thus  $I^*u = u$ , as desired.

- (e) Suppose  $T \in \mathcal{L}(V, W)$  and  $S \in \mathcal{L}(W, U)$ . If  $v \in V$  and  $u \in U$ , then

$$\langle v, (ST)^*u \rangle = \langle STv, u \rangle = \langle Tv, S^*u \rangle = \langle v, T^*(S^*u) \rangle.$$

Thus  $(ST)^*u = T^*(S^*u)$ , as desired. ■

The next result shows the relationship between the null space and the range of a linear map and its adjoint. The symbol  $\iff$  used in the proof means “if and only if”; this symbol could also be read to mean “is equivalent to”.

### 7.7 Null space and range of $T^*$

Suppose  $T \in \mathcal{L}(V, W)$ . Then

- (a)  $\text{null } T^* = (\text{range } T)^\perp$ ;
- (b)  $\text{range } T^* = (\text{null } T)^\perp$ ;
- (c)  $\text{null } T = (\text{range } T^*)^\perp$ ;
- (d)  $\text{range } T = (\text{null } T^*)^\perp$ .

**Proof** We begin by proving (a). Let  $w \in W$ . Then

$$\begin{aligned} w \in \text{null } T^* &\iff T^*w = 0 \\ &\iff \langle v, T^*w \rangle = 0 \text{ for all } v \in V \\ &\iff \langle Tv, w \rangle = 0 \text{ for all } v \in V \\ &\iff w \in (\text{range } T)^\perp. \end{aligned}$$

Thus  $\text{null } T^* = (\text{range } T)^\perp$ , proving (a).

If we take the orthogonal complement of both sides of (a), we get (d), where we have used 6.51. Replacing  $T$  with  $T^*$  in (a) gives (c), where we have used 7.6(c). Finally, replacing  $T$  with  $T^*$  in (d) gives (b). ■

### 7.8 Definition conjugate transpose

The **conjugate transpose** of an  $m$ -by- $n$  matrix is the  $n$ -by- $m$  matrix obtained by interchanging the rows and columns and then taking the complex conjugate of each entry.

### 7.9 Example

The conjugate transpose of the matrix  $\begin{pmatrix} 2 & 3+4i & 7 \\ 6 & 5 & 8i \end{pmatrix}$  is the matrix

$$\begin{pmatrix} 2 & 6 \\ 3-4i & 5 \\ 7 & -8i \end{pmatrix}.$$

If  $\mathbf{F} = \mathbf{R}$ , then the conjugate transpose of a matrix is the same as its **transpose**, which is the matrix obtained by interchanging the rows and columns.

*The adjoint of a linear map does not depend on a choice of basis. This explains why this book emphasizes adjoints of linear maps instead of conjugate transposes of matrices.*

The next result shows how to compute the matrix of  $T^*$  from the matrix of  $T$ .

**Caution:** Remember that the result below applies only when we are dealing with orthonormal bases. With respect to nonorthonormal bases, the matrix of  $T^*$  does not necessarily equal the conjugate transpose of the matrix of  $T$ .

### 7.10 The matrix of $T^*$

Let  $T \in \mathcal{L}(V, W)$ . Suppose  $e_1, \dots, e_n$  is an orthonormal basis of  $V$  and  $f_1, \dots, f_m$  is an orthonormal basis of  $W$ . Then

$$\mathcal{M}(T^*, (f_1, \dots, f_m), (e_1, \dots, e_n))$$

is the conjugate transpose of

$$\mathcal{M}(T, (e_1, \dots, e_n), (f_1, \dots, f_m)).$$

**Proof** In this proof, we will write  $\mathcal{M}(T)$  instead of the longer expression  $\mathcal{M}(T, (e_1, \dots, e_n), (f_1, \dots, f_m))$ ; we will also write  $\mathcal{M}(T^*)$  instead of  $\mathcal{M}(T^*, (f_1, \dots, f_m), (e_1, \dots, e_n))$ .

Recall that we obtain the  $k^{\text{th}}$  column of  $\mathcal{M}(T)$  by writing  $Te_k$  as a linear combination of the  $f_j$ 's; the scalars used in this linear combination then become the  $k^{\text{th}}$  column of  $\mathcal{M}(T)$ . Because  $f_1, \dots, f_m$  is an orthonormal basis of  $W$ , we know how to write  $Te_k$  as a linear combination of the  $f_j$ 's (see 6.30):

$$Te_k = \langle Te_k, f_1 \rangle f_1 + \cdots + \langle Te_k, f_m \rangle f_m.$$

Thus the entry in row  $j$ , column  $k$ , of  $\mathcal{M}(T)$  is  $\langle Te_k, f_j \rangle$ .

Replacing  $T$  with  $T^*$  and interchanging the roles played by the  $e$ 's and  $f$ 's, we see that the entry in row  $j$ , column  $k$ , of  $\mathcal{M}(T^*)$  is  $\langle T^* f_k, e_j \rangle$ , which equals  $\langle f_k, T e_j \rangle$ , which equals  $\overline{\langle T e_j, f_k \rangle}$ , which equals the complex conjugate of the entry in row  $k$ , column  $j$ , of  $\mathcal{M}(T)$ . In other words,  $\mathcal{M}(T^*)$  is the conjugate transpose of  $\mathcal{M}(T)$ . ■

## Self-Adjoint Operators

Now we switch our attention to operators on inner product spaces. Thus instead of considering linear maps from  $V$  to  $W$ , we will be focusing on linear maps from  $V$  to  $V$ ; recall that such linear maps are called operators.

### 7.11 Definition self-adjoint

An operator  $T \in \mathcal{L}(V)$  is called **self-adjoint** if  $T = T^*$ . In other words,  $T \in \mathcal{L}(V)$  is self-adjoint if and only if

$$\langle Tv, w \rangle = \langle v, Tw \rangle$$

for all  $v, w \in V$ .

**7.12 Example** Suppose  $T$  is the operator on  $\mathbf{F}^2$  whose matrix (with respect to the standard basis) is

$$\begin{pmatrix} 2 & b \\ 3 & 7 \end{pmatrix}.$$

Find all numbers  $b$  such that  $T$  is self-adjoint.

**Solution** The operator  $T$  is self-adjoint if and only if  $b = 3$  (because  $\mathcal{M}(T) = \mathcal{M}(T^*)$  if and only if  $b = 3$ ; recall that  $\mathcal{M}(T^*)$  is the conjugate transpose of  $\mathcal{M}(T)$ —see 7.10).

You should verify that the sum of two self-adjoint operators is self-adjoint and that the product of a real scalar and a self-adjoint operator is self-adjoint.

A good analogy to keep in mind (especially when  $\mathbf{F} = \mathbf{C}$ ) is that the adjoint on  $\mathcal{L}(V)$  plays a role similar to complex conjugation on  $\mathbf{C}$ . A complex number  $z$  is real if and only if  $z = \bar{z}$ ; thus a self-adjoint operator ( $T = T^*$ ) is analogous to a real number.

Some mathematicians use the term **Hermitian** instead of self-adjoint, honoring French mathematician Charles Hermite, who in 1873 published the first proof that  $e$  is not a zero of any polynomial with integer coefficients.

We will see that the analogy discussed above is reflected in some important properties of self-adjoint operators, beginning with eigenvalues in the next result.

If  $\mathbf{F} = \mathbf{R}$ , then by definition every eigenvalue is real, so the next result is interesting only when  $\mathbf{F} = \mathbf{C}$ .

### 7.13 Eigenvalues of self-adjoint operators are real

Every eigenvalue of a self-adjoint operator is real.

**Proof** Suppose  $T$  is a self-adjoint operator on  $V$ . Let  $\lambda$  be an eigenvalue of  $T$ , and let  $v$  be a nonzero vector in  $V$  such that  $Tv = \lambda v$ . Then

$$\lambda \|v\|^2 = \langle \lambda v, v \rangle = \langle Tv, v \rangle = \langle v, Tv \rangle = \langle v, \lambda v \rangle = \bar{\lambda} \|v\|^2.$$

Thus  $\lambda = \bar{\lambda}$ , which means that  $\lambda$  is real, as desired. ■

The next result is false for real inner product spaces. As an example, consider the operator  $T \in \mathcal{L}(\mathbf{R}^2)$  that is a counterclockwise rotation of  $90^\circ$  around the origin; thus  $T(x, y) = (-y, x)$ . Obviously  $Tv$  is orthogonal to  $v$  for every  $v \in \mathbf{R}^2$ , even though  $T \neq 0$ .

### 7.14 Over $\mathbf{C}$ , $Tv$ is orthogonal to $v$ for all $v$ only for the 0 operator

Suppose  $V$  is a complex inner product space and  $T \in \mathcal{L}(V)$ . Suppose

$$\langle Tv, v \rangle = 0$$

for all  $v \in V$ . Then  $T = 0$ .

**Proof** We have

$$\begin{aligned} \langle Tu, w \rangle &= \frac{\langle T(u + w), u + w \rangle - \langle T(u - w), u - w \rangle}{4} \\ &\quad + \frac{\langle T(u + iw), u + iw \rangle - \langle T(u - iw), u - iw \rangle}{4} i \end{aligned}$$

for all  $u, w \in V$ , as can be verified by computing the right side. Note that each term on the right side is of the form  $\langle Tv, v \rangle$  for appropriate  $v \in V$ . Thus our hypothesis implies that  $\langle Tu, w \rangle = 0$  for all  $u, w \in V$ . This implies that  $T = 0$  (take  $w = Tu$ ). ■

*The next result provides another example of how self-adjoint operators behave like real numbers.*

The next result is false for real inner product spaces, as shown by considering any operator on a real inner product space that is not self-adjoint.

### 7.15 Over $\mathbf{C}$ , $\langle Tv, v \rangle$ is real for all $v$ only for self-adjoint operators

Suppose  $V$  is a complex inner product space and  $T \in \mathcal{L}(V)$ . Then  $T$  is self-adjoint if and only if

$$\langle Tv, v \rangle \in \mathbf{R}$$

for every  $v \in V$ .

**Proof** Let  $v \in V$ . Then

$$\langle Tv, v \rangle - \overline{\langle Tv, v \rangle} = \langle Tv, v \rangle - \langle v, Tv \rangle = \langle Tv, v \rangle - \langle T^*v, v \rangle = \langle (T - T^*)v, v \rangle.$$

If  $\langle Tv, v \rangle \in \mathbf{R}$  for every  $v \in V$ , then the left side of the equation above equals 0, so  $\langle (T - T^*)v, v \rangle = 0$  for every  $v \in V$ . This implies that  $T - T^* = 0$  (by 7.14). Hence  $T$  is self-adjoint.

Conversely, if  $T$  is self-adjoint, then the right side of the equation above equals 0, so  $\langle Tv, v \rangle = \overline{\langle Tv, v \rangle}$  for every  $v \in V$ . This implies that  $\langle Tv, v \rangle \in \mathbf{R}$  for every  $v \in V$ , as desired. ■

On a real inner product space  $V$ , a nonzero operator  $T$  might satisfy  $\langle Tv, v \rangle = 0$  for all  $v \in V$ . However, the next result shows that this cannot happen for a self-adjoint operator.

### 7.16 If $T = T^*$ and $\langle Tv, v \rangle = 0$ for all $v$ , then $T = 0$

Suppose  $T$  is a self-adjoint operator on  $V$  such that

$$\langle Tv, v \rangle = 0$$

for all  $v \in V$ . Then  $T = 0$ .

**Proof** We have already proved this (without the hypothesis that  $T$  is self-adjoint) when  $V$  is a complex inner product space (see 7.14). Thus we can assume that  $V$  is a real inner product space. If  $u, w \in V$ , then

$$7.17 \quad \langle Tu, w \rangle = \frac{\langle T(u + w), u + w \rangle - \langle T(u - w), u - w \rangle}{4};$$

this is proved by computing the right side using the equation

$$\langle Tw, u \rangle = \langle w, Tu \rangle = \langle Tu, w \rangle,$$

where the first equality holds because  $T$  is self-adjoint and the second equality holds because we are working in a real inner product space.

Each term on the right side of 7.17 is of the form  $\langle Tv, v \rangle$  for appropriate  $v$ . Thus  $\langle Tu, w \rangle = 0$  for all  $u, w \in V$ . This implies that  $T = 0$  (take  $w = Tu$ ). ■

## Normal Operators

### 7.18 Definition *normal*

- An operator on an inner product space is called ***normal*** if it commutes with its adjoint.
- In other words,  $T \in \mathcal{L}(V)$  is normal if

$$TT^* = T^*T.$$

Obviously every self-adjoint operator is normal, because if  $T$  is self-adjoint then  $T^* = T$ .

---

**7.19 Example** Let  $T$  be the operator on  $\mathbf{F}^2$  whose matrix (with respect to the standard basis) is

$$\begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix}.$$

Show that  $T$  is not self-adjoint and that  $T$  is normal.

**Solution** This operator is not self-adjoint because the entry in row 2, column 1 (which equals 3) does not equal the complex conjugate of the entry in row 1, column 2 (which equals  $-3$ ).

The matrix of  $TT^*$  equals

$$\begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 2 & 3 \\ -3 & 2 \end{pmatrix}, \text{ which equals } \begin{pmatrix} 13 & 0 \\ 0 & 13 \end{pmatrix}.$$

Similarly, the matrix of  $T^*T$  equals

$$\begin{pmatrix} 2 & 3 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix}, \text{ which equals } \begin{pmatrix} 13 & 0 \\ 0 & 13 \end{pmatrix}.$$

Because  $TT^*$  and  $T^*T$  have the same matrix, we see that  $TT^* = T^*T$ . Thus  $T$  is normal.

---

*The next result implies that  $\text{null } T = \text{null } T^*$  for every normal operator  $T$ .*

In the next section we will see why normal operators are worthy of special attention.

The next result provides a simple characterization of normal operators.

### 7.20 $T$ is normal if and only if $\|Tv\| = \|T^*v\|$ for all $v$

An operator  $T \in \mathcal{L}(V)$  is normal if and only if

$$\|Tv\| = \|T^*v\|$$

for all  $v \in V$ .

**Proof** Let  $T \in \mathcal{L}(V)$ . We will prove both directions of this result at the same time. Note that

$$\begin{aligned} T \text{ is normal} &\iff T^*T - TT^* = 0 \\ &\iff \langle (T^*T - TT^*)v, v \rangle = 0 \quad \text{for all } v \in V \\ &\iff \langle T^*Tv, v \rangle = \langle TT^*v, v \rangle \quad \text{for all } v \in V \\ &\iff \|Tv\|^2 = \|T^*v\|^2 \quad \text{for all } v \in V, \end{aligned}$$

where we used 7.16 to establish the second equivalence (note that the operator  $T^*T - TT^*$  is self-adjoint). The equivalence of the first and last conditions above gives the desired result. ■

Compare the next corollary to Exercise 2. That exercise states that the eigenvalues of the adjoint of each operator are equal (as a set) to the complex conjugates of the eigenvalues of the operator. The exercise says nothing about eigenvectors, because an operator and its adjoint may have different eigenvectors. However, the next corollary implies that a normal operator and its adjoint have the same eigenvectors.

### 7.21 For $T$ normal, $T$ and $T^*$ have the same eigenvectors

Suppose  $T \in \mathcal{L}(V)$  is normal and  $v \in V$  is an eigenvector of  $T$  with eigenvalue  $\lambda$ . Then  $v$  is also an eigenvector of  $T^*$  with eigenvalue  $\bar{\lambda}$ .

**Proof** Because  $T$  is normal, so is  $T - \lambda I$ , as you should verify. Using 7.20, we have

$$0 = \|(T - \lambda I)v\| = \|(T - \lambda I)^*v\| = \|(T^* - \bar{\lambda}I)v\|.$$

Hence  $v$  is an eigenvector of  $T^*$  with eigenvalue  $\bar{\lambda}$ , as desired. ■

Because every self-adjoint operator is normal, the next result applies in particular to self-adjoint operators.

## 7.22 Orthogonal eigenvectors for normal operators

Suppose  $T \in \mathcal{L}(V)$  is normal. Then eigenvectors of  $T$  corresponding to distinct eigenvalues are orthogonal.

**Proof** Suppose  $\alpha, \beta$  are distinct eigenvalues of  $T$ , with corresponding eigenvectors  $u, v$ . Thus  $Tu = \alpha u$  and  $Tv = \beta v$ . From 7.21 we have  $T^*v = \bar{\beta}v$ . Thus

$$\begin{aligned} (\alpha - \beta)\langle u, v \rangle &= \langle \alpha u, v \rangle - \langle u, \bar{\beta}v \rangle \\ &= \langle Tu, v \rangle - \langle u, T^*v \rangle \\ &= 0. \end{aligned}$$

Because  $\alpha \neq \beta$ , the equation above implies that  $\langle u, v \rangle = 0$ . Thus  $u$  and  $v$  are orthogonal, as desired. ■

## EXERCISES 7.A

---

- 1 Suppose  $n$  is a positive integer. Define  $T \in \mathcal{L}(\mathbf{F}^n)$  by

$$T(z_1, \dots, z_n) = (0, z_1, \dots, z_{n-1}).$$

Find a formula for  $T^*(z_1, \dots, z_n)$ .

- 2 Suppose  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ . Prove that  $\lambda$  is an eigenvalue of  $T$  if and only if  $\bar{\lambda}$  is an eigenvalue of  $T^*$ .
- 3 Suppose  $T \in \mathcal{L}(V)$  and  $U$  is a subspace of  $V$ . Prove that  $U$  is invariant under  $T$  if and only if  $U^\perp$  is invariant under  $T^*$ .
- 4 Suppose  $T \in \mathcal{L}(V, W)$ . Prove that
- $T$  is injective if and only if  $T^*$  is surjective;
  - $T$  is surjective if and only if  $T^*$  is injective.

- 5 Prove that

$$\dim \text{null } T^* = \dim \text{null } T + \dim W - \dim V$$

and

$$\dim \text{range } T^* = \dim \text{range } T$$

for every  $T \in \mathcal{L}(V, W)$ .

- 6** Make  $\mathcal{P}_2(\mathbf{R})$  into an inner product space by defining

$$\langle p, q \rangle = \int_0^1 p(x)q(x) dx.$$

Define  $T \in \mathcal{L}(\mathcal{P}_2(\mathbf{R}))$  by  $T(a_0 + a_1x + a_2x^2) = a_1x$ .

- (a) Show that  $T$  is not self-adjoint.  
 (b) The matrix of  $T$  with respect to the basis  $(1, x, x^2)$  is

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

This matrix equals its conjugate transpose, even though  $T$  is not self-adjoint. Explain why this is not a contradiction.

- 7** Suppose  $S, T \in \mathcal{L}(V)$  are self-adjoint. Prove that  $ST$  is self-adjoint if and only if  $ST = TS$ .
- 8** Suppose  $V$  is a real inner product space. Show that the set of self-adjoint operators on  $V$  is a subspace of  $\mathcal{L}(V)$ .
- 9** Suppose  $V$  is a complex inner product space with  $V \neq \{0\}$ . Show that the set of self-adjoint operators on  $V$  is not a subspace of  $\mathcal{L}(V)$ .
- 10** Suppose  $\dim V \geq 2$ . Show that the set of normal operators on  $V$  is not a subspace of  $\mathcal{L}(V)$ .
- 11** Suppose  $P \in \mathcal{L}(V)$  is such that  $P^2 = P$ . Prove that there is a subspace  $U$  of  $V$  such that  $P = P_U$  if and only if  $P$  is self-adjoint.
- 12** Suppose that  $T$  is a normal operator on  $V$  and that 3 and 4 are eigenvalues of  $T$ . Prove that there exists a vector  $v \in V$  such that  $\|v\| = \sqrt{2}$  and  $\|Tv\| = 5$ .
- 13** Give an example of an operator  $T \in \mathcal{L}(\mathbf{C}^4)$  such that  $T$  is normal but not self-adjoint.
- 14** Suppose  $T$  is a normal operator on  $V$ . Suppose also that  $v, w \in V$  satisfy the equations

$$\|v\| = \|w\| = 2, \quad Tv = 3v, \quad Tw = 4w.$$

Show that  $\|T(v + w)\| = 10$ .

**15** Fix  $u, x \in V$ . Define  $T \in \mathcal{L}(V)$  by

$$Tv = \langle v, u \rangle x$$

for every  $v \in V$ .

- (a) Suppose  $\mathbf{F} = \mathbf{R}$ . Prove that  $T$  is self-adjoint if and only if  $u, x$  is linearly dependent.
- (b) Prove that  $T$  is normal if and only if  $u, x$  is linearly dependent.

**16** Suppose  $T \in \mathcal{L}(V)$  is normal. Prove that

$$\text{range } T = \text{range } T^*.$$

**17** Suppose  $T \in \mathcal{L}(V)$  is normal. Prove that

$$\text{null } T^k = \text{null } T \quad \text{and} \quad \text{range } T^k = \text{range } T$$

for every positive integer  $k$ .

- 18** Prove or give a counterexample: If  $T \in \mathcal{L}(V)$  and there exists an ortho-normal basis  $e_1, \dots, e_n$  of  $V$  such that  $\|Te_j\| = \|T^*e_j\|$  for each  $j$ , then  $T$  is normal.
- 19** Suppose  $T \in \mathcal{L}(\mathbf{C}^3)$  is normal and  $T(1, 1, 1) = (2, 2, 2)$ . Suppose  $(z_1, z_2, z_3) \in \text{null } T$ . Prove that  $z_1 + z_2 + z_3 = 0$ .
- 20** Suppose  $T \in \mathcal{L}(V, W)$  and  $\mathbf{F} = \mathbf{R}$ . Let  $\Phi_V$  be the isomorphism from  $V$  onto the dual space  $V'$  given by Exercise 17 in Section 6.B, and let  $\Phi_W$  be the corresponding isomorphism from  $W$  onto  $W'$ . Show that if  $\Phi_V$  and  $\Phi_W$  are used to identify  $V$  and  $W$  with  $V'$  and  $W'$ , then  $T^*$  is identified with the dual map  $T'$ . More precisely, show that  $\Phi_V \circ T^* = T' \circ \Phi_W$ .
- 21** Fix a positive integer  $n$ . In the inner product space of continuous real-valued functions on  $[-\pi, \pi]$  with inner product

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)g(x) dx,$$

let

$$V = \text{span}(1, \cos x, \cos 2x, \dots, \cos nx, \sin x, \sin 2x, \dots, \sin nx).$$

- (a) Define  $D \in \mathcal{L}(V)$  by  $Df = f'$ . Show that  $D^* = -D$ . Conclude that  $D$  is normal but not self-adjoint.
- (b) Define  $T \in \mathcal{L}(V)$  by  $Tf = f''$ . Show that  $T$  is self-adjoint.

## 7.B The Spectral Theorem

Recall that a diagonal matrix is a square matrix that is 0 everywhere except possibly along the diagonal. Recall also that an operator on  $V$  has a diagonal matrix with respect to a basis if and only if the basis consists of eigenvectors of the operator (see 5.41).

The nicest operators on  $V$  are those for which there is an *orthonormal* basis of  $V$  with respect to which the operator has a diagonal matrix. These are precisely the operators  $T \in \mathcal{L}(V)$  such that there is an orthonormal basis of  $V$  consisting of eigenvectors of  $T$ . Our goal in this section is to prove the Spectral Theorem, which characterizes these operators as the normal operators when  $\mathbf{F} = \mathbf{C}$  and as the self-adjoint operators when  $\mathbf{F} = \mathbf{R}$ . The Spectral Theorem is probably the most useful tool in the study of operators on inner product spaces.

Because the conclusion of the Spectral Theorem depends on  $\mathbf{F}$ , we will break the Spectral Theorem into two pieces, called the Complex Spectral Theorem and the Real Spectral Theorem. As is often the case in linear algebra, complex vector spaces are easier to deal with than real vector spaces. Thus we present the Complex Spectral Theorem first.

### The Complex Spectral Theorem

The key part of the Complex Spectral Theorem (7.24) states that if  $\mathbf{F} = \mathbf{C}$  and  $T \in \mathcal{L}(V)$  is normal, then  $T$  has a diagonal matrix with respect to some orthonormal basis of  $V$ . The next example illustrates this conclusion.

---

**7.23 Example** Consider the normal operator  $T \in \mathcal{L}(\mathbf{C}^2)$  from Example 7.19, whose matrix (with respect to the standard basis) is

$$\begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix}.$$

As you can verify,  $\frac{(i,1)}{\sqrt{2}}, \frac{(-i,1)}{\sqrt{2}}$  is an orthonormal basis of  $\mathbf{C}^2$  consisting of eigenvectors of  $T$ , and with respect to this basis the matrix of  $T$  is the diagonal matrix

$$\begin{pmatrix} 2+3i & 0 \\ 0 & 2-3i \end{pmatrix}.$$

---

In the next result, the equivalence of (b) and (c) is easy (see 5.41). Thus we prove only that (c) implies (a) and that (a) implies (c).

## 7.24 Complex Spectral Theorem

Suppose  $\mathbf{F} = \mathbf{C}$  and  $T \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $T$  is normal.
- (b)  $V$  has an orthonormal basis consisting of eigenvectors of  $T$ .
- (c)  $T$  has a diagonal matrix with respect to some orthonormal basis of  $V$ .

**Proof** First suppose (c) holds, so  $T$  has a diagonal matrix with respect to some orthonormal basis of  $V$ . The matrix of  $T^*$  (with respect to the same basis) is obtained by taking the conjugate transpose of the matrix of  $T$ ; hence  $T^*$  also has a diagonal matrix. Any two diagonal matrices commute; thus  $T$  commutes with  $T^*$ , which means that  $T$  is normal. In other words, (a) holds.

Now suppose (a) holds, so  $T$  is normal. By Schur's Theorem (6.38), there is an orthonormal basis  $e_1, \dots, e_n$  of  $V$  with respect to which  $T$  has an upper-triangular matrix. Thus we can write

$$7.25 \quad \mathcal{M}(T, (e_1, \dots, e_n)) = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ & \ddots & \vdots \\ 0 & & a_{n,n} \end{pmatrix}.$$

We will show that this matrix is actually a diagonal matrix.

We see from the matrix above that

$$\|Te_1\|^2 = |a_{1,1}|^2$$

and

$$\|T^*e_1\|^2 = |a_{1,1}|^2 + |a_{1,2}|^2 + \cdots + |a_{1,n}|^2.$$

Because  $T$  is normal,  $\|Te_1\| = \|T^*e_1\|$  (see 7.20). Thus the two equations above imply that all entries in the first row of the matrix in 7.25, except possibly the first entry  $a_{1,1}$ , equal 0.

Now from 7.25 we see that

$$\|Te_2\|^2 = |a_{2,2}|^2$$

(because  $a_{1,2} = 0$ , as we showed in the paragraph above) and

$$\|T^*e_2\|^2 = |a_{2,2}|^2 + |a_{2,3}|^2 + \cdots + |a_{2,n}|^2.$$

Because  $T$  is normal,  $\|Te_2\| = \|T^*e_2\|$ . Thus the two equations above imply that all entries in the second row of the matrix in 7.25, except possibly the diagonal entry  $a_{2,2}$ , equal 0.

Continuing in this fashion, we see that all the nondiagonal entries in the matrix 7.25 equal 0. Thus (c) holds. ■

## The Real Spectral Theorem

We will need a few preliminary results, which apply to both real and complex inner product spaces, for our proof of the Real Spectral Theorem.

You could guess that the next result is true and even discover its proof by thinking about quadratic polynomials with real coefficients. Specifically, suppose  $b, c \in \mathbf{R}$  and  $b^2 < 4c$ . Let  $x$  be a real number. Then

$$x^2 + bx + c = \left(x + \frac{b}{2}\right)^2 + \left(c - \frac{b^2}{4}\right) > 0.$$

*This technique of completing the square can be used to derive the quadratic formula.*

In particular,  $x^2 + bx + c$  is an invertible real number (a convoluted way of saying that it is not 0). Replacing the real number  $x$  with a self-adjoint operator (recall the analogy between real numbers and self-adjoint operators), we are led to the result below.

### 7.26 Invertible quadratic expressions

Suppose  $T \in \mathcal{L}(V)$  is self-adjoint and  $b, c \in \mathbf{R}$  are such that  $b^2 < 4c$ . Then

$$T^2 + bT + cI$$

is invertible.

**Proof** Let  $v$  be a nonzero vector in  $V$ . Then

$$\begin{aligned} \langle (T^2 + bT + cI)v, v \rangle &= \langle T^2v, v \rangle + b\langle Tv, v \rangle + c\langle v, v \rangle \\ &= \langle Tv, Tv \rangle + b\langle Tv, v \rangle + c\|v\|^2 \\ &\geq \|Tv\|^2 - |b|\|Tv\|\|v\| + c\|v\|^2 \\ &= \left(\|Tv\| - \frac{|b|\|v\|}{2}\right)^2 + \left(c - \frac{b^2}{4}\right)\|v\|^2 \\ &> 0, \end{aligned}$$

where the third line above holds by the Cauchy–Schwarz Inequality (6.15). The last inequality implies that  $(T^2 + bT + cI)v \neq 0$ . Thus  $T^2 + bT + cI$  is injective, which implies that it is invertible (see 3.69). ■

We know that every operator, self-adjoint or not, on a finite-dimensional nonzero complex vector space has an eigenvalue (see 5.21). Thus the next result tells us something new only for real inner product spaces.

### 7.27 Self-adjoint operators have eigenvalues

Suppose  $V \neq \{0\}$  and  $T \in \mathcal{L}(V)$  is a self-adjoint operator. Then  $T$  has an eigenvalue.

**Proof** We can assume that  $V$  is a real inner product space, as we have already noted. Let  $n = \dim V$  and choose  $v \in V$  with  $v \neq 0$ . Then

$$v, T v, T^2 v, \dots, T^n v$$

cannot be linearly independent, because  $V$  has dimension  $n$  and we have  $n + 1$  vectors. Thus there exist real numbers  $a_0, \dots, a_n$ , not all 0, such that

$$0 = a_0 v + a_1 T v + \cdots + a_n T^n v.$$

Make the  $a$ 's the coefficients of a polynomial, which can be written in factored form (see 4.17) as

$$\begin{aligned} a_0 + a_1 x + \cdots + a_n x^n \\ = c(x^2 + b_1 x + c_1) \cdots (x^2 + b_M x + c_M)(x - \lambda_1) \cdots (x - \lambda_m), \end{aligned}$$

where  $c$  is a nonzero real number, each  $b_j$ ,  $c_j$ , and  $\lambda_j$  is real, each  $b_j^2$  is less than  $4c_j$ ,  $m + M \geq 1$ , and the equation holds for all real  $x$ . We then have

$$\begin{aligned} 0 &= a_0 v + a_1 T v + \cdots + a_n T^n v \\ &= (a_0 I + a_1 T + \cdots + a_n T^n)v \\ &= c(T^2 + b_1 T + c_1 I) \cdots (T^2 + b_M T + c_M I)(T - \lambda_1 I) \cdots (T - \lambda_m I)v. \end{aligned}$$

By 7.26, each  $T^2 + b_j T + c_j I$  is invertible. Recall also that  $c \neq 0$ . Thus the equation above implies that  $m > 0$  and

$$0 = (T - \lambda_1 I) \cdots (T - \lambda_m I)v.$$

Hence  $T - \lambda_j I$  is not injective for at least one  $j$ . In other words,  $T$  has an eigenvalue. ■

The next result shows that if  $U$  is a subspace of  $V$  that is invariant under a self-adjoint operator  $T$ , then  $U^\perp$  is also invariant under  $T$ . Later we will show that the hypothesis that  $T$  is self-adjoint can be replaced with the weaker hypothesis that  $T$  is normal (see 9.30).

## 7.28 Self-adjoint operators and invariant subspaces

Suppose  $T \in \mathcal{L}(V)$  is self-adjoint and  $U$  is a subspace of  $V$  that is invariant under  $T$ . Then

- (a)  $U^\perp$  is invariant under  $T$ ;
- (b)  $T|_U \in \mathcal{L}(U)$  is self-adjoint;
- (c)  $T|_{U^\perp} \in \mathcal{L}(U^\perp)$  is self-adjoint.

**Proof** To prove (a), suppose  $v \in U^\perp$ . Let  $u \in U$ . Then

$$\langle Tv, u \rangle = \langle v, Tu \rangle = 0,$$

where the first equality above holds because  $T$  is self-adjoint and the second equality above holds because  $U$  is invariant under  $T$  (and hence  $Tu \in U$ ) and because  $v \in U^\perp$ . Because the equation above holds for each  $u \in U$ , we conclude that  $Tv \in U^\perp$ . Thus  $U^\perp$  is invariant under  $T$ , completing the proof of (a).

To prove (b), note that if  $u, v \in U$ , then

$$\langle (T|_U)u, v \rangle = \langle Tu, v \rangle = \langle u, Tv \rangle = \langle u, (T|_U)v \rangle.$$

Thus  $T|_U$  is self-adjoint.

Now (c) follows from replacing  $U$  with  $U^\perp$  in (b), which makes sense by (a). ■

We can now prove the next result, which is one of the major theorems in linear algebra.

## 7.29 Real Spectral Theorem

Suppose  $\mathbf{F} = \mathbf{R}$  and  $T \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $T$  is self-adjoint.
- (b)  $V$  has an orthonormal basis consisting of eigenvectors of  $T$ .
- (c)  $T$  has a diagonal matrix with respect to some orthonormal basis of  $V$ .

**Proof** First suppose (c) holds, so  $T$  has a diagonal matrix with respect to some orthonormal basis of  $V$ . A diagonal matrix equals its transpose. Hence  $T = T^*$ , and thus  $T$  is self-adjoint. In other words, (a) holds.

We will prove that (a) implies (b) by induction on  $\dim V$ . To get started, note that if  $\dim V = 1$ , then (a) implies (b). Now assume that  $\dim V > 1$  and that (a) implies (b) for all real inner product spaces of smaller dimension.

Suppose (a) holds, so  $T \in \mathcal{L}(V)$  is self-adjoint. Let  $u$  be an eigenvector of  $T$  with  $\|u\| = 1$  (7.27 guarantees that  $T$  has an eigenvector, which can then be divided by its norm to produce an eigenvector with norm 1). Let  $U = \text{span}(u)$ . Then  $U$  is a 1-dimensional subspace of  $V$  that is invariant under  $T$ . By 7.28(c), the operator  $T|_{U^\perp} \in \mathcal{L}(U^\perp)$  is self-adjoint.

By our induction hypothesis, there is an orthonormal basis of  $U^\perp$  consisting of eigenvectors of  $T|_{U^\perp}$ . Adjoining  $u$  to this orthonormal basis of  $U^\perp$  gives an orthonormal basis of  $V$  consisting of eigenvectors of  $T$ , completing the proof that (a) implies (b).

We have proved that (c) implies (a) and that (a) implies (b). Clearly (b) implies (c), completing the proof. ■

**7.30 Example** Consider the self-adjoint operator  $T$  on  $\mathbf{R}^3$  whose matrix (with respect to the standard basis) is

$$\begin{pmatrix} 14 & -13 & 8 \\ -13 & 14 & 8 \\ 8 & 8 & -7 \end{pmatrix}.$$

As you can verify,

$$\frac{(1, -1, 0)}{\sqrt{2}}, \frac{(1, 1, 1)}{\sqrt{3}}, \frac{(1, 1, -2)}{\sqrt{6}}$$

is an orthonormal basis of  $\mathbf{R}^3$  consisting of eigenvectors of  $T$ , and with respect to this basis, the matrix of  $T$  is the diagonal matrix

$$\begin{pmatrix} 27 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & -15 \end{pmatrix}.$$

If  $\mathbf{F} = \mathbf{C}$ , then the Complex Spectral Theorem gives a complete description of the normal operators on  $V$ . A complete description of the self-adjoint operators on  $V$  then easily follows (they are the normal operators on  $V$  whose eigenvalues all are real; see Exercise 6).

If  $\mathbf{F} = \mathbf{R}$ , then the Real Spectral Theorem gives a complete description of the self-adjoint operators on  $V$ . In Chapter 9, we will give a complete description of the normal operators on  $V$  (see 9.34).

## EXERCISES 7.B

---

- 1 True or false (and give a proof of your answer): There exists  $T \in \mathcal{L}(\mathbf{R}^3)$  such that  $T$  is not self-adjoint (with respect to the usual inner product) and such that there is a basis of  $\mathbf{R}^3$  consisting of eigenvectors of  $T$ .
- 2 Suppose that  $T$  is a self-adjoint operator on a finite-dimensional inner product space and that 2 and 3 are the only eigenvalues of  $T$ . Prove that  $T^2 - 5T + 6I = 0$ .
- 3 Give an example of an operator  $T \in \mathcal{L}(\mathbf{C}^3)$  such that 2 and 3 are the only eigenvalues of  $T$  and  $T^2 - 5T + 6I \neq 0$ .
- 4 Suppose  $\mathbf{F} = \mathbf{C}$  and  $T \in \mathcal{L}(V)$ . Prove that  $T$  is normal if and only if all pairs of eigenvectors corresponding to distinct eigenvalues of  $T$  are orthogonal and

$$V = E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T),$$

where  $\lambda_1, \dots, \lambda_m$  denote the distinct eigenvalues of  $T$ .

- 5 Suppose  $\mathbf{F} = \mathbf{R}$  and  $T \in \mathcal{L}(V)$ . Prove that  $T$  is self-adjoint if and only if all pairs of eigenvectors corresponding to distinct eigenvalues of  $T$  are orthogonal and

$$V = E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T),$$

where  $\lambda_1, \dots, \lambda_m$  denote the distinct eigenvalues of  $T$ .

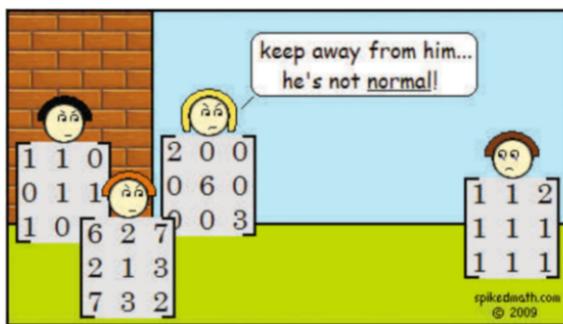
- 6 Prove that a normal operator on a complex inner product space is self-adjoint if and only if all its eigenvalues are real.  
*[The exercise above strengthens the analogy (for normal operators) between self-adjoint operators and real numbers.]*
- 7 Suppose  $V$  is a complex inner product space and  $T \in \mathcal{L}(V)$  is a normal operator such that  $T^9 = T^8$ . Prove that  $T$  is self-adjoint and  $T^2 = T$ .
- 8 Give an example of an operator  $T$  on a complex vector space such that  $T^9 = T^8$  but  $T^2 \neq T$ .
- 9 Suppose  $V$  is a complex inner product space. Prove that every normal operator on  $V$  has a square root. (An operator  $S \in \mathcal{L}(V)$  is called a *square root* of  $T \in \mathcal{L}(V)$  if  $S^2 = T$ .)

- 10** Give an example of a real inner product space  $V$  and  $T \in \mathcal{L}(V)$  and real numbers  $b, c$  with  $b^2 < 4c$  such that  $T^2 + bT + cI$  is not invertible.  
*[The exercise above shows that the hypothesis that  $T$  is self-adjoint is needed in 7.26, even for real vector spaces.]*
- 11** Prove or give a counterexample: every self-adjoint operator on  $V$  has a cube root. (An operator  $S \in \mathcal{L}(V)$  is called a **cube root** of  $T \in \mathcal{L}(V)$  if  $S^3 = T$ .)
- 12** Suppose  $T \in \mathcal{L}(V)$  is self-adjoint,  $\lambda \in \mathbf{F}$ , and  $\epsilon > 0$ . Suppose there exists  $v \in V$  such that  $\|v\| = 1$  and

$$\|Tv - \lambda v\| < \epsilon.$$

Prove that  $T$  has an eigenvalue  $\lambda'$  such that  $|\lambda - \lambda'| < \epsilon$ .

- 13** Give an alternative proof of the Complex Spectral Theorem that avoids Schur's Theorem and instead follows the pattern of the proof of the Real Spectral Theorem.
- 14** Suppose  $U$  is a finite-dimensional real vector space and  $T \in \mathcal{L}(U)$ . Prove that  $U$  has a basis consisting of eigenvectors of  $T$  if and only if there is an inner product on  $U$  that makes  $T$  into a self-adjoint operator.
- 15** Find the matrix entry below that is covered up.



## 7.C Positive Operators and Isometries

### Positive Operators

#### 7.31 Definition *positive operator*

An operator  $T \in \mathcal{L}(V)$  is called **positive** if  $T$  is self-adjoint and

$$\langle Tv, v \rangle \geq 0$$

for all  $v \in V$ .

If  $V$  is a complex vector space, then the requirement that  $T$  is self-adjoint can be dropped from the definition above (by 7.15).

---

#### 7.32 Example *positive operators*

- (a) If  $U$  is a subspace of  $V$ , then the orthogonal projection  $P_U$  is a positive operator, as you should verify.
  - (b) If  $T \in \mathcal{L}(V)$  is self-adjoint and  $b, c \in \mathbf{R}$  are such that  $b^2 < 4c$ , then  $T^2 + bT + cI$  is a positive operator, as shown by the proof of 7.26.
- 

#### 7.33 Definition *square root*

An operator  $R$  is called a **square root** of an operator  $T$  if  $R^2 = T$ .

---

**7.34 Example** If  $T \in \mathcal{L}(\mathbf{F}^3)$  is defined by  $T(z_1, z_2, z_3) = (z_3, 0, 0)$ , then the operator  $R \in \mathcal{L}(\mathbf{F}^3)$  defined by  $R(z_1, z_2, z_3) = (z_2, z_3, 0)$  is a square root of  $T$ .

---

The characterizations of the positive operators in the next result correspond to characterizations of the nonnegative numbers among  $\mathbf{C}$ . Specifically, a complex number  $z$  is nonnegative if and only if it has a nonnegative square root, corresponding to condition (c). Also,  $z$  is nonnegative if and only if it has a real square root, corresponding to condition (d). Finally,  $z$  is nonnegative if and only if there exists a complex number  $w$  such that  $z = \bar{w}w$ , corresponding to condition (e).

*The positive operators correspond to the numbers  $[0, \infty)$ , so better terminology would use the term nonnegative instead of positive. However, operator theorists consistently call these the positive operators, so we will follow that custom.*

### 7.35 Characterization of positive operators

Let  $T \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $T$  is positive;
- (b)  $T$  is self-adjoint and all the eigenvalues of  $T$  are nonnegative;
- (c)  $T$  has a positive square root;
- (d)  $T$  has a self-adjoint square root;
- (e) there exists an operator  $R \in \mathcal{L}(V)$  such that  $T = R^*R$ .

**Proof** We will prove that (a)  $\Rightarrow$  (b)  $\Rightarrow$  (c)  $\Rightarrow$  (d)  $\Rightarrow$  (e)  $\Rightarrow$  (a).

First suppose (a) holds, so that  $T$  is positive. Obviously  $T$  is self-adjoint (by the definition of a positive operator). To prove the other condition in (b), suppose  $\lambda$  is an eigenvalue of  $T$ . Let  $v$  be an eigenvector of  $T$  corresponding to  $\lambda$ . Then

$$0 \leq \langle Tv, v \rangle = \langle \lambda v, v \rangle = \lambda \langle v, v \rangle.$$

Thus  $\lambda$  is a nonnegative number. Hence (b) holds.

Now suppose (b) holds, so that  $T$  is self-adjoint and all the eigenvalues of  $T$  are nonnegative. By the Spectral Theorem (7.24 and 7.29), there is an orthonormal basis  $e_1, \dots, e_n$  of  $V$  consisting of eigenvectors of  $T$ . Let  $\lambda_1, \dots, \lambda_n$  be the eigenvalues of  $T$  corresponding to  $e_1, \dots, e_n$ ; thus each  $\lambda_j$  is a nonnegative number. Let  $R$  be the linear map from  $V$  to  $V$  such that

$$Re_j = \sqrt{\lambda_j} e_j$$

for  $j = 1, \dots, n$  (see 3.5). Then  $R$  is a positive operator, as you should verify. Furthermore,  $R^2 e_j = \lambda_j e_j = Te_j$  for each  $j$ , which implies that  $R^2 = T$ . Thus  $R$  is a positive square root of  $T$ . Hence (c) holds.

Clearly (c) implies (d) (because, by definition, every positive operator is self-adjoint).

Now suppose (d) holds, meaning that there exists a self-adjoint operator  $R$  on  $V$  such that  $T = R^2$ . Then  $T = R^*R$  (because  $R^* = R$ ). Hence (e) holds.

Finally, suppose (e) holds. Let  $R \in \mathcal{L}(V)$  be such that  $T = R^*R$ . Then  $T^* = (R^*R)^* = R^*(R^*)^* = R^*R = T$ . Hence  $T$  is self-adjoint. To complete the proof that (a) holds, note that

$$\langle Tv, v \rangle = \langle R^*Rv, v \rangle = \langle Rv, Rv \rangle \geq 0$$

for every  $v \in V$ . Thus  $T$  is positive. ■

Each nonnegative number has a unique nonnegative square root. The next result shows that positive operators enjoy a similar property.

*Some mathematicians also use the term **positive semidefinite operator**, which means the same as positive operator.*

### 7.36 Each positive operator has only one positive square root

Every positive operator on  $V$  has a unique positive square root.

**Proof** Suppose  $T \in \mathcal{L}(V)$  is positive. Suppose  $v \in V$  is an eigenvector of  $T$ . Thus there exists  $\lambda \geq 0$  such that  $Tv = \lambda v$ .

Let  $R$  be a positive square root of  $T$ . We will prove that  $Rv = \sqrt{\lambda}v$ . This will imply that the behavior of  $R$  on the eigenvectors of  $T$  is uniquely determined. Because there is a basis of  $V$  consisting of eigenvectors of  $T$  (by the Spectral Theorem), this will imply that  $R$  is uniquely determined.

To prove that  $Rv = \sqrt{\lambda}v$ , note that the Spectral Theorem asserts that there is an orthonormal basis  $e_1, \dots, e_n$  of  $V$  consisting of eigenvectors of  $R$ . Because  $R$  is a positive operator, all its eigenvalues are nonnegative. Thus there exist nonnegative numbers  $\lambda_1, \dots, \lambda_n$  such that  $Re_j = \sqrt{\lambda_j}e_j$  for  $j = 1, \dots, n$ .

Because  $e_1, \dots, e_n$  is a basis of  $V$ , we can write

$$v = a_1e_1 + \dots + a_ne_n$$

for some numbers  $a_1, \dots, a_n \in \mathbf{F}$ . Thus

$$Rv = a_1\sqrt{\lambda_1}e_1 + \dots + a_n\sqrt{\lambda_n}e_n$$

and hence

$$R^2v = a_1\lambda_1e_1 + \dots + a_n\lambda_ne_n.$$

But  $R^2 = T$ , and  $Tv = \lambda v$ . Thus the equation above implies

$$a_1\lambda e_1 + \dots + a_n\lambda e_n = a_1\lambda_1e_1 + \dots + a_n\lambda_ne_n.$$

The equation above implies that  $a_j(\lambda - \lambda_j) = 0$  for  $j = 1, \dots, n$ . Hence

$$v = \sum_{\{j: \lambda_j = \lambda\}} a_j e_j,$$

and thus

$$Rv = \sum_{\{j: \lambda_j = \lambda\}} a_j \sqrt{\lambda}e_j = \sqrt{\lambda}v,$$

as desired. ■

## Isometries

Operators that preserve norms are sufficiently important to deserve a name:

### 7.37 Definition *isometry*

- An operator  $S \in \mathcal{L}(V)$  is called an *isometry* if

$$\|Sv\| = \|v\|$$

for all  $v \in V$ .

- In other words, an operator is an isometry if it preserves norms.

The Greek word *isos* means equal; the Greek word *metron* means measure. Thus *isometry* literally means equal measure.

For example,  $\lambda I$  is an isometry whenever  $\lambda \in \mathbf{F}$  satisfies  $|\lambda| = 1$ . We will see soon that if  $\mathbf{F} = \mathbf{C}$ , then the next example includes all isometries.

**7.38 Example** Suppose  $\lambda_1, \dots, \lambda_n$  are scalars with absolute value 1 and  $S \in \mathcal{L}(V)$  satisfies  $Se_j = \lambda_j e_j$  for some orthonormal basis  $e_1, \dots, e_n$  of  $V$ . Show that  $S$  is an isometry.

**Solution** Suppose  $v \in V$ . Then

$$7.39 \quad v = \langle v, e_1 \rangle e_1 + \cdots + \langle v, e_n \rangle e_n$$

and

$$7.40 \quad \|v\|^2 = |\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_n \rangle|^2,$$

where we have used 6.30. Applying  $S$  to both sides of 7.39 gives

$$\begin{aligned} Sv &= \langle v, e_1 \rangle Se_1 + \cdots + \langle v, e_n \rangle Se_n \\ &= \lambda_1 \langle v, e_1 \rangle e_1 + \cdots + \lambda_n \langle v, e_n \rangle e_n. \end{aligned}$$

The last equation, along with the equation  $|\lambda_j| = 1$ , shows that

$$7.41 \quad \|Sv\|^2 = |\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_n \rangle|^2.$$

Comparing 7.40 and 7.41 shows that  $\|v\| = \|Sv\|$ . In other words,  $S$  is an isometry.

The next result provides several conditions that are equivalent to being an isometry. The equivalence of (a) and (b) shows that an operator is an isometry if and only if it preserves inner products. The equivalence of (a) and (c) [or (d)] shows that an operator is an isometry if and only if the list of columns of its matrix with respect to every [or some] basis is orthonormal. Exercise 10 implies that in the previous sentence we can replace “columns” with “rows”.

An isometry on a real inner product space is often called an **orthogonal operator**. An isometry on a complex inner product space is often called a **unitary operator**. We use the term **isometry** so that our results can apply to both real and complex inner product spaces.

## 7.42 Characterization of isometries

Suppose  $S \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $S$  is an isometry;
- (b)  $\langle Su, Sv \rangle = \langle u, v \rangle$  for all  $u, v \in V$ ;
- (c)  $Se_1, \dots, Se_n$  is orthonormal for every orthonormal list of vectors  $e_1, \dots, e_n$  in  $V$ ;
- (d) there exists an orthonormal basis  $e_1, \dots, e_n$  of  $V$  such that  $Se_1, \dots, Se_n$  is orthonormal;
- (e)  $S^*S = I$ ;
- (f)  $SS^* = I$ ;
- (g)  $S^*$  is an isometry;
- (h)  $S$  is invertible and  $S^{-1} = S^*$ .

**Proof** First suppose (a) holds, so  $S$  is an isometry. Exercises 19 and 20 in Section 6.A show that inner products can be computed from norms. Because  $S$  preserves norms, this implies that  $S$  preserves inner products, and hence (b) holds. More precisely, if  $V$  is a real inner product space, then for every  $u, v \in V$  we have

$$\begin{aligned}\langle Su, Sv \rangle &= (\|Su + Sv\|^2 - \|Su - Sv\|^2)/4 \\ &= (\|S(u + v)\|^2 - \|S(u - v)\|^2)/4 \\ &= (\|u + v\|^2 - \|u - v\|^2)/4 \\ &= \langle u, v \rangle,\end{aligned}$$

where the first equality comes from Exercise 19 in Section 6.A, the second equality comes from the linearity of  $S$ , the third equality holds because  $S$  is an isometry, and the last equality again comes from Exercise 19 in Section 6.A. If  $V$  is a complex inner product space, then use Exercise 20 in Section 6.A instead of Exercise 19 to obtain the same conclusion. In either case, we see that (b) holds.

Now suppose (b) holds, so  $S$  preserves inner products. Suppose that  $e_1, \dots, e_n$  is an orthonormal list of vectors in  $V$ . Then we see that the list  $Se_1, \dots, Se_n$  is orthonormal because  $\langle Se_j, Se_k \rangle = \langle e_j, e_k \rangle$ . Thus (c) holds.

Clearly (c) implies (d).

Now suppose (d) holds. Let  $e_1, \dots, e_n$  be an orthonormal basis of  $V$  such that  $Se_1, \dots, Se_n$  is orthonormal. Thus

$$\langle S^* Se_j, e_k \rangle = \langle e_j, e_k \rangle$$

for  $j, k = 1, \dots, n$  [because the term on the left equals  $\langle Se_j, Se_k \rangle$  and  $(Se_1, \dots, Se_n)$  is orthonormal]. All vectors  $u, v \in V$  can be written as linear combinations of  $e_1, \dots, e_n$ , and thus the equation above implies that  $\langle S^* Su, v \rangle = \langle u, v \rangle$ . Hence  $S^* S = I$ ; in other words, (e) holds.

Now suppose (e) holds, so that  $S^* S = I$ . In general, an operator  $S$  need not commute with  $S^*$ . However,  $S^* S = I$  if and only if  $SS^* = I$ ; this is a special case of Exercise 10 in Section 3.D. Thus  $SS^* = I$ , showing that (f) holds.

Now suppose (f) holds, so  $SS^* = I$ . If  $v \in V$ , then

$$\|S^* v\|^2 = \langle S^* v, S^* v \rangle = \langle SS^* v, v \rangle = \langle v, v \rangle = \|v\|^2.$$

Thus  $S^*$  is an isometry, showing that (g) holds.

Now suppose (g) holds, so  $S^*$  is an isometry. We know that (a)  $\Rightarrow$  (e) and (a)  $\Rightarrow$  (f) because we have shown (a)  $\Rightarrow$  (b)  $\Rightarrow$  (c)  $\Rightarrow$  (d)  $\Rightarrow$  (e)  $\Rightarrow$  (f). Using the implications (a)  $\Rightarrow$  (e) and (a)  $\Rightarrow$  (f) but with  $S$  replaced with  $S^*$  [and using the equation  $(S^*)^* = S$ ], we conclude that  $SS^* = I$  and  $S^* S = I$ . Thus  $S$  is invertible and  $S^{-1} = S^*$ ; in other words, (h) holds.

Now suppose (h) holds, so  $S$  is invertible and  $S^{-1} = S^*$ . Thus  $S^* S = I$ . If  $v \in V$ , then

$$\|Sv\|^2 = \langle Sv, Sv \rangle = \langle S^* Sv, v \rangle = \langle v, v \rangle = \|v\|^2.$$

Thus  $S$  is an isometry, showing that (a) holds.

We have shown (a)  $\Rightarrow$  (b)  $\Rightarrow$  (c)  $\Rightarrow$  (d)  $\Rightarrow$  (e)  $\Rightarrow$  (f)  $\Rightarrow$  (g)  $\Rightarrow$  (h)  $\Rightarrow$  (a), completing the proof. ■

The previous result shows that every isometry is normal [see (a), (e), and (f) of 7.42]. Thus characterizations of normal operators can be used to give descriptions of isometries. We do this in the next result in the complex case and in Chapter 9 in the real case (see 9.36).

### 7.43 Description of isometries when $\mathbf{F} = \mathbf{C}$

Suppose  $V$  is a complex inner product space and  $S \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $S$  is an isometry.
- (b) There is an orthonormal basis of  $V$  consisting of eigenvectors of  $S$  whose corresponding eigenvalues all have absolute value 1.

**Proof** We have already shown (see Example 7.38) that (b) implies (a).

To prove the other direction, suppose (a) holds, so  $S$  is an isometry. By the Complex Spectral Theorem (7.24), there is an orthonormal basis  $e_1, \dots, e_n$  of  $V$  consisting of eigenvectors of  $S$ . For  $j \in \{1, \dots, n\}$ , let  $\lambda_j$  be the eigenvalue corresponding to  $e_j$ . Then

$$|\lambda_j| = \|\lambda_j e_j\| = \|Se_j\| = \|e_j\| = 1.$$

Thus each eigenvalue of  $S$  has absolute value 1, completing the proof. ■

## EXERCISES 7.C

- 1 Prove or give a counterexample: If  $T \in \mathcal{L}(V)$  is self-adjoint and there exists an orthonormal basis  $e_1, \dots, e_n$  of  $V$  such that  $\langle Te_j, e_j \rangle \geq 0$  for each  $j$ , then  $T$  is a positive operator.

- 2 Suppose  $T$  is a positive operator on  $V$ . Suppose  $v, w \in V$  are such that

$$Tv = w \quad \text{and} \quad Tw = v.$$

Prove that  $v = w$ .

- 3 Suppose  $T$  is a positive operator on  $V$  and  $U$  is a subspace of  $V$  invariant under  $T$ . Prove that  $T|_U \in \mathcal{L}(U)$  is a positive operator on  $U$ .
- 4 Suppose  $T \in \mathcal{L}(V, W)$ . Prove that  $T^*T$  is a positive operator on  $V$  and  $TT^*$  is a positive operator on  $W$ .

- 5 Prove that the sum of two positive operators on  $V$  is positive.
- 6 Suppose  $T \in \mathcal{L}(V)$  is positive. Prove that  $T^k$  is positive for every positive integer  $k$ .
- 7 Suppose  $T$  is a positive operator on  $V$ . Prove that  $T$  is invertible if and only if

$$\langle Tv, v \rangle > 0$$

for every  $v \in V$  with  $v \neq 0$ .

- 8 Suppose  $T \in \mathcal{L}(V)$ . For  $u, v \in V$ , define  $\langle u, v \rangle_T$  by

$$\langle u, v \rangle_T = \langle Tu, v \rangle.$$

Prove that  $\langle \cdot, \cdot \rangle_T$  is an inner product on  $V$  if and only if  $T$  is an invertible positive operator (with respect to the original inner product  $\langle \cdot, \cdot \rangle$ ).

- 9 Prove or disprove: the identity operator on  $\mathbf{F}^2$  has infinitely many self-adjoint square roots.
- 10 Suppose  $S \in \mathcal{L}(V)$ . Prove that the following are equivalent:
- (a)  $S$  is an isometry;
  - (b)  $\langle S^*u, S^*v \rangle = \langle u, v \rangle$  for all  $u, v \in V$ ;
  - (c)  $S^*e_1, \dots, S^*e_m$  is an orthonormal list for every orthonormal list of vectors  $e_1, \dots, e_m$  in  $V$ ;
  - (d)  $S^*e_1, \dots, S^*e_n$  is an orthonormal basis for some orthonormal basis  $e_1, \dots, e_n$  of  $V$ .
- 11 Suppose  $T_1, T_2$  are normal operators on  $\mathcal{L}(\mathbf{F}^3)$  and both operators have 2, 5, 7 as eigenvalues. Prove that there exists an isometry  $S \in \mathcal{L}(\mathbf{F}^3)$  such that  $T_1 = S^*T_2S$ .
- 12 Give an example of two self-adjoint operators  $T_1, T_2 \in \mathcal{L}(\mathbf{F}^4)$  such that the eigenvalues of both operators are 2, 5, 7 but there does not exist an isometry  $S \in \mathcal{L}(\mathbf{F}^4)$  such that  $T_1 = S^*T_2S$ . Be sure to explain why there is no isometry with the required property.
- 13 Prove or give a counterexample: if  $S \in \mathcal{L}(V)$  and there exists an orthonormal basis  $e_1, \dots, e_n$  of  $V$  such that  $\|Se_j\| = 1$  for each  $e_j$ , then  $S$  is an isometry.
- 14 Let  $T$  be the second derivative operator in Exercise 21 in Section 7.A. Show that  $-T$  is a positive operator.

## 7.D *Polar Decomposition and Singular Value Decomposition*

### Polar Decomposition

Recall our analogy between  $\mathbf{C}$  and  $\mathcal{L}(V)$ . Under this analogy, a complex number  $z$  corresponds to an operator  $T$ , and  $\bar{z}$  corresponds to  $T^*$ . The real numbers ( $z = \bar{z}$ ) correspond to the self-adjoint operators ( $T = T^*$ ), and the nonnegative numbers correspond to the (badly named) positive operators.

Another distinguished subset of  $\mathbf{C}$  is the unit circle, which consists of the complex numbers  $z$  such that  $|z| = 1$ . The condition  $|z| = 1$  is equivalent to the condition  $\bar{z}z = 1$ . Under our analogy, this would correspond to the condition  $T^*T = I$ , which is equivalent to  $T$  being an isometry (see 7.42). In other words, the unit circle in  $\mathbf{C}$  corresponds to the isometries.

Continuing with our analogy, note that each complex number  $z$  except 0 can be written in the form

$$z = \left(\frac{z}{|z|}\right)|z| = \left(\frac{z}{|z|}\right)\sqrt{\bar{z}z},$$

where the first factor, namely,  $z/|z|$ , is an element of the unit circle. Our analogy leads us to guess that each operator  $T \in \mathcal{L}(V)$  can be written as an isometry times  $\sqrt{T^*T}$ . That guess is indeed correct, as we now prove after defining the obvious notation, which is justified by 7.36.

#### 7.44 Notation $\sqrt{T}$

If  $T$  is a positive operator, then  $\sqrt{T}$  denotes the unique positive square root of  $T$ .

Now we can state and prove the Polar Decomposition, which gives a beautiful description of an arbitrary operator on  $V$ . Note that  $T^*T$  is a positive operator for every  $T \in \mathcal{L}(V)$ , and thus  $\sqrt{T^*T}$  is well defined.

#### 7.45 Polar Decomposition

Suppose  $T \in \mathcal{L}(V)$ . Then there exists an isometry  $S \in \mathcal{L}(V)$  such that

$$T = S\sqrt{T^*T}.$$

**Proof** If  $v \in V$ , then

$$\begin{aligned}\|Tv\|^2 &= \langle Tv, Tv \rangle = \langle T^*Tv, v \rangle \\ &= \langle \sqrt{T^*T} \sqrt{T^*T}v, v \rangle \\ &= \langle \sqrt{T^*T}v, \sqrt{T^*T}v \rangle \\ &= \|\sqrt{T^*T}v\|^2.\end{aligned}$$

Thus

$$\mathbf{7.46} \quad \|Tv\| = \|\sqrt{T^*T}v\|$$

for all  $v \in V$ .

Define a linear map  $S_1 : \text{range } \sqrt{T^*T} \rightarrow \text{range } T$  by

$$\mathbf{7.47} \quad S_1(\sqrt{T^*T}v) = Tv.$$

The idea of the proof is to extend  $S_1$  to an isometry  $S \in \mathcal{L}(V)$  such that  $T = S\sqrt{T^*T}$ . Now for the details.

First we must check that  $S_1$  is well defined. To do this, suppose  $v_1, v_2 \in V$  are such that  $\sqrt{T^*T}v_1 = \sqrt{T^*T}v_2$ . For the definition given by 7.47 to make sense, we must show that  $Tv_1 = Tv_2$ . Note that

$$\begin{aligned}\|Tv_1 - Tv_2\| &= \|T(v_1 - v_2)\| \\ &= \|\sqrt{T^*T}(v_1 - v_2)\| \\ &= \|\sqrt{T^*T}v_1 - \sqrt{T^*T}v_2\| \\ &= 0,\end{aligned}$$

where the second equality holds by 7.46. The equation above shows that  $Tv_1 = Tv_2$ , so  $S_1$  is indeed well defined. You should verify that  $S_1$  is a linear map.

We see from 7.47 that  $S_1$  maps  $\text{range } \sqrt{T^*T}$  onto  $\text{range } T$ . Clearly 7.46 and 7.47 imply that

$$\|S_1u\| = \|u\|$$

for all  $u \in \text{range } \sqrt{T^*T}$ .

*The rest of the proof extends  $S_1$  to an isometry  $S$  on all of  $V$ .*

In particular,  $S_1$  is injective. Thus from the Fundamental Theorem of Linear Maps (3.22), applied to  $S_1$ , we have

$$\dim \text{range } \sqrt{T^*T} = \dim \text{range } T.$$

This implies that  $\dim(\text{range } \sqrt{T^*T})^\perp = \dim(\text{range } T)^\perp$  (see 6.50). Thus orthonormal bases  $e_1, \dots, e_m$  of  $(\text{range } \sqrt{T^*T})^\perp$  and  $f_1, \dots, f_m$  of  $(\text{range } T)^\perp$  can be chosen; the key point here is that these two orthonormal bases have the same length (denoted  $m$ ). Now define a linear map  $S_2 : (\text{range } \sqrt{T^*T})^\perp \rightarrow (\text{range } T)^\perp$  by

$$S_2(a_1e_1 + \cdots + a_m e_m) = a_1f_1 + \cdots + a_m f_m.$$

For all  $w \in (\text{range } \sqrt{T^*T})^\perp$ , we have  $\|S_2w\| = \|w\|$  (from 6.25).

Now let  $S$  be the operator on  $V$  that equals  $S_1$  on  $\text{range } \sqrt{T^*T}$  and equals  $S_2$  on  $(\text{range } \sqrt{T^*T})^\perp$ . More precisely, recall that each  $v \in V$  can be written uniquely in the form

**7.48**

$$v = u + w,$$

where  $u \in \text{range } \sqrt{T^*T}$  and  $w \in (\text{range } \sqrt{T^*T})^\perp$  (see 6.47). For  $v \in V$  with decomposition as above, define  $Sv$  by

$$Sv = S_1u + S_2w.$$

For each  $v \in V$  we have

$$S(\sqrt{T^*T}v) = S_1(\sqrt{T^*T}v) = Tv,$$

so  $T = S\sqrt{T^*T}$ , as desired. All that remains is to show that  $S$  is an isometry. However, this follows easily from two uses of the Pythagorean Theorem: if  $v \in V$  has decomposition as in 7.48, then

$$\|Sv\|^2 = \|S_1u + S_2w\|^2 = \|S_1u\|^2 + \|S_2w\|^2 = \|u\|^2 + \|w\|^2 = \|v\|^2;$$

the second equality holds because  $S_1u \in \text{range } T$  and  $S_2w \in (\text{range } T)^\perp$ . ■

The Polar Decomposition (7.45) states that each operator on  $V$  is the product of an isometry and a positive operator. Thus we can write each operator on  $V$  as the product of two operators, each of which comes from a class that we can completely describe and that we understand reasonably well. The isometries are described by 7.43 and 9.36; the positive operators are described by the Spectral Theorem (7.24 and 7.29).

Specifically, consider the case  $\mathbf{F} = \mathbf{C}$ , and suppose  $T = S\sqrt{T^*T}$  is a Polar Decomposition of an operator  $T \in \mathcal{L}(V)$ , where  $S$  is an isometry. Then there is an orthonormal basis of  $V$  with respect to which  $S$  has a diagonal matrix, and there is an orthonormal basis of  $V$  with respect to which  $\sqrt{T^*T}$  has a diagonal matrix. **Warning:** there may not exist an orthonormal basis that simultaneously puts the matrices of both  $S$  and  $\sqrt{T^*T}$  into these nice diagonal forms. In other words,  $S$  may require one orthonormal basis and  $\sqrt{T^*T}$  may require a different orthonormal basis.

## Singular Value Decomposition

The eigenvalues of an operator tell us something about the behavior of the operator. Another collection of numbers, called the singular values, is also useful. Recall that eigenspaces and the notation  $E$  are defined in 5.36.

### 7.49 Definition *singular values*

Suppose  $T \in \mathcal{L}(V)$ . The *singular values* of  $T$  are the eigenvalues of  $\sqrt{T^*T}$ , with each eigenvalue  $\lambda$  repeated  $\dim E(\lambda, \sqrt{T^*T})$  times.

The singular values of  $T$  are all nonnegative, because they are the eigenvalues of the positive operator  $\sqrt{T^*T}$ .

### 7.50 Example

Define  $T \in \mathcal{L}(\mathbf{F}^4)$  by

$$T(z_1, z_2, z_3, z_4) = (0, 3z_1, 2z_2, -3z_4).$$

Find the singular values of  $T$ .

**Solution** A calculation shows  $T^*T(z_1, z_2, z_3, z_4) = (9z_1, 4z_2, 0, 9z_4)$ , as you should verify. Thus

$$\sqrt{T^*T}(z_1, z_2, z_3, z_4) = (3z_1, 2z_2, 0, 3z_4),$$

and we see that the eigenvalues of  $\sqrt{T^*T}$  are 3, 2, 0 and

$$\dim E(3, \sqrt{T^*T}) = 2, \dim E(2, \sqrt{T^*T}) = 1, \dim E(0, \sqrt{T^*T}) = 1.$$

Hence the singular values of  $T$  are 3, 3, 2, 0.

Note that  $-3$  and  $0$  are the only eigenvalues of  $T$ . Thus in this case, the collection of eigenvalues did not pick up the number 2 that appears in the definition (and hence the behavior) of  $T$ , but the collection of singular values does include 2.

Each  $T \in \mathcal{L}(V)$  has  $\dim V$  singular values, as can be seen by applying the Spectral Theorem and 5.41 [see especially part (e)] to the positive (hence self-adjoint) operator  $\sqrt{T^*T}$ . For example, the operator  $T$  defined in Example 7.50 on the four-dimensional vector space  $\mathbf{F}^4$  has four singular values (they are 3, 3, 2, 0), as we saw above.

The next result shows that every operator on  $V$  has a clean description in terms of its singular values and two orthonormal bases of  $V$ .

### 7.51 Singular Value Decomposition

Suppose  $T \in \mathcal{L}(V)$  has singular values  $s_1, \dots, s_n$ . Then there exist orthonormal bases  $e_1, \dots, e_n$  and  $f_1, \dots, f_n$  of  $V$  such that

$$Tv = s_1 \langle v, e_1 \rangle f_1 + \cdots + s_n \langle v, e_n \rangle f_n$$

for every  $v \in V$ .

**Proof** By the Spectral Theorem applied to  $\sqrt{T^*T}$ , there is an orthonormal basis  $e_1, \dots, e_n$  of  $V$  such that  $\sqrt{T^*T}e_j = s_j e_j$  for  $j = 1, \dots, n$ .

We have

$$v = \langle v, e_1 \rangle e_1 + \cdots + \langle v, e_n \rangle e_n$$

for every  $v \in V$  (see 6.30). Apply  $\sqrt{T^*T}$  to both sides of this equation, getting

$$\sqrt{T^*T}v = s_1 \langle v, e_1 \rangle e_1 + \cdots + s_n \langle v, e_n \rangle e_n$$

for every  $v \in V$ . By the Polar Decomposition (see 7.45), there is an isometry  $S \in \mathcal{L}(V)$  such that  $T = S\sqrt{T^*T}$ . Apply  $S$  to both sides of the equation above, getting

$$Tv = s_1 \langle v, e_1 \rangle Se_1 + \cdots + s_n \langle v, e_n \rangle Se_n$$

for every  $v \in V$ . For each  $j$ , let  $f_j = Se_j$ . Because  $S$  is an isometry,  $f_1, \dots, f_n$  is an orthonormal basis of  $V$  (see 7.42). The equation above now becomes

$$Tv = s_1 \langle v, e_1 \rangle f_1 + \cdots + s_n \langle v, e_n \rangle f_n$$

for every  $v \in V$ , completing the proof. ■

When we worked with linear maps from one vector space to a second vector space, we considered the matrix of a linear map with respect to a basis of the first vector space and a basis of the second vector space. When dealing with operators, which are linear maps from a vector space to itself, we almost always use only one basis, making it play both roles.

The Singular Value Decomposition allows us a rare opportunity to make good use of two different bases for the matrix of an operator. To do this, suppose  $T \in \mathcal{L}(V)$ . Let  $s_1, \dots, s_n$  denote the singular values of  $T$ , and let  $e_1, \dots, e_n$  and  $f_1, \dots, f_n$  be orthonormal bases of  $V$  such that the Singular Value Decomposition 7.51 holds. Because  $Te_j = s_j f_j$  for each  $j$ , we have

$$\mathcal{M}(T, (e_1, \dots, e_n), (f_1, \dots, f_n)) = \begin{pmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_n \end{pmatrix}.$$

In other words, every operator on  $V$  has a diagonal matrix with respect to some orthonormal bases of  $V$ , provided that we are permitted to use two different bases rather than a single basis as customary when working with operators.

Singular values and the Singular Value Decomposition have many applications (some are given in the exercises), including applications in computational linear algebra. To compute numeric approximations to the singular values of an operator  $T$ , first compute  $T^*T$  and then compute approximations to the eigenvalues of  $T^*T$  (good techniques exist for approximating eigenvalues of positive operators). The nonnegative square roots of these (approximate) eigenvalues of  $T^*T$  will be the (approximate) singular values of  $T$ . In other words, the singular values of  $T$  can be approximated without computing the square root of  $T^*T$ . The next result helps justify working with  $T^*T$  instead of  $\sqrt{T^*T}$ .

### 7.52 Singular values without taking square root of an operator

Suppose  $T \in \mathcal{L}(V)$ . Then the singular values of  $T$  are the nonnegative square roots of the eigenvalues of  $T^*T$ , with each eigenvalue  $\lambda$  repeated  $\dim E(\lambda, T^*T)$  times.

**Proof** The Spectral Theorem implies that there are an orthonormal basis  $e_1, \dots, e_n$  and nonnegative numbers  $\lambda_1, \dots, \lambda_n$  such that  $T^*Te_j = \lambda_j e_j$  for  $j = 1, \dots, n$ . It is easy to see that  $\sqrt{T^*T}e_j = \sqrt{\lambda_j}e_j$  for  $j = 1, \dots, n$ , which implies the desired result. ■

## EXERCISES 7.D

---

- 1 Fix  $u, x \in V$  with  $u \neq 0$ . Define  $T \in \mathcal{L}(V)$  by

$$Tv = \langle v, u \rangle x$$

for every  $v \in V$ . Prove that

$$\sqrt{T^*T}v = \frac{\|x\|}{\|u\|} \langle v, u \rangle u$$

for every  $v \in V$ .

- 2 Give an example of  $T \in \mathcal{L}(\mathbf{C}^2)$  such that 0 is the only eigenvalue of  $T$  and the singular values of  $T$  are 5, 0.

- 3 Suppose  $T \in \mathcal{L}(V)$ . Prove that there exists an isometry  $S \in \mathcal{L}(V)$  such that

$$T = \sqrt{TT^*} S.$$

- 4 Suppose  $T \in \mathcal{L}(V)$  and  $s$  is a singular value of  $T$ . Prove that there exists a vector  $v \in V$  such that  $\|v\| = 1$  and  $\|Tv\| = s$ .
- 5 Suppose  $T \in \mathcal{L}(\mathbf{C}^2)$  is defined by  $T(x, y) = (-4y, x)$ . Find the singular values of  $T$ .
- 6 Find the singular values of the differentiation operator  $D \in \mathcal{P}(\mathbf{R}^2)$  defined by  $Dp = p'$ , where the inner product on  $\mathcal{P}(\mathbf{R}^2)$  is as in Example 6.33.
- 7 Define  $T \in \mathcal{L}(\mathbf{F}^3)$  by

$$T(z_1, z_2, z_3) = (z_3, 2z_1, 3z_2).$$

Find (explicitly) an isometry  $S \in \mathcal{L}(\mathbf{F}^3)$  such that  $T = S\sqrt{T^*T}$ .

- 8 Suppose  $T \in \mathcal{L}(V)$ ,  $S \in \mathcal{L}(V)$  is an isometry, and  $R \in \mathcal{L}(V)$  is a positive operator such that  $T = SR$ . Prove that  $R = \sqrt{T^*T}$ .  
*[The exercise above shows that if we write  $T$  as the product of an isometry and a positive operator (as in the Polar Decomposition 7.45), then the positive operator equals  $\sqrt{T^*T}$ .]*
- 9 Suppose  $T \in \mathcal{L}(V)$ . Prove that  $T$  is invertible if and only if there exists a unique isometry  $S \in \mathcal{L}(V)$  such that  $T = S\sqrt{T^*T}$ .
- 10 Suppose  $T \in \mathcal{L}(V)$  is self-adjoint. Prove that the singular values of  $T$  equal the absolute values of the eigenvalues of  $T$ , repeated appropriately.
- 11 Suppose  $T \in \mathcal{L}(V)$ . Prove that  $T$  and  $T^*$  have the same singular values.
- 12 Prove or give a counterexample: if  $T \in \mathcal{L}(V)$ , then the singular values of  $T^2$  equal the squares of the singular values of  $T$ .
- 13 Suppose  $T \in \mathcal{L}(V)$ . Prove that  $T$  is invertible if and only if 0 is not a singular value of  $T$ .
- 14 Suppose  $T \in \mathcal{L}(V)$ . Prove that  $\dim \text{range } T$  equals the number of nonzero singular values of  $T$ .
- 15 Suppose  $S \in \mathcal{L}(V)$ . Prove that  $S$  is an isometry if and only if all the singular values of  $S$  equal 1.

- 16** Suppose  $T_1, T_2 \in \mathcal{L}(V)$ . Prove that  $T_1$  and  $T_2$  have the same singular values if and only if there exist isometries  $S_1, S_2 \in \mathcal{L}(V)$  such that  $T_1 = S_1 T_2 S_2$ .

- 17** Suppose  $T \in \mathcal{L}(V)$  has singular value decomposition given by

$$Tv = s_1 \langle v, e_1 \rangle f_1 + \cdots + s_n \langle v, e_n \rangle f_n$$

for every  $v \in V$ , where  $s_1, \dots, s_n$  are the singular values of  $T$  and  $e_1, \dots, e_n$  and  $f_1, \dots, f_n$  are orthonormal bases of  $V$ .

- (a) Prove that if  $v \in V$ , then

$$T^*v = s_1 \langle v, f_1 \rangle e_1 + \cdots + s_n \langle v, f_n \rangle e_n.$$

- (b) Prove that if  $v \in V$ , then

$$T^*Tv = s_1^2 \langle v, e_1 \rangle e_1 + \cdots + s_n^2 \langle v, e_n \rangle e_n.$$

- (c) Prove that if  $v \in V$ , then

$$\sqrt{T^*Tv} = s_1 \langle v, e_1 \rangle e_1 + \cdots + s_n \langle v, e_n \rangle e_n.$$

- (d) Suppose  $T$  is invertible. Prove that if  $v \in V$ , then

$$T^{-1}v = \frac{\langle v, f_1 \rangle e_1}{s_1} + \cdots + \frac{\langle v, f_n \rangle e_n}{s_n}$$

for every  $v \in V$ .

- 18** Suppose  $T \in \mathcal{L}(V)$ . Let  $\hat{s}$  denote the smallest singular value of  $T$ , and let  $s$  denote the largest singular value of  $T$ .

- (a) Prove that  $\hat{s}\|v\| \leq \|Tv\| \leq s\|v\|$  for every  $v \in V$ .  
 (b) Suppose  $\lambda$  is an eigenvalue of  $T$ . Prove that  $\hat{s} \leq |\lambda| \leq s$ .

- 19** Suppose  $T \in \mathcal{L}(V)$ . Show that  $T$  is uniformly continuous with respect to the metric  $d$  on  $V$  defined by  $d(u, v) = \|u - v\|$ .

- 20** Suppose  $S, T \in \mathcal{L}(V)$ . Let  $s$  denote the largest singular value of  $S$ , let  $t$  denote the largest singular value of  $T$ , and let  $r$  denote the largest singular value of  $S + T$ . Prove that  $r \leq s + t$ .



# CHAPTER 8

*Hypatia, the 5<sup>th</sup> century Egyptian mathematician and philosopher, as envisioned around 1900 by Alfred Seifert.*

## *Operators on Complex Vector Spaces*

In this chapter we delve deeper into the structure of operators, with most of the attention on complex vector spaces. An inner product does not help with this material, so we return to the general setting of a finite-dimensional vector space. To avoid some trivialities, we will assume that  $V \neq \{0\}$ . Thus our assumptions for this chapter are as follows:

### 8.1 Notation $\mathbf{F}, V$

- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  denotes a finite-dimensional nonzero vector space over  $\mathbf{F}$ .

### LEARNING OBJECTIVES FOR THIS CHAPTER

- generalized eigenvectors and generalized eigenspaces
- characteristic polynomial and the Cayley–Hamilton Theorem
- decomposition of an operator
- minimal polynomial
- Jordan Form

## 8.A Generalized Eigenvectors and Nilpotent Operators

### Null Spaces of Powers of an Operator

We begin this chapter with a study of null spaces of powers of an operator.

#### 8.2 Sequence of increasing null spaces

Suppose  $T \in \mathcal{L}(V)$ . Then

$$\{0\} = \text{null } T^0 \subset \text{null } T^1 \subset \cdots \subset \text{null } T^k \subset \text{null } T^{k+1} \subset \cdots .$$

**Proof** Suppose  $k$  is a nonnegative integer and  $v \in \text{null } T^k$ . Then  $T^k v = 0$ , and hence  $T^{k+1}v = T(T^k v) = T(0) = 0$ . Thus  $v \in \text{null } T^{k+1}$ . Hence  $\text{null } T^k \subset \text{null } T^{k+1}$ , as desired. ■

The next result says that if two consecutive terms in this sequence of subspaces are equal, then all later terms in the sequence are equal.

#### 8.3 Equality in the sequence of null spaces

Suppose  $T \in \mathcal{L}(V)$ . Suppose  $m$  is a nonnegative integer such that  $\text{null } T^m = \text{null } T^{m+1}$ . Then

$$\text{null } T^m = \text{null } T^{m+1} = \text{null } T^{m+2} = \text{null } T^{m+3} = \cdots .$$

**Proof** Let  $k$  be a positive integer. We want to prove that

$$\text{null } T^{m+k} = \text{null } T^{m+k+1}.$$

We already know from 8.2 that  $\text{null } T^{m+k} \subset \text{null } T^{m+k+1}$ .

To prove the inclusion in the other direction, suppose  $v \in \text{null } T^{m+k+1}$ . Then

$$T^{m+1}(T^k v) = T^{m+k+1}v = 0.$$

Hence

$$T^k v \in \text{null } T^{m+1} = \text{null } T^m.$$

Thus  $T^{m+k}v = T^m(T^k v) = 0$ , which means that  $v \in \text{null } T^{m+k}$ . This implies that  $\text{null } T^{m+k+1} \subset \text{null } T^{m+k}$ , completing the proof. ■

The proposition above raises the question of whether there exists a non-negative integer  $m$  such that  $\text{null } T^m = \text{null } T^{m+1}$ . The proposition below shows that this equality holds at least when  $m$  equals the dimension of the vector space on which  $T$  operates.

#### 8.4 Null spaces stop growing

Suppose  $T \in \mathcal{L}(V)$ . Let  $n = \dim V$ . Then

$$\text{null } T^n = \text{null } T^{n+1} = \text{null } T^{n+2} = \dots .$$

**Proof** We need only prove that  $\text{null } T^n = \text{null } T^{n+1}$  (by 8.3). Suppose this is not true. Then, by 8.2 and 8.3, we have

$$\{0\} = \text{null } T^0 \subsetneq \text{null } T^1 \subsetneq \dots \subsetneq \text{null } T^n \subsetneq \text{null } T^{n+1},$$

where the symbol  $\subsetneq$  means “contained in but not equal to”. At each of the strict inclusions in the chain above, the dimension increases by at least 1. Thus  $\dim \text{null } T^{n+1} \geq n + 1$ , a contradiction because a subspace of  $V$  cannot have a larger dimension than  $n$ . ■

Unfortunately, it is not true that  $V = \text{null } T \oplus \text{range } T$  for each  $T \in \mathcal{L}(V)$ . However, the following result is a useful substitute.

#### 8.5 $V$ is the direct sum of $\text{null } T^{\dim V}$ and $\text{range } T^{\dim V}$

Suppose  $T \in \mathcal{L}(V)$ . Let  $n = \dim V$ . Then

$$V = \text{null } T^n \oplus \text{range } T^n.$$

**Proof** First we show that

$$8.6 \quad (\text{null } T^n) \cap (\text{range } T^n) = \{0\}.$$

Suppose  $v \in (\text{null } T^n) \cap (\text{range } T^n)$ . Then  $T^n v = 0$ , and there exists  $u \in V$  such that  $v = T^n u$ . Applying  $T^n$  to both sides of the last equation shows that  $T^n v = T^{2n} u$ . Hence  $T^{2n} u = 0$ , which implies that  $T^n u = 0$  (by 8.4). Thus  $v = T^n u = 0$ , completing the proof of 8.6.

Now 8.6 implies that  $\text{null } T^n + \text{range } T^n$  is a direct sum (by 1.45). Also,

$$\dim(\text{null } T^n \oplus \text{range } T^n) = \dim \text{null } T^n + \dim \text{range } T^n = \dim V,$$

where the first equality above comes from 3.78 and the second equality comes from the Fundamental Theorem of Linear Maps (3.22). The equation above implies that  $\text{null } T^n \oplus \text{range } T^n = V$ , as desired. ■

---

**8.7 Example** Suppose  $T \in \mathcal{L}(\mathbf{F}^3)$  is defined by

$$T(z_1, z_2, z_3) = (4z_2, 0, 5z_3).$$

For this operator,  $\text{null } T + \text{range } T$  is not a direct sum of subspaces, because  $\text{null } T = \{(z_1, 0, 0) : z_1 \in \mathbf{F}\}$  and  $\text{range } T = \{(z_1, 0, z_3) : z_1, z_3 \in \mathbf{F}\}$ . Thus  $\text{null } T \cap \text{range } T \neq \{0\}$  and hence  $\text{null } T + \text{range } T$  is not a direct sum. Also note that  $\text{null } T + \text{range } T \neq \mathbf{F}^3$ .

However, we have  $T^3(z_1, z_2, z_3) = (0, 0, 125z_3)$ . Thus we see that  $\text{null } T^3 = \{(z_1, z_2, 0) : z_1, z_2 \in \mathbf{F}\}$  and  $\text{range } T^3 = \{(0, 0, z_3) : z_3 \in \mathbf{F}\}$ . Hence  $\mathbf{F}^3 = \text{null } T^3 \oplus \text{range } T^3$ .

---

## Generalized Eigenvectors

Unfortunately, some operators do not have enough eigenvectors to lead to a good description. Thus in this subsection we introduce the concept of generalized eigenvectors, which will play a major role in our description of the structure of an operator.

To understand why we need more than eigenvectors, let's examine the question of describing an operator by decomposing its domain into invariant subspaces. Fix  $T \in \mathcal{L}(V)$ . We seek to describe  $T$  by finding a “nice” direct sum decomposition

$$V = U_1 \oplus \cdots \oplus U_m,$$

where each  $U_j$  is a subspace of  $V$  invariant under  $T$ . The simplest possible nonzero invariant subspaces are 1-dimensional. A decomposition as above where each  $U_j$  is a 1-dimensional subspace of  $V$  invariant under  $T$  is possible if and only if  $V$  has a basis consisting of eigenvectors of  $T$  (see 5.41). This happens if and only if  $V$  has an eigenspace decomposition

$$8.8 \quad V = E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T),$$

where  $\lambda_1, \dots, \lambda_m$  are the distinct eigenvalues of  $T$  (see 5.41).

The Spectral Theorem in the previous chapter shows that if  $V$  is an inner product space, then a decomposition of the form 8.8 holds for every normal operator if  $\mathbf{F} = \mathbf{C}$  and for every self-adjoint operator if  $\mathbf{F} = \mathbf{R}$  because operators of those types have enough eigenvectors to form a basis of  $V$  (see 7.24 and 7.29).

Sadly, a decomposition of the form 8.8 may not hold for more general operators, even on a complex vector space. An example was given by the operator in 5.43, which does not have enough eigenvectors for 8.8 to hold. Generalized eigenvectors and generalized eigenspaces, which we now introduce, will remedy this situation.

### 8.9 Definition *generalized eigenvector*

Suppose  $T \in \mathcal{L}(V)$  and  $\lambda$  is an eigenvalue of  $T$ . A vector  $v \in V$  is called a **generalized eigenvector** of  $T$  corresponding to  $\lambda$  if  $v \neq 0$  and

$$(T - \lambda I)^j v = 0$$

for some positive integer  $j$ .

Although  $j$  is allowed to be an arbitrary integer in the equation

$$(T - \lambda I)^j v = 0$$

in the definition of a generalized eigenvector, we will soon prove that every generalized eigenvector satisfies this equation with  $j = \dim V$ .

*Note that we do not define the concept of a generalized eigenvalue, because this would not lead to anything new. Reason: if  $(T - \lambda I)^j$  is not injective for some positive integer  $j$ , then  $T - \lambda I$  is not injective, and hence  $\lambda$  is an eigenvalue of  $T$ .*

### 8.10 Definition *generalized eigenspace*, $G(\lambda, T)$

Suppose  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ . The **generalized eigenspace** of  $T$  corresponding to  $\lambda$ , denoted  $G(\lambda, T)$ , is defined to be the set of all generalized eigenvectors of  $T$  corresponding to  $\lambda$ , along with the 0 vector.

Because every eigenvector of  $T$  is a generalized eigenvector of  $T$  (take  $j = 1$  in the definition of generalized eigenvector), each eigenspace is contained in the corresponding generalized eigenspace. In other words, if  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ , then

$$E(\lambda, T) \subset G(\lambda, T).$$

The next result implies that if  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ , then  $G(\lambda, T)$  is a subspace of  $V$  (because the null space of each linear map on  $V$  is a subspace of  $V$ ).

### 8.11 Description of generalized eigenspaces

Suppose  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ . Then  $G(\lambda, T) = \text{null}(T - \lambda I)^{\dim V}$ .

**Proof** Suppose  $v \in \text{null}(T - \lambda I)^{\dim V}$ . The definitions imply  $v \in G(\lambda, T)$ . Thus  $G(\lambda, T) \supset \text{null}(T - \lambda I)^{\dim V}$ .

Conversely, suppose  $v \in G(\lambda, T)$ . Thus there is a positive integer  $j$  such that

$$v \in \text{null}(T - \lambda I)^j.$$

From 8.2 and 8.4 (with  $T - \lambda I$  replacing  $T$ ), we get  $v \in \text{null}(T - \lambda I)^{\dim V}$ . Thus  $G(\lambda, T) \subset \text{null}(T - \lambda I)^{\dim V}$ , completing the proof. ■

### 8.12 Example

Define  $T \in \mathcal{L}(\mathbf{C}^3)$  by

$$T(z_1, z_2, z_3) = (4z_2, 0, 5z_3).$$

- (a) Find all eigenvalues of  $T$ , the corresponding eigenspaces, and the corresponding generalized eigenspaces.
- (b) Show that  $\mathbf{C}^3$  is the direct sum of generalized eigenspaces corresponding to the distinct eigenvalues of  $T$ .

#### Solution

- (a) A routine use of the definition of eigenvalue shows that the eigenvalues of  $T$  are 0 and 5. The corresponding eigenspaces are easily seen to be  $E(0, T) = \{(z_1, 0, 0) : z_1 \in \mathbf{C}\}$  and  $E(5, T) = \{(0, 0, z_3) : z_3 \in \mathbf{C}\}$ .

Note that this operator  $T$  does not have enough eigenvectors to span its domain  $\mathbf{C}^3$ .

We have  $T^3(z_1, z_2, z_3) = (0, 0, 125z_3)$  for all  $z_1, z_2, z_3 \in \mathbf{C}$ . Thus 8.11 implies that  $G(0, T) = \{(z_1, z_2, 0) : z_1, z_2 \in \mathbf{C}\}$ .

We have  $(T - 5I)^3(z_1, z_2, z_3) = (-125z_1 + 300z_2, -125z_2, 0)$ . Thus 8.11 implies that  $G(5, T) = \{(0, 0, z_3) : z_3 \in \mathbf{C}\}$ .

- (b) The results in part (a) show that  $\mathbf{C}^3 = G(0, T) \oplus G(5, T)$ .

One of our major goals in this chapter is to show that the result in part (b) of the example above holds in general for operators on finite-dimensional complex vector spaces; we will do this in 8.21.

We saw earlier (5.10) that eigenvectors corresponding to distinct eigenvalues are linearly independent. Now we prove a similar result for generalized eigenvectors.

### 8.13 Linearly independent generalized eigenvectors

Let  $T \in \mathcal{L}(V)$ . Suppose  $\lambda_1, \dots, \lambda_m$  are distinct eigenvalues of  $T$  and  $v_1, \dots, v_m$  are corresponding generalized eigenvectors. Then  $v_1, \dots, v_m$  is linearly independent.

**Proof** Suppose  $a_1, \dots, a_m$  are complex numbers such that

$$\mathbf{8.14} \quad 0 = a_1v_1 + \cdots + a_mv_m.$$

Let  $k$  be the largest nonnegative integer such that  $(T - \lambda_1 I)^k v_1 \neq 0$ . Let

$$w = (T - \lambda_1 I)^k v_1.$$

Thus

$$(T - \lambda_1 I)w = (T - \lambda_1 I)^{k+1}w = 0,$$

and hence  $Tw = \lambda_1 w$ . Thus  $(T - \lambda I)w = (\lambda_1 - \lambda)w$  for every  $\lambda \in \mathbb{F}$  and hence

$$\mathbf{8.15} \quad (T - \lambda I)^n w = (\lambda_1 - \lambda)^n w$$

for every  $\lambda \in \mathbb{F}$ , where  $n = \dim V$ .

Apply the operator

$$(T - \lambda_1 I)^k (T - \lambda_2 I)^n \cdots (T - \lambda_m I)^n$$

to both sides of 8.14, getting

$$\begin{aligned} 0 &= a_1(T - \lambda_1 I)^k (T - \lambda_2 I)^n \cdots (T - \lambda_m I)^n v_1 \\ &= a_1(T - \lambda_2 I)^n \cdots (T - \lambda_m I)^n w \\ &= a_1(\lambda_1 - \lambda_2)^n \cdots (\lambda_1 - \lambda_m)^n w, \end{aligned}$$

where we have used 8.11 to get the first equation above and 8.15 to get the last equation above.

The equation above implies that  $a_1 = 0$ . In a similar fashion,  $a_j = 0$  for each  $j$ , which implies that  $v_1, \dots, v_m$  is linearly independent. ■

## Nilpotent Operators

### 8.16 Definition *nilpotent*

An operator is called *nilpotent* if some power of it equals 0.

### 8.17 Example *nilpotent operators*

- (a) The operator  $N \in \mathcal{L}(\mathbf{F}^4)$  defined by

$$N(z_1, z_2, z_3, z_4) = (z_3, z_4, 0, 0)$$

is nilpotent because  $N^2 = 0$ .

- (b) The operator of differentiation on  $\mathcal{P}_m(\mathbf{R})$  is nilpotent because the  $(m+1)^{\text{st}}$  derivative of every polynomial of degree at most  $m$  equals 0. Note that on this space of dimension  $m+1$ , we need to raise the nilpotent operator to the power  $m+1$  to get the 0 operator.

The Latin word *nil* means nothing or zero; the Latin word *potent* means power. Thus *nilpotent* literally means zero power.

The next result shows that we never need to use a power higher than the dimension of the space.

### 8.18 Nilpotent operator raised to dimension of domain is 0

Suppose  $N \in \mathcal{L}(V)$  is nilpotent. Then  $N^{\dim V} = 0$ .

**Proof** Because  $N$  is nilpotent,  $G(0, N) = V$ . Thus 8.11 implies that  $\text{null } N^{\dim V} = V$ , as desired. ■

Given an operator  $T$  on  $V$ , we want to find a basis of  $V$  such that the matrix of  $T$  with respect to this basis is as simple as possible, meaning that the matrix contains many 0's.

If  $V$  is a complex vector space, a proof of the next result follows easily from Exercise 7, 5.27, and 5.32. But the proof given here uses simpler ideas than needed to prove 5.27, and it works for both real and complex vector spaces.

The next result shows that if  $N$  is nilpotent, then we can choose a basis of  $V$  such that the matrix of  $N$  with respect to this basis has more than half of its entries equal to 0. Later in this chapter we will do even better.

### 8.19 Matrix of a nilpotent operator

Suppose  $N$  is a nilpotent operator on  $V$ . Then there is a basis of  $V$  with respect to which the matrix of  $N$  has the form

$$\begin{pmatrix} 0 & * \\ & \ddots \\ 0 & 0 \end{pmatrix};$$

here all entries on and below the diagonal are 0's.

**Proof** First choose a basis of  $\text{null } N$ . Then extend this to a basis of  $\text{null } N^2$ . Then extend to a basis of  $\text{null } N^3$ . Continue in this fashion, eventually getting a basis of  $V$  (because 8.18 states that  $\text{null } N^{\dim V} = V$ ).

Now let's think about the matrix of  $N$  with respect to this basis. The first column, and perhaps additional columns at the beginning, consists of all 0's, because the corresponding basis vectors are in  $\text{null } N$ . The next set of columns comes from basis vectors in  $\text{null } N^2$ . Applying  $N$  to any such vector, we get a vector in  $\text{null } N$ ; in other words, we get a vector that is a linear combination of the previous basis vectors. Thus all nonzero entries in these columns lie above the diagonal. The next set of columns comes from basis vectors in  $\text{null } N^3$ . Applying  $N$  to any such vector, we get a vector in  $\text{null } N^2$ ; in other words, we get a vector that is a linear combination of the previous basis vectors. Thus once again, all nonzero entries in these columns lie above the diagonal. Continue in this fashion to complete the proof. ■

## EXERCISES 8.A

---

- 1 Define  $T \in \mathcal{L}(\mathbf{C}^2)$  by

$$T(w, z) = (z, 0).$$

Find all generalized eigenvectors of  $T$ .

- 2 Define  $T \in \mathcal{L}(\mathbf{C}^2)$  by

$$T(w, z) = (-z, w).$$

Find the generalized eigenspaces corresponding to the distinct eigenvalues of  $T$ .

- 3** Suppose  $T \in \mathcal{L}(V)$  is invertible. Prove that  $G(\lambda, T) = G\left(\frac{1}{\lambda}, T^{-1}\right)$  for every  $\lambda \in \mathbf{F}$  with  $\lambda \neq 0$ .
- 4** Suppose  $T \in \mathcal{L}(V)$  and  $\alpha, \beta \in \mathbf{F}$  with  $\alpha \neq \beta$ . Prove that

$$G(\alpha, T) \cap G(\beta, T) = \{0\}.$$

- 5** Suppose  $T \in \mathcal{L}(V)$ ,  $m$  is a positive integer, and  $v \in V$  is such that  $T^{m-1}v \neq 0$  but  $T^m v = 0$ . Prove that

$$v, T v, T^2 v, \dots, T^{m-1} v$$

is linearly independent.

- 6** Suppose  $T \in \mathcal{L}(\mathbf{C}^3)$  is defined by  $T(z_1, z_2, z_3) = (z_2, z_3, 0)$ . Prove that  $T$  has no square root. More precisely, prove that there does not exist  $S \in \mathcal{L}(\mathbf{C}^3)$  such that  $S^2 = T$ .
- 7** Suppose  $N \in \mathcal{L}(V)$  is nilpotent. Prove that 0 is the only eigenvalue of  $N$ .
- 8** Prove or give a counterexample: The set of nilpotent operators on  $V$  is a subspace of  $\mathcal{L}(V)$ .
- 9** Suppose  $S, T \in \mathcal{L}(V)$  and  $ST$  is nilpotent. Prove that  $TS$  is nilpotent.
- 10** Suppose that  $T \in \mathcal{L}(V)$  is not nilpotent. Let  $n = \dim V$ . Show that  $V = \text{null } T^{n-1} \oplus \text{range } T^{n-1}$ .
- 11** Prove or give a counterexample: If  $V$  is a complex vector space and  $\dim V = n$  and  $T \in \mathcal{L}(V)$ , then  $T^n$  is diagonalizable.
- 12** Suppose  $N \in \mathcal{L}(V)$  and there exists a basis of  $V$  with respect to which  $N$  has an upper-triangular matrix with only 0's on the diagonal. Prove that  $N$  is nilpotent.
- 13** Suppose  $V$  is an inner product space and  $N \in \mathcal{L}(V)$  is normal and nilpotent. Prove that  $N = 0$ .
- 14** Suppose  $V$  is an inner product space and  $N \in \mathcal{L}(V)$  is nilpotent. Prove that there exists an orthonormal basis of  $V$  with respect to which  $N$  has an upper-triangular matrix.
- [If  $F = \mathbf{C}$ , then the result above follows from Schur's Theorem (6.38) without the hypothesis that  $N$  is nilpotent. Thus the exercise above needs to be proved only when  $\mathbf{F} = \mathbf{R}$ .]

- 15 Suppose  $N \in \mathcal{L}(V)$  is such that  $\text{null } N^{\dim V - 1} \neq \text{null } N^{\dim V}$ . Prove that  $N$  is nilpotent and that

$$\dim \text{null } N^j = j$$

for every integer  $j$  with  $0 \leq j \leq \dim V$ .

- 16 Suppose  $T \in \mathcal{L}(V)$ . Show that

$$V = \text{range } T^0 \supset \text{range } T^1 \supset \cdots \supset \text{range } T^k \supset \text{range } T^{k+1} \supset \cdots.$$

- 17 Suppose  $T \in \mathcal{L}(V)$  and  $m$  is a nonnegative integer such that

$$\text{range } T^m = \text{range } T^{m+1}.$$

Prove that  $\text{range } T^k = \text{range } T^m$  for all  $k > m$ .

- 18 Suppose  $T \in \mathcal{L}(V)$ . Let  $n = \dim V$ . Prove that

$$\text{range } T^n = \text{range } T^{n+1} = \text{range } T^{n+2} = \cdots.$$

- 19 Suppose  $T \in \mathcal{L}(V)$  and  $m$  is a nonnegative integer. Prove that

$$\text{null } T^m = \text{null } T^{m+1} \quad \text{if and only if} \quad \text{range } T^m = \text{range } T^{m+1}.$$

- 20 Suppose  $T \in \mathcal{L}(\mathbf{C}^5)$  is such that  $\text{range } T^4 \neq \text{range } T^5$ . Prove that  $T$  is nilpotent.

- 21 Find a vector space  $W$  and  $T \in \mathcal{L}(W)$  such that  $\text{null } T^k \subsetneq \text{null } T^{k+1}$  and  $\text{range } T^k \supsetneq \text{range } T^{k+1}$  for every positive integer  $k$ .

## 8.B Decomposition of an Operator

### Description of Operators on Complex Vector Spaces

We saw earlier that the domain of an operator might not decompose into eigenspaces, even on a finite-dimensional complex vector space. In this section we will see that every operator on a finite-dimensional complex vector space has enough generalized eigenvectors to provide a decomposition.

We observed earlier that if  $T \in \mathcal{L}(V)$ , then  $\text{null } T$  and  $\text{range } T$  are invariant under  $T$  [see 5.3, parts (c) and (d)]. Now we show that the null space and the range of each polynomial of  $T$  is also invariant under  $T$ .

#### 8.20 The null space and range of $p(T)$ are invariant under $T$

Suppose  $T \in \mathcal{L}(V)$  and  $p \in \mathcal{P}(\mathbb{F})$ . Then  $\text{null } p(T)$  and  $\text{range } p(T)$  are invariant under  $T$ .

**Proof** Suppose  $v \in \text{null } p(T)$ . Then  $p(T)v = 0$ . Thus

$$((p(T))(Tv)) = T(p(T)v) = T(0) = 0.$$

Hence  $Tv \in \text{null } p(T)$ . Thus  $\text{null } p(T)$  is invariant under  $T$ , as desired.

Suppose  $v \in \text{range } p(T)$ . Then there exists  $u \in V$  such that  $v = p(T)u$ . Thus

$$Tv = T(p(T)u) = p(T)(Tu).$$

Hence  $Tv \in \text{range } p(T)$ . Thus  $\text{range } p(T)$  is invariant under  $T$ , as desired. ■

The following major result shows that every operator on a complex vector space can be thought of as composed of pieces, each of which is a nilpotent operator plus a scalar multiple of the identity. Actually we have already done the hard work in our discussion of the generalized eigenspaces  $G(\lambda, T)$ , so at this point the proof is easy.

#### 8.21 Description of operators on complex vector spaces

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Let  $\lambda_1, \dots, \lambda_m$  be the distinct eigenvalues of  $T$ . Then

- (a)  $V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T)$ ;
- (b) each  $G(\lambda_j, T)$  is invariant under  $T$ ;
- (c) each  $(T - \lambda_j I)|_{G(\lambda_j, T)}$  is nilpotent.

**Proof** Let  $n = \dim V$ . Recall that  $G(\lambda_j, T) = \text{null}(T - \lambda_j I)^n$  for each  $j$  (by 8.11). From 8.20 [with  $p(z) = (z - \lambda_j)^n$ ], we get (b). Obviously (c) follows from the definitions.

We will prove (a) by induction on  $n$ . To get started, note that the desired result holds if  $n = 1$ . Thus we can assume that  $n > 1$  and that the desired result holds on all vector spaces of smaller dimension.

Because  $V$  is a complex vector space,  $T$  has an eigenvalue (see 5.21); thus  $m \geq 1$ . Applying 8.5 to  $T - \lambda_1 I$  shows that

$$8.22 \quad V = G(\lambda_1, T) \oplus U,$$

where  $U = \text{range}(T - \lambda_1 I)^n$ . Using 8.20 [with  $p(z) = (z - \lambda_1)^n$ ], we see that  $U$  is invariant under  $T$ . Because  $G(\lambda_1, T) \neq \{0\}$ , we have  $\dim U < n$ . Thus we can apply our induction hypothesis to  $T|_U$ .

None of the generalized eigenvectors of  $T|_U$  correspond to the eigenvalue  $\lambda_1$ , because all generalized eigenvectors of  $T$  corresponding to  $\lambda_1$  are in  $G(\lambda_1, T)$ . Thus each eigenvalue of  $T|_U$  is in  $\{\lambda_2, \dots, \lambda_m\}$ .

By our induction hypothesis,  $U = G(\lambda_2, T|_U) \oplus \dots \oplus G(\lambda_m, T|_U)$ . Combining this information with 8.22 will complete the proof if we can show that  $G(\lambda_k, T|_U) = G(\lambda_k, T)$  for  $k = 2, \dots, m$ .

Thus fix  $k \in \{2, \dots, m\}$ . The inclusion  $G(\lambda_k, T|_U) \subset G(\lambda_k, T)$  is clear.

To prove the inclusion in the other direction, suppose  $v \in G(\lambda_k, T)$ . By 8.22, we can write  $v = v_1 + u$ , where  $v_1 \in G(\lambda_1, T)$  and  $u \in U$ . Our induction hypothesis implies that

$$u = v_2 + \dots + v_m,$$

where each  $v_j$  is in  $G(\lambda_j, T|_U)$ , which is a subset of  $G(\lambda_j, T)$ . Thus

$$v = v_1 + v_2 + \dots + v_m,$$

Because generalized eigenvectors corresponding to distinct eigenvalues are linearly independent (see 8.13), the equation above implies that each  $v_j$  equals 0 except possibly when  $j = k$ . In particular,  $v_1 = 0$  and thus  $v = u \in U$ . Because  $v \in U$ , we can conclude that  $v \in G(\lambda_k, T|_U)$ , completing the proof. ■

As we know, an operator on a complex vector space may not have enough eigenvectors to form a basis of the domain. The next result shows that on a complex vector space there are enough generalized eigenvectors to do this.

### 8.23 A basis of generalized eigenvectors

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Then there is a basis of  $V$  consisting of generalized eigenvectors of  $T$ .

**Proof** Choose a basis of each  $G(\lambda_j, T)$  in 8.21. Put all these bases together to form a basis of  $V$  consisting of generalized eigenvectors of  $T$ . ■

### Multiplicity of an Eigenvalue

If  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ , then the decomposition of  $V$  provided by 8.21 can be a powerful tool. The dimensions of the subspaces involved in this decomposition are sufficiently important to get a name.

#### 8.24 Definition *multiplicity*

- Suppose  $T \in \mathcal{L}(V)$ . The ***multiplicity*** of an eigenvalue  $\lambda$  of  $T$  is defined to be the dimension of the corresponding generalized eigenspace  $G(\lambda, T)$ .
- In other words, the multiplicity of an eigenvalue  $\lambda$  of  $T$  equals  $\dim \text{null}(T - \lambda I)^{\dim V}$ .

The second bullet point above is justified by 8.11.

---

#### 8.25 Example Suppose $T \in \mathcal{L}(\mathbf{C}^3)$ is defined by

$$T(z_1, z_2, z_3) = (6z_1 + 3z_2 + 4z_3, 6z_2 + 2z_3, 7z_3).$$

The matrix of  $T$  (with respect to the standard basis) is

$$\begin{pmatrix} 6 & 3 & 4 \\ 0 & 6 & 2 \\ 0 & 0 & 7 \end{pmatrix}.$$

The eigenvalues of  $T$  are 6 and 7, as follows from 5.32. You can verify that the generalized eigenspaces of  $T$  are as follows:

$$G(6, T) = \text{span}((1, 0, 0), (0, 1, 0)) \quad \text{and} \quad G(7, T) = \text{span}((10, 2, 1)).$$

Thus the eigenvalue 6 has multiplicity 2 and the eigenvalue 7 has multiplicity 1.

The direct sum  $\mathbf{C}^3 = G(6, T) \oplus G(7, T)$  is the decomposition promised by 8.21. A basis of  $\mathbf{C}^3$  consisting of generalized eigenvectors of  $T$ , as promised by 8.23, is

$$(1, 0, 0), (0, 1, 0), (10, 2, 1).$$


---

In Example 8.25, the sum of the multiplicities of the eigenvalues of  $T$  equals 3, which is the dimension of the domain of  $T$ . The next result shows that this always happens on a complex vector space.

### 8.26 Sum of the multiplicities equals $\dim V$

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Then the sum of the multiplicities of all the eigenvalues of  $T$  equals  $\dim V$ .

**Proof** The desired result follows from 8.21 and the obvious formula for the dimension of a direct sum (see 3.78 or Exercise 16 in Section 2.C). ■

The terms **algebraic multiplicity** and **geometric multiplicity** are used in some books. In case you encounter this terminology, be aware that the algebraic multiplicity is the same as the multiplicity defined here and the geometric multiplicity is the dimension of the corresponding eigenspace. In other words, if  $T \in \mathcal{L}(V)$  and  $\lambda$  is an eigenvalue of  $T$ , then

$$\text{algebraic multiplicity of } \lambda = \dim \text{null}(T - \lambda I)^{\dim V} = \dim G(\lambda, T),$$

$$\text{geometric multiplicity of } \lambda = \dim \text{null}(T - \lambda I) = \dim E(\lambda, T).$$

Note that as defined above, the algebraic multiplicity also has a geometric meaning as the dimension of a certain null space. The definition of multiplicity given here is cleaner than the traditional definition that involves determinants; 10.25 implies that these definitions are equivalent.

## Block Diagonal Matrices

To interpret our results in matrix form, we make the following definition, generalizing the notion of a diagonal matrix.

*Often we can understand a matrix better by thinking of it as composed of smaller matrices.*

If each matrix  $A_j$  in the definition below is a 1-by-1 matrix, then we actually have a diagonal matrix.

### 8.27 Definition *block diagonal matrix*

A **block diagonal matrix** is a square matrix of the form

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix},$$

where  $A_1, \dots, A_m$  are square matrices lying along the diagonal and all the other entries of the matrix equal 0.

**8.28 Example** The 5-by-5 matrix

$$A = \begin{pmatrix} (4) & 0 & 0 & 0 & 0 \\ 0 & \begin{pmatrix} 2 & -3 \\ 0 & 2 \end{pmatrix} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \begin{pmatrix} 1 & 7 \\ 0 & 1 \end{pmatrix} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

is a block diagonal matrix with

$$A = \begin{pmatrix} A_1 & & 0 \\ & A_2 & \\ 0 & & A_3 \end{pmatrix},$$

where

$$A_1 = (4), \quad A_2 = \begin{pmatrix} 2 & -3 \\ 0 & 2 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 7 \\ 0 & 1 \end{pmatrix}.$$

Here the inner matrices in the 5-by-5 matrix above are blocked off to show how we can think of it as a block diagonal matrix.

Note that in the next result we get many more zeros in the matrix of  $T$  than are needed to make it upper triangular.

**8.29 Block diagonal matrix with upper-triangular blocks**

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Let  $\lambda_1, \dots, \lambda_m$  be the distinct eigenvalues of  $T$ , with multiplicities  $d_1, \dots, d_m$ . Then there is a basis of  $V$  with respect to which  $T$  has a block diagonal matrix of the form

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix},$$

where each  $A_j$  is a  $d_j$ -by- $d_j$  upper-triangular matrix of the form

$$A_j = \begin{pmatrix} \lambda_j & * & \\ & \ddots & \\ 0 & & \lambda_j \end{pmatrix}.$$

**Proof** Each  $(T - \lambda_j I)|_{G(\lambda_j, T)}$  is nilpotent [see 8.21(c)]. For each  $j$ , choose a basis of  $G(\lambda_j, T)$ , which is a vector space with dimension  $d_j$ , such that the matrix of  $(T - \lambda_j I)|_{G(\lambda_j, T)}$  with respect to this basis is as in 8.19. Thus the matrix of  $T|_{G(\lambda_j, T)}$ , which equals  $(T - \lambda_j I)|_{G(\lambda_j, T)} + \lambda_j I|_{G(\lambda_j, T)}$ , with respect to this basis will look like the desired form shown above for  $A_j$ .

Putting the bases of the  $G(\lambda_j, T)$ 's together gives a basis of  $V$  [by 8.21(a)]. The matrix of  $T$  with respect to this basis has the desired form. ■

The 5-by-5 matrix in 8.28 is of the form promised by 8.29, with each of the blocks itself an upper-triangular matrix that is constant along the diagonal of the block. If  $T$  is an operator on a 5-dimensional vector space whose matrix is as in 8.28, then the eigenvalues of  $T$  are 4, 2, 1 (as follows from 5.32), with multiplicities 1, 2, 2.

### 8.30 Example

Suppose  $T \in \mathcal{L}(\mathbb{C}^3)$  is defined by

$$T(z_1, z_2, z_3) = (6z_1 + 3z_2 + 4z_3, 6z_2 + 2z_3, 7z_3).$$

The matrix of  $T$  (with respect to the standard basis) is

$$\begin{pmatrix} 6 & 3 & 4 \\ 0 & 6 & 2 \\ 0 & 0 & 7 \end{pmatrix},$$

which is an upper-triangular matrix but is not of the form promised by 8.29.

As we saw in Example 8.25, the eigenvalues of  $T$  are 6 and 7 and the corresponding generalized eigenspaces are

$$G(6, T) = \text{span}((1, 0, 0), (0, 1, 0)) \quad \text{and} \quad G(7, T) = \text{span}((10, 2, 1)).$$

We also saw that a basis of  $\mathbb{C}^3$  consisting of generalized eigenvectors of  $T$  is

$$(1, 0, 0), (0, 1, 0), (10, 2, 1).$$

The matrix of  $T$  with respect to this basis is

$$\begin{pmatrix} \begin{pmatrix} 6 & 3 \\ 0 & 6 \\ 0 & 0 \end{pmatrix} & 0 \\ 0 & \begin{pmatrix} 7 \end{pmatrix} \end{pmatrix},$$

which is a matrix of the block diagonal form promised by 8.29.

When we discuss the Jordan Form in Section 8.D, we will see that we can find a basis with respect to which an operator  $T$  has a matrix with even more 0's than promised by 8.29. However, 8.29 and its equivalent companion 8.21 are already quite powerful. For example, in the next subsection we will use 8.21 to show that every invertible operator on a complex vector space has a square root.

## Square Roots

Recall that a square root of an operator  $T \in \mathcal{L}(V)$  is an operator  $R \in \mathcal{L}(V)$  such that  $R^2 = T$  (see 7.33). Every complex number has a square root, but not every operator on a complex vector space has a square root. For example, the operator on  $\mathbf{C}^3$  in Exercise 6 in Section 8.A has no square root. The noninvertibility of that operator is no accident, as we will soon see. We begin by showing that the identity plus any nilpotent operator has a square root.

### 8.31 Identity plus nilpotent has a square root

Suppose  $N \in \mathcal{L}(V)$  is nilpotent. Then  $I + N$  has a square root.

**Proof** Consider the Taylor series for the function  $\sqrt{1+x}$ :

$$8.32 \quad \sqrt{1+x} = 1 + a_1x + a_2x^2 + \dots.$$

Because  $a_1 = 1/2$ , the formula above shows that  $1+x/2$  is a good estimate for  $\sqrt{1+x}$  when  $x$  is small.

We will not find an explicit formula for the coefficients or worry about whether the infinite sum converges because we will use this equation only as motivation.

Because  $N$  is nilpotent,  $N^m = 0$  for some positive integer  $m$ . In 8.32, suppose we replace  $x$  with  $N$  and 1 with  $I$ . Then the infinite sum on the right side becomes a finite sum (because  $N^j = 0$  for all  $j \geq m$ ). In other words, we guess that there is a square root of  $I + N$  of the form

$$I + a_1N + a_2N^2 + \dots + a_{m-1}N^{m-1}.$$

Having made this guess, we can try to choose  $a_1, a_2, \dots, a_{m-1}$  such that the operator above has its square equal to  $I + N$ . Now

$$\begin{aligned} & (I + a_1N + a_2N^2 + a_3N^3 + \dots + a_{m-1}N^{m-1})^2 \\ &= I + 2a_1N + (2a_2 + a_1^2)N^2 + (2a_3 + 2a_1a_2)N^3 + \dots \\ & \quad + (2a_{m-1} + \text{terms involving } a_1, \dots, a_{m-2})N^{m-1}. \end{aligned}$$

We want the right side of the equation above to equal  $I + N$ . Hence choose  $a_1$  such that  $2a_1 = 1$  (thus  $a_1 = 1/2$ ). Next, choose  $a_2$  such that  $2a_2 + a_1^2 = 0$  (thus  $a_2 = -1/8$ ). Then choose  $a_3$  such that the coefficient of  $N^3$  on the right side of the equation above equals 0 (thus  $a_3 = 1/16$ ). Continue in this fashion for  $j = 4, \dots, m-1$ , at each step solving for  $a_j$  so that the coefficient of  $N^j$  on the right side of the equation above equals 0. Actually we do not care about the explicit formula for the  $a_j$ 's. We need only know that some choice of the  $a_j$ 's gives a square root of  $I + N$ . ■

The previous lemma is valid on real and complex vector spaces. However, the next result holds only on complex vector spaces. For example, the operator of multiplication by  $-1$  on the 1-dimensional real vector space  $\mathbf{R}$  has no square root.

### 8.33 Over $\mathbf{C}$ , invertible operators have square roots

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$  is invertible. Then  $T$  has a square root.

**Proof** Let  $\lambda_1, \dots, \lambda_m$  be the distinct eigenvalues of  $T$ . For each  $j$ , there exists a nilpotent operator  $N_j \in \mathcal{L}(G(\lambda_j, T))$  such that  $T|_{G(\lambda_j, T)} = \lambda_j I + N_j$  [see 8.21(c)]. Because  $T$  is invertible, none of the  $\lambda_j$ 's equals 0, so we can write

$$T|_{G(\lambda_j, T)} = \lambda_j \left( I + \frac{N_j}{\lambda_j} \right)$$

for each  $j$ . Clearly  $N_j/\lambda_j$  is nilpotent, and so  $I + N_j/\lambda_j$  has a square root (by 8.31). Multiplying a square root of the complex number  $\lambda_j$  by a square root of  $I + N_j/\lambda_j$ , we obtain a square root  $R_j$  of  $T|_{G(\lambda_j, T)}$ .

A typical vector  $v \in V$  can be written uniquely in the form

$$v = u_1 + \cdots + u_m,$$

where each  $u_j$  is in  $G(\lambda_j, T)$  (see 8.21). Using this decomposition, define an operator  $R \in \mathcal{L}(V)$  by

$$Rv = R_1 u_1 + \cdots + R_m u_m.$$

You should verify that this operator  $R$  is a square root of  $T$ , completing the proof. ■

By imitating the techniques in this section, you should be able to prove that if  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$  is invertible, then  $T$  has a  $k^{\text{th}}$  root for every positive integer  $k$ .

## EXERCISES 8.B

---

- Suppose  $V$  is a complex vector space,  $N \in \mathcal{L}(V)$ , and 0 is the only eigenvalue of  $N$ . Prove that  $N$  is nilpotent.
- Give an example of an operator  $T$  on a finite-dimensional real vector space such that 0 is the only eigenvalue of  $T$  but  $T$  is not nilpotent.

- 3 Suppose  $T \in \mathcal{L}(V)$ . Suppose  $S \in \mathcal{L}(V)$  is invertible. Prove that  $T$  and  $S^{-1}TS$  have the same eigenvalues with the same multiplicities.
- 4 Suppose  $V$  is an  $n$ -dimensional complex vector space and  $T$  is an operator on  $V$  such that  $\text{null } T^{n-2} \neq \text{null } T^{n-1}$ . Prove that  $T$  has at most two distinct eigenvalues.
- 5 Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Prove that  $V$  has a basis consisting of eigenvectors of  $T$  if and only if every generalized eigenvector of  $T$  is an eigenvector of  $T$ .
- [For  $\mathbf{F} = \mathbf{C}$ , the exercise above adds an equivalence to the list in 5.41.]
- 6 Define  $N \in \mathcal{L}(\mathbf{F}^5)$  by

$$N(x_1, x_2, x_3, x_4, x_5) = (2x_2, 3x_3, -x_4, 4x_5, 0).$$

Find a square root of  $I + N$ .

- 7 Suppose  $V$  is a complex vector space. Prove that every invertible operator on  $V$  has a cube root.
- 8 Suppose  $T \in \mathcal{L}(V)$  and 3 and 8 are eigenvalues of  $T$ . Let  $n = \dim V$ . Prove that  $V = (\text{null } T^{n-2}) \oplus (\text{range } T^{n-2})$ .
- 9 Suppose  $A$  and  $B$  are block diagonal matrices of the form

$$A = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & & 0 \\ & \ddots & \\ 0 & & B_m \end{pmatrix},$$

where  $A_j$  has the same size as  $B_j$  for  $j = 1, \dots, m$ . Show that  $AB$  is a block diagonal matrix of the form

$$AB = \begin{pmatrix} A_1 B_1 & & 0 \\ & \ddots & \\ 0 & & A_m B_m \end{pmatrix}.$$

- 10 Suppose  $\mathbf{F} = \mathbf{C}$  and  $T \in \mathcal{L}(V)$ . Prove that there exist  $D, N \in \mathcal{L}(V)$  such that  $T = D + N$ , the operator  $D$  is diagonalizable,  $N$  is nilpotent, and  $DN = ND$ .
- 11 Suppose  $T \in \mathcal{L}(V)$  and  $\lambda \in \mathbf{F}$ . Prove that for every basis of  $V$  with respect to which  $T$  has an upper-triangular matrix, the number of times that  $\lambda$  appears on the diagonal of the matrix of  $T$  equals the multiplicity of  $\lambda$  as an eigenvalue of  $T$ .

## 8.C Characteristic and Minimal Polynomials

### The Cayley–Hamilton Theorem

The next definition associates a polynomial with each operator on  $V$  if  $\mathbf{F} = \mathbf{C}$ . For  $\mathbf{F} = \mathbf{R}$ , the corresponding definition will be given in the next chapter.

#### 8.34 Definition *characteristic polynomial*

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Let  $\lambda_1, \dots, \lambda_m$  denote the distinct eigenvalues of  $T$ , with multiplicities  $d_1, \dots, d_m$ . The polynomial

$$(z - \lambda_1)^{d_1} \cdots (z - \lambda_m)^{d_m}$$

is called the *characteristic polynomial* of  $T$ .

**8.35 Example** Suppose  $T \in \mathcal{L}(\mathbf{C}^3)$  is defined as in Example 8.25. Because the eigenvalues of  $T$  are 6, with multiplicity 2, and 7, with multiplicity 1, we see that the characteristic polynomial of  $T$  is  $(z - 6)^2(z - 7)$ .

#### 8.36 Degree and zeros of characteristic polynomial

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Then

- the characteristic polynomial of  $T$  has degree  $\dim V$ ;
- the zeros of the characteristic polynomial of  $T$  are the eigenvalues of  $T$ .

**Proof** Clearly part (a) follows from 8.26 and part (b) follows from the definition of the characteristic polynomial. ■

Most texts define the characteristic polynomial using determinants (the two definitions are equivalent by 10.25). The approach taken here, which is considerably simpler, leads to the following easy proof of the Cayley–Hamilton Theorem. In the next chapter, we will see that this result also holds on real vector spaces (see 9.24).

#### 8.37 Cayley–Hamilton Theorem

Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Let  $q$  denote the characteristic polynomial of  $T$ . Then  $q(T) = 0$ .

English mathematician Arthur Cayley (1821–1895) published three math papers before completing his undergraduate degree in 1842. Irish mathematician William Rowan Hamilton (1805–1865) was made a professor in 1827 when he was 22 years old and still an undergraduate!

**Proof** Let  $\lambda_1, \dots, \lambda_m$  be the distinct eigenvalues of the operator  $T$ , and let  $d_1, \dots, d_m$  be the dimensions of the corresponding generalized eigenspaces  $G(\lambda_1, T), \dots, G(\lambda_m, T)$ . For each  $j \in \{1, \dots, m\}$ , we know that  $(T - \lambda_j I)|_{G(\lambda_j, T)}$  is nilpotent. Thus we have  $(T - \lambda_j I)^{d_j}|_{G(\lambda_j, T)} = 0$  (by 8.18).

Every vector in  $V$  is a sum of vectors in  $G(\lambda_1, T), \dots, G(\lambda_m, T)$  (by 8.21). Thus to prove that  $q(T) = 0$ , we need only show that  $q(T)|_{G(\lambda_j, T)} = 0$  for each  $j$ .

Thus fix  $j \in \{1, \dots, m\}$ . We have

$$q(T) = (T - \lambda_1 I)^{d_1} \cdots (T - \lambda_m I)^{d_m}.$$

The operators on the right side of the equation above all commute, so we can move the factor  $(T - \lambda_j I)^{d_j}$  to be the last term in the expression on the right. Because  $(T - \lambda_j I)^{d_j}|_{G(\lambda_j, T)} = 0$ , we conclude that  $q(T)|_{G(\lambda_j, T)} = 0$ , as desired. ■

## The Minimal Polynomial

In this subsection we introduce another important polynomial associated with each operator. We begin with the following definition.

### 8.38 Definition monic polynomial

A **monic polynomial** is a polynomial whose highest-degree coefficient equals 1.

---

**8.39 Example** The polynomial  $2 + 9z^2 + z^7$  is a monic polynomial of degree 7.

---

### 8.40 Minimal polynomial

Suppose  $T \in \mathcal{L}(V)$ . Then there is a unique monic polynomial  $p$  of smallest degree such that  $p(T) = 0$ .

**Proof** Let  $n = \dim V$ . Then the list

$$I, T, T^2, \dots, T^{n^2}$$

is not linearly independent in  $\mathcal{L}(V)$ , because the vector space  $\mathcal{L}(V)$  has dimension  $n^2$  (see 3.61) and we have a list of length  $n^2 + 1$ . Let  $m$  be the smallest positive integer such that the list

$$\mathbf{8.41} \quad I, T, T^2, \dots, T^m$$

is linearly dependent. The Linear Dependence Lemma (2.21) implies that one of the operators in the list above is a linear combination of the previous ones. Because  $m$  was chosen to be the smallest positive integer such that the list above is linearly dependent, we conclude that  $T^m$  is a linear combination of  $I, T, T^2, \dots, T^{m-1}$ . Thus there exist scalars  $a_0, a_1, a_2, \dots, a_{m-1} \in \mathbf{F}$  such that

$$\mathbf{8.42} \quad a_0 I + a_1 T + a_2 T^2 + \cdots + a_{m-1} T^{m-1} + T^m = 0.$$

Define a monic polynomial  $p \in \mathcal{P}(\mathbf{F})$  by

$$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_{m-1} z^{m-1} + z^m.$$

Then 8.42 implies that  $p(T) = 0$ .

To prove the uniqueness part of the result, note that the choice of  $m$  implies that no monic polynomial  $q \in \mathcal{P}(\mathbf{F})$  with degree smaller than  $m$  can satisfy  $q(T) = 0$ . Suppose  $q \in \mathcal{P}(\mathbf{F})$  is a monic polynomial with degree  $m$  and  $q(T) = 0$ . Then  $(p - q)(T) = 0$  and  $\deg(p - q) < m$ . The choice of  $m$  now implies that  $q = p$ , completing the proof. ■

The last result justifies the following definition.

### 8.43 Definition *minimal polynomial*

Suppose  $T \in \mathcal{L}(V)$ . Then the **minimal polynomial** of  $T$  is the unique monic polynomial  $p$  of smallest degree such that  $p(T) = 0$ .

The proof of the last result shows that the degree of the minimal polynomial of each operator on  $V$  is at most  $(\dim V)^2$ . The Cayley–Hamilton Theorem (8.37) tells us that if  $V$  is a complex vector space, then the minimal polynomial of each operator on  $V$  has degree at most  $\dim V$ . This remarkable improvement also holds on real vector spaces, as we will see in the next chapter.

Suppose you are given the matrix (with respect to some basis) of an operator  $T \in \mathcal{L}(V)$ . You could program a computer to find the minimal polynomial of  $T$  as follows: Consider the system of linear equations

$$\mathbf{8.44} \quad a_0\mathcal{M}(I) + a_1\mathcal{M}(T) + \cdots + a_{m-1}\mathcal{M}(T)^{m-1} = -\mathcal{M}(T)^m$$

*Think of this as a system of  $(\dim V)^2$  linear equations in  $m$  variables  $a_0, a_1, \dots, a_{m-1}$ .*

for successive values of  $m = 1, 2, \dots$  until this system of equations has a solution  $a_0, a_1, a_2, \dots, a_{m-1}$ . The scalars

$a_0, a_1, a_2, \dots, a_{m-1}, 1$  will then be the

coefficients of the minimal polynomial of  $T$ . All this can be computed using a familiar and fast (for a computer) process such as Gaussian elimination.

**8.45 Example** Let  $T$  be the operator on  $\mathbf{C}^5$  whose matrix (with respect to the standard basis) is

$$\begin{pmatrix} 0 & 0 & 0 & 0 & -3 \\ 1 & 0 & 0 & 0 & 6 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Find the minimal polynomial of  $T$ .

**Solution** Because of the large number of 0's in this matrix, Gaussian elimination is not needed here. Simply compute powers of  $\mathcal{M}(T)$ , and then you will notice that there is clearly no solution to 8.44 until  $m = 5$ . Do the computations and you will see that the minimal polynomial of  $T$  equals  $z^5 - 6z + 3$ .

The next result completely characterizes the polynomials that when applied to an operator give the 0 operator.

#### 8.46 $q(T) = 0$ implies $q$ is a multiple of the minimal polynomial

Suppose  $T \in \mathcal{L}(V)$  and  $q \in \mathcal{P}(\mathbf{F})$ . Then  $q(T) = 0$  if and only if  $q$  is a polynomial multiple of the minimal polynomial of  $T$ .

**Proof** Let  $p$  denote the minimal polynomial of  $T$ .

First we prove the easy direction. Suppose  $q$  is a polynomial multiple of  $p$ . Thus there exists a polynomial  $s \in \mathcal{P}(\mathbf{F})$  such that  $q = ps$ . We have

$$q(T) = p(T)s(T) = 0 s(T) = 0,$$

as desired.

To prove the other direction, now suppose  $q(T) = 0$ . By the Division Algorithm for Polynomials (4.8), there exist polynomials  $s, r \in \mathcal{P}(\mathbf{F})$  such that

**8.47**

$$q = ps + r$$

and  $\deg r < \deg p$ . We have

$$0 = q(T) = p(T)s(T) + r(T) = r(T).$$

The equation above implies that  $r = 0$  (otherwise, dividing  $r$  by its highest-degree coefficient would produce a monic polynomial that when applied to  $T$  gives 0; this polynomial would have a smaller degree than the minimal polynomial, which would be a contradiction). Thus 8.47 becomes the equation  $q = ps$ . Hence  $q$  is a polynomial multiple of  $p$ , as desired. ■

The next result is stated only for complex vector spaces, because we have not yet defined the characteristic polynomial when  $\mathbf{F} = \mathbf{R}$ . However, the result also holds for real vector spaces, as we will see in the next chapter.

#### 8.48 Characteristic polynomial is a multiple of minimal polynomial

Suppose  $\mathbf{F} = \mathbf{C}$  and  $T \in \mathcal{L}(V)$ . Then the characteristic polynomial of  $T$  is a polynomial multiple of the minimal polynomial of  $T$ .

**Proof** The desired result follows immediately from the Cayley–Hamilton Theorem (8.37) and 8.46. ■

We know (at least when  $\mathbf{F} = \mathbf{C}$ ) that the zeros of the characteristic polynomial of  $T$  are the eigenvalues of  $T$  (see 8.36). Now we show that the minimal polynomial has the same zeros (although the multiplicities of these zeros may differ).

#### 8.49 Eigenvalues are the zeros of the minimal polynomial

Let  $T \in \mathcal{L}(V)$ . Then the zeros of the minimal polynomial of  $T$  are precisely the eigenvalues of  $T$ .

**Proof** Let

$$p(z) = a_0 + a_1z + a_2z^2 + \cdots + a_{m-1}z^{m-1} + z^m$$

be the minimal polynomial of  $T$ .

First suppose  $\lambda \in \mathbf{F}$  is a zero of  $p$ . Then  $p$  can be written in the form

$$p(z) = (z - \lambda)q(z),$$

where  $q$  is a monic polynomial with coefficients in  $\mathbf{F}$  (see 4.11). Because  $p(T) = 0$ , we have

$$0 = (T - \lambda I)(q(T)v)$$

for all  $v \in V$ . Because the degree of  $q$  is less than the degree of the minimal polynomial  $p$ , there exists at least one vector  $v \in V$  such that  $q(T)v \neq 0$ . The equation above thus implies that  $\lambda$  is an eigenvalue of  $T$ , as desired.

To prove the other direction, now suppose  $\lambda \in \mathbf{F}$  is an eigenvalue of  $T$ . Thus there exists  $v \in V$  with  $v \neq 0$  such that  $Tv = \lambda v$ . Repeated applications of  $T$  to both sides of this equation show that  $T^j v = \lambda^j v$  for every nonnegative integer  $j$ . Thus

$$\begin{aligned} 0 &= p(T)v = (a_0I + a_1T + a_2T^2 + \cdots + a_{m-1}T^{m-1} + T^m)v \\ &= (a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_{m-1}\lambda^{m-1} + \lambda^m)v \\ &= p(\lambda)v. \end{aligned}$$

Because  $v \neq 0$ , the equation above implies that  $p(\lambda) = 0$ , as desired. ■

The next three examples show how our results can be useful in finding minimal polynomials and in understanding why eigenvalues of some operators cannot be exactly computed.

**8.50 Example** Find the minimal polynomial of the operator  $T \in \mathcal{L}(\mathbf{C}^3)$  in Example 8.30.

**Solution** In Example 8.30 we noted that the eigenvalues of  $T$  are 6 and 7. Thus by 8.49, the minimal polynomial of  $T$  is a polynomial multiple of  $(z - 6)(z - 7)$ .

In Example 8.35, we saw that the characteristic polynomial of  $T$  is  $(z - 6)^2(z - 7)$ . Thus by 8.48 and the paragraph above, the minimal polynomial of  $T$  is either  $(z - 6)(z - 7)$  or  $(z - 6)^2(z - 7)$ . A simple computation shows that

$$(T - 6I)(T - 7I) \neq 0.$$

Thus the minimal polynomial of  $T$  is  $(z - 6)^2(z - 7)$ .

---

**8.51 Example** Find the minimal polynomial of the operator  $T \in \mathcal{L}(\mathbf{C}^3)$  defined by  $T(z_1, z_2, z_3) = (6z_1, 6z_2, 7z_3)$ .

**Solution** It is easy to see that for this operator  $T$ , the eigenvalues of  $T$  are 6 and 7, and the characteristic polynomial of  $T$  is  $(z - 6)^2(z - 7)$ .

Thus as in the previous example, the minimal polynomial of  $T$  is either  $(z - 6)(z - 7)$  or  $(z - 6)^2(z - 7)$ . A simple computation shows that  $(T - 6I)(T - 7I) = 0$ . Thus the minimal polynomial of  $T$  is  $(z - 6)(z - 7)$ .

---

**8.52 Example** What are the eigenvalues of the operator in Example 8.45?

**Solution** From 8.49 and the solution to Example 8.45, we see that the eigenvalues of  $T$  equal the solutions to the equation

$$z^5 - 6z + 3 = 0.$$

Unfortunately, no solution to this equation can be computed using rational numbers, roots of rational numbers, and the usual rules of arithmetic (a proof of this would take us considerably beyond linear algebra). Thus we cannot find an exact expression for any eigenvalue of  $T$  in any familiar form, although numeric techniques can give good approximations for the eigenvalues of  $T$ . The numeric techniques, which we will not discuss here, show that the eigenvalues for this particular operator are approximately

$$-1.67, \quad 0.51, \quad 1.40, \quad -0.12 + 1.59i, \quad -0.12 - 1.59i.$$

The nonreal eigenvalues occur as a pair, with each the complex conjugate of the other, as expected for a polynomial with real coefficients (see 4.15).

---

## EXERCISES 8.C

---

- 1 Suppose  $T \in \mathcal{L}(\mathbf{C}^4)$  is such that the eigenvalues of  $T$  are 3, 5, 8. Prove that  $(T - 3I)^2(T - 5I)^2(T - 8I)^2 = 0$ .
- 2 Suppose  $V$  is a complex vector space. Suppose  $T \in \mathcal{L}(V)$  is such that 5 and 6 are eigenvalues of  $T$  and that  $T$  has no other eigenvalues. Prove that  $(T - 5I)^{n-1}(T - 6I)^{n-1} = 0$ , where  $n = \dim V$ .
- 3 Give an example of an operator on  $\mathbf{C}^4$  whose characteristic polynomial equals  $(z - 7)^2(z - 8)^2$ .

- 4 Give an example of an operator on  $\mathbf{C}^4$  whose characteristic polynomial equals  $(z - 1)(z - 5)^3$  and whose minimal polynomial equals  $(z - 1)(z - 5)^2$ .
- 5 Give an example of an operator on  $\mathbf{C}^4$  whose characteristic and minimal polynomials both equal  $z(z - 1)^2(z - 3)$ .
- 6 Give an example of an operator on  $\mathbf{C}^4$  whose characteristic polynomial equals  $z(z - 1)^2(z - 3)$  and whose minimal polynomial equals  $z(z - 1)(z - 3)$ .
- 7 Suppose  $V$  is a complex vector space. Suppose  $T \in \mathcal{L}(V)$  is such that  $P^2 = P$ . Prove that the characteristic polynomial of  $P$  is  $z^m(z - 1)^n$ , where  $m = \dim \text{null } P$  and  $n = \dim \text{range } P$ .
- 8 Suppose  $T \in \mathcal{L}(V)$ . Prove that  $T$  is invertible if and only if the constant term in the minimal polynomial of  $T$  is nonzero.
- 9 Suppose  $T \in \mathcal{L}(V)$  has minimal polynomial  $4 + 5z - 6z^2 - 7z^3 + 2z^4 + z^5$ . Find the minimal polynomial of  $T^{-1}$ .
- 10 Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$  is invertible. Let  $p$  denote the characteristic polynomial of  $T$  and let  $q$  denote the characteristic polynomial of  $T^{-1}$ . Prove that

$$q(z) = \frac{1}{p(0)} z^{\dim V} p\left(\frac{1}{z}\right)$$

for all nonzero  $z \in \mathbf{C}$ .

- 11 Suppose  $T \in \mathcal{L}(V)$  is invertible. Prove that there exists a polynomial  $p \in \mathcal{P}(\mathbf{F})$  such that  $T^{-1} = p(T)$ .
- 12 Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Prove that  $V$  has a basis consisting of eigenvectors of  $T$  if and only if the minimal polynomial of  $T$  has no repeated zeros.  
*[For complex vector spaces, the exercise above adds another equivalence to the list given by 5.41.]*
- 13 Suppose  $V$  is an inner product space and  $T \in \mathcal{L}(V)$  is normal. Prove that the minimal polynomial of  $T$  has no repeated zeros.
- 14 Suppose  $V$  is a complex inner product space and  $S \in \mathcal{L}(V)$  is an isometry. Prove that the constant term in the characteristic polynomial of  $S$  has absolute value 1.

**15** Suppose  $T \in \mathcal{L}(V)$  and  $v \in V$ .

- (a) Prove that there exists a unique monic polynomial  $p$  of smallest degree such that  $p(T)v = 0$ .
- (b) Prove that  $p$  divides the minimal polynomial of  $T$ .

**16** Suppose  $V$  is an inner product space and  $T \in \mathcal{L}(V)$ . Suppose

$$a_0 + a_1 z + a_2 z^2 + \cdots + a_{m-1} z^{m-1} + z^m$$

is the minimal polynomial of  $T$ . Prove that

$$\overline{a_0} + \overline{a_1} z + \overline{a_2} z^2 + \cdots + \overline{a_{m-1}} z^{m-1} + z^m$$

is the minimal polynomial of  $T^*$ .

**17** Suppose  $\mathbf{F} = \mathbf{C}$  and  $T \in \mathcal{L}(V)$ . Suppose the minimal polynomial of  $T$  has degree  $\dim V$ . Prove that the characteristic polynomial of  $T$  equals the minimal polynomial of  $T$ .

**18** Suppose  $a_0, \dots, a_{n-1} \in \mathbf{C}$ . Find the minimal and characteristic polynomials of the operator on  $\mathbf{C}^n$  whose matrix (with respect to the standard basis) is

$$\begin{pmatrix} 0 & -a_0 \\ 1 & 0 & -a_1 \\ & 1 & \ddots & -a_2 \\ & & \ddots & \vdots \\ & & & 0 & -a_{n-2} \\ & & & & 1 & -a_{n-1} \end{pmatrix}.$$

[The exercise above shows that every monic polynomial is the characteristic polynomial of some operator.]

**19** Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Suppose that with respect to some basis of  $V$  the matrix of  $T$  is upper triangular, with  $\lambda_1, \dots, \lambda_n$  on the diagonal of this matrix. Prove that the characteristic polynomial of  $T$  is  $(z - \lambda_1) \cdots (z - \lambda_n)$ .

**20** Suppose  $V$  is a complex vector space and  $V_1, \dots, V_m$  are nonzero subspaces of  $V$  such that  $V = V_1 \oplus \cdots \oplus V_m$ . Suppose  $T \in \mathcal{L}(V)$  and each  $V_j$  is invariant under  $T$ . For each  $j$ , let  $p_j$  denote the characteristic polynomial of  $T|_{V_j}$ . Prove that the characteristic polynomial of  $T$  equals  $p_1 \cdots p_m$ .

## 8.D Jordan Form

We know that if  $V$  is a complex vector space, then for every  $T \in \mathcal{L}(V)$  there is a basis of  $V$  with respect to which  $T$  has a nice upper-triangular matrix (see 8.29). In this section we will see that we can do even better—there is a basis of  $V$  with respect to which the matrix of  $T$  contains 0's everywhere except possibly on the diagonal and the line directly above the diagonal.

We begin by looking at two examples of nilpotent operators.

**8.53 Example** Let  $N \in \mathcal{L}(\mathbb{F}^4)$  be the nilpotent operator defined by

$$N(z_1, z_2, z_3, z_4) = (0, z_1, z_2, z_3).$$

If  $v = (1, 0, 0, 0)$ , then  $N^3v, N^2v, Nv, v$  is a basis of  $\mathbb{F}^4$ . The matrix of  $N$  with respect to this basis is

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The next example of a nilpotent operator has more complicated behavior than the example above.

**8.54 Example** Let  $N \in \mathcal{L}(\mathbb{F}^6)$  be the nilpotent operator defined by

$$N(z_1, z_2, z_3, z_4, z_5, z_6) = (0, z_1, z_2, 0, z_4, 0).$$

Unlike the nice behavior of the nilpotent operator of the previous example, for this nilpotent operator there does not exist a vector  $v \in \mathbb{F}^6$  such that  $N^5v, N^4v, N^3v, N^2v, Nv, v$  is a basis of  $\mathbb{F}^6$ . However, if we take  $v_1 = (1, 0, 0, 0, 0, 0)$ ,  $v_2 = (0, 0, 0, 1, 0, 0)$ , and  $v_3 = (0, 0, 0, 0, 0, 1)$ , then  $N^2v_1, Nv_1, v_1, Nv_2, v_2, v_3$  is a basis of  $\mathbb{F}^6$ . The matrix of  $N$  with respect to this basis is

$$\begin{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} & 0 & 0 & 0 \\ 0 & \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} & 0 \\ 0 & 0 & \begin{pmatrix} 0 \end{pmatrix} \end{pmatrix}.$$

Here the inner matrices are blocked off to show that we can think of the 6-by-6 matrix above as a block diagonal matrix consisting of a 3-by-3 block with 1's on the line above the diagonal and 0's elsewhere, a 2-by-2 block with 1 above the diagonal and 0's elsewhere, and a 1-by-1 block containing 0.

Our next result shows that every nilpotent operator  $N \in \mathcal{L}(V)$  behaves similarly to the previous example. Specifically, there is a finite collection of vectors  $v_1, \dots, v_n \in V$  such that there is a basis of  $V$  consisting of the vectors of the form  $N^k v_j$ , as  $j$  varies from 1 to  $n$  and  $k$  varies (in reverse order) from 0 to the largest nonnegative integer  $m_j$  such that  $N^{m_j} v_j \neq 0$ . For the matrix interpretation of the next result, see the first part of the proof of 8.60.

### 8.55 Basis corresponding to a nilpotent operator

Suppose  $N \in \mathcal{L}(V)$  is nilpotent. Then there exist vectors  $v_1, \dots, v_n \in V$  and nonnegative integers  $m_1, \dots, m_n$  such that

- (a)  $N^{m_1} v_1, \dots, N v_1, v_1, \dots, N^{m_n} v_n, \dots, N v_n, v_n$  is a basis of  $V$ ;
- (b)  $N^{m_1+1} v_1 = \dots = N^{m_n+1} v_n = 0$ .

**Proof** We will prove this result by induction on  $\dim V$ . To get started, note that the desired result obviously holds if  $\dim V = 1$  (in that case, the only nilpotent operator is the 0 operator, so take  $v_1$  to be any nonzero vector and  $m_1 = 0$ ). Now assume that  $\dim V > 1$  and that the desired result holds on all vector spaces of smaller dimension.

Because  $N$  is nilpotent,  $N$  is not injective. Thus  $N$  is not surjective (by 3.69) and hence  $\text{range } N$  is a subspace of  $V$  that has a smaller dimension than  $V$ . Thus we can apply our induction hypothesis to the restriction operator  $N|_{\text{range } N} \in \mathcal{L}(\text{range } N)$ . [We can ignore the trivial case  $\text{range } N = \{0\}$ , because in that case  $N$  is the 0 operator and we can choose  $v_1, \dots, v_n$  to be any basis of  $V$  and  $m_1 = \dots = m_n = 0$  to get the desired result.]

By our induction hypothesis applied to  $N|_{\text{range } N}$ , there exist vectors  $v_1, \dots, v_n \in \text{range } N$  and nonnegative integers  $m_1, \dots, m_n$  such that

$$\mathbf{8.56} \quad N^{m_1} v_1, \dots, N v_1, v_1, \dots, N^{m_n} v_n, \dots, N v_n, v_n$$

is a basis of  $\text{range } N$  and

$$N^{m_1+1} v_1 = \dots = N^{m_n+1} v_n = 0.$$

Because each  $v_j$  is in  $\text{range } N$ , for each  $j$  there exists  $u_j \in V$  such that  $v_j = Nu_j$ . Thus  $N^{k+1} u_j = N^k v_j$  for each  $j$  and each nonnegative integer  $k$ . We now claim that

$$\mathbf{8.57} \quad N^{m_1+1} u_1, \dots, Nu_1, u_1, \dots, N^{m_n+1} u_n, \dots, Nu_n, u_n$$

is a linearly independent list of vectors in  $V$ . To verify this claim, suppose that some linear combination of 8.57 equals 0. Applying  $N$  to that linear combination, we get a linear combination of 8.56 equal to 0. However, the list 8.56 is linearly independent, and hence all the coefficients in our original linear combination of 8.57 equal 0 except possibly the coefficients of the vectors

$$N^{m_1+1}u_1, \dots, N^{m_n+1}u_n,$$

which equal the vectors

$$N^{m_1}v_1, \dots, N^{m_n}v_n.$$

Again using the linear independence of the list 8.56, we conclude that those coefficients also equal 0, completing our proof that the list 8.57 is linearly independent.

Now extend 8.57 to a basis

$$\mathbf{8.58} \quad N^{m_1+1}u_1, \dots, Nu_1, u_1, \dots, N^{m_n+1}u_n, \dots, Nu_n, u_n, w_1, \dots, w_p$$

of  $V$  (which is possible by 2.33). Each  $Nw_j$  is in range  $N$  and hence is in the span of 8.56. Each vector in the list 8.56 equals  $N$  applied to some vector in the list 8.57. Thus there exists  $x_j$  in the span of 8.57 such that  $Nw_j = Nx_j$ . Now let

$$u_{n+j} = w_j - x_j.$$

Then  $Nu_{n+j} = 0$ . Furthermore,

$$N^{m_1+1}u_1, \dots, Nu_1, u_1, \dots, N^{m_n+1}u_n, \dots, Nu_n, u_n, u_{n+1}, \dots, u_{n+p}$$

spans  $V$  because its span contains each  $x_j$  and each  $u_{n+j}$  and hence each  $w_j$  (and because 8.58 spans  $V$ ).

Thus the spanning list above is a basis of  $V$  because it has the same length as the basis 8.58 (where we have used 2.42). This basis has the required form, completing the proof. ■

French mathematician Camille Jordan (1838–1922) first published a proof of 8.60 in 1870.

In the next definition, the diagonal of each  $A_j$  is filled with some eigenvalue  $\lambda_j$  of  $T$ , the line directly above the diagonal of  $A_j$  is filled with 1's, and all

other entries in  $A_j$  are 0 (to understand why each  $\lambda_j$  is an eigenvalue of  $T$ , see 5.32). The  $\lambda_j$ 's need not be distinct. Also,  $A_j$  may be a 1-by-1 matrix ( $\lambda_j$ ) containing just an eigenvalue of  $T$ .

### 8.59 Definition *Jordan basis*

Suppose  $T \in \mathcal{L}(V)$ . A basis of  $V$  is called a ***Jordan basis*** for  $T$  if with respect to this basis  $T$  has a block diagonal matrix

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_p \end{pmatrix},$$

where each  $A_j$  is an upper-triangular matrix of the form

$$A_j = \begin{pmatrix} \lambda_j & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & \ddots & 1 \\ & & & \lambda_j \end{pmatrix}.$$

### 8.60 Jordan Form

Suppose  $V$  is a complex vector space. If  $T \in \mathcal{L}(V)$ , then there is a basis of  $V$  that is a Jordan basis for  $T$ .

**Proof** First consider a nilpotent operator  $N \in \mathcal{L}(V)$  and the vectors  $v_1, \dots, v_n \in V$  given by 8.55. For each  $j$ , note that  $N$  sends the first vector in the list  $N^{m_j} v_j, \dots, Nv_j, v_j$  to 0 and that  $N$  sends each vector in this list other than the first vector to the previous vector. In other words, 8.55 gives a basis of  $V$  with respect to which  $N$  has a block diagonal matrix, where each matrix on the diagonal has the form

$$\begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & \ddots & 1 \\ & & & 0 \end{pmatrix}.$$

Thus the desired result holds for nilpotent operators.

Now suppose  $T \in \mathcal{L}(V)$ . Let  $\lambda_1, \dots, \lambda_m$  be the distinct eigenvalues of  $T$ . We have the generalized eigenspace decomposition

$$V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T),$$

where each  $(T - \lambda_j I)|_{G(\lambda_j, T)}$  is nilpotent (see 8.21). Thus some basis of each  $G(\lambda_j, T)$  is a Jordan basis for  $(T - \lambda_j I)|_{G(\lambda_j, T)}$  (see previous paragraph). Put these bases together to get a basis of  $V$  that is a Jordan basis for  $T$ . ■

## EXERCISES 8.D

---

- 1 Find the characteristic polynomial and the minimal polynomial of the operator  $N$  in Example 8.53.
- 2 Find the characteristic polynomial and the minimal polynomial of the operator  $N$  in Example 8.54.
- 3 Suppose  $N \in \mathcal{L}(V)$  is nilpotent. Prove that the minimal polynomial of  $N$  is  $z^{m+1}$ , where  $m$  is the length of the longest consecutive string of 1's that appears on the line directly above the diagonal in the matrix of  $N$  with respect to any Jordan basis for  $N$ .
- 4 Suppose  $T \in \mathcal{L}(V)$  and  $v_1, \dots, v_n$  is a basis of  $V$  that is a Jordan basis for  $T$ . Describe the matrix of  $T$  with respect to the basis  $v_n, \dots, v_1$  obtained by reversing the order of the  $v$ 's.
- 5 Suppose  $T \in \mathcal{L}(V)$  and  $v_1, \dots, v_n$  is a basis of  $V$  that is a Jordan basis for  $T$ . Describe the matrix of  $T^2$  with respect to this basis.
- 6 Suppose  $N \in \mathcal{L}(V)$  is nilpotent and  $v_1, \dots, v_n$  and  $m_1, \dots, m_n$  are as in 8.55. Prove that  $N^{m_1}v_1, \dots, N^{m_n}v_n$  is a basis of null  $N$ .  
*[The exercise above implies that  $n$ , which equals  $\dim \text{null } N$ , depends only on  $N$  and not on the specific Jordan basis chosen for  $N$ .]*
- 7 Suppose  $p, q \in \mathcal{P}(\mathbf{C})$  are monic polynomials with the same zeros and  $q$  is a polynomial multiple of  $p$ . Prove that there exists  $T \in \mathcal{L}(\mathbf{C}^{\deg q})$  such that the characteristic polynomial of  $T$  is  $q$  and the minimal polynomial of  $T$  is  $p$ .
- 8 Suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . Prove that there does not exist a direct sum decomposition of  $V$  into two proper subspaces invariant under  $T$  if and only if the minimal polynomial of  $T$  is of the form  $(z - \lambda)^{\dim V}$  for some  $\lambda \in \mathbf{C}$ .



*Euclid explaining geometry (from The School of Athens, painted by Raphael around 1510).*

# Operators on Real Vector Spaces

In the last chapter we learned about the structure of an operator on a finite-dimensional complex vector space. In this chapter, we will use our results about operators on complex vector spaces to learn about operators on real vector spaces.

Our assumptions for this chapter are as follows:

## 9.1 Notation $\mathbf{F}$ , $V$

- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  denotes a finite-dimensional nonzero vector space over  $\mathbf{F}$ .

### LEARNING OBJECTIVES FOR THIS CHAPTER

- complexification of a real vector space
- complexification of an operator on a real vector space
- operators on finite-dimensional real vector spaces have an eigenvalue or a 2-dimensional invariant subspace
- characteristic polynomial and the Cayley–Hamilton Theorem
- description of normal operators on a real inner product space
- description of isometries on a real inner product space

## 9.A Complexification

### Complexification of a Vector Space

As we will soon see, a real vector space  $V$  can be embedded, in a natural way, in a complex vector space called the complexification of  $V$ . Each operator on  $V$  can be extended to an operator on the complexification of  $V$ . Our results about operators on complex vector spaces can then be translated to information about operators on real vector spaces.

We begin by defining the complexification of a real vector space.

#### 9.2 Definition complexification of $V$ , $V_C$

Suppose  $V$  is a real vector space.

- The **complexification** of  $V$ , denoted  $V_C$ , equals  $V \times V$ . An element of  $V_C$  is an ordered pair  $(u, v)$ , where  $u, v \in V$ , but we will write this as  $u + iv$ .
- Addition on  $V_C$  is defined by

$$(u_1 + iv_1) + (u_2 + iv_2) = (u_1 + u_2) + i(v_1 + v_2)$$

for  $u_1, v_1, u_2, v_2 \in V$ .

- Complex scalar multiplication on  $V_C$  is defined by

$$(a + bi)(u + iv) = (au - bv) + i(av + bu)$$

for  $a, b \in \mathbf{R}$  and  $u, v \in V$ .

Motivation for the definition above of complex scalar multiplication comes from usual algebraic properties and the identity  $i^2 = -1$ . If you remember the motivation, then you do not need to memorize the definition above.

We think of  $V$  as a subset of  $V_C$  by identifying  $u \in V$  with  $u + i0$ . The construction of  $V_C$  from  $V$  can then be thought of as generalizing the construction of  $\mathbf{C}^n$  from  $\mathbf{R}^n$ .

#### 9.3 $V_C$ is a complex vector space.

Suppose  $V$  is a real vector space. Then with the definitions of addition and scalar multiplication as above,  $V_C$  is a complex vector space.

The proof of the result above is left as an exercise for the reader. Note that the additive identity of  $V_C$  is  $0 + i0$ , which we write as just 0.

Probably everything that you think should work concerning complexification does work, usually with a straightforward verification, as illustrated by the next result.

#### 9.4 Basis of $V$ is basis of $V_{\mathbb{C}}$

Suppose  $V$  is a real vector space.

- (a) If  $v_1, \dots, v_n$  is a basis of  $V$  (as a real vector space), then  $v_1, \dots, v_n$  is a basis of  $V_{\mathbb{C}}$  (as a complex vector space).
- (b) The dimension of  $V_{\mathbb{C}}$  (as a complex vector space) equals the dimension of  $V$  (as a real vector space).

**Proof** To prove (a), suppose  $v_1, \dots, v_n$  is a basis of the real vector space  $V$ . Then  $\text{span}(v_1, \dots, v_n)$  in the complex vector space  $V_{\mathbb{C}}$  contains all the vectors  $v_1, \dots, v_n, iv_1, \dots, iv_n$ . Thus  $v_1, \dots, v_n$  spans the complex vector space  $V_{\mathbb{C}}$ .

To show that  $v_1, \dots, v_n$  is linearly independent in the complex vector space  $V_{\mathbb{C}}$ , suppose  $\lambda_1, \dots, \lambda_n \in \mathbb{C}$  and

$$\lambda_1 v_1 + \cdots + \lambda_n v_n = 0.$$

Then the equation above and our definitions imply that

$$(\operatorname{Re} \lambda_1)v_1 + \cdots + (\operatorname{Re} \lambda_n)v_n = 0 \quad \text{and} \quad (\operatorname{Im} \lambda_1)v_1 + \cdots + (\operatorname{Im} \lambda_n)v_n = 0.$$

Because  $v_1, \dots, v_n$  is linearly independent in  $V$ , the equations above imply  $\operatorname{Re} \lambda_1 = \cdots = \operatorname{Re} \lambda_n = 0$  and  $\operatorname{Im} \lambda_1 = \cdots = \operatorname{Im} \lambda_n = 0$ . Thus we have  $\lambda_1 = \cdots = \lambda_n = 0$ . Hence  $v_1, \dots, v_n$  is linearly independent in  $V_{\mathbb{C}}$ , completing the proof of (a). ■

Clearly (b) follows immediately from (a). ■

#### Complexification of an Operator

Now we can define the complexification of an operator.

#### 9.5 Definition *complexification of $T$ , $T_{\mathbb{C}}$*

Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . The *complexification* of  $T$ , denoted  $T_{\mathbb{C}}$ , is the operator  $T_{\mathbb{C}} \in \mathcal{L}(V_{\mathbb{C}})$  defined by

$$T_{\mathbb{C}}(u + iv) = Tu + iTv$$

for  $u, v \in V$ .

You should verify that if  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ , then  $T_C$  is indeed in  $\mathcal{L}(V_C)$ . The key point here is that our definition of complex scalar multiplication can be used to show that  $T_C(\lambda(u + iv)) = \lambda T_C(u + iv)$  for all  $u, v \in V$  and all **complex** numbers  $\lambda$ .

The next example gives a good way to think about the complexification of a typical operator.

**9.6 Example** Suppose  $A$  is an  $n$ -by- $n$  matrix of real numbers. Define  $T \in \mathcal{L}(\mathbf{R}^n)$  by  $Tx = Ax$ , where elements of  $\mathbf{R}^n$  are thought of as  $n$ -by-1 column vectors. Identifying the complexification of  $\mathbf{R}^n$  with  $\mathbf{C}^n$ , we then have  $T_C z = Az$  for each  $z \in \mathbf{C}^n$ , where again elements of  $\mathbf{C}^n$  are thought of as  $n$ -by-1 column vectors.

In other words, if  $T$  is the operator of matrix multiplication by  $A$  on  $\mathbf{R}^n$ , then the complexification  $T_C$  is also matrix multiplication by  $A$  but now acting on the larger domain  $\mathbf{C}^n$ .

The next result makes sense because 9.4 tells us that a basis of a real vector space is also a basis of its complexification. The proof of the next result follows immediately from the definitions.

### 9.7 Matrix of $T_C$ equals matrix of $T$

Suppose  $V$  is a real vector space with basis  $v_1, \dots, v_n$  and  $T \in \mathcal{L}(V)$ . Then  $\mathcal{M}(T) = \mathcal{M}(T_C)$ , where both matrices are with respect to the basis  $v_1, \dots, v_n$ .

The result above and Example 9.6 provide complete insight into complexification, because once a basis is chosen, every operator essentially looks like Example 9.6. Complexification of an operator could have been defined using matrices, but the approach taken here is more natural because it does not depend on the choice of a basis.

We know that every operator on a nonzero finite-dimensional complex vector space has an eigenvalue (see 5.21) and thus has a 1-dimensional invariant subspace. We have seen an example [5.8(a)] of an operator on a nonzero finite-dimensional real vector space with no eigenvalues and thus no 1-dimensional invariant subspaces. However, we now show that an invariant subspace of dimension 1 or 2 always exists. Notice how complexification leads to a simple proof of this result.

## 9.8 Every operator has an invariant subspace of dimension 1 or 2

Every operator on a nonzero finite-dimensional vector space has an invariant subspace of dimension 1 or 2.

**Proof** Every operator on a nonzero finite-dimensional complex vector space has an eigenvalue (5.21) and thus has a 1-dimensional invariant subspace.

Hence assume  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . The complexification  $T_C$  has an eigenvalue  $a + bi$  (by 5.21), where  $a, b \in \mathbf{R}$ . Thus there exist  $u, v \in V$ , not both 0, such that  $T_C(u + iv) = (a + bi)(u + iv)$ . Using the definition of  $T_C$ , the last equation can be rewritten as

$$Tu + iTv = (au - bv) + (av + bu)i.$$

Thus

$$Tu = au - bv \quad \text{and} \quad Tv = av + bu.$$

Let  $U$  equal the span in  $V$  of the list  $u, v$ . Then  $U$  is a subspace of  $V$  with dimension 1 or 2. The equations above show that  $U$  is invariant under  $T$ , completing the proof. ■

## The Minimal Polynomial of the Complexification

Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Repeated application of the definition of  $T_C$  shows that

$$9.9 \quad (T_C)^n(u + iv) = T^n u + iT^n v$$

for every positive integer  $n$  and all  $u, v \in V$ .

Notice that the next result implies that the minimal polynomial of  $T_C$  has real coefficients.

## 9.10 Minimal polynomial of $T_C$ equals minimal polynomial of $T$

Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Then the minimal polynomial of  $T_C$  equals the minimal polynomial of  $T$ .

**Proof** Let  $p \in \mathcal{P}(\mathbf{R})$  denote the minimal polynomial of  $T$ . From 9.9 it is easy to see that  $p(T_C) = (p(T))_C$ , and thus  $p(T_C) = 0$ .

Suppose  $q \in \mathcal{P}(\mathbf{C})$  is a monic polynomial such that  $q(T_C) = 0$ . Then  $(q(T_C))(u) = 0$  for every  $u \in V$ . Letting  $r$  denote the polynomial whose  $j^{\text{th}}$  coefficient is the real part of the  $j^{\text{th}}$  coefficient of  $q$ , we see that  $r$  is a monic polynomial and  $r(T) = 0$ . Thus  $\deg q = \deg r \geq \deg p$ .

The conclusions of the two previous paragraphs imply that  $p$  is the minimal polynomial of  $T_C$ , as desired. ■

## Eigenvalues of the Complexification

Now we turn to questions about the eigenvalues of the complexification of an operator. Again, everything that we expect to work indeed works easily.

We begin with a result showing that the real eigenvalues of  $T_C$  are precisely the eigenvalues of  $T$ . We give two different proofs of this result. The first proof is more elementary, but the second proof is shorter and gives some useful insight.

### 9.11 Real eigenvalues of $T_C$

Suppose  $V$  is a real vector space,  $T \in \mathcal{L}(V)$ , and  $\lambda \in \mathbf{R}$ . Then  $\lambda$  is an eigenvalue of  $T_C$  if and only if  $\lambda$  is an eigenvalue of  $T$ .

**Proof 1** First suppose  $\lambda$  is an eigenvalue of  $T$ . Then there exists  $v \in V$  with  $v \neq 0$  such that  $Tv = \lambda v$ . Thus  $T_C v = \lambda v$ , which shows that  $\lambda$  is an eigenvalue of  $T_C$ , completing one direction of the proof.

To prove the other direction, suppose now that  $\lambda$  is an eigenvalue of  $T_C$ . Then there exist  $u, v \in V$  with  $u + iv \neq 0$  such that

$$T_C(u + iv) = \lambda(u + iv).$$

The equation above implies that  $Tu = \lambda u$  and  $Tv = \lambda v$ . Because  $u \neq 0$  or  $v \neq 0$ , this implies that  $\lambda$  is an eigenvalue of  $T$ , completing the proof. ■

**Proof 2** The (real) eigenvalues of  $T$  are the (real) zeros of the minimal polynomial of  $T$  (by 8.49). The real eigenvalues of  $T_C$  are the real zeros of the minimal polynomial of  $T_C$  (again by 8.49). These two minimal polynomials are the same (by 9.10). Thus the eigenvalues of  $T$  are precisely the real eigenvalues of  $T_C$ , as desired. ■

Our next result shows that  $T_C$  behaves symmetrically with respect to an eigenvalue  $\lambda$  and its complex conjugate  $\bar{\lambda}$ .

### 9.12 $T_C - \lambda I$ and $T_C - \bar{\lambda} I$

Suppose  $V$  is a real vector space,  $T \in \mathcal{L}(V)$ ,  $\lambda \in \mathbf{C}$ ,  $j$  is a nonnegative integer, and  $u, v \in V$ . Then

$$(T_C - \lambda I)^j(u + iv) = 0 \quad \text{if and only if} \quad (T_C - \bar{\lambda} I)^j(u - iv) = 0.$$

**Proof** We will prove this result by induction on  $j$ . To get started, note that if  $j = 0$  then (because an operator raised to the power 0 equals the identity operator) the result claims that  $u + iv = 0$  if and only if  $u - iv = 0$ , which is clearly true.

Thus assume by induction that  $j \geq 1$  and the desired result holds for  $j - 1$ . Suppose  $(T_C - \lambda I)^j(u + iv) = 0$ . Then

$$9.13 \quad (T_C - \lambda I)^{j-1}((T_C - \lambda I)(u + iv)) = 0.$$

Writing  $\lambda = a + bi$ , where  $a, b \in \mathbf{R}$ , we have

$$9.14 \quad (T_C - \lambda I)(u + iv) = (Tu - au + bv) + i(Tv - av - bu)$$

and

$$9.15 \quad (T_C - \bar{\lambda} I)(u - iv) = (Tu - au + bv) - i(Tv - av - bu).$$

Our induction hypothesis, 9.13, and 9.14 imply that

$$(T_C - \bar{\lambda} I)^{j-1}((Tu - au + bv) - i(Tv - av - bu)) = 0.$$

Now the equation above and 9.15 imply that  $(T_C - \bar{\lambda} I)^j(u - iv) = 0$ , completing the proof in one direction.

The other direction is proved by replacing  $\lambda$  with  $\bar{\lambda}$ , replacing  $v$  with  $-v$ , and then using the first direction. ■

An important consequence of the result above is the next result, which states that if a number is an eigenvalue of  $T_C$ , then its complex conjugate is also an eigenvalue of  $T_C$ .

### 9.16 Nonreal eigenvalues of $T_C$ come in pairs

Suppose  $V$  is a real vector space,  $T \in \mathcal{L}(V)$ , and  $\lambda \in \mathbf{C}$ . Then  $\lambda$  is an eigenvalue of  $T_C$  if and only if  $\bar{\lambda}$  is an eigenvalue of  $T_C$ .

**Proof** Take  $j = 1$  in 9.12. ■

By definition, the eigenvalues of an operator on a real vector space are real numbers. Thus when mathematicians sometimes informally mention the complex eigenvalues of an operator on a real vector space, what they have in mind is the eigenvalues of the complexification of the operator.

Recall that the multiplicity of an eigenvalue is defined to be the dimension of the generalized eigenspace corresponding to that eigenvalue (see 8.24). The next result states that the multiplicity of an eigenvalue of a complexification equals the multiplicity of its complex conjugate.

### 9.17 Multiplicity of $\lambda$ equals multiplicity of $\bar{\lambda}$

Suppose  $V$  is a real vector space,  $T \in \mathcal{L}(V)$ , and  $\lambda \in \mathbf{C}$  is an eigenvalue of  $T_C$ . Then the multiplicity of  $\lambda$  as an eigenvalue of  $T_C$  equals the multiplicity of  $\bar{\lambda}$  as an eigenvalue of  $T_C$ .

**Proof** Suppose  $u_1 + iv_1, \dots, u_m + iv_m$  is a basis of the generalized eigenspace  $G(\lambda, T_C)$ , where  $u_1, \dots, u_m, v_1, \dots, v_m \in V$ . Then using 9.12, routine arguments show that  $u_1 - iv_1, \dots, u_m - iv_m$  is a basis of the generalized eigenspace  $G(\bar{\lambda}, T_C)$ . Thus both  $\lambda$  and  $\bar{\lambda}$  have multiplicity  $m$  as eigenvalues of  $T_C$ . ■

### 9.18 Example

Suppose  $T \in \mathcal{L}(\mathbf{R}^3)$  is defined by

$$T(x_1, x_2, x_3) = (2x_1, x_2 - x_3, x_2 + x_3).$$

The matrix of  $T$  with respect to the standard basis of  $\mathbf{R}^3$  is  $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix}$ .

As you can verify, 2 is an eigenvalue of  $T$  with multiplicity 1 and  $T$  has no other eigenvalues.

If we identify the complexification of  $\mathbf{R}^3$  with  $\mathbf{C}^3$ , then the matrix of  $T_C$  with respect to the standard basis of  $\mathbf{C}^3$  is the matrix above. As you can verify, the eigenvalues of  $T_C$  are 2,  $1+i$ , and  $1-i$ , each with multiplicity 1. Thus the nonreal eigenvalues of  $T_C$  come as a pair, with each the complex conjugate of the other and with the same multiplicity, as expected by 9.17.

We have seen an example [5.8(a)] of an operator on  $\mathbf{R}^2$  with no eigenvalues. The next result shows that no such example exists on  $\mathbf{R}^3$ .

### 9.19 Operator on odd-dimensional vector space has eigenvalue

Every operator on an odd-dimensional real vector space has an eigenvalue.

**Proof** Suppose  $V$  is a real vector space with odd dimension and  $T \in \mathcal{L}(V)$ . Because the nonreal eigenvalues of  $T_C$  come in pairs with equal multiplicity (by 9.17), the sum of the multiplicities of all the nonreal eigenvalues of  $T_C$  is an even number.

Because the sum of the multiplicities of all the eigenvalues of  $T_C$  equals the (complex) dimension of  $V_C$  (by Theorem 8.26), the conclusion of the paragraph above implies that  $T_C$  has a real eigenvalue. Every real eigenvalue of  $T_C$  is also an eigenvalue of  $T$  (by 9.11), giving the desired result. ■

## Characteristic Polynomial of the Complexification

In the previous chapter we defined the characteristic polynomial of an operator on a finite-dimensional complex vector space (see 8.34). The next result is a key step toward defining the characteristic polynomial for operators on finite-dimensional real vector spaces.

### 9.20 Characteristic polynomial of $T_C$

Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Then the coefficients of the characteristic polynomial of  $T_C$  are all real.

**Proof** Suppose  $\lambda$  is a nonreal eigenvalue of  $T_C$  with multiplicity  $m$ . Then  $\bar{\lambda}$  is also an eigenvalue of  $T_C$  with multiplicity  $m$  (by 9.17). Thus the characteristic polynomial of  $T_C$  includes factors of  $(z - \lambda)^m$  and  $(z - \bar{\lambda})^m$ . Multiplying together these two factors, we have

$$(z - \lambda)^m(z - \bar{\lambda})^m = (z^2 - 2(\operatorname{Re} \lambda)z + |\lambda|^2)^m.$$

The polynomial above on the right has real coefficients.

The characteristic polynomial of  $T_C$  is the product of terms of the form above and terms of the form  $(z - t)^d$ , where  $t$  is a real eigenvalue of  $T_C$  with multiplicity  $d$ . Thus the coefficients of the characteristic polynomial of  $T_C$  are all real. ■

Now we can define the characteristic polynomial of an operator on a finite-dimensional real vector space to be the characteristic polynomial of its complexification.

### 9.21 Definition *Characteristic polynomial*

Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Then the **characteristic polynomial** of  $T$  is defined to be the characteristic polynomial of  $T_C$ .

---

**9.22 Example** Suppose  $T \in \mathcal{L}(\mathbf{R}^3)$  is defined by

$$T(x_1, x_2, x_3) = (2x_1, x_2 - x_3, x_2 + x_3).$$

As we noted in 9.18, the eigenvalues of  $T_C$  are 2,  $1 + i$ , and  $1 - i$ , each with multiplicity 1. Thus the characteristic polynomial of the complexification  $T_C$  is  $(z - 2)(z - (1 + i))(z - (1 - i))$ , which equals  $z^3 - 4z^2 + 6z - 4$ . Hence the characteristic polynomial of  $T$  is also  $z^3 - 4z^2 + 6z - 4$ .

---

In the next result, the eigenvalues of  $T$  are all real (because  $T$  is an operator on a real vector space).

### 9.23 Degree and zeros of characteristic polynomial

Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Then

- (a) the coefficients of the characteristic polynomial of  $T$  are all real;
- (b) the characteristic polynomial of  $T$  has degree  $\dim V$ ;
- (c) the eigenvalues of  $T$  are precisely the real zeros of the characteristic polynomial of  $T$ .

**Proof** Part (a) holds because of 9.20.

Part (b) follows from 8.36(a).

Part (c) holds because the real zeros of the characteristic polynomial of  $T$  are the real eigenvalues of  $T_C$  [by 8.36(a)], which are the eigenvalues of  $T$  (by 9.11). ■

In the previous chapter, we proved the Cayley–Hamilton Theorem (8.37) for complex vector spaces. Now we can also prove it for real vector spaces.

### 9.24 Cayley–Hamilton Theorem

Suppose  $T \in \mathcal{L}(V)$ . Let  $q$  denote the characteristic polynomial of  $T$ . Then  $q(T) = 0$ .

**Proof** We have already proved this result when  $V$  is a complex vector space. Thus assume that  $V$  is a real vector space.

The complex case of the Cayley–Hamilton Theorem (8.37) implies that  $q(T_C) = 0$ . Thus we also have  $q(T) = 0$ , as desired. ■

---

**9.25 Example** Suppose  $T \in \mathcal{L}(\mathbf{R}^3)$  is defined by

$$T(x_1, x_2, x_3) = (2x_1, x_2 - x_3, x_2 + x_3).$$

As we saw in 9.22, the characteristic polynomial of  $T$  is  $z^3 - 4z^2 + 6z - 4$ . Thus the Cayley–Hamilton Theorem implies that  $T^3 - 4T^2 + 6T - 4I = 0$ , which can also be verified by direct calculation.

---

We can now prove another result that we previously knew only in the complex case.

## 9.26 Characteristic polynomial is a multiple of minimal polynomial

Suppose  $T \in \mathcal{L}(V)$ . Then

- (a) the degree of the minimal polynomial of  $T$  is at most  $\dim V$ ;
- (b) the characteristic polynomial of  $T$  is a polynomial multiple of the minimal polynomial of  $T$ .

**Proof** Part (a) follows immediately from the Cayley–Hamilton Theorem.

Part (b) follows from the Cayley–Hamilton Theorem and 8.46. ■

## EXERCISES 9.A

---

- 1 Prove 9.3.
- 2 Verify that if  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ , then  $T_C \in \mathcal{L}(V_C)$ .
- 3 Suppose  $V$  is a real vector space and  $v_1, \dots, v_m \in V$ . Prove that  $v_1, \dots, v_m$  is linearly independent in  $V_C$  if and only if  $v_1, \dots, v_m$  is linearly independent in  $V$ .
- 4 Suppose  $V$  is a real vector space and  $v_1, \dots, v_m \in V$ . Prove that  $v_1, \dots, v_m$  spans  $V_C$  if and only if  $v_1, \dots, v_m$  spans  $V$ .
- 5 Suppose that  $V$  is a real vector space and  $S, T \in \mathcal{L}(V)$ . Show that  $(S + T)_C = S_C + T_C$  and that  $(\lambda T)_C = \lambda T_C$  for every  $\lambda \in \mathbf{R}$ .
- 6 Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Prove that  $T_C$  is invertible if and only if  $T$  is invertible.
- 7 Suppose  $V$  is a real vector space and  $N \in \mathcal{L}(V)$ . Prove that  $N_C$  is nilpotent if and only if  $N$  is nilpotent.
- 8 Suppose  $T \in \mathcal{L}(\mathbf{R}^3)$  and 5, 7 are eigenvalues of  $T$ . Prove that  $T_C$  has no nonreal eigenvalues.
- 9 Prove there does not exist an operator  $T \in \mathcal{L}(\mathbf{R}^7)$  such that  $T^2 + T + I$  is nilpotent.
- 10 Give an example of an operator  $T \in \mathcal{L}(\mathbf{C}^7)$  such that  $T^2 + T + I$  is nilpotent.

- 11** Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Suppose there exist  $b, c \in \mathbf{R}$  such that  $T^2 + bT + cI = 0$ . Prove that  $T$  has an eigenvalue if and only if  $b^2 \geq 4c$ .
- 12** Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Suppose there exist  $b, c \in \mathbf{R}$  such that  $b^2 < 4c$  and  $T^2 + bT + cI$  is nilpotent. Prove that  $T$  has no eigenvalues.
- 13** Suppose  $V$  is a real vector space,  $T \in \mathcal{L}(V)$ , and  $b, c \in \mathbf{R}$  are such that  $b^2 < 4c$ . Prove that  $\text{null}(T^2 + bT + cI)^j$  has even dimension for every positive integer  $j$ .
- 14** Suppose  $V$  is a real vector space with  $\dim V = 8$ . Suppose  $T \in \mathcal{L}(V)$  is such that  $T^2 + T + I$  is nilpotent. Prove that  $(T^2 + T + I)^4 = 0$ .
- 15** Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$  has no eigenvalues. Prove that every subspace of  $V$  invariant under  $T$  has even dimension.
- 16** Suppose  $V$  is a real vector space. Prove that there exists  $T \in \mathcal{L}(V)$  such that  $T^2 = -I$  if and only if  $V$  has even dimension.
- 17** Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$  satisfies  $T^2 = -I$ . Define complex scalar multiplication on  $V$  as follows: if  $a, b \in \mathbf{R}$ , then
- $$(a + bi)v = av + bTv.$$
- (a) Show that the complex scalar multiplication on  $V$  defined above and the addition on  $V$  makes  $V$  into a complex vector space.
- (b) Show that the dimension of  $V$  as a complex vector space is half the dimension of  $V$  as a real vector space.
- 18** Suppose  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Prove that the following are equivalent:
- (a) All the eigenvalues of  $T_C$  are real.
- (b) There exists a basis of  $V$  with respect to which  $T$  has an upper-triangular matrix.
- (c) There exists a basis of  $V$  consisting of generalized eigenvectors of  $T$ .
- 19** Suppose  $V$  is a real vector space with  $\dim V = n$  and  $T \in \mathcal{L}(V)$  is such that  $\text{null } T^{n-2} \neq \text{null } T^{n-1}$ . Prove that  $T$  has at most two distinct eigenvalues and that  $T_C$  has no nonreal eigenvalues.

## 9.B Operators on Real Inner Product Spaces

We now switch our focus to the context of inner product spaces. We will give a complete description of normal operators on real inner product spaces; a key step in the proof of this result (9.34) requires the result from the previous section that an operator on a finite-dimensional real vector space has an invariant subspace of dimension 1 or 2 (9.8).

After describing the normal operators on real inner product spaces, we will use that result to give a complete description of isometries on such spaces.

### Normal Operators on Real Inner Product Spaces

The Complex Spectral Theorem (7.24) gives a complete description of normal operators on complex inner product spaces. In this subsection we will give a complete description of normal operators on real inner product spaces.

We begin with a description of the operators on 2-dimensional real inner product spaces that are normal but not self-adjoint.

#### 9.27 Normal but not self-adjoint operators

Suppose  $V$  is a 2-dimensional real inner product space and  $T \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $T$  is normal but not self-adjoint.
- (b) The matrix of  $T$  with respect to every orthonormal basis of  $V$  has the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix},$$

with  $b \neq 0$ .

- (c) The matrix of  $T$  with respect to some orthonormal basis of  $V$  has the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix},$$

with  $b > 0$ .

**Proof** First suppose (a) holds, so that  $T$  is normal but not self-adjoint. Let  $e_1, e_2$  be an orthonormal basis of  $V$ . Suppose

$$9.28 \quad \mathcal{M}(T, (e_1, e_2)) = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

Then  $\|Te_1\|^2 = a^2 + b^2$  and  $\|T^*e_1\|^2 = a^2 + c^2$ . Because  $T$  is normal,  $\|Te_1\| = \|T^*e_1\|$  (see 7.20); thus these equations imply that  $b^2 = c^2$ . Thus  $c = b$  or  $c = -b$ . But  $c \neq b$ , because otherwise  $T$  would be self-adjoint, as can be seen from the matrix in 9.28. Hence  $c = -b$ , so

$$9.29 \quad \mathcal{M}(T, (e_1, e_2)) = \begin{pmatrix} a & -b \\ b & d \end{pmatrix}.$$

The matrix of  $T^*$  is the transpose of the matrix above. Use matrix multiplication to compute the matrices of  $TT^*$  and  $T^*T$  (do it now). Because  $T$  is normal, these two matrices are equal. Equating the entries in the upper-right corner of the two matrices you computed, you will discover that  $bd = ab$ . Now  $b \neq 0$ , because otherwise  $T$  would be self-adjoint, as can be seen from the matrix in 9.29. Thus  $d = a$ , completing the proof that (a) implies (b).

Now suppose (b) holds. We want to prove that (c) holds. Choose an orthonormal basis  $e_1, e_2$  of  $V$ . We know that the matrix of  $T$  with respect to this basis has the form given by (b), with  $b \neq 0$ . If  $b > 0$ , then (c) holds and we have proved that (b) implies (c). If  $b < 0$ , then, as you should verify, the matrix of  $T$  with respect to the orthonormal basis  $e_1, -e_2$  equals  $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ , where  $-b > 0$ ; thus in this case we also see that (b) implies (c).

Now suppose (c) holds, so that the matrix of  $T$  with respect to some orthonormal basis has the form given in (c) with  $b > 0$ . Clearly the matrix of  $T$  is not equal to its transpose (because  $b \neq 0$ ). Hence  $T$  is not self-adjoint. Now use matrix multiplication to verify that the matrices of  $TT^*$  and  $T^*T$  are equal. We conclude that  $TT^* = T^*T$ . Hence  $T$  is normal. Thus (c) implies (a), completing the proof. ■

The next result tells us that a normal operator restricted to an invariant subspace is normal. This will allow us to use induction on  $\dim V$  when we prove our description of normal operators (9.34).

### 9.30 Normal operators and invariant subspaces

Suppose  $V$  is an inner product space,  $T \in \mathcal{L}(V)$  is normal, and  $U$  is a subspace of  $V$  that is invariant under  $T$ . Then

- (a)  $U^\perp$  is invariant under  $T$ ;
- (b)  $U$  is invariant under  $T^*$ ;
- (c)  $(T|_U)^* = (T^*)|_U$ ;
- (d)  $T|_U \in \mathcal{L}(U)$  and  $T|_{U^\perp} \in \mathcal{L}(U^\perp)$  are normal operators.

**Proof** First we will prove (a). Let  $e_1, \dots, e_m$  be an orthonormal basis of  $U$ . Extend to an orthonormal basis  $e_1, \dots, e_m, f_1, \dots, f_n$  of  $V$  (this is possible by 6.35). Because  $U$  is invariant under  $T$ , each  $Te_j$  is a linear combination of  $e_1, \dots, e_m$ . Thus the matrix of  $T$  with respect to the basis  $e_1, \dots, e_m, f_1, \dots, f_n$  is of the form

$$\mathcal{M}(T) = \begin{pmatrix} e_1 & \dots & e_m & f_1 & \dots & f_n \\ \vdots & & & & & \\ e_1 & & & A & & B \\ \vdots & & & & & \\ f_1 & & & & & \\ \vdots & & & 0 & & C \\ f_n & & & & & \end{pmatrix};$$

here  $A$  denotes an  $m$ -by- $m$  matrix,  $0$  denotes the  $n$ -by- $m$  matrix of all 0's,  $B$  denotes an  $m$ -by- $n$  matrix,  $C$  denotes an  $n$ -by- $n$  matrix, and for convenience the basis has been listed along the top and left sides of the matrix.

For each  $j \in \{1, \dots, m\}$ ,  $\|Te_j\|^2$  equals the sum of the squares of the absolute values of the entries in the  $j^{\text{th}}$  column of  $A$  (see 6.25). Hence

$$9.31 \quad \sum_{j=1}^m \|Te_j\|^2 = \begin{array}{l} \text{the sum of the squares of the absolute} \\ \text{values of the entries of } A. \end{array}$$

For each  $j \in \{1, \dots, m\}$ ,  $\|T^*e_j\|^2$  equals the sum of the squares of the absolute values of the entries in the  $j^{\text{th}}$  rows of  $A$  and  $B$ . Hence

$$9.32 \quad \sum_{j=1}^m \|T^*e_j\|^2 = \begin{array}{l} \text{the sum of the squares of the absolute} \\ \text{values of the entries of } A \text{ and } B. \end{array}$$

Because  $T$  is normal,  $\|Te_j\| = \|T^*e_j\|$  for each  $j$  (see 7.20); thus

$$\sum_{j=1}^m \|Te_j\|^2 = \sum_{j=1}^m \|T^*e_j\|^2.$$

This equation, along with 9.31 and 9.32, implies that the sum of the squares of the absolute values of the entries of  $B$  equals 0. In other words,  $B$  is the matrix of all 0's. Thus

$$9.33 \quad \mathcal{M}(T) = \begin{pmatrix} e_1 & \dots & e_m & f_1 & \dots & f_n \\ \vdots & & & A & & 0 \\ e_m & & & & & \\ f_1 & & & & & \\ \vdots & & & 0 & & C \\ f_n & & & & & \end{pmatrix}.$$

This representation shows that  $Tf_k$  is in the span of  $f_1, \dots, f_n$  for each  $k$ . Because  $f_1, \dots, f_n$  is a basis of  $U^\perp$ , this implies that  $Tv \in U^\perp$  whenever  $v \in U^\perp$ . In other words,  $U^\perp$  is invariant under  $T$ , completing the proof of (a).

To prove (b), note that  $\mathcal{M}(T^*)$ , which is the conjugate transpose of  $\mathcal{M}(T)$ , has a block of 0's in the lower left corner (because  $\mathcal{M}(T)$ , as given above, has a block of 0's in the upper right corner). In other words, each  $T^*e_j$  can be written as a linear combination of  $e_1, \dots, e_m$ . Thus  $U$  is invariant under  $T^*$ , completing the proof of (b).

To prove (c), let  $S = T|_U \in \mathcal{L}(U)$ . Fix  $v \in U$ . Then

$$\begin{aligned} \langle Su, v \rangle &= \langle Tu, v \rangle \\ &= \langle u, T^*v \rangle \end{aligned}$$

for all  $u \in U$ . Because  $T^*v \in U$  [by (b)], the equation above shows that  $S^*v = T^*v$ . In other words,  $(T|_U)^* = (T^*)|_U$ , completing the proof of (c).

To prove (d), note that  $T$  commutes with  $T^*$  (because  $T$  is normal) and that  $(T|_U)^* = (T^*)|_U$  [by (c)]. Thus  $T|_U$  commutes with its adjoint and hence is normal. Interchanging the roles of  $U$  and  $U^\perp$ , which is justified by (a), shows that  $T|_{U^\perp}$  is also normal, completing the proof of (d). ■

*Note that if an operator  $T$  has a block diagonal matrix with respect to some basis, then the entry in each 1-by-1 block on the diagonal of this matrix is an eigenvalue of  $T$ .*

Our next result shows that normal operators on real inner product spaces come close to having diagonal matrices. Specifically, we get block diagonal matrices, with each block having size at most 2-by-2.

We cannot expect to do better than the next result, because on a real inner product space there exist normal operators that do not have a diagonal matrix with respect to any basis. For example, the operator  $T \in \mathcal{L}(\mathbf{R}^2)$  defined by  $T(x, y) = (-y, x)$  is normal (as you should verify) but has no eigenvalues; thus this particular  $T$  does not have even an upper-triangular matrix with respect to any basis of  $\mathbf{R}^2$ .

### 9.34 Characterization of normal operators when $\mathbf{F} = \mathbf{R}$

Suppose  $V$  is a real inner product space and  $T \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $T$  is normal.
- (b) There is an orthonormal basis of  $V$  with respect to which  $T$  has a block diagonal matrix such that each block is a 1-by-1 matrix or a 2-by-2 matrix of the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix},$$

with  $b > 0$ .

**Proof** First suppose (b) holds. With respect to the basis given by (b), the matrix of  $T$  commutes with the matrix of  $T^*$  (which is the transpose of the matrix of  $T$ ), as you should verify (use Exercise 9 in Section 8.B for the product of two block diagonal matrices). Thus  $T$  commutes with  $T^*$ , which means that  $T$  is normal, completing the proof that (b) implies (a).

Now suppose (a) holds, so  $T$  is normal. We will prove that (b) holds by induction on  $\dim V$ . To get started, note that our desired result holds if  $\dim V = 1$  (trivially) or if  $\dim V = 2$  [if  $T$  is self-adjoint, use the Real Spectral Theorem (7.29); if  $T$  is not self-adjoint, use 9.27].

Now assume that  $\dim V > 2$  and that the desired result holds on vector spaces of smaller dimension. Let  $U$  be a subspace of  $V$  of dimension 1 that is invariant under  $T$  if such a subspace exists (in other words, if  $T$  has an eigenvector, let  $U$  be the span of this eigenvector). If no such subspace exists, let  $U$  be a subspace of  $V$  of dimension 2 that is invariant under  $T$  (an invariant subspace of dimension 1 or 2 always exists by 9.8).

If  $\dim U = 1$ , choose a vector in  $U$  with norm 1; this vector will be an orthonormal basis of  $U$ , and of course the matrix of  $T|_U \in \mathcal{L}(U)$  is a 1-by-1 matrix. If  $\dim U = 2$ , then  $T|_U \in \mathcal{L}(U)$  is normal (by 9.30) but not self-adjoint (otherwise  $T|_U$ , and hence  $T$ , would have an eigenvector by 7.27). Thus we can choose an orthonormal basis of  $U$  with respect to which the matrix of  $T|_U \in \mathcal{L}(U)$  has the required form (see 9.27).

Now  $U^\perp$  is invariant under  $T$  and  $T|_{U^\perp}$  is a normal operator on  $U^\perp$  (by 9.30). Thus by our induction hypothesis, there is an orthonormal basis of  $U^\perp$  with respect to which the matrix of  $T|_{U^\perp}$  has the desired form. Adjoining this basis to the basis of  $U$  gives an orthonormal basis of  $V$  with respect to which the matrix of  $T$  has the desired form. Thus (b) holds. ■

## Isometries on Real Inner Product Spaces

As we will see, the next example is a key building block for isometries on real inner product spaces. Also, note that the next example shows that an isometry on  $\mathbf{R}^2$  may have no eigenvalues.

---

**9.35 Example** Let  $\theta \in \mathbf{R}$ . Then the operator on  $\mathbf{R}^2$  of counterclockwise rotation (centered at the origin) by an angle of  $\theta$  is an isometry, as is geometrically obvious. The matrix of this operator with respect to the standard basis is

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

If  $\theta$  is not an integer multiple of  $\pi$ , then no nonzero vector of  $\mathbf{R}^2$  gets mapped to a scalar multiple of itself, and hence the operator has no eigenvalues.

---

The next result shows that every isometry on a real inner product space is composed of pieces that are rotations on 2-dimensional subspaces, pieces that equal the identity operator, and pieces that equal multiplication by  $-1$ .

### 9.36 Description of isometries when $\mathbf{F} = \mathbf{R}$

Suppose  $V$  is a real inner product space and  $S \in \mathcal{L}(V)$ . Then the following are equivalent:

- (a)  $S$  is an isometry.
- (b) There is an orthonormal basis of  $V$  with respect to which  $S$  has a block diagonal matrix such that each block on the diagonal is a 1-by-1 matrix containing 1 or  $-1$  or is a 2-by-2 matrix of the form

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix},$$

with  $\theta \in (0, \pi)$ .

**Proof** First suppose (a) holds, so  $S$  is an isometry. Because  $S$  is normal, there is an orthonormal basis of  $V$  with respect to which  $S$  has a block diagonal matrix such that each block is a 1-by-1 matrix or a 2-by-2 matrix of the form

**9.37** 
$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix},$$

with  $b > 0$  (by 9.34).

If  $\lambda$  is an entry in a 1-by-1 matrix along the diagonal of the matrix of  $S$  (with respect to the basis mentioned above), then there is a basis vector  $e_j$  such that  $Se_j = \lambda e_j$ . Because  $S$  is an isometry, this implies that  $|\lambda| = 1$ . Thus  $\lambda = 1$  or  $\lambda = -1$ , because these are the only real numbers with absolute value 1.

Now consider a 2-by-2 matrix of the form 9.37 along the diagonal of the matrix of  $S$ . There are basis vectors  $e_j, e_{j+1}$  such that

$$Se_j = ae_j + be_{j+1}.$$

Thus

$$1 = \|e_j\|^2 = \|Se_j\|^2 = a^2 + b^2.$$

The equation above, along with the condition  $b > 0$ , implies that there exists a number  $\theta \in (0, \pi)$  such that  $a = \cos \theta$  and  $b = \sin \theta$ . Thus the matrix 9.37 has the required form, completing the proof in this direction.

Conversely, now suppose (b) holds, so there is an orthonormal basis of  $V$  with respect to which the matrix of  $S$  has the form required by the theorem. Thus there is a direct sum decomposition

$$V = U_1 \oplus \cdots \oplus U_m,$$

where each  $U_j$  is a subspace of  $V$  of dimension 1 or 2. Furthermore, any two vectors belonging to distinct  $U$ 's are orthogonal, and each  $S|_{U_j}$  is an isometry mapping  $U_j$  into  $U_j$ . If  $v \in V$ , we can write

$$v = u_1 + \cdots + u_m,$$

where each  $u_j$  is in  $U_j$ . Applying  $S$  to the equation above and then taking norms gives

$$\begin{aligned}\|Sv\|^2 &= \|Su_1 + \cdots + Su_m\|^2 \\ &= \|Su_1\|^2 + \cdots + \|Su_m\|^2 \\ &= \|u_1\|^2 + \cdots + \|u_m\|^2 \\ &= \|v\|^2.\end{aligned}$$

Thus  $S$  is an isometry, and hence (a) holds. ■

## EXERCISES 9.B

---

- 1 Suppose  $S \in \mathcal{L}(\mathbf{R}^3)$  is an isometry. Prove that there exists a nonzero vector  $x \in \mathbf{R}^3$  such that  $S^2x = x$ .
- 2 Prove that every isometry on an odd-dimensional real inner product space has 1 or  $-1$  as an eigenvalue.
- 3 Suppose  $V$  is a real inner product space. Show that

$$\langle u + iv, x + iy \rangle = \langle u, x \rangle + \langle v, y \rangle + (\langle v, x \rangle - \langle u, y \rangle)i$$

for  $u, v, x, y \in V$  defines a complex inner product on  $V_C$ .

- 4 Suppose  $V$  is a real inner product space and  $T \in \mathcal{L}(V)$  is self-adjoint. Show that  $T_C$  is a self-adjoint operator on the inner product space  $V_C$  defined by the previous exercise.
- 5 Use the previous exercise to give a proof of the Real Spectral Theorem (7.29) via complexification and the Complex Spectral Theorem (7.24).
- 6 Give an example of an operator  $T$  on an inner product space such that  $T$  has an invariant subspace whose orthogonal complement is not invariant under  $T$ .  
*[The exercise above shows that 9.30 can fail without the hypothesis that  $T$  is normal.]*
- 7 Suppose  $T \in \mathcal{L}(V)$  and  $T$  has a block diagonal matrix

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix}$$

with respect to some basis of  $V$ . For  $j = 1, \dots, m$ , let  $T_j$  be the operator on  $V$  whose matrix with respect to the same basis is a block diagonal matrix with blocks the same size as in the matrix above, with  $A_j$  in the  $j^{\text{th}}$  block, and with all the other blocks on the diagonal equal to identity matrices (of the appropriate size). Prove that  $T = T_1 \cdots T_m$ .

- 8 Suppose  $D$  is the differentiation operator on the vector space  $V$  in Exercise 21 in Section 7.A. Find an orthonormal basis of  $V$  such that the matrix of the normal operator  $D$  has the form promised by 9.34.

# CHAPTER 10



*British mathematician and pioneer computer scientist Ada Lovelace (1815–1852), as painted by Alfred Chalon in this 1840 portrait.*

## Trace and Determinant

Throughout this book our emphasis has been on linear maps and operators rather than on matrices. In this chapter we pay more attention to matrices as we define the trace and determinant of an operator and then connect these notions to the corresponding notions for matrices. The book concludes with an explanation of the important role played by determinants in the theory of volume and integration.

Our assumptions for this chapter are as follows:

### 10.1 Notation $\mathbf{F}, V$

- $\mathbf{F}$  denotes  $\mathbf{R}$  or  $\mathbf{C}$ .
- $V$  denotes a finite-dimensional nonzero vector space over  $\mathbf{F}$ .

### LEARNING OBJECTIVES FOR THIS CHAPTER

- change of basis and its effect upon the matrix of an operator
- trace of an operator and of a matrix
- determinant of an operator and of a matrix
- determinants and volume

## 10.A Trace

For our study of the trace and determinant, we will need to know how the matrix of an operator changes with a change of basis. Thus we begin this chapter by developing the necessary material about change of basis.

### Change of Basis

With respect to every basis of  $V$ , the matrix of the identity operator  $I \in \mathcal{L}(V)$  is the diagonal matrix with 1's on the diagonal and 0's elsewhere. We also use the symbol  $I$  for the name of this matrix, as shown in the next definition.

#### 10.2 Definition *identity matrix, $I$*

Suppose  $n$  is a positive integer. The  $n$ -by- $n$  diagonal matrix

$$\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

is called the *identity matrix* and is denoted  $I$ .

Note that we use the symbol  $I$  to denote the identity operator (on all vector spaces) and the identity matrix (of all possible sizes). You should always be able to tell from the context which particular meaning of  $I$  is intended. For example, consider the equation  $\mathcal{M}(I) = I$ ; on the left side  $I$  denotes the identity operator, and on the right side  $I$  denotes the identity matrix.

If  $A$  is a square matrix (with entries in  $\mathbf{F}$ , as usual) with the same size as  $I$ , then  $AI = IA = A$ , as you should verify.

#### 10.3 Definition *invertible, inverse, $A^{-1}$*

A square matrix  $A$  is called *invertible* if there is a square matrix  $B$  of the same size such that  $AB = BA = I$ ; we call  $B$  the *inverse* of  $A$  and denote it by  $A^{-1}$ .

*Some mathematicians use the terms **nonsingular**, which means the same as invertible, and **singular**, which means the same as noninvertible.*

The same proof as used in 3.54 shows that if  $A$  is an invertible square matrix, then there is a unique matrix  $B$  such that  $AB = BA = I$  (and thus the notation  $B = A^{-1}$  is justified).

In Section 3.C we defined the matrix of a linear map from one vector space to another with respect to two bases—one basis of the first vector space and another basis of the second vector space. When we study operators, which are linear maps from a vector space to itself, we almost always use the same basis for both vector spaces (after all, the two vector spaces in question are equal). Thus we usually refer to the matrix of an operator with respect to a basis and display at most one basis because we are using one basis in two capacities.

The next result is one of the unusual cases in which we use two different bases even though we have operators from a vector space to itself. It is just a convenient restatement of 3.43 (with  $U$  and  $W$  both equal to  $V$ ), but now we are being more careful to include the various bases explicitly in the notation. The result below holds because we defined matrix multiplication to make it true—see 3.43 and the material preceding it.

#### 10.4 The matrix of the product of linear maps

Suppose  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  and  $w_1, \dots, w_n$  are all bases of  $V$ . Suppose  $S, T \in \mathcal{L}(V)$ . Then

$$\begin{aligned} \mathcal{M}(ST, (u_1, \dots, u_n), (w_1, \dots, w_n)) &= \\ \mathcal{M}(S, (v_1, \dots, v_n), (w_1, \dots, w_n))\mathcal{M}(T, (u_1, \dots, u_n), (v_1, \dots, v_n)). \end{aligned}$$

The next result deals with the matrix of the identity operator  $I$  with respect to two different bases. Note that the  $k^{\text{th}}$  column of the matrix  $\mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n))$  consists of the scalars needed to write  $u_k$  as a linear combination of  $v_1, \dots, v_n$ .

#### 10.5 Matrix of the identity with respect to two bases

Suppose  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  are bases of  $V$ . Then the matrices  $\mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n))$  and  $\mathcal{M}(I, (v_1, \dots, v_n), (u_1, \dots, u_n))$  are invertible, and each is the inverse of the other.

**Proof** In 10.4, replace  $w_j$  with  $u_j$ , and replace  $S$  and  $T$  with  $I$ , getting

$$I = \mathcal{M}(I, (v_1, \dots, v_n), (u_1, \dots, u_n))\mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n)).$$

Now interchange the roles of the  $u$ 's and  $v$ 's, getting

$$I = \mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n))\mathcal{M}(I, (v_1, \dots, v_n), (u_1, \dots, u_n)).$$

These two equations give the desired result. ■

**10.6 Example** Consider the bases  $(4, 2), (5, 3)$  and  $(1, 0), (0, 1)$  of  $\mathbb{F}^2$ . Obviously

$$\mathcal{M}(I, ((4, 2), (5, 3)), ((1, 0), (0, 1))) = \begin{pmatrix} 4 & 5 \\ 2 & 3 \end{pmatrix},$$

because  $I(4, 2) = 4(1, 0) + 2(0, 1)$  and  $I(5, 3) = 5(1, 0) + 3(0, 1)$ .

The inverse of the matrix above is

$$\begin{pmatrix} \frac{3}{2} & -\frac{5}{2} \\ -1 & 2 \end{pmatrix},$$

as you should verify. Thus 10.5 implies that

$$\mathcal{M}(I, ((1, 0), (0, 1)), ((4, 2), (5, 3))) = \begin{pmatrix} \frac{3}{2} & -\frac{5}{2} \\ -1 & 2 \end{pmatrix}.$$

Now we can see how the matrix of  $T$  changes when we change bases. In the result below, we have two different bases of  $V$ . Recall that the notation  $\mathcal{M}(T, (u_1, \dots, u_n))$  is shorthand for  $\mathcal{M}(T, (u_1, \dots, u_n), (u_1, \dots, u_n))$

### 10.7 Change of basis formula

Suppose  $T \in \mathcal{L}(V)$ . Let  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  be bases of  $V$ . Let  $A = \mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n))$ . Then

$$\mathcal{M}(T, (u_1, \dots, u_n)) = A^{-1} \mathcal{M}(T, (v_1, \dots, v_n)) A.$$

**Proof** In 10.4, replace  $w_j$  with  $u_j$  and replace  $S$  with  $I$ , getting

$$10.8 \quad \mathcal{M}(T, (u_1, \dots, u_n)) = A^{-1} \mathcal{M}(T, (u_1, \dots, u_n), (v_1, \dots, v_n)),$$

where we have used 10.5.

Again use 10.4, this time replacing  $w_j$  with  $v_j$ . Also replace  $T$  with  $I$  and replace  $S$  with  $T$ , getting

$$\mathcal{M}(T, (u_1, \dots, u_n), (v_1, \dots, v_n)) = \mathcal{M}(T, (v_1, \dots, v_n)) A.$$

Substituting the equation above into 10.8 gives the desired result. ■

## Trace: A Connection Between Operators and Matrices

Suppose  $T \in \mathcal{L}(V)$  and  $\lambda$  is an eigenvalue of  $T$ . Let  $n = \dim V$ . Recall that we defined the multiplicity of  $\lambda$  to be the dimension of the generalized eigenspace  $G(\lambda, T)$  (see 8.24) and that this multiplicity equals  $\dim \text{null}(T - \lambda I)^n$  (see 8.11). Recall also that if  $V$  is a complex vector space, then the sum of the multiplicities of all the eigenvalues of  $T$  equals  $n$  (see 8.26).

In the definition below, the sum of the eigenvalues “with each eigenvalue repeated according to its multiplicity” means that if  $\lambda_1, \dots, \lambda_m$  are the distinct eigenvalues of  $T$  (or of  $T_C$  if  $V$  is a real vector space) with multiplicities  $d_1, \dots, d_m$ , then the sum is

$$d_1\lambda_1 + \cdots + d_m\lambda_m.$$

Or if you prefer to list the eigenvalues with each repeated according to its multiplicity, then the eigenvalues could be denoted  $\lambda_1, \dots, \lambda_n$  (where the index  $n$  equals  $\dim V$ ) and the sum is

$$\lambda_1 + \cdots + \lambda_n.$$

### 10.9 Definition trace of an operator

Suppose  $T \in \mathcal{L}(V)$ .

- If  $\mathbf{F} = \mathbf{C}$ , then the *trace* of  $T$  is the sum of the eigenvalues of  $T$ , with each eigenvalue repeated according to its multiplicity.
- If  $\mathbf{F} = \mathbf{R}$ , then the *trace* of  $T$  is the sum of the eigenvalues of  $T_C$ , with each eigenvalue repeated according to its multiplicity.

The trace of  $T$  is denoted by  $\text{trace } T$ .

### 10.10 Example

Suppose  $T \in \mathcal{L}(\mathbf{C}^3)$  is the operator whose matrix is

$$\begin{pmatrix} 3 & -1 & -2 \\ 3 & 2 & -3 \\ 1 & 2 & 0 \end{pmatrix}.$$

Then the eigenvalues of  $T$  are  $1, 2 + 3i$ , and  $2 - 3i$ , each with multiplicity 1, as you can verify. Computing the sum of the eigenvalues, we find that  $\text{trace } T = 1 + (2 + 3i) + (2 - 3i)$ ; in other words,  $\text{trace } T = 5$ .

The trace has a close connection with the characteristic polynomial. Suppose  $\lambda_1, \dots, \lambda_n$  are the eigenvalues of  $T$  (or of  $T_C$  if  $V$  is a real vector space) with each eigenvalue repeated according to its multiplicity. Then by definition (see 8.34 and 9.21), the characteristic polynomial of  $T$  equals

$$(z - \lambda_1) \cdots (z - \lambda_n).$$

Expanding the polynomial above, we can write the characteristic polynomial of  $T$  in the form

$$10.11 \quad z^n - (\lambda_1 + \cdots + \lambda_n)z^{n-1} + \cdots + (-1)^n(\lambda_1 \cdots \lambda_n).$$

The expression above immediately leads to the following result.

### 10.12 Trace and characteristic polynomial

Suppose  $T \in \mathcal{L}(V)$ . Let  $n = \dim V$ . Then trace  $T$  equals the negative of the coefficient of  $z^{n-1}$  in the characteristic polynomial of  $T$ .

Most of the rest of this section is devoted to discovering how to compute trace  $T$  from the matrix of  $T$  (with respect to an arbitrary basis).

Let's start with the easiest situation. Suppose  $V$  is a complex vector space,  $T \in \mathcal{L}(V)$ , and we choose a basis of  $V$  as in 8.29. With respect to that basis,  $T$  has an upper-triangular matrix with the diagonal of the matrix containing precisely the eigenvalues of  $T$ , each repeated according to its multiplicity. Thus trace  $T$  equals the sum of the diagonal entries of  $\mathcal{M}(T)$  with respect to that basis.

The same formula works for the operator  $T \in \mathcal{L}(\mathbb{C}^3)$  in Example 10.10 whose trace equals 5. In that example, the matrix is not in upper-triangular form. However, the sum of the diagonal entries of the matrix in that example equals 5, which is the trace of the operator  $T$ .

At this point you should suspect that trace  $T$  equals the sum of the diagonal entries of the matrix of  $T$  with respect to an arbitrary basis. Remarkably, this suspicion turns out to be true. To prove it, we start by making the following definition.

### 10.13 Definition trace of a matrix

The **trace** of a square matrix  $A$ , denoted  $\text{trace } A$ , is defined to be the sum of the diagonal entries of  $A$ .

Now we have defined the trace of an operator and the trace of a square matrix, using the same word “trace” in two different contexts. This would be bad terminology unless the two concepts turn out to be essentially the same. As we will see, it is indeed true that  $\text{trace } T = \text{trace } \mathcal{M}(T, (v_1, \dots, v_n))$ , where  $v_1, \dots, v_n$  is an arbitrary basis of  $V$ . We will need the following result for the proof.

### 10.14 Trace of $AB$ equals trace of $BA$

If  $A$  and  $B$  are square matrices of the same size, then

$$\text{trace}(AB) = \text{trace}(BA).$$

**Proof** Suppose

$$A = \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{n,1} & \dots & A_{n,n} \end{pmatrix}, \quad B = \begin{pmatrix} B_{1,1} & \dots & B_{1,n} \\ \vdots & & \vdots \\ B_{n,1} & \dots & B_{n,n} \end{pmatrix}.$$

The  $j^{\text{th}}$  term on the diagonal of  $AB$  equals

$$\sum_{k=1}^n A_{j,k} B_{k,j}.$$

Thus

$$\begin{aligned} \text{trace}(AB) &= \sum_{j=1}^n \sum_{k=1}^n A_{j,k} B_{k,j} \\ &= \sum_{k=1}^n \sum_{j=1}^n B_{k,j} A_{j,k} \\ &= \sum_{k=1}^n k^{\text{th}} \text{ term on the diagonal of } BA \\ &= \text{trace}(BA), \end{aligned}$$

as desired. ■

Now we can prove that the sum of the diagonal entries of the matrix of an operator is independent of the basis with respect to which the matrix is computed.

### 10.15 Trace of matrix of operator does not depend on basis

Let  $T \in \mathcal{L}(V)$ . Suppose  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  are bases of  $V$ . Then

$$\text{trace } \mathcal{M}(T, (u_1, \dots, u_n)) = \text{trace } \mathcal{M}(T, (v_1, \dots, v_n)).$$

**Proof** Let  $A = \mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n))$ . Then

$$\begin{aligned} \text{trace } \mathcal{M}(T, (u_1, \dots, u_n)) &= \text{trace}\left(A^{-1}(\mathcal{M}(T, (v_1, \dots, v_n))A)\right) \\ &= \text{trace}\left((\mathcal{M}(T, (v_1, \dots, v_n))A)A^{-1}\right) \\ &= \text{trace } \mathcal{M}(T, (v_1, \dots, v_n)), \end{aligned}$$

where the first equality comes from 10.7 and the second equality follows from 10.14. The third equality completes the proof. ■

The result below, which is the most important result in this section, states that the trace of an operator equals the sum of the diagonal entries of the matrix of the operator. This theorem does not specify a basis because, by the result above, the sum of the diagonal entries of the matrix of an operator is the same for every choice of basis.

### 10.16 Trace of an operator equals trace of its matrix

Suppose  $T \in \mathcal{L}(V)$ . Then  $\text{trace } T = \text{trace } \mathcal{M}(T)$ .

**Proof** As noted above,  $\text{trace } \mathcal{M}(T)$  is independent of which basis of  $V$  we choose (by 10.15). Thus to show that

$$\text{trace } T = \text{trace } \mathcal{M}(T)$$

for every basis of  $V$ , we need only show that the equation above holds for some basis of  $V$ .

As we have already discussed, if  $V$  is a complex vector space, then choosing the basis as in 8.29 gives the desired result. If  $V$  is a real vector space, then applying the complex case to the complexification  $T_C$  (which is used to define  $\text{trace } T$ ) gives the desired result. ■

If we know the matrix of an operator on a complex vector space, the result above allows us to find the sum of all the eigenvalues without finding any of the eigenvalues, as shown by the next example.

**10.17 Example** Consider the operator on  $\mathbf{C}^5$  whose matrix is

$$\begin{pmatrix} 0 & 0 & 0 & 0 & -3 \\ 1 & 0 & 0 & 0 & 6 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

No one can find an exact formula for any of the eigenvalues of this operator. However, we do know that the sum of the eigenvalues equals 0, because the sum of the diagonal entries of the matrix above equals 0.

We can use 10.16 to give easy proofs of some useful properties about traces of operators by shifting to the language of traces of matrices, where certain properties have already been proved or are obvious. The proof of the next result is an example of this technique. The eigenvalues of  $S + T$  are not, in general, formed from adding together eigenvalues of  $S$  and eigenvalues of  $T$ . Thus the next result would be difficult to prove without using 10.16.

### 10.18 Trace is additive

Suppose  $S, T \in \mathcal{L}(V)$ . Then  $\text{trace}(S + T) = \text{trace } S + \text{trace } T$ .

**Proof** Choose a basis of  $V$ . Then

$$\begin{aligned} \text{trace}(S + T) &= \text{trace } \mathcal{M}(S + T) \\ &= \text{trace}(\mathcal{M}(S) + \mathcal{M}(T)) \\ &= \text{trace } \mathcal{M}(S) + \text{trace } \mathcal{M}(T) \\ &= \text{trace } S + \text{trace } T, \end{aligned}$$

where again the first and last equalities come from 10.16; the third equality is obvious from the definition of the trace of a matrix. ■

The techniques we have developed have the following curious consequence. A generalization of this result to infinite-dimensional vector spaces has important consequences in modern physics, particularly in quantum theory.

*The statement of the next result does not involve traces, although the short proof uses traces. Whenever something like this happens in mathematics, we can be sure that a good definition lurks in the background.*

### 10.19 The identity is not the difference of $ST$ and $TS$

There do not exist operators  $S, T \in \mathcal{L}(V)$  such that  $ST - TS = I$ .

**Proof** Suppose  $S, T \in \mathcal{L}(V)$ . Choose a basis of  $V$ . Then

$$\begin{aligned}\text{trace}(ST - TS) &= \text{trace}(ST) - \text{trace}(TS) \\ &= \text{trace } \mathcal{M}(ST) - \text{trace } \mathcal{M}(TS) \\ &= \text{trace}(\mathcal{M}(S)\mathcal{M}(T)) - \text{trace}(\mathcal{M}(T)\mathcal{M}(S)) \\ &= 0,\end{aligned}$$

where the first equality comes from 10.18, the second equality comes from 10.16, the third equality comes from 3.43, and the fourth equality comes from 10.14. Clearly the trace of  $I$  equals  $\dim V$ , which is not 0. Because  $ST - TS$  and  $I$  have different traces, they cannot be equal. ■

## EXERCISES 10.A

---

- 1 Suppose  $T \in \mathcal{L}(V)$  and  $v_1, \dots, v_n$  is a basis of  $V$ . Prove that the matrix  $\mathcal{M}(T, (v_1, \dots, v_n))$  is invertible if and only if  $T$  is invertible.
- 2 Suppose  $A$  and  $B$  are square matrices of the same size and  $AB = I$ . Prove that  $BA = I$ .
- 3 Suppose  $T \in \mathcal{L}(V)$  has the same matrix with respect to every basis of  $V$ . Prove that  $T$  is a scalar multiple of the identity operator.
- 4 Suppose  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  are bases of  $V$ . Let  $T \in \mathcal{L}(V)$  be the operator such that  $Tv_k = u_k$  for  $k = 1, \dots, n$ . Prove that
$$\mathcal{M}(T, (v_1, \dots, v_n)) = \mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n)).$$
- 5 Suppose  $B$  is a square matrix with complex entries. Prove that there exists an invertible square matrix  $A$  with complex entries such that  $A^{-1}BA$  is an upper-triangular matrix.
- 6 Give an example of a real vector space  $V$  and  $T \in \mathcal{L}(V)$  such that  $\text{trace}(T^2) < 0$ .
- 7 Suppose  $V$  is a real vector space,  $T \in \mathcal{L}(V)$ , and  $V$  has a basis consisting of eigenvectors of  $T$ . Prove that  $\text{trace}(T^2) \geq 0$ .

- 8** Suppose  $V$  is an inner product space and  $v, w \in V$ . Define  $T \in \mathcal{L}(V)$  by  $Tu = \langle u, v \rangle w$ . Find a formula for trace  $T$ .
- 9** Suppose  $P \in \mathcal{L}(V)$  satisfies  $P^2 = P$ . Prove that

$$\text{trace } P = \dim \text{range } P.$$

- 10** Suppose  $V$  is an inner product space and  $T \in \mathcal{L}(V)$ . Prove that

$$\text{trace } T^* = \overline{\text{trace } T}.$$

- 11** Suppose  $V$  is an inner product space. Suppose  $T \in \mathcal{L}(V)$  is a positive operator and  $\text{trace } T = 0$ . Prove that  $T = 0$ .
- 12** Suppose  $V$  is an inner product space and  $P, Q \in \mathcal{L}(V)$  are orthogonal projections. Prove that  $\text{trace}(PQ) \geq 0$ .
- 13** Suppose  $T \in \mathcal{L}(\mathbf{C}^3)$  is the operator whose matrix is

$$\begin{pmatrix} 51 & -12 & -21 \\ 60 & -40 & -28 \\ 57 & -68 & 1 \end{pmatrix}.$$

Someone tells you (accurately) that  $-48$  and  $24$  are eigenvalues of  $T$ . Without using a computer or writing anything down, find the third eigenvalue of  $T$ .

- 14** Suppose  $T \in \mathcal{L}(V)$  and  $c \in \mathbf{F}$ . Prove that  $\text{trace}(cT) = c \text{ trace } T$ .
- 15** Suppose  $S, T \in \mathcal{L}(V)$ . Prove that  $\text{trace}(ST) = \text{trace}(TS)$ .
- 16** Prove or give a counterexample: if  $S, T \in \mathcal{L}(V)$ , then  $\text{trace}(ST) = (\text{trace } S)(\text{trace } T)$ .
- 17** Suppose  $T \in \mathcal{L}(V)$  is such that  $\text{trace}(ST) = 0$  for all  $S \in \mathcal{L}(V)$ . Prove that  $T = 0$ .
- 18** Suppose  $V$  is an inner product space with orthonormal basis  $e_1, \dots, e_n$  and  $T \in \mathcal{L}(V)$ . Prove that

$$\text{trace}(T^*T) = \|Te_1\|^2 + \dots + \|Te_n\|^2.$$

Conclude that the right side of the equation above is independent of which orthonormal basis  $e_1, \dots, e_n$  is chosen for  $V$ .

- 19** Suppose  $V$  is an inner product space. Prove that

$$\langle S, T \rangle = \text{trace}(ST^*)$$

defines an inner product on  $\mathcal{L}(V)$ .

- 20** Suppose  $V$  is a complex inner product space and  $T \in \mathcal{L}(V)$ . Let  $\lambda_1, \dots, \lambda_n$  be the eigenvalues of  $T$ , repeated according to multiplicity. Suppose

$$\begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{n,1} & \dots & A_{n,n} \end{pmatrix}$$

is the matrix of  $T$  with respect to some orthonormal basis of  $V$ . Prove that

$$|\lambda_1|^2 + \dots + |\lambda_n|^2 \leq \sum_{k=1}^n \sum_{j=1}^n |A_{j,k}|^2.$$

- 21** Suppose  $V$  is an inner product space. Suppose  $T \in \mathcal{L}(V)$  and

$$\|T^*v\| \leq \|Tv\|$$

for every  $v \in V$ . Prove that  $T$  is normal.

[The exercise above fails on infinite-dimensional inner product spaces, leading to what are called hyponormal operators, which have a well-developed theory.]

## 10.B Determinant

### Determinant of an Operator

Now we are ready to define the determinant of an operator. Notice that the definition below mimics the approach we took when defining the trace, with the product of the eigenvalues replacing the sum of the eigenvalues.

#### 10.20 Definition determinant of an operator, $\det T$

Suppose  $T \in \mathcal{L}(V)$ .

- If  $\mathbf{F} = \mathbf{C}$ , then the **determinant** of  $T$  is the product of the eigenvalues of  $T$ , with each eigenvalue repeated according to its multiplicity.
- If  $\mathbf{F} = \mathbf{R}$ , then the **determinant** of  $T$  is the product of the eigenvalues of  $T_{\mathbf{C}}$ , with each eigenvalue repeated according to its multiplicity.

The determinant of  $T$  is denoted by  $\det T$ .

If  $\lambda_1, \dots, \lambda_m$  are the distinct eigenvalues of  $T$  (or of  $T_{\mathbf{C}}$  if  $V$  is a real vector space) with multiplicities  $d_1, \dots, d_m$ , then the definition above implies

$$\det T = \lambda_1^{d_1} \cdots \lambda_m^{d_m}.$$

Or if you prefer to list the eigenvalues with each repeated according to its multiplicity, then the eigenvalues could be denoted  $\lambda_1, \dots, \lambda_n$  (where the index  $n$  equals  $\dim V$ ) and the definition above implies

$$\det T = \lambda_1 \cdots \lambda_n.$$

#### 10.21 Example Suppose $T \in \mathcal{L}(\mathbf{C}^3)$ is the operator whose matrix is

$$\begin{pmatrix} 3 & -1 & -2 \\ 3 & 2 & -3 \\ 1 & 2 & 0 \end{pmatrix}.$$

Then the eigenvalues of  $T$  are  $1, 2 + 3i$ , and  $2 - 3i$ , each with multiplicity 1, as you can verify. Computing the product of the eigenvalues, we find that  $\det T = 1 \cdot (2 + 3i) \cdot (2 - 3i)$ ; in other words,  $\det T = 13$ .

The determinant has a close connection with the characteristic polynomial. Suppose  $\lambda_1, \dots, \lambda_n$  are the eigenvalues of  $T$  (or of  $T_C$  if  $V$  is a real vector space) with each eigenvalue repeated according to its multiplicity. Then the expression for the characteristic polynomial of  $T$  given by 10.11 gives the following result.

### 10.22 Determinant and characteristic polynomial

Suppose  $T \in \mathcal{L}(V)$ . Let  $n = \dim V$ . Then  $\det T$  equals  $(-1)^n$  times the constant term of the characteristic polynomial of  $T$ .

Combining the result above and 10.12, we have the following result.

### 10.23 Characteristic polynomial, trace, and determinant

Suppose  $T \in \mathcal{L}(V)$ . Then the characteristic polynomial of  $T$  can be written as

$$z^n - (\text{trace } T)z^{n-1} + \cdots + (-1)^n(\det T).$$

We turn now to some simple but important properties of determinants. Later we will discover how to calculate  $\det T$  from the matrix of  $T$  (with respect to an arbitrary basis).

The crucial result below has an easy proof due to our definition.

### 10.24 Invertible is equivalent to nonzero determinant

An operator on  $V$  is invertible if and only if its determinant is nonzero.

**Proof** First suppose  $V$  is a complex vector space and  $T \in \mathcal{L}(V)$ . The operator  $T$  is invertible if and only if 0 is not an eigenvalue of  $T$ . Clearly this happens if and only if the product of the eigenvalues of  $T$  is not 0. Thus  $T$  is invertible if and only if  $\det T \neq 0$ , as desired.

Now consider the case where  $V$  is a real vector space and  $T \in \mathcal{L}(V)$ . Again,  $T$  is invertible if and only if 0 is not an eigenvalue of  $T$ , which happens if and only if 0 is not an eigenvalue of  $T_C$  (because  $T_C$  and  $T$  have the same real eigenvalues by 9.11). Thus again we see that  $T$  is invertible if and only if  $\det T \neq 0$ . ■

Some textbooks take the result below as the definition of the characteristic polynomial and then have our definition of the characteristic polynomial as a consequence.

### 10.25 Characteristic polynomial of $T$ equals $\det(zI - T)$

Suppose  $T \in \mathcal{L}(V)$ . Then the characteristic polynomial of  $T$  equals  $\det(zI - T)$ .

**Proof** First suppose  $V$  is a complex vector space. If  $\lambda, z \in \mathbf{C}$ , then  $\lambda$  is an eigenvalue of  $T$  if and only if  $z - \lambda$  is an eigenvalue of  $zI - T$ , as can be seen from the equation

$$-(T - \lambda I) = (zI - T) - (z - \lambda)I.$$

Raising both sides of this equation to the  $\dim V$  power and then taking null spaces of both sides shows that the multiplicity of  $\lambda$  as an eigenvalue of  $T$  equals the multiplicity of  $z - \lambda$  as an eigenvalue of  $zI - T$ .

Let  $\lambda_1, \dots, \lambda_n$  denote the eigenvalues of  $T$ , repeated according to multiplicity. Thus for  $z \in \mathbf{C}$ , the paragraph above shows that the eigenvalues of  $zI - T$  are  $z - \lambda_1, \dots, z - \lambda_n$ , repeated according to multiplicity. The determinant of  $zI - T$  is the product of these eigenvalues. In other words,

$$\det(zI - T) = (z - \lambda_1) \cdots (z - \lambda_n).$$

The right side of the equation above is, by definition, the characteristic polynomial of  $T$ , completing the proof when  $V$  is a complex vector space.

Now suppose  $V$  is a real vector space. Applying the complex case to  $T_C$  gives the desired result. ■

## Determinant of a Matrix

Our next task is to discover how to compute  $\det T$  from the matrix of  $T$  (with respect to an arbitrary basis). Let's start with the easiest situation. Suppose  $V$  is a complex vector space,  $T \in \mathcal{L}(V)$ , and we choose a basis of  $V$  as in 8.29. With respect to that basis,  $T$  has an upper-triangular matrix with the diagonal of the matrix containing precisely the eigenvalues of  $T$ , each repeated according to its multiplicity. Thus  $\det T$  equals the product of the diagonal entries of  $\mathcal{M}(T)$  with respect to that basis.

When dealing with the trace in the previous section, we discovered that the formula (trace = sum of diagonal entries) that worked for the upper-triangular matrix given by 8.29 also worked with respect to an arbitrary basis. Could that also work for determinants? In other words, is the determinant of an operator equal to the product of the diagonal entries of the matrix of the operator with respect to an arbitrary basis?

Unfortunately, the determinant is more complicated than the trace. In particular,  $\det T$  need not equal the product of the diagonal entries of  $\mathcal{M}(T)$  with respect to an arbitrary basis. For example, the operator in Example 10.21 has determinant 13 but the product of the diagonal entries of its matrix equals 0.

For each square matrix  $A$ , we want to define the determinant of  $A$ , denoted  $\det A$ , so that  $\det T = \det \mathcal{M}(T)$  regardless of which basis is used to compute  $\mathcal{M}(T)$ . We begin our search for the correct definition of the determinant of a matrix by calculating the determinants of some special operators.

**10.26 Example** Suppose  $a_1, \dots, a_n \in \mathbf{F}$ . Let

$$A = \begin{pmatrix} 0 & & & & a_n \\ a_1 & 0 & & & \\ & a_2 & 0 & & \\ & & \ddots & \ddots & \\ & & & a_{n-1} & 0 \end{pmatrix};$$

here all entries of the matrix are 0 except for the upper-right corner and along the line just below the diagonal. Suppose  $v_1, \dots, v_n$  is a basis of  $V$  and  $T \in \mathcal{L}(V)$  is such that  $\mathcal{M}(T, (v_1, \dots, v_n)) = A$ . Find the determinant of  $T$ .

**Solution** First assume  $a_j \neq 0$  for each  $j = 1, \dots, n - 1$ . Note that the list  $v_1, Tv_1, T^2v_1, \dots, T^{n-1}v_1$  equals  $v_1, a_1v_2, a_1a_2v_3, \dots, a_1 \cdots a_{n-1}v_n$ .

*Computing the minimal polynomial is often an efficient method of finding the characteristic polynomial, as is done in this example.*

Thus  $v_1, Tv_1, \dots, T^{n-1}v_1$  is linearly independent (because the  $a$ 's are all nonzero). Hence if  $p$  is a monic polynomial with degree at most  $n - 1$ , then  $p(T)v_1 \neq 0$ . Thus the minimal polynomial of  $T$  cannot have degree less than  $n$ .

As you should verify,  $T^n v_j = a_1 \cdots a_n v_j$  for each  $j$ . Thus we have  $T^n = a_1 \cdots a_n I$ . Hence  $z^n - a_1 \cdots a_n$  is the minimal polynomial of  $T$ . Because  $n = \dim V$  and the characteristic polynomial is a polynomial multiple of the minimal polynomial (9.26), this implies that  $z^n - a_1 \cdots a_n$  is also the characteristic polynomial of  $T$ .

Thus 10.22 implies that

$$\det T = (-1)^{n-1} a_1 \cdots a_n.$$

If some  $a_j$  equals 0, then  $Tv_j = 0$  for some  $j$ , which implies that 0 is an eigenvalue of  $T$  and hence  $\det T = 0$ . In other words, the formula above also holds if some  $a_j$  equals 0.

Thus in order to have  $\det T = \det \mathcal{M}(T)$ , we will have to make the determinant of the matrix in Example 10.26 equal to  $(-1)^{n-1} a_1 \cdots a_n$ . However, we do not yet have enough evidence to make a reasonable guess about the proper definition of the determinant of an arbitrary square matrix.

To compute the determinants of a more complicated class of operators, we introduce the notion of permutation.

**10.27 Definition *permutation*,  $\text{perm } n$**

- A **permutation** of  $(1, \dots, n)$  is a list  $(m_1, \dots, m_n)$  that contains each of the numbers  $1, \dots, n$  exactly once.
- The set of all permutations of  $(1, \dots, n)$  is denoted  $\text{perm } n$ .

For example,  $(2, 3, 4, 5, 1) \in \text{perm } 5$ . You should think of an element of  $\text{perm } n$  as a rearrangement of the first  $n$  integers.

---

**10.28 Example** Suppose  $a_1, \dots, a_n \in \mathbf{F}$  and  $v_1, \dots, v_n$  is a basis of  $V$ . Consider a permutation  $(p_1, \dots, p_n) \in \text{perm } n$  that can be obtained as follows: break  $(1, \dots, n)$  into lists of consecutive integers and in each list move the first term to the end of that list. For example, taking  $n = 9$ , the permutation

$$(2, 3, 1, 5, 6, 7, 4, 9, 8)$$

is obtained from  $(1, 2, 3), (4, 5, 6, 7), (8, 9)$  by moving the first term of each of these lists to the end, producing  $(2, 3, 1), (5, 6, 7, 4), (9, 8)$ , and then putting these together to form the permutation displayed above.

Let  $T \in \mathcal{L}(V)$  be the operator such that

$$Tv_k = a_k v_{p_k}$$

for  $k = 1, \dots, n$ . Find  $\det T$ .

**Solution** This generalizes Example 10.26, because if  $(p_1, \dots, p_n)$  is the permutation  $(2, 3, \dots, n, 1)$ , then our operator  $T$  is the same as the operator  $T$  in Example 10.26.

With respect to the basis  $v_1, \dots, v_n$ , the matrix of the operator  $T$  is a block diagonal matrix

$$A = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_M \end{pmatrix},$$

where each block is a square matrix of the form of the matrix in 10.26.

Correspondingly, we can write  $V = V_1 \oplus \cdots \oplus V_M$ , where each  $V_j$  is invariant under  $T$  and each  $T|_{V_j}$  is of the form of the operator in 10.26. Because  $\det T = (\det T|_{V_1}) \cdots (\det T|_{V_M})$  (because the dimensions of the generalized eigenspaces in the  $V_j$  add up to  $\dim V$ ), we have

$$\det T = (-1)^{n_1-1} \cdots (-1)^{n_M-1} a_1 \cdots a_n,$$

where  $V_j$  has dimension  $n_j$  (and correspondingly each  $A_j$  has size  $n_j$ -by- $n_j$ ) and we have used the result from 10.26.

The number  $(-1)^{n_1-1} \cdots (-1)^{n_M-1}$  that appears above is called the sign of the corresponding permutation  $(p_1, \dots, p_n)$ , denoted  $\text{sign}(p_1, \dots, p_n)$  [this is a temporary definition that we will change to an equivalent definition later, when we define the sign of an arbitrary permutation].

To put this into a form that does not depend on the particular permutation  $(p_1, \dots, p_n)$ , let  $A_{j,k}$  denote the entry in row  $j$ , column  $k$ , of the matrix  $A$  from Example 10.28. Thus

$$A_{j,k} = \begin{cases} 0 & \text{if } j \neq p_k; \\ a_k & \text{if } j = p_k. \end{cases}$$

Example 10.28 shows that we want

$$10.29 \quad \det A = \sum_{(m_1, \dots, m_n) \in \text{perm } n} (\text{sign}(m_1, \dots, m_n)) A_{m_1,1} \cdots A_{m_n,n};$$

note that each summand is 0 except the one corresponding to the permutation  $(p_1, \dots, p_n)$  [which is why it does not matter that the sign of the other permutations is not yet defined].

We can now guess that  $\det A$  should be defined by 10.29 for an arbitrary square matrix  $A$ . This will turn out to be correct. We will now dispense with the motivation and begin the more formal approach. First we will need to define the sign of an arbitrary permutation.

### 10.30 Definition sign of a permutation

- The **sign** of a permutation  $(m_1, \dots, m_n)$  is defined to be 1 if the number of pairs of integers  $(j, k)$  with  $1 \leq j < k \leq n$  such that  $j$  appears after  $k$  in the list  $(m_1, \dots, m_n)$  is even and  $-1$  if the number of such pairs is odd.
- In other words, the sign of a permutation equals 1 if the natural order has been changed an even number of times and equals  $-1$  if the natural order has been changed an odd number of times.

**10.31 Example** *sign of permutation*

- The only pair of integers  $(j, k)$  with  $j < k$  such that  $j$  appears after  $k$  in the list  $(2, 1, 3, 4)$  is  $(1, 2)$ . Thus the permutation  $(2, 1, 3, 4)$  has sign  $-1$ .
- In the permutation  $(2, 3, \dots, n, 1)$ , the only pairs  $(j, k)$  with  $j < k$  that appear with changed order are  $(1, 2), (1, 3), \dots, (1, n)$ ; because we have  $n - 1$  such pairs, the sign of this permutation equals  $(-1)^{n-1}$  (note that the same quantity appeared in Example 10.26).

The next result shows that interchanging two entries of a permutation changes the sign of the permutation.

**10.32 Interchanging two entries in a permutation**

Interchanging two entries in a permutation multiplies the sign of the permutation by  $-1$ .

**Proof** Suppose we have two permutations, where the second permutation is obtained from the first by interchanging two entries. If the two interchanged entries were in their natural order in the first permutation, then they no longer are in the second permutation, and vice versa, for a net change (so far) of 1 or  $-1$  (both odd numbers) in the number of pairs not in their natural order.

Consider each entry between the two interchanged entries. If an intermediate entry was originally in the natural order with respect to both interchanged entries, then it is now in the natural order with respect to neither interchanged entry. Similarly, if an intermediate entry was originally in the natural order with respect to neither of the interchanged entries, then it is now in the natural order with respect to both interchanged entries. If an intermediate entry was originally in the natural order with respect to exactly one of the interchanged entries, then that is still true. Thus the net change for each intermediate entry in the number of pairs not in their natural order is 2,  $-2$ , or 0 (all even numbers).

Some texts use the term **signum**, which means the same as sign.

For all the other entries, there is no change in the number of pairs not in their natural order. Thus the total net change in the number of pairs not in their natural order is an odd number. Thus the sign of the second permutation equals  $-1$  times the sign of the first permutation. ■

Our motivation for the next definition comes from 10.29.

**10.33 Definition determinant of a matrix,  $\det A$** 

Suppose  $A$  is an  $n$ -by- $n$  matrix

$$A = \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{n,1} & \dots & A_{n,n} \end{pmatrix}.$$

The **determinant** of  $A$ , denoted  $\det A$ , is defined by

$$\det A = \sum_{(m_1, \dots, m_n) \in \text{perm } n} (\text{sign}(m_1, \dots, m_n)) A_{m_1,1} \cdots A_{m_n,n}.$$

**10.34 Example determinants**

- If  $A$  is the 1-by-1 matrix  $[A_{1,1}]$ , then  $\det A = A_{1,1}$ , because  $\text{perm } 1$  has only one element, namely (1), which has sign 1.
- Clearly  $\text{perm } 2$  has only two elements, namely (1, 2), which has sign 1, and (2, 1), which has sign  $-1$ . Thus

$$\det \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} = A_{1,1}A_{2,2} - A_{2,1}A_{1,2}.$$

*The set  $\text{perm } 3$  contains six elements. In general,  $\text{perm } n$  contains  $n!$  elements. Note that  $n!$  rapidly grows large as  $n$  increases.*

To make sure you understand this process, you should now find the formula for the determinant of an arbitrary 3-by-3 matrix using just the definition given above.

**10.35 Example** Compute the determinant of an upper-triangular matrix

$$A = \begin{pmatrix} A_{1,1} & & * \\ & \ddots & \\ 0 & & A_{n,n} \end{pmatrix}.$$

**Solution** The permutation  $(1, 2, \dots, n)$  has sign 1 and thus contributes a term of  $A_{1,1} \cdots A_{n,n}$  to the sum defining  $\det A$  in 10.33. Any other permutation  $(m_1, \dots, m_n) \in \text{perm } n$  contains at least one entry  $m_j$  with  $m_j > j$ , which means that  $A_{m_j,j} = 0$  (because  $A$  is upper triangular). Thus all the other terms in the sum in 10.33 make no contribution.

Hence  $\det A = A_{1,1} \cdots A_{n,n}$ . In other words, the determinant of an upper-triangular matrix equals the product of the diagonal entries.

Suppose  $V$  is a complex vector space,  $T \in \mathcal{L}(V)$ , and we choose a basis of  $V$  as in 8.29. With respect to that basis,  $T$  has an upper-triangular matrix with the diagonal of the matrix containing precisely the eigenvalues of  $T$ , each repeated according to its multiplicity. Thus Example 10.35 tells us that  $\det T = \det \mathcal{M}(T)$ , where the matrix is with respect to that basis.

Our goal is to prove that  $\det T = \det \mathcal{M}(T)$  for every basis of  $V$ , not just the basis from 8.29. To do this, we will need to develop some properties of determinants of matrices. The result below is the first of the properties we will need.

### 10.36 Interchanging two columns in a matrix

Suppose  $A$  is a square matrix and  $B$  is the matrix obtained from  $A$  by interchanging two columns. Then

$$\det A = -\det B.$$

**Proof** Think of the sum defining  $\det A$  in 10.33 and the corresponding sum defining  $\det B$ . The same products of  $A_{j,k}$ 's appear in both sums, although they correspond to different permutations. The permutation corresponding to a given product of  $A_{j,k}$ 's when computing  $\det B$  is obtained by interchanging two entries in the corresponding permutation when computing  $\det A$ , thus multiplying the sign of the permutation by  $-1$  (see 10.32). Hence we see that  $\det A = -\det B$ . ■

If  $T \in \mathcal{L}(V)$  and the matrix of  $T$  (with respect to some basis) has two equal columns, then  $T$  is not injective and hence  $\det T = 0$ . Although this comment makes the next result plausible, it cannot be used in the proof, because we do not yet know that  $\det T = \det \mathcal{M}(T)$  for every choice of basis.

### 10.37 Matrices with two equal columns

If  $A$  is a square matrix that has two equal columns, then  $\det A = 0$ .

**Proof** Suppose  $A$  is a square matrix that has two equal columns. Interchanging the two equal columns of  $A$  gives the original matrix  $A$ . Thus from 10.36 (with  $B = A$ ), we have

$$\det A = -\det A,$$

which implies that  $\det A = 0$ . ■

Recall from 3.44 that if  $A$  is an  $n$ -by- $n$  matrix

$$A = \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{n,1} & \dots & A_{n,n} \end{pmatrix},$$

then we can think of the  $k^{\text{th}}$  column of  $A$  as an  $n$ -by-1 matrix denoted  $A_{\cdot,k}$ :

$$A_{\cdot,k} = \begin{pmatrix} A_{1,k} \\ \vdots \\ A_{n,k} \end{pmatrix}.$$

*Some books define the determinant to be the function defined on the square matrices that is linear as a function of each column separately and that satisfies 10.38 and  $\det I = 1$ . To prove that such a function exists and that it is unique takes a nontrivial amount of work.*

Note that  $A_{j,k}$ , with two subscripts, denotes an entry of  $A$ , whereas  $A_{\cdot,k}$ , with a dot as a placeholder and one subscript, denotes a column of  $A$ . This notation allows us to write  $A$  in the form

$$(A_{\cdot,1} \ \dots \ A_{\cdot,n}),$$

which will be useful.

The next result shows that a permutation of the columns of a matrix changes the determinant by a factor of the sign of the permutation.

### 10.38 Permuting the columns of a matrix

Suppose  $A = (A_{\cdot,1} \ \dots \ A_{\cdot,n})$  is an  $n$ -by- $n$  matrix and  $(m_1, \dots, m_n)$  is a permutation. Then

$$\det(A_{\cdot,m_1} \ \dots \ A_{\cdot,m_n}) = (\text{sign}(m_1, \dots, m_n)) \det A.$$

**Proof** We can transform the matrix  $(A_{\cdot,m_1} \ \dots \ A_{\cdot,m_n})$  into  $A$  through a series of steps. In each step, we interchange two columns and hence multiply the determinant by  $-1$  (see 10.36). The number of steps needed equals the number of steps needed to transform the permutation  $(m_1, \dots, m_n)$  into the permutation  $(1, \dots, n)$  by interchanging two entries in each step. The proof is completed by noting that the number of such steps is even if  $(m_1, \dots, m_n)$  has sign 1, odd if  $(m_1, \dots, m_n)$  has sign  $-1$  (this follows from 10.32, along with the observation that the permutation  $(1, \dots, n)$  has sign 1). ■

The next result about determinants will also be useful.

### 10.39 Determinant is a linear function of each column

Suppose  $k, n$  are positive integers with  $1 \leq k \leq n$ . Fix  $n$ -by-1 matrices  $A_{\cdot,1}, \dots, A_{\cdot,n}$  except  $A_{\cdot,k}$ . Then the function that takes an  $n$ -by-1 column vector  $A_{\cdot,k}$  to

$$\det(A_{\cdot,1} \ \dots \ A_{\cdot,k} \ \dots \ A_{\cdot,n})$$

is a linear map from the vector space of  $n$ -by-1 matrices with entries in  $\mathbf{F}$  to  $\mathbf{F}$ .

**Proof** The linearity follows easily from 10.33, where each term in the sum contains precisely one entry from the  $k^{\text{th}}$  column of  $A$ . ■

Now we are ready to prove one of the key properties about determinants of square matrices. This property will enable us to connect the determinant of an operator with the determinant of its matrix. Note that this proof is considerably more complicated than the proof of the corresponding result about the trace (see 10.14).

*The result below was first proved in 1812 by French mathematicians Jacques Binet and Augustin-Louis Cauchy.*

### 10.40 Determinant is multiplicative

Suppose  $A$  and  $B$  are square matrices of the same size. Then

$$\det(AB) = \det(BA) = (\det A)(\det B).$$

**Proof** Write  $A = (A_{\cdot,1} \ \dots \ A_{\cdot,n})$ , where each  $A_{\cdot,k}$  is an  $n$ -by-1 column of  $A$ . Also write

$$B = \begin{pmatrix} B_{1,1} & \dots & B_{1,n} \\ \vdots & & \vdots \\ B_{n,1} & \dots & B_{n,n} \end{pmatrix} = (B_{\cdot,1} \ \dots \ B_{\cdot,n}),$$

where each  $B_{\cdot,k}$  is an  $n$ -by-1 column of  $B$ . Let  $e_k$  denote the  $n$ -by-1 matrix that equals 1 in the  $k^{\text{th}}$  row and 0 elsewhere. Note that  $Ae_k = A_{\cdot,k}$  and  $Be_k = B_{\cdot,k}$ . Furthermore,  $B_{\cdot,k} = \sum_{m=1}^n B_{m,k}e_m$ .

First we will prove  $\det(AB) = (\det A)(\det B)$ . As we observed earlier (see 3.49), the definition of matrix multiplication easily implies that  $AB = (AB_{\cdot,1} \ \dots \ AB_{\cdot,n})$ . Thus

$$\begin{aligned}
\det(AB) &= \det(AB_{\cdot,1} \dots AB_{\cdot,n}) \\
&= \det(A(\sum_{m_1=1}^n B_{m_1,1}e_{m_1}) \dots A(\sum_{m_n=1}^n B_{m_n,n}e_{m_n})) \\
&= \det(\sum_{m_1=1}^n B_{m_1,1}Ae_{m_1} \dots \sum_{m_n=1}^n B_{m_n,n}Ae_{m_n}) \\
&= \sum_{m_1=1}^n \dots \sum_{m_n=1}^n B_{m_1,1} \dots B_{m_n,n} \det(Ae_{m_1} \dots Ae_{m_n}),
\end{aligned}$$

where the last equality comes from repeated applications of the linearity of  $\det$  as a function of one column at a time (10.39). In the last sum above, all terms in which  $m_j = m_k$  for some  $j \neq k$  can be ignored, because the determinant of a matrix with two equal columns is 0 (by 10.37). Thus instead of summing over all  $m_1, \dots, m_n$  with each  $m_j$  taking on values  $1, \dots, n$ , we can sum just over the permutations, where the  $m_j$ 's have distinct values. In other words,

$$\begin{aligned}
\det(AB) &= \sum_{(m_1, \dots, m_n) \in \text{perm } n} B_{m_1,1} \dots B_{m_n,n} \det(Ae_{m_1} \dots Ae_{m_n}) \\
&= \sum_{(m_1, \dots, m_n) \in \text{perm } n} B_{m_1,1} \dots B_{m_n,n} (\text{sign}(m_1, \dots, m_n)) \det A \\
&= (\det A) \sum_{(m_1, \dots, m_n) \in \text{perm } n} (\text{sign}(m_1, \dots, m_n)) B_{m_1,1} \dots B_{m_n,n} \\
&= (\det A)(\det B),
\end{aligned}$$

where the second equality comes from 10.38.

In the paragraph above, we proved that  $\det(AB) = (\det A)(\det B)$ . Interchanging the roles of  $A$  and  $B$ , we have  $\det(BA) = (\det B)(\det A)$ . The last equation can be rewritten as  $\det(BA) = (\det A)(\det B)$ , completing the proof. ■

*Note the similarity of the proof of the next result to the proof of the analogous result about the trace (see 10.15).*

Now we can prove that the determinant of the matrix of an operator is independent of the basis with respect to which the matrix is computed.

#### 10.41 Determinant of matrix of operator does not depend on basis

Let  $T \in \mathcal{L}(V)$ . Suppose  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  are bases of  $V$ . Then

$$\det \mathcal{M}(T, (u_1, \dots, u_n)) = \det \mathcal{M}(T, (v_1, \dots, v_n)).$$

**Proof** Let  $A = \mathcal{M}(I, (u_1, \dots, u_n), (v_1, \dots, v_n))$ . Then

$$\begin{aligned}\det \mathcal{M}(T, (u_1, \dots, u_n)) &= \det(A^{-1}(\mathcal{M}(T, (v_1, \dots, v_n))A)) \\ &= \det((\mathcal{M}(T, (v_1, \dots, v_n))A)A^{-1}) \\ &= \det \mathcal{M}(T, (v_1, \dots, v_n)),\end{aligned}$$

where the first equality follows from 10.7 and the second equality follows from 10.40. The third equality completes the proof. ■

The result below states that the determinant of an operator equals the determinant of the matrix of the operator. This theorem does not specify a basis because, by the result above, the determinant of the matrix of an operator is the same for every choice of basis.

#### 10.42 Determinant of an operator equals determinant of its matrix

Suppose  $T \in \mathcal{L}(V)$ . Then  $\det T = \det \mathcal{M}(T)$ .

**Proof** As noted above, 10.41 implies that  $\det \mathcal{M}(T)$  is independent of which basis of  $V$  we choose. Thus to show that  $\det T = \det \mathcal{M}(T)$  for every basis of  $V$ , we need only show that the result holds for some basis of  $V$ .

As we have already discussed, if  $V$  is a complex vector space, then choosing a basis of  $V$  as in 8.29 gives the desired result. If  $V$  is a real vector space, then applying the complex case to the complexification  $T_C$  (which is used to define  $\det T$ ) gives the desired result. ■

If we know the matrix of an operator on a complex vector space, the result above allows us to find the product of all the eigenvalues without finding any of the eigenvalues.

**10.43 Example** Suppose  $T$  is the operator on  $\mathbb{C}^5$  whose matrix is

$$\left( \begin{array}{ccccc} 0 & 0 & 0 & 0 & -3 \\ 1 & 0 & 0 & 0 & 6 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array} \right).$$

No one knows an exact formula for any of the eigenvalues of this operator. However, we do know that the product of the eigenvalues equals  $-3$ , because the determinant of the matrix above equals  $-3$ .

We can use 10.42 to give easy proofs of some useful properties about determinants of operators by shifting to the language of determinants of matrices, where certain properties have already been proved or are obvious. We carry out this procedure in the next result.

### 10.44 Determinant is multiplicative

Suppose  $S, T \in \mathcal{L}(V)$ . Then

$$\det(ST) = \det(TS) = (\det S)(\det T).$$

**Proof** Choose a basis of  $V$ . Then

$$\begin{aligned}\det(ST) &= \det \mathcal{M}(ST) \\ &= \det(\mathcal{M}(S)\mathcal{M}(T)) \\ &= (\det \mathcal{M}(S))(\det \mathcal{M}(T)) \\ &= (\det S)(\det T),\end{aligned}$$

where the first and last equalities come from 10.42 and the third equality comes from 10.40.

In the paragraph above, we proved that  $\det(ST) = (\det S)(\det T)$ . Interchanging the roles of  $S$  and  $T$ , we have  $\det(TS) = (\det T)(\det S)$ . Because multiplication of elements of  $\mathbf{F}$  is commutative, the last equation can be rewritten as  $\det(TS) = (\det S)(\det T)$ , completing the proof. ■

### The Sign of the Determinant

We proved the basic results of linear algebra before introducing determinants in this final chapter. Although determinants have value as a research tool in more advanced subjects, they play little role in basic linear algebra (when the subject is done right).

*Most applied mathematicians agree that determinants should rarely be used in serious numeric calculations.*

Determinants do have one important application in undergraduate mathematics, namely, in computing certain volumes and integrals. In this subsection we interpret the meaning of the sign of

the determinant on a real vector space. Then in the final subsection we will use the linear algebra we have learned to make clear the connection between determinants and these applications. Thus we will be dealing with a part of analysis that uses linear algebra.

We will begin with some purely linear algebra results that will also be useful when investigating volumes. Our setting will be inner product spaces. Recall that an isometry on an inner product space is an operator that preserves norms. The next result shows that every isometry has determinant with absolute value 1.

### 10.45 Isometries have determinant with absolute value 1

Suppose  $V$  is an inner product space and  $S \in \mathcal{L}(V)$  is an isometry. Then  $|\det S| = 1$ .

**Proof** First consider the case where  $V$  is a complex inner product space. Then all the eigenvalues of  $S$  have absolute value 1 (see the proof of 7.43). Thus the product of the eigenvalues of  $S$ , counting multiplicity, has absolute value one. In other words,  $|\det S| = 1$ , as desired.

Now suppose  $V$  is a real inner product space. We present two different proofs in this case.

Proof 1: With respect to the inner product on the complexification  $V_{\mathbb{C}}$  given by Exercise 3 in Section 9.B, it is easy to see that  $S_{\mathbb{C}}$  is an isometry on  $V_{\mathbb{C}}$ . Thus by the complex case that we have already done, we have  $|\det S_{\mathbb{C}}| = 1$ . By definition of the determinant on real vector spaces, we have  $\det S = \det S_{\mathbb{C}}$  and thus  $|\det S| = 1$ , completing the proof.

Proof 2: By 9.36, there is an orthonormal basis of  $V$  with respect to which  $\mathcal{M}(S)$  is a block diagonal matrix, where each block on the diagonal is a 1-by-1 matrix containing 1 or  $-1$  or a 2-by-2 matrix of the form

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix},$$

with  $\theta \in (0, \pi)$ . Note that the determinant of each 2-by-2 matrix of the form above equals 1 (because  $\cos^2 \theta + \sin^2 \theta = 1$ ). Thus the determinant of  $S$ , which is the product of the determinants of the blocks (see Exercise 6), is the product of 1's and  $-1$ 's. Hence,  $|\det S| = 1$ , as desired. ■

The Real Spectral Theorem 7.29 states that a self-adjoint operator  $T$  on a real inner product space has an orthonormal basis consisting of eigenvectors. With respect to such a basis, the number of times each eigenvalue appears on the diagonal of  $\mathcal{M}(T)$  is its multiplicity. Thus  $\det T$  equals the product of its eigenvalues, counting multiplicity (of course, this holds for every operator, self-adjoint or not, on a complex vector space).

Recall that if  $V$  is an inner product space and  $T \in \mathcal{L}(V)$ , then  $T^*T$  is a positive operator and hence has a unique positive square root, denoted  $\sqrt{T^*T}$  (see 7.35 and 7.36). Because  $\sqrt{T^*T}$  is positive, all its eigenvalues are non-negative (again, see 7.35), and hence  $\det \sqrt{T^*T} \geq 0$ . These considerations play a role in next example.

---

**10.46 Example** Suppose  $V$  is a real inner product space and  $T \in \mathcal{L}(V)$  is invertible (and thus  $\det T$  is either positive or negative). Attach a geometric meaning to the sign of  $\det T$ .

**Solution** First we consider an isometry  $S \in \mathcal{L}(V)$ . By 10.45, the determinant of  $S$  equals 1 or  $-1$ . Note that

$$\{v \in V : Sv = -v\}$$

*We are not formally defining the phrase “reverses direction” because these comments are meant only as an intuitive aid to our understanding.*

is the eigenspace  $E(-1, S)$ . Thinking geometrically, we could say that this is the subspace on which  $S$  reverses direction. An examination of proof 2 of 10.45 shows that  $\det S = 1$  if this subspace has even dimension and  $\det S = -1$  if this subspace has odd dimension.

Returning to our arbitrary invertible operator  $T \in \mathcal{L}(V)$ , by the Polar Decomposition (7.45) there is an isometry  $S \in \mathcal{L}(V)$  such that

$$T = S\sqrt{T^*T}.$$

Now 10.44 tells us that

$$\det T = (\det S)(\det \sqrt{T^*T}).$$

The remarks just before this example pointed out that  $\det \sqrt{T^*T} \geq 0$ . Thus whether  $\det T$  is positive or negative depends on whether  $\det S$  is positive or negative. As we saw in the paragraph above, this depends on whether the subspace on which  $S$  reverses direction has even or odd dimension.

Because  $T$  is the product of  $S$  and an operator that never reverses direction (namely,  $\sqrt{T^*T}$ ), we can reasonably say that whether  $\det T$  is positive or negative depends on whether  $T$  reverses vectors an even or an odd number of times.

---

## Volume

The next result will be a key tool in our investigation of volume. Recall that our remarks before Example 10.46 pointed out that  $\det \sqrt{T^*T} \geq 0$ .

$$10.47 \quad |\det T| = \det \sqrt{T^*T}$$

Suppose  $V$  is an inner product space and  $T \in \mathcal{L}(V)$ . Then

$$|\det T| = \det \sqrt{T^*T}.$$

### Proof

By the Polar Decomposition (7.45), there is an isometry  $S \in \mathcal{L}(V)$  such that

*Another proof of this result is suggested in Exercise 8.*

$$T = S\sqrt{T^*T}.$$

Thus

$$\begin{aligned} |\det T| &= |\det S| \det \sqrt{T^*T} \\ &= \det \sqrt{T^*T}, \end{aligned}$$

where the first equality follows from 10.44 and the second equality follows from 10.45. ■

Now we turn to the question of volume in  $\mathbf{R}^n$ . Fix a positive integer  $n$  for the rest of this subsection. We will consider only the real inner product space  $\mathbf{R}^n$ , with its standard inner product.

We would like to assign to each subset  $\Omega$  of  $\mathbf{R}^n$  its  $n$ -dimensional volume (when  $n = 2$ , this is usually called area instead of volume). We begin with boxes, where we have a good intuitive notion of volume.

### 10.48 Definition *box*

A **box** in  $\mathbf{R}^n$  is a set of the form

$$\{(y_1, \dots, y_n) \in \mathbf{R}^n : x_j < y_j < x_j + r_j \text{ for } j = 1, \dots, n\},$$

where  $r_1, \dots, r_n$  are positive numbers and  $(x_1, \dots, x_n) \in \mathbf{R}^n$ . The numbers  $r_1, \dots, r_n$  are called the **side lengths** of the box.

You should verify that when  $n = 2$ , a box is a rectangle with sides parallel to the coordinate axes, and that when  $n = 3$ , a box is a familiar 3-dimensional box with sides parallel to the coordinate axes.

The next definition fits with our intuitive notion of volume, because we define the volume of a box to be the product of the side lengths of the box.

### 10.49 Definition *volume of a box*

The **volume** of a box  $B$  in  $\mathbf{R}^n$  with side lengths  $r_1, \dots, r_n$  is defined to be  $r_1 \cdots r_n$  and is denoted by  $\text{volume } B$ .

*Readers familiar with outer measure will recognize that concept here.*

To define the volume of an arbitrary set  $\Omega \subset \mathbf{R}^n$ , the idea is to write  $\Omega$  as a subset of a union of many small boxes, then add up the volumes of these small

boxes. As we approximate  $\Omega$  more accurately by unions of small boxes, we get a better estimate of volume  $\Omega$ .

### 10.50 Definition *volume*

Suppose  $\Omega \subset \mathbf{R}^n$ . Then the **volume** of  $\Omega$ , denoted  $\text{volume } \Omega$ , is defined to be the infimum of

$$\text{volume } B_1 + \text{volume } B_2 + \cdots,$$

where the infimum is taken over all sequences  $B_1, B_2, \dots$  of boxes in  $\mathbf{R}^n$  whose union contains  $\Omega$ .

We will work only with an intuitive notion of volume. Our purpose in this book is to understand linear algebra, whereas notions of volume belong to analysis (although volume is intimately connected with determinants, as we will soon see). Thus for the rest of this section we will rely on intuitive notions of volume rather than on a rigorous development, although we shall maintain our usual rigor in the linear algebra parts of what follows. Everything said here about volume will be correct if appropriately interpreted—the intuitive approach used here can be converted into appropriate correct definitions, correct statements, and correct proofs using the machinery of analysis.

### 10.51 Notation $T(\Omega)$

For  $T$  a function defined on a set  $\Omega$ , define  $T(\Omega)$  by

$$T(\Omega) = \{Tx : x \in \Omega\}.$$

For  $T \in \mathcal{L}(\mathbf{R}^n)$  and  $\Omega \subset \mathbf{R}^n$ , we seek a formula for volume  $T(\Omega)$  in terms of  $T$  and volume  $\Omega$ . We begin by looking at positive operators.

### 10.52 Positive operators change volume by factor of determinant

Suppose  $T \in \mathcal{L}(\mathbf{R}^n)$  is a positive operator and  $\Omega \subset \mathbf{R}^n$ . Then

$$\text{volume } T(\Omega) = (\det T)(\text{volume } \Omega).$$

**Proof** To get a feeling for why this result is true, first consider the special case where  $\lambda_1, \dots, \lambda_n$  are positive numbers and  $T \in \mathcal{L}(\mathbf{R}^n)$  is defined by

$$T(x_1, \dots, x_n) = (\lambda_1 x_1, \dots, \lambda_n x_n).$$

This operator stretches the  $j^{\text{th}}$  standard basis vector by a factor of  $\lambda_j$ . If  $B$  is a box in  $\mathbf{R}^n$  with side lengths  $r_1, \dots, r_n$ , then  $T(B)$  is a box in  $\mathbf{R}^n$  with side lengths  $\lambda_1 r_1, \dots, \lambda_n r_n$ . The box  $T(B)$  thus has volume  $\lambda_1 \cdots \lambda_n r_1 \cdots r_n$ , whereas the box  $\Omega$  has volume  $r_1 \cdots r_n$ . Note that  $\det T = \lambda_1 \cdots \lambda_n$ . Thus

$$\text{volume } T(B) = (\det T)(\text{volume } B)$$

for every box  $B$  in  $\mathbf{R}^n$ . Because the volume of  $\Omega$  is approximated by sums of volumes of boxes, this implies that  $\text{volume } T(\Omega) = (\det T)(\text{volume } \Omega)$ .

Now consider an arbitrary positive operator  $T \in \mathcal{L}(\mathbf{R}^n)$ . By the Real Spectral Theorem (7.29), there exist an orthonormal basis  $e_1, \dots, e_n$  of  $\mathbf{R}^n$  and nonnegative numbers  $\lambda_1, \dots, \lambda_n$  such that  $Te_j = \lambda_j e_j$  for  $j = 1, \dots, n$ . In the special case where  $e_1, \dots, e_n$  is the standard basis of  $\mathbf{R}^n$ , this operator is the same one as defined in the paragraph above. For an arbitrary orthonormal basis  $e_1, \dots, e_n$ , this operator has the same behavior as the one in the paragraph above—it stretches the  $j^{\text{th}}$  basis vector in an orthonormal basis by a factor of  $\lambda_j$ . Your intuition about volume should convince you that volume behaves the same with respect to each orthonormal basis. That intuition, and the special case of the paragraph above, should convince you that  $T$  multiplies volume by a factor of  $\lambda_1 \cdots \lambda_n$ , which again equals  $\det T$ . ■

Our next tool is the following result, which states that isometries do not change volume.

### 10.53 An isometry does not change volume

Suppose  $S \in \mathcal{L}(\mathbf{R}^n)$  is an isometry and  $\Omega \subset \mathbf{R}^n$ . Then

$$\text{volume } S(\Omega) = \text{volume } \Omega.$$

**Proof** For  $x, y \in \mathbf{R}^n$ , we have

$$\begin{aligned}\|Sx - Sy\| &= \|S(x - y)\| \\ &= \|x - y\|.\end{aligned}$$

In other words,  $S$  does not change the distance between points. That property alone may be enough to convince you that  $S$  does not change volume.

However, if you need stronger persuasion, consider the complete description of isometries on real inner product spaces provided by 9.36. According to 9.36,  $S$  can be decomposed into pieces, each of which is the identity on some subspace (which clearly does not change volume) or multiplication by  $-1$  on some subspace (which again clearly does not change volume) or a rotation on a 2-dimensional subspace (which again does not change volume). Or use 9.36 in conjunction with Exercise 7 in Section 9.B to write  $S$  as a product of operators, each of which does not change volume. Either way, you should be convinced that  $S$  does not change volume. ■

Now we can prove that an operator  $T \in \mathcal{L}(\mathbf{R}^n)$  changes volume by a factor of  $|\det T|$ . Note the huge importance of the Polar Decomposition in the proof.

#### 10.54 $T$ changes volume by factor of $|\det T|$

Suppose  $T \in \mathcal{L}(\mathbf{R}^n)$  and  $\Omega \subset \mathbf{R}^n$ . Then

$$\text{volume } T(\Omega) = |\det T|(\text{volume } \Omega).$$

**Proof** By the Polar Decomposition (7.45), there is an isometry  $S \in \mathcal{L}(V)$  such that

$$T = S\sqrt{T^*T}.$$

If  $\Omega \subset \mathbf{R}^n$ , then  $T(\Omega) = S(\sqrt{T^*T}(\Omega))$ . Thus

$$\begin{aligned}\text{volume } T(\Omega) &= \text{volume } S(\sqrt{T^*T}(\Omega)) \\ &= \text{volume } \sqrt{T^*T}(\Omega) \\ &= (\det \sqrt{T^*T})(\text{volume } \Omega) \\ &= |\det T|(\text{volume } \Omega),\end{aligned}$$

where the second equality holds because volume is not changed by the isometry  $S$  (by 10.53), the third equality holds by 10.52 (applied to the positive operator  $\sqrt{T^*T}$ ), and the fourth equality holds by 10.47. ■

The result that we just proved leads to the appearance of determinants in the formula for change of variables in multivariable integration. To describe this, we will again be vague and intuitive.

Throughout this book, almost all the functions we have encountered have been linear. Thus please be aware that the functions  $f$  and  $\sigma$  in the material below are not assumed to be linear.

The next definition aims at conveying the idea of the integral; it is not intended as a rigorous definition.

### 10.55 Definition *integral*, $\int_{\Omega} f$

If  $\Omega \subset \mathbf{R}^n$  and  $f$  is a real-valued function on  $\Omega$ , then the *integral* of  $f$  over  $\Omega$ , denoted  $\int_{\Omega} f$  or  $\int_{\Omega} f(x) dx$ , is defined by breaking  $\Omega$  into pieces small enough that  $f$  is almost constant on each piece. On each piece, multiply the (almost constant) value of  $f$  by the volume of the piece, then add up these numbers for all the pieces, getting an approximation to the integral that becomes more accurate as  $\Omega$  is divided into finer pieces.

Actually,  $\Omega$  in the definition above needs to be a reasonable set (for example, open or measurable) and  $f$  needs to be a reasonable function (for example, continuous or measurable), but we will not worry about those technicalities. Also, notice that the  $x$  in  $\int_{\Omega} f(x) dx$  is a dummy variable and could be replaced with any other symbol.

Now we define the notions of differentiable and derivative. Notice that in this context, the derivative is an operator, not a number as in one-variable calculus. The uniqueness of  $T$  in the definition below is left as Exercise 9.

### 10.56 Definition *differentiable, derivative*, $\sigma'(x)$

Suppose  $\Omega$  is an open subset of  $\mathbf{R}^n$  and  $\sigma$  is a function from  $\Omega$  to  $\mathbf{R}^n$ . For  $x \in \Omega$ , the function  $\sigma$  is called *differentiable* at  $x$  if there exists an operator  $T \in \mathcal{L}(\mathbf{R}^n)$  such that

$$\lim_{y \rightarrow 0} \frac{\|\sigma(x + y) - \sigma(x) - Ty\|}{\|y\|} = 0.$$

If  $\sigma$  is differentiable at  $x$ , then the unique operator  $T \in \mathcal{L}(\mathbf{R}^n)$  satisfying the equation above is called the *derivative* of  $\sigma$  at  $x$  and is denoted by  $\sigma'(x)$ .

If  $n = 1$ , then the derivative in the sense of the definition above is the operator on  $\mathbf{R}$  of multiplication by the derivative in the usual sense of one-variable calculus.

The idea of the derivative is that for  $x$  fixed and  $\|y\|$  small,

$$\sigma(x + y) \approx \sigma(x) + (\sigma'(x))(y);$$

because  $\sigma'(x) \in \mathcal{L}(\mathbf{R}^n)$ , this makes sense.

Suppose  $\Omega$  is an open subset of  $\mathbf{R}^n$  and  $\sigma$  is a function from  $\Omega$  to  $\mathbf{R}^n$ . We can write

$$\sigma(x) = (\sigma_1(x), \dots, \sigma_n(x)),$$

where each  $\sigma_j$  is a function from  $\Omega$  to  $\mathbf{R}$ . The partial derivative of  $\sigma_j$  with respect to the  $k^{\text{th}}$  coordinate is denoted  $D_k \sigma_j$ . Evaluating this partial derivative at a point  $x \in \Omega$  gives  $D_k \sigma_j(x)$ . If  $\sigma$  is differentiable at  $x$ , then the matrix of  $\sigma'(x)$  with respect to the standard basis of  $\mathbf{R}^n$  contains  $D_k \sigma_j(x)$  in row  $j$ , column  $k$  (this is left as an exercise). In other words,

$$10.57 \quad \mathcal{M}(\sigma'(x)) = \begin{pmatrix} D_1 \sigma_1(x) & \dots & D_n \sigma_1(x) \\ \vdots & & \vdots \\ D_1 \sigma_n(x) & \dots & D_n \sigma_n(x) \end{pmatrix}.$$

Now we can state the change of variables integration formula. Some additional mild hypotheses are needed for  $f$  and  $\sigma'$  (such as continuity or measurability), but we will not worry about them because the proof below is really a pseudoproof that is intended to convey the reason the result is true.

The result below is called a change of variables formula because you can think of  $y = \sigma(x)$  as a change of variables, as illustrated by the two examples that follow the proof.

### 10.58 Change of variables in an integral

Suppose  $\Omega$  is an open subset of  $\mathbf{R}^n$  and  $\sigma : \Omega \rightarrow \mathbf{R}^n$  is differentiable at every point of  $\Omega$ . If  $f$  is a real-valued function defined on  $\sigma(\Omega)$ , then

$$\int_{\sigma(\Omega)} f(y) dy = \int_{\Omega} f(\sigma(x)) |\det \sigma'(x)| dx.$$

**Proof** Let  $x \in \Omega$  and let  $\Gamma$  be a small subset of  $\Omega$  containing  $x$  such that  $f$  is approximately equal to the constant  $f(\sigma(x))$  on the set  $\sigma(\Gamma)$ .

Adding a fixed vector [such as  $\sigma(x)$ ] to each vector in a set produces another set with the same volume. Thus our approximation for  $\sigma$  near  $x$  using the derivative shows that

$$\text{volume } \sigma(\Gamma) \approx \text{volume}[(\sigma'(x))(\Gamma)].$$

Using 10.54 applied to the operator  $\sigma'(x)$ , this becomes

$$\text{volume } \sigma(\Gamma) \approx |\det \sigma'(x)|(\text{volume } \Gamma).$$

Let  $y = \sigma(x)$ . Multiply the left side of the equation above by  $f(y)$  and the right side by  $f(\sigma(x))$  [because  $y = \sigma(x)$ , these two quantities are equal], getting

$$f(y) \text{volume } \sigma(\Gamma) \approx f(\sigma(x))|\det \sigma'(x)|(\text{volume } \Gamma).$$

Now break  $\Omega$  into many small pieces and add the corresponding versions of the equation above, getting the desired result. ■

The key point when making a change of variables is that the factor of  $|\det \sigma'(x)|$  must be included when making a substitution  $y = f(x)$ , as in the right side of 10.58. We finish up by illustrating this point with two important examples.

### 10.59 Example polar coordinates

Define  $\sigma: \mathbf{R}^2 \rightarrow \mathbf{R}^2$  by

$$\sigma(r, \theta) = (r \cos \theta, r \sin \theta),$$

where we have used  $r, \theta$  as the coordinates instead of  $x_1, x_2$  for reasons that will be obvious to everyone familiar with polar coordinates (and will be a mystery to everyone else). For this choice of  $\sigma$ , the matrix of partial derivatives corresponding to 10.57 is

$$\begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix},$$

as you should verify. The determinant of the matrix above equals  $r$ , thus explaining why a factor of  $r$  is needed when computing an integral in polar coordinates.

For example, note the extra factor of  $r$  in the following familiar formula involving integrating a function  $f$  over a disk in  $\mathbf{R}^2$ :

$$\int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} f(x, y) dy dx = \int_0^{2\pi} \int_0^1 f(r \cos \theta, r \sin \theta) r dr d\theta.$$


---

---

**10.60 Example spherical coordinates**

Define  $\sigma: \mathbf{R}^3 \rightarrow \mathbf{R}^3$  by

$$\sigma(\rho, \varphi, \theta) = (\rho \sin \varphi \cos \theta, \rho \sin \varphi \sin \theta, \rho \cos \varphi),$$

where we have used  $\rho, \theta, \varphi$  as the coordinates instead of  $x_1, x_2, x_3$  for reasons that will be obvious to everyone familiar with spherical coordinates (and will be a mystery to everyone else). For this choice of  $\sigma$ , the matrix of partial derivatives corresponding to 10.57 is

$$\begin{pmatrix} \sin \varphi \cos \theta & \rho \cos \varphi \cos \theta & -\rho \sin \varphi \sin \theta \\ \sin \varphi \sin \theta & \rho \cos \varphi \sin \theta & \rho \sin \varphi \cos \theta \\ \cos \varphi & -\rho \sin \varphi & 0 \end{pmatrix},$$

as you should verify. The determinant of the matrix above equals  $\rho^2 \sin \varphi$ , thus explaining why a factor of  $\rho^2 \sin \varphi$  is needed when computing an integral in spherical coordinates.

For example, note the extra factor of  $\rho^2 \sin \varphi$  in the following familiar formula involving integrating a function  $f$  over a ball in  $\mathbf{R}^3$ :

$$\begin{aligned} & \int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} f(x, y, z) dz dy dx \\ &= \int_0^{2\pi} \int_0^\pi \int_0^1 f(\rho \sin \varphi \cos \theta, \rho \sin \varphi \sin \theta, \rho \cos \varphi) \rho^2 \sin \varphi d\rho d\varphi d\theta. \end{aligned}$$


---

## EXERCISES 10.B

---

- 1 Suppose  $V$  is a real vector space. Suppose  $T \in \mathcal{L}(V)$  has no eigenvalues. Prove that  $\det T > 0$ .
- 2 Suppose  $V$  is a real vector space with even dimension and  $T \in \mathcal{L}(V)$ . Suppose  $\det T < 0$ . Prove that  $T$  has at least two distinct eigenvalues.
- 3 Suppose  $T \in \mathcal{L}(V)$  and  $n = \dim V > 2$ . Let  $\lambda_1, \dots, \lambda_n$  denote the eigenvalues of  $T$  (or of  $T_C$  if  $V$  is a real vector space), repeated according to multiplicity.
  - (a) Find a formula for the coefficient of  $z^{n-2}$  in the characteristic polynomial of  $T$  in terms of  $\lambda_1, \dots, \lambda_n$ .
  - (b) Find a formula for the coefficient of  $z$  in the characteristic polynomial of  $T$  in terms of  $\lambda_1, \dots, \lambda_n$ .

- 4 Suppose  $T \in \mathcal{L}(V)$  and  $c \in \mathbf{F}$ . Prove that  $\det(cT) = c^{\dim V} \det T$ .
- 5 Prove or give a counterexample: if  $S, T \in \mathcal{L}(V)$ , then  $\det(S + T) = \det S + \det T$ .
- 6 Suppose  $A$  is a block upper-triangular matrix

$$A = \begin{pmatrix} A_1 & & * \\ & \ddots & \\ 0 & & A_m \end{pmatrix},$$

where each  $A_j$  along the diagonal is a square matrix. Prove that

$$\det A = (\det A_1) \cdots (\det A_m).$$

- 7 Suppose  $A$  is an  $n$ -by- $n$  matrix with real entries. Let  $S \in \mathcal{L}(\mathbf{C}^n)$  denote the operator on  $\mathbf{C}^n$  whose matrix equals  $A$ , and let  $T \in \mathcal{L}(\mathbf{R}^n)$  denote the operator on  $\mathbf{R}^n$  whose matrix equals  $A$ . Prove that  $\text{trace } S = \text{trace } T$  and  $\det S = \det T$ .
- 8 Suppose  $V$  is an inner product space and  $T \in \mathcal{L}(V)$ . Prove that

$$\det T^* = \overline{\det T}.$$

Use this to prove that  $|\det T| = \det \sqrt{T^* T}$ , giving a different proof than was given in 10.47.

- 9 Suppose  $\Omega$  is an open subset of  $\mathbf{R}^n$  and  $\sigma$  is a function from  $\Omega$  to  $\mathbf{R}^n$ . Suppose  $x \in \Omega$  and  $\sigma$  is differentiable at  $x$ . Prove that the operator  $T \in \mathcal{L}(\mathbf{R}^n)$  satisfying the equation in 10.56 is unique.  
[This exercise shows that the notation  $\sigma'(x)$  is justified.]
- 10 Suppose  $T \in \mathcal{L}(\mathbf{R}^n)$  and  $x \in \mathbf{R}^n$ . Prove that  $T$  is differentiable at  $x$  and  $T'(x) = T$ .
- 11 Find a suitable hypothesis on  $\sigma$  and then prove 10.57.
- 12 Let  $a, b, c$  be positive numbers. Find the volume of the ellipsoid

$$\left\{ (x, y, z) \in \mathbf{R}^3 : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} < 1 \right\}$$

by finding a set  $\Omega \subset \mathbf{R}^3$  whose volume you know and an operator  $T \in \mathcal{L}(\mathbf{R}^3)$  such that  $T(\Omega)$  equals the ellipsoid above.

# *Photo Credits*

---

- page 1: Pierre Louis Dumesnil; 1884 copy by Nils Forsberg/Public domain image from *Wikimedia*.
- page 27: George M. Bergman/Archives of the Mathematisches Forschungsinstitut Oberwolfach.
- page 51: Gottlieb Biermann; photo by A. Wittmann/Public domain image from *Wikimedia*.
- page 117: Mostafa Azizi/Public domain image from *Wikimedia*.
- page 131: Hans-Peter Postel/Public domain image from *Wikimedia*.
- page 163: Public domain image from *Wikimedia*.
- page 203: Public domain image from *Wikimedia*. Original painting is in Tate Britain.
- page 224: Spiked Math.
- page 241: Public domain image from *Wikimedia*.
- page 275: Public domain image from *Wikimedia*. Original fresco is in the Vatican.
- page 295: Public domain image from *Wikimedia*.

# Symbol Index

$A^{-1}$ , 296	$\mathcal{P}(\mathbf{F})$ , 30
$A_{j,\cdot}$ , 76	$\pi$ , 97
$A_{j,k}$ , 70	$\mathcal{P}_m(\mathbf{F})$ , 31
$A_{\cdot,k}$ , 76	$p(T)$ , 143
$A^t$ , 109	$P_U$ , 195
<b>C</b> , 2	<b>R</b> , 2
$\deg$ , 31	$\text{Re}$ , 118
$\Delta$ , 179	$\sigma'$ , 327
$\det$ , 307, 314	$\subsetneq$ , 243
$\dim$ , 44	$\tilde{T}$ , 97
$\oplus$ , 21	$\sqrt{T}$ , 233
$D_k$ , 328	$T'$ , 103
$E(\lambda, T)$ , 155	$T^*$ , 204
<b>F</b> , 4	$T^{-1}$ , 80
$\mathbf{F}^\infty$ , 13	$T(\Omega)$ , 324
$\mathbf{F}^{m,n}$ , 73	$T_C$ , 277
$\mathbf{F}^n$ , 6	$T^m$ , 143
$\mathbf{F}^S$ , 14	$T _U$ , 132, 137
$G(\lambda, T)$ , 245	$T/U$ , 137
$I$ , 52, 296	$U^\perp$ , 193
$\iff$ , 207	$U^0$ , 104
$\text{Im}$ , 118	$\langle u, v \rangle$ , 166
$-\infty$ , 31	$V$ , 16
$\int_\Omega f$ , 327	$\ v\ $ , 168
$\mathcal{L}(V)$ , 86	$V'$ , 101
$\mathcal{L}(V, W)$ , 52	$V/U$ , 95
$\mathcal{M}(T)$ , 70, 146	$-v$ , 15
$\mathcal{M}(v)$ , 84	$V_C$ , 276
$\text{perm}$ , 311	$v + U$ , 94
	$\bar{z}$ , 118
	$ z $ , 118

# *Index*

- absolute value, 118  
addition  
    in quotient space, 96  
    of complex numbers, 2  
    of functions, 14  
    of linear maps, 55  
    of matrices, 72  
    of subspaces, 20  
    of vectors, 12  
    of vectors in  $\mathbf{F}^n$ , 7  
additive inverse  
    in  $\mathbf{C}$ , 3, 4  
    in  $\mathbf{F}^n$ , 9  
    in vector space, 12, 15  
additivity, 52  
adjoint of a linear map, 204  
affine subset, 94  
algebraic multiplicity, 255  
annihilator of a subspace, 104  
Apollonius's Identity, 179  
associativity, 3, 12, 56  
  
backward shift, 53, 59, 81, 86, 140  
basis, 39  
    of eigenvectors, 157, 218,  
        221, 224, 268  
    of generalized eigenvectors,  
        254  
Binet, Jacques, 317  
Blake, William, 203  
block diagonal matrix, 255  
box in  $\mathbf{R}^n$ , 323  
Cauchy, Augustin-Louis, 171, 317  
Cauchy–Schwarz Inequality, 172  
Cayley, Arthur, 262  
Cayley–Hamilton Theorem  
    on complex vector space, 261  
    on real vector space, 284  
change of basis, 298  
change of variables in integral,  
    328  
characteristic polynomial  
    on complex vector space, 261  
    on real vector space, 283  
characteristic value, 134  
Christina, Queen of Sweden, 1  
closed under addition, 18  
closed under scalar multiplication,  
    18  
column rank of a matrix, 111  
commutativity, 3, 7, 12, 25, 56, 75,  
    79, 144, 212  
complex conjugate, 118  
complex number, 2  
Complex Spectral Theorem, 218  
complex vector space, 13  
complexification  
    of a vector space, 276  
    of an operator, 277  
conjugate symmetry, 166

- conjugate transpose of a matrix, 207  
 coordinate, 6  
 cube root of an operator, 224  
 cubic formula, 124  
 degree of a polynomial, 31  
 derivative, 327  
 Descartes, René, 1  
 determinant  
     of a matrix, 314  
     of an operator, 307  
 diagonal matrix, 155  
 diagonal of a square matrix, 147  
 diagonalizable, 156  
 differentiable, 327  
 differentiation linear map, 53, 56,  
     59, 61, 62, 69, 72, 78,  
     144, 190, 248, 294  
 dimension, 44  
     of a sum of subspaces, 47  
 direct sum, 21, 42, 93  
     of a subspace and its  
         orthogonal complement,  
         194  
     of null  $T^n$  and range  $T^n$ , 243  
 distributive property, 3, 12, 16, 56,  
     79  
 Division Algorithm for  
     Polynomials, 121  
 division of complex numbers, 4  
 dot product, 164  
 double dual space, 116  
 dual  
     of a basis, 102  
     of a linear map, 103  
     of a vector space, 101  
 eigenspace, 155  
 eigenvalue of an operator, 134  
 eigenvector, 134  
 Euclid, 275  
 Euclidean inner product, 166  
 factor of a polynomial, 122  
 Fibonacci, 131  
 Fibonacci sequence, 161  
 field, 10  
 finite-dimensional vector space, 30  
*Flatland*, 6  
 Fundamental Theorem of Algebra,  
     124  
 Fundamental Theorem of Linear  
     Maps, 63  
 Gauss, Carl Friedrich, 51  
 generalized eigenspace, 245  
 generalized eigenvector, 245  
 geometric multiplicity, 255  
 Gram, Jørgen, 182  
 Gram–Schmidt Procedure, 182  
 graph of a linear map, 98  
 Halmos, Paul, 27  
 Hamilton, William, 262  
 harmonic function, 179  
 Hermitian, 209  
 homogeneity, 52  
 homogeneous system of linear  
     equations, 65, 90  
 Hypatia, 241  
 identity map, 52, 56  
 identity matrix, 296  
 image, 62  
 imaginary part, 118  
 infinite-dimensional vector space,  
     31  
 inhomogeneous system of linear  
     equations, 66, 90  
 injective, 60  
 inner product, 166

- inner product space, 167
- integral, 327
- invariant subspace, 132
- inverse
  - of a linear map, 80
  - of a matrix, 296
- invertible linear map, 80
- invertible matrix, 296
- isometry, 228, 292, 321
- isomorphic vector spaces, 82
- isomorphism, 82
- Jordan basis, 273
- Jordan Form, 273
- Jordan, Camille, 272
- kernel, 59
- Khayyám, Omar, 117
- Laplacian, 179
- length of list, 5
- Leonardo of Pisa, 131
- linear combination, 28
- Linear Dependence Lemma, 34
- linear functional, 101, 187
- linear map, 52
- linear span, 29
- linear subspace, 18
- linear transformation, 52
- linearly dependent, 33
- linearly independent, 32
- list, 5
  - of vectors, 28
- Lovelace, Ada, 295
- matrix, 70
  - multiplication, 75
  - of linear map, 70
  - of nilpotent operator, 249
  - of operator, 146
  - of product of linear maps, 75,
- 297
- of  $T'$ , 110
- of  $T^*$ , 208
- of vector, 84
- minimal polynomial, 263, 279
- minimizing distance, 198
- monic polynomial, 262
- multiplication, *see* product
- multiplicity of an eigenvalue, 254
- Newton, Isaac, 203
- nilpotent operator, 248, 271
- nonsingular matrix, 296
- norm, 164, 168
- normal operator, 212, 287
- null space, 59
  - of powers of an operator, 242
  - of  $T'$ , 106
  - of  $T^*$ , 207
- one-to-one, 60
- onto, 62
- operator, 86
- orthogonal
  - complement, 193
  - operator, 229
  - projection, 195
  - vectors, 169
- orthonormal
  - basis, 181
  - list, 180
- parallel affine subsets, 94
- Parallelogram Equality, 174
- permutation, 311
- photo credits, 333
- point, 13
- polar coordinates, 329
- Polar Decomposition, 233
- polynomial, 30
- positive operator, 225
- positive semidefinite operator, 227

- product
  - of complex numbers, 2
  - of linear maps, 55
  - of matrices, 75
  - of polynomials, 144
  - of scalar and linear map, 55
  - of scalar and vector, 12
  - of scalar and vector in  $\mathbf{F}^n$ , 10
  - of vector spaces, 91
- Pythagorean Theorem, 170
- quotient
  - map, 97
  - operator, 137
  - space, 95
- range, 61
  - of powers of an operator, 251
  - of  $T'$ , 107
  - of  $T^*$ , 207
- rank of a matrix, 112
- Raphael, 275
- real part, 118
- Real Spectral Theorem, 221
- real vector space, 13
- restriction operator, 137
- Riesz Representation Theorem, 188
- Riesz, Frigyes, 187
- row rank of a matrix, 111
- scalar, 4
- scalar multiplication, 10, 12
  - in quotient space, 96
  - of linear maps, 55
  - of matrices, 73
- Schmidt, Erhard, 182
- School of Athens*, 275
- Schur's Theorem, 186
- Schur, Issai, 186
- Schwarz, Hermann, 171
- self-adjoint operator, 209
- sign of a permutation, 312
- signum, 313
- singular matrix, 296
- Singular Value Decomposition, 237
- singular values, 236
- span, 29
- spans, 30
- Spectral Theorem, 218, 221
- spherical coordinates, 330
- square root of an operator, 223, 225, 233, 259
- standard basis, 39
- subspace, 18
- subtraction of complex numbers, 4
- sum, *see* addition
- Supreme Court, 174
- surjective, 62
- trace
  - of a matrix, 300
  - of an operator, 299
- transpose of a matrix, 109, 207
- Triangle Inequality, 173
- tuple, 5
- unitary operator, 229
- upper-triangular matrix, 147, 256
- vector, 8, 13
- vector space, 12
- volume, 324
- zero of a polynomial, 122

**A B S T R A C T**

---

**A L G E B R A**

---

---

THIRD EDITION

---

---

**DAVID S. DUMMIT**

---

---

**RICHARD M. FOOTE**

*Dedicated to our families  
especially  
Janice, Evan, and Krysta  
and  
Zsuzsanna, Peter, Karoline, and Alexandra*

## Frequently Used Notation

$f^{-1}(A)$	the inverse image or preimage of $A$ under $f$
$a \mid b$	$a$ divides $b$
$(a, b)$	the greatest common divisor of $a, b$ also the ideal generated by $a, b$
$ A ,  x $	the order of the set $A$ , the order of the element $x$
$\mathbb{Z}, \mathbb{Z}^+$	the integers, the positive integers
$\mathbb{Q}, \mathbb{Q}^+$	the rational numbers, the positive rational numbers
$\mathbb{R}, \mathbb{R}^+$	the real numbers, the positive real numbers
$\mathbb{C}, \mathbb{C}^\times$	the complex numbers, the nonzero complex numbers
$\mathbb{Z}/n\mathbb{Z}$	the integers modulo $n$
$(\mathbb{Z}/n\mathbb{Z})^\times$	the (multiplicative group of) invertible integers modulo $n$
$A \times B$	the direct or Cartesian product of $A$ and $B$
$H \leq G$	$H$ is a subgroup of $G$
$\mathbb{Z}_n$	the cyclic group of order $n$
$D_{2n}$	the dihedral group of order $2n$
$S_n, S_\Omega$	the symmetric group on $n$ letters, and on the set $\Omega$
$A_n$	the alternating group on $n$ letters
$Q_8$	the quaternion group of order 8
$V_4$	the Klein 4-group
$\mathbb{F}_N$	the finite field of $N$ elements
$GL_n(F), GL(V)$	the general linear groups
$SL_n(F)$	the special linear group
$A \cong B$	$A$ is isomorphic to $B$
$C_G(A), N_G(A)$	the centralizer, and normalizer in $G$ of $A$
$Z(G)$	the center of the group $G$
$G_s$	the stabilizer in the group $G$ of $s$
$\langle A \rangle, \langle x \rangle$	the group generated by the set $A$ , and by the element $x$
$G = \langle \dots   \dots \rangle$	generators and relations (a presentation) for $G$
$\ker \varphi, \text{im } \varphi$	the kernel, and the image of the homomorphism $\varphi$
$N \trianglelefteq G$	$N$ is a normal subgroup of $G$
$gH, Hg$	the left coset, and right coset of $H$ with coset representative $g$
$ G : H $	the index of the subgroup $H$ in the group $G$
$\text{Aut}(G)$	the automorphism group of the group $G$
$Syl_p(G)$	the set of Sylow $p$ -subgroups of $G$
$n_p$	the number of Sylow $p$ -subgroups of $G$
$[x, y]$	the commutator of $x, y$
$H \rtimes K$	the semidirect product of $H$ and $K$
$\mathbb{H}$	the real Hamilton Quaternions
$R^\times$	the multiplicative group of units of the ring $R$
$R[x], R[x_1, \dots, x_n]$	polynomials in $x$ , and in $x_1, \dots, x_n$ with coefficients in $R$
$RG, FG$	the group ring of the group $G$ over the ring $R$ , and over the field $F$
$\mathcal{O}_K$	the ring of integers in the number field $K$
$\varprojlim A_i, \varinjlim A_i$	the direct, and the inverse limit of the family of groups $A_i$
$\mathbb{Z}_p, \mathbb{Q}_p$	the $p$ -adic integers, and the $p$ -adic rationals
$A \oplus B$	the direct sum of $A$ and $B$

$LT(f)$ , $LT(I)$	the leading term of the polynomial $f$ , the ideal of leading terms
$M_n(R)$ , $M_{n \times m}(R)$	the $n \times n$ , and the $n \times m$ matrices over $R$
$M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$	the matrix of the linear transformation $\varphi$ with respect to bases $\mathcal{B}$ (domain) and $\mathcal{E}$ (range)
$\text{tr}(A)$	the trace of the matrix $A$
$\text{Hom}_R(A, B)$	the $R$ -module homomorphisms from $A$ to $B$
$\text{End}(M)$	the endomorphism ring of the module $M$
$\text{Tor}(M)$	the torsion submodule of $M$
$\text{Ann}(M)$	the annihilator of the module $M$
$M \otimes_R N$	the tensor product of modules $M$ and $N$ over $R$
$\mathcal{T}^k(M)$ , $\mathcal{T}(M)$	the $k^{\text{th}}$ tensor power, and the tensor algebra of $M$
$\mathcal{S}^k(M)$ , $\mathcal{S}(M)$	the $k^{\text{th}}$ symmetric power, and the symmetric algebra of $M$
$\bigwedge^k(M)$ , $\bigwedge(M)$	the $k^{\text{th}}$ exterior power, and the exterior algebra of $M$
$m_T(x)$ , $c_T(x)$	the minimal, and characteristic polynomial of $T$
$\text{ch}(F)$	the characteristic of the field $F$
$K/F$	the field $K$ is an extension of the field $F$
$[K : F]$	the degree of the field extension $K/F$
$F(\alpha)$ , $F(\alpha, \beta)$ , etc.	the field generated over $F$ by $\alpha$ or $\alpha, \beta$ , etc.
$m_{\alpha, F}(x)$	the minimal polynomial of $\alpha$ over the field $F$
$\text{Aut}(K)$	the group of automorphisms of a field $K$
$\text{Aut}(K/F)$	the group of automorphisms of a field $K$ fixing the field $F$
$\text{Gal}(K/F)$	the Galois group of the extension $K/F$
$\mathbb{A}^n$	affine $n$ -space
$k[\mathbb{A}^n]$ , $k[V]$	the coordinate ring of $\mathbb{A}^n$ , and of the affine algebraic set $V$
$\mathcal{Z}(I)$ , $\mathcal{Z}(f)$	the locus or zero set of $I$ , the locus of an element $f$
$\mathcal{I}(A)$	the ideal of functions that vanish on $A$
$\text{rad } I$	the radical of the ideal $I$
$\text{Ass}_R(M)$	the associated primes for the module $M$
$\text{Supp}(M)$	the support of the module $M$
$D^{-1}R$	the ring of fractions (localization) of $R$ with respect to $D$
$R_P$ , $R_f$	the localization of $R$ at the prime ideal $P$ , and at the element $f$
$\mathcal{O}_{v, V}$ , $\mathbb{T}_{v, V}$	the local ring, and the tangent space of the variety $V$ at the point $v$
$\mathfrak{m}_{v, V}$	the unique maximal ideal of $\mathcal{O}_{v, V}$
$\text{Spec } R$ , $\text{mSpec } R$	the prime spectrum, and the maximal spectrum of $R$
$\mathcal{O}_X$	the structure sheaf of $X = \text{Spec } R$
$\mathcal{O}(U)$	the ring of sections on an open set $U$ in $\text{Spec } R$
$\mathcal{O}_P$	the stalk of the structure sheaf at $P$
$\text{Jac } R$	the Jacobson radical of the ring $R$
$\text{Ext}_R^n(A, B)$	the $n^{\text{th}}$ cohomology group derived from $\text{Hom}_R$
$\text{Tor}_n^R(A, B)$	the $n^{\text{th}}$ cohomology group derived from the tensor product over $R$
$A^G$	the fixed points of $G$ acting on the $G$ -module $A$
$H^n(G, A)$	the $n^{\text{th}}$ cohomology group of $G$ with coefficients in $A$
$\text{Res}$ , $\text{Cor}$	the restriction, and corestriction maps on cohomology
$\text{Stab}(1 \trianglelefteq A \trianglelefteq G)$	the stability group of the series $1 \trianglelefteq A \trianglelefteq G$
$  \theta  $	the norm of the character $\theta$
$\text{Ind}_H^G(\psi)$	the character of the representation $\psi$ induced from $H$ to $G$

# ABSTRACT ALGEBRA

## Third Edition

**David S. Dummit**  
*University of Vermont*

**Richard M. Foote**  
*University of Vermont*



John Wiley & Sons, Inc.

<b>ASSOCIATE PUBLISHER</b>	<b>Laurie Rosatone</b>
<b>ASSISTANT EDITOR</b>	<b>Jennifer Battista</b>
<b>FREELANCE DEVELOPMENTAL EDITOR</b>	<b>Anne Scanlan-Rohrer</b>
<b>SENIOR MARKETING MANAGER</b>	<b>Julie Z. Lindstrom</b>
<b>SENIOR PRODUCTION EDITOR</b>	<b>Ken Santor</b>
<b>COVER DESIGNER</b>	<b>Michael Jung</b>

This book was typeset using the Y&Y TeX System with DVIWindo. The text was set in Times Roman using *MathTime* from Y&Y, Inc. Titles were set in OceanSans. This book was printed by Malloy Inc. and the cover was printed by Phoenix Color Corporation.

This book is printed on acid-free paper.

Copyright © 2004 John Wiley and Sons, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201)748-6011, fax (201)748-6008, E-mail: PERMREQ@WILEY.COM.

To order books or for customer service please call 1-800-CALL WILEY (225-5945).

**ISBN 0-471-43334-9**

**WIE 0-471-45234-3**

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

*Dedicated to our families  
especially  
Janice, Evan, and Krysta  
and  
Zsuzsanna, Peter, Karoline, and Alexandra*

# Contents

Preface xi

Preliminaries 1

- 0.1 Basics 1
- 0.2 Properties of the Integers 4
- 0.3  $\mathbb{Z}/n\mathbb{Z}$  : The Integers Modulo  $n$  8

## Part I – GROUP THEORY 13

Chapter 1 Introduction to Groups 16

- 1.1 Basic Axioms and Examples 16
- 1.2 Dihedral Groups 23
- 1.3 Symmetric Groups 29
- 1.4 Matrix Groups 34
- 1.5 The Quaternion Group 36
- 1.6 Homomorphisms and Isomorphisms 36
- 1.7 Group Actions 41

Chapter 2 Subgroups 46

- 2.1 Definition and Examples 46
- 2.2 Centralizers and Normalizers, Stabilizers and Kernels 49
- 2.3 Cyclic Groups and Cyclic Subgroups 54
- 2.4 Subgroups Generated by Subsets of a Group 61
- 2.5 The Lattice of Subgroups of a Group 66

<b>Chapter 3</b>	<b>Quotient Groups and Homomorphisms</b>	<b>73</b>
3.1	Definitions and Examples	73
3.2	More on Cosets and Lagrange's Theorem	89
3.3	The Isomorphism Theorems	97
3.4	Composition Series and the Hölder Program	101
3.5	Transpositions and the Alternating Group	106

<b>Chapter 4</b>	<b>Group Actions</b>	<b>112</b>
4.1	Group Actions and Permutation Representations	112
4.2	Groups Acting on Themselves by Left Multiplication—Cayley's Theorem	118
4.3	Groups Acting on Themselves by Conjugation—The Class Equation	122
4.4	Automorphisms	133
4.5	The Sylow Theorems	139
4.6	The Simplicity of $A_n$	149

<b>Chapter 5</b>	<b>Direct and Semidirect Products and Abelian Groups</b>	<b>152</b>
5.1	Direct Products	152
5.2	The Fundamental Theorem of Finitely Generated Abelian Groups	158
5.3	Table of Groups of Small Order	167
5.4	Recognizing Direct Products	169
5.5	Semidirect Products	175

<b>Chapter 6</b>	<b>Further Topics in Group Theory</b>	<b>188</b>
6.1	$p$ -groups, Nilpotent Groups, and Solvable Groups	188
6.2	Applications in Groups of Medium Order	201
6.3	A Word on Free Groups	215

## Part II – RING THEORY 222

<b>Chapter 7</b>	<b>Introduction to Rings</b>	<b>223</b>
7.1	Basic Definitions and Examples	223
7.2	Examples: Polynomial Rings, Matrix Rings, and Group Rings	233
7.3	Ring Homomorphisms and Quotient Rings	239
7.4	Properties of Ideals	251
7.5	Rings of Fractions	260
7.6	The Chinese Remainder Theorem	265

**Chapter 8 Euclidean Domains, Principal Ideal Domains and Unique Factorization Domains 270**

- 8.1 Euclidean Domains 270
- 8.2 Principal Ideal Domains (P.I.D.s) 279
- 8.3 Unique Factorization Domains (U.F.D.s) 283

**Chapter 9 Polynomial Rings 295**

- 9.1 Definitions and Basic Properties 295
- 9.2 Polynomial Rings over Fields I 299
- 9.3 Polynomial Rings that are Unique Factorization Domains 303
- 9.4 Irreducibility Criteria 307
- 9.5 Polynomial Rings over Fields II 313
- 9.6 Polynomials in Several Variables over a Field and Gröbner Bases 315

**Part III – MODULES AND VECTOR SPACES 336**

**Chapter 10 Introduction to Module Theory 337**

- 10.1 Basic Definitions and Examples 337
- 10.2 Quotient Modules and Module Homomorphisms 345
- 10.3 Generation of Modules, Direct Sums, and Free Modules 351
- 10.4 Tensor Products of Modules 359
- 10.5 Exact Sequences—Projective, Injective, and Flat Modules 378

**Chapter 11 Vector Spaces 408**

- 11.1 Definitions and Basic Theory 408
- 11.2 The Matrix of a Linear Transformation 415
- 11.3 Dual Vector Spaces 431
- 11.4 Determinants 435
- 11.5 Tensor Algebras, Symmetric and Exterior Algebras 441

**Chapter 12 Modules over Principal Ideal Domains 456**

- 12.1 The Basic Theory 458
- 12.2 The Rational Canonical Form 472
- 12.3 The Jordan Canonical Form 491

**Chapter 13 Field Theory 510**

- 13.1 Basic Theory of Field Extensions 510
- 13.2 Algebraic Extensions 520
- 13.3 Classical Straightedge and Compass Constructions 531
- 13.4 Splitting Fields and Algebraic Closures 536
- 13.5 Separable and Inseparable Extensions 545
- 13.6 Cyclotomic Polynomials and Extensions 552

**Chapter 14 Galois Theory 558**

- 14.1 Basic Definitions 558
- 14.2 The Fundamental Theorem of Galois Theory 567
- 14.3 Finite Fields 585
- 14.4 Composite Extensions and Simple Extensions 591
- 14.5 Cyclotomic Extensions and Abelian Extensions over  $\mathbb{Q}$  596
- 14.6 Galois Groups of Polynomials 606
- 14.7 Solvable and Radical Extensions: Insolvability of the Quintic 625
- 14.8 Computation of Galois Groups over  $\mathbb{Q}$  640
- 14.9 Transcendental Extensions, Inseparable Extensions, Infinite Galois Groups 645

**Part V – AN INTRODUCTION TO COMMUTATIVE RINGS,  
ALGEBRAIC GEOMETRY, AND  
HOMOLOGICAL ALGEBRA 655**

**Chapter 15 Commutative Rings and Algebraic Geometry 656**

- 15.1 Noetherian Rings and Affine Algebraic Sets 656
- 15.2 Radicals and Affine Varieties 673
- 15.3 Integral Extensions and Hilbert's Nullstellensatz 691
- 15.4 Localization 706
- 15.5 The Prime Spectrum of a Ring 731

**Chapter 16 Artinian Rings, Discrete Valuation Rings, and  
Dedekind Domains 750**

- 16.1 Artinian Rings 750
- 16.2 Discrete Valuation Rings 755
- 16.3 Dedekind Domains 764

**Chapter 17 Introduction to Homological Algebra and Group Cohomology 776**

- 17.1 Introduction to Homological Algebra—Ext and Tor 777
- 17.2 The Cohomology of Groups 798
- 17.3 Crossed Homomorphisms and  $H^1(G, A)$  814
- 17.4 Group Extensions, Factor Sets and  $H^2(G, A)$  824

**Part VI – INTRODUCTION TO THE REPRESENTATION THEORY OF FINITE GROUPS 839**

**Chapter 18 Representation Theory and Character Theory 840**

- 18.1 Linear Actions and Modules over Group Rings 840
- 18.2 Wedderburn's Theorem and Some Consequences 854
- 18.3 Character Theory and the Orthogonality Relations 864

**Chapter 19 Examples and Applications of Character Theory 880**

- 19.1 Characters of Groups of Small Order 880
- 19.2 Theorems of Burnside and Hall 886
- 19.3 Introduction to the Theory of Induced Characters 892

**Appendix I: Cartesian Products and Zorn's Lemma 905**

**Appendix II: Category Theory 911**

**Index 919**

# Preface to the Third Edition

The principal change from the second edition is the addition of Gröbner bases to this edition. The basic theory is introduced in a new Section 9.6. Applications to solving systems of polynomial equations (elimination theory) appear at the end of this section, rounding it out as a self-contained foundation in the topic. Additional applications and examples are then woven into the treatment of affine algebraic sets and  $k$ -algebra homomorphisms in Chapter 15. Although the theory in the latter chapter remains independent of Gröbner bases, the new applications, examples and computational techniques significantly enhance the development, and we recommend that Section 9.6 be read either as a segue to or in parallel with Chapter 15. A wealth of exercises involving Gröbner bases, both computational and theoretical in nature, have been added in Section 9.6 and Chapter 15. Preliminary exercises on Gröbner bases can (and should, as an aid to understanding the algorithms) be done by hand, but more extensive computations, and in particular most of the use of Gröbner bases in the exercises in Chapter 15, will likely require computer assisted computation.

Other changes include a streamlining of the classification of simple groups of order 168 (Section 6.2), with the addition of a uniqueness proof via the projective plane of order 2. Some other proofs or portions of the text have been revised slightly. A number of new exercises have been added throughout the book, primarily at the ends of sections in order to preserve as much as possible the numbering schemes of earlier editions. In particular, exercises have been added on free modules over noncommutative rings (10.3), on Krull dimension (15.3), and on flat modules (10.5 and 17.1).

As with previous editions, the text contains substantially more than can normally be covered in a one year course. A basic introductory (one year) course should probably include Part I up through Section 5.3, Part II through Section 9.5, Sections 10.1, 10.2, 10.3, 11.1, 11.2 and Part IV. Chapter 12 should also be covered, either before or after Part IV. Additional topics from Chapters 5, 6, 9, 10 and 11 may be interspersed in such a course, or covered at the end as time permits.

Sections 10.4 and 10.5 are at a slightly higher level of difficulty than the initial sections of Chapter 10, and can be deferred on a first reading for those following the text sequentially. The latter section on properties of exact sequences, although quite long, maintains coherence through a parallel treatment of three basic functors in respective subsections.

Beyond the core material, the third edition provides significant flexibility for students and instructors wishing to pursue a number of important areas of modern algebra,

either in the form of independent study or courses. For example, well integrated one-semester courses for students with some prior algebra background might include the following: Section 9.6 and Chapters 15 and 16; or Chapters 10 and 17; or Chapters 5, 6 and Part VI. Each of these would also provide a solid background for a follow-up course delving more deeply into one of many possible areas: algebraic number theory, algebraic topology, algebraic geometry, representation theory, Lie groups, etc.

The choice of new material and the style for developing and integrating it into the text are in consonance with a basic theme in the book: the power and beauty that accrues from a rich interplay between different areas of mathematics. The emphasis throughout has been to motivate the introduction and development of important algebraic concepts using as many examples as possible. We have not attempted to be encyclopedic, but have tried to touch on many of the central themes in elementary algebra in a manner suggesting the very natural development of these ideas.

A number of important ideas and results appear in the exercises. This is not because they are not significant, rather because they did not fit easily into the flow of the text but were too important to leave out entirely. Sequences of exercises on one topic are prefaced with some remarks and are structured so that they may be read without actually doing the exercises. In some instances, new material is introduced first in the exercises—often a few sections before it appears in the text—so that students may obtain an easier introduction to it by doing these exercises (e.g., Lagrange’s Theorem appears in the exercises in Section 1.7 and in the text in Section 3.2). All the exercises are within the scope of the text and hints are given [in brackets] where we felt they were needed. Exercises we felt might be less straightforward are usually phrased so as to provide the answer to the exercise; as well many exercises have been broken down into a sequence of more routine exercises in order to make them more accessible.

We have also purposely minimized the functorial language in the text in order to keep the presentation as elementary as possible. We have refrained from providing specific references for additional reading when there are many fine choices readily available. Also, while we have endeavored to include as many fundamental topics as possible, we apologize if for reasons of space or personal taste we have neglected any of the reader’s particular favorites.

We are deeply grateful to and would like here to thank the many students and colleagues around the world who, over more than 15 years, have offered valuable comments, insights and encouragement—their continuing support and interest have motivated our writing of this third edition.

David Dummit  
Richard Foote  
June, 2003