THE MACMILLAN COMPANY
NEW YORK · CHICAGO
DALLAS · ATLANTA · SAN FRANCISCO

THE MACMILLAN COMPANY
OF CANADA, LIMITED
TORONTO

THEORY OF NUMBERS

B. M. STEWART
Associate Professor of Mathematics
Michigan State College

THE MACMILLAN COMPANY *New York*

PRINTED IN THE UNITED STATES OF AMERICA

First printing

PREFACE

The present work makes no pretense at being more than a textbook; the subject matter is classical and the only attempt at originality is in the choice of topics and the manner of presentation. A conscious effort has been made to have the style that of a classroom lecture and to divide the material in such a way that the basic parts of each chapter may be presented in a fifty-minute period, although in many cases the details will have to be assigned for outside study. Accompanying every chapter is a set of exercises of varied difficulty, some designed to illustrate the lesson topics and others, to extend them and open new horizons.

In mimeographed form this material has been used for a number of years at Michigan State College. The background for its production was that available books seemed to be either too short or too long, too easy or too advanced, for a short course, involving twenty-five to thirty lessons, offered to a mixed group of undergraduate and beginning graduate students.

It is perhaps characteristic of the subject matter that certain topics will appeal especially to one person and not to another. To a beginner it may be difficult to decide whether a topic which does not interest him is required for later chapters. Therefore, as a footnote to the title of each chapter, there is a brief discussion as to whether, in the author's opinion, the chapter is a basic one, containing minimum essentials, or is of a supplementary nature, containing material that might be omitted from a short course.

For example, the experienced reader will recognize the slide rule modulo 29 to be a supplementary item as far as the theory of numbers is concerned; but the author hopes that many teachers will agree that here is a device that may pique student interest and whether class time can be devoted to a topic is relatively unimportant once interest is aroused.

In a few places where a part of a supplementary chapter is required in a later chapter, the necessary cross reference is indicated in the introductory footnote. It is hoped that these notes may help the inexperienced reader not to spend too much time on relatively unimportant topics at the expense of basic materials.

The quotations with which the chapters begin are meant only for pleasant meditation, but in places they are almost in context and reflect, in part, the attitude which the author would hope to implant in his student readers.

Acknowledgments and references are to be found at various places in the text, but two of more personal nature demand relating here.

It was Professor A. J. Kempner who first showed me that mathematics could be a delightful, creative, challenging pursuit. "Just for the fun of it," he said as he set me to reading Carmichael's monograph. I still recall the kindly way he sent me off to read for myself in Dickson's *History* how Descartes had anticipated me by several hundred years in devising some multipli-perfect numbers.

It was Professor C. C. MacDuffee whose forbearance and encouragement made life as a graduate student brighter. In retrospect this book seems little more than an elaboration of his explanations; yet it is my hope that these lessons will in some way attract new minds to mathematics and number theory and thus extend the chain of indebtedness.

B. M. Stewart

Lansing, Michigan

CONTENTS

Chapter 1

PRELIMINARY CONSIDERATIONS

Chapter 2

NUMBER THEORY IN THE GAME OF SOLITAIRE

Chapter 3

MATHEMATICAL INDUCTION

Chapter 4

REPRESENTATION OF THE INTEGERS

Chapter 10

THE FUNCTION $E(p, n)$

Chapter 11

GROUPS OF TRANSFORMATIONS; MATRICES AND DETERMINANTS

Chapter 12

DIOPHANTINE EQUATIONS OF THE FIRST DEGREE

Chapter 13

MORE DIOPHANTINE EQUATIONS OF THE FIRST DEGREE

Chapter 14

PYTHAGOREAN TRIPLETS

Chapter 20

CONGRUENCES OF HIGHER DEGREE

Chapter 21

EXPONENTS, PRIMITIVE ROOTS, AND INDICES

Chapter 22

QUADRATIC RESIDUES AND LEGENDRE'S SYMBOL

Chapter 23

THE QUADRATIC RECIPROCITY LAW

Chapter 30

RATIONAL NUMBERS

Chapter 31

DECIMAL REPRESENTATION OF RATIONAL NUMBERS

Chapter 32

CONTINUED FRACTIONS

Chapter 33

THE FUNDAMENTAL THEOREM RECONSIDERED

THEORY OF NUMBERS

▶*Die ganzen Zahlen hat Gott gemacht, alles andere ist Menschenwerk.*

—L. KRONECKER

CHAPTER 1[*]

PRELIMINARY CONSIDERATIONS

1.1. Prerequisites. In this textbook on the theory of numbers we shall assume that the student has completed the usual work in college algebra, so that he is well acquainted with such topics as symbols of grouping, exponents, and factoring. We want to make use of inequalities and absolute value and we include below a brief review of these essential notions. Although we shall make occasional use of ideas from analytic geometry, beginning calculus, and the theory of equations, we shall try to make the exposition of such topics self-contained. The whole point of these preliminary remarks is to encourage the reader of modest background and to point out that our real subject matter is more directly related to grade school arithmetic than it is to advanced courses.

However, our work will demand a maturity of attitude and an ingenuity of mind that will do more than test ability at sums. In fact we shall find that one of the attractions of our subject (and it has attracted mathematicians, both amateur and professional, for over 2000 years) is that some of its problems can be stated in such simple form that the celebrated man in the street can understand what is wanted, and yet the concentrated efforts of generations of

[*] Note to the reader and instructor: Chapter 1 is a basic chapter, with essential orientation, terminology, and references.

workers have as yet failed to yield solutions. Naturally in an introductory textbook we shall present topics that have been well worked out and problems for which definite methods of attack can be suggested, but even in our somewhat elementary situations the need for ingenuity and special methods will manifest itself in many problems and will provide a real challenge to interested and serious students.

Our purpose is to study some of the properties of the ordinary *integers*, i.e., the positive and negative whole numbers and zero:

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots \quad .$$

We shall assume that the arithmetic of these numbers is well known to each student; that the *addition* and *multiplication tables* and the *rules of exponents* are perfectly understood; and that the *representation* of these numbers *using the base* 10 is completely mastered, so that a symbol like 7203 is immediately recognized as an abbreviation for

$$7(10)^3 + 2(10)^2 + 0(10) + 3 \quad .$$

We shall assume that the algebraic symbolism in postulates like the following is readily interpreted and that the laws here expressed are willingly granted (but in a later lesson we shall indicate how these laws may be proved as theorems on the basis, of course, of other still simpler postulates):

The *associative* laws for addition and multiplication:

$$(a + b) + c = a + (b + c), \quad (ab)c = a(bc).$$

The *commutative* laws for addition and multiplication:

$$a + b = b + a, \quad ab = ba.$$

The *distributive* law relating addition and multiplication:

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

The law that *subtraction* is always *possible*:

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

For all these laws it is understood that a,b,c are *any* integers, *not necessarily different* and that the equation $a + x = b$ is always solvable for x, an integer, without inventing any new numbers.

We shall assume that the student has sufficient maturity to appreciate statements like the following: "If a and b are given integers, the equation $ax = b$ is, in general, impossible of solution in integers." For example, simple possible and impossible cases are $2x = 4$ and $2x = 3$, respectively. Of course the student will be expected at times to deal with numbers like 3/2, but the point is that he must not admit their use in problems where only the integers are under consideration.

The student must learn to appreciate the warning: "The statement that a given equation is possible, or impossible, of solution is meaningless until the system in which solutions are sought has been specified." From such warnings the student must learn to state theorems, problems, and solutions with precision.

We need also the concept of *inequality* where we write $a < b$ and read that "a is less than b" (equivalently we write $b > a$ and read that "b is greater than a") if and only if there exists a *positive* number p such that $a + p = b$. For example, $-3 < 5$ because $-3 + 8 = 5$, with $p = 8$ a positive number; but $-5 < -3$ because $-5 + 2 = -3$, with $p = 2$ a positive number. If it seems to the reader that the listing $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$ is a natural way in which to order the integers, then he already has a good intuitive idea of the meaning of $a < b$, realizing that it is equivalent to saying that in the above listing, a occurs "to the left" of b. The symbol $a \leqq b$ will indicate that a is *either* equal to b *or* less than b. For example, $x > 0$ is a convenient way to say that "x is positive," $x < 0$ a way to say that "x is negative," while $x \geqq 0$ is a way of saying that "x is non-negative" meaning that x is either positive or zero. Again, if x is an integer and is either $-2, -1, 0, 1, 2$, or 3, then an easy way to indicate this last restriction is to write that $-3 < x < 4$, for this is understood to mean that x must satisfy *both* the conditions of being less than 4 and greater than -3; an equivalent statement would be $-2 \leqq x \leqq 3$.

Two important rules about inequalities are as follows:

If $a < b$, then $a + c < b + c$ for any number c.

If $a < b$, then $ac < bc$ for any positive number c, but $ac > bc$ for any negative number c.

The reader will soon see why the last rule of inequality has two cases if he contrasts the rule of signs for multiplication by a negative number with the rule for multiplication by a positive number.

At times we find it convenient to speak of the *absolute value* (or *numerical value*) of a number, using the symbol $|a|$ whose definition is as follows:

$$|0| = 0; \quad \text{if } a > 0, \text{ then } |a| = a; \quad \text{if } a < 0, \text{ then } |a| = -a$$

For example, $|6| = 6$ and $|-10| = 10$.

Two important rules about absolute value are as follows:

$$|ab| = |a|\,|b|; \quad |a + b| \leqq |a| + |b|.$$

The reader can establish these rules by considering the various cases that arise according as one of a or b is zero, according as a and b have like or unlike signs, and according as $|a|$ or $|b|$ is the greater or that $|a| = |b|$.

Finally, we shall assume that the student is acquainted with the *division algorithm*, and to this topic we devote the next section.

1.2. The division algorithm. An algorithm is a step-by-step process, complete in a finite number of steps, for solving a given problem. By the division algorithm we mean that process with which the student became familiar in arithmetic, where he was given, say, the dividend 712, the divisor 13, and was asked to find the quotient and the remainder. By a long division he found

$$\begin{array}{r|l} & \;\;54 = q \\ \hline b = 13 & 712 = a \\ & 650 \\ \hline & \;\;\;62 \\ & \;\;\;52 \\ \hline & \;\;\;10 = r \end{array}$$

and concluded that the quotient is $q = 54$ and the remainder is $r = 10$, with the process ending at this point because $10 < 13$. Certain steps of the long division work are tentative; for example, to find that the first part of the quotient is 50, not 40 or 60, may require a student who does not know the multiples of 13 to make several trials, but not more than nine (fewer, we hope!).

In general, the division algorithm is that process, complete in a finite number of steps, by which for any given integer a (the dividend) and any given *non-zero* integer b (the divisor), we find the values of *the* integer q (the quotient) and *the* non-negative integer r (the remainder) such that

$$a = qb + r, \quad 0 \leqq r < |b|.$$

We consider first the case when b is a positive integer, for then from the standard ordering of the integers

$$\cdots < -3 < -2 < -1 < 0 < +1 < +2 < +3 < \ldots$$

we obtain a standard ordering of the multiples of b:

$$\cdots < -3b < -2b < -b < 0 < b < 2b < 3b < \ldots.$$

Then any given integer a must either be a certain one of the numbers of this list, so that $a = qb + 0$; or a must fall within a certain one of the intervals, say $qb < a < (q + 1)b$, and then $a = qb + r$, $0 < r < b$; combining the two cases we have the *existence* of q and r, exactly as described in the division algorithm. When a is positive the long division process allows us to *find* q and r in a finite number of steps, but the case when a is negative requires special consideration. For example, if $a = -712$ and $b = 13$, then from our previous example we have $a = -54b - 10$, but this remainder is negative; however, by subtracting and adding 13, we find $a = -55b + 3$ and with $q = -55$ and $r = 3$ we satisfy the requirement $0 \leqq r < b$. In general, if for $a > 0$ we have found $a = Qb$, then for $-a$ we may write $-a = qb + r$ where we have set $q = -Q$ and $r = 0$; but if $a = Qb + R$, $0 < R < b$, then for $-a$ we may write $-a = qb + r$ where we have set $q = -(Q + 1)$ and $r = b - R$; in both circumstances r satisfies the requirement $0 \leqq r < b$.

If b is negative, then $|b|$ is positive, and hence we may use the previous arguments to find Q and r so that $a = Q|b| + r$, $0 \leqq r < |b|$. Then noting that $|b| = -b$ and taking $q = -Q$, we may write $a = qb + r$, $0 \leqq r < |b|$.

This concludes the description of the division algorithm except for the remark that the italicizing of "... *the* integer q and *the* integer r ..." was purposeful, because for a given pair of integers a and b the corresponding q and r are indeed unique.

For if we suppose that there are two sets of solutions, say, $a = qb + r$, $0 \leqq r < |b|$ and $a = q_1b + r_1$, $0 \leqq r_1 < |b|$; then when we equate the two expressions for a and rearrange the result we have $(q - q_1)b = r_1 - r$, so that $r_1 - r$ is a multiple of b. But the inequalities which r and r_1 satisfy allow us to deduce that $-b < r_1 - r < b$. Hence the only suitable multiple of b is 0. But $r_1 - r = 0$ shows $r_1 = r$; then since $b \neq 0$, $(q - q_1)b = 0$ implies $q - q_1 = 0$ or $q = q_1$. Thus the quotient and remainder in the division algorithm are both unique.

1.3. Related material.

In studying, teaching, and writing about this subject, we are much in debt to most of the following authors and their books. For convenience of reference we list them here and suggest that they be used for collateral reading and further research.

Carmichael, R. D., *Theory of Numbers*. Mathematical Monographs, No. 13. New York, Wiley, 1914.

Dickson, L. E., *Introduction to the Theory of Numbers*. Chicago, University of Chicago Press, 1931.

------, *Modern Elementary Theory of Numbers.* Chicago, University of Chicago Press, 1939.

------, *History of the Theory of Numbers.* Carnegie Institution, Vol. I, 1919; Vol. II, 1920; Vol. III, 1923.

Hardy, G. H., and Wright, E. M., *An Introduction to the Theory of Numbers.* Oxford, Clarendon Press, 1938.

MacDuffee, C. C., *Introduction to Abstract Algebra* (Chapter I). New York, Wiley, 1940.

Ore, O., *Number Theory and Its History.* New York, McGraw-Hill, 1948.

Uspensky, J. V., and Heaslet, M. H., *Elementary Number Theory.* New York, McGraw-Hill, 1939.

Wright, H. N., *First Course in the Theory of Numbers.* New York, Wiley, 1939.

Any large library will provide many other reference books in English, French, and German. We are especially fond of the well-written books of E. Landau, such as his *Grundlagen der Analysis* and *Vorlesungen über Zahlentheorie* (Vol. I) which are available in Chelsea reprints. The first of these is now available in English translation as *Foundations of Analysis*. New York, Chelsea, 1951.

1.4. The position of our subject. From Carl Friedrich Gauss (1777–1855), the "Prince of Mathematicians," we have the saying: "Mathematics is the queen of the sciences, but arithmetic is the queen of mathematics." By arithmetic Gauss meant our subject, theory of numbers, and his attitude in regard to the queenship was based on two almost opposite attributes: on one hand, the theory of numbers is the purest of pure mathematics, many of its problems being of interest only in themselves and not for any applications they may have; on the other hand, most of the number systems used in more practical branches of mathematics have the integers as their basic building blocks. So it is perhaps not too surprising that there have been some very remarkable interchanges of ideas, problems, and solutions between the ivory towers of number theory and the laboratories of applied mathematics.

Historically our subject has important roots in the works of Diophantus of the third century in whose honor we term the search for integer solutions of a given equation the solving of a "Diophantine equation." Early in our work we will encounter the names of Pythagoras and Euclid—yes, the Euclid of fame in geometry, the latter an important contributor to our subject as early as 300 B.C.

From these earliest men to the present there is hardly a mathematician of note who has not contributed in some way to number theory. As we pursue the subject we find some names such as those of Fermat, Euler, Legendre, Gauss, Eisenstein, and Jacobi occurring often; but other men of lesser fame are remembered too, sometimes for just one particular theorem.

Dickson's monumental "History of the Theory of Numbers" is a mine of historical and factual information that the student will find particularly valuable if he makes some little discovery of his own (and one of the nice features of our subject is how soon the student can explore for himself) and wonders whether it has been published before. Our subject is not a dead one, many famous problems are still being attacked and new ones are being proposed, and the most recent journals carry articles and problems that concern our course directly. At certain places in the development we will suggest some generalizations of the subject, and then, for the interested student, whole new fields of exploration and study will be opened.

EXERCISES

EX. *1.1.* For any integer x, prove that $x^2 \geqq 0$.

EX. *1.2.* Supposing x to be an integer interpret the following statements (i.e., find all solutions in integers):
(a) $0 < x < 9$; (b) $x^2 < 9$; (c) $x^2 \leqq 9$; (d) $|x| < 9$

EX. *1.3* Supposing x to be any integer show that:
(a) $f(x) = x^2 - 4x + 5$ satisfies $f(x) > 0$;
(b) $g(x) = x^2 - 5x + 6$ satisfies $g(x) \geqq 0$.

EX. *1.4.* Prove the two rules about inequalities given in the text.

EX. *1.5.* If a, b, c are integers with $ac > bc$ and $c > 0$, does it follow that $a > b$?

EX. *1.6.* If $a < b$ and $b < c$, prove that $a < c$.

EX. *1.7.* Prove the two rules about absolute value given in the text.

EX. *1.8.* In each of the following cases find integers q and r such that $a = qb + r$, $0 \leqq r < |b|$:
(a) $a = 7143$, $b = 17$; (b) $a = -2047$, $b = 130$; (c) $a = -6080$, $b = -42$.

> ►*It is well in order to aid the understanding and memory to choose intermediate truths (which are called* lemmas, *since they appear to be a digression) which will shorten the major proof and yet appear memorable and worthy in themselves of being demonstrated and there is real art in this.*
>
> — G. W. LEIBNITZ

CHAPTER 2*

NUMBER THEORY IN THE GAME OF SOLITAIRE

2.1. Parity. Almost every textbook in the theory of numbers leads off with a few interesting problems and games whose solution depends in some way on the properties of integers. In line with this tradition, but desiring not to duplicate the usual examples, we present here a description of the game of Solitaire. But first we shall give the bit of number theory that is required in the mathematical theory associated with the game.

By the division algorithm when $b = 2$ and a is an integer, the possible remainders are $r = 0$ and $r = 1$; when $r = 0$, the integer a is called *even* and has the form $a = 2q$; when $r = 1$, the integer a is called *odd* and has the form $a = 2q + 1$. If two given integers s and t are both even, or both odd, then s and t are said to be of the *same parity*; but if one of s and t is even, and the other odd, then s and t are said to be of *different parity*.

*Except for the concept of parity in **2.1**, Chapter **2** is a supplementary chapter (albeit a favorite of the author) which may be omitted in a short course.

The observation which we shall need and which we shall call a "lemma," meaning a tool theorem or subordinate theorem useful in proving other more interesting or more important theorems, is as follows:

Lemma: The difference $s - t$ of two given integers s and t is even if and only if s and t are of the same parity.

Proof: The four possible cases are as follows:

$$\begin{aligned} 2S - 2T &= 2(S - T), \\ (2S + 1) - 2T &= 2(S - T) + 1, \\ (2S + 1) - (2T + 1) &= 2(S - T), \\ 2S - (2T + 1) &= 2(S - T - 1) + 1. \end{aligned}$$

2.2. The game of Solitaire. Of ancient origin is the game which we are about to describe, although the first mathematical mention of it seems to be by Leibnitz. The game of Solitaire is played upon a field, of arbitrary but fixed shape, consisting of squares arranged in rows and columns. On certain of these squares appear playing pieces, at most one piece to a square. A move, or jump, is possible when on three adjacent squares A,B,C of a row or column (but *not* a diagonal) there are pieces on A and B, but none on C. The jump consists in moving the piece on A to C and removing the piece on B from play. The object of the game is by a succession of jumps (of course at least one square must be empty initially so that the game can begin) to leave the remaining pieces in some stated configuration upon the field (usually to leave only one piece on the field).

The object of a mathematical study of Solitaire is to show that some proposed games of Solitaire are impossible of solution or that the final outcome is limited in some way. Since we anticipate using the theory of integers, it is natural to write down equations which describe the progress of the game and which have as variables the number of pieces and the number of jumps, for neither fractional pieces nor fractional jumps are allowed, and this procedure will surely lead to a Diophantine problem. Such an analysis is possible if we first label the squares along one set of diagonals, say those running from upper left to lower right, in a systematic manner: first, say, a diagonal colored green, the next colored purple, the next tan, and the next green, and then all the rest in cyclic manner: purple, tan, green, purple, tan, green, etc. For example, in Figure 1 such a labeling has been carried out for a rectangular 7-by-5 field with the obvious abbreviations: G for green, P for purple, and T for tan.

G	P	T	G	P
T	G	P	T	G
P	T	G	P	T
G	P	T	G	P
T	G	P	T	G
P	T	G	P	T
G	P	T	G	P

FIGURE 1

With such a labeling it becomes possible to assert that every jump ending on a square of one color increases by one the number of pieces on squares of that color and decreases by one the number of pieces on each of the other two colors. Let G,P,T be given the new meaning of indicating, respectively, the number of pieces initially present on squares of green, purple, or tan color. Let g,p,t indicate, respectively, the number of jumps ending on squares of green, purple, or tan color. Let G',P',T' indicate, respectively, the number of pieces finally present on squares of green, purple, or tan color.

Then using our previous observation about the effect of each jump, we find that these *integers* must satisfy the following system of equations:

$$(2.1) \qquad G + g - p - t = G', \quad P - g + p - t = P', \quad T - g - p + t = T'.$$

By any of the usual methods, such as adding the equations in pairs, we can show the system (2.1) to be equivalent to the following system of equations:

$$(2.2) \qquad 2g = (P + T) - (P' + T'), \quad 2p = (T + G) - (T' + G'), \quad 2t = (G + P) - (G' + P').$$

Inasmuch as all the variables are integers, we are in a position to apply the lemma given in **2.1**,* and to conclude that *one set of necessary* conditions for the game of Solitaire to be possible is that the initial and final distribution of the pieces are such that

$$(2.3) \qquad \left.\begin{array}{l} P + T \text{ and } P' + T' \text{ are of the same parity,} \\ T + G \text{ and } T' + G' \text{ are of the same parity,} \\ G + P \text{ and } G' + P' \text{ are of the same parity.} \end{array}\right\}$$

This set of conditions (2.3) is sometimes powerful enough in itself to decide that a proposed game of Solitaire is impossible of solution. Another set of necessary conditions is obtained by labeling with colors the diagonals that run from upper right to lower left. If either set of conditions fails to be satisfied, then the game is impossible. That

*A bold-face reference, as by **2.1**, is to the first section of Chapter 2.

these two sets of conditions are necessary, but not sufficient, to guarantee a solution may be shown by the reader by studying a very small playing field or one with widely separated pieces. When the two sets of conditions are satisfied, one can sometimes show the existence of a solution to the game by the perhaps non-mathematical, but amusing, process of carrying out a suitable sequence of jumps.

Let us apply this theory to the 7-by-5 field shown in Figure **1**, supposing that initially all squares except the top left one are occupied, so that $G = 11$, $P = 12$, $T = 11$. Suppose that only one piece is to be left at the end of the game, so that $(G',P',T',)$ is either (1,0,0) or (0,1,0) or (0,0,1). Both the first and last of these suggested endings violate (*2.3*), so to attempt to leave the final piece on the lower left square, for example, is to attack an impossible game. However, the ending (0,1,0) is compatible with the conditions (*2.3*).

Now let us employ a set of labels on the other diagonals, say red, blue, and yellow, indicated by R,B, and Y, respectively, as in Figure 2. Then using the same type of symbolism as before, we find that the proposed game has $R = 12, B = 11$, $Y = 11$. Conditions analogous to (*2.3*) then require $(R',B',Y') = (1,0,0)$, or otherwise, the game will be impossible.

B	*R*	*Y*	*B*	*R*
R	*Y*	*B*	*R*	*Y*
Y	*B*	*R*	*Y*	*B*
B	*R*	*Y*	*B*	*R*
R	*Y*	*B*	*R*	*Y*
Y	*B*	*R*	*Y*	*B*
B	*R*	*Y*	*B*	*R*

FIGURE 2

Combining these two sets of conditions we see that if the proposed game is possible, beginning with only the upper left square empty and closing with but one piece on the field, then the final piece *must* be on one of the squares marked X in Figure 3, for these are the only squares carrying *both* a purple and a red label.

To the uninitiated there must surely be something of black magic in such assertions, and we feel the fascinating lure of number theory when we see that the whole matter depends essentially on setting the problem in such a way that we can apply the trivial lemma of **2.1**.

	X			*X*
	X			*X*
	X			*X*

FIGURE 3

2.3. "Red Cross" Solitaire. To illustrate a somewhat more difficult variation of Solitaire, let us consider the 7-by-5 field and study the possibility of having but one square initially empty and the final five pieces in the position shown in Figure 4. Since we first proposed this game as a means of attracting attention to a charity drive, we have taken the liberty of calling it "Red Cross" Solitaire.

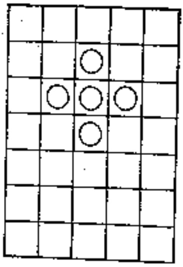


FIGURE 4

The complete field has $P = 12, G = 12, T = 11$. In the "Red Cross" ending, $P' = 2, G' = 1, T' = 2$. Hence if but one square is empty initially, the parity conditions (*2.3*) can be satisfied only if the empty square is a P-square. Similarly, with reference to the other set of diagonals, the empty square must be a B-square. Thus a necessary condition is that the initial configuration be as in Case 1 or Case 2 of Figure 5. That either of these conditions is also sufficient we show by producing the play-by-play solution.

The reader should, of course, not deny himself the fun of trying this game, but if he tires from failures to reach the desired ending, he can trace through the plays indicated in Figure 6. In that figure the first two rows show how both cases can be brought into the same form, so that the remainder of the solution, in the other rows of the figure, is the same for the two cases. In each diagram a "plus" circle shows the piece which is to make the jump and a "minus" circle shows the piece which is to be removed from the field. The arrows connecting the diagrams show the sequence in which the jumps are to be performed.

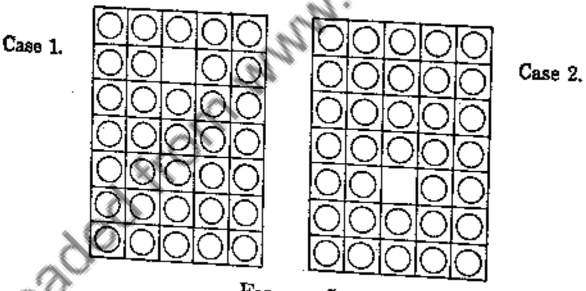


FIGURE 5

EXERCISES

EX. *2.1.* For the Solitaire game described in **2.2**, show that at least some of the endings suggested in Figure 3 are possible by actually carrying

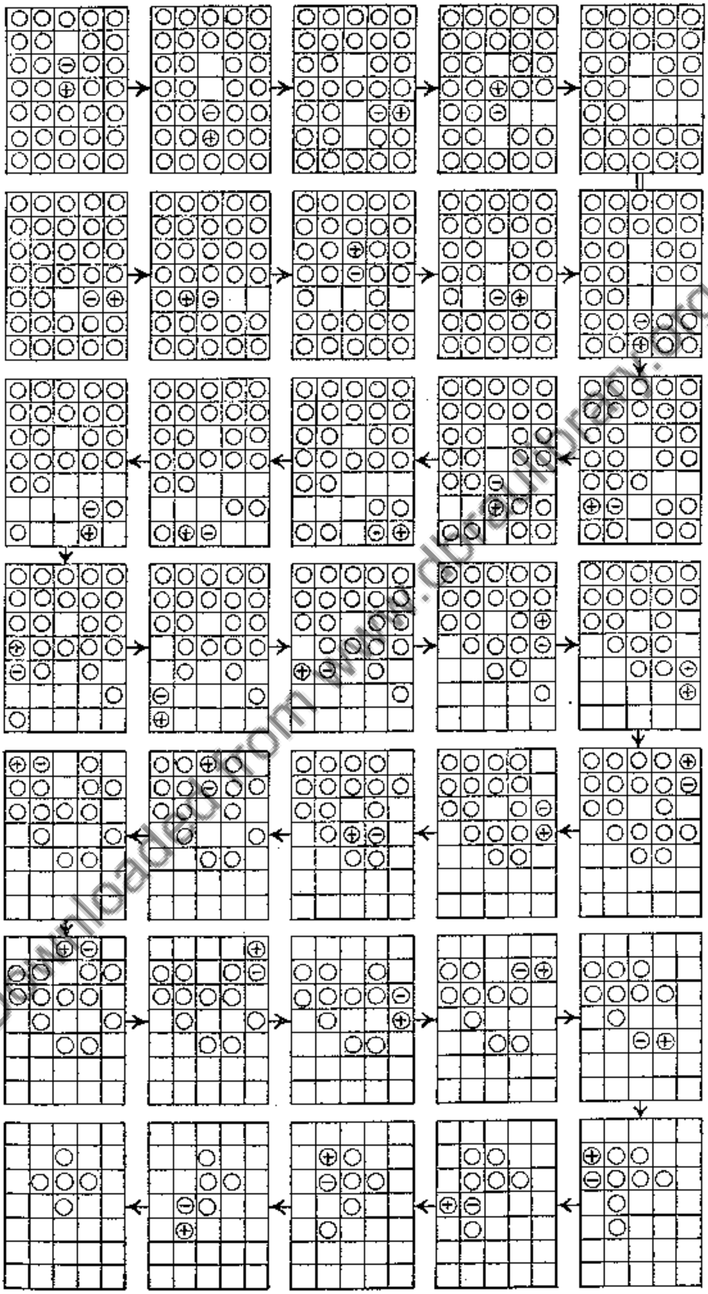


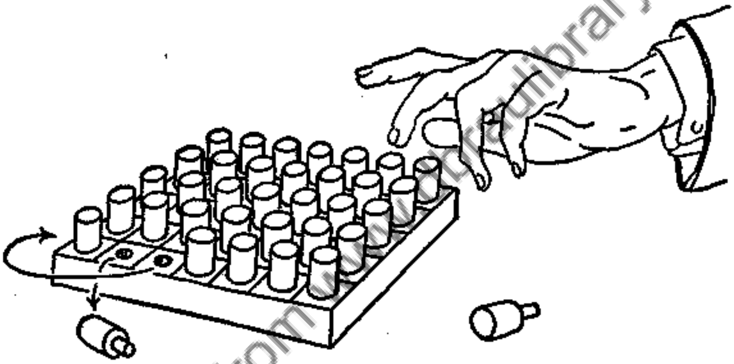
FIGURE 6

out the game. (In fact, two well-planned attacks, leaving the last three moves variable, can be found to prove that all six endings are possible.)

EX. *2.2.* On the 7-by-5 field, as shown in Figures 1 and 2, prove that if only the upper middle square is initially empty, it is then impossible to end the game with but one piece on the field.

EX. *2.3.* Show that if Solitaire on any field is to end with but one piece on the field, then G,P,T must not all be of the same parity; and if one of these three is *exceptional*, by being of opposite parity to the other two, then the final piece must be on a square of exceptional color.

EX. *2.4.* In the lemma of **2.1**, is it correct to replace the word "difference" by the word "sum"?



▶ *Every word mathematicians use conveys a determinate idea and by accurate definitions they excite the same ideas in the mind of the reader that were in the mind of the writer . . . then they premise a few principles . . . and from these plain, simple principles they have raised most astonishing speculations.*

—JOHN ADAMS

CHAPTER 3*

MATHEMATICAL INDUCTION

3.1. The axiom of mathematical induction. One of the most useful and most powerful tools of mathematics is the principle or axiom of mathematical induction which is intimately related to the theory of numbers. Here is a proposition so basic that no proof is expected, but a proposition that we are willing to grant as an essential characteristic of the positive integers.

It will be assumed in the statement of the axiom and in the examples and exercises which follow that the student is familiar with the use of the symbol . . ., called the *ellipsis*, which when interposed between two numbers, either in a list of numbers or in an expression involving operations on the numbers, stands for *all numbers of the same type which intervene* between the two given numbers. Often it is necessary to give more than the end numbers, or to insert some formula, or to give a description in words so that it will be absolutely clear just what type of number is intended in the unwritten ellipsis.

One precise statement of the axiom of mathematical induction is as follows:

If a set M of positive integers is such that

(I) M contains the integer 1; and

*Chapter 3 is a basic chapter.

(II) on the assumption that M contains all the integers $1, 2, \ldots, n$, it can be proved that M contains the integer $n + 1$;

then the set M contains all positive integers.

It certainly seems that anyone who has considered the process of counting should be willing to grant that the set M described in the axiom is such that no positive integer is omitted from the set.

The uses of this axiom are manifold in all branches of mathematics. In a later chapter we shall indicate how the basic laws of addition and multiplication may be established with little more than mathematical induction as background; and in later problems we will become increasingly aware of the possibilities of using this scheme of proof. For the present, then, perhaps one simple example, discussed in detail, with suffice.

3.2. An example using mathematical induction. Consider the meaning of, and some way of establishing, the following formula:

$$1 + 3 + 5 + \ldots + (2n - 1) = n^2$$

where it is evident that the formula has meaning only for positive integral values of n. By substituting special values of n, say $n = 1$, 2, 3, we find that the formula states that

$$1 = 1^2, \quad 1 + 3 = 4 = 2^2, \quad 1 + 3 + 5 = 9 = 3^2,$$

in these respective cases. We begin to see that this formula is not like the usual function notation, at least on the left side, for the left side continually changes form as n changes. In words the proposition seems to be as follows: "The sum of the first n odd numbers is equal to the square of the integer n." But the formula is apparently an infinity of formulas, changing as n changes, and it is evidently hopeless to prove such a proposition, in the way that laboratory sciences "prove laws," by checking the first thousand cases. What we need is the definite procedure of mathematical induction by which the complete proof for the infinity of cases is made to depend upon just *two* steps, namely: (I) the usually very easy task of checking the formula in the case $n = 1$ (or sometimes the first positive integer for which the formula has meaning is the integer that is tested here); this may be called the *basis* for the induction; and (II) the making of the *hypothesis* H that the formula is correct in all the cases $1,2,\ldots,n$, and then the *proving* that, as a consequence of the hypothesis H and previously known theorems, the formula is correct in the case $n + 1$; this step may be called the *core of the induction proof.* Finally, if (I) *and* (II) have been established, we can apply the axiom of mathematical induction and make the *conclusion* that the formula under consideration is true for *all* positive integers (or for all positive integers beginning with the smallest integer than can be used in step (I)).

In the example proposed above the complete proof by induction should read somewhat as follows:

Problem: Prove that $1 + 3 + 5 + \ldots + (2n - 1) = n^2$.

Proof: We shall use an induction proof on n.

(I) When $n = 1$, the formula is true because $1 = 1^2$.

(II) We make the hypothesis H that the formula is correct in each of the cases $1,2,\ldots,n$; in particular, this includes the case n so that we are assuming that $1 + 3 + 5 + \ldots + (2n - 1) = n^2$. Let us then consider the next case where n is replaced by $n + 1$. On the left side of the formula *one more summand* will appear, and since $2(n + 1) - 1 = 2n + 1$ is the next odd number following $2n - 1$, we find that we have to consider the following sum:
$1 + 3 + 5 + \ldots + (2n - 1) + (2n + 1) = n^2 + (2n + 1) = (n + 1)^2$. The first equality is justified by the hypothesis H, the second equality is justified by a well-known factoring formula, and reading from first to last we find that we have established the truth of the formula in the case $n + 1$.

From (I), (II), and the principle of mathematical induction, the given formula is correct for *all* positive integers n.

3.3. Words of caution. In the application of the principle of mathematical induction the student must be careful to establish *both* (I) and (II). The situation has been likened to that which arises in the children's game of lining up toy soldiers (or cards, or dominoes) so that if one falls he will knock over the next. If either no soldier is pushed over (so that (I) fails) or if some soldier is A.W.O.L. (so that (II) fails), then the complete line will not fall.

For example, (I) can be established for the *false* formula

$$1 + 3 + 5 + \ldots + (2n - 1) = n^3 - 5n^2 + 11n - 6$$

using either $n = 1$, 2, or 3; but (II) *cannot* be established for this formula. In fact, a *false* formula, correct for the first thousand cases (!), but not correct thereafter, is easily given, as follows:

$$1+3+\ldots+(2n-1)=n^2+(n-1)(n-2)\ldots(n-999)(n-1000).$$

As another example, we can establish (II) for the *false* formula

$$1+3+\ldots+(2n-1)=n^2+5,$$

i.e., *if* the formula were true in the cases $1,2,\ldots,n$, we could prove it true in the case $n+1$, yet this formula is not correct for any value of n, so (I) fails to be true in a "big" way.

In carrying out the step (II), which is the core of the proof and sometimes difficult, the student should be on the alert for some way of rewriting the formula in the case $n+1$ so as to involve one, or even several, of the formulas for the cases $1,2,\ldots,n$, so that the hypothesis H can be actively employed, for surely one will have to use the hypothesis H in some way before being able to complete step (II).

Finally, we should note that this principle of induction is primarily a method of proof for a known or suspected formula, and it is not in itself a tool for discovering such formulas. Thus in the problems that close this lesson, some of the fun of mathematical investigation is lost because the formulas are stated without asking the student to uncover them for himself. But until the principle of mathematical induction is completely mastered there is not much point in guessing at formulas that one cannot rigorously establish.

EXERCISES

By mathematical induction establish the following formulas:

EX. *3.1.* $1+2+3+\ldots+n=n(n+1)/2$.

EX. *3.2.* $(x-1)(1+x+x^2+\ldots+x^n)=x^{n+1}-1$.

EX. *3.3.* $1^2+2^2+3^2+\ldots+n^2=n(n+1)(2n+1)/6$.

EX. *3.4.* $1^2+3^2+5^2+\ldots+(2n-1)^2=(4n^3-n)/3$.

EX. *3.5.* $1^3+2^3+3^3+\ldots+n^3=n^2(n+1)^2/4$. (Compare with EX. *3.1.*)

EX. *3.6.* Using EX. *3.5* derive a formula for $1^3+3^3+5^3+\ldots+(2k-1)^3$.

EX. *3.7* Use the abbreviation $r!$, read "r-factorial," to mean $r!=1\cdot2\cdot3\ldots r$

when $r > 0$; and define $0! = 1$. Define $\binom{n}{r} = n!/r!(n-r)!$ for $0 \leqq r \leqq n$. Use mathematical induction on n to establish the "binomial theorem" which follows:

$$(x+y)^n = \binom{n}{n}x^ny^0 + \binom{n}{n-1}x^{n-1}y^1 + \dots$$
$$+ \binom{n}{r}x^ry^{n-r} + \dots + \binom{n}{0}x^0y^n.$$

EX. *3.8.* Define $S(k,n) = 1^k + 2^k + 3^k + \dots + n^k$. For a fixed positive integer k, use mathematical induction on n to prove that

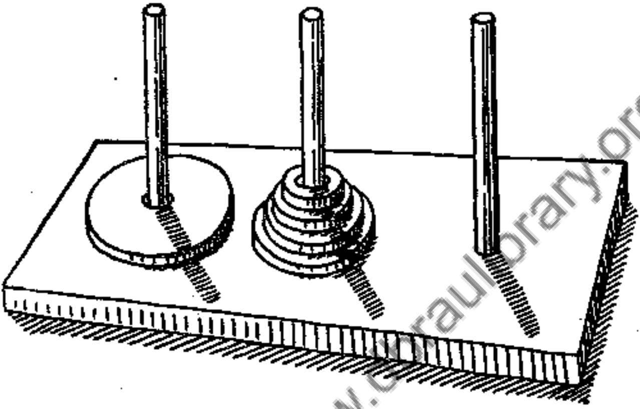
$$\binom{k+1}{k}S(k,n) + \dots + \binom{k+1}{2}S(2,n) + \binom{k+1}{1}S(1,n) =$$
$$(n+1)^{k+1} - (n+1).$$

(*Hint:* Make good use of EX. *3.7* with special values of x and y.)

EX. *3.9.* Use the recursion formula given in EX. *3.8* to compute, in succession, $S(1,n)$, $S(2,n)$, $S(3,n)$, and $S(4,n)$.

EX. *3.10.* It is said that at the time of the creation, there were placed in one of those incredible temples at Hanoi in Indo-China some 64 golden washers or disks, no two the same in size, all set on one of three golden needles; and the priesthood of the temple were set busy moving the disks, one at a time, to *any one* of the needles, subject always to the condition, which held also at the outset, that no disk be placed above a smaller disk. The needles were farther apart than the outer diameter of the largest disk. The priests were to aim at arranging all the disks on *another one* of the needles and they were pledged both to move one disk every minute and to make their moves so that the goal would be achieved in the least number of moves. When the appointed task was completed, there would come the day of doom for many, but of reward for the faithful. Naturally, some of the unfaithful, as they watched the ceaseless activity at the temple, the shifts by night and day, and the sage noddings of the wise men who directed the laborers, were much concerned as to just how soon to expect the judgment day. We could have aided them considerably, for by assuming that there are n disks in the problem, we can show by mathematical induction that the minimum number of moves is given by $2^n - 1$. And a few minutes of translating $2^{64} - 1$ minutes into years will bring considerable comfort to the most unfaithful. (E. Lucas.)

EX. *3.11.* One worshipper at the temple of Hanoi (see EX. *3.10*) suggested that it would be easier for the priests if they arranged the three needles in a row and limited themselves to moving each disk to an *adjacent* needle. But it was discovered that the suggestion was influenced by a mundane sect of mathematical inductors who had discovered that to move n disks *from one end needle to the other end needle*, under this new restriction, would require $3^n - 1$ moves.



▶*Arithmetic has a very great and compelling effect, elevating the soul to reason about abstract number, and if visible or tangible objects are obtruding upon the argument, refusing to be satisfied.* —PLATO

CHAPTER 4[*]

REPRESENTATION OF THE INTEGERS

4.1. Representation with the base b. It is clear that if we can find some convenient way of representing any positive integer a, then the symbol $-a$ can be used for the companion negative integer, and the symbol 0 can be used for the zero, and we will then have a way of representing all the integers—positive, negative, and zero.

Our scheme of representation requires us to use exponents, so it would be well for the student to review the basic definitions and rules for exponents and, in particular, to recall that it is convenient to define $b^0 = 1$ when $b \neq 0$.

Representation theorem: If b is a fixed positive integer with $1 < b$, then for any given positive integer a, a non-negative integer n can be found and a set of $n + 1$ integers: $a_0,a_1,\ldots,a_n$, such that a may be represented uniquely in the following form:

$$a = a_0 + a_1b + a_2b^2 + \ldots + a_nb^n$$

with $0 \leqq a_i < b$ for $i \neq n$ and $0 < a_n < b$.

Proof: (A) We show that at least one representation is possible,

[*] Chapter 4 is a supplementary chapter, but easy and something of a diversion. The ideas are used somewhat in later chapters, particularly Chapter 10.

using mathematical induction on a. Let M be the set of all positive integers a for which the theorem holds.

(I) M contains 1, for we may take $n = 0$ and $a_0 = 1$ and satisfy $0 < a_n < b$ inasmuch as a major premise of the theorem is that $1 < b$.

(II) If we assume that M contains all the integers $1,2,\ldots,a$, then we can prove that M contains $a + 1$. Since $1 < b$, it follows that $1 < b < b^2 < b^3 < \ldots$, and hence there exists an integer n with $0 \leqq n$ such that $b^n \leqq a + 1 < b^{n+1}$. By the division algorithm there exist integers a_n and r such that $a + 1 = a_nb^n + r$ with $0 \leqq r < b^n$. Here $0 < a_n$, because $0 = b^n - b^n < a + 1 - r = a_nb^n$; and $a_n < b$, because $a_nb^n \leqq a + 1 < b^{n+1}$. Either we have $r = 0$, so that $a + 1 = a_nb^n$ with $0 \leqq n$ and with $0 = a_i$ for $i \neq n$ and $0 < a_n < b$, completing the proof for this case; or we have $0 < r$, but since we also have the fact that $r < b^n \leqq a + 1$, it follows that in this case the induction hypothesis may be applied to r and we may write $r = a_0 + a_1b + \ldots + a_kb^k$ for some $k \geqq 0$, with $0 \leqq a_i < b$ when $i \neq k$ and with $0 < a_k < b$. Since $b^k \leqq a_kb^k \leqq r < b^n$, we find $k < n$. Hence

$$a + 1 = a_0 + a_1b + \ldots + a_kb^k + (0)b^{k+1} + \ldots + (0)b^{n-1} + a_nb^n$$

is a representation of the desired type for $a + 1$, completing the proof for this case.

By (I), and (II), and the principle of mathematical induction we draw the conclusion that M contains all the positive integers and the proof of part (A) is complete.

(B) We may show the representation to be unique by showing how two different representations would lead to a contradiction. Suppose that $a = a_0 + a_1b + \ldots + a_nb^n = c_0 + c_1b + \ldots + c_tb^t$ with $0 \leqq a_i < b$ when $i \neq n$ and $0 < a_n < b$, and $0 \leqq c_i < b$ when $i \neq t$ and $0 < c_t < b$. Unless $n = t$ and $a_i = c_i$ for $i = 0,1,\ldots,n$, it follows by subtraction that $0 = d_0 + d_1b + \ldots + d_kb^k$ with $k > 0$ and with $d_i = a_i - c_i$ and $-b < d_i < b$ for $i = 0,1,\ldots,k$, and $d_k \neq 0$. Since $d_k \neq 0$, there is a smallest subscript $i < k$, such that $d_i \neq 0$.

Then from $\quad 0 = d_ib^i + d_{i+1}b^{i+1} + \ldots + d_kb^k$
we find that $\quad d_i = -b(d_{i+1} + \ldots + d_kb^{k-i-1})$
so that d_i is a multiple of b. But $|d_i| < b$, hence $d_i = 0$. This contradiction establishes the uniqueness of the representation.

This completes the proof of the theorem.

4.2. The choice of a base b. The important implication of the theorem presented in the preceding section is that in addition to the symbol 0 only $b - 1$ other symbols are required (allowing repetitions, of course) to represent *any* integer a, a total of b symbols! Anthropologically and geographically and historically speaking, by far the most important choice of the base b is the choice which we call "ten," motivated, it is sure, by the usual supply of fingers among the primates. Here, modified from Hindu-Arabic sources, the set of b symbols is as follows:

$$0,\quad 1,\quad 1+1=2,\quad 2+1=3,\quad 3+1=4,\quad 4+1=5,$$
$$5+1=6,\quad 6+1=7,\quad 7+1=8,\quad 8+1=9.$$

The next following integer is $9 + 1 = b = 0 + (1)b$ with $n = 1$.

But our theorem shows that this customary choice of the base b is by no means necessary, and indeed the history of number representation among various peoples in various parts of the world at various times in history reveals uses of the bases "five," "twenty," "sixty," and several others.

Any departure from the usual representation cannot result in any really different theorems about the integers. However, it is still true that for certain problems the choice of some other base than "ten" may make the proof of a theorem shorter or more easily understood.

For example, using the notation of the preceding section, we may desire to reduce the value of n, and this may be accomplished by increasing b; of course, this would require additional symbols. Quite widely advocated is the adoption of the "dozen" or "twelve" system, wherein we might write $9 + 1 = x$, $x + 1 = L$, and $L + 1 = b$.

On the other hand, we may desire to reduce the number of symbols needed, even at the sacrifice of using larger values of n. The extreme case, in this direction, is the "binary" or "two" system, which has been especially useful to mathematicians and designers of some types of computing machines; in this system the only symbols required are 0 and 1, since $1 + 1 = b$.

To avoid writing the powers of b and the $+$ signs we shall agree to adopt the following *positional notation*:

$$a = a_0 + a_1b + \ldots + a_nb^n = (a_n \ldots a_1a_0)_b$$

where a_i occurs in the $i + 1$ position, counting from the *right*. The only danger is that this symbol may be interpreted as the product of the coefficients, so be wary! Of course the b-subscript specifying the base may be dropped if the context indicates the value of b.

4.3. Representation with the base "six." Had we been born in a certain mountain village on the border between France and Spain where, so the anthropologists tell us, inbreeding has resulted in a whole community of people with six fingers on each hand, it is conceivable that in kindergarten we might have learned to count in terms of "sixes," i.e., 0, 1, 2, 3, 4, 5, then "six" which in the positional notation would be written 10, followed by 11, 12, 13, 14, 15, then 20, 21, etc. We would have learned "addition" and "multiplication" tables like the following:

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	10
2	2	3	4	5	10	11
3	3	4	5	10	11	12
4	4	5	10	11	12	13
5	5	10	11	12	13	14

×	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	10	12	14
3	0	3	10	13	20	23
4	0	4	12	20	24	32
5	0	5	14	23	32	41

Having once learned these tables we could have progressed easily to more complicated arithmetic—speaking of unit digits, b-digits, b^2-digits, etc., carrying, borrowing, and arranging our harder multiplications, for example, by virtue of the associative and distributive laws, in rows and columns. For example, to multiple 3204 by 513 we would write

$$\begin{array}{r} 3204 \\ 513 \\ \hline 14020 \\ 3204 \\ 24432 \\ \hline 2533300. \end{array}$$

But, having learned as we did, we will probably find the last example more convincing, or at least feel that it has been properly checked, if we first convert the multiplicand and multiplier to the base "ten" as follows:

$$(3204)_6 = 4 + 0\cdot 6 + 2\cdot 6^2 + 3\cdot 6^3 = (724)_{10};$$
$$(513)_6 = 3 + 6 + 5\cdot 6^2 = (189)_{10};$$

and then carry out the problem as follows:

$$\begin{array}{r} 724 \\ 189 \\ \hline 6516 \\ 5792 \\ 724 \\ \hline 136836. \end{array}$$

Our answer of $(136836)_{10}$ we convert to a representation in terms of the base 6 by repeated applications of the division algorithm with the divisor always 6, as follows:

$$\begin{array}{rrl} & 136836 & \\ \text{quotients} \left\{ \begin{array}{r} 22806 \\ 3801 \\ 633 \\ 105 \\ 17 \\ 2 \\ 0 \end{array} \right. & & \left. \begin{array}{r} 0 \\ 0 \\ 3 \\ 3 \\ 3 \\ 5 \\ 2 \end{array} \right\} \text{remainders.} \end{array}$$

Hence we find $(136836)_{10} = (2533300)_6$ and this checks our previous work in the "six" system.

To justify the last procedure for taking a number given in a known system, say with the base 10, and converting it, by repeated divisions in the *known* system, to a representation with a different base b, perhaps b equal to 6, we observe that if $a = a_0 + a_1b + \ldots + a_nb^n$, then $a = (a_1 + a_2b + \ldots + a_nb^{n-1})b + a_0$; and since $0 \leqq a_0 < b$, we see that the division algorithm $a = qb + r$, $0 \leqq r < b$ must yield $r = a_0$ and $q = a_1 + a_2b + \ldots + a_nb^{n-1}$. Similarly, $q = q_1b + r_1$ yields $r_1 = a_1$; $q_1 = q_2b + r_2$ yields $r_2 = a_2$; etc.

EXERCISES

In the following five exercises (1) carry out the indicated operations, using the tables given in 4.3, entirely within the "six" system; then (2) convert the given numbers to the base "ten" and carry out the operations in the familiar ten system; (3) convert the answers obtained in step (2) to the base "six"; and (4) check the results obtained in step (3) with those obtained in (1).

EX. *4.1.* Add $(3542)_6$ to $(1135)_6$.

EX. *4.2.* Subtract $(3025)_6$ from $(11111)_6$.

EX. *4.3.* Multiply $(234)_6$ by $(531)_6$.

EX. *4.4.* Divide $a = (3014)_6$ by $c = (12)_6$ to find q and r so that $a = qc + r$, $0 \leqq r < c$.

EX. *4.5.* Use the square root extraction process to find the square root of $(24041)_6$.

EX. *4.6.* Use the representation of integers in the "binary" or "two" system to show that a properly chosen set of n weights will suffice to weigh objects of weights $1,2,3,\ldots,2^n - 1$, if the weights are put in one pan of a balance and the object to be weighed in the other pan.

EX. *4.7.* If the base $b = 2K + 1$ with $K > 0$, discuss the possibility of representing any positive integer a in the form $a = a_0 + a_1b + \ldots + a_nb^n$ with $-K \leqq a_i \leqq K$, $i \neq n$, and $0 < a_n \leqq K$.

> ▶*Euclid is the only man to whom there ever came, or can ever come again, the glory of having successfully incorporated in his own writings all the essential parts of the accumulated mathematical knowledge of his time.*
>
> —D. E. SMITH

CHAPTER 5*

THE EUCLID ALGORITHM

5.1. Classification of the integers by divisibility properties. If $c = ab$, then c is called a *multiple* of b, and b, a *divisor* or *factor* of c. The *zero* is exceptional from this point of view since 0 is a multiple of every integer.

If $ab = 1$, then b is a *unit*. The only units among the ordinary integers are $+1$ and -1.

If p is not a unit and if $p = ab$ implies that either a or b must be a unit, then p is a *prime*. The first five prime integers are 2,3,5,7, and 11.

An integer which is neither zero, a unit, nor a prime, is said to be *composite*. The first five composite integers are 4,6,8,9, and 10.

Thus on the basis of rather simple divisibility properties the integers are separated into four mutually exclusive categories. It is the purpose of this and the following chapter to show that the primes are fundamental building blocks in terms of which all the composite integers may be conveniently and uniquely represented. This program will be initiated by studying two ideas that are very useful in themselves and which provide the key ideas for the next chapter, where our program will be finally achieved.

*Chapter 5 is a basic chapter.

5.2. Definition of a greatest common divisor. If $a = Ad$ and $b = Bd$, then d is called a *common divisor* of a and b.

Given the integers a and b, if there exists an integer d such that

(1) d is a common divisor of a and b;

(2) every common divisor of a and b is a divisor of d;

then d is called a *greatest common divisor* of a and b, and is designated by $d = (a,b)$.

The student reader will probably consider himself already acquainted with this idea, since without much effort he can recognize that $(15,21) = 3$. Upon greater consideration he can see that his success depends on factoring the numbers into prime factors: $15 = 3\cdot 5$ and $21 = 3\cdot 7$ and then selecting the correct powers of common prime factors. In this simple case the selection is easy and (as we shall show later) the factorizations are unique and 3 is indeed a correct greatest common divisor. The remarkable thing about the argument of the next section is that it in no way depends upon factoring a and b into prime factors (a task which may be formidable for large numbers), and yet it proves the existence of, and provides a direct construction for, a greatest common divisor.

5.3. The Euclid algorithm for finding a greatest common divisor. If we are given a pair of positive integers a and b, it is merely a matter of notation to assume $0 < b \leqq a$. By the division algorithm we may write $a = qb + r$ with $0 \leqq r < b$. If $r = 0$, we stop with the one equation $a = qb$. In the more important case where $r \neq 0$, we apply the division algorithm repeatedly, say $k + 2$ times, to obtain the following sequence of equations which is universally known as the *Euclid algorithm:*

$$a = qb + r, \quad b = q_1r + r_1, \quad r = q_2r_1 + r_2, \quad \ldots,$$

$$r_{k-2} = q_kr_{k-1} + r_k, \quad r_{k-1} = q_{k+1}r_k + 0,$$

$$b > r > r_1 > r_2 > \ldots > r_k > r_{k+1} = 0.$$

Since the remainders form a decreasing sequence of positive integers, it follows that in a finite number of steps the process will terminate, say as indicated with $r_{k+1} = 0$. It is understood, of course, that k will vary depending on a and b.

We are now in a position to discuss in a constructive way the problem of the existence of a greatest common divisor.

Theorem: For any given pair of positive integers a and b, a greatest common divisor $d = (a,b)$

(A) exists,

(B) is unique except for a unit factor,

(C) is such that there exist integers x and y for which $d = ax + by$,

(D) is such that the integers d, x, y can be found in a finite number of steps by the Euclid algorithm.

Proof: (A) We consider the Euclid algorithm described above. If $r = 0$, so that the algorithm consists of the one equation $a = qb$, it is clear that b satisfies both requirements (*1*) and (*2*) of the definition of a greatest common divisor, so we take $d = b$. If $r \neq 0$, we shall show that r_k, the last non-zero remainder in the Euclid algorithm, is a greatest common divisor of a and b (in case $r_1 = 0$, we agree to define $r_0 = r$). We must show that r_k possesses properties (*1*) and (*2*) of the definition given in the preceding section.

Proof of (1): From $r_{k-1} = q_{k+1}r_k$ it follows that r_k divides r_{k-1}. Then from $r_{k-2} = q_k r_{k-1} + r_k$ it follows that r_k divides r_{k-2}. And tracing *back*, equation by equation in the algorithm, we find that r_k divides b and finally that r_k divides a. Thus r_k is a common divisor of a and b.

Proof of (2): Let D be any common divisor of a and b. We rearrange the first equation of the Euclid algorithm to see from $r = a - qb$ that D divides r. Then the rearranged second equation $r_1 = b - q_1 r$ shows that D divides r_1. And moving *forward*, equation by equation in the algorithm, we find finally from $r_k = r_{k-2} - q_k r_{k-1}$ that D divides r_k. Thus every common divisor of a and b divides r_k.

Since we have shown that r_k possesses properties (*1*) and (*2*), it follows from the definition of d that $r_k = d$. Hence at least one integer $d = (a,b)$ exists.

(B) Let d and d' be two greatest common divisors of a and b. By property (*2*) it follows on the one hand that $d = kd'$ and on the other hand that $d' = md$. Then $d = kmd$ and since $d \neq 0$, it follows that $km = 1$, hence k and m are units. Conversely, if m is a unit and d is a greatest common divisor of a and b, then $d' = md$ is also a greatest common divisor of a and b (see EX. *5.1*).

Thus the greatest common divisor is unique only up to a unit factor, and $-d$ has to be considered as equally acceptable an answer as d. In number systems which we shall investigate later, where still more units are available, even greater freedom in the selection of $d = (a,b)$ is to be expected. The use of the adjective "greatest" is merely a hangover from the simplest case where one considers only positive integers.

(C) and (D) Furthermore, if $r \neq 0$, by straightforward (even if lengthy) successive eliminations of $r_{k-1}, r_{k-2}, \ldots, r_1, r$ from the system of equations of the algorithm, beginning with

$$r_k = r_{k-2} - q_k r_{k-1} = (r_{k-4} - q_{k-2}r_{k-3}) - q_k(r_{k-3} - q_{k-1}r_{k-2}),$$

etc., etc., we discover, in a finite number of steps, suitable integers x and y such that $r_k = ax + by$. In case $r = 0$, then $d = b$, so we can take $x = 0$ and $y = 1$ to have $d = ax + by$.

This completes the proof of the theorem.

Before presenting an example to illustrate the theorem, we note that if a and b are positive and $d = (a,b)$, then $d = (-a,b)$ and $d = (-a,-b)$, because the divisors of a and $-a$ are the same. If $a \neq 0$, then $a = (a,0)$. Hence the symbol (a,b) is meaningful in every case except $(0,0)$ where it is certainly meaningless, since, as was noted at the beginning of this chapter, every integer is a divisor of 0.

For the example $a = 2210$ and $b = 493$, the Euclid algorithm may be written as follows:

$$\begin{array}{rrrl}
 & 4 = q & & \\
b = 493 & 2210 = a & & \\
 & 1972 & 2 = q_1 & \\
r = & 238 & 493 & \\
 & & 476 & 14 = q_2 \\
r_1 = & & 17 & 238 \\
 & & & 238 \\
r_2 = & & & 0.
\end{array}$$

Since $r_2 = 0$, we know that $r_1 = 17 = (2210,493)$. To find x and y we have only to eliminate r from the equations, as follows:

$$17 = b - 2r = b - 2(a - 4b) = -2a + 9b,$$

hence we may take $x = -2$ and $y = +9$. In performing such an elimination it is convenient to retain letters for the r's, a, and b, and to substitute actual numbers only for the q's.

In closing this lesson we mention that if the greatest common divisor of a and b is a unit, $(a,b) = \pm 1$, then a and b are said to be *relatively prime.* Thus 15 and 8 are relatively prime, although neither is a prime.

EXERCISES

EX. *5.1.* If $d = (a,b)$ and m is a unit, show that $d' = md$ is also a greatest common divisor of a and b.

EX. *5.2.* Use the Euclid algorithm to find $d = (11951,11063)$ and to find x and y such that $d = 11951x + 11063y$.

EX. *5.3.* Prove that p and q are relatively prime if and only if there exist integers s and t such that $1 = ps + qt$.

EX. *5.4.* If $d = (a,b)$, $a = Ad$, $b = Bd$, prove that A and B are relatively prime, using EX. *5.3.*

EX. *5.5.* If $d = (a,b) = ax + by$, prove that x and y are relatively prime, using EX. 5.3.

EX. *5.6.* If $(a,b) = 1$ and $(a,c) = 1$, show that $(a,bc) = 1$, using EX. *5.3.*

EX. *5.7.* If $(a,b) = 1$, show that $(a^s,b^t) = 1$, using EX. *5.6.*

EX. *5.8.* If $(a,b) = 1$, show that $(a + ub,b) = 1$ for any integer u.

EX. *5.9.* If $(a,b) = 1$, show that $(a + b, a - b) = 2$, or 1, according as a and b are of the same, or opposite, parity.

EX. *5.10.* Prove that $(ka,kb) = k(a,b)$, for any integer $k \neq 0$.

EX. *5.11.* State the definition for $d = (a,b,c)$, extending that given in **5.2**, and prove that $d = (a,b,c) = ((a,b),c)$.

EX. *5.12.* If $(a,b,c) = \pm 1$, then a,b,c are said to form a relatively prime triple. Prove by examples that a relatively prime triple $(a,b,c) = 1$ can occur with none, one, two, or three of the pairs a,b or b,c or c,a being relatively prime pairs.

EX. *5.13.* Give a new proof of EX. *5.7* using EX. *5.3* and the binomial theorem.

► *In the beginning everything is self-evident, and it is hard to see whether one self-evident proposition follows from another or not. Obviousness is always the enemy to correctness. Hence we must invent new and even difficult symbolism in which nothing is obvious.* —BERTRAND RUSSELL

CHAPTER 6*

THE FUNDAMENTAL THEOREM OF ARITHMETIC

6.1. The fundamental lemma. We lean heavily on the results of the preceding chapter to establish the following basic lemma:

Fundamental lemma: If a prime p divides a product ab, then p must divide at least one of the integers a or b.

Proof: Suppose p divides b, then the lemma is true. Next, suppose p does not divide b; then $(p,b) = 1$, because the only divisors of the prime p are $+p$, $-p$, $+1$, -1. Hence by the theorem of the preceding lesson there exist integers x and y such that $1 = bx + py$. Multiplying by a, we find $a = abx + apy$. Since by hypothesis p divides ab and since obviously p divides p, it follows from the last equation that p divides a, which completes the proof.

Corollary: If a prime p divides a product $a_1a_2 \ldots a_n$, then p must divide at least one of the factors $a_1, a_2, \ldots, a_n$.

The proof of the corollary is left as one of the exercises.

6.2. The fundamental theorem. We are now in a position to

*Chapter 6 is a basic lesson.

establish what is justly described as the fundamental theorem of the arithmetic of ordinary integers.

The fundamental theorem of arithmetic: Any given positive integer n, other than 1, can be written uniquely as follows:

$$n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k}$$

where k is a positive integer, where each p_i is a prime integer, where each a_i is a positive integer, and where $1 < p_1 < p_2 < \ldots < p_k$.

(It is understood that the choice of k and the p_i and the a_i will vary with different n. We shall refer to this representation as "writing n in standard form.")

Proof: (A) We shall show that there exists at least one such representation by making an induction argument on n. Let M be the set of all positive integers $n \geqq 2$ for which the theorem holds.

(I) M contains 2, for 2 is itself a prime.

(II) Suppose that M contains the integers $2,3,\ldots,n$. Then we can show that M must contain $n + 1$. If $n + 1$ is a prime, then the desired representation is already found. If $n + 1$ is composite, then $n + 1 = bc$ with $1 < b$ and $1 < c$, hence with $c < bc$ and $b < bc$. Thus since $1 < b < n + 1$ and $1 < c < n + 1$, it follows that the induction hypothesis applies to both b and c. By combining the representations for b and c, grouping like primes together and rearranging the combined set of primes in natural order with new labels, if necessary, we arrive at a representation of $n + 1$ of the desired form. By (I), (II), and the principle of mathematical induction, it follows that M contains all positive integers $n \geqq 2$.

(B) Suppose that there exist two standard representations for a given integer n, say

$$n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k} = q_1^{b_1}q_2^{b_2}\ldots q_m^{b_m}$$

where the p_i and q_i are primes and $1 < p_1 < p_2 < \ldots < p_k$ and $1 < q_1 < q_2 < \ldots < q_m$. It will be no essential restriction to suppose $m \geqq k$. By the fundamental lemma and corollary of the preceding section it follows that the *prime* p_1 must divide some factor q_i, and since q_i is itself a *prime* that $p_1 = q_i$; but $q_i \geqq q_1$, hence $p_1 \geqq q_1$. But similarly, the corollary shows that the prime q_1 must divide some factor p_j, and since p_j is a prime, it follows that $q_1 = p_j$; however, $p_j \geqq p_1$, hence $q_1 \geqq p_1$. Thus it now follows that $p_1 = q_1$.

Now suppose that $b_1 \geqq a_1$, then $p_1^{a_1} = q_1^{a_1}$ may be divided out of the equation which we are studying to leave the following equation:

$$p_2^{a_2} \ldots p_k^{a_k} = q_1^{b_1 - a_1} q_2^{b_2} \ldots q_m^{b_m}.$$

If $b_1 > a_1$, the prime q_1, by the same arguments as before, must equal some p_j, $j \geqq 2$; but since $q_1 = p_1 < p_j$, $j \geqq 2$, we have arrived at a contradiction; therefore $b_1 = a_1$. A similar argument suffices if we suppose initially that $a_1 \geqq b_1$.

Repeating this same kind of argument, we are step by step led to the following conclusions: $p_2 = q_2$, $a_2 = b_2$; $p_3 = q_3$, $a_3 = b_3$; $\ldots$; $p_k = q_k$, $a_k = b_k$. At this stage the equation being studied reduces (in case $m > k$) to the following:

$$1 = q_{k+1}^{b_{k+1}} \ldots q_m^{b_m}$$

but this is a contradiction, since a prime q is not a divisor of 1. Hence $m = k$, and the proof of the uniqueness of the representation is complete.

In many texts the fundamental theorem is stated in this way: "Every positive integer, except 1, can be represented uniquely as a product of primes, *except for order*." By making the rather natural agreement to collect like primes and to arrange the primes in ascending order, we have replaced the italicized phrase by the condition $1 < p_1 < p_2 < \ldots < p_k$.

6.3. Critique. The theorem in **6.2** is called the fundamental theorem of arithmetic because in the further study of the theory of numbers, we use this unique factorization at almost every stage of the development. The need for presenting a proof of this theorem, however obvious the result may seem, will be apparent in our last chapter where we shall describe, briefly, systems of *algebraic integers* (our present system of natural integers being a special case) which contain a zero, units, primes, and composite integers; yet in some of these systems the fundamental theorem fails. In the higher algebra courses one of the chief concerns is to provide a remedy for this anomaly. One reason for the failure is that in some of these systems the fundamental lemma of **6.1** is lacking; since that fundamental lemma depended in its turn on the concept of a greatest common divisor introduced in Chapter 5, it may not be too surprising to learn that the remedy which is applied in the higher courses is to supply a suitable generalization of the notion of a greatest common divisor. It is true that Zermelo has shown that the fundamental theorem of **6.2** for the natural integers can be proved by an induction argument without using the lemma of **6.1** and without using the greatest common divisor theorem; but we have preferred to present here the traditional order of proof because it does suggest in a better way what later generalizations should be made.

However, proceeding from Zermelo's proof, or working back from our proof above of the fundamental theorem, we discover that because of the unique factorization all the possible divisors of a number are immediately obtainable from the standard form. For if

$$n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k},$$

then every possible positive divisor s of n is obtained by considering

$$s = p_1^{b_1}p_2^{b_2}\ldots p_k^{b_k} \quad \text{where } 0 \leqq b_i \leqq a_i.$$

Thus the greatest common divisor of two given integers can be found by writing each of these integers in standard form, selecting those primes which are *common* factors, say $P_1, P_2, \ldots, P_t$, and forming $d = P_1^{m_1}P_2^{m_2}\ldots P_t^{m_t}$, where m_i is the *minimum* exponent of P_i as one compares the exponents of P_i in the two given integers.

For example, if $a = 2520 = 2^3 \cdot 3^2 \cdot 5 \cdot 7$ and $b = 4950 = 2 \cdot 3^2 \cdot 5^2 \cdot 11$, then $d = (a,b) = 2 \cdot 3^2 \cdot 5 = 90$.

Theoretically this construction of d is very easy, but practically it depends upon finding the standard representation, and as we shall show in the next chapter this assignment may be very difficult. Hence, as we have tried to emphasize earlier, the Euclid algorithm for finding d is, in general, to be preferred, for it avoids completely the question of finding prime factors.

6.4. Least common multiple. If $m = qa = rb$, then m is called a *common multiple* of a and b.

If *(1)* m is a common multiple of a and b; and

(2) m is a divisor of every common multiple of a and b;

then m is called a *least common multiple* of a and b, and is designated by $m = [a,b]$.

(These definitions should be compared carefully with those in **5.2** in order to appreciate their "dual" nature.)

Directly from the definitions and from the notion of unique factorization into primes, it follows that if a and b are written in standard form, then the standard form for m is

$$m = Q_1^{M_1}Q_2^{M_2}\ldots Q_u^{M_u}$$

where the Q's include *all* prime factors of both a and b and where M_i is the *maximum* exponent of Q_i as one compares the exponents of Q_i in a and b.

For example, referring to the factorizations in the example of **6.3**, we see that $[2520,4950] = 2^3 \cdot 3^2 \cdot 5^2 \cdot 7 \cdot 11 = 138600$.

This method of constructing $[a,b]$ depends on finding prime factors. A way of avoiding this difficulty is suggested by the identity in EX. 6.5.

EXERCISES

EX. *6.1.* Prove the corollary in **6.1**, using the lemma in **6.1** and induction on n.

EX. *6.2.* Using the same type of argument as in the proof of the lemma in **6.1**, show that if $(a,b) = 1$ and a divides bc, then a must divide c.

EX. *6.3.* Find the standard representations for $a = 2625$ and $b = 24633$.

EX. *6.4.* Find $m = [2625,24633]$.

EX. *6.5.* If $d = (a,b)$ and $m = [a,b]$, prove that $md = ab$, using the standard forms suggested in **6.3** and **6.4**.

EX. *6.6.* Give a proof that $md = ab$, using EX. *5.4* and EX. *6.2*.

EX. *6.7.* Using the identity given in EX. *6.5* and EX. *6.6*, describe a method for finding $m = [a,b]$, not depending upon finding the prime factors of a and b.

EX. *6.8.* Prove that $[a,b] = ab$ if and only if a and b are relatively prime.

EX. *6.9.* Show that $m = [a,b]$ is unique only up to a unit factor.

EX. *6.10.* Prove that $[ka,kb] = k[a,b]$.

EX. *6.11.* Extend the definition of **6.4** to the case of the least common multiple $m = [a,b,c]$ of three given integers and prove that $m = [a,b,c] = [[a,b],c]$.

EX. *6.12.* Produce examples to show that $(a,b,c)[a,b,c]$ can be less than or equal to abc.

EX. *6.13.* Show that $(a,b,c)[a,b,c]$ cannot be greater than abc.

EX. *6.14.* Show that $(a,b,c,)\ [a,b,c,] = abc$ if and only if a,b,c are relatively prime *in pairs*.

▶*For an easy way to reach the mountain top, many a traveller buys his ticket and takes the funicular. But some like a stiff climb over rocks and across streams, and such an ascent has its advantages if the heart is good and the muscles are strong.*

—W. F. OSGOOD

CHAPTER 7*

PRIME AND COMPOSITE INTEGERS

7.1. Some questions. Motivated by the fundamental theorem discussed in the previous lesson, it is natural for us to ask questions like the following:

(*1*) How can one prepare a list of prime and composite integers which are $\leq n$, where n is a given integer?

(*2*) How can one determine whether a given integer n is prime or composite?

(*3*) Are there infinitely many distinct primes?

The answer to the last question being "Yes," we then ask:

(*4*) Is it possible to give a formula for the nth prime?

(5) Is it possible to find a polynomial $f(x)$ which will represent only primes for all integral values of x?

(*6*) Are there infinitely many "prime twins," i.e., pairs of integers, k and $k + 2$, both of which are primes?

(*7*) Are there arbitrarily long sequences of integers, all of which are composite?

7.2. The sieve of Eratosthenes. Of ancient origin is the device of preparing a list of prime numbers less than a given limit by

*Chapter 7 is a basic chapter.

writing down all the integers up to that limit and then in a systematic way eliminating all the composite integers. One such device is ascribed to Eratosthenes (276–194 B.C.).

For example, with a limit of $n = 100$, we first set down a list of the integers from 2 to 100. Recognizing that 2 is a prime, but that all proper multiples of 2 are composite, we cross out 4,6,8,. . .,100. The next number not crossed out is 3, which must be a prime for the only possible proper factor is 2, and 3 is not a multiple of 2 else it would have been crossed out. Recognizing that all proper multiples of 3 are composite, we cross out 6,9,12,. . .,99—although it is not actually necessary to cross out 6,12,18,. . .,96 again, since they are already crossed out, being multiples of 2. The next number not crossed out is 5; this number must be a prime, for if it were composite, it would have to have as a proper factor a prime less than 5, namely, either 2 or 3; but since 5 is not crossed out, it is not a multiple of 2 or 3. Crossing out all multiples of 5, not previously crossed out, namely: 25,35,55,65,85,95, we find, by the same reasoning as before, that the next number not crossed out must be a prime; it is 7. The only multiples of 7, not previously crossed out, are 49,77,91, and these we now cancel. Now, unless we have been analyzing the sieve process carefully, we are due for a surprise—*all* the *remaining* numbers which have survived the sieve are *primes*! The sieve appears as follows:

	2	3	~~4~~	5	~~6~~	7	~~8~~	~~9~~	~~10~~	11	~~12~~	13	~~14~~	~~15~~	~~16~~	17	~~18~~	19	~~20~~
~~21~~	~~22~~	23	~~24~~	~~25~~	~~26~~	~~27~~	~~28~~	29	~~30~~	31	~~32~~	~~33~~	~~34~~	~~35~~	~~36~~	37	~~38~~	~~39~~	~~40~~
41	~~42~~	43	~~44~~	~~45~~	~~46~~	47	~~48~~	~~49~~	~~50~~	~~51~~	~~52~~	53	~~54~~	~~55~~	~~56~~	~~57~~	~~58~~	59	~~60~~
61	~~62~~	~~63~~	~~64~~	~~65~~	~~66~~	67	~~68~~	~~69~~	~~70~~	71	~~72~~	73	~~74~~	~~75~~	~~76~~	~~77~~	~~78~~	79	~~80~~
~~81~~	~~82~~	83	~~84~~	~~85~~	~~86~~	~~87~~	~~88~~	89	~~90~~	~~91~~	~~92~~	~~93~~	~~94~~	~~95~~	~~96~~	97	~~98~~	~~99~~	~~100~~

We are always sure to reach the end of the sieve process when we have crossed out the proper multiples of p' where p' is the largest prime such that $p' \leqq \sqrt{n}$. This follows because if $s = ab$ is composite at least one (and in fact usually just one) of the factors a and b must be $\leqq \sqrt{s}$; otherwise if $a > \sqrt{s}$ and $b > \sqrt{s}$, we would find $s = ab > (\sqrt{s})^2 = s$, an obvious contradiction. Hence if s is not crossed out when the proper multiples of p' (and of all the smaller primes) have been eliminated, then s must be a prime. For not having been crossed out in this or any of the previous steps, s can have no non-unit factor $\leqq \sqrt{n}$; and since $s \leqq n$, s can have no non-unit factor $\leqq \sqrt{s}$; but as we have just demonstrated above, such an s cannot be composite.

Thus one answer to questions (*1*) and (*2*) has been provided. It is, of course, not too satisfactory an answer. For example, if we let the function $\pi(x)$ indicate the number of positive prime integers less than or equal to x, then our answer to (*1*) demands the making of $\pi(\sqrt{n})$ sieving steps and our answer to (*2*) demands perhaps the making of as many as $\pi(\sqrt{n})$ divisibility tests. Thus to prove that a number like $2^{127} - 1$ is a prime, would not be feasible by this method.

In general, no really satisfactory test has been found to answer the question (*2*) whether a given integer is prime or composite, but in the course of this book, we will point out various criteria which give impractical complete answers (like the above sieve process) or incomplete practical answers.

Essentially by the sieve method, but also with the aid of other theorems and, in recent years, with the aid of advanced mechanical computers, various mathematicians have prepared extensive tables of primes and of factors. The tables in the usual handbooks will serve for ordinary problems. For more extended numerical investigations, the student should become acquainted with the work of D. N. Lehmer:

> *Factor table for the first ten millions containing the smallest factor of every number not divisible by 2, 3, 5 and 7 between the limits 0 and 10,017,000.* Carnegie Institution of Washington Publication 105, 1909.
>
> *List of prime numbers from 1* to *10,006,721.* Carnegie Institution of Washington Publication 165, 1914.

Useful as they are, such tables are, of course, inadequate to handle problems like the one proposed above concerning

$$2^{127} - 1 = 170{,}141{,}183{,}460{,}469{,}231{,}731{,}687{,}303{,}715{,}884{,}105{,}727$$

for this number is considerably beyond the range of existing tables. Yet by clever devices, Lucas was able to show in 1876 that this number (known as M_{127}) is indeed a prime, and until recent years it remained the largest number known to be a prime (the student should compare this remark with the theorem of the next section which will show the *existence* of infinitely many primes).

A "too easy" arbitrarily large composite number n is provided by increasing x in $n = 2^x$. But the largest "genuine" composite number (meaning a number which is known to be composite, but whose factors are not known!) is $M_{257} = 2^{257} - 1$ whose compositeness was proved by Lehmer and Kraitchik.

7.3. The number of primes is infinite. For the theorem used as the title of this section and providing the answer to question (*3*), many, many proofs have been given, some simple, some erudite. We will present three of reasonable simplicity.

Proof 1 using $p! + 1$. If we recall the definition of $n!$ (see EX. *3.7*), then it is especially easy to describe Euclid's proof that there are infinitely many distinct primes. For suppose that the prime p is the largest prime. We shall show that this supposition is false by studying the number $M = p! + 1 = (1\cdot2\cdot3\cdot\ldots\cdot p) + 1$. Evidently M is not divisible by any of the numbers $2,3,4,\ldots,p$ because there is a remainder 1 in each of these cases; hence M is not divisible by any prime $\leqq p$. However, by the fundamental theorem in **6.2**, M is either (A) itself a prime or (B) is a product of primes. In either case we see that there must exist a prime larger than the prime p. Hence there is no largest prime p. Hence there must be infinitely many distinct primes.

In the preceding proof the student is cautioned to note the possibility of either (A) or (B). For example, $3! + 1 = 7$, a prime; but $5! + 1 = 121$, a composite number. However, 121 is the square of 11 and 11 is a prime larger than 5, so the proof is as correct in this case as in the former. In neither case is the prime uncovered by the proof necessarily the *next* prime, witness the two examples just given.

Proof 2 using integers of the form $6x - 1$. We can show that there are infinitely many primes among the integers of the arithmetic progression A: 5,11,17,23,29,35,. . ., the general form for an integer of this sequence being $6x - 1$. For if we suppose that $P_1,P_2,\ldots,P_k$ are the first k primes belonging to A, arranged in natural order, then we can prove the existence of a still larger prime belonging to A. Consider the integer $M = 6P_1P_2\ldots P_k - 1$, and its standard form as in **6.2**. Since M is odd and not a multiple of 3, it follows that all the prime factors of M are of the form $6x + 1$ or $6x - 1$, for there are no odd primes >3 of the form $6x \pm 3$, all of these latter numbers (except 3) being obviously composite. However, the product of any number of primes of the form $6x+1$ is again a number of the form $6x+1$. In order for M to have the form $6x-1$, as it does, M must have *at least one* prime factor p of the form $6x-1$. However, this prime p must be larger than P_k, because none of $P_1, P_2, \ldots, P_k$ is a factor of M, since each of these when tried as a factor leaves a remainder of -1. Hence P_k is not the largest prime in A, and A must contain infinitely many primes. The result just given is a special case of the celebrated theorem of Dirichlet that if a and b are given *relatively prime* integers, then the arithmetic progression made up of all integers of the type $ax+b$ contains infinitely many primes (see Dickson, *Modern Elementary Theory of Numbers*).

Proof 3 using generalized Fermat numbers. Let a be any fixed positive integer with $a \geqq 2$. Then if a^s+1 is a prime, it is necessary, but not sufficient, that s have the form $s = 2^x$. To prove this remark we need to observe that if q is odd, and $q > 1$, then a^q+1 has the non-trivial factor $a+1$ and $a^{2^x q}+1$ has the non-trivial factor $a^{2^x}+1$. Both of these results follow by letting $q = 2n+1$ and $r = a$ or $r = a^{2^x}$ in the identity which follows:

$$r^{2n+1}+1 = (r+1)(r^{2n}-r^{2n-1}+\ldots+r^2-r+1)\ .$$

This last relation may be established by induction on n (see EX. 7.5).

Hence if looking for primes of the form a^s+1, we need examine only the numbers $F_m = a^{2^m}+1$, which we shall describe as generalized Fermat numbers, in memory of an incorrect but provocative conjecture of Fermat, who believed in the case when $a = 2$ that all the numbers F_m were primes (however, a hundred years later Euler showed that $F_5 = 2^{32}+1$ is composite).

We shall show, for each a, that there are infinitely many distinct primes to be found *among the factors* of the F_m. Our method is to show that *any two* members F_m and F_n of the sequence of generalized Fermat numbers going with a fixed a, *have at most a prime factor 2 in common*, for then the infinite sequence of F_m must have an infinite number of distinct primes appearing as prime factors.

We begin by making repeated applications of the well-known identity $x^2-1 = (x-1)(x+1)$ to obtain the following relations:

$$F_m - 2 = a^{2^m} - 1 = (a^{2^{m-1}}-1)F_{m-1} = (a^{2^{m-2}}-1)F_{m-2}F_{m-1} = \\ \ldots = (a-1)F_0F_1\ldots F_{m-2}F_{m-1}.$$

From this last result it follows readily that $(F_n, F_m - 2) = F_n$ for every

$n < m$. Let $d = (F_n, F_m)$. Since d divides F_n, it follows that d divides $F_m - 2$; but since d also divides F_m, then d can divide $F_m - 2$ only if d divides 2. If a is even, each F_i is odd and hence $d = 1$; but if a is odd, each F_i is even and hence $d = 2$. This completes the proof.

7.4. Distribution of the primes. All the tables and all the known theorems indicate that the primes occur in a very irregular way within the sequence of all integers. For example, as far as any tables have been extended there always occur, now and again, "prime twins," i.e., a pair of successive odd integers x and $x + 2$, both of which are primes, such as 101 and 103, 107 and 109, 137 and 139, etc. But as yet no complete answer is available to question *(6)* as to whether there are infinitely many prime twins.

Question *(7)* is an easier one with a positive answer, for it can be shown that there are arbitrarily long sequences of integers all of which are composite. Thus if given the integer n, we have but to consider the n numbers running from $(n + 1)! + 2$ to $(n + 1)! + n + 1$ to have at hand a sequence of n successive, composite integers. Actually sequences of n composite numbers usually occur much earlier in the tables; for example, there are 13 composite numbers from 114 to 126.

Many amateur mathematicians have sought formulas which would answer question *(4)* and show directly what integer is the nth prime; or which would show the $n + 1$ prime, if one knew the nth prime. Most professional mathematicians who have worked on this problem say that the weight of evidence is to the effect that no such formulas can be found. Perhaps the greatest progress has been made in the study of the function $\pi(x)$, giving the number of primes less than or equal to x. Of course, if the exact form of $\pi(x)$ were known, the previous problems could be answered at once. But the progress of which we speak is of a different kind and belongs to what might be called the advanced theory of numbers, where analytic methods based on infinite series and various integrals of the calculus have made it possible to estimate the value of $\pi(x)$ for "sufficiently large" values of x.*

Of a somewhat different, but still rather fruitless, nature is the

*For example, see Trygve Nagell, *Introduction to Number Theory*, Chapter VIII. New York, Wiley, 1951.

search for formulas, like Fermat's incorrect one, which will yield only primes, even if they won't give all the primes. The person who first studied the function $f(x) = x^2 - x + 41$ must have been excited as he substituted $x = 1,2,\ldots,40$ to find that he obtained forty primes. Had he been a laboratory scientist, he might have shouted "Eureka!" But being only a mathematician, he substituted $x = 41$, and then went out for some coffee.

Of a similar exciting and then disappointing nature is the function $f(x) = x^2 - 79x + 1601$.

In view of these examples it may be of interest to answer question (5) in a definitely negative way and to prove that a polynomial $f(x)$ which is not a constant and which has integer coefficients cannot be prime for all integral values of x, and is composite for infinitely many integral values of x. The proof demands only a little familiarity with the properties of polynomials.

Since $f(x)$ is not a constant, $|f(k)| > 1$, for some integer k. Set $y = f(k)$ and consider $f(ty + k)$. There are several ways of showing that $f(ty + k) = yQ + f(k)$, where Q is a polynomial in t,y,k with integer coefficients (see EX. *7.9*). Hence $f(ty + k) = y(Q + 1)$ is divisible by $y = f(k)$ for all values of t. Since $f(x)$ is not a constant, $f(ty + k)$ increases in absolute value for t sufficiently large; therefore, for such sufficiently large values of t, the complementary factor $Q + 1$ is not a unit, and hence $f(ty + k)$ is composite. Since $ty + k$ becomes arbitrarily large with t, the latter having, say, the same sign as y, it follows that $f(x)$ fails to represent just primes in an infinity of cases and in fact for all x of the form $x = ty + k$ when t is sufficiently large.

EXERCISES

EX. *7.1.* Show that $\pi(\sqrt{210}) = 6$, listing the six primes concerned.

EX. *7.2.* Apply the sieve process to only the interval 190 to 210 (recall that just $\pi(\sqrt{210})$ steps are required) and find all primes and all prime twins in this interval.

EX. *7.3.* Modify Euclid's proof that there are infinitely many primes by supposing the kth prime to be the largest and using $M = (p_1p_2\ldots p_k) + 1$, where $p_1,p_2,\ldots,p_k$ are the first k primes, to arrive at a contradiction.

EX. *7.4.* By a slight variation of *Proof 2* in **7.3**, show that there are infinitely many primes in the arithmetic progression 3,7,11,15,... of integers of the form $4x - 1$.

EX. *7.5.* By induction on n, prove that

$$r^{2n+1} + 1 = (r + 1)(1 - r + r^2 - \dots + r^{2n}) .$$

EX. *7.6.* Prove that $a^s - 1$ is composite if $a > 2$ and $s > 1$ (see EX. *3.2*).

EX. *7.7.* Prove that $2^s - 1$ is composite if s is composite (see EX. *3.2*).

EX. *7.8.* Show that there can be no prime triplets, i.e., three successive odd integers, each a prime.

EX. *7.9.* If $f(x) = a_0 + a_1x + \dots + a_nx^n$ with integer coefficients, use EX. *3.7* to show that $f(ty + k) = yQ + f(k)$ where Q is a polynomial in t,y,k having integer coefficients.

EX. *7.10.* Give a different proof of EX. *7.9*, using the division algorithm for polynomials with y as the divisor and R, free from y, as the remainder; then set $y = 0$ to show $R = f(k)$.

EX. *7.11.* Illustrate EX. *7.9* when $f(x) = x^2 - 79x + 1601$, showing that $Q = t(ty + 2k - 79)$. With $k = 1, y = f(1), t = 1$, show that $f(1524) = 1523 \cdot 1447$.

EX. *7.12.* A sieve for odd numbers. Consider the set C of all numbers $C(r,s) = 2rs + r + s$, where r and s are positive integers. Prove that an odd number $p = 2K + 1$ is a prime if and only if K is not in the set C. (For convenience the numbers of C may be arranged in rows and columns; the elements of the rth row are then the terms of an arithmetic progression with first term $3r + 1$ and common difference $2r + 1$.)

EX. *7.13.* Using the notation of EX. *7.3*, show that the $P_{k+1} - 2$ integers following M are composite.

▶ *The essence of mathematics is its freedom.*

—G. CANTOR

CHAPTER 8[*]

THE NUMBER-THEORETIC FUNCTIONS $\tau(n)$ AND $\sigma(n)$

8.1. $\tau(n)$, the number of divisors of n. Let us seek a function $\tau(n)$ to give the *number* of positive integer divisors of any given positive integer n. As we shall discover, such a function must be of a very different character from the functions usually studied in algebra or analysis, for it depends in a critical way not only upon the value of n, but also upon the standard representation of n, as in **6.2**, and the standard representation changes radically as we pass from n to $n + 1$. Hence we shall describe $\tau(n)$, and any other function whose range depends upon the standard form of n, as a *number-theoretic* function, the adjective intended to emphasize the special nature of the function.

If n is written in standard form as

$$n = p_1^{a_1}p_2^{a_2}\dots p_k^{a_k}$$

then all the positive integer divisors of n are given, without repetition, by the form

$$d = p_1^{b_1}p_2^{b_2}\dots p_k^{b_k}$$

where for each value of i, the b_i runs independently through the following range of values: $b_i = 0,1,2,\dots,a_i$.

[*] Chapter 8 is a basic chapter.

Now, by a well-known combinatorial principle, it follows that if b_1 can be chosen in $a_1 + 1$ ways, if b_2 can be chosen in $a_2 + 1$ ways, $\ldots$, and if b_k can be chosen in $a_k + 1$ ways, then $b_1, b_2, \ldots, b_k$ all together can be selected in a number of ways given by the product $(a_1 + 1)(a_2 + 1)\ldots(a_k + 1)$.

Hence we find that the number $\tau(n)$ of positive integer divisors of n is given exactly by

$$\tau(n) = (a_1 + 1)(a_2 + 1)\ldots(a_k + 1).$$

(This covers all cases where $n > 1$, see **6.2**, and it is easy to confirm that $\tau(1) = 1$.)

For example, $2520 = 2^3 \cdot 3^2 \cdot 5 \cdot 7$, hence

$$\tau(2520) = (3 + 1)(2 + 1)(1 + 1)(1 + 1) = 48,$$

so 2520 has exactly 48 distinct positive integer divisors.

8.2 $\sigma(n)$, the sum of the divisors of n. It is clear from the preceding discussion that the *sum* $\sigma(n)$ of all the distinct positive integer divisors of a given positive integer $n > 1$ is given by the following product:

$$\sigma(n) = (1 + p_1 + p_1^2 + \ldots + p_1^{a_1}) \\ (1 + p_2 + \ldots + p_2^{a_2})\ldots(1 + p_k + \ldots + p_k^{a_k})$$

because in this product each of the divisors d of n, described in the previous section, appears once and only once as a summand, when the product has been expanded.

(When $n = 1$, we see directly that $\sigma(1) = 1$.)

With the aid of EX. *3.2*, inasmuch as each $p_i > 1$, we find that $\sigma(n)$ can also be written in the following form:

$$\sigma(n) = \frac{p_1^{a_1+1} - 1}{p_1 - 1}\,\frac{p_2^{a_2+1} - 1}{p_2 - 1}\ldots\frac{p_k^{a_k+1} - 1}{p_k - 1}$$

For example, $2520 = 2^3 \cdot 3^2 \cdot 5 \cdot 7$, hence we find that

$$\sigma(2520) = \frac{2^4 - 1}{2 - 1}\,\frac{3^3 - 1}{3 - 1}\,\frac{5^2 - 1}{5 - 1}\,\frac{7^2 - 1}{7 - 1} = 15 \cdot 13 \cdot 6 \cdot 8 = 9360,$$

so the sum of all the positive integer divisors of 2520 is exactly 9360.

8.3. Perfect numbers. Number mysticism has played an important part in the history of the theory of numbers. One vestige of the influence of numerology is in the use of the adjectives *deficient*, *perfect*, and *abundant* to describe integers for which $\sigma(n) < 2n$, $\sigma(n) = 2n$, and $\sigma(n) > 2n$, respectively.

A rather fascinating, but unfinished, chapter of the theory concerns the determination of all perfect numbers: the first part concerns the discovery of all *even* perfect numbers with a creditable part by Euclid and a doubtful portion by Mersenne; the second, unfinished part concerns the as yet unsolved problem as to whether there are any *odd* perfect numbers.

Let $n = 2^{k-1}A$ be an even perfect number, $k \geqq 2$ and A odd. By definition we must have $\sigma(n) = 2n$ or $\sigma(2^{k-1}A) = 2^kA$. Since $(2^{k-1},A) = 1$, we may apply EX. *8.2* to see that $\sigma(2^{k-1}A) = \sigma(2^{k-1})\sigma(A)$. By the formula in **8.2**, we know that $\sigma(2^{k-1}) = 2^k - 1$, hence we arrive at the following condition:

$$(2^k - 1)\sigma(A) = 2^kA.$$

Let us write $\sigma(A) = A + X$, where X is the sum of all the positive divisors of A which are less than A. Then the condition displayed above reduces to the form $(2^k - 1)X = A$. This implies that X is a divisor of A; moreover, since $k \geqq 2$, X is less than A; thus X, which is supposed to be the sum of all divisors of A less than A, must include X itself. But $X = X + Y$ implies $Y = 0$, so A has *only one* divisor X less than A; however, the only integers with this property are primes; therefore A is an odd prime, $X = 1$, and A must have the form $2^k - 1$. We have thus arrived at Euclid's conclusion: the only possible even perfect numbers must have the form $n = 2^{k-1}p$ where $p = 2^k - 1$ is an odd prime.

Conversely, to complete the argument, we must check that every such number *is* a perfect number; but this check is very easy, for by section **8.2** we find

$$\sigma(n) = \sigma(2^{k-1}p) = (2^k - 1)(p + 1) = p2^k = 2n.$$

The result just given suggests a search for all primes of the form $M_k = 2^k - 1$, since each prime so discovered will provide a corresponding perfect number $P_k = 2^{k-1}M_k$. The study of numbers of the type M_k is known as the study of *Mersenne numbers* after Mersenne (1588–1648) who made a number of correct and several incorrect statements about which ones of these numbers are composite and which prime. As EX. *7.7* shows, M_k is composite if k is composite, hence the search for Mersenne primes (and for even perfect numbers) is narrowed to the case where k is prime. Thus far only 12 Mersenne primes are known, although the search has been completed among all primes $k \leqq 257$. The simplest cases are as follows:

$$M_2 = 3,\ P_2 = 6;\quad M_3 = 7,\ P_3 = 28;$$
$$M_5 = 31,\ P_5 = 796;\quad M_7 = 127,\ P_7 = 8128.$$

The next Mersenne primes are M_{13}, M_{17}, M_{19}, M_{31}, M_{61}, M_{89}, M_{107}, and M_{127}. Concerning the last of these we have already made some remarks at the close of **7.2**.

The best results about odd perfect numbers are of the type that if there are any such numbers they must have more than a certain number of distinct prime factors. Imperfect though such results may be, they are yet sufficient to indicate that if any odd perfect numbers exist they will be large numbers and not found by mere guesswork.

8.4. Multiplicative number-theoretic functions. A number-theoretic function $f(n)$ is defined to be *multiplicative* if and only if $f(ab) = f(a)f(b)$ for all positive integers a,b for which $(a,b) = 1$.

For example, both $\tau(n)$ and $\sigma(n)$ are multiplicative, as may be seen from their formulas developed above, see EX. *8.2*. The functions $f(n) = 1$ and $f(n) = n$ are other simple examples of multiplicative functions.

If $f(n)$ is multiplicative, it follows since $(a,1) = 1$, that $f(a) = f(a\cdot 1) = f(a)f(1)$ and hence that $f(1) = 1$. For $n > 1$ we can apply the fundamental theorem to write n in standard form as $n = p_1{}^{a_1}p_2{}^{a_2}\ldots p_k{}^{a_k}$ where the p_i are distinct primes. Since the factors $p_i{}^{a_i}$ are relatively prime in pairs, it follows that an expression for a function $f(n)$, which is known to be multiplicative, can be found by investigating the value of $f(p^a)$ where p^a is a power of a prime.

Of special interest is the following theorem which shows how new multiplicative functions may be generated from known multiplicative functions.

Theorem: If $f(n)$ is a multiplicative number-theoretic function, then so also is $F(n)$ where $F(n) = \Sigma f(d)$, summed over all positive divisors d of n.

Proof: By definition

$F(a) = \Sigma f(d)$ summed over the set S of all divisors d of a,
$F(b) = \Sigma f(d')$ summed over the set S' of all divisors d' of b,
$F(ab) = \Sigma f(d'')$ summed over the set S'' of all divisors d'' of ab.

Let S^* be the set of all numbers of the form dd' where d ranges over S and d' over S'. The relation $d''k'' = ab$ shows that any prime factor of d'' must divide either a or b, hence d'' has the form $d'' = dd'$ where d divides a and d' divides b. Thus every number in S'' is in S^*. Conversely, every number in S^* is in S'', for from $a = dk$, $b = d'k'$, we find $ab = dd'kk'$, so that dd' is a divisor of ab. However, the sets S'' and S^* are not, in general, identical, for S'' is defined in such a way as to have no duplications, whereas S^* may have duplications.

We can show that S^* has no duplications under the assumption that $(a,b) = 1$. For it is clear that (d,d') divides (a,b), hence if $(a,b) = 1$, then $(d,d') = 1$ for all d in S and all d' in S'. Consequently an equality $dd' = d_1d_1'$ implies since $(d,d_1') = 1$ that d divides d_1, and since $(d_1,d') = 1$ that d_1 divides d; hence $d = d_1$ and $d' = d_1'$. Thus the factors d and d' of a number of S^* are uniquely determined.

By hypothesis $f(n)$ is multiplicative; under the assumption that $(a,b) = 1$, we have seen that $(d,d') = 1$; combining these remarks we may write $f(d)f(d') = f(dd')$. Under the hypothesis that $(a,b) = 1$, we have identified the sets S^* and S'', so we may write

$$F(a)F(b) = \sum_S f(d) \sum_{S'} f(d') = \sum_{S^*} f(dd') = \sum_{S''} f(d'') = F(ab).$$

This is, of course, precisely the requirement which shows $F(n)$ to be multiplicative.

We have mentioned above that $f(n) = 1$ is a multiplicative function. This is easy to see since $f(ab) = 1 = 1 \cdot 1 = f(a)f(b)$ for all a,b, including, therefore, the cases when $(a,b) = 1$. By applying the theorem we know that $F(n) = \Sigma f(d)$ must be multiplicative. However, $\Sigma f(d) = 1 + 1 + \ldots + 1$ with as many summands as there are positive integral divisors of n; hence $F(n)$ is what we have previously called $\tau(n)$. By this approach we know a priori that $\tau(n)$ is multiplicative. Consequently by investigating

$$\tau(p^a) = f(1) + f(p) + f(p^2) + \ldots + f(p^a) = a + 1$$

we are able to conclude that for $n = p_1^{a_1}p_2^{a_1}\ldots p_k^{a_k}$,

$$\tau(n) = (a_1 + 1)(a_2 + 1)\ldots(a_k + 1).$$

This develops the formula for $\tau(n)$ in a way entirely independent of that given in **3.1**.

Similarly, we may show that $f(n) = n$ is a multiplicative function, for $f(ab) = ab = f(a)f(b)$ for all a,b, including the required cases where $(a,b) = 1$. It follows from the theorem that $F(n) = \Sigma f(d) = \Sigma d$ is multiplicative and we see that $F(n)$ is what we have previously called $\sigma(n)$. By this approach we know a priori that $\sigma(n)$ is multiplicative. We therefore investigate

$$\sigma(p^a) = f(1) + f(p) + f(p^2) + \ldots + f(p^a)$$
$$= 1 + p + p^2 + \ldots + p^a = \frac{(p^{a+1} - 1)}{(p - 1)}$$

and conclude that for $n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k}$,

$$\sigma(n) = \frac{p_1^{a_1+1} - 1}{p_1 - 1}\,\frac{p_2^{a_2+1} - 1}{p_2 - 1}\,\ldots\,\frac{p_k^{a_k+1} - 1}{p_k - 1},$$

which agrees with the result established by different reasoning in **8.2**.

In the exercises and in later chapters we will make further use of this theorem on multiplicative functions.

EXERCISES

EX. *8.1.* Compute $\tau(4950)$ and $\sigma(4950)$.

EX. *8.2.* Assuming the formulas in **8.1** and **8.2**, prove that $\tau(ab) = \tau(a)\tau(b)$ and $\sigma(ab) = \sigma(a)\sigma(b)$ whenever $(a,b) = 1$.

EX. *8.3.* Show $\tau(n)$ is odd if and only if n is a square.

EX. *8.4.* Show $\tau(x) = q$ has infinitely many solutions x for every given integer $q > 1$. Use **7.3**.

EX. *8.5.* Make a table of values of $\sigma(p^a) = (p^{a+1} - 1)/(p - 1)$ where p is a prime and $\sigma(p^a) < 100$.

EX. *8.6.* Find all solutions of $\sigma(x) = 72$. Use EX. *8.5*.

EX. *8.7.* Let n be called *multipli-perfect* if $\sigma(n) = kn$, where k is an integer with $k \geqq 3$. Prove that $n = 120$ and $n = 672$ are multipli-perfect numbers.

EX. *8.8.* Use the definition in EX. *8.7* and show that $n = 14{,}182{,}439{,}040$ is a multipli-perfect number. (Descartes.)

EX. *8.9.* Find a common property of $\sigma(81)$, $\sigma(343)$, $\sigma(400)$.

EX. *8.10.* Let a pair of positive integers A and B be called *amicable* if $\sigma(A) = A + B = \sigma(B)$. Prove that 220 and 284 are amicable.

EX. *8.11.* Prove that the *product* of all the positive divisors of n is given by $n^{\tau(n)/2}$. Compare with EX. *8.3*.

EX. *8.12.* Use the theorem in **8.4** and develop a formula for $\tau_1(n) = \Sigma\tau(d)$, summed over the positive divisors d of n.

EX. *8.13* Develop a formula for $\tau_2(n) = \Sigma\tau_1(d)$, where $\tau_1(n)$ is defined in EX. *8.12*.

EX. *8.14.* , If $n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k}$ is in standard form and if s is a fixed integer, define $f(n) = s^k$ for $n > 1$ and $f(1) = 1$. Prove that $f(n)$ is multiplicative.

EX. *8.15.* Develop a formula for $F(n) = \Sigma f(d)$ where $f(n)$ is defined in EX. *8.14*.

> ▶ *If a man's wit be wandering, let him study mathematics; for in demonstrations, if his wit be called away never so little, he must begin again.* —LORD BACON

*CHAPTER 9**

THE BRACKET FUNCTION

9.1. Definition of the bracket function. With any real number x we may associate a uniquely determined *integer*, called the "integral part of x" and designated $[x]$ which may be read "bracket x," by requiring that $[x]$ be an integer for which

$$[x] \leqq x < [x] + 1.$$

For example: $[14/3] = 4$, $[-7] = -7$, $[\sqrt{10}] = 3$, $[-\sqrt{10}] = -4$, $[\pi] = 3$, $[-\pi] = -4$.

As a consequence of the definition it follows immediately that

$$x = [x] + \theta \quad \text{with} \quad 0 \leqq \theta < 1.$$

Then the following properties of the bracket function may be readily established.

B.1: If m is an integer, $[x + m] = [x] + m$.

Proof: From $x = [x] + \theta$, $0 \leqq \theta < 1$, it follows that $x + m = [x] + m + \theta$; since $[x] + m$ is an integer, **B.1** must hold.

B.2: $[x] + [-x] = 0$, or -1, according as x is, or is not, an integer.

*Chapter 9 is a supplementary chapter, prerequisite, however, for Chapter 10.

Proof: If x is an integer, $[x] = x$, and $[-x] = -x$, hence $[x] + [-x] = x - x = 0$. If x is not an integer, then $x = [x] + \theta, 0 < \theta < 1$; hence $-x = -[x] - \theta = -1 - [x] + (1 - \theta)$ with $0 < 1 - \theta < 1$ and with $-1 - [x]$ an integer. Therefore $[-x] = -1 - [x]$, or otherwise expressed, $[x] + [-x] = -1$.

B.3: $[x + y] \geqq [x] + [y]$.

Proof: Let $x = [x] + \theta_1, 0 \leqq \theta_1 < 1; y = [y] + \theta_2, 0 \leqq \theta_2 < 1$. Then $x + y = [x] + [y] + \theta_1 + \theta_2$ with $[x] + [y]$ an integer and with $0 \leqq \theta_1 + \theta_2 < 2$. Either $0 \leqq \theta_1 + \theta_2 < 1$ and $[x + y] = [x] + [y]$; or $1 \leqq \theta_1 + \theta_2 < 2$ and $[x + y] = [x] + [y] + 1$. But in either case, **B.3** holds.

B.4: If n is a positive integer, then $[[x]/n] = [x/n]$.

Proof: Let $x = [x] + \theta, 0 \leqq \theta < 1$. By the division algorithm find q and r so that $[x] = qn + r, 0 \leqq r \leqq n - 1$. Then $[x]/n = q + r/n$ with q an integer and with $0 \leqq r/n < 1$; hence $[[x]/n] = q$. But $x/n = (qn + r + \theta)/n = q + (r + \theta)/n$ with q an integer and with $0 \leqq (r + \theta)/n \leqq (n - 1 + \theta)/n < 1$; hence $[x/n] = q$. Comparing these two results we find that we have established **B.4**.

9.2. An interesting exercise. For the purpose of the following exercise we shall assume the reader has had some experience with real numbers and knows that these numbers may be divided into two classes: the rational numbers, or ordinary fractions, of the form p/q where p and q are integers with $q \neq 0$; and the irrational numbers which are those real numbers which are not rational, such as $\sqrt{2}$ and π. These concepts are explained in detail in a later chapter.

Exercise: If a and b are positive irrational numbers such that $1/a + 1/b = 1$, then the two series $[an]$ and $[bn]$ for $n = 1,2, \ldots$ represent *all* positive integers *without repetition*.

For example: If $a = \sqrt{2}$, then $b = 2 + \sqrt{2}$. The a-series begins $[a] = 1, [2a] = 2, [3a] = 4, [4a] = 5, [5a] = 7, [6a] = 8, [7a] = 9, [8a] = 11$, etc.; while the b-series begins with $[b] = 3, [2b] = 6, [3b] = 10$, etc., exactly complementing the a-list.

Proof: Since a and b are positive with $1/a + 1/b = 1$, it follows that both $a > 1$ and $b > 1$.

(A) We shall show that there is *no repetition* of integers in the series $[an]$ and $[bn]$, $n = 1,2, \ldots$.

(A.1) In the series $[an]$ there is no repetition, for we have

$$[an] < an < a(n+1) - 1 < [a(n+1)];$$

the first inequality follows since a and (therefore) an are irrational; the second inequality follows from $1 < a$. Similarly, there is no repetition in the series $[bn]$.

(A.2) For all positive integers n and m we can show $[an] \neq [bm]$. For if we suppose the integer $x = [an] = [bm]$, then

$$an - 1 < x < an \quad \text{or} \quad n - 1/a < x/a < n,$$
$$bm - 1 < x < bm \quad \text{or} \quad m - 1/b < x/b < m.$$

Adding the latter inequalities of each line and using $1/a + 1/b = 1$, we find $n + m - 1 < x < n + m$, so that the *integer* x lies between two successive integers, which is an obvious contradiction.

(B) Next we can show that no integer x is omitted in *both* sequences $[an]$ and $[bm]$. For if we suppose x to be such an integer, then there must exist integers n and m such that

$$an < [an] + 1 \leqq x \leqq [a(n+1)] - 1 < a(n+1) - 1$$

or

$$n < x/a < n + 1 - 1/a,$$

$$bm < [bm] + 1 \leqq x \leqq [b(m+1)] - 1 < b(m+1) - 1$$

or

$$m < x/b < m + 1 - 1/b.$$

Adding the latter inequalities of each line and using $1/a + 1/b = 1$, we find $n + m < x < n + m + 1$, which is a contradiction.

Establishing (A) and (B) completes the proof of the exercise.

9.3. $\tau(n)$ and the bracket function. The following theorem shows a connection between $\tau(n)$ and the bracket function.

Theorem: For any positive integer n, the following equation holds:

$$\tau(1) + \tau(2) + \ldots + \tau(n) = [n/1] + [n/2] + \ldots + [n/n].$$

Proof: The proof makes use of the interesting device of counting in two different ways the number of solutions in positive integers d of the set of equations

$$dx = 1,\ dx = 2,\ dx = 3,\ \ldots,\ dx = n.$$

First, consider the set of equations in the order just given. Certainly in the equation $dx = k$, each factor of k leads to a suitable d, and k has exactly $\tau(k)$ factors. As $k = 1,2,\ldots,n$, we find that the total number of solutions d agrees with the left side of the equation which we are trying to establish.

Secondly, consider the equations with $x = 1,2,\ldots,n$, and with k restricted to the range $1 \leqq k \leqq n$. Then the equations and corresponding solutions may be grouped as follows:

$$\begin{array}{lll}
d1 = k;\ k = 1,2,\ldots,1[n/1]; & d = 1,2,\ldots,[n/1]; \\
d2 = k;\ k = 2,4,\ldots,2[n/2]; & d = 1,2,\ldots,[n/2]; \\
d3 = k;\ k = 3,6,\ldots,3[n/3]; & d = 1,2,\ldots,[n/3]; \\
\ldots\ ;\ \ldots \qquad\qquad ; & \ldots \qquad\qquad ; \\
dn = k;\ k = n = \quad n[n/n]; & d = 1 = \quad [n/n].
\end{array}$$

From this point of view we find that the total number of solutions d agrees with the right side of the initial equation.

For example, let us consider $n = 10$;

$$\tau(1) + \tau(2) + \ldots + \tau(10) =$$
$$1 + 2 + 2 + 3 + 2 + 4 + 2 + 4 + 3 + 4 = 27,$$
$$[10/1] + [10/2] + \ldots + [10/10] =$$
$$10 + 5 + 3 + 2 + 2 + 1 + 1 + 1 + 1 + 1 = 27.$$

EXERCISES

EX. *9.1.* For all real x from -2 to $+2$ draw the graphs of (a) $y = [x]$; (b) $y = [x] + [-x]$; (c) $y = [x]^2$; (d) $y = [x^2]$.

EX. *9.2.* Illustrate **9.2** when $a = \sqrt{3}$.

EX. *9.3.* Illustrate **9.3** when $n = 20$.

EX. *9.4.* Prove that for any positive integer q and any real x

$$[x] + [x + 1/q] + [x + 2/q] + \ldots + [x + (q-1)/q] = [qx]$$

(Hint: let $x = [x] + \theta$, $0 \leqq \theta < 1$, and consider $s = [q\theta] \leqq q\theta < s + 1$.)

EX. *9.5.* Use **9.3** to establish that

$$\tau(n) = 1 + \Sigma([n/d] - [(n-1)/d])$$

where the summation runs from $d = 1$ to $d = n - 1$.

EX. *9.6.* Prove that

$$\tau(n) = [\sqrt{n}] - [\sqrt{n-1}] + 2\Sigma([n/d] - [(n-1)/d])$$

where the summation runs from $d = 1$ to $d = [\sqrt{n-1}]$.

EX. *9.7.* Whereas $[x]$ represents the "largest integer less than or equal to x," show that $[x + 1/2]$ is a "nearest integer to x."

EX. *9.8.* Prove that $f(x) = |x - [x + \frac{1}{2}]|$ has the property $f(x + m) = f(x)$ for every integer m.

EX. *9.9.* For values of $k = 0,1,2$, draw the graphs of

$$2^k y_k = |2^k x - [2^k x + \tfrac{1}{2}]|$$

from $x = 0$ to $x = 2$. Then draw the graph of $y = y_0 + y_1 + y_2$.

> ▶ *Every common mechanic has something to say in his craft about good and evil, useful and useless, but these practical considerations never enter into the purview of the mathematician.* —ARISTIPPUS, *the Cyrenaic*

CHAPTER 10[*]

THE FUNCTION $E(p,n)$

10.1. Definition and evaluation of $E(p,n)$. By $E(p,n)$ we mean the exponent of the prime p in the standard form of $n!$.

For example, $E(2,4) = 3$, $E(3,4) = 1$, and $E(5,4) = 0$, because $4! = 1\cdot 2\cdot 3\cdot 4 = 24 = 2^3 3$. But what we desire is a formula for determining with some rapidity the value, for example, of $E(3,101)$, where it is not practical to begin by actually writing $101!$ in standard form.

We shall show, using the bracket function of Chapter 9, that

$$(10.1)\quad E(p,n) = [n/p] + [n/p^2] + [n/p^3] + \dots + [n/p^k],$$

where
$$p^k \leqq n < p^{k+1}.$$

Proof: The integer $[n/p]$ shows the number of integers $\leqq n$ which are multiples of p and contribute at least 1 to $E(p,n)$. Then $[n/p^2]$ shows the number of integers $\leqq n$ which are multiples of p^2; and the fact that all numbers of this type are included in the previous set is properly counted since these multiples of p^2 contribute at least 2 to $E(p,n)$ and they are here counted at least twice, once in $[n/p]$ and again in $[n/p^2]$. Similarly, $[n/p^3]$ shows the number of integers $\leqq n$ which are multiples of p^3, and the fact that all multiples of this

[*] Chapter 10 is a supplementary chapter, requiring previous reading in Chapters 4 and 9.

type are included in both the previous sets is properly counted for they each contribute at least 3 to $E(p,n)$ and they are here counted at least three times, once in $[n/p]$, again in $[n/p^2]$, and again in $[n/p^3]$. Finally, $[n/p^k]$ shows the number of integers $\leq n$ which are multiples of p^k, and that each of these integers contributes k to $E(p,n)$ is properly counted, for each of these integers is a member of all the previous sets and is counted exactly k times: once in $[n/p]$, again in $[n/p^2], \ldots$, and again in $[n/p^k]$. No further increments to $E(p,n)$ can occur since $n < p^{k+1}$, so the proof of formula *(10.1)* is complete.

For purposes of computation we can improve *(10.1)* by using **B.4** of **9.1**. Let us set $n_i = [n/p^i]$. Then

$$n_{i+1} = [n/p^{i+1}] = [(n/p^i)/p]$$

and by **B.4** we may write

$$n_{i+1} = [[n/p^i]/p] = [n_i/p].$$

From *(10.1)* we have $E(p,n) = n_1 + n_2 + \ldots + n_k$, hence with the aid of the present observations we may write

$$(10.2) \quad E(p,n) = [n/p] + [n_1/p] + [n_2/p] + \ldots + [n_{k-1}/p]$$

where it is understood that, although by definition $n_i = [n/p^i]$, we shall here use $n_1 = [n/p]$ and $n_{i+1} = [n_i/p]$.

For example, using *(10.1)* we may write

$$E(3{,}101) = [101/3] + [101/9] + [101/27] + 101/81] = 33 + 11 + 3 + 1 = 48,$$

or using *(10.2)* we may write, recursively,

$$E(3{,}101) = [101/3] + [33/3] + [11/3] + [3/3] = 33 + 11 + 3 + 1 = 48.$$

10.2. $E(p,n)$ and representation of n to the base p. If n is written in the base p as in Chapter 4, say

$$n = (a_k \ldots a_1 a_0)_p = a_0 + a_1 p + a_2 p^2 + \ldots a_k p^k,$$

it is evident, since each a_i satisfies $0 \leqq a_i < p$, that the computation of the n_i, defined in the previous section, can be explicitly indicated, as follows:

$$n_1 = [n/p] = a_1 + a_2 p + \ldots + a_k p^{k-1};$$
$$n_2 = [n_1/p] = a_2 + a_3 p + \ldots + a_k p^{k-2}; \ldots;$$
$$n_{k-1} = a_{k-1} + a_k p; \; n_k = a_k.$$

However, these computations can be avoided because of the following relations:

$$n = n_1p + a_0;\ n_1 = n_2p + a_1;\ n_2 = n_3p + a_2;$$
$$\ldots;\ n_{k-1} = n_kp + a_{k-1};\ n_k = a_k;$$

for if these equations are added together we find, employing (*10.2*), that

$$n + E(p,n) = pE(p,n) + a_0 + a_1 + \ldots + a_k.$$

If we solve this last equation for $E(p,n)$ we obtain the following useful formula, due to Legendre:

$$(10.3) \qquad E(p,n) = (n - (a_0 + a_1 + \ldots + a_k))/(p - 1).$$

For example: since $(101)_{10} = 81 + 2\cdot 9 + 2 = (10202)_3$, we find with the aid of (*10.3*) that

$$E(3,101) = (101 - (2 + 0 + 2 + 0 + 1))/(3 - 1) = 96/2 = 48.$$

10.3. The equation $E(p,n) = m$. In this section we find the solutions n, if any exist, of the equation $E(p,n) = m$, where p is a given prime and m a given integer.

To this end we begin by considering the increment

$$\Delta E(p,n) = E(p,n) - E(p,n - 1).$$

Lemma: If $n = (a_k\ldots a_1a_0)_p$ and $a_0 \neq 0$, then $\Delta E(p,n) = 0$; but if $a_0 = a_1 = \ldots = a_{s-1} = 0$, while $a_s \neq 0$, then $\Delta E(p,n) = s$.

(In other words the increment $\Delta E(p,n)$ is the same as the number of terminal zeros in the representation of n to the base p.)

Proof: The proof follows almost at once from the definition of $E(p,n)$. For if $a_0 \neq 0$, so that n is not a multiple of p, then $n!$ contains no more factors p than does $(n - 1)!$. But if $a_0 = a_1 = \ldots = a_{s-1} = 0$, then $n = a_sp^s + \ldots + a_kp^k$, with $a_s \neq 0$, is divisible by p^s, but by no higher power of p, hence $n!$ contains exactly s more factors p than does $(n - 1)!$.

Corollary 1: As n increases, $E(p,n)$ is non-decreasing.

Proof: The *Lemma* shows that $\Delta E(p,n)$ is never negative.

Corollary 2: If there is one solution n to the equation $E(p,n) = m$, then there are exactly p solutions, namely,

$$n_i = n - a_0 + i,\ i = 0,1,\ldots,p - 1.$$

Proof: By the lemma $\Delta E(p,n_i) = 0$ for $i = 1,2,\ldots,p - 1$, because for this range of i we find n_i has a non-zero "units digit"; hence

$E(p,n_i) = E(p,n) = m$, for $i = 0,1,\ldots,p-1$. By the same reasoning $\Delta E(p,n_0) > 0$ and $\Delta E(p,n_p) > 0$; hence by *Corollary 1* there are no other values of n satisfying $E(p,n) = m$.

The fact that solutions occur in sets of p consecutive integers is also obvious from *(10.3)* for in the numerator the a_0 in n cancels with the a_0 of the sum of the "digits" leaving the expression for $E(p,n)$ actually independent of the value of a_0.

Corollary 3: There exist values of m for which $E(p,n) = m$ has no solution.

Proof: Since *Corollary 1* shows $E(p,n)$ to be non-decreasing and since there exist integers, say n^*, with $a_0 = a_1 = \ldots = a_{s-1} = 0$ and $a_s \neq 0$ with $s > 1$, for which the lemma shows $\Delta E(p,n^*) = s$, it is clear that each such saltus provides $s - 1$ values of m, namely, $m_i = E(p,n^*) - i$, for $i = 1,2,\ldots,s-1$, for which $E(p,n) = m_i$ has no solution.

The first such exceptional m for which solutions are lacking is $m = p$, which arises by taking $n^* = (100)_p = p^2$ with $s = 2$ and computing $m = E(p,n^*) - 1$ as in the proof of *Corollary 3*.

But, to return to the original problem, we must provide a way of deciding, when m is a *given* integer, whether there is a solution of $E(p,n) = m$; and if there is a solution, a way of finding it (of course, *Corollary 2*, will then provide p solutions). In the light of the *Lemma* and *Corollaries* the following rule is a natural and effective answer to these problems.

Rule to solve $E(p,n) = m$: Given the integer m, we consider

$$x = (p-1)(m+1) = (a_k \ldots a_1 a_0)_p.$$

Either $E(p,x) = m$ and p solutions can be found as in *Corollary 2*. Or $E(p,x) < m$ and we can form, in succession, the integers

$$x_1 = x + p - a_0,\ x_2 = x_1 + p,\ \ldots,\ x_i = x_{i-1} + p$$

and if s_i is the number of terminal zeros in x_i written to the base p, we can compute, in succession, the values of

$$E(p,x_1), E(p,x_2), \ldots, E(p,x_i)$$

with the aid of the formula

$$E(p,x_j) = E(p,x) + s_1 + s_2 + \ldots + s_j.$$

We continue this process (it will never be necessary to proceed beyond x_{k+1}) until either $E(p,x_i) = m$ and p solutions can be found, or until $E(p,x_{i-1}) < m < E(p,x_i)$ in which case there is no solution to $E(p,n) = m$.

(Note, however, in the eventuality of no solution, that x_i is still of some interest since x_i is the smallest integer such that p^m will divide $x_i!$ even though the exponent of p in $x_i!$ is larger than m.)

Proof of the rule: From *(10.3)* since $a_k \neq 0$ we see that $E(p,n) < n/(p-1)$; hence $E(p,x) < m + 1$ and $E(p,x) \leqq m$. The x_j of the *Rule* are selected, according to the *Lemma*, as the succession of integers following x for which $\Delta E(p,x_j) > 0$; and the formula for $E(p,x_j)$ also follows directly from the *Lemma*. To show that the process described in the rule will terminate in at most $k + 1$ steps, we define $X = x + k(p-1)$ and note, since $0 \leqq a_0 < p$, that $X < x_{k+1} = x + (k+1)p - a_0$. Since $x < p^{k+1}$ and $k(p-1) \leqq kp \leqq p^k$, we find $X < 2p^{k+1}$ or

$$X \leqq p^{k+1} + (p-1)(p^k + p^{k-1} + \dots + p + 1).$$

Therefore using *(10.3)* we can see that

$$E(p,X) \geqq ((m + 1 + k)(p-1) - (1 + (k+1)(p-1)))/(p-1) = m - 1/(p-1),$$

but since $E(p,X)$ is an integer, we conclude that $E(p,X) \geqq m$. Finally, combining these remarks with *Corollary 1*, we have $m \leqq E(p,X) \leqq E(p,x_{k+1})$.

Example 1: $p = 5$, $m = 100{,}000$. By the rule we compute $x = 4(100001) = 400004$. Then as in **6.3** we write x to the base 5, the result being as follows: $x = (100300004)_5$. By *(10.3)* we compute $E(p,x) = (400004 - (1 + 3 + 4))/4 = 99999$. Since $x_1 = x + p - a_0 = x + 5 - 4 = x + 1 = (100300010)_5$ has $s_1 = 1$, we find $E(p,x_1) = E(p,x) + 1 = m$; so the proposed problem $E(5,n) = 100000$ has 5 solutions: $x_1 = 400005$, and $x_1 + t$, $t = 1,2,3,4$.

Example 2: $p = 3$, $m = 1{,}000{,}000$. We compute $x = 2(1000001) = 2000002 = (10202121111011)_3$ and then find

$$E(p,x) = (2000002 - 14)/2 = 999994.$$

In succession we find

$$x_1 = x + 2 = (10202121111020)_3,\quad s_1 = 1,\ E(p,x_1) = 999995;$$
$$x_2 = x_1 + 3 = (10202121111100)_3,\quad s_2 = 2,\ E(p,x_2) = 999997;$$
$$x_3 = x_2 + 3 = (10202121111110)_3,\quad s_3 = 1,\ E(p,x_3) = 999998;$$
$$x_4 = x_3 + 3 = (10202121111120)_3,\quad s_4 = 1,\ E(p,x_4) = 999999;$$
$$x_5 = x_4 + 3 = (10202121111200)_3,\quad s_5 = 2,\ E(p,x_5) = 1000001.$$

Hence there is no solution to the equation $E(p,n) = m$. But the smallest integer N such that $N!$ is divisible by p^m is $N = x_5 = x + 14 = 2000016$.

EXERCISES

EX. *10.1.* Use (*10.3*) and compute $E(10,10202)$ with all the given numbers and all computations to the base 3. Check the result with the example following (*10.3*).

EX. *10.2.* Compute $E(5,101)$ by each of (*10.1*), (*10.2*), (*10.3*).

EX. *10.3.* Show that $E(2,n) \geqq E(5,n)$ for all n.

EX. *10.4.* Use EX. *10.2* and EX. *10.3* to answer the question: "In how many zeros does 101! end?"

EX. *10.5.* Compute $E(3,1001)$ by (*10.2*) and (*10.3*).

EX. *10.6.* Use **B.3** of **9.1** and (*10.1*) to prove that $(a+b)!/a!b!$ is an integer. Compare with EX. *3.7*.

EX. *10.7.* Use EX. *10.6* to prove that the product of b consecutive positive integers is divisible by $b!$.

EX. *10.8.* Define, recursively: $T_1 = 1$, and $T_i = pT_{i-1} + 1$ for $i > 1$. Then use (*10.3*) and EX. *3.2* to show, as did Kempner, that

(*10.4*) $$E(p,n) = a_1T_1 + a_2T_2 + \ldots + a_kT_k.$$

EX. *10.9.* Referring to the definitions in EX. *10.8* show that

$$T_s - i = (p-1)(T_{s-1} + T_{s-2} + \ldots + T_{s-i+1}) + pT_{s-i},$$
$$i = 2,3,\ldots,s-1.$$

EX. *10.10.* By repeated use of the division algorithm show that if $T_k \leqq m < T_{k+1}$, then there exists an integer j, $1 \leqq j \leqq k$, such that m can be written uniquely in the form

$$m = B_kT_k + B_{k-1}T_{k-1} + \ldots + B_jT_j$$

where the $B_k, B_{k-1}, \ldots, B_{j+1}, B_j$, are integers such that

$$0 < B_k < p; 0 \leqq B_i < p, \ i = k-1,\ldots,j+1; \ 0 < B_j \leqq p.$$

EX. *10.11.* Using the notation of EX. *10.10* and (*10.4*), show that if $B_j < p$, then $E(p,n) = m$ is solved by $n = B_jp^j + B_{j+1}p^{j+1} + \ldots + B_kp^k$. Using Corollary 3 and EX. *10.9*, show that if $B_j = p$, then $E(p,n) = m$ has no solution. (Kempner.)

▶ *Now this establishment of correspondence between two aggregates and investigation of the propositions that are carried over by the correspondence may be called the central idea of modern mathematics.* —W. K. CLIFFORD

CHAPTER 11

GROUPS OF TRANSFORMATIONS; MATRICES, AND DETERMINANTS

11.1. Transformations. Given a set S of elements, $x,y,\ldots$, then a *transformation* T of S is a rule which makes correspond to each x of S a unique element x' of S, which shall be written as follows: $x' = xT$ and read "x' is the T-transform of x." S will be called the *domain* of T; the set S' of all T-transforms will be called the *range* of T and will ordinarily be a *proper* subset of S.

For example, let S be the set of all integers and let T be defined by $x^2 = xT$. Then T is a transformation of S, for the square of an integer x is a unique integer x^2. The range S' of T is certainly a proper subset of S, since 2, for example, is in S, but not in S'.

Two transformations T and U of S will be defined to be *equal*, written $T = U$, if and only if $xT = xU$ for *every* x in S. In other words, $T = U$ if and only if T and U transform S in exactly the same way.

The *product* TU of two transformations T and U of S is a transformation of S defined by

(11.1) $$x(TU) = (xT)U \quad \text{for every } x \text{ of } S.$$

*Chapter 11 is in many respects a basic chapter, introducing concepts which are used in some later chapters. However, if the reader is already acquainted with determinants, especially Cramer's rule and multiplication of determinants, this chapter may be put on the supplementary list.

Since T is assumed to be a transformation of S, xT is a uniquely determined element of S; and then since U is a transformation of S and xT is in S, we see that $(xT)U$ is a uniquely determined element of S for every x of S, so TU as defined by *(11.1)* is indeed a transformation of S.

For example, let S be the set of all integers and let T and U be defined by $x^2 = xT$ and $x + 3 = xU$. By *(11.1)* we find $x(TU) = (xT)U = (x^2)U = x^2 + 3$; however, since $x(UT) = (xU)T = (x+3)T = (x + 3)^2 = x^2 + 6x + 9$, we find that $x(TU) = x(UT)$ only for $x = -1$, *not for all* x of S and therefore $TU \neq UT$.

The example just given illustrates that the operation of forming the product of two transformations of a set S is *not, in general*, a *commutative* operation. This negative observation lends more interest to the following positive result.

Q.1: The operation of forming the product of transformations is associative: thus if U,V,W are any transformations of a set S, then $(UV)W = U(VW)$.

Proof: The proof stems directly from the definition of the equality of two transformations and several applications, indicated by appropriate parentheses, of the definition *(11.1)*:
$x((UV)W) = (x(UV))W = ((xU)V)W = (xU)(VW) = x(U(VW))$. This holds for every x of S, hence $(UV)W = U(VW)$.

The most obvious transformation of all is the one which transforms each element of S into itself; it is indicated by the letter I and called the *identity* transformation; its defining property is that $x = xI$ for every x of S.

Q.2: For every transformation T of S, $IT = T = TI$.

Proof: By the defining property of I and by *(11.1)* we find $x(IT) = (xI)T = xT = (xT)I = x(TI)$, for every x of S, hence $IT = T = TI$.

A transformation T of S will be said to *have an inverse* U if and only if there exists a transformation U of S such that $TU = UT = I$.

For example, if S is the set of all integers and k is a fixed integer, then the transformation T defined by $x + k = xT$ has an inverse U defined by $x - k = xU$; for by *(11.1)* we find $x(TU) = (xT)U = (x + k)U = (x + k) - k = x = xI$ for every x in S, hence $TU = I$. Similarly, we may show $UT = I$.

Q.3: If a transformation T of S has an inverse U, then U is unique.

Proof: Suppose T has two inverses U and V, so that $TU = UT = I = TV = VT$. Then by **Q.1** and **Q.2** we find that $U = UI = U(TV) = (UT)V = IV = V$.

A transformation T of S will be said to be *one-to-one* if $x' = xT$ has one and only one solution x for *any* assigned element x' of S. In other words, the range S' of T contains *every* element of S *without repetition.*

For example, if S is the set of all integers and T is defined by $x^2 = xT$, then T is *not* one-to-one, for there is no solution x to $2 = xT$ within the set S. For another example, if S is the set of all real numbers and T is defined by $x^3 - x = xT$, then T is *not* one-to-one, for although S contains a solution x for every equation $x' = xT$, there are some cases where there is more than one solution: for example, $0 = xT$ has three solutions: $x = -1, 0, +1$.

Q.4: If T is a transformation of S, then T has an inverse if and only if T is one-to-one.

Proof: (A) **If** T is one-to-one, we may define U by saying that $x'U = x$ if and only if $x' = xT$; then U *is* a transformation of S, because since T is one-to-one we know there is one and only one x for every x' in S; and furthermore we have both $x(TU) = (xT)U = (x')U = x = xI$ for every x in S, hence $TU = I$; and $x'(UT) = (x'U)T = xT = x' = x'I$ for every x' in $S' = S$, hence $UT = I$. Thus U is an inverse of T, and by **Q.3**, U is *the* inverse of T.

(B) If T has an inverse U so that $TU = UT = I$, then for any x' of S, we find $x = x'U$ is *a* solution of $x' = xT$, since $xT = (x'U)T = x'(UT) = x'I = x'$; and there is *only one* solution, for if $xT = yT$, then we find

$$x = xI = x(TU) = (xT)U = (yT)U = y(TU) = yI = y$$

Hence T is one-to-one. This completes the proof of **Q.4**.

11.2. Groups of transformations. A set G of one-to-one transformations $T, U, \ldots$ of a set S is said to form a *group* if G has the following properties:

H.1: closure: if T and U are in G, then TU is in G;

H.2: identity: I is in G;

H.3: inverses: if T is in G. then the inverse of T is in G.

For example, let S be the set of all integers and let G be the set of *all* transformations of the type T_k defined by $x + k = xT_k$ where k is a fixed integer; then we can show that G is a group by demonstrating each of the properties **H.1**, **H.2**, **H.3**. But first we should check that the transformations T_k are one-to-one; however, T_k is precisely the example given in **11.1** of a transformation with an inverse, and hence by **Q.4**, T_k is one-to-one. Moreover, the inverse of T_k was shown to be T_{-k} and since G contains all transformations of this type, it follows that G contains T_{-k} for every integer k, and hence **H.3** is satisfied. **H.1** is satisfied for if T_k and T_m are in G, we can show that T_kT_m is in G; for

$$x(T_kT_m) = (xT_k)T_m = (x + k)T_m = (x + k) + m =$$
$$x + (k + m) = xT_{k+m}$$

for every x in S, and hence $T_kT_m = T_{k+m}$; since G contains all transformations of this type, it follows that G contains $T_kT_m = T_{k+m}$ for all integers k and m. Finally, **H.2** is satisfied, for $I = T_0$ since $xI = x = x + 0 = xT_0$ for every x in S; and since G contains all transformations of this type, G contains $I = T_0$. Thus G is a group, known as the "group of all translations" of S or as the "additive group" of S.

The student who continues into higher courses in algebra will discover that one of the most important mathematical systems is that of a group, of which the transformation groups discussed here are the primitive and, according to a celebrated theorem of Cayley, the essential examples. In the remainder of this chapter and in the next chapter we will show how these notions apply rather directly to the theory of numbers.

11.3. Linear transformations, matrices, and determinants. Consider the set S_2 of all ordered pairs (x,y) of integers x,y, defining $(x,y) = (u,v)$ if and only if $x = u$ and $y = v$. This set S_2 is called the "set of all lattice points of 2-space."

A transformation T of S_2 defined by

$$(ax + by, cx + dy) = (x,y)T \tag{11.2}$$

where a,b,c,d are fixed integers, is called a *linear transformation* of S_2.

If we write $(x',y') = (x,y)' = (x,y)T$, then the equations $x' = ax + by$, $y' = cx + dy$ give another way of describing T. It is evident that T is completely determined by specifying the values of a,b,c,d and ordering them correctly, hence a convenient representation for T is to write

$$T = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \tag{11.3}$$

where the 2-by-2 square array on the right is called a *2-by-2 matrix* and a,b,c,d are called the *elements* of the matrix. The first column of T determines x' and the second, y'; the first row of T shows the coefficients of x and the second, the coefficients of y.

Let U be a linear transformation of S_2 defined by $(x,y)U = (a_1x + b_1y, c_1x + d_1y)$ and represented, according to (11.3), by the matrix

$$U = \begin{pmatrix} a_1 & c_1 \\ b_1 & d_1 \end{pmatrix}.$$

M.1: If T and U are linear transformations of S_2, then $T = U$ if and only if $a = a_1$, $b = b_1$, $c = c_1$, $d = d_1$. (In other words, $T = U$ if and only if the matrices representing T and U have their corresponding elements equal.)

Proof: By definition $T = U$ if and only if $(x,y)T = (x,y)U$ for every (x,y) in S_2. In particular we may take $(x,y) = (1,0)$ to see that

$$(a,c) = (1,0)T = (1,0)U = (a_1,c_1)$$

if and only if $a = a_1$, $c = c_1$; and we may take $(x,y) = (0,1)$ to see that

$$(b,d) = (0,1)T = (0,1)U = (b_1,d_1)$$

if and only if $b = b_1$, $d = d_1$.

M.2: If T and U are linear transformations of S_2, then TU is a linear transformation of S_2, and the matrix representing TU is given by

$$TU,= \begin{pmatrix} aa_1 + cb_1 & ac_1 + cd_1 \\ ba_1 + db_1 & bc_1 + dd_1 \end{pmatrix} \tag{11.4}$$

Proof: By definition (11.1) we find

$$\begin{aligned}(x,y)(TU) &= ((x,y)T)U = (ax + by, cx + dy)U \\ &= ((ax + by)a_1 + (cx + dy)b_1, (ax + by)c_1(cx + dy)d_1) \\ &= ((aa_1 + cb_1)x + (ba_1 + db_1)y, (ac_1 + cd_1)x + (bc_1 + dd_1)y)\end{aligned}$$

Since a,b,c,d,a_1,b_1,c_1,d_1 are integers, so are $aa_1 + cb_1, ba_1 + db_1, ac_1 + cd_1, bc_1 + dd_1$, and TU is seen to have the correct form to be a linear transformation, and by *(11.3)* the matrix corresponding to TU is precisely that given in *(11.4)*.

At first it would seem difficult to memorize *(11.4)*, but there is an easy device called "matric multiplication" or "row-by-column multiplication" which is defined purposely in such a way as to give exactly the result *(11.4)*.

Rule for matric multiplication: To find the element in the ith row and jth column of the matrix TU, find the sum of the products of corresponding elements of the ith row of T and the jth column of U. Thus we find

$$(11.4)' \qquad \begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} a_1 & c_1 \\ b_1 & d_1 \end{pmatrix} = \begin{pmatrix} aa_1 + cb_1 & ac_1 + cd_1 \\ ba_1 + db_1 & bc_1 + dd_1 \end{pmatrix}.$$

For example, to find the element in the first row and second column of TU, take the first row of T, namely: (a,c) and the second column of U, namely: (c_1,d_1), multiply together their first elements to obtain ac_1 and multiply their second elements to obtain cd_1, and finally take the sum of these products to obtain $ac_1 + cd_1$.

Such a complicated rule for finding the product of two matrices, if presented by itself, might seem highly artificial; but in view of the preceding discussion of products of transformations, applied in particular to products of linear transformations, the rule has quite a satisfactory motivation, resulting in a product-preserving one-to-one correspondence between linear transformations and matrices.

Just as the product of transformations is, in general, not a commutative operation, so the product of matrices might be expected not, in general, to be commutative. To illustrate this remark we compute, by *(11.4)'*, the following products:

$$\begin{pmatrix} 2 & 3 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 8 & 3 \\ -1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix}.$$

By **M.1** we see that these results are not equal.

However, the following result is true, and by our previous work is easily established.

M.3: Multiplication of matrices is an associative operation.

Proof: Each matrix corresponds to a linear transformation and each product of two matrices corresponds to the product of the corresponding linear transformations, as we have previously seen in (*11.3*), (*11.4*), *11.4*)′. However, by **Q.1** in **11.1**, the product of linear transformations is an associative operation; hence the multiplication of matrices is also an associative operation.

Of course, this theorem can also be proved by direct computation from the product rule (*11.4*)′, but the proof here given is much more elegant—particularly if the theory of matrices is extended, as it can be, to other than 2-by-2 matrices.

M.4: The identity transformation is a linear transformation and the corresponding matrix is given by

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Proof: By definition $(x,y)I = (x,y)$, hence

$$x' = x = 1 \cdot x + 0 \cdot y,\ y' = y = 0 \cdot x + 1 \cdot y,$$

so that I is seen to be a linear transformation, and by (*11.3*) the proper matrix is the one displayed above.

Since the following relation $TI = T = IT$ holds by **Q.2** for all linear transformations, it follows by (*11.3*), (*11.4*), (*11.4*)′ that the same relation holds for matrices.

M.5: A linear transformation T of S_2, represented by the matrix (*11.3*), has an inverse if and only if $ad - bc$ is a unit, namely $+1$ or -1; and the inverse transformation is a linear transformation.

Proof: Consider the relations $x' = ax + by$, $y' = cx + dy$. By elimination we find

(*11.5*) $(ad - bc)x = dx' - by', \quad (ad - bc)y = -cx' + ay'.$

Let $t = ad - bc$. By **Q.4** of **11.1**, we know that T has an inverse if and only if T is one-to-one, i.e., if and only if $(x',y') = (x,y)T$ has a unique solution (x,y) for every (x',y') of S_2. Examining (*11.5*), we see that in order that there be a unique solution, allowing *fractions*, it is necessary and sufficient that $t \neq 0$. But in order for every pair of integers x',y' to determine a *unique* pair of *integers* x,y, we can show that it is necessary and sufficient that $t = +1$ or -1.

If t does not divide a, or if t does not divide b, then if in (*11.5*) we substitute $x' = 0$, $y' = 1$, we find $tx = -b$, or $ty = a$; hence in the one case, y cannot be an integer, or in the other case, x cannot be an integer. Similarly, if t does not divide c or d, we may substitute $x' = 1, y' = 0$ in (*11.5*) to obtain $tx = d, ty = -c$; so that in one case, y cannot be an integer, or in the other case, x cannot be an integer. Thus if T is one-by-one, t must divide a,b,c,d. Hence t^2 must divide $ad - bc = t$, say $t^2u = t$; then $tu = 1$, so that t must be a unit and either $t = +1$ or $t = -1$.

Conversely, if $t = \pm 1$, then $t^2 = +1$, hence from (*11.5*) we find

$$x = tdx' - tby', \quad y = -tcx' + tay'. \tag{11.6}$$

From (*11.6*) it is clear that every pair of integers x',y' leads to a unique pair of integers x,y; hence T is one-to-one and has an inverse. Moreover, the form of (*11.6*) is such that the inverse of T is seen to be a linear transformation U represented by a matrix

$$U = \begin{pmatrix} td & -tc \\ -tb & ta \end{pmatrix}. \tag{11.7}$$

Definition: If T is the matrix given in (*11.3*), then the function $d(T) = ad - bc$ is called the *determinant* of T.

In terms of this definition, **M.5** may be reworded as follows. A linear transformation T of S_2 has an inverse U if and only if the matrix representing T has a unit determinant $t = d(T) = \pm 1$. Moreover, $d(U) = (td)(ta) - (-tb)(-tc) = t^2(ab - bc) = t = d(T)$.

M.6: The set G of all linear transformations T of S_2, such that $d(T) = \pm 1$, form a group, called the "lattice group."

Proof: We must show that G has properties **H.1, H.2, H.3. H.1** is satisfied because if T_1 and T_2 are linear transformations of S_2 with $d(T_1) = \pm 1$ and $d(T_2) = \pm 1$, then on the one hand we know by **M.2** that T_1T_2 is a *linear* transformation; and on the other hand we know by **M.5** that T_1 has an inverse U_1 so that $T_1U_1 = U_1T_1 = I$, and that T_2 has an inverse U_2 so that $T_2U_2 = U_2T_2 = I$, whence we can show that U_2U_1 is the inverse of T_1T_2, for using **Q.1** or **M.3**, we have

$$(T_1T_2)(U_2U_1) = T_1(T_2(U_2U_1)) = T_1((T_2U_2)U_1) = T_1(IU_1) = T_1U_1 = I$$

and similarly, $(U_2U_1)(T_1T_2) = I$; but since T_1T_2 *has an inverse*, it follows by **M.5** that $d(T_1T_2) = \pm 1$. Since G includes *all* linear transformations of S_2 of unit determinant, if follows that G satisfies property **H.1**.

H.2 is satisfied because we noted in **M.4** that I is a linear transformation and we note now that $d(I) = 1\cdot 1 - 0\cdot 0 = 1$, so that G contains I.

H.3 is satisfied because we noted in **M.5** that the inverse U of a linear transformation T is a linear transformation and we noted in the paragraph preceding **M.6** that $d(U) = d(T)$; since G includes *all* linear transformations of unit determinant, it follows that G includes U. This completes the proof of **M.6**.

M.7: If T and U are linear transformations of S_2, then $d(TU) = d(T)d(U)$.

Proof: Using the notation of **M.2**, we find by direct computation, omitting a few terms which cancel, that

$$\begin{aligned} d(TU) &= (aa_1 + cb_1)(bc_1 + dd_1) - (ba_1 + db_1)(ac_1 + cd_1) \\ &= aa_1dd_1 + cb_1bc_1 - ba_1cd_1 - db_1ac_1 \\ &= (ad - bc)(a_1d_1 - b_1c_1) = d(T)d(U). \end{aligned}$$

11.4. Significance of the lattice group for Diophantine problems. We have already indicated in Chapter 1, that if in a given problem we restrict our attention to those solutions which are integers, then we can agree to describe this restriction by calling our problem a *Diophantine* problem.

If a Diophantine problem involves two variables x and y, then a very powerful simplifying device may often be to replace the two given variables by suitably chosen linear combinations with integer coefficients of two new variables x' and y', so chosen that every pair of integer values of the first two variables will determine a unique pair of integer values for the two new variables, and conversely. In other words we would like to use a *linear transformation* that is *completely reversible in integers.*

But such a transformation is a linear transformation of S_2 and according to **M.5**, it is of necessity one belonging to the lattice group discussed in **M.6**, i.e., a linear transformation of unit determinant and hence possessing an inverse. Moreover from the closure property of the lattice group, the product of any number of these transformations, that is to say the application in convenient sequences of any number of these simplifying transformations, is equal to another of the same kind, that is to say, can be accomplished by a single such transformation.

We observe from EX. *5.3* that if we are given a and b, then c and d can be found so that $ad - bc = \pm 1$, if and only if a and b are relatively prime. But this still leaves a good deal of freedom, as we shall show in the next chapter; and from the many corresponding linear transformations that are, as we have shown in **M.5**, *completely reversible in integers*, we can frequently find one or more that will greatly simplify the form of a Diophantine problem, *without losing or gaining* even one extra solution. This technique is well illustrated in Chapter 13.

EXERCISES

EX. *11.1.* Show that the rule T defined by $xT = x(x + 1)/2$ is a transformation of the set S of all integers, but that T does not have an inverse.

EX. *11.2.* If S is the set of all integers and T and U are defined by $xT = |x|$ and $xU = x + 2$, find and compare TU and UT.

EX. *11.3.* If $T_1,T_2,\ldots,T_n$ are transformations of a group G and $U_1,U_2,\ldots,U_n$ are the corresponding inverses, prove by induction that the inverse of $T_1T_2\ldots T_n$ is $U_n\ldots U_2U_1$ (the inverse of a product is the product of the inverses *in reverse order*).

EX. *11.4.* A group G is called *commutative* if $T_1T_2 = T_2T_1$ for *every* pair of transformations T_1 and T_2 in G. Show that the translation group in **11.2** is commutative. Show that the lattice group in **11.3** is not commutative.

EX. *11.5.* Given the matrices

$$T = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix}, \quad U = \begin{pmatrix} 3 & 1 \\ 2 & -5 \end{pmatrix},$$

find TU, UT, $T^2 = TT$, and the determinants of the five matrices.

EX. *11.6.* Prove **M.3** by direct computation.

EX. *11.7.* Although, in general, for matrices $TU \neq UT$, show that, always, for determinants $d(TU) = d(UT)$.

EX. *11.8.* For matrices T and U as in **M.2** and for any integer q make the following definitions:

$$T + U = \begin{pmatrix} a + a_1 & c + c_1 \\ b + b_1 & d + d_1 \end{pmatrix}, \quad qT = \begin{pmatrix} qa & qc \\ qb & qd \end{pmatrix}.$$

Show that $d(qI - T) = q^2 - (a + d)q + t$.
Show that $T^2 - (a + d)T + tI = 0 \cdot I$.

Exercises

EX. *11.9* Show that the four linear transformations of S_2 corresponding to the following matrices, form a group, the "cyclic group of order 4":

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad A^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A^3 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

EX *11.10.* Show that the four linear transformations of S_2 corresponding to the following matrices, form a group, the "four-group":

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

EX. *11.11.* Using **M.7**, give a different proof that the lattice group of **M.6** has closure.

EX. *11.12.* Show that the set G of *all* linear transformations T of S_2 with $d(T) = +1$ is a group, the "direct" lattice group.

EX. *11.13.* For every fixed pair of integers k,m define a *translation* T_{km} of S_2 by the following rule: $(x,y)T_{km} = (x + k,y + m)$. Show that the set G of *all* translations of S_2 is a commutative group.

> ▶*Whether he drew much or little from the work of his predecessors, it is certain that the* ARITHMETICA *of Diophantus has exercised a profound influence on the development of number theory.* —R. D. CARMICHAEL

CHAPTER 12*

DIOPHANTINE EQUATIONS OF THE FIRST DEGREE

12.1. The equation $ax + by = n$. The problem which we propose here is that of finding all pairs of integers x,y which satisfy the equation

$$ax + by = n, \tag{12.1}$$

where a,b,n are given integers.

Because we restrict our attention to *integer solutions* we shall describe the problem as a *Diophantine problem.* (However, this is a modern agreement since Diophantus would have sought all fractional solutions.) Sometimes such a problem is called a problem in *indeterminate analysis* because when there are not as many equations as unknowns there may be infinitely many solutions, and when a solution is not unique, it is frequently described as indeterminate.

Theorem: The equation $ax + by = n$ *has a solution* in integers, say x^*,y^*, if and only if $d = (a,b)$ divides n. In case there is a solution, *every* solution is given by

$$x = x^* + Bt, \quad y = y^* - At, \tag{12.2}$$

where A and B are defined by $a = Ad$ and $b = Bd$, and t is an arbitrary integer.

*Chapter 12 is a basic chapter.

Proof: (A) If there exist integers x^*,y^* such that $ax^* + by^* = n$, then every common divisor of a and b, including d, is a divisor of the left side of the equation and hence a divisor of n.

(B) Conversely, suppose $d = (a,b)$ is a divisor of n, say $n = Nd$. By the Euclid algorithm as in **5.3** we can find integers X and Y such that $aX + bY = d$. Then $x^* = XN$, $y^* = YN$ provides one solution of the equation (*12.1*), because

$$n = Nd = a(NX) + b(NY) = ax^* + by^*.$$

(C) Let x^*,y^* be one solution, so that $ax^* + by^* = n$. Then as shown in (A) we must have $n = Nd$ where $d = (a,b)$ and $a = Ad$, $b = Bd$. As in (B) we know $d = aX + bY$ for properly chosen integers X and Y; hence dividing by d, we find $1 = AX + BY$. Therefore as in EX. *5.3*, we have $(A,B) = 1$. Let x,y be any solution of (*12.1*) so that $ax + by = n$. By subtraction and rearrangement we find $a(x - x^*) = b(y^* - y)$. If we substitute $a = Ad$, $b = Bd$, and cancel d, we find that $A(x - x^*) = B(y^* - y)$. Since A divides $B(y^* - y)$ but $(A,B) = 1$, it follows by EX. *6.2* that A divides $y^* - y$, say $At = y^* - y$, then $y = y^* - At$. With this agreement, we discover upon substitution and canceling A, that $x - x^* = Bt$, or $x = x^* + Bt$. Thus we have found that every possible solution of (*12.1*) can be written in the form (*12.2*).

(*D*) Finally, we need to show that every pair of integers of the form (*12.2*) is a solution of (*12.1*), but this is easily verified by direct substitution, since for every value of t we find

$$ax + by = a(x^* + Bt) + b(y^* - At) = ax^* + by^* + AdBt - BdAt = n.$$

Corollary 1: The basic solution, denoted by x^*,y^* in (*12.2*), may be chosen in just one way, say X^*,Y^*, such that $0 \leqq X^* < |B|$.

Proof: Since all solutions of (*12.1*) are given by (*12.2*), we need show only that there is one and just one value of the integral parameter t such that $0 \leqq X^* = x^* + Bt < |B|$. Using fractions and supposing that $B > 0$, we find that

$$-x^*/B \leqq t < (|B| - x^*)/B = 1 - x^*/B,$$

hence t is the unique integer in the interval from $-x^*/B$ to $1 - x^*/B$. Similarly, when $B < 0$, t is the unique integer satisfying $-x^*/B \geqq t > -1 - x^*/B$.

Corollary 2: All solutions of $ax + by = n$ in *positive* integers x,y, if there are any, can be found by solving $x > 0$, $y > 0$, simultaneously, with x,y given by (*12.2*), to find those values of t which are suitable.

If we assume that a and b are both positive, there will be at most a finite number of solutions of $ax + by = n$ in positive integers, so this restriction makes the problem a little more interesting and is considered the "cricket" requirement for most problems of this kind.

Example: Find the smallest positive integer m such that

$$533x + 299y = 10000 + m$$

has a solution in integers, and for this value of m find how many solutions the equation has in positive integers.

We begin by the Euclid algorithm to find $d = (a,b)$ where $a = 533$, $b = 299$, realizing that this work should be preserved to help find x^*,y^*.

```
            1
      -----------
  299 | 533
      | 299      1
      ---------------
        234 | 299
            | 234      3
            ---------------
               65 | 234
                  | 195      1
                  ---------------
                     39 | 65
                        | 39      1
                        --------------
                          26 | 39
                             | 26      2
                             --------------
                               13 | 26
                                  | 26
                                  ---------
                                      0
```

Since $d = 13$ we must find the smallest positive value of m such that $n = 10000 + m$ is divisible by 13, for this is the implication of our theorem. Since $10000 = 769 \cdot 13 + 3$, the smallest suitable value is $m = 10$. Then $n = 10010$.

From the calculations of the algorithm we find

$$\begin{aligned} d = 13 &= 39 - 26 = (234 - 3\cdot 65) - (65 - 39) \\ &= 234 - 4\cdot 65 + (234 - 3\cdot 65) = 2\cdot 234 - 7\cdot 65 \\ &= 2\cdot 234 - 7(b - 234) = 9\cdot 234 - 7b = 9(a - b) - 7b = 9a - 16b, \end{aligned}$$

hence $X = 9$, $Y = -16$.

Since $a = 533 = 41d$, $b = 299 = 23d$, $n = 10010 = 770d$, we have $A = 41, B = 23, N = 770$. Therefore the complete solution may be based on the particular solution

$$x^* = NX = 6930, \quad y^* = NY = -12320,$$

and would appear as follows:

$$x = 6930 + 23t, \quad y = -12320 - 41t.$$

However, *Corollary 1* suggests that the solution may be put in a much more convenient form. We therefore solve $0 \leqq 6930 + 23t < 23$, for $t = -301$, and compute $X^* = 7$, $Y^* = 21$. With this new basic solution the complete solution is as follows:

$$x = 7 + 23t, \quad y = 21 - 41t$$

(there is no harm in retaining t as the symbol for the parameter, if we agree from this point on to use the new, more convenient solution).

Finally, to determine the solutions in positive integers, we must consider simultaneously the inequalities

$$x = 7 + 23t > 0, \quad y = 21 - 41t > 0.$$

These inequalities simplify, if we use fractions, to the form

$$-7/23 < t < 21/41;$$

so that one and only one suitable *integer* value of t, namely, $t = 0$, can be found. Thus the only solution of $533x + 299y = 10010$ in positive integers is $x = 7, y = 21$.

12.2. Computation of X,Y so that $d = (a,b) = aX + bY$. Since the solution of (*12.1*) has been shown in the preceding section to depend upon finding X,Y so that $d = (a,b) = aX + bY$, it may be desirable, both for theoretical and computational purposes, to have a mechanical scheme for performing eliminations from the equations of the Euclid algorithm.

Toward this end, we reconsider the Euclid algorithm and after setting $r_{-2} = a$ and $r_{-1} = b$, we write the following equations:

(*12.3*) $$r_{i-2} = q_i r_{i-1} + r_i, \qquad i = 0,1,2,\ldots,k;$$

where $$0 < r_k < r_{k-1} < \ldots < r_1 < r_0 < |b| \leqq |a|$$

and where $r_{k-1} = q_{k+1}r_k$, so that $d = r_k = (a,b)$ as in 5.3.

Theorem: If x_i and y_i are defined to be a solution of

(*12.4*) $$(-1)^i r_i = ax_i - by_i, \qquad \text{for } -1 \leqq i \leqq k,$$

where the r_i are determined by the Euclid algorithm (*12.3*), then solutions x_i, y_i may be entered, recursively, in the following chart:

q		q_0	q_1	$\dots$	q_k
x	x_{-1}	x_0	x_1	$\dots$	x_k
y	y_{-1}	y_0	y_1	$\dots$	y_k

by remembering $x_{-1} = 0$, $y_{-1} = 1$, $x_0 = 1$, $y_0 = q_0$, and then computing, in succession, the values of

(*12.5*) $x_{i+1} = x_{i-1} + x_i q_{i+1}, \quad y_{i+1} = y_{i-1} + y_i q_{i+1}, \qquad i = 1,2,\dots,k.$

In particular, when $i = k$, $X = (-1)^k x_k$ and $Y = (-1)^{k+1} y_k$ provide a solution of $d = r_k = (a,b) = aX + bY$.

Proof: When $i = -1$, we have $(-1)^{-1} r_{-1} = -b = a(0) - b(1)$ so that $x_{-1} = 0$ and $y_{-1} = 1$ are solutions of (*12.4*).

When $i = 0$, we have $(-1)^0 r_0 = r_0 = a(1) - bq_0$ so that $x_0 = 1$ and $y_0 = q_0$ are solutions of (*12.4*).

To establish (*12.5*) we may use an incomplete induction on i with i running from 1 to k.

(I) When $i = 1$, we have $(-1)^1 r_1 = -r_1 = q_1 r_0 - b$ or $-r_1 = q_1(a - bq_0) - b$, so that $x_1 = q_1$, $y_1 = 1 + q_0 q_1$ are solutions of (*12.4*). These solutions may be rewritten as $x_1 = x_{-1} + x_0 q_1$ and $y_1 = y_{-1} + y_0 q_1$ in agreement with (*12.5*).

(II) Suppose as the induction hypothesis that (*12.5*) provides correct solutions of (*12.4*) for values of $i = 1,2,\dots,j$, where $j \leqq k - 1$. Then starting from (*12.3*) and using the induction hypothesis we may write

$$r_{j+1} = r_{j-1} - q_{j+1} r_j = \{(ax_{j-1} - by_{j-1}) + q_{j+1}(ax_j - by_j)\}(-1)^{j+1}$$

Then

$$(-1)^{j+1} r_{j+1} = a(x_{j-1} + x_j q_{j+1}) - b(y_{j-1} + y_j q_{j+1})$$

so that we find (*12.4*) satisfied when we choose

$$x_{j+1} = x_{j-1} + x_j q_{j+1}, \quad y_{j+1} = y_{j-1} + y_j q_{j+1}$$

but these are exactly the solutions specified by (*12.5*) when $i = j + 1$.

From (I) and (II) it follows that (*12.5*) gives correct solutions of (*12.4*) for $i = 1,2,\dots,k$, completing the proof of the theorem.

For the example given in **12.1** the computations of this theorem would appear as follows:

q		1	1	3	1	1
x	0	1	1	4	5	9
y	1	1	2	7	9	16

For example: $y_3 = y_1 + y_2q_3 = 2 + 7(1) = 9$.

In a later chapter we shall find that the recursion formulas (*12.5*) are fundamental computational devices in the study of continued fractions.

12.3. Other algorithms. In the discussion of the preceding sections we have supposed that a "standard" Euclid algorithm has been used, i.e., an algorithm in which the remainders are "least positive" remainders. Lamé has shown that the number of divisions in such a standard algorithm will not exceed five times the number of digits in the smaller number b (with computations in the base 10).

But there is no compulsion to use least positive remainders. Thus if $a = qb + r$, $0 < r < b$, we also have the possibility of using $a = (q + 1)b + r'$, where $|r'| = b - r$ satisfies $0 < |r'| < b$. If one, say r^*, of r and $|r'|$ is the smaller, then a second step of the algorithm, $b = q_1r^* + r_1$, $0 \leqq |r_1| < r^*$ would seem to have the possibility of a smaller remainder than in the standard algorithm and hence the discovery in fewer steps of $d = (a,b)$. An algorithm which at each step uses that one of the remainders which is smallest in absolute value has been shown by Kronecker to be at least as short as any other algorithm (and often, as examples show, there is considerable gain over the standard algorithm in the use of this "least absolute value" algorithm). Further discussion of these matters may be found in the book by Uspensky and Heaslet listed in **1.3.**

Even if some Euclid algorithm other than the standard one is used, a list of the successive quotients can still be used in (*12.5*) to find X,Y solving $d = aX + bY$, for there was nothing about the derivation of those formulas to require positive remainders. But, of course, different algorithms may lead to different solutions X,Y.

For example, consider the following algorithm:

$$\begin{array}{rrrr}
 & 2 & & \\
\cline{1-2}
299 & \multicolumn{1}{|r}{533} & & \\
 & \multicolumn{1}{|r}{598} & -5 & \\
\cline{2-3}
 & -65 & \multicolumn{1}{|r}{299} & \\
 & & \multicolumn{1}{|r}{325} & 3 \\
\cline{3-4}
 & & -26 & \multicolumn{1}{|r}{-65} \\
 & & & \multicolumn{1}{|r}{-78} \\
\cline{4-4}
 & & & 13.
\end{array}$$

Then using (*12.5*) with $q_0 = 2$, $q_1 = -5$, $q_2 = 3$, we find

q		$+2$	-5	$+3$
x	0	1	-5	-14
y	1	$+2$	-9	-25

hence we may take $X = -14$, $Y = 25$. Inasmuch as $-14 = 9 - 23$, $25 = -16 + 41$, this solution is seen to be compatible, in the light of (*12.2*), with the solution $X = 9$, $Y = -16$ previously obtained.

Still another computational device may be suggested, based essentially on the least absolute value algorithm. To avoid trivial cases suppose that b does not divide a and that $0 < b < |a|$. By the division algorithm compute

(*1*) $$a - qb = r, \quad 0 < r < b;$$

(*2*) $$a - (q + 1)b = -r', \quad 0 < r' = b - r < b.$$

Let $D = (r,r')$. From (*1*) and (*2*) it is clear that $d = (a,b)$ divides both r and r', hence d divides D. But $b = r + r'$ and $a = qb + r = (q + 1)r + r'$, so it is clear that $D = (r,r')$ divides both a and b, hence D divides d. Therefore $D = \pm d$. Thus the problem of finding $d = (a,b)$ may be replaced by the problem of finding $d = (r,r')$, where both $0 < r < b < |a|$ and $0 < r' < b < |a|$ so that smaller numbers are involved. Perhaps by inspection d as well as s and t can be found so that $d = sr - tr'$. Then if we add s times (*1*) to t times (*2*) we have

(*3*) $$(s + t)a - (sq + t(q + 1))b = d,$$

so that $X = s + t$, $Y = -(s + t)q - t$ solves $aX + bY = d$.

If the numbers r and r' are not sufficiently small to provide obvious solutions of d,s,t, the process can be repeated beginning with r and r'.

The process will terminate as usual in a finite number of steps, because of the decreasing non-negative character of the numbers involved, with the discovery of a trivial case where one member of the pair is 0 and the other is d.

In practice there is no need to memorize the formulas for X and Y. For example, to compute $d = (533,299)$ we proceed step by step as indicated by the arrows:

$$\begin{array}{lll} \left.\begin{array}{l}(1)\ 533-1\cdot299=234\\(2)\ \underline{533}-2\cdot299=\underline{-65}\end{array}\right\}\rightarrow & \left.\begin{array}{l}(1)'\ 234-3\cdot65=39\\(2)'\ \underline{234}-4\cdot65=\underline{-26}\end{array}\right\}\rightarrow & \begin{array}{l}\text{but here,}\\ \text{obviously (?),}\end{array}\\ (3)\left\{\begin{array}{l}9\cdot533-16\cdot299=13\\ X=9,\ Y=-16\end{array}\right. \leftarrow & \left\{\begin{array}{l}(3)'\ 2\cdot234-7\cdot65=13\\ s=2,\ t=7\end{array}\right. \leftarrow & \left\{\begin{array}{l}39-26=13=(39,26)\\ s'=1,\ t'=1\end{array}\right.\end{array}$$

EXERCISES

EX. *12.1.* Show that $213x + 441y = 10002$ has solutions in integers but none where both x and y are positive integers.

EX. *12.2.* Show that if $ax + by = n$ has any solutions in integers we may assume the problem reduced to the form $Ax + By = N$ where $(A,B) = 1$ and $B > 0$.

In the following three exercises find $d = (3713,1343)$ and find one solution X,Y of $d = 3713X + 1343Y$:

EX. *12.3.* Use the "standard" algorithm and (*12.5*).

EX. *12.4.* Use the "least absolute value" algorithm and (*12.5*).

EX. *12.5.* Use the (*1*),(*2*),(*3*) method at the end of 12.3.

EX. *12.6.* Obtain a formula for *all* solutions of the equation $d = (3713,1343) = 3713X + 1343Y$.

EX. *12.7.* Consider $ax + by = c$ where a and b are fixed positive integers such that $(a,b) = 1$. If c is a positive integer and $K(c)$ denotes the number of solutions of the equation in positive integers x,y, and if s,t is any solution of $bt - as = 1$, show that

$$K(c) = [tc/a] - [sc/b] - E(c),$$

where $E(c) = 1$ or 0 according as c is or is not a multiple of a, and $[x]$ is the bracket-function of Chapter 9. (P. Barlow, 1811.)

EX. *12.8.* With the notation of EX. *12.7*, if T is a given positive integer, show that all integers c such that $K(c) = T$ must be in the interval from $L(T) = (T - 1)ab + a + b$ to $U(T) = (T + 1)ab$ and that both $L(T)$ and $U(T)$ are values of c which satisfy $K(c) = T$. (A. D. Wheeler, 1860.)

EX. *12.9.* Using the notation of the preceding two exercises let $N(T)$ be the number of values of c such that $K(c) = T$. Prove that $N(T)$ is independent of T, if T is a positive integer.

EX. *12.10.* Using EX. *12.8*, show that the values of c such that $K(c) = T$ are consecutive only when $a = 1$ or $b = 1$.

> ►*When you come to a hard or dreary passage, pass it over; and then come back to it after you have seen its importance or found the need for it further on.* —G. CHRYSTAL

CHAPTER 13*

MORE DIOPHANTINE EQUATIONS OF THE FIRST DEGREE

13.1. The equation $ax + by + cz = n$. The problem of this section is to find all triples of integers x,y,z, if there are any, which will satisfy the equation

$$ax + by + cz = n, \tag{13.1}$$

where a,b,c,n are given integers.

It is very easy to prove that (*13.1*) has *a solution* in integers if and only if $d = (a,b,c)$ is a divisor of n, for we can parallel every step of parts (*A*) and (*B*) of the proof in **12.1**. But it is not quite so simple to find formulas which will give *every* solution.

We will resort to the device discussed in **11.4** setting $y = AY + BZ$ and $z = CY + DZ$, where the integers A,B,C,D are at our disposal, except that in the light of **M.5** of **11.3**, we insist that $AD - BC = \pm 1$, so that the transformation is completely reversible in integers. Upon substituting for y and z in (*13.1*) we obtain

$$ax + (bA + cC)Y + (bB + cD)Z = n. \tag{13.2}$$

It is our intention to choose B and D in such a way that $(B,D) = 1$ and that $(bB + cD) = 0$, so that the equation (*13.2*) will reduce to one involving *only two unknowns*, x and Y, which we can solve com-

[*] Chapter 13 is a supplementary chapter.

pletely as in **12.1**: and adding *any* value of Z, we will have *all* solutions x,Y,Z of *(13.2)*; then since $(B,D) = 1$, we will be able to find A and C so that $AD - BC = 1$, and our transformation will therefore be completely reversible in integers and we will be able to pass back to *all* solutions x,y,z of *(13.1)*.

The details in this program are as follows: let $(b,c) = k$, then $b = Rk$ and $c = Sk$, with $(R,S) = 1$. If we take $B = S$ and $D = -R$, then $(B,D) = 1$, and $bB + cD = RkB + SkD = RkS - SkR = 0$ as desired. By inspection or by the Euclid algorithm or its modifica- tions as suggested in **12.2** and **12.3**, we then determine A and C so that $AD - BC = \pm 1$. Next we solve the *depressed* equation

$$(13.3) \qquad ax + (bA + cC)Y = n$$

by the method of **12.1** obtaining x and Y as functions of the integral parameter t. Finally we take Z as an arbitrary integral parameter and solve for

$$y = AY + BZ, \quad z = CY + DZ.$$

In our final answer x involves the parameter t, while y and z involve both the parameters t and Z. The presence of two arbitrary parameters in the answer is not unexpected, if we consider that the original equation is doubly indeterminate.

13.2. The example of Customer Jones. Manor Mouse Jones was given \$103 by his wife to exactly cover the cost, including 3% sales tax, of some items A, @ \$7; some items B, @ \$3; and some items C, @ \$15. Alas, that poor Jones! When he got to the store he could only recall the names of the items and that he was to get at least one of each. Find the probability, if Jones spent all his money for certain numbers of items A, B, C, and paid the tax, that he would get exactly his wife's order.

Stripped of inessentials, the Jones problem is that of solving the Diophantine equation

$$7x + 3y + 15z = 100, \quad a = 7, \quad b = 3, \quad c = 15, \quad n = 100,$$

for the number of solutions in positive integers.

Since $(a,b,c) = 1$ which divides n, the problem has solutions in integers. Since $(b,c) = 3 = k$, we find $B = 5$, $D = -1$ and by inspection choose $A = 1$, $C = 0$, so that $AD - BC = -1$. Since $bA + cC = 3$, we find the depressed equation like *(13.3)* to be

$$7x + 3Y = 100.$$

Since $7(1) + 3(-2) = 1$, a basic solution $x^* = 100$, $Y^* = -200$ is easily found; then the general solution is

$$x = 100 + 3t, \quad Y = -200 - 7t;$$

but this is needlessly complicated, so we use $t = T - 33$ to write

$$x = 1 + 3T, \quad Y = 31 - 7T.$$

Then from $y = AY + BZ$, $z = CY + DZ$, we have the complete solution of the original equation given by

$$x = 1 + 3T, \quad y = 31 - 7T + 5Z, \quad z = -Z.$$

We can check our work by direct substitution:

$7(1 + 3T) + 3(31 - 7T + 5Z) + 15(-Z) = 100$, for all T and Z.

Finally, to find all the solutions in positive integers we must study the inequalities $x > 0$, $y > 0$, $z > 0$, requiring these to hold simultaneously; furthermore, the parameters T and Z must be limited to integer values. First we find $T \geqq 0$, $31 + 5Z > 7T \geqq 0$, $Z < 0$, and then from $0 > Z > 31/5$, we conclude that $Z = -1, -2, -3, -4, -5$, or -6.

When $Z = -1$, $0 \leqq 7T < 26$ limits T to be 0,1,2,3.
When $Z = -2$, $0 \leqq 7T < 21$ limits T to be 0,1,2.
When $Z = -3$, $0 \leqq 7T < 16$ limits T to be 0,1,2.
When $Z = -4$, $0 \leqq 7T < 11$ limits T to be 0,1.
When $Z = -5$, $0 \leqq 7T < 6$ limits T to be 0.
When $Z = -6$, $0 \leqq 7T < 1$ limits T to be 0.

Thus there are fourteen solutions in positive integers. Hence (assuming Jones to be something of a mathematician and able to compute this figure) the required probability is $1/14$, one success out of fourteen possibilities. Jones had better return to his spouse for written information!

As an example of one of the solutions referred to above, we may take $Z = -2$, $T = 2$, then $x = 7$, $y = 7$, $z = 2$, and we check that $7(7) + 3(7) + 15(2) = 49 + 21 + 30 = 100$.

13.3. One equation in four or more unknowns. It is reasonably clear that the procedure explained in **13.1** will depress a Diophantine equation of the first degree involving, say, four variables to one involving only three variables and then the procedure in **13.1** will apply directly to complete the solution.

For example, let us consider the equation

$$14x + 6y + 30z + 90w = 200.$$

Since $(14,6,30,90) = 2$ which divides 200, there are solutions in integers (see EX. *13.5*). In fact, the equation might as well be replaced immediately by the equivalent equation

$$7x + 3y + 15z + 45w = 100.$$

Then the transformation $z = -2Z + 3W$, $w = Z - W$, has a determinant -1 and depresses the equation to

$$7x + 3y + 15Z = 100.$$

Availing ourselves of the complete solution to the latter equation, as found in the preceding section, we have in terms of integral parameters U, V, W, a complete solution of the given equation, expressed as follows:

$$x = 1 + 3U, \quad y = 31 - 7U + 5V, \quad z = 2V + 3W, \quad w = -V - W.$$

If we desire only positive solutions we are led to the following inequalities:

$$0 \leqq 7U < 31 + 5V; \quad -2V/3 < W < -V; \quad -6 \leqq V \leqq -1;$$

for which there are only four solutions:

$$V = -4,\ U = 0 \text{ or } 1,\ W = 3; \quad V = -5,\ U = 0,\ W = 4;$$
$$V = -6,\ U = 0,\ W = 5;$$
$$(x,y,z,w) = (1,11,1,1); \quad (4,4,1,1); \quad (1,6,2,1); \quad (1,1,3,1).$$

13.4. Systems of Diophantine equations of the first degree. If a system of two (or more) Diophantine equations of the first degree must be solved, the plan is to obtain the complete solution in integers (if any such solution exists) of one of these equations, and then to substitute this solution into the second equation to obtain a new second equation involving the parameters of the solution of the first equation. Inasmuch as the parameters must take on only integer values, a complete solution of this new second equation represents a complete solution of the system. (More equations can be handled by further successive substitutions.)

It should be noted that a given Diophantine system may be indeterminate (having solutions involving one or more parameters), determinate (having one and only one solution), or inconsistent (having no solution in integers).

To one who is familiar with the various elimination procedures used in the theory of equations which reduce a system of m linear equations in n unknowns, $m \leqq n$, to equations in $n - m + 1$ unknowns, the procedure for Diophantine equations may seem familiar, but awkward. But it will be recognized as a procedure which at every step collects together every possible solution in integers. If one were interested in fractional solutions, say, this method would actually work (for the parameters could then be allowed to be fractions), but would be needlessly complicated.

For an example, let us consider the Diophantine system

$$\begin{cases} 7x + 3y + 15z = 100, \\ \ \ x + 5y + \ \ 3z = 120. \end{cases}$$

As in **13.2** the complete solution of the first equation is

$$x = 1 + 3T, \quad y = 31 - 7T + 5Z, \quad z = -Z;$$

and if we substitute these results in the second equation, we find that the integers T and Z must satisfy

$$16T - 11Z = 18.$$

Since $16(-2) - 11(-3) = 1$, this new equation is solved completely by $T = -36 + 11m$, $Z = -54 + 16m$; or as a matter of convenience we may set $m = M + 3$ and obtain the answer

$$T = -3 + 11M, \quad Z = -6 + 16M.$$

Returning to the variables of the original system, we find the complete solution of the system in integers to be a one-parameter solution, given as follows:

$$x = -8 + 33M, \quad y = 22 + 3M, \quad z = 6 - 16M.$$

We shall want to return to this type of problem again after the notion of congruences has been introduced, for the theory of congruences will be found to simplify many of the procedures explained above.

EXERCISES

EX. *13.1.* Find the complete solution in integers of $3x + 7y + 10z = 102$ and show that there are just twenty solutions in positive integers.

EX. *13.2.* Find the complete solution in integers of the system:

$$3x + 7y + 10z = 102, \ 2x + 3y + 4z = 46;$$

and determine whether there are solutions in positive integers.

EX. *13.3.* Reconsider the problem of Customer Jones, in **13.2**, supposing the prices were A @ \$13, B @ \$7, C @ \$18.

EX. *13.4.* Using the notation of **13.1**, show that the coefficient of Y in (*13.3*) is either $+k$ or $-k$.

EX. *13.5.* Prove by induction on $k \geqq 2$ that

$$a_1x_1 + a_2x_2 + \ldots + a_kx_k = n$$

has a solution in integers if and only if $d = (a_1, a_2, \ldots, a_k)$ is a divisor of n.

EX. *13.6.* If a rooster is worth 5 coins, if a hen is worth 3 coins, and if three chicks are worth 1 coin, how many roosters, hens, chicks, 100 in all and at least some of each kind, will be worth 100 coins. (The Chinese problem of "One Hundred Fowls.")

EX. *13.7.* Show that the number of solutions in positive integers of

$$x + 2y + 3z = n$$

is given by $1 + [n(n-6)/12]$, where the brackets indicate the bracket-function of Chapter 9.

> ▶*In mathematics, as in other fields, to find oneself lost in wonder at some manifestation is frequently the half of a new discovery.*
>
> —P. G. L. DIRICHLET

CHAPTER 14*

PYTHAGOREAN TRIPLETS

14.1. The Diophantine equation $x^2 + y^2 = z^2$. As an introduction to the subject of quadratic Diophantine equations it is natural to try to find the complete solution in integers of the Pythagorean equation $x^2 + y^2 = z^2$, for, as every student of geometry and trigonometry knows, the variables x,y,z can be interpreted as the sides and hypotenuse of a right triangle, and it is particularly convenient for "nice" problems or for the drawing of a right angle to have a fund of whole number solutions such as the well-known 3,4,5 and 5,12,13. But our object here is a bit more profound—we wish to find formulas exhibiting *all* integral solutions of the equation.

To begin with we will observe that if x,y,z is an integral solution of $x^2 + y^2 = z^2$ and if $(x,y,z) = d$, $x = Xd$, $y = Yd$, $z = Zd$, then $X^2 + Y^2 = Z^2$ with $(X,Y,Z) = 1$. Conversely, if $(x,y,z) = 1$ and $x^2 + y^2 = z^2$, then for any integer k, the integers $X = xk$, $Y = yk$, $Z = zk$ satisfy the relation $X^2 + Y^2 = Z^2$. If we describe a solution for which $(x,y,z) = 1$ as a "primitive triplet" and a solution for which $(x,y,z) = d > 1$ as an "imprimitive triplet," then the situation which we have just investigated may be described as follows:

Every primitive triplet generates a family of imprimitive triplets; and conversely, every imprimitive triplet may be obtained from a

*Section 14.1 is a basic section, while section 14.2 is of a supplementary nature.

properly chosen primitive triplet. Hence to find *all* solutions in integers of the Pythagorean equation it will suffice to find *all primitive* solutions.

In the following discussion we shall need two lemmas:

L.1: Given $(a,b) = 1$, then $(a^s,b^t) = 1$ for all positive integers s and t.

Proof: Given $(a,b) = 1$, we know that there exist integers x and y so that $ax + by = 1$. Then there also exist integers X and Y so that $a^sX + b^tY = 1$. For we have $(ax + by)^{s+t} = 1$, and by EX. *3.7* we know that $(ax + by)^{s+t}$ can be written as the sum of $t + 1$ terms each involving a^s as a factor and of s terms each involving b^t as a factor. But $a^sX + b^tY = 1$ implies $(a^s,b^t) = 1$, as was to be proved. (See also EX. *5.3* and EX. *5.7*.)

L.2: The square of an even number is a multiple of 4; the square of an odd number is one more than a multiple of 8.

Proof: (A) $(2k)^2 = 4k^2$.

(B) $(2k + 1)^2 = 4k^2 + 4k + 1 = 4k(k + 1) + 1$. Since either k or $k + 1$ is even it follows that $(2k + 1)^2 = 8m + 1$.

We shall begin by assuming that the equation $x^2 + y^2 = z^2$ does have some primitive solutions for which $(x,y,z) = 1$ and we shall seek to describe these solutions more completely. For convenience we have divided the argument into eight numbered steps as follows:

(*1*) Not only must $(x,y,z) = 1$, but also we must have $(x,y) = 1$, $(x,z) = 1$, $(y,z) = 1$. For example, suppose $(x,y) = d$. Then from $x^2 + y^2 = z^2$, it follows that d^2 divides z^2. If p is a prime factor of d, then by the *Fuadamental Lemma* of **6.1**, p must divide z; but then p must divide $(x,y,z) = 1$ which is a contradiction; hence d must be without prime factors; in other words, $d = 1$. Similarly, $(x,z) = 1$ and $(y,z) = 1$.

(*2*) The integers x and y must be of opposite parity. For if x and y were both even, we would have a contradiction of $(x,y) = 1$ which was established in step (*1*). And if x and y were both odd we could apply **L.2** and have $x^2 + y^2 = 8X + 1 + 8Y + 1 = 8(X + Y) + 2$; but according to **L.2** there exists no integer z with a square which is 2 more than a multiple of 8. We may, therefore, assume x to be even, say $x = 2X$, and y to be odd. Then, of course, z must be odd, and we may define new variables r and s, which will be integers, as follows:

$$z - y = 2s,\ z + y = 2r; \quad \text{or} \quad y = r - s,\ z = r + s.$$

(3) The new variables r and s must be of opposite parity and relatively prime. Suppose $(r,s) = d$; then from the last equations of step (2), it follows that d divides both y and z; but by step (1) we know that $(y,z) = 1$, hence $d = 1$. Furthermore, unless r and s are of different parity, y and z will not be odd, as agreed upon in step (2).

(4) Using the new variables we may replace the original equation by a new equation of simpler structure. We rewrite $x^2 + y^2 = z^2$ as $x^2 = z^2 - y^2 = (z + y)(z - y)$. By substitution we obtain $4X^2 = 4rs$ which we simplify to the form $rs = X^2$.

(5) Steps (3) and (4) imply $r = u^2$, $s = v^2$, $X = uv$, where u and v are integers that are relatively prime and of opposite parity. Suppose $(r,X) = m$, $r = um$, $X = vm$, then $(u,v) = 1$ as in EX. 5.4. Also by **L.1**, we have $(u,v^2) = 1$. The equation $rs = X^2$ obtained in step (4) takes the form $ums = v^2m^2$, or $us = v^2m$. Since $(u,v^2) = 1$, it follows from EX. 6.2 that $m = wu$; then $s = v^2w$ and $r = u^2w$. Therefore w divides both r and s; however, by step (3) we know $(r,s) = 1$, hence $w = 1$ and $r = u^2$, $s = v^2$, $X = uv$. Also by step (3), r and s must be of opposite parity, hence by **L.2** we see that u and v must be of opposite parity.

(6) Hence all primitive solutions of $x^2 + y^2 = z^2$, *if there are any*, must have the following form:

$$x = 2uv, \quad y = u^2 - v^2, \quad z = u^2 + v^2$$

where $(u,v) = 1$ and where u and v are of opposite parity.

This step is, of course, merely a summary of steps (2) and (5), with emphasis upon the possibility that there may be no primitive solutions of the original equation.

(7) Every set of integers x,y,z defined as in step (6) *is a solution* of the Pythagorean equation. This check is easy since by elementary algebra we find that

$$(2uv)^2 + (u^2 - v^2)^2 = (u^2 + v^2)^2.$$

(8) Every set of integers x,y,z defined as in step (6) is a *primitive* solution. Let $(2uv, u^2 - v^2, u^2 + v^2) = d$. Then since $(u + v)^2 = (u^2 + v^2) + (2uv)$ and $(u - v)^2 = (u^2 + v^2) - (2uv)$, it follows that d divides $(u + v)^2$ and $(u - v)^2$. Let p be a prime factor of d. Then by **6.1**, p must divide both $u + v$ and $u - v$. Hence p divides both $2u = (u + v) + (u - v)$ and $2v = (u + v) - (u - v)$. Therefore p divides $(2u,2v)$. However $(2u,2v) = 2(u,v) = 2$, because in step (6) we agree to take $(u,v) = 1$. Therefore p divides 2; but $p \neq 2$ for in step (6) we agree to take u and v of opposite parity so that $u^2 + v^2$

is odd, hence d is odd, and p must be odd; hence d has no prime factors and $d = 1$. So $(x,y,z) = 1$ and the x,y,z given in step (6) do form a primitive triplet.

Thus the formulas in step (6) represent all primitive solutions of the Pythagorean equation and from these primitive solutions all solutions can be generated as explained in the preliminary remarks. This completes the solution of the Diophantine equation $x^2 + y^2 = z^2$.

In step (6) by insisting that $0 < v < u$ we can make all of x,y,z positive. A short table of examples follows:

v	u	x	y	z	v	u	x	y	z
1	2	4	3	5	2	5	20	21	29
1	4	8	15	17	2	7	28	45	53
1	6	12	35	37	2	9	36	77	85
1	8	16	63	65	3	4	24	7	25
2	3	12	5	13	3	8	48	55	73

14.2. The inradius of Pythagorean triplets. Let us consider a Pythagorean triplet (x,y,z) of positive integers x,y,z such that $x^2 + y^2 = z^2$ and let r designate the radius of the inscribed circle of the corresponding right triangle as in Figure 7. Let us call r the "inradius" of the triplet.

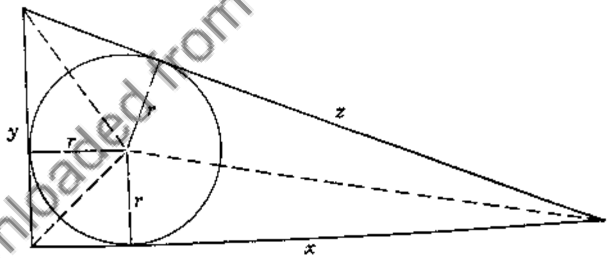


FIGURE 7

R.1: Given the Pythagorean triplet x,y,z, then its inradius r is an integer.

Proof: There are two rather obvious ways to express the area A of the triangle leading to the following equations:

$$r(x + y + z) = 2A = xy.$$

By the discussion in **14.1** we know that all positive integer solutions of $x^2 + y^2 = z^2$ have the form

$$x = k2uv,\ y = k(u^2 - v^2),\ z = k(u^2 + v^2),$$

where k,u,v are positive integers with $(u,v) = 1$, with u and v of different parity, and with $u > v$. By substitution in the displayed equation we obtain the relation $2rk(u^2 + uv) = 2k^2uv(u^2 - v^2)$ which simplifies readily to $r = kv(u - v)$, thus proving r to be an integer.

With certain agreements we can write every positive integer r uniquely in the following form

$$(14.1) \qquad r = 2^a p_1^{a_1} p_2^{a_2} \dots p_n^{a_n}, \qquad a \geqq 0, \quad n \geqq 0,$$

where p_i is an odd prime, $a_i \geqq 1$, and $2 < p_1 < p_2 < \dots < p_n$. If $n = 0$, it is understood that $r = 2^a$. If $a = 0$, $2^0 = 1$.

We will let the number-theoretic function $P(r)$ represent the number of distinct positive primitive Pythagorean triplets having r as the corresponding inradius. (We consider triplets corresponding to congruent triangles to be the same: e.g., $x = 3$, $y = 4$, $z = 5$ is considered the same as $x = 4$, $y = 3$, $z = 5$.)

R.2: If r is given by *(14.1)*, then $P(r) = 2^n$.

Proof: The integers x,y,z as used in the proof of **R.1** will form a primitive triplet if and only if $k = 1$. Then the equation for r takes the form $r = v(u - v)$, where the second factor, $u - v$, must be odd, since u and v are of different parity. Furthermore, from $(u,v) = 1$ it follows that $(v,u - v) = 1$ as in EX. *5.8*. Conversely, if $r = VU$ where V and U are positive integers with U odd and with $(V,U) = 1$, then the equations $V = v$, $U = u - v$ may be solved for $u = U + V$, $v = V$ where u and v are positive integers such that $u > v$, $(u,v) = 1$, and u and v are of opposite parity. Hence all positive primitive Pythagorean triplets having r as inradius are found by factoring r in all possible ways as a product VU of two relatively prime factors V and U of which U is odd. That this procedure leads to no repetitions follows from EX. *14.1*.

Since U must be odd, either $U = 1$ or U contains odd prime factors. If r is given by *(14.1)* and if U has the factor p_i, then U must have the factor $p_i^{a_i}$ in order that $r = VU$ with $(V,U) = 1$. Hence the number of choices of U, and, therefore, the value of $P(r)$, is exactly the same as the number $\tau(r')$ of factors of $r' = p_1p_2 \dots p_n$. But by the formula developed in Chapter 8 we find

$$\tau(r') = (e_1 + 1)(e_2 + 1) \dots (e_n + 1) = 2^n$$

since each exponent e_i in r' has the value 1. Since $P(r) = \tau(r')$, this completes the proof of **R.2**.

In particular, we have shown that $P(r)$ is positive for every r and that the range of $P(r)$ consists exactly of all powers of 2, including $2^0 = 1$.

As an example, consider $r = 15$ for which $n = 2$ so that $P(15) = 4$. In tabular form the solutions are as follows:

V	U	u	v	x	y	z
15	1	16	15	480	31	481
5	3	8	5	80	39	89

V	U	u	v	x	y	z
3	5	8	3	48	55	73
1	15	16	1	32	255	257

Let $N(r)$ represent the total number of distinct positive Pythagorean triplets, *not necessarily primitive*, having r as the corresponding inradius.

R.3: If r is given by *(14.1)*, then

$$N(r) = (a + 1)(2a_1 + 1)(2a_2 + 1)\ldots(2a_n + 1).$$

Proof: From the formula $r = kv(u - v)$ in **R.1** we see that k must be a divisor of r. Then $d = r/k$ is the inradius of a primitive Pythagorean triplet. Conversely, if d is a divisor of r, say $r = kd$, then a Pythagorean triplet of inradius r can be found by magnifying a primitive Pythagorean triplet of inradius d by the factor of proportionality k; and distinct solutions for d lead to distinct solutions for r; also different values of d lead to distinct solutions for r. It follows that the desired function $N(r)$ can be obtained as follows:

$$N(r) = \sum P(d)$$

where the summation is extended over all positive divisors d of r.

This is a perfect occasion to apply the theorem on multiplicative functions developed in **8.3**, for we can show $P(d)$ to be a multiplicative function. In fact by **R.2** we know $P(d) = 2^{\nu(d)}$ where $\nu(d)$ is the number of distinct odd prime factors of d. When $(a,b) = 1$, the integers a and b can have no odd prime factors in common, hence $\nu(ab) = \nu(a) + \nu(b)$. Therefore if $(a,b) = 1$, then

$$P(ab) = 2^{\nu(ab)} = 2^{\nu(a)+\nu(b)} = 2^{\nu(a)}2^{\nu(b)} = P(a)P(b),$$

so that $P(d)$ is a multiplicative function.

It follows from **8.3** that $N(r) = \Sigma P(d)$ is also a multiplicative function, so to find the precise form for $N(r)$ we need only to investigate $N(p^a)$ for primes p.

When $p = 2$, the only even prime, we find

$$N(2^a) = P(1) + P(2) + P(2^2) + \ldots + P(2^a) =$$
$$2^0 + 2^0 + 2^0 + \ldots + 2^0 = a + 1.$$

When p is odd, we find

$$N(p^a) = P(1) + P(p) + P(p^2) + \ldots + P(p^a) = 2^0 + 2^1 + 2^1 + \ldots + 2^1 = 2a + 1.$$

Combining these results with the fact that $N(r)$ is multiplicative, we arrive at the formula displayed in **R.3.**

For a numerical example, we take $r = 15$. Here $a = 0$, $a_1 = 1$, $a_2 = 1$, hence $N(15) = (0 + 1)(2 + 1)(2 + 1) = 9$. Then we compute

$k = 15$, $d = 1$, $P(1) = 1$ with (4,3,5) leading to (60,45,75);

$k = 5$, $d = 3$, $P(3) = 2$ with (8,15,17) leading to (40,75,85), and with (24,7,25) leading to (120,35,125);

$k = 3$, $d = 5$, $P(5) = 2$ with (12,35,37) leading to (36,105,111), and with (60,11,61) leading to (180,33,183);

$k = 1$, $d = 15$, $P(15) = 4$ leading to the four primitive triplets given in the previous example.

EXERCISES

EX. *14.1.* Show that distinct values of u,v as in step (6) of 14.1 lead to distinct Pythagorean triplets.

EX. *14.2.* Find all primitive Pythagorean triplets having $x = 60$.

EX. *14.3.* Show that it is impossible to find a primitive Pythagorean triplet with side T where T is even but not a multiple of 4. Show that all other integers T can be the side of at least one primitive Pythagorean triplet.

EX. *14.4.* Let T be written in the form (*14.1*) and let $S(T)$ indicate the number of primitive Pythagorean triplets of side T. Show that $S(T) = 2^{n-1}, 0, 2^n$, according as $a = 0$, $a = 1$, $a > 1$. (E. Bahier.)

EX. *14.5.* Follow the method of 14.1 and obtain the complete solution of the Diophantine equation $x^2 + 2y^2 = z^2$, supplying the proofs of the following steps:

(0) To find all solutions it will suffice to find all primitive solutions for which $(x,y,z) = 1$.

(1) Primitive solutions must have $(x,y) = 1$, $(x,z) = 1$, $(y,z) = 1$.

(2) Both x and z must be odd and y even. Define $y = 2Y$, $z + x = 2r$, $z - x = 2s$.

(3) From $(x,z) = 1$ it follows that $(r,s) = 1$.

(4) $x^2 + 2y^2 = z^2$ becomes $2Y^2 = rs$.

(5) Either (I) $r = 2R^2$, $s = S^2$, with $(2R,S) = 1$; or (II) $r = S^2$, $s = 2R^2$, with $(2R,S) = 1$.

(6) Every primitive solution must have the following form

$$x = \pm(2R^2 - S^2), \quad y = 2RS, \quad z = 2R^2 + S^2, \quad (2R,S) = 1.$$

(*7*) Every x,y,z of the form (*6*) is a *solution* of $x^2 + 2y^2 = z^2$.

(*8*) Every x,y,z of the form (*6*) is a *primitive* solution.

EX. *14.6*. Graph the function $P(r)$ in **R.2** of **14.2** for values of r from 1 through 30.

EX. *14.7*. Find the eight smallest solutions of $P(r) = 8$.

EX. *14.8*. Graph the function $N(r)$ in **R.3** of **14.2** for values of r from 1 through 30.

EX. *14.9*. Find the eight smallest solutions of $N(r) = 6$.

EX. *14.10*. Show that $N(r) = t$ has a solution r for every positive integer t, but that the solution is unique if and only if t is a power of 2, including $2^0 = 1$.

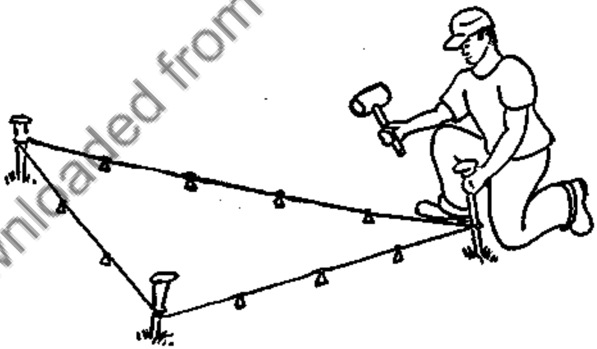
EX. *14.11*. Investigate the meaning and origin of the word "harpedonaptae."

EX. *14.12*. Let r_x,r_y,r_z designate the radii of the three escribed circles of the triangle corresponding to a Pythagorean triplet. Prove that r_x, r_y, r_z are all integers. If r is the inradius, show that $rr_z = r_xr_y$ and $r_z = r + r_x + r_y$.

EX. *14.13*. Prove that one side "x" of a Pythagorean triplet always has an extra factor 4 as compared with the other side "y." Let $N_x(r)$, $N_y(r)$, $N_z(r)$ indicate the total numbers of Pythagorean triplets such that $r = r_x$, $r = r_y$, $r = r_z$, respectively. Beginning with EX. *14.12*, prove that

$$N(r) = N_x(r) + N_y(r) + N_z(r) + 1,$$

where $N(r)$ is defined as in **14.2**.



▶*I have found a*
ingly beautiful the

CHAPTER 15*

FERMAT'S METHOD OF DES

15.1. Fermat's "last theorem." On the m
Diophantus, Fermat wrotê that the Diophantine

$$x^n + y^n = z^n$$

is impossible of solution in *positive* integers x,y,z fo
he had found a truly remarkable way of proving
that, unfortunately, the margin was not large e
writing out the proof. (The restriction "posi
essential, for otherwise $x = 0$, $y = z$ is a solution
certainly a trivial solution.)

This general problem is still unsolved and is
Fermat's "last theorem." By special methods th
solved as far as $n = 616$. Despite its special and
this problem has been the source of some of th
and analysis, as efforts to solve Fermat's problem
methods and exposed hidden pitfalls of older met

It is easy to show that the problem will be co
can be shown that Fermat's conjecture is true f
$n = p$ and for $n = 4$. Suppose that n is com
$n = kp$ where p is an odd prime for which Fermat

*The author regards sections **15.1**, **15.2**, **15.3** as basic and
as supplementary.

is known to be true; then the theorem is also true for n; for if we suppose the theorem not true for n, then there exist positive integers x,y,z such that $x^n + y^n = z^n$, but this is a contradiction of the assumption that the theorem is true for p, inasmuch as it shows $X = x^k$, $Y = y^k$, $Z = z^k$ to be positive integers for which

$$X^p + Y^p = Z^p.$$

Similarly, if the theorem is true for $n = 4$, then the theorem is true for $n = 2^k$, $k \geqq 2$.

The easiest case in which we can prove the "last theorem" is the case of $n = 4$ where we can employ a method known as Fermat's "method of descent" which may possibly have been the "remarkable way" which he mentioned.

15.2. Fermat's "method of descent." If a proposition $P(n)$ is true for some positive integers, then there is a least positive integer for which $P(n)$ is true. (It is explained in a later chapter how this assertion is really just another version of mathematical induction.) But suppose it can be shown that the assumed truth of $P(n)$ always implies the truth of $P(n')$ where n' is a positive integer less than n. Then a contradiction has been reached and the proposition $P(n)$ must be false. This method of proof, depending as it does on descending from the positive integer n to the smaller positive integer n', has long been given the name: the "method of descent."

Usually we employ this method to prove the falsity of a given proposition. But sometimes we can use the method in a positive way, showing that the "descent" is possible until we reach a certain type of integer; if for this special type of integer the proposition is true, then we can reverse the argument and "ascend" to all solutions of the proposition.

Both these uses of the "method of descent" will be illustrated in the next sections.

15.3. The relation $x^4 + y^4 = z^2$ is impossible in positive integers. To prove the proposition used as the title of this section we shall use the method of descent.

Let us assume that the equation $x^4 + y^4 = z^2$ does have some solutions in positive integers and let x,y,z be a specific one of these solutions. If $(x,y) = d$, then $x = Xd$, $y = Yd$, and $z = Zd^2$; and furthermore $X^4 + Y^4 = Z^2$. Hence if we describe the problem we are studying as $P(z)$ and if $d > 1$, then we have already shown that the truth of $P(z)$ implies the truth of $P(Z)$ where Z is a positive integer less than z. But if $d = 1$, more argument will be required. However, if $(x,y) = 1$ it follows that $(x^2,y^2,z) = 1$, hence x^2,y^2,z is a primitive Pythagorean triplet. Therefore by the results of **14.1** we know that we may write $x^2 = 2rs$, $y^2 = r^2 - s^2$, $z = r^2 + s^2$, with $(r,s) = 1$ and with r and s of different parity. But we must not choose r to be even for then s and y are odd and the relation $y^2 + s^2 = r^2$ is a contradiction of **L.2** of **14.1**. Hence we have $(r,2s) = 1$. Then as in step (5) of **14.1** we argue from $(r,2s) = 1$ and $x^2 = 2rs$ that $r = R^2$ and $s = 2S^2$. The relation $(2S^2)^2 + y^2 = (R^2)^2$ and the condition $(2S^2,y,R^2) = 1$ show that $2S^2,y,R^2$ is another primitive Pythagorean triplet, so again using **14.1** we write $2S^2 = 2uv$, $y = u^2 - v^2$, $R^2 = u^2 + v^2$, with $(u,v) = 1$ and u and v of opposite parity. Again as in step (5) of **14.1** we see from $(u,v) = 1$ and $S^2 = uv$ that $u = U^2$ and $v = V^2$. Therefore $R^2 = U^4 + V^4$ and we have arrived at another solution of the equation of this section; in other words the truth of $P(z)$ implies the truth of $P(R)$ and (what is of critical concern for the method) R is a positive integer less than z because

$$R < R^4 + 4S^4 = r^2 + s^2 = z.$$

Hence as explained in **15.2** we have successfully demonstrated the descent and are therefore caught in a contradiction; the only way out of the contradiction is in the decision that $x^4 + y^4 = z^2$ is *impossible* in positive integers.

Corollary: Fermat's "last theorem" is true for $n = 4$. For if $x^4 + y^4 = z^4$ in positive integers, we would have $x^4 + y^4 = (z^2)^2$ contradicting the principal result of this section.

15.4. The relation $x^4 - 8y^4 = z^2$ is impossible in positive integers. To establish the proposition used as the title of this section we shall use an argument that is seen on closer inspection to be merely a rephrasing of the method of descent.

Let us assume that the equation $x^4 - 8y^4 = z^2$ does have some solutions in positive integers and that among all these x,y,z is a solution with a minimum value of x. If $(x,y) = d$, then $x = Xd$, $y = Yd$, and $z = Zd^2$ with X,Y,Z providing a solution with $X < x$, unless $d = 1$. Hence we assume $(x,y) = 1$. But also $(x,2y) = 1$; for if x were even, z^2 would be a multiple of 8 and z would be a multiple of 4; but this would require y to be even, contradicting $(x,y) = 1$. Hence $(z,2y^2,x^2) = 1$ and if we write the given equation in the form $z^2 + 2(2y^2)^2 = (x^2)^2$, we find that $z,2y^2,x^2$ is a primitive solution of the equation studied in EX. *14.5.* Therefore we may write $z = \pm(2R^2 - S^2)$, $2y^2 = 2RS$, $x^2 = 2R^2 + S^2$, $(2R,S) = 1$. Then since $(R,S) = 1$ and $y^2 = RS$ we may write $R = u^2$, $S = v^2$, with $(2u^2,v^2) = 1$. But then $x^2 = 2(u^2)^2 + (v^2)^2$ so that v^2,u^2,x is also a primitive solution of the equation studied in EX. *14.5.* Therefore we may write $v^2 = \pm(2M^2 - N^2)$, $u^2 = 2MN$, $x = 2M^2 + N^2$, $(2M,N) = 1$. From $(2M,N) = 1$ and $u^2 = 2MN$ we may write $M = 2Y^2$, $N = X^2$, $(2Y,X) = 1$. Since v and N are odd, it follows from **L.2** of **14.1** that $v^2 = 2M^2 - N^2 = 8Y^4 - N^2$ is impossible, and it is the second case $v^2 = N^2 - 2M^2 = X^4 - 8Y^4$ which must hold. But this is an equation of the same type with which we started, and it has a solution in which $X \leqq X^2 = N < N^2 + 2M^2 = x$. Hence we have reached a contradiction of the minimal property supposedly enjoyed by x. The only resolution of this contradiction is in the theorem that the equation $x^4 - 8y^4 = z^2$ has no solution in positive integers.

Corollary: The relation $x^4 + 2y^4 = z^2$ is impossible in positive integers.

Proof: If we suppose that there exist positive integers x,y,z such that $x^4 + 2y^4 = z^2$, then we may write

$$(x^4 - 2y^4)^2 = (x^4 + 2y^4)^2 - 8y^4 = z^4 - 8y^4,$$

and inasmuch as we can show that $x^4 - 2y^4 \neq 0$, it follows that one of the triplets z, y, $\pm(x^4 - 2y^4)$ provides a solution in positive integers that contradicts the principal theorem of this section.

The missing detail may be treated as follows: suppose there exist positive integers x and y such that $x^4 - 2y^4 = 0$ or $x^4 = 2y^4$. If $(x,y) = d$, then $x = Xd$, $y = Yd$, $(X,Y) = 1$, and $X^4 = 2Y^4$. Hence X must be even, say $X = 2S$ and $Y^4 = 8S^4$; but then Y must be even, which contradicts $(X,Y) = 1$.

15.5. Chains of solutions of $x^4 - 2y^4 = z^2$. In contrast to the preceding result about the equation $x^4 + 2y^4 = z^2$, it appears that the equation $x^4 - 2y^4 = z^2$ does have some solutions in positive integers; for example, we find by inspection that $x = 3$, $y = 2$, $z = 7$ is a primitive solution, and from a primitive solution as many other solutions as desired can be obtained by taking $X = kx$, $Y = ky$, $Z = k^2z$ where k is any integer. Conversely, any solution of the equation is a kind of multiple of a primitive solution: for if x,y,z is a solution and $(x,y) = d$, then $x = Xd$, $y = Yd$, $z = Zd^2$, $(X,Y) = 1$, and $X^4 - 2Y^4 = Z^2$, with $(X,Y,Z) = 1$. Let us, therefore, seek all primitive solutions.

If $(x,y,z) = 1$ and $x^4 - 2y^4 = z^2$, then $(z,y^2,x^2) = 1$ and $z^2 + 2(y^2)^2 = (x^2)^2$ so that z,y^2,x^2 is a primitive solution of the equation studied in EX. *14.5*, and we can write

$$z = \pm(2R^2 - S^2), \quad y^2 = 2RS, \quad x^2 = 2R^2 + S^2, \quad (2R,S) = 1.$$

It follows from $(2R,S) = 1$ and $y^2 = 2RS$ that $R = 2u^2$, $S = v^2$, $(2u,v) = 1$. Then $(v^2,2u^2,x) = 1$ and $(v^2)^2 + 2(2u^2)^2 = x^2$ so that v^2, $2u^2$, x is another primitive solution of the equation in EX. *14.5* and we can write

$$v^2 = \pm(2P^2 - Q^2), \quad 2u^2 = 2PQ, \quad x = 2P^2 + Q^2, \quad (2P,Q) = 1.$$

From $(2P,Q) = 1$ and $u^2 = PQ$ it follows that $P = a^2$, $Q = b^2$, $(2a,b) = 1$. We note that both v and b are odd and then proceed to consider the two cases for v^2.

Case 1: $v^2 = b^4 - 2a^4$. Since v and b are odd, this case is possible only if a is even; but then $y = 2uv = 2abv$, so that y is a multiple of 4. Since $a \leq ab = u < 2uv = y$, it is possible in this case to descend to a solution of the original equation with smaller "y" value. If this were the only kind of descent, then we would have a proof, just as in **15.3** and **15.4**, that the given equation has no solution. But we note that this kind of descent will fail as soon as we reach a "y" which is not a multiple of 4. We must turn our attention, therefore, to the other case for v^2.

Case 2: $v^2 = 2a^4 - b^4$. Since v and b are odd, this case is possible only if a is odd; but then $y = 2uv = 2abv$ is a multiple of 2, but not a multiple of 4.

Thus every solution of $x^4 - 2y^4 = z^2$ is a member of a chain of solutions of this same equation, descending with respect to "y," and ending at a primitive solution of an equation $2a^4 - b^4 = v^2$ whose complete solution we will now attempt, hoping to build backward from the solutions of this latter equation to solutions of the given problem.

Matters of oddness and evenness as in **L.2** of **14.1** show that a primitive solution of $2a^4 - b^4 = v^2$ must have a,b,v all odd and

$(b,v) = 1$. If we set $b^2 + v = 2T$ and $b^2 - v = 2U$, then $b^2 = T + U$ and $v = T - U$ and it follows that $(T,U) = 1$. Our equation takes the new form

$$2a^4 = b^4 + v^2 = (T + U)^2 + (T - U)^2 = 2T^2 + 2U^2,\ a^4 = T^2 + U^2.$$

Hence T,U,a^2 form a primitive Pythagorean triplet, although not necessarily a triplet of positive integers for U may be negative or 0 (however, this last case occurs only when $T = v = b = 1$). With this understanding we may write

$$T = m^2 - n^2, \quad U = 2mn, \quad a^2 = m^2 + n^2, \quad (m,n) = 1,$$

with m and n of opposite parity, but n not necessarily positive (in particular, we shall be interested in the case $n = 0$ and $m = 1$ corresponding to a minimum positive value of a). Since $b^2 = T + U = m^2 - n^2 + 2mn$ and b is odd, it follows that m must be odd, rather than n, so we have $(m,2n) = 1$. Since $(m,n,a) = 1$ and $m^2 + n^2 = a^2$ we see that m,n,a is a primitive Pythagorean triplet so we set

$$m = g^2 - h^2, \quad n = 2gh, \quad a = g^2 + h^2, \quad (g,h) = 1,$$

with g and h of opposite parity. But we also have $(b,n,m + n) = 1$ and $b^2 + 2n^2 = (m + n)^2$ so that $b,n,m + n$ is a primitive solution of the equation in EX. *14.5*, hence we set

$$b = \pm(2A^2 - B^2), \quad n = 2AB, \quad m + n = 2A^2 + B^2, \quad (2A,B) = 1.$$

By comparison we have $gh = AB$ and $g^2 - h^2 = 2A^2 + B^2 - 2AB$ and these equations give us the clue to how to proceed.

Let $(g,B) = D$ with $g = Da_1$, $B = Db_1$, and $(a_1,b_1) = 1$. Then from $gh = AB$ we have $a_1h = Ab_1$ and since $(a_1,b_1) = 1$ we find $h = Eb_1$, $A = Ea_1$. Since $(g,h) = 1$ it follows that $(D,E) = 1$. From $g^2 - h^2 = 2A^2 + B^2 - 2AB$ we find by substitution and rearrangement $E^2(2a_1^2 + b_1^2) - 2EDa_1b_1 = D^2(a_1^2 - b_1^2)$. If we multiply both sides of this equation by $2a_1^2 + b_1^2$ and then add to each side the term $D^2a_1^2b_1^2$ we find that

$$\{E(2a_1^2 + b_1^2) - Da_1b_1\}^2 = D^2(2a_1^4 - b_1^4).$$

Hence it follows (see EX. *15.1*) that $2a_1^4 - b_1^4$ must be a perfect square, say $2a_1^4 - b_1^4 = v_1^2$.

Ordinarily $a_1 \leq Da_1 = g < g^2 + h^2 = a$, so that it is usually possible to descend from a solution with a given "a" to a solution with a smaller "a." The one exception is the case $g = 1$, $h = 0$, when we find $a_1 = a = 1$. Thus a descent from any primitive solution of $2a^4 - b^4 = v^2$ to the basic solution $a = 1$, $b = 1$, $v = 1$ is always possible.

It remains to show the method of ascent, starting from 1,1,1 or any known solution a_1,b_1,v_1. From the last displayed equation we find, upon taking square roots, that

$$E(2a_1^2 + b_1^2) = D(a_1b_1 \pm v_1).$$

Since we must have $(D,E) = 1$, we find $K_1 = (2a_1^2 + b_1^2, a_1b_1 + v_1)$ and $K_2 = (2a_1^2 + b_1^2, a_1b_1 - v_1)$, and then in the first case we take $D = (2a_1^2 + b_1^2)/K_1$, $E = (a_1b_1 + v_1)/K_1$ and in the second case, $D = (2a_1^2 + b_1^2)/K_2$, $E = (a_1b_1 - v_1)/K_2$. Then we compute, for whichever case we desire, the values of $g = Da_1$, $h = Eb_1$, $A = Ea_1$, $B = Db_1$. Finally we compute

$$a = g^2 + h^2, \quad b = \pm(2A^2 - B^2), \quad v = (g^2 - h^2 - 2gh)^2 - 8(gh)^2.$$

For example, after 1,1,1 the next solution is $a = 13$, $b = 1$, $v = 239$, there being only one case in this first step of ascent.

Perhaps the following symbols will help indicate the sense in which the preceding formulas represent the complete set of primitive solutions of $2a^4 - b^4 = v^2$. Set $A(0) = (1,1,1)$ and $A(1) = (13,1,239)$; then let $A(n;i_2,i_3,\ldots,i_n)$ for $n \geqq 2$, with each $i_j = 1$ or 2, indicate a solution a,b,v which is a member of the "nth generation" with the following "geneology"—that according as $i_j = 1$ or 2, the member (when $j = n$) or its "ancestor" of the jth generation (when $2 \leqq j \leqq n - 1$) was a "male" (K_1) or "female" (K_2) "offspring" of $A(j - 1; i_2,\ldots,i_{j-1})$.

Using this terminology we find in the second generation

$$A(2;1) = (2165017, 2372159, 3503833734241),$$
$$A(2;2) = (1525,1343,2750257).$$

In the third generation there would be four members: $A(3;1,1)$, $A(3;1,2)$, $A(3;2,1)$, $A(3;2,2)$; but we shall not bother to compute the values of a,b,v for these, because we are worn out with computing and checking the values of a,b,v for the second generation!

Of course this *recursive* complete solution is a very different article from an *explicit* complete solution like that for primitive Pythagorean triplets; but since every primitive solution of $2a^4 - b^4 = v^2$ occupies a definite place $A(n;i_2,\ldots,i_n)$ in the "family tree," the description of our formulas as providing a *complete* solution seems justified.

Now let us return to complete the solution of the original problem $x^4 - 2y^4 = z^2$.

It follows from the discussion preceding *Case 1* and *Case 2* that we can start from any primitive solution of $2a^4 - b^4 = v^2$ (whose complete solution has just been described) and can find a primitive solution of $x^4 - 2y^4 = z^2$; or we can continue from any primitive solution, say x',y',z' of the latter equation to find another primitive solution by exactly the same formulas, providing we set $b = x'$, $a = y'$, $v = z'$; the necessary formulas are as follows:

$$x = 2a^4 + b^4, \quad y = 2abv, \quad z = \pm(8a^4b^4 - v^4).$$

For example, from 1,1,1 solving $2a^4 - b^4 = v^2$ we find 3,2,7 solving $x^4 - 2y^4 = z^2$; then with $b = 3$, $a = 2$, $v = 7$, we ascend to the solution $x = 113$, $y = 84$, $z = 7967$; and so on, as far as we care to go in this particular chain.

The complete set of solutions of $x^4 - 2y^4 = z^2$ can be described as follows: use the symbols $S(0;t,k)$, $S(1;t,k)$ and $S(n;i_2,i_3,\ldots,i_n;t,k)$ with $n \geqq 2$ and $i_j = 1$ or 2, and with $t \geqq 1$ and $k \geqq 1$. Here the 0, the 1, and the $n;i_2,i_3,\ldots,i_n$ refer to the solution $A(0)$, $A(1)$, and $A(n;i_2,i_3,\ldots,i_n)$, respectively, of $2a^4 - b^4 = v^2$ from which the chain of solutions of $x^4 - 2y^4 = z^2$ originates; the t indicates the "generation" of x,y,z in the chain of primitive solutions of $x^4 - 2y^4 = z^2$; and the k indicates a solution of x',y',z' obtained from a primitive solution x,y,z by setting $x' = kx$, $y' = ky$, $z' = k^2z$.

For example: $S(0;1,1) = (3,2,7)$, $S(0;1,2) = (6,4,28)$,

$$S(0;1,3) = (9,6,63); \quad S(0;2,1) = (113,84,7967);$$
$$S(1;1,1) = (57123,\ 6214,\ 3262580153).$$

EXERCISES

EX. *15.1.* If m,n,k are given integers with $(m,n) = 1$, $mn \neq 0$, and $k = 2$, show that there exist non-zero integers x and y such that $mx^k = ny^k$ if and only if there are integers M and N such that $m = M^k$ and $n = N^k$.

EX. *15.2.* If there exist non-zero relatively prime integers x and y such that

$$0 = a_0x^n + a_1x^{n-1}y + \ldots + a_rx^{n-r}y^r + \ldots + a_{n-1}xy^{n-1} + a_ny^n,$$

where the a_i are integers with $a_0a_n \neq 0$, show that x must divide a_n and that y must divide a_0.

EX. *15.3.* Apply EX. *15.1* or EX. *15.2* to show that the Diophantine equations $x^2 = 2y^2$, $x^3 = 2y^3$, $x^4 = 2y^4$ are impossible of solution in non-zero integers x and y.

EX. *15.4.* Use Fermat's method of descent to show that $x^4 + 4y^4 = z^2$ is impossible of solution in positive integers.

EX. *15.5.* As a corollary to EX. *15.4* show that $x^4 - y^4 = z^2$ is impossible in positive integers.

EX. *15.6.* As a corollary to EX. *15.5* show that the area of the right triangle corresponding to a Pythagorean triplet cannot be a perfect square.

EX. *15.7.* Make a diagram showing the interrelated family trees for the 56 solutions A and S described in **15.5** for $n = 0,1,2,3$; $t = 1,2$; $k = 1,2,3$.

> ▶*Beyond the real world, though perceptibly connected with it, mathematicians have created intellectually an ideal world, which they attempt to develop into the most perfect of all worlds, and which is being explored in every direction.*
>
> —A. PRINGSHEIM

CHAPTER 16*

EULER'S PHI-FUNCTION

16.1. More about multiplicative functions. As a tool for later use in this lesson, we need a theorem which is essentially a converse to that given in **8.4**, so we shall use the same definitions and notations.

Theorem: If $F(n)$ and $f(n)$ are number-theoretic functions such that

(1) $F(n)$ is multiplicative, and

(2) $F(n) = \Sigma f(d)$, summed over all the positive divisors d of n, then $f(n)$ is multiplicative.

Proof: By (1), $F(1) = 1$; by (2) $F(1) = f(1)$; hence $f(1) = 1$.

When $(a,b) = 1$, we have shown in **8.4** that the set S'' of all positive divisors d'' of ab is exactly the same as the set S^* formed of integers dd' where d runs over the set S of positive divisors d of a and where d' runs over the set S' of positive divisors d' of b. Furthermore we know $(d,d') = 1$.

We shall make an induction proof on $n = ab$, where we assume, of course, that $(a,b) = 1$. The fundamental theorem guarantees that there is such a representation for every positive integer n.

(I) When $ab = 1$, then $a = 1 = b$. Since we have noted above that $f(1) = 1$, it follows in this case that $f(ab) = f(a)f(b)$.

*Chapter 16 is a basic chapter, except for **16.4** which is supplementary.

(II) The induction hypothesis regarding $n = ab$ with $(a,b) = 1$ will be that $f(dd') = f(d)f(d')$ if $(d,d') = 1$ and $dd' < ab$.

From (*1*) we know that $F(ab) = F(a)F(b)$. From (2) and the remarks above about the sets S^* and S'' it follows that

$$\sum_{S} f(d) \sum_{S'} f(d') = \sum_{S''} f(d'') = \sum_{S^*} f(dd')$$

By the induction hypothesis it follows that the expanded product on the left contains, with possibly one exception, exactly the same summands as does the sum on the right. But this forces the remaining terms on each side to be the same, namely, $f(a)f(b) = f(ab)$.

No matter what particular representation $n = ab$ with $(a,b) = 1$ is chosen, arguments (I) and (II) are valid; since there are for each n only a finite number of these representations, it follows that the induction argument is complete and that $f(n)$ is multiplicative.

Corollary: For a prime p, we have $f(p^a) = F(p^a) - F(p^{a-1})$.

Proof: The sum (2) for $F(p^a)$ contains just one more term, namely, $f(p^a)$, than does the sum (2) for $F(p^{a-1})$.

By combining the theorem and corollary, it follows that if $F(n)$ is multiplicative, then the exact formula for $f(n)$ is readily found. (Even if $F(n)$ is not multiplicative, a formula for $f(n)$ is known. See the exercises of this lesson.)

16.2. Definition and formula for Euler's phi-function. The Euler phi-function (sometimes called the totient function) is a widely used number-theoretic function, almost always indicated by $\phi(n)$. For $n = 1$, we define $\phi(1) = 1$, and when $n > 1$, we define $\phi(n)$ to be the *number of positive integers less than n* and *relatively prime* to n.

For example, since the only positive integers less than 12 and relatively prime to 12 are 1,5,7,11, it follows that $\phi(12) = 4$. Similarly, $\phi(1) = 1$, $\phi(2) = 1$, $\phi(3) = 2$, $\phi(4) = 2$, $\phi(5) = 4$, $\phi(6) = 2$, etc. But we desire a formula which will allow us to compute the value of $\phi(n)$ directly from the standard form of n, without actually listing all the numbers less than n and relatively prime to n.

In this lesson we shall give two derivations of the formula for $\phi(n)$; in a later lesson we shall give yet another derivation.

For the first derivation we shall begin by stating and proving the theorem which is the correct generalization of the following example: $12 = \phi(1) + \phi(2) + \phi(3) + \phi(4) + \phi(6) + \phi(12) = 1 + 1 + 2 + 2 + 2 + 4.$

Theorem: For any positive integer n, $n = \Sigma\phi(d)$, where the summation extends over all the positive divisors d of n.

Proof: The theorem is obvious for $n = 1$, since $1 = \phi(1)$. Consider $n > 1$. For every positive integer $x \leqq n$, $(x,n) = d$, where d is a uniquely determined divisor of n. On this basis alone the n numbers $1,2,\ldots,n$ are divided into mutually exclusive d-classes. From $(x,n) = d$ we have $x = kd$, $n = d'd$, with $(k,d') = 1$ and with $k \leqq d'$ since $x \leqq n$. The case $k = d'$ is exceptional for from the condition $(k,d') = 1$ this case can arise only when $d' = 1$. Hence in all cases we find that there are exactly $\phi(d')$ choices for k, and hence $\phi(d')$ integers x which belong to the d-class. Thus by the use of the d-classes we have found $n = \Sigma\phi(d')$ where $d'd = n$ and the summation is over all divisors d of n. However the set of numbers $\{d'\}$ is simply the set $\{d\}$ in another order, hence we are justified in writing $n = \Sigma\phi(d') = \Sigma\phi(d)$ which completes the proof.

Thus in the example given above,

$\phi(12) = 4$ indicates 4 integers in the 1-class: 1,5,7,11;
$\phi(6) = 2$ indicates 2 integers in the 2-class: 2,10;
$\phi(4) = 2$ indicates 2 integers in the 3-class: 3,9;
$\phi(3) = 2$ indicates 2 integers in the 4-class: 4,8;
$\phi(2) = 1$ indicates 1 integer in the 6-class: 6;
$\phi(1) = 1$ indicates 1 integer in the 12-class: 12.

Of course, this theorem is "tailor-made" so that the theorem and corollary in **16.1** may be applied to obtain the following result.

Theorem: If n is written in standard form as

$$n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k}$$

where each p_i is a prime, $1 < p_1 < p_2 < \ldots < p_k$, and $a_i \geqq 1$, then

$$\phi(n) = n\left(\frac{p_1 - 1}{p_1}\right)\left(\frac{p_2 - 1}{p_2}\right)\ldots\left(\frac{p_k - 1}{p_k}\right).$$

Proof: We have observed before in **8.4** that the function $F(n) = n$ is multiplicative. Since we have just shown that $n = \Sigma\phi(d)$, summed over all the positive divisors d of n, it follows from the theorem in **16.1** that $\phi(n)$ is multiplicative. From the corollary in **16.1** we see that

$$\phi(p^a) = F(p^a) - F(p^{a-1}) = p^a - p^{a-1} = p^a\left(\frac{p-1}{p}\right).$$

Combining these results we arrive at the formula for $\phi(n)$ displayed above.

For example: since $12 = 2^2 3$, $\phi(12) = 12(1/2)(2/3) = 4$; and since $8316 = 2^2 3^3 7(11)$, it follows that

$$\phi(8316) = 2^2 3^3 7(11)(1/2)(2/3)(6/7)(10/11) = 2^4 3^3 5 = 2160.$$

Interpreting this last example, we know that there are 2160 positive integers less than 8316 and relatively prime to 8316; and we have obtained this figure of 2160 in a way far more satisfactory than mere counting.

16.3. Combinatorial logic. To obtain another independent derivation of the formula for $\phi(n)$ we shall first need to make a digression and discuss certain topics in combinatorial logic.

Let S be a set of objects. Let A be a set of objects in S possessing a certain property or attribute; without confusion this property itself can also be designated by the letter A. Let AB indicate the set of objects in S possessing *both* properties A and B; if there are no objects of this description let the set be described as the *empty* or *null* set and designated by the symbol 0. Let A' indicate all the objects in S *not* possessing the property A. Let $A + B$ indicate the set of all objects in S possessing *either* property A *or* property B. Let $N(A)$ indicate the *number* of objects in the set S possessing the property A.

The following results are then obtained by formal logic:

(*1*): $N(A + B) = N(A) + N(B) - N(AB)$, for the $N(AB)$ objects possessing both properties A and B are included once in $N(A)$ and again in $N(B)$, hence we must subtract $N(AB)$ from $N(A) + N(B)$ to obtain the correct count for $N(A + B)$.

(2): $N((A + B)C) = N(AC + BC)$, for both $(A + B)C$ and $AC + BC$ contain exactly the same subsets $AB'C + ABC + A'BC$.

(3): $N(AA) = N(A)$, for $AA = A$.

(4): $N(S) = N(A) + N(A')$, for AA' is an empty set and $A + A' = S$ hence it follows from (*1*) that $N(S) = N(A + A') = N(A) + N(A') - N(AA') = N(A) + N(A')$.

Theorem: If $A_1, A_2, \ldots, A_k$ are k properties possessed by various elements of S then

$$N(A_1 + A_2 + \ldots + A_k) = \sum_{i=1}^{k} N(A_i) - \sum_{i,j=1;\ i<j}^{k} N(A_iA_j) + \ldots + (-1)^{k+1} N(A_1A_2 \ldots A_k),$$

where the general term on the right of this formula is

$$(-1)^{r+1}\sum N(A_{i_1}A_{i_2}\ldots A_{i_r}), \qquad 1 \leqq r \leqq k,$$

with this last summation being extended over the $\binom{k}{r}$ combinations of $1,2,\ldots,k$ taken r at a time.

Thus, for example, if $k = 3$, we have

$$\begin{aligned} N(A_1 + A_2 + A_3) = {} & N(A_1) + N(A_2) + N(A_3) \\ & -N(A_2A_3) - N(A_1A_3) - N(A_1A_2) + N(A_1A_2A_3). \end{aligned}$$

Proof: The proof of the theorem is by induction on k.

(I) When $k = 1$, the theorem is true, reducing to the trivial observation that $N(A_1) = N(A_1)$.

(II) We shall assume that the theorem is true for every case involving k or fewer attributes and examine the case of $k + 1$ attributes. With the aid of *(1)* and *(2)* we may write

$$\begin{aligned} N(A_1 + A_2 + \ldots + A_k + A_{k+1}) &= N(A_1 + A_2 + \ldots + A_k) + N(A_{k+1}) \\ - N((A_1 + \ldots + A_k)A_{k+1}) &= N(A_1 + A_2 + \ldots + A_k) + N(A_{k+1}) \\ & \quad - N(A_1A_{k+1} + A_2A_{k+1} + \ldots + A_kA_{k+1}). \end{aligned}$$

Now we may apply the induction hypothesis to each of these sets, for they each involve not more than k attributes. In this application when a situation like $(A_iA_{k+1})(A_jA_{k+1})$ arises, we may use *(3)* to write this term in the simpler form $A_iA_jA_{k+1}$. Using these observations we now find that

$$\begin{aligned} N(A_1 + \ldots + A_k + A_{k+1}) = {} & \sum_{i=1}^{k} N(A_i) - \sum_{i,j=1;\ i<j}^{k} N(A_iA_j) \\ & + \ldots + (-1)^{k+1}N(A_1A_2\ldots A_k) + N(A_{k+1}) - \\ \left\{\sum_{i=1}^{k} N(A_iA_{k+1}) - \sum_{i,j=1;\ i<j}^{k} N(A_iA_jA_{k+1}) \right. & \left. + \ldots + (-1)^{k+1}N(A_1\ldots A_kA_{k+1})\right\} \\ = {} & \sum_{i=1}^{k+1} N(A_i) - \sum_{i,j=1;\ i<j}^{k+1} N(A_iA_j) + \ldots + (-1)^{k+2}N(A_1A_2\ldots A_kA_{k+1}). \end{aligned}$$

But this last result is precisely the form the theorem should take in the case $k + 1$.

By (I), (II), and the principle of mathematical induction the theorem is always true.

Corollary: If $N' = N((A_1 + A_2 + \ldots + A_k)')$ indicates the number of elements of S possessing *none* of the properties $A_1, A_2, \ldots,$ or A_k, then

$$N' = N(S) - \sum_{i=1}^{k} N(A_i) + \sum_{i,j=1;\ i<j}^{k} N(A_iA_j) - \ldots + (-1)^k N(A_1A_2\ldots A_k).$$

Proof: The corollary follows readily from *(4)* and the theorem.

Theorem: If n is written in standard form as

$$n = p_1{}^{a_1}p_2{}^{a_2}\ldots p_k{}^{a_k},$$

where each p_i is a prime, $1 < p_1 < p_2 < \ldots < p_k$, and $a_i \geqq 1$, then

$$\phi(n) = n(1 - 1/p_1)(1 - 1/p_2)\ldots(1 - 1/p_k).$$

Proof: We intend to use the corollary above. We let S be the set of integers: $1,2,3,\ldots,n$. We let A_i be the property that an integer is divisible by the prime p_i. Then $N(A_i)$, the number of integers in the set S divisible by the prime p_i, is given by $N(A_i) = n/p_i$. And $N(A_iA_j)$, the number of integers in S divisible by both p_i and p_j, is given by $N(A_iA_j) = n/p_ip_j$; etc. Since $\phi(n)$ is the number of integers in S *not divisible* by any of the primes $p_1,p_2,\ldots,p_j$, it follows from the corollary above that

$$\begin{aligned}\phi(n) = N' &= N(S) - \sum N(A_i) + \sum N(A_iA_j) - \ldots \\ &\qquad\qquad + (-1)^k N(A_1A_2\ldots A_k) \\ &= n - \sum n/p_i + \sum n/p_ip_j - \ldots + (-1)^k n/p_1p_2\ldots p_k \\ &= n(1 - 1/p_1)(1 - 1/p_2)\ldots(1 - 1/p_k),\end{aligned}$$

since the expansion of this last product is seen to contain exactly the summands listed in the previous line, correct even to plus and minus signs.

16.4 Inversion of the Euler phi-function. Let us consider the inverse problem: given a, find all solutions n of $\phi(n) = a$.

When $a = 1$ there are exactly two solutions: $n = 1$ and $n = 2$, for if $n > 2$, since both 1 and $n - 1$ are less than n and relatively prime to n, it follows that $\phi(n) > 1$. In fact, if $(i,n) = 1$ with $0 < i < n$, then $(n - i,n) = 1$ and $0 < n - i < n$; furthermore $n - i \neq i$, for otherwise we would have $n = 2i$ and $(i,n) = i$ which would contradict $(i,n) = 1$ when $n > 2$; hence if $n > 2$, the integers less than n and relatively prime to n occur in pairs and $\phi(n)$ must be even. (This result also appears directly from the formula in **16.2**, see EX. *16.1*).

Let us suppose $n > 2$ and write n in a slightly modified standard form, suggested by the formula in **16.2**:

$$(16.1) \qquad n = \Pi A_i^{k_i+1}\Pi B_j$$

where the A's and B's are distinct prime factors of n and $k_i \geqq 1$ (the Π is the usual abbreviation for the product of terms of the kind following the symbol; of course, in certain n there may be no prime factors of type A, or none of type B). Then by **16.2** we have

$$\phi(n) = n\Pi(1 - 1/A_i)\Pi(1 - 1/B_j)$$

which on simplification gives us the following formula:

$$(16.2) \qquad \phi(n) = \Pi A_i^{k_i}\Pi(A_i - 1)\Pi(B_j - 1).$$

We attack our problem by finding all ways of writing a in the form *(16.2)* and then we can pass back to the solutions n in the form *(16.1)*. If a is odd and $a > 1$, there are no solutions, since $\phi(n)$ must be either 1 or even. Having previously discussed the case $a = 1$, we suppose now that a is an even positive integer.

First write a in standard form as $a = P_1^{a_1}P_2^{a_2}\ldots P_m^{a_m}$ where the P_i are primes, $2 = P_1 < P_2 < \ldots < P_m$, and the $a_i \geqq 1$. Then form the following $(a_1 + 1)(a_2 + 1)\ldots(a_m + 1)$ numbers in lexicographic order:

$$C_{b_1,b_2,\ldots,b_m} = 1 + P_1^{b_1}P_2^{b_2}\ldots P_m^{b_m}, \qquad 0 \leqq b_i \leqq a_i,$$

understanding, as usual, that $P_i^0 = 1$.

Then classify the C's as follows:

(1) Discard C's that are not primes.

(2) If C is a prime P_i, call it A. Order the A's by magnitude: $A_1 < A_2 < \ldots < A_u$.

(3) If C is a prime, but not a P_i, call it B. Order the B's by magnitude: $B_1 < B_2 < \ldots < B_v$.

Next, proceeding in lexicographic fashion, form all the following sets of exponents $L(N)$ and the corresponding numbers N, as follows:

$$L(N): \quad k_1,k_2,\ldots,k_u; \quad s_1,s_2,\ldots,s_u; \quad t_1,t_2,\ldots,t_v$$

under the restrictions $t_j = 0$ or 1; $s_i = 0$ or 1; if $s_i = 0$, then $k_i = 0$; if $s_i = 1$, then $0 \leqq k_i \leqq a_i$;

$$N = \Pi A_i^{k_i}\Pi(A_i - 1)^{s_i}\Pi(B_j - 1)^{t_j}$$

where the products run over all the A_i and B_j found in (2) and *(3)*.

Each set of exponents $L(N)$ for which $N = a$ gives a solution

$$n = \Pi A_i^{k_i+s_i}\Pi B_j^{t_j}$$

of the equation $\phi(n) = a$; and all solutions are found by this method.

In discussion of this rule let us note that the C's include *all possible numbers* such that $C - 1$ will divide a, which is the requirement suggested by the form (*16.2*). But the A_i and B_j corresponding to the $A_i - 1$ and $B_j - 1$ in (*16.2*) must be *primes*, so we throw out the C's which are not primes. The distinction between the B's which are not factors of $\phi(n) = a$ and *cannot be repeated* factors of n, and the A's which are factors of $\phi(n) = a$ and *can be repeated* factors of n, is explained by considering the equations (*16.1*) and (*16.2*).

For example: $a = 72 = 2^3 3^2$, $P_1 = 2$, $a_1 = 3$, $P_2 = 3$, $a_2 = 2$.
$C_{00} = 1 + 1 = 2 = A_1$; $C_{01} = 1 + 3 = 4$, out; $C_{02} = 1 + 9 = 10$, out;
$C_{10} = 1 + 2 = 3 = A_2$; $C_{11} = 1 + 6 = 7 = B_2$; $C_{12} = 1 + 18 = 19 = B_4$;
$C_{20} = 1 + 4 = 5 = B_1$; $C_{21} = 1 + 12 = 13 = B_3$; $C_{22} = 1 + 36 = 37 = B_5$;
$C_{30} = 1 + 8 = 9$, out; $C_{31} = 1 + 24 = 25$, out; $C_{32} = 1 + 72 = 73 = B_6$.

We must then systematically find all solutions of

$$N = 2^{k_1}3^{k_2}1^{s_1}2^{s_2}4^{t_1}6^{t_2}(12)^{t_3}(18)^{t_4}(36)^{t_5}(72)^{t_6} = a = 72$$

subject to the restrictions $t_j = 0,1$; $s_i = 0 = k_i$; $s_1 = 1$, $0 \leqq k_1 \leqq 3$; $s_2 = 1$, $0 \leqq k_2 \leqq 2$. In tabular form the solutions are as follows:

$L(N)$:	k_1	k_2	s_1	s_2	t_1	t_2	t_3	t_4	t_5	t_6	a	n
	0	0	0	0	0	0	0	0	0	1	72	73
	0	0	0	0	0	1	1	0	0	0	6·12	7·13
	0	0	0	0	1	0	0	1	0	0	4·18	5·19
	0	0	0	1	0	0	0	0	1	0	2·36	3·37
	0	0	1	0	0	0	0	0	0	1	1·72	2·73
	0	0	1	0	0	1	1	0	0	0	1·6·12	2·7·13
	0	0	1	0	1	0	0	1	0	0	1·4·18	2·5·19
	0	0	1	1	0	0	0	0	1	0	1·2·36	2·3·37
	0	1	0	1	0	0	1	0	0	0	3·2·12	9·13
	0	1	1	1	0	0	1	0	0	0	3·1·2·12	2·9·13
	0	2	0	1	1	0	0	0	0	0	9·2·4	27·5
	0	2	1	1	1	0	0	0	0	0	9·1·2·4	2·27·5
	1	0	1	0	0	0	0	0	1	0	2·1·36	4·37
	1	0	1	1	0	0	0	1	0	0	2·1·2·18	4·3·19
	1	1	1	1	0	1	0	0	0	0	2·3·1·2·6	4·9·7
	2	0	1	0	0	0	0	1	0	0	4·1·18	8·19
	2	2	1	1	0	0	0	0	0	0	4·9·1·2	8·27

Hence $\phi(n) = 72$ has exactly 17 solutions, namely, $n = 73,91,95,111,117,135,146,148,152,182,190,216,222,228,234,252,270$.

Concerning the inversion problem of this section there is an interesting conjecture by Carmichael to the effect that if, for a given a, the equation $\phi(n) = a$ has *any* solutions, then it has *at least two* solutions.

EXERCISES

EX. *16.1.* Show from the formula in **16.2** that $\phi(n)$ is even for $n > 2$.

EX. *16.2.* Find $\phi(72)$, $\phi(210)$, and $\phi(p^t)$ where p is a prime.

EX. *16.3.* Show that the *sum* of all integers less than n and relatively prime to n is given by $n\phi(n)/2$ for $n > 2$.

EX. *16.4.* Prove that if $(a,b) = 1$, then $\phi(ab) = \phi(a)\phi(b)$, using the formula for $\phi(n)$ in **16.3**.

EX. *16.5.* Construct a table (46 entries) of all values of $A^k(A - 1) < 5000$ where A is a prime and $k \geqq 1$.

EX. *16.6.* Find all solutions of $\phi(n) = 60$, using EX. *16.5* or *16.4*.

EX. *16.7.* Show that the numbers $a = 242,244,246,248$ form a sequence of four consecutive even numbers such that there are no solutions of $\phi(n) = a$.

EX. *16.8.* Find a sequence of five consecutive even numbers a (less than 1000) for which there are no solutions to $\phi(n) = a$.

EX. *16.9.* If $n > 1$ is written in standard form as $n = p_1^{a_1}p^{a_2}\ldots p_k^{a_k}$ where each p_i is a prime, $1 < p_1 < p_2 < \ldots < p_k$, and $a_i \geqq 1$, define the Mobius function $\mu(n)$ as follows: if any $a_i > 1$, define $\mu(n) = 0$; if every $a_i = 1$, define $\mu(n) = (-1)^k$. For $n = 1$, define $\mu(1) = 1$. Prove that $\mu(n)$ is multiplicative.

EX. *16.10.* Use **8.4** to prove that $G(n) = \Sigma\mu(d)$, summed over the positive divisors d of n, is multiplicative, and that $G(n) = 0$ when $n > 1$.

EX. *16.11.* If $F(n)$ and $f(n)$ are number-theoretic functions such that (*1*): $F(n) = \Sigma f(d)$, summed over the positive divisors d of n, prove that (*2*): $f(n) = \Sigma\mu(d')F(d)$, summed over the positive divisors d of n, where $dd' = n$ and where $\mu(n)$ is the Mobius function of the preceding exercises. (*Hint*: show by induction that (*1*) completely determines $f(n)$; then show that (*2*) solves (*1*), employing the properties of $G(n)$ in EX. *16.10*.)

EX. *16.12.* Use EX. *16.11* and the first theorem of **16.2** to give a derivation of the $\phi(n)$ formula not depending upon a priori determination that $\phi(n)$ is multiplicative.

▶ *The invention of the symbol ≡ by Gauss affords a striking example of the advantage which may be derived from an appropriate notation and marks an epoch in the development of the science of arithmetic.*

—G. B. MATHEWS

CHAPTER 17*

INTRODUCTION TO THE CONGRUENCE NOTATION

17.1. Definition of congruence modulo *m*. Let m be a fixed positive integer, then we shall define the integer a to be congruent to the integer b modulo m, written

$$a \equiv b \bmod m$$

and read "a is congruent to b mod m," if and only if

$$a - b = km$$

where k is an integer.

For example:

$$17 \equiv 2 \bmod 5, \text{ because } 17 - 2 = (3)5;$$

$$-8 \equiv 2 \bmod 5, \text{ because } -8 - 2 = (-2)5;$$

$$17 \equiv -8 \bmod 5, \text{ because } 17 - (-8) = (5)5.$$

This notation is due to Gauss, and, as we shall see in this and later chapters, the comment quoted beneath the chapter head is well justified. Remembering Gauss, we shall denote the following series of theorems about the congruence notation by **G.1, G.2,** etc.

G.1: We find $a \equiv b \bmod m$ if and only if a and b have the same remainder R, $O \leqq R < m$, when divided by m.

Proof: If $a \equiv b \bmod m$, so that $a - b = km$, with k an integer,

[*] Chapter 17 is a basic chapter.

and if $b = qm + R, O \leqq R < m$, then $a = b + km = (q + k)m + R$ has the same remainder as b. Conversely, if $a = Qm + R$, $b = qm + R$, $O \leqq R < m$, then $a - b = (Q - c)m$, with $Q - q$ an integer, hence $a \equiv b \bmod m$.

17.2 Congruence modulo *m* is an equivalence relation. Within a mathematical system there may be various relations between the elements of the system. If $a,b, \ldots$ are elements of a mathematical system S, then we say that a relation **E** between the elements, written $a\mathbf{E}b$ and read "a is **E** to b" or written $a\not\mathbf{E}b$ and read "a is not **E** to b," is an *equivalence relation* if and only if **E** satisfies the following requirements:

E.1: **E** is *determinative:* for any two elements a,b in S, either $a \mathbf{E} b$ or $a\not\mathbf{E}b$, but not both of these.

E.2: **E** is *reflexive:* $a\mathbf{E}a$, for every element a in S.

E.3: **E** is *symmetric:* if $a \mathbf{E} b$, then $b \mathbf{E} a$, for all a,b, in S.

E.4: **E** is *transitive:* if $a \mathbf{E} b$ and $b \mathbf{E} c$, then $a\mathbf{E}c$, for all a,b,c in S.

If S is the set of all integers, then one of the most striking examples of an equivalence relation, other than the ordinary equality, is the notion of congruence modulo m.

G.2: Congruence modulo m is an equivalence relation for integers.

Proof: **E.1:** By its very definition congruence is *determinative* for the difference $a - b$ of any two given integers a and b either is or is not a multiple of m; or we may refer to **G.1** and comment that either a and b do have the same remainder R, $0 \leqq R < m$, when divided by m, or they do not have the same remainder.

E.2: Congruence is *reflexive* because for every integer a we have $a - a = (0)m$, and 0 is an integer, hence $a \equiv a \bmod m$.

E.3: Congruence is *symmetric*, for if $a \equiv b \bmod m$ so that $a - b = km$, where k is an integer, then $b - a = (-k)m$ with $-k$ an integer, hence $b \equiv a \bmod m$.

E.4: Congruence is *transitive*, because $a \equiv b$, and $b \equiv c \bmod m$ imply $a - b = km$ and $b - c = Km$, where k and K are integers;

then by addition we find $a - c = (k + K)m$ with $k + K$ an integer, hence $a \equiv c \bmod m$.

Every equivalence relation of a set divides the set into mutually exclusive classes of "equal" elements. In this case we see, from **G.1** and **G.2**, that under congruence modulo m, all the integers are divided into exactly m classes, corresponding to the possible remainders $R = 0,1,\ldots,m - 1$, and we see that each "R-class" contains infinitely many integers, namely, $qm + R$, where $q = 0, \pm 1, \pm 2, \ldots$.

These classes are commonly called *residue classes*, where we are using the word "residue" is the same sense as "remainder." Any set of m numbers, one and only one from each residue class, constitutes a *complete residue system.*

17.3. Addition and multiplication of residue classes. We shall define the "sum" of the a-class, modulo m, and the b-class, mod m, to be the residue class containing $a + b$. Similarly, we define the "product" of the a-class and the b-class, mod m, to be the class containing ab.

When in a mathematical system a new operation is defined, there are two of its properties to be investigated and established before the new operation can be regarded as very useful.

First we must ask if the operation is "closed" by which we mean to require that the result of the operation be an element of the system.

Secondly we must check whether the operation is "well defined" by which we mean to refer to the particular equivalence relation **E** being used in the system and to require that if each of the elements on which the operation is performed be replaced by an equivalent element and the operation be performed anew, then the second result must be equivalent to the first result.

For the operations with residue classes which we have defined above, it is clear that both operations are "closed"—the result in each case being a certain residue class. It is now necessary to show that these operations are "well defined."

Thus if a is replaced by an "equal" element A: i.e., any member of the a-class; and if b is replaced by an "equal" element B: i.e., any member of the b-class, it is necessary to show that the definitions are of such a nature that the "sum" and "product" found by using a and b are "equal," respectively, to the "sum" and "product" found by using A and B. Otherwise, the proposed definitions are useless.

G.3: Addition and multiplication of residue classes modulo m, defined by

$$(a\text{-class}) + (b\text{-class}) = ((a + b)\text{-class}),$$
$$(a\text{-class})(b\text{-class}) = (ab\text{-class}),$$

are well defined.

Proof: We must show that if $a \equiv A$, and $b \equiv B \bmod m$, then $a + b \equiv A + B$, and $ab \equiv AB \bmod m$. By hypothesis we have $a - A = km$ and $b - B = Km$, where k and K are integers, whence by addition we find $(a + b) - (A + B) = (k + K)m$, and since $k + K$ is an integer, it follows that $a + b \equiv A + B \bmod m$. We may write the equations resulting from the hypothesis in the form $a = A + km$ and $b = B + Km$, whence by multiplication we find

$$ab = AB + AKm + kmB + kmKm,$$
$$ab - AB = (AK + kB + kmK)m,$$

and since $AK + kB + kmK$ is an integer, it follows that $ab \equiv AB \bmod m$. Thus both addition and multiplication of residue classes are well defined operations.

By way of illustration let us consider a complete residue system modulo 5, with the corresponding classes partly indicated as shown below:

$$0 \equiv \ldots, -10, -5, 0, 5, 10, \ldots \bmod 5$$
$$1 \equiv \ldots, -9, -4, 1, 6, 11, \ldots$$
$$2 \equiv \ldots, -8, -3, 2, 7, 12, \ldots$$
$$3 \equiv \ldots, -7, -2, 3, 8, 13, \ldots$$
$$4 \equiv \ldots, -6, -1, 4, 9, 14, \ldots$$

Theorem G.3 implies that any member of the 2-class added to any member of the 4-class must give a result in the 1-class, for $2 + 4 = 6 \equiv 1 \bmod 5$. Similarly, any member of the 2-class multiplied by any member of the 4-class must give a result in the 3-class, for $2 \cdot 4 = 8 \equiv 3 \bmod 5$. The complete addition and multiplication tables mod 5 are as follows:

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

As another exercise in the use of **G.3** we seek the remainder when $x = 2^{73} + (14)^3$ is divided by 11. With the aid of the congruence notation this problem, which would otherwise seem a fearsome one, is easily solved. First, $2^4 = 16 \equiv 5 \bmod 11$; then by **G.3**, $2^8 = (2^4)^2 \equiv 5^2 \equiv 3 \bmod 11$; $2^{10} = 2^8 4 \equiv 3 \cdot 4 = 12 \equiv 1 \bmod 11$; therefor, $2^{70} = (2^{10})^7 \equiv 1^7 = 1 \bmod 11$; hence, $2^{73} = 2^{70}8 \equiv 8 \bmod 11$. Secondly, $(14)^3 \equiv 3^3 = 27 \equiv 5 \bmod 11$. Finally, $x = 2^{73} + (14)^3 \equiv 8 + 5 = 13 \equiv 2 \bmod 11$. As we have shown in **G.1** this is just another way of saying that there is an integer k so that $x = 11k + 2$ and that the desired remainder is 2. We have avoided completely the actual computation of x and k and this is a remarkable gain that can be exploited in many ways.

17.4. Casting out nines. To introduce this section we shall establish a theorem from which the main results of the section follow readily.

G.4: Given the polynomial

$$f(x) = a_0 + a_1x + \ldots + a_nx^n,$$

where the coefficients are integers and given that x and y are integers such that $x \equiv y \bmod m$, then

$$f(x) \equiv f(y), \bmod m.$$

Proof: The result in **G.4** is a corollary to **G.3**. Thus $x \equiv y \bmod m$ and **G.3** together imply that $x^i \equiv y^i \bmod m$; from **G.2** we have $a_i \equiv a_i \bmod m$; hence by **G.3**, $a_ix^i \equiv a_iy^i \bmod m$; summing with respect to i, employing **G.3**, we find that $f(x) \equiv f(y) \bmod m$.

As an immediate application of **G.4** we can prove the following theorem:

G.4: Any positive integer written to the base 10 is congruent to the sum of its digits modulo 9.

Proof: If the polynomial $f(x)$ in **G.4** is restricted by insisting that $0 \leqq a_i < 10$, $i = 0,1,\ldots,n - 1$; $0 < a_n < 10$; then $F = f(10)$ is a suitable way of representing any desired positive integer. If in **G.4** we take $m = 9$, $x = 10$, and $y = 1$, we have $F = f(10) \equiv f(1) \bmod 9$; but $f(1) = a_0 + a_1 + \ldots + a_n$ is the sum of the digits of F, so this completes the proof.

By repeatedly applying **G.4** we may find the least positive residue

R of F mod 9, so that $F \equiv R$ mod 9 and $0 \leq R < 9$. For example, by **G.5**, $3275 \equiv 3 + 2 + 7 + 5 = 17$; then applying **G.5** again, we find $17 \equiv 1 + 7 = 8$; hence by the transitive property in **G.2**, we know $3275 \equiv 8$ mod 9. Actually part of this work is superfluous since the $2 + 7$ occurring in the first step is a multiple of 9 and may be "cast out" immediately.

By properly interpreting **G.3**, **G.4**, **G.5**, we have at hand the results frequently taught (without proof) in elementary school arithmetic, under the name "casting out nines," as a check on operations with integers.

First, for each number F used in a problem we compute its least positive residue F' mod 9 by repeated use of **G.5** and casting out of nines. Next, the assigned operations are performed on the given integers F,G,H, etc., until the result X is obtained. Then the same operations are performed on F',G',H', etc., until the result X' is found. Now as **G.3** and **G.4** show, if all the computations are correctly performed, we must find $X \equiv X'$ mod 9.

Conversely, however, if $X \equiv X'$ mod 9, it does *not* follow that the answer X is correct (although this erroneous conclusion is sometimes taught), for it is clear from the congruence point of view that any number of errors involving multiples of 9 may have been made, and that these errors will escape the supposed check. Thus the casting out of nines affords only a *partial* check on the accuracy of arithmetical calculations; it will detect errors if the errors are not multiples of 9.

This method of checking is illustrated in the following problems:

$$\begin{array}{rll}
3275 & \equiv 8 \text{ mod } 9 & \\
\underline{676} & \equiv \underline{1} & \\
19650 & \quad 8 & \\
22925 & & \text{partial} \\
\underline{19650} & & \text{check} \\
2213900 & \equiv 8 &
\end{array}
\qquad\qquad
\begin{array}{rrcl}
 & 1635 & \equiv & 6 \text{ mod } 9 \\
 & 173 & \equiv & 2 \\
 & 9919 & \equiv & 1 \\
 & 325 & \equiv & 1 \\
 & \underline{617} & \equiv & \underline{5} \\
6 \equiv & 12669 & & 15 \equiv 6
\end{array}$$

partial check

If the integers are expressed in the base b and if the operations are carried out in that base, then it follows by the same arguments as used above with $m = b - 1$, $x = b$, $y = 1$, that a partial check on the operations may be obtained by "casting out $(b - 1)$'s." The following examples illustrate operations in the base "6" checked by "casting out fives:"

$$\begin{array}{rcl}
3211 & \equiv & 2 \bmod 5\\
\underline{531} & \equiv & \underline{4}\\
3211 & & 12 \equiv 3\\
14033 & & \swarrow\!\!\nearrow\\
\underline{24455} & & \text{partial}\\
3033441 \equiv 3 & & \text{check}
\end{array}
\qquad\qquad
\begin{array}{rcl}
1534 & \equiv & 3 \bmod 5\\
152 & \equiv & 3\\
5515 & \equiv & 1\\
325 & \equiv & 0\\
\underline{514} & \equiv & \underline{0}\\
\end{array}$$

$$\underbrace{2 \equiv 13332 \qquad 11 \equiv 2}_{\text{partial check}}$$

EXERCISES

EX. *17.1.* If $(a,m) = d$ and $A \equiv a \bmod m$, show that $(A,m) = d$. State this result in terms of residue classes.

EX. *17.2.* Construct addition and multiplication tables for the residue classes mod 6. *Compare* the addition table mod 6 with the addition table mod 5 (given in 17.3); *contrast* the multiplication table mod 6 with the multiplication table mod 5.

EX. *17.3.* Find the remainder when 3^{40} is divided by 23.

EX. *17.4.* Prove that $M_{37} = 2^{37} - 1$ is divisible by 223.

EX. *17.5.* Prove that an integer is divisible by 3, or by 9, if and only if the sum of its digits is divisible by 3, or by 9, respectively.

EX. *17.6.* Using the notation in **G.4** and **G.5**, prove that an integer $F = f(10)$ is divisible by 11 if and only if $f(-1)$, the alternating sum of its digits, is divisible by 11.

EX. *17.7.* Show that a representation of an integer F in the base "1000," say $F = g(1000)$, can be obtained from the usual representation in the base "10," say $F = f(10)$, by grouping the digits of the latter representation in suitable triplets. Prove that F is divisible by 7, 11, or 13 if and only if the alternating sum $g(-1)$ is divisible by 7, 11, or 13, respectively.

EX *17.8.* Apply the tests of **EX.** *17.5*, **EX** *17.6*, **EX.** *17.7*, to the integer 847,963,207.

EX. *17.9.* Show that if a is odd, then $a^{2n} \equiv 1 \bmod 2^{n+2}$.

EX. *17.10.* Compute $X = (4353)^3 + 1734$ and check (partially!) by casting out nines.

The following exercises use terms defined in Chapter 11. Let m be a fixed positive integer and i, an integer. Let R_i indicate a transformation of the points p of a circle S in which the circle is rotated about its center through an angle whose measure in degrees is $i(360)/m$.

EX. *17.11.* Understanding that one revolution is represented by 360 degrees, show that $R_i = R_j$ if and only if $i \equiv j \bmod m$.

EX. *17.12.* Show that $R_iR_j = R_t$ if and only if $i + j \equiv t \bmod m$.

EX. *17.13.* Prove that the set G_m of all rotations R_i is a transformation group (called the *cyclic group of order m*).

EX. *17.14.* Show that G_m can be represented by m rotations and that the multiplication table for G_m can be represented by the addition table for residue classes modulo m.

▶*Some derive the same sort of stimulus from a technical monograph as others find in a detective novel—to which, in fact, it may bear some sort of esoteric resemblance. It is all a question of taste, and taste is a thing which no one person can decide for another.*

—R. CURLE

CHAPTER 18*

THE EULER-FERMAT THEOREMS

18.1. The restricted cancellation law. For the ordinary integers we know that if $a \neq 0$, then $ab = ac$ implies $b = c$, hence "cancellation" or the "cancellation law" is valid for all non-zero integers.

For residue classes mod m, it is natural to ask: "If a is not in the 0-class, does it follow from $ab \equiv ac \bmod m$, that $b \equiv c \bmod m$?"

An immediate answer of "No!" is provided by the following example:

$(2)(1) \equiv (2)(3) \bmod 4$ with $2 \not\equiv 0 \bmod 4$ and yet $1 \not\equiv 3 \bmod 4$.

As the nearest substitute for this anomaly we have the following theorem:

G.6: If $ab \equiv ac \bmod m$, $d = (a,m)$, $m = m_1d$, then $b \equiv c \bmod m_1$.

Proof: By hypothesis with $ab - ac = km$, $d = (a,m)$, $m = m_1d$, $a = a_1d$, we find $a_1(b - c) = km_1$. But since $(a_1,m_1) = 1$, it follows that $k = Ka_1$, whence $b - c = Km_1$ and therefore $b \equiv c \bmod m_1$.

Corollary: The cancellation law mod m is valid for a-classes for which $(a,m) = 1$.

Proof: In G.6, if $(a,m) = 1$, then $m_1 = m$. It only remains to prove (as in EX. *17.1*) that if $(a,m) = 1$, then every member A of the

*Chapter 13 is a basic chapter.

a-class has the property $(A,m) = 1$. But if $A \equiv a \bmod m$, then $A = a + km$. Let $(A,m) = (a + km,m) = d$; then d divides m and also divides $a + km$, hence d divides a; but this implies that d divides $(a,m) = 1$, and hence $d = 1$.

Conversely, if $(a,m) = d$, where $1 < d < m$, then it is possible to find a case where the cancellation law fails. For if $a = a_1d$, $m = m_1d$, consider how $da_1 \equiv d(a_1 + m_1) \bmod m$ and $d \not\equiv 0 \bmod m$, yet $a_1 \not\equiv a_1 + m_1 \bmod m$; because if $a_1 \equiv a_1 + m_1 \bmod m$, we would have $m_1 \equiv 0 \bmod m$; but m cannot be a divisor of m_1, for with $d > 1$ we have $0 < m_1 = m/d < m$.

18.2. Various residue systems. As we showed in 17.2 there are exactly m residue classes under congruence modulo m, and any set of m integers, one and only one from each residue class, constitutes, by definition, a *complete residue system.*

If the integers $x_1,x_2,\ldots,x_m$ of a complete residue system also satisfy the added condition $0 \leqq x_i < m$, we have what is called a *least positive residue system.*

If m is odd and $0 \leqq |x_i| \leqq [m/2]$, we have an *absolutely least residue system:* if m is even, we modify this definition by allowing $+m/2$, but deleting $-m/2$.

Thus if $m = 6$:

a complete residue system is $3,4,5,6,7,8$;
the least positive system is $0,1,2,3,4,5$;
the absolutely least system is $-2,-1,0,1,2,3$.

A *reduced residue system* contains, by definition, just those members of a complete residue system for which the cancellation law is valid. Hence by the corollary to **G.6** and by the definition of $\phi(m)$ it follows that a reduced residue system contains exactly $\phi(m)$ classes, determined by integers a for which $(a,m) = 1$.

For example, when $m = 6$, a reduced residue system contains just two classes represented, say, by 1 and 5.

This relation between the Euler phi-function and the validity of the cancellation law for congruence leads us naturally to the next section.

18.3. Another development of the formula for $\phi(n)$. In this section we derive the formula for $\phi(n)$ in a way very different from that employed in Chapter 16 and involving an instructive use of the concepts of complete and reduced residue systems. We begin with a series of six lemmas.

L.1: If $(m,n) = 1$ and if $r_1,r_2,\ldots,r_m$ and $s_1,s_2,\ldots,s_n$ are complete residue systems mod m and mod n, respectively, then the set $\{nr_1 + ms_j\}$ is a set of mn integers forming a complete residue system mod mn.

Proof: (A) The set $\{nr_i + ms_j\}$ does contain mn integers, for there are m choices for i and n choices for j.

(B) We must show that no two of these numbers are congruent mod mn. Suppose $nr_i + ms_j \equiv nr_k + ms_t$ mod mn. Then it follows that $nr_i \equiv nr_k$ mod m; but since $(m,n) = 1$, we may use **G.6** to write $r_i \equiv r_k$ mod m; but since the r's form a *complete* residue system mod m, it follows that $i = k$. Similarly, we have $ms_j \equiv ms_t$ mod n, whence $s_j \equiv s_t$ mod n, whence $j = t$.

Since parts (A) and (B) fulfill the two requirements for a complete residue system mod mn, the proof of **L.1** is complete.

For example, if $m = 3$, $r_1 = 0$, $r_2 = 1$, $r_3 = 2$; and if $n = 4$, $s_1 = 0$, $s_2 = 1$, $s_3 = 2$, $s_4 = 3$; then in lexicographic order the integers of the set $\{nr_i + ms_j\}$ are as follows:

$$0, \quad 3, \quad 6, \quad 9, \quad 4, \quad 7, \quad 10, \quad 13, \quad 8, \quad 11, \quad 14, \quad 17$$

and these are readily checked as forming a *complete* residue system mod 12, albeit not a least positive residue system.

L.2: If $(m,n) = 1$ and if both $(r,m) = 1$ and $(s,n) = 1$, then $(nr + ms,mn) = 1$.

Proof: Let $(nr + ms,mn) = d$ and let p be a prime dividing d. By the *Fundamental Lemma* in **6.1** since p divides mn, p must divide, say, m; then p does not divide n, for $(m,n) = 1$; but p divides $nr + ms$ and hence divides nr; however, not being a divisor of n, p must divide r; hence p divides $(r,m) = 1$; but this is a contradiction. It must be that d has no prime divisors; in other words, $d = 1$.

L.3: If $(m,n) = 1$ and $(a,mn) = 1$, then $a = nr + ms$ where $(r,m) = 1$ and $(s,n) = 1$.

Proof: Since $(m,n) = 1$ there exist integers x and y such that $1 = mx + ny$; hence there exist integers $r = ay$ and $s = ax$ such that $a = nr + ms$. Suppose $(r,m) = d$; then since d divides both r and m

it follows that d divides a; hence d divides $(a,mn) = 1$; hence $d = 1$. Similarly, we may show $(s,n) = 1$.

L.4: If $(m,n) = 1$ and if $r_1,r_2,\ldots,r_{\phi(m)}$ and $s_1,s_2,\ldots,s_{\phi(n)}$ are reduced residue systems mod m and mod n, respectively, then the set $\{nr_1 + ms_j\}$ is a set of $\phi(m)\phi(n)$ integers forming a reduced residue system mod mn.

Proof: (A) There are $\phi(m)\phi(n)$ integers in the set $\{nr_i + ms_j\}$ for there are $\phi(m)$ choices for i and $\phi(n)$ choices for j.

(B) No two of the integers in the set $\{nr_i + ms_j\}$ are congruent mod mn; for each of the reduced residue systems is part of a complete residue system; and the property in question has been proved for complete residue systems in **L.1.**

(C) Each integer $nr_i + ms_j$ is relatively prime to mn, for since the r's and s's form reduced residue systems we have $(r_i,m) = 1$ and $(s_j,n) = 1$; and the required result follows from **L.2.**

(D) Every integer a relatively prime to mn occurs in one of the classes represented by some $nr_i + ms_j$; for this is the implication of **L.3** and EX. *17.4.*

Then (A),(B),(C),(D) together show that the $\phi(m)\phi(n)$ integers of the set $\{nr_i + ms_j\}$ constitute an entire reduced residue system mod mn.

For example, if $m = 3, r_1 = 1, r_2 = 2$; and if $n = 4, s_1 = 1, s_2 = 3$; then in lexicographic order the integers of the set $\{nr_i + ms_j\}$ are as follows: 7, 13, 11, 17; if we note that $13 \equiv 1$ and $17 \equiv 5 \bmod 12$, it is easy to check that we have here a reduced residue system mod 12.

L.5: If $(m,n) = 1$, then $\phi(mn) = \phi(m)\phi(n)$.

Proof: A reduced residue system mod mn contains exactly $\phi(mn)$ integers; but by **L.4**, if $(m,n) = 1$, a reduced residue system mod mn contains $\phi(m)\phi(n)$ integers; hence if $(m,n) = 1$, we have $\phi(mn) = \phi(m)\phi(n)$.

The important point about the proof just given is that it is entirely independent of a priori knowledge of a formula for $\phi(n)$. Hence the property expressed by **L.5**, usually described as the "multiplicative property," can be put to use as part of an entirely different derivation of the formula for $\phi(n)$ originally developed in Chapter 16.

L.6: If p is a prime, then $\phi(p^a) = p^a - p^{a-1}$.

Proof: The proof is made by the simple process of counting the positive integers less than p^a and relatively prime to p^a. The integers k such that $1 \leqq k \leqq p^a$ and such that $(k,p^a) > 1$ must of necessity be multiples of p, so they may be listed as follows: $p,2p,3p,\ldots,(p^{a-1}-1)p,(p^{a-1})p = p^a$; hence they are p^{a-1} in number. All other numbers x with $1 \leqq x < p^a$ are $p^a - p^{a-1}$ in number and have the property $(x,p^a) = 1$; thus we have shown that $\phi(p^a) = p^a - p^{a-1}$.

Theorem: If $n > 1$ is written in standard form as

$$n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k},$$

then

$$\phi(n) = n\left(\frac{p_1 - 1}{p_1}\right)\left(\frac{p_2 - 1}{p_2}\right)\cdots\left(\frac{p_k - 1}{p_k}\right).$$

Proof: Since $p_1^{a_1},p_2^{a_2},\ldots,p_k^{a_k}$ involve distinct primes, we apply **L.5** repeatedly to see that

$$\phi(n) = \phi(p_1^{a_1})\phi(p_2^{a_2})\ldots\phi(p_k^{a_k}).$$

To each $\phi(p_i^{a_i})$ we apply **L.6** to find

$$\phi(p_i^{a_i}) = p_i^{a_i} - p_i^{a_i-1} = p_i^{a_i}\left(\frac{p_i - 1}{p_i}\right).$$

Then the given formula follows immediately by substituting these results for $i = 1,2,\ldots,k$, and rearranging the product in an obvious way.

18.4. The Euler-Fermat theorems.

Lemma: If $r_i,r_2,\ldots,r_{\phi(m)}$ form a reduced residue system mod m and if $(a,m) = 1$, then $ar_1,ar_2,\ldots,ar_{\phi(m)}$ also form a reduced residue system mod m.

Proof: (A) There are $\phi(m)$ numbers in the set $ar_1,ar_2,\ldots,ar_{\phi(m)}$.

(B) Each ar_i is relatively prime to m, for from $(r_i,m) = 1$ and $(a,m) = 1$ it follows that $(ar_i,m) = 1$ by EX. 5.6.

(C) No two distinct ar_i and ar_k are congruent mod m, for from $ar_i \equiv ar_k$ mod m, since $(a,m) = 1$, it would follow from **G.6** that $r_i \equiv r_k$ mod m; but since the r's form a reduced residue system mod m, the last congruence can hold only if $i = k$.

The proof is now complete, for (A),(B),(C) together show that $ar_1,ar_2,\ldots,ar_{\phi(m)}$ satisfy all the requirements to form a reduced residue system mod m.

G.7: Euler's theorem: If $(a,m) = 1$, then $a^{\phi(m)} \equiv 1 \bmod m$.

Proof: Let $r_1, r_2, \ldots, r_{\phi(m)}$ be a reduced residue system mod m. Since $(a,m) = 1$, it follows from the *Lemma* above that $ar_1, ar_2, \ldots, ar_{\phi(m)}$ is also a reduced residue system mod m. Therefore each ar_i is congruent mod m to one and only one r_j. Multiplying these congruences together using **G.3**, we find *upon rearranging the r's on the right in natural order* that

$$a^{\phi(m)} r_1 r_2 \ldots r_{\phi(m)} \equiv r_1 r_2 \ldots r_{\phi(m)} \bmod m.$$

Since $(r_i,m) = 1$, we may employ **G.6** repeatedly to "cancel" $r_1 r_2, \ldots, r_{\phi(m)}$ and find $a^{\phi(m)} \equiv 1 \bmod m$.

For example, since $\phi(9) = 6$ with 1,2,4,5,7,8 each relatively prime to 9, it follows that

$$1 \equiv 1^6 \equiv 2^6 \equiv 4^6 \equiv 5^6 \equiv 7^6 \equiv 8^6 \bmod 9.$$

Corollary: Fermat's theorem: If p is a prime, then for any integer a, we have $a^p \equiv a \bmod p$.

Proof: Since all integers x such that $x \not\equiv 0 \bmod p$ satisfy $(x,p) = 1$ and since $\phi(p) = p - 1$, it follows from **G.7** that for these x we have $x^{p-1} \equiv 1 \bmod p$. Multiplying each side by x, we find $x^p \equiv x \bmod p$. But this latter congruence is satisfied also by $x \equiv 0 \bmod p$, hence the proof is complete.

For example, since 7 is a prime, it follows that $1^7 \equiv 1$, $2^7 \equiv 2$, $3^7 \equiv 3$, $4^7 \equiv 4$, $5^7 \equiv 5$, $6^7 \equiv 6$, $7^7 \equiv 7$, mod 7. Similarly, we know $(10)^6 \equiv 1$, $(113)^6 \equiv 1$, mod 7, etc.

EXERCISES

EX. *18.1.* Illustrate **L.1** and **L.4** with $m = 3$ and $n = 5$.

EX. *18.2.* Find the absolutely least residue systems mod 3 and mod 5. Show that "u and v of the same, or different, parity" is equivalent to "$u \equiv v$, or $u \not\equiv v$, mod 2." Show that in every primitive Pythagorean triplet one and only one member is a multiple of 3; —— of 4; —— of 5.

EX. *18.3.* Use Fermat's theorem to show that every prime, except 2 and 5, divides infinitely many of the integers: 9,99,999,9999, ...

EX. *18.4.* Show that for every integer n, the number $n^{13} - n$ is divisible by 2730.

EX. *18.5.* Study the binomial coefficient $p!/(p-r)!r!$ where p is a prime and $0 < r < p$, and prove directly that

$$(a + b)^p \equiv a^p + b^p \bmod p$$

(but don't let any freshmen observe this heresy!).

Exercises

EX. *18.6.* Give an independent proof of Fermat's theorem using EX. *18.5* and mathematical induction.

EX. *18.7.* Define Carmichael's lambda-function as follows: $\lambda(1) = \phi(1)$, $\lambda(2) = \phi(2)$, $\lambda(4) = \phi(4)$; $\lambda(2^a) = \frac{1}{2}\phi(2^a)$, $a > 2$; $\lambda(p^a) = \phi(p^a)$ if p is an odd prime; and if m is written in standard form as $m = 2^a p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$ where p_i is an odd prime, then $\lambda(m) = [\lambda 2^a), \lambda(p_1^{a_1}), \lambda(p_2^{a_2}), \dots, \lambda(p_k^{a_k})]$, where the brackets indicate least common multiple. Use EX. *17.9* and **G.7** to prove that if $(a,m) = 1$, then $a^{\lambda(m)} \equiv 1$ mod m.

EX. *18.8.* Show that $\lambda(m)$ is a divisor of $\phi(m)$. Compare $\lambda(m)$ and $\phi(m)$ when $m = 2^6 \cdot 3 \cdot 5 \cdot 7 \cdot 17 \cdot 19$.

For the following exercises define an "abstract group" to consist of a set G of elements $a,b,\dots$, with an equivalence relation and an ordered binary operation $ab = c$ which is (*1*) closed, (*2*) associative, (*3*) has an identity, and (*4*) has for each element an inverse. Note that such a system differs from the "transformation group" defined in Chapter 11, for the elements are not required to be transformations, the operation is not necessarily that of forming the product of two transformations, and it is therefore necessary to postulate (or with examples, to prove), not automatically have, the associative property.

EX. *18.9.* Show that all integers form a group under the operation of addition.

EX. *18.10.* Show that all the residue classes mod m form a group under the operation of addition of residue classes. Compare with EX. *17.14.*

EX. *18.11.* Show that all the residue classes of a reduced residue system mod m form a group under the operation of multiplication of residue classes.

EX. *18.12.* Let S be the residue classes of a reduced residue system mod m. For each r-class in S define a transformation T_r of S as follows: $aT_r \equiv ar$ mod m. Show that the set G of all $\phi(m)$ transformations T_r forms a transformation group. Compare G with the group in EX. *18.11.*

EX. *18.13.* Let S be the set of all residue classes mod m. Define a transformation $T(r,b)$ of S as follows: $aT(r,b) \equiv ar + b$ mod m, where $(r,m) = 1$, $1 \leqq r < m$, $0 \leqq b < m$. Show that the set G of all $m\phi(m)$ transformations $T(r,b)$ forms a transformation group.

▶*Excellence is evident in full and adequate solutions to problems; for whatsoever theorem solves the most complicated problem of the kind, does with a due reduction reach all the subordinate cases.* —E. HALLEY

CHAPTER 19*

LINEAR CONGRUENCES

19.1 Theory of congruences. Let

$$F(x) = a_0x^n + a_1x^{n-1} + \dots + a_n, \qquad n \geqq 1,$$

be a polynomial with integers as coefficients and with $a_0 \not\equiv 0 \bmod m$; then $F(x) \equiv 0 \bmod m$ will be said to be a congruence of degree $n \bmod m$.

If there exists an integer x_1 such that $F(x_1) \equiv 0 \bmod m$, it would be natural to define x_1 to be a solution of the congruence. However, our earlier theorems show that if X_1 is any integer such that $X_1 \equiv x_1 \bmod m$, then we also have $F(X_1) \equiv 0 \bmod m$. Thus if one solution can be found, then infinitely many others can be obtained, but related to each other in an obvious manner. To avoid this trivial duplication we therefore agree to speak in terms of residue classes and we define the x_1-*residue class* to be a solution of $F(x) \equiv 0 \bmod m$ if and only if $F(x_1) \equiv 0 \bmod m$.

If x_1 and x_2 are solutions of $F(x) \equiv 0 \bmod m$, they will be considered as distinct solutions if and only if $x_1 \not\equiv x_2 \bmod m$. Hence by the number of solutions of a congruence mod m we shall mean the maximum number of solutions incongruent in pairs.

According to this definition there cannot be more than m solutions for any given congruence, since there are only m different residue

*Chapter 19 is a basic chapter except for sections **19.4** and **19.5** which are supplementary.

classes to be considered. If m is small, this implies that all the solutions may be found by direct substitution.

In elementary algebra courses most of our readers will have studied, at least in an introductory way, the "theory of equations" of the complex number system, beginning with linear equations and progressing to quadratics, cubics, etc. It is therefore natural that here we propose a study of the "theory of congruences," starting with the linear case and continuing to congruences of higher degree. Many points of difference between the two theories will appear.

As explained above, for a congruence a "solution" will mean a "residue class," so each solution will actually involve infinitely many integers; and "distinct" solutions are defined to be "incongruent" solutions. In contrast a solution of a polynomial equation over the complex number system is individual; and distinct solutions are unequal solutions. However, the wider view of an "equivalence relation" which we have been emphasizing makes this situation readily understandable, for congruence of integers mod m and equality of complex numbers are two different equivalence relations: the first has infinitely many elements in each equivalence class, the second has only one element in each equivalence class.

A congruence may have no solution. For example, witness $x^2 \equiv 3$ mod 5, for trying in turn each of the five possibilities: 0,1,2,3,4, we fail to find a solution. In contrast, over the complex number system, every polynomial equation with coefficients in the system has a solution within the system.

Again, a congruence may have more distinct solutions than its degree. Consider the example $x^2 \equiv 1$ mod 8 which is of degree 2, but has four incongruent solutions: 1,3,5,7. In contrast, a polynomial equation over the complex number system of degree n has at most n distinct solutions.

But the most striking difference is that we shall be able to give an explicit method for solving any congruence of any degree and any modulus m. (Of course, as explained above, one such "method" would be to substitute, in turn, each of the integers of a complete residue system, say, $0,1,2,\ldots,m-1$, and while this method is complete in a finite number of steps, it is not practical for large values of m.) In contrast, no comparable method can be found in the theory of equations for complex numbers for equations of degree greater than 4.

19.2. Linear congruences in one unknown. For emphasis we repeat the remarks above that in the following theorems the words "unique solution" must be interpreted carefully, not to mean one integer, but to mean one residue class; so that there may be two solutions, unequal in the usual sense, but "equal" in the sense of being "congruent." Thus $2x \equiv 1 \bmod 5$ is said to have the "unique" solution 3; although 8 and 13 are other solutions, they are not counted as different, because each of these is congruent to 3 mod 5.

G.8: If $(a,m) = 1$, then $ax \equiv b \bmod m$ has a unique solution.

Proof: For a first proof, we appeal to Euler's theorem **G.7** and to **G.3** and multiply each side of the given congruence by $a^{\phi(m)-1}$ to find that we must have $x \equiv ba^{\phi(m)-1} \bmod m$.

For a second proof, we return to the Euclid algorithm, for since $(a,m) = 1$, we know there exist integers s and t such that $as + mt = 1$ and then $a(sb) + m(tb) = b$. Hence we find that $x = ab$ is a solution of the given congruence, for we have $ax - b = (-tb)m$ which implies $ax \equiv b \bmod m$. If X is any other solution, so that $aX \equiv b \bmod m$, then we see by **G.2** that $aX \equiv ax \bmod m$; then since $(a,m) = 1$, we may apply **G.6** to see that $X \equiv x \bmod m$, so the solution is unique.

Of course the main idea of the first proof given above is to produce explicitly an integer A such that $aA \equiv 1 \bmod m$, for then, when multiplied by A, the congruence $ax \equiv b \bmod m$ will take the solved form $x \equiv Ab \bmod m$. The process is comparable to that used in solving a linear equation and A is called a "reciprocal" or "inverse" of a; however, A must *not* be written in the form $1/a$, for we are dealing strictly with integers. From Euler's theorem an explicit value for A is $a^{\phi(m)-1}$; but often in problems a suitable value of A can be found by inspection (note that the "uniqueness feature" of **G.8** guarantees that $a^{\phi(m)-1} \equiv A \bmod m$ for any A such that $aA \equiv 1 \bmod m$).

The second method of proof shows that our congruence problem is equivalent to the Diophantine equation $ax + my = b$ studied in **12.1** and suggests an entirely new method of solving the Diophantine equation when $(a,m) = 1$.

For example, let us solve $17x + 11y = 16$ by the congruence method. Considered mod 11 this problem becomes $6x \equiv 5 \bmod 11$. We then seek A so that $6A \equiv 1 \bmod 11$; by Euler's theorem we know, since $\phi(11) = 10$, that $A \equiv 6^9 \bmod 11$; however, by inspection, we can see immediately that $A \equiv 2 \bmod 11$. Then $x \equiv 2\cdot 5 \equiv 10 \bmod 11$, or $x = 10 + 11k$, where k is any integer. Substituting this value of x and solving for y we find that $11y = 16 - 170 - 17(11k)$, whence $y = -14 - 17k$, which completes the solution.

In general, using the first proof of **G.8**, we can write the solution of $ax + my = b$, with $(a,m) = 1$, explicitly, as follows:

$$x = ba^{\phi(m)-1} + km, \quad y = -b(a^{\phi(m)} - 1)/m - ka;$$

or, more conveniently, if A is any solution of $aA \equiv 1 \bmod m$, we may write the solution as follows:

$$x = bA + km, \quad y = -b(aA - 1)/m - ka,$$

where, in both cases, k is an arbitrary integral parameter.

G.8.1: If $(a,m) = d$, then $ax \equiv b \bmod m$ has no solution when d is not a divisor of b; but if d divides b, there are exactly d solutions.

Proof: Since the congruence is equivalent to $ax + mk = b$ in integers x and k, the existence of solutions x and k requires that $d = (a,m)$ divide b. Suppose then that this requirement is satisfied and let $a = a_1d$, $m = m_1d$, $b = b_1d$; then according to **G.6** the congruence $ax \equiv b \bmod m$ reduces to $a_1x \equiv b_1 \bmod m_1$. But $(a_1,m_1) = 1$, hence **G.8** is applicable and hence this new congruence has a unique solution mod m_1, say $X \equiv b_1a_1^{\phi(m_1)-1} \bmod m_1$, to be explicit. Then X, $X + m_1$, $X + 2m_1$, $\ldots$, $X + (d-1)m_1$ make up exactly d solutions mod m of $ax \equiv b \bmod m$. Any other solution must have the form $X + sm_1$ and must be congruent mod m to one of the solutions listed; for if we set $s = qd + r$, $0 \leqq r < d$, then $X + sm_1 = X + (qd + r)m_1 = X + qm + rm_1 \equiv X + rm_1 \bmod m$, and $X + rm_1$ is in the list. The solutions listed are distinct mod m, for they are of the form $X + r_im_1$ where the r_i form a complete residue system mod d. Then if $X + r_im_1 \equiv X + r_jm_1 \bmod m$, we find by **G.3** that $r_im_1 \equiv r_jm_1 \bmod m$ and then by **G.6** that $r_i \equiv r_j \bmod d$, hence $i = j$.

For example, we solve $39x \equiv 65 \bmod 52$ as follows: since $a = 39$, $m = 52$, $(a,m) = 13$, $b = 65 = 5\cdot 13$, there must be 13 distinct solutions; the "reduced" congruence is $3x \equiv 5 \bmod 4$, with the unique solution $x \equiv 3 \bmod 4$; hence $x = 3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, 47, 51$, are the 13 distinct solutions of the original congruence.

As a related example, we note that $39x \equiv 64 \bmod 52$ has *no* solution, since 64 is *not* a multiple of 13.

19.3 The Chinese remainder theorem. Problems of the kind which can be solved by the following theorem were solved by the Chinese in ancient times, and in honor of these early contributions we term our theorem the "Chinese remainder theorem," although the notion of congruence enables us to state the theorem and solution and make the proof in a much more condensed and convenient form than was available to these ancients.

G.9: If $m_1, m_2, \ldots, m_k$ are given moduli, relatively prime in pairs, then the system of linear congruences

$$x \equiv a_1 \bmod m_1, \quad x \equiv a_2 \bmod m_2, \quad \ldots, \quad x \equiv a_k \bmod m_k$$

where a_i are given remainders, has a unique solution modulo m, where $m = m_1 m_2 \ldots m_k$.

Proof: If we define M_i by requiring $m_i M_i = m$, then since the m_i are relatively prime in pairs, it follows that $(M_i, m_i) = 1$, and hence by **G.8** there exists an integer x_i such that $M_i x_i \equiv 1 \bmod m_i$, for $i = 1, 2, \ldots, k$. Then a solution x of the given system of congruences is provided by

$$x = \sum_{i=1}^{k} M_i x_i a_i.$$

For if we substitute x in any of the given congruences, say the ith congruence, we find that M_j, for every $j \neq i$, contains m_i as a factor so that $M_j \equiv 0 \bmod m_i$, $j \neq i$; but $M_i x_i \equiv 1 \bmod m_i$, hence $x \equiv a_i \bmod m_i$, as required.

If X is another solution, then $X \equiv a_i \bmod m_i$, for $i = 1, 2, \ldots, k$; then by **G.2**, $X \equiv x \bmod m_i$, for $i = 1, 2, \ldots, k$. Since the m_i are relatively prime in pairs, we may use Ex. *19.3* of this chapter to conclude that $X \equiv x \bmod m = m_1 m_2 \ldots m_k$, which completes the proof of **G.9**.

Just for fun let's do the "Chinese remaining problem."

A band of 17 pirates upon dividing their doubloons in equal portions found 3 coins remaining which they agreed they ought to give to their Chinese cook, Wun Tu. But 6 of the pirates were killed in a brawl, and now when the total fortune was divided equally among them, there were 4 coins left over which they considered giving to Wun Tu. In a shipwreck that followed only 6 of the pirates, the coins, and the cook were saved; this time an equal division left a remainder of 5 coins for the cook. Wearying of his masters' niggardliness, Wun Tu took advantage of his culinary position to concoct a potent mushroom stew so that the entire fortune in doubloons became his own. With the aid of the Chinese remainder theorem we are to find the two smallest numbers of coins which may have been the fortune of the Chinese remaining.

Stripped of embellishment, our problem is to find the two smallest positive solutions of the system of congruences:

$$x \equiv 3 \bmod 17, \quad x \equiv 4 \bmod 11, \quad x \equiv 5 \bmod 6.$$

Since 17, 11, 6 are relatively prime in pairs, the theorem **G.9** may be applied. Here we have

$$\begin{array}{llll} m_1 = 17, & M_1 = 66, & 66x_1 \equiv -2x_1 \equiv 1 \bmod 17, & \text{so } x_1 = 8; \\ m_2 = 11, & M_2 = 102, & 102x_2 \equiv \quad 3x_2 \equiv 1 \bmod 11, & \text{so } x_2 = 4; \\ m_3 = 6, & M_3 = 187, & 187x_3 \equiv \quad\ \ x_3 \equiv 1 \bmod 6, & \text{so } x_3 = 1. \end{array}$$

Then the complete solution is given by

$$\begin{aligned} X = x + mt &= M_1x_1a_1 + M_2x_2a_2 + M_3x_3a_3 + m_1m_2m_3t \\ &= (66)(8)(3) + (102)(4)(4) + (187)(1)(5) + (17)(11)(6)t \\ &= 4151 + 1122t. \end{aligned}$$

Quite symbolically, Wun Tu's fortune depends upon multiples of 1122 (!). The desired solutions are found by taking $t = -3$ and $t = -2$ with the results $x = 785$ and $x = 1907$, respectively.

19.4. Systems of n linear congruences in n unknowns. In contrast to the system studied in **19.3** involving *one* unknown, but k *different* moduli, we here study a system of n linear congruences in n unknowns, say, $x_1, x_2, \ldots, x_n$, all with the *same* modulus m. If we set

$$L_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - c_i,$$

where the a_{ij} and the c_i are given integers with $i,j = 1,2,\ldots,n$, then the system which we propose to study may be indicated by

$$L_i \equiv 0 \bmod m, \qquad i = 1,2,\ldots,n.$$

Two such systems will be said to be *equivalent* if they have exactly the same solutions.

G.10: If $L_j' = L_j$ when $j \neq k$, but

$$L_k' = b_1L_1 + b_2L_2 + \ldots + b_kL_k + \ldots + b_nL_n,$$

where the b's are integers, then if $(b_k,m) = 1$, the system $L_i' \equiv 0 \bmod m$, $i = 1,2,\ldots,n$, is equivalent to the system $L_i \equiv 0 \bmod m$, $i = 1,2,\ldots,n$; but if $(b_k,m) > 1$, the primed system may have extraneous solutions which will not satisfy the original system.

Proof: (A) Obviously each solution of the original system is a solution of the primed system, for the only congruence which is different is the kth one; and since by hypothesis $L_i \equiv 0 \bmod m$ for $i = 1,2,\ldots,n$, it follows by substitution that $L_k' \equiv 0 \bmod m$.

(B) Conversely, each solution of the primed system is a solution of all the congruences of the original system except possibly the kth one, because $L_i = L_i'$ when $i \neq k$. For the same reason the congruence $L_k' \equiv 0 \bmod m$ takes the simplified form $b_kL_k \equiv 0 \bmod m$. Hence by **G.6** if $(b_k,m) = 1$, we find that we must have $L_k \equiv 0 \bmod m$, so that the two systems are equivalent. But if $(b_k,m) = d > 1$, then $b_kL_k \equiv 0 \bmod m$ may be satisfied by having $L_k \equiv s(m/d) \bmod m$ for $s = 1,2,\ldots,d-1$, as well as by having $L_k \equiv 0 \bmod m$; such a situation shows plainly that a solution of the primed system need not be a solution of the original system and an attempt to replace the original system by the primed system may, when $(b_k,m) > 1$, introduce extraneous solutions.

The suggestion is strong that by repeated application of **G.10**, with suitable and cautious choice of the multipliers b, we may be able to replace a given system by an equivalent system of the type

$$A_ix_i \equiv B_i \bmod m, \qquad i = 1,2,\ldots,n$$

so that the solutions, if any exist, may be found by applying **G.8**. Since no solutions can be lost by the method, it is perhaps easier to rely on a check by substitution to delete extraneous solutions than it is to avoid the introduction of the extraneous solutions. Following the terminology of the theory of equations we may designate such a method as is proposed here as "elimination by addition and subtraction." The method is illustrated in the following example. (If the method seems haphazard, the student who is acquainted with the theory of determinants can substitute an explicit method and tests for the existence of solutions and the exclusion of extraneous solutions, see EX. *19.11.*)

Given $L_1 = 2x + 11y - 5$ and $L_2 = x + 3y$, solve the system $L_1 \equiv 0$, $L_2 \equiv 0$, mod 15.

To eliminate first x and then y we consider

$$L_1' = L_1 - 2L_2 = 5y - 5 \quad \text{and} \quad L_2' = 3L_1' - 5L_2 = -5x + 15.$$

By **G.10** the system $L_1' \equiv 0$, $L_2 \equiv 0$, mod 15 is equivalent to $L_1 \equiv 0$, $L_2 \equiv 0$, mod 15, because $(1,15) = 1$; but the system $L_1' \equiv 0$, $L_2' \equiv 0$, mod 15 may have solutions extraneous to those of the system $L_1' \equiv 0$,

$L_2 \equiv 0$, mod 15, because $(5,15) = 5 > 1$. In fact, $L_1' \equiv 0$ mod 15, or $5y \equiv 5$ mod 15, has solutions $y = 1,4,7,10,13$; and $L_2' \equiv 0$ mod 15, or $10x \equiv 0$ mod 15, has solutions $x = 0,3,6,9,12$; so that the system $L_1' \equiv 0$, $L_2' \equiv 0$, mod 15 has a grand total of 25 solutions (pairing each value of x with each value of y). However, a check reveals that only 5 of these solutions solve the original system, namely:

$$(x,y) = (0,10); (3,4); (6,13); (9,7); (12.1).$$

The theorem **G.10** guarantees that this is the complete set of solutions of the original problem.

19.5. A cipher based on congruences. The ideas of the preceding section have been made the basis of an interesting, flexible, and, in a certain sense, unbreakable cipher.

To begin with we select for the modulus m any prime just a little larger than the number of letters in the alphabet required for the messages. Thus with the usual 26-letter English alphabet in mind, we might select $m = 29$.

Then we adjoin to the alphabet a sufficient number of useful symbols so that we can establish a one-to-one correspondence (C) between residue classes mod m and symbols of the (enlarged) alphabet. For example, to the alphabet $A,B,\ldots,Z$ we might adjoin the symbols &, ., ?, supposing that we have $m = 29$, and assign to these "letters" the residue classes $1,2,\ldots,26$ and $27,28,0$, respectively.

The purpose in choosing a *prime* modulus is to afford us a great variety of ways to choose n^2 integers a_{ij} so that the system of congruences

$$(E) \qquad c_i \equiv a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \text{ mod } m, \quad i = 1,2,\ldots,n$$

will have a unique solution for any selection of the c_i (see Ex. *19.11* for the exact conditions), namely:

$$(D) \qquad x_i \equiv d_{i1}c_1 + d_{i2}c_2 + \ldots + d_{in}c_n \text{ mod } m, \quad i = 1,2,\ldots,n.$$

The integers a_{ij} which are selected are said to form the *encipherment-matrix*; the integers d_{ij}, which can be computed if the encipherment-matrix is known, are said to form the *decipherment-matrix*.

A given message is enciphered in the following way:

(*1*) it is divided into groups of n letters according to some plan (P), perhaps just successive groups;

(2) the letters of a group are designated in order as $x_1,x_2,\ldots,x_n$ and each is given its numerical equivalent according to the plan (C);

(3) using the agreed upon encipherment-matrix the values of $c_1,c_2,\ldots,c_n$ are computed mod m from the congruences (E);
(4) the numerical values of $c_1,c_2,\ldots,c_n$ are replaced by their letter equivalents according to the plan (C) and this group of n letters is one group of the enciphered message.

A ciphered message is reduced to "clear" in the following way:
(1) the message is divided into successive groups of n letters; the letters of a group are designated in order as $c_1,c_2,\ldots,c_n$ and are replaced by their numerical equivalents according to the plan (C);
(2) using the computed decipherment-matrix the values of $x_1,x_2,\ldots,x_n$ are computed mod m from the congruences (D);
(3) the numerical values of $x_1,x_2,\ldots,x_n$ are replaced by their letter equivalents according to the plan (C);
(4) the various groups of n letters are arranged in proper position by reversing the plan (P).

For example, let us take $m = 29$ and the correspondence (C) suggested above; let us take $n = 3$ and the following congruences for encipherment:

$$(E)\qquad \begin{aligned} c_1 &\equiv x_1 + 5x_2 + 6x_3,\\ c_2 &\equiv 8x_1 + 2x_2 + 4x_3,\\ c_3 &\equiv 9x_1 + 7x_2 + 3x_3, \quad \bmod 29. \end{aligned}$$

Then if we follow the standard plan (P) a message such as *HE DIED* becomes first (8,5,4)(9,5,4) and enciphers by (E) as (28,3,3)(0,11,12) so that the cipher message is *.CC?KL.*

If we solve the system of congruences (E) by the method in **19.4** we obtain (see EX. *19.12*) the following formulas for decipherment:

$$(D)\qquad \begin{aligned} x_1 &\equiv 13c_1 + 17c_2 + 19c_3,\\ x_2 &\equiv 14c_1 + 25c_2 + 5c_3,\\ x_3 &\equiv 25c_1 + 25c_2 + 4c_3, \quad \bmod 29. \end{aligned}$$

This cipher is unbreakable in the following sense: even if the "enemy" knows the method of encipherment, the correspondence (C), and the plan (P), he has a poor chance of guessing the key encipherment-matrix, because there exists a matrix of integers a_{ij} such that *any* message of only n letters will, when enciphered by the congruences based on that matrix, take the form $c_1,c_2,\ldots,c_n$ of the cipher message which the "enemy" is trying to break.

For example, the correct decipherment of *.CC* by the congruences (D) above is *HED*. But it is easy to find an encipherment-matrix for which the decipherment of .CC would be, say, THE. We would have to choose the a_{ij} so that

$$\begin{aligned} a_{11}20 + a_{12}8 + a_{13}5 &\equiv 28, \\ a_{21}20 + a_{22}8 + a_{23}5 &\equiv 3, \\ a_{31}20 + a_{32}8 + a_{33}5 &\equiv 3, \quad \text{mod } 29, \end{aligned}$$

and so that the new system (E') would have a unique solution (D'). One suitable choice is

$$(E') \qquad \begin{aligned} c_1 &\equiv x_1 + x_2 \qquad\qquad , \\ c_2 &\equiv x_1 + 4x_2 + 25x_3, \\ c_3 &\equiv 2x_1 + 28x_2 \qquad\quad , \quad \text{mod } 29. \end{aligned}$$

Since in practice the value of n must be fairly small, the almost inevitable recursion of certain combinations of letters in exactly the same position in the n-groups would, however, probably allow the skilled cryptographer to break the cipher, particularly if he had several long messages of more than n letters to study.

EXERCISES

EX. *19.1.* Solve $513x \equiv -17$ mod 1163.

EX. *19.2.* Solve $66x \equiv 121$ mod 737.

EX. *19.3.* If $X \equiv x$ mod r and if $X \equiv x$ mod s and if $(r,s) = 1$, prove that $X \equiv x$ mod rs. By induction establish the result needed in the last step of the proof of **G.9**.

EX. *19.4.* If $X \equiv x$ mod r and if $X \equiv x$ mod s, prove that $X \equiv x$ mod $[r,s]$ (see EX. *6.5*).

EX. *19.5.* If $x \equiv a$ mod r and $x \equiv b$ mod s, prove that $a \equiv b$ mod (r,s).

EX. *19.6* Use EX. *19.4* and EX. *19.5* to generalize **G.9** to "The system $x \equiv a_i$ mod m_i, $i = 1,2,\ldots,k$, has a solution if and only if (m_i,m_j) divides $a_i - a_j$ for $i,j = 1,2,\ldots,k$; if the solution exists, it is unique mod $[m_1,m_2,\ldots,m_k]$."

EX. *19.7.* Find the least two positive integers with the remainders 2,3,2, when divided by 3,5,7, respectively. (Sun-Tsu, first century.)

EX. *19.8.* Find a number having remainders 2,3,4,5, when divided by 3,4,5,6, respectively. (Brahmegupta, seventh century.) (First apply **G.9** to the first three conditions.)

EX. *19.9.* Solve the following system, mod 29:

$$\begin{aligned} 2x - 4y + z &\equiv 3, \\ x + 5y - z &\equiv 2, \\ 3x - y + 2z &\equiv 1. \end{aligned}$$

EX. *19.10.* Solve the system in EX. *19.9*, mod 24.

EX. *19.11.* This exercise is intended for students who are familiar with the theory of determinants. Modify Cramer's rule so that it will apply to the solution of a system of n linear congruence in n unknowns, mod m; and apply **G.10** to prove that there will be a *unique* solution of the system *when*

$$(D,m) = (A_{11},m) = (A_{22},m) = \ldots = (A_{nn},m) = 1.$$

where D is the determinant of the a_{ij} and where A_{ii} is the cofactor of a_{ii}. Apply **G.8.1** to show that *in general* the number of solutions may vary from *none* to $(D,m)^n$.

EX. *19.12.* Obtain the solution (D) mod 29 given in **19.5**, and decipher the following cipher message: *WV&VLJ.*

> ▶ *The higher arithmetic presents us with an inexhaustible store of interesting truths—of truths, too, which are not isolated, but stand in a close internal connection, and between which, as our knowledge increases, we are continually discovering new and sometimes wholly unexpected ties. A great part of its theories derives an additional charm from the peculiarity that important propositions, with the impress of simplicity upon them, are often easily discoverable by induction, and yet are of so profound a character that we cannot find their demonstration till after many vain attempts; and even then, when we do succeed, it is often by some tedious and artificial process, while the simpler methods may long remain concealed.* —C. F. GAUSS

CHAPTER 20*

CONGRUENCES OF HIGHER DEGREE

20.1. Preliminary considerations. Let us now direct our attention to congruences of any degree. We will use the notation introduced in **19.1**, letting

$$F(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n, \qquad n \geqq 1,$$

be a polynomial with integers as coefficients and with $a_0 \not\equiv 0 \bmod m$; then we will consider the congruence $F(x) \equiv 0 \bmod m$.

As explained in the previous lesson, when m is *small*, all the solutions can be obtained by direct trial of integers from a complete residue system mod m; and according to our agreement about distinct solutions, there can never be more than m solutions. But our object

[*] Chapter 20 is a basic chapter.

here is to obtain a method, more suitable than mere trial, when m is *large.*

We must bear in mind that even if the given congruence is of degree n, there may be more than n solutions; however, a later theorem will show that this anomaly can arise only when the modulus m is composite.

To preserve the continuity of the following chain of theorems— **G.11, G.12, G.13, G.14**—we shall present all the theorems and proofs, and then begin an example in whose solution we can illustrate all the theorems.

20.2. Reduction of the solution of congruences mod m to the solution of congruences mod p^s where p is a prime. In this section we shall use the Chinese remainder theorem to prove the following theorem:

G.11: If $m > 1$ is written in standard form as $m = p_1^{s_1}p_2^{s_2}\ldots p_k^{s_k}$ where p_i is a prime and $1 < p_1 < p_2 < \ldots < p_k$, then the solution of $F(x) \equiv 0 \bmod m$ depends upon the solution of $F(x) \equiv 0 \bmod p_i^{s_i}$, for $i = 1,2,\ldots,k$.

Proof: Obviously, if $F(x) \equiv 0 \bmod m$, then $F(x) \equiv 0 \bmod p_i^{s_i}$ for $i = 1,2,\ldots,k$; so every solution of the given congruence mod m is a solution of the several congruences mod $p_i^{s_i}$.

Conversely, suppose that all solutions of the congruences $F(x) \equiv 0 \bmod p_i^{s_i}$ can be found. Let us suppose that integers $x_1, x_2, \ldots, x_k$ have been found so that

$$F(x_1) \equiv 0 \bmod p_1^{s_1}, F(x_2) \equiv 0 \bmod p_2^{s_2},\ldots,F(x_k) \equiv 0 \bmod p_k^{s_k}.$$

Then since the $p_i^{s_i}$, $p_j^{s_j}$ are relatively prime in pairs, we are in a position to apply the Chinese remainder theorem, **G.9**, and to find an integer x such that

$$x \equiv x_1 \bmod p_1^{s_1}, \quad x \equiv x_2 \bmod p_2^{s_2},\ldots, \quad x \equiv x_k \bmod p_k^{s_k}.$$

Then since $F(x) \equiv F(x_i) \equiv 0 \bmod p_i^{s_i}$, for $i = 1,2,\ldots,k$, it follows from EX. *19.3* that $F(x) \equiv 0 \bmod m$. Moreover, **G.9** asserts that the x which has just been found is unique mod m. Hence we have shown that each distinct set of solutions $x_1,x_2,\ldots,x_k$ of the system of several congruences leads to a distinct solution of the given congruence mod m. Thus if there are T_i incongruent solutions x_i of $F(x) \equiv 0 \bmod p_i^{s_i}$, then there will be $T = T_1T_2\ldots T_k$ incongruent solutions x of

$F(x) \equiv 0 \bmod m$. It should be noted, that if any $T_i = 0$, then, of course, $T = 0$ so that there is *no* solution mod m.

20.3. Reduction of the solution of congruences mod p^s to solutions mod p. In this section we show that the solution of a congruence mod p^s, where p is a prime and $s > 1$, can be reduced to the solution of a congruence mod p^{s-1}, hence by repeated applications of this process the solution can be reduced to the solution of a congruence mod p.

To carry out the next proof we need first to make a slight digression and consider certain consequences of the binomial theorem of EX. 3.7. We note that if a and b are integers and if n is an integer, $n \geqq 2$, then

(*20.1*) $$(a + b)^n = a^n + na^{n-1}b + b^2Q_n(a,b)$$

where $Q_n(a,b)$ is an *integer*, depending on n,a, and b.

By repeated application of (*20.1*) we find that if $n \geqq 2$, then

$$\begin{aligned} F(a+b) &= a_0(a+b)^n + a_1(a+b)^{n-1} + \ldots + a_{n-2}(a+b)^2 + a_{n-1}(a+b) + a_n \\ &= (a_0a^n + a_1a^{n-1} + \ldots + a_{n-2}a^2 + a_{n-1}a + a_n) \\ &\quad + b\{na_0a^{n-1} + (n-1)a_1a^{n-2} + \ldots + 2a_{n-2}a + a_{n-1}\} \\ &\quad + b^2\{a_0Q_n(a,b) + a_1Q_{n-1}(a,b) + \ldots + a_{n-2}Q_2(a,b)\}. \end{aligned}$$

Let us define a new function, $F'(x)$, read "F-prime of x" and called the "derivative of $F(x)$," derived from $F(x)$ according to the following formula:

(*20.2*) $$F'(x) = na_0x^{n-1} + (n-1)a_1x^{n-2} + \ldots + 2a_{n-2}x + a_{n-1}, \quad n \geqq 1.$$

In terms of $F'(x)$ we find that $F(a + b)$ may be written

(*20.3*) $$F(a + b) = F(a) + bF'(a) + b^2Q.$$

where $Q = a_0Q_n(a,b) + a_1Q_{n-1}(a,b) + \ldots + a_{n-2}Q_2(a,b)$ if $n \geqq 2$

and $Q = 0$ if $n = 1$.

For example,
$$\begin{aligned} F(x) &= 2x^3 + 3x^2 + 5x + 7, \\ F'(x) &= 6x^2 + 6x + 5, \\ F(a+b) &= (2a^3 + 3a^2 + 5a + 7) + b(6a^2 + 6a + 5) + b^2Q, \\ Q &= 6a + 2b + 3. \end{aligned}$$

In the application which we shall make of (*20.3*) we shall not need to know the exact value of Q, but merely that Q is an *integer*.

C.12. If $s > 1$ the solution of $F(x) \equiv 0 \bmod p^s$, where p is a prime, depends upon the solution of $F(x) \equiv 0 \bmod p^{s-1}$.

Proof: We begin by observing that each solution x of $F(x) \equiv 0 \bmod p^s$ is obviously a solution of $F(x) \equiv 0 \bmod p^{s-1}$. Consequently all solutions of $F(x) \equiv 0 \bmod p^s$ must be included among* the solutions of $F(x) \equiv 0 \bmod p^{s-1}$. In other words, if x is a solution of $F(x) \equiv 0 \bmod p^s$, it must be possible for us to find a solution X of $F(x) \equiv 0 \bmod p^{s-1}$ so that $x \equiv X \bmod p^{s-1}$; i.e., x must have the form $x = X + tp^{s-1}$ for a suitably chosen integer t.

We will suppose then that all solutions X of $F(x) \equiv 0 \bmod p^{s-1}$ have been found and we shall check each of these, in turn, to see if one or more integers t can be found so that $x = X + tp^{s-1}$ will be a solution of $F(x) \equiv 0 \bmod p^s$, for we are certain from the above discussion that this is the only way solutions of the latter congruence can arise.

In the attempt to find suitable values of t we may use *(20.3)* for this equation allows us to write

$$F(x) = F(X + tp^{s-1}) = F(X) + tp^{s-1}F'(X) + t^2(p^{s-1})^2Q$$

where Q is an integer. Since we are seeking solutions x of $F(x) \equiv 0 \bmod p^s$, and since for $s > 1$ it is clear that $(p^{s-1})^2 \equiv 0 \bmod p^s$, we are led to the following restriction on t:

$$F(X) + tp^{s-1}F'(X) \equiv 0 \bmod p^s.$$

However, by hypothesis $F(X) \equiv 0 \bmod p^{s-1}$ so there exists an integer M so that $F(X) = Mp^{s-1}$. Therefore the congruence restriction on t may be replaced by the following congruence mod p:

$$M + tF'(X) \equiv 0 \bmod p. \tag{20.4}$$

To the congruence *(20.4)* we may apply all the results of **G.8.1**, as follows:

there is *one* solution t if $F'(X) \not\equiv 0 \bmod p$;
there is *no* solution t if $F'(X) \equiv 0 \bmod p$ and $M \not\equiv 0 \bmod p$;
there are p solutions t if $F'(X) \equiv 0 \bmod p$ and $M \equiv 0 \bmod p$.

Using these results we have at hand a definite method when $s > 1$ of discovering every possible solution of $F(x) \equiv 0 \bmod p^s$ if we have previously found every solution of $F(x) \equiv 0 \bmod p^{s-1}$, so this completes the proof of **G.12**.

20.4. Modulo p, a prime, only congruences of degree less than p need be considered.

By repeated application of **G.12**, we see that solving $F(x) \equiv 0 \bmod p^s$ reduces to solving $F(x) \equiv 0 \bmod p$.

*However, the phrase "included among" must be interpreted carefully; there may be more solutions mod p^s than mod p^{s-1}, because integers congruent mod p^{s-1} may be incongruent mod p^s.

Next by using Fermat's theorem we are able to make a significant reduction in the number of congruences that need be considered. Whereas it was obvious from the start that the coefficients of the congruence are limited by the number of residue classes, it will now appear that for a prime modulus the *degree* of the congruence can also be limited.

G.13: If p is a prime, $F(x) \equiv 0 \bmod p$, may be replaced by a congruence of degree less than p.

Proof: By the division algorithm for polynomials we may write

$$F(x) = A(x)(x^p - x) + R(x)$$

where the degree of $R(x)$ is less than p. Since Fermat's theorem **G.7** shows $x^p - x \equiv 0 \bmod p$ for every integer x, it follows that $F(x) \equiv R(x) \bmod p$ for every integer x. Hence the solutions of $F(x) \equiv 0 \bmod p$ and $R(x) \equiv 0 \bmod p$ are exactly the same.

Since the leading coefficient a of $R(x)$ mod p satisfies $a \not\equiv 0 \bmod p$, we have $(a,p) = 1$, so by **G.8** there exists an integer b so that $ab \equiv 1$ mod p. Then $R(x)$ may be replaced by $bR(x)$ with a leading coefficient 1, and $R(x) \equiv 0 \bmod p$ and $bR(x) \equiv 0 \bmod p$ have the same solutions. Having agreed to make the leading coefficient 1, we cannot further specify the coefficients that may appear in $R(x)$ and each may be chosen in p ways. The degree of $R(x)$ may vary from 1 to $p - 1$, and, combining this fact with the previous observation about the coefficients, we find that there are a total of

$$p + p^2 + \dots + p^{p-1} = p(p^{p-1} - 1)/(p - 1)$$

congruences mod p that need be considered. Any other congruence mod p may be reduced to one of these.

For example, if $p = 3$, there are just 12 congruences that need be considered corresponding to the following $R(x)$ of degrees 1 and 2 and with leading coefficient 1:

$$x, \quad x + 1, \quad x + 2, \quad x^2, \quad x^2 + 1, \quad x^2 + 2, \quad x^2 + x,$$
$$x^2 + x + 1, \quad x^2 + x + 2, \quad x^2 + 2x, \quad x^2 + 2x + 1, \quad x^2 + 2x + 2.$$

Any other congruence mod 3 is reducible to one of the forms $R(x) \equiv 0$ mod 3. For example, $2x^4 + x^3 + x + 7 \equiv 0 \bmod 3$, reduces by **G.13** to $2x^2 + 2x + 1 \equiv 0 \bmod 3$, and if multiplied by 2 reduces to $x^2 + x + 2 \equiv 0 \bmod 3$, corresponding to one of the "standard" congruences mod 3 listed above.

20.5. Lagrange's theorem. The anomaly that a congruence of degree n may have more solutions than its degree can appear only when the modulus m is composite, for the following theorem due to Lagrange shows that the ordinary rule of the theory of equations for complex numbers holds for the theory of congruences when the modulus is a prime.

G.14: If p is a prime, the number of incongruent solutions of $F(x) \equiv 0 \bmod p$ is never more than the degree of the congruence.

Proof: The proof is by induction on the degree n of the congruences.

(I) When $n = 1$, the congruences take the form $ax \equiv b \bmod p$ with $a \not\equiv 0 \bmod p$, and by **G.8** such a congruence has just one solution, for we have here, since p is a prime, that $(a,p) = 1$.

(II) Suppose the theorem has been established for all congruences of degree $<n + 1$. Consider a congruence $F(x) \equiv 0 \bmod p$ of degree $n + 1$; and suppose, if such a thing be possible, that the congruence has $n + 2$ incongruent solutions. Let r be one of these solutions on which we fix attention and let s be any one of the other $n + 1$ solutions. By the division algorithm we may write

$$F(x) = (x - r)Q(x) + R,$$

where R is an integer and $Q(x)$ is of degree n and has integers as coefficients. Since by hypothesis $F(r) \equiv 0 \bmod p$, it follows by substitution that $R \equiv 0 \bmod p$. Also by hypothesis $F(s) \equiv 0 \bmod p$. Combining these observations we see that $(s - r)Q(s) \equiv 0 \bmod p$. Since s and r are incongruent mod p and since p is a prime, we have $(s - r,p) = 1$, hence we may invoke **G.6** to assert that $Q(s) \equiv 0 \bmod p$. (It is exactly at this point that the argument will break down if the modulus m is composite, for then it is possible to have $s - r \not\equiv 0 \bmod m$, $(s - r)Q(s) \equiv 0 \bmod m$, without forcing $Q(s) \equiv 0 \bmod m$.) But $Q(x)$ is of degree n and s is *any one* of $n + 1$ incongruent residues for each of which we have just proved that $Q(s) \equiv 0 \bmod p$. This is a contradiction of the induction hypothesis. Hence a congruence of degree $n + 1$ must have at most $n + 1$ incongruent solutions mod p.

By (I), (II), and the principle of mathematical induction the proof of **G.14** is complete.

20.6. An example. To illustrate the preceding theorems **G.11**, **G.12**, **G.13**, **G.14**, we propose to find the complete solution of

$$F(x) = x^7 - 14x - 2 \equiv 0 \bmod 1323.$$

Here we have $m = 1323 = 3^3 7^2$, so we begin as **G.11** suggests and consider separately

$$(1) \quad F(x) \equiv 0 \bmod 49; \qquad (2) \quad F(x) \equiv 0 \bmod 27.$$

To solve (*1*) we begin as **G.12** suggests and consider $F(x) \equiv 0 \bmod 7$; but to this problem we may apply the reductions suggested in **G.13**, such as $x^7 \equiv x$, $14 \equiv 0$, mod 7, to find that the congruence reduces to $x - 2 \equiv 0 \bmod 7$ with the unique solution $X \equiv 2 \bmod 7$.

Now we are ready to apply **G.12** and (*20.4*) so we compute $M = F(2)/7 = (128 - 28 - 2)/7 = 14$ and $F'(x) = 7x^6 - 14$. Inasmuch as $M \equiv 0 \bmod 7$ and $F'(2) \equiv 0 \bmod 7$, there are 7 suitable values for t solving $M + tF'(X) \equiv 0 \bmod 7$; hence from $x = X + 7t$, we find $x = 2, 9, 16, 23, 30, 37, 44$—the complete solution of $F(x) \equiv 0 \bmod 49$.

To solve (2) we begin by considering $F(x) \equiv 0 \bmod 3$, a congruence which reduces readily to the form $2x \equiv 2 \bmod 3$ with the unique solution $X \equiv 1 \bmod 3$.

Then $M = F(1)/3 = -5 \equiv 1 \bmod 3$ and $F'(1) = -7 \equiv 2 \bmod 3$, so that (*20.4*) becomes $1 + 2t \equiv 0 \bmod 3$ with just one solution, $t \equiv 1 \bmod 3$. Hence there is just one solution, $x = X + 3t = 4$ of $F(x) \equiv 0 \bmod 9$.

We must apply **G.12** and (*20.4*) once more, now with $X = 4$. Then

$$M = F(4)/9 = 16326/9 = 1814 \equiv 2 \bmod 3,$$
$$F'(4) = 7(4)^6 - 14 \equiv 1 + 1 \equiv 2 \bmod 3,$$

so that (*20.4*) becomes $2 + 2t \equiv 0 \bmod 3$ with just one solution $t \equiv 2 \bmod 3$. Hence there is just one solution $x = X + 9t = 4 + 9(2) = 22$, of $F(x) \equiv 0 \bmod 27$.

To finish the problem we need to apply **G.11** which means we must solve *several* problems of the form

$$x \equiv a \bmod 27, \quad x \equiv b \bmod 49.$$

For this purpose we use **G.8** to solve

$$49x_1 \equiv 1 \bmod 27, \quad \text{for } x_1 \equiv 16 \bmod 27;$$
$$27x_2 \equiv 1 \bmod 49, \quad \text{for } x_2 \equiv 20 \bmod 49;$$

and then we apply **G.9** to write the solution

$$x \equiv (49)(16)a + (27)(20)b \bmod 1323$$

of the given pair of congruences.

In the present case with $a = 22$ and a variety of values for b we find that $x \equiv 49 + 540\, b \bmod 1323$.

As we give b, in turn, the values 2,9,16,23,30,37,44, we find

$$x \equiv 1129,\ 940,\ 751,\ 562,\ 373,\ 184,\ -5, \bmod 1323.$$

These seven solutions represent the complete solution of

$$x^7 - 14x - 2 \equiv 0 \bmod 1323.$$

20.7. Wilson's theorem. We are now in a position to present one of the complete, but impractical, tests, mentioned in **6.2**, for deciding whether a given integer n is a prime.

Wilson's theorem: A necessary and sufficient condition that n be a prime is that $(n - 1)! \equiv -1 \bmod n$.

Proof: (*A*) If p is a prime, then by Euler's theorem **G.7** there are $p - 1$ solutions $x = 1,2,\ldots,p - 1$ of the congruence

$$G(x) = x^{p-1} - 1 \equiv 0 \bmod p.$$

On the other hand the congruence

$$H(x) = (x - 1)(x - 2)\ldots(x - (p - 1)) \equiv 0 \bmod p$$

also has $p - 1$ solutions: $x = 1,2,\ldots,p - 1$. Both $G(x)$ and $H(x)$ are of degree $p - 1$ and they have the same leading term x^{p-1}. It follows that $F(x) = G(x) - H(x) \equiv 0 \bmod p$ is a congruence of degree at most $p - 2$ having $p - 1$ incongruent solutions. But this is a contradiction of Lagrange's theorem **G.14**, unless every coefficient of $F(x)$ is a multiple of p (so that $F(x)$ is not of degree $\geqq 1$, mod p); but, in this latter circumstance, $F(x) \equiv 0 \bmod p$ is also satisfied by $x = 0$. Hence we find

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

If p is odd, $p - 1$ is even, so $(-1)^{p-1} \equiv +1 \bmod p$. If p is even, then $p = 2$, and $(-1)^{p-1} = -1 \equiv +1 \bmod 2$. Thus for *every* prime p we find

$$(p - 1)! \equiv -1 \bmod p.$$

(B) Conversely, if n is composite, then n has at least one divisor d, with $1 < d < n$, so that d divides $(n - 1)!$ and $(n - 1)! \equiv 0 \bmod d$. It is therefore impossible that

$$(n - 1)! \equiv -1 \bmod n$$

for this latter congruence would imply $(n - 1)! \equiv -1 \bmod d$, a patent contradiction.

EXERCISES

EX. *20.1.* (a) By substitution from absolutely least residue systems find all solutions of

$$x^3 + 3x^2 + 31x + 23 \equiv 0$$

mod 5 and mod 7.

(b) Using the results of (a) and **G.11**, find all solutions of the given congruence mod 35.

(c) Discuss the numbers of solutions mod 5, mod 7, and mod 35 as illustrations of **G.14**.

EX. *20.2.* (a) solve

$$x^3 + 3x^2 + x + 3 \equiv 0 \text{ mod } 5 \qquad \text{(two solutions)}.$$

(b) Apply **G.12** and solve the same congruence mod 25 (six solutions).

(c) Apply **G.12** again and solve the same congruence mod 125 (eleven solutions).

EX. *20.3.* Solve

$$x^3 + 64x^2 + x + 30 \equiv \text{mod } 216.$$

EX. *20.4.* Solve $x^3 \equiv 13 \text{ mod } 490$.

EX. *20.5.* Show that

$$x^3 + x + 3 \equiv (x - 1)^2x \text{ mod } 2;$$
$$x^3 + x + 3 \equiv (x - 2)^2(x + 4) \text{ mod } 13;$$

and then solve $x^3 + x + 3 \equiv 0 \text{ mod } 676$.

> ►*Many of the greatest masters were first attracted to mathematical inquiry by problems relating to numbers, and no one can glance at the periodicals of the present day which contain questions for solution without noticing how singular a charm such problems continue to exert.* —J. W. L. GLAISHER

CHAPTER 21*

EXPONENTS, PRIMITIVE ROOTS, AND INDICES

21.1. The exponent of *a* modulo *m*. The general object of this lesson is to pursue further the implications of Euler's theorem, with our results culminating, in case the modulus is a prime, in a remarkable analogue of the theory of logarithms. The new theory ties in with the preceding chapters in that it enables us to find in a new way the solutions (and, first of all, to decide whether there are solutions) of congruences of the type $x^n \equiv b \bmod p$, sometimes called "pure" congruences.

Since Euler's theorem shows that $a^{\phi(m)} \equiv 1 \bmod m$ whenever $(a,m) = 1$, it follows that for such an a and m there must exist a *least positive* exponent e such that $a^e \equiv 1 \bmod m$. We shall describe this least exponent by saying "e is the exponent of a modulo m" or that "a belongs to e modulo m." It is important to note that the definition concerns only integers a satisfying $(a,m) = 1$.

For example, with $a = 3$ and $m = 13$, we investigate the powers of a and find $3 \equiv 3$, $3^2 \equiv 9$, $3^3 \equiv 27 \equiv 1 \bmod 13$, so we say that "3 belongs to 3 mod 13." Without such an investigation we might have applied Euler's theorem to assert correctly that $3^{12} \equiv 1 \bmod 13$; but,

*Chapter 21 is a basic chapter, except for **21.4** which is a supplementary section.

as we have just seen, it would have been wrong to conclude from this that 12 is the *least* positive exponent which will serve our purpose.

G.15: If a belongs to e mod m, and if $a^k \equiv 1$ mod m, then e divides k.

Proof: Since e is a minimal positive exponent such that $a^e \equiv 1$ mod m, it follows that $k \geqq e$. Suppose $k = qe + r$, with $0 \leqq r < e$. Then

$$a^r \equiv (a^e)^q a^r \equiv a^{qe+r} \equiv a^k \equiv 1 \text{ mod } m$$

But this is a contradiction of the minimal property of e, unless $r = 0$; but if $r = 0$, then $k = qe$, and e divides k.

G.15.1: If a belongs to e mod m, then e divides $\phi(m)$.

Proof: By Euler's theorem $a^{\phi(m)} \equiv 1$ mod m, hence by **G.15** we find that e must divide $\phi(m)$.

G.15.2: If $a^s \equiv a^t$ mod m, then $s \equiv t$ mod e.

Proof: It is no restriction to assume $s \geqq t$. Then $(a,m) = 1$ implies $(a^t,m) = 1$, so that we may apply the cancellation law **G.6** to the given congruence $a^s \equiv a^t$ mod m to conclude that $a^{s-t} \equiv 1$ mod m. Hence by **G.15** it follows that e divides $s - t$, but this is equivalent to writing $s \equiv t$ mod e.

G.16: If p is a prime and if d is a positive divisor of $p - 1$, then $x^d \equiv 1$ mod p has d distinct solutions.

Proof: Since d divides $p - 1$ we can write $p - 1 = kd$ and $x^{p-1} - 1 = (x^d - 1)Q(x)$ where $Q(x) = x^{(k-1)d} + x^{(k-2)d} + \ldots + x^d + 1$ has integral coefficients and is of degree $p - 1 - d$. Let D be the number of distinct solutions of $x^d - 1 \equiv 0$ mod p. By Euler's theorem, since p is a prime, there are $p - 1$ distinct solutions of $x^{p-1} - 1 \equiv 0$ mod p. Every solution r of $x^{p-1} - 1 \equiv 0$ mod p that is not a solution of $x^d - 1 \equiv 0$ mod p must be a solution of $Q(x) \equiv 0$ mod p, because

$$0 \equiv r^{p-1} - 1 \equiv (r^d - 1)Q(r) \text{ mod } p$$

with $r^d - 1 \not\equiv 0$ mod p, implies, by virtue of **G.6**, that $Q(r) \equiv 0$ mod p. Hence $Q(x) \equiv 0$ mod p must have $p - 1 - D$ solutions. However, by Lagrange's theorem, **G.14**, we know with regard to $x^d - 1 \equiv 0$ mod p that $D \leqq d$ and with regard to $Q(x) \equiv 0$ mod p that $p - 1 - D \leqq p - 1 - d$, since the number of solutions of a congruence with a prime modulus is at most equal to its degree. But the second of these inequalities is equivalent to $d \leqq D$, and when coupled with the first inequality, this shows that $D = d$.

G.17: If p is a prime and if e is a positive divisor of $p - 1$, then the number of residue classes belonging to e modulo p is given by $\phi(e)$.

Proof: Let the divisors of $p - 1$ be arranged in order:

$$1 = d_1 < d_2 < \ldots < d_k < d_{k+1} < \ldots < d_{\tau(p-1)} = p - 1$$

where $\tau(n)$ is the number-theoretic function described in **8.1.** The proof will be by "limited induction" on k, i.e., an induction type of proof limited to the integers k for which $1 \leqq k \leqq \tau(p - 1)$.

(I) The theorem is true for $k = 1$, because $d_1 = 1$, $\phi(1) = 1$, and only the 1-class belongs to 1 mod p.

(II) Let us assume that the theorem is correct for $d_1,d_2,\ldots,d_k$, where k is limited to the range $1 \leqq k < \tau(p - 1)$, and let us consider the next case involving d_{k+1}. We shall divide the argument into several steps:

(A) By **G.16** there are exactly d_{k+1} solutions of the congruence $x^{d_{k+1}} \equiv 1 \bmod p$.

(B) Every *proper* divisor $d_1',d_2',\ldots,d_{\tau(d_{k+1})-1}'$ of d_{k+1} is a divisor of $p - 1$ less than d_{k+1}, and hence is included in the list $d_1,d_2,\ldots,d_k$ to which the hypothesis of induction applies, hence there are $\phi(d_i')$ residue classes belonging to d_i' mod p; since d_i' is a divisor of d_{k+1} it follows that every one of the $\phi(d_i')$ residue classes belonging to d_i' mod p is a solution of $x^{d_{k+1}} \equiv 1 \bmod p$.

(C) From steps (A) and (B) it follows that the number of residue classes belonging to d_{k+1} mod p, *not* to some *proper* divisor of d_{k+1}, is given by s where

$$s = d_{k+1} - \{\phi(d_1') + \phi(d_2') + \ldots + \phi(d_{\tau(d_{k+1})-1}')\}.$$

(D) From the theorem in **16.3** we know that

$$d_{k+1} = \phi(d_1') + \phi(d_2') + \ldots + \phi(d_{\tau(d_{k+1})-1}') + \phi(d_{\tau(d_{k+1})}')$$

where $d_{\tau(d_{k+1})}' = d_{k+1}$.

(E) Substituting from (D) into (C) we find $s = \phi(d_{k+1})$.

Hence if the theorem is true for $d_1,d_2,\ldots,d_k$ where $1 \leqq k < \tau(p - 1)$, then the theorem is true for d_{k+1}.

Then from (I), (II), and the principle of mathematical induction it follows that the theorem is true for all the $\tau(p - 1)$ divisors of $p - 1$, which completes the proof.

An example illustrating the theorem is given at the close of the next section.

21.2. Primitive roots. By definition, if p is a prime and if a belongs to $p - 1$ mod p, then a is called a *primitive root* mod p. The terminology results, of course, from comparing the congruence $x^{p-1} \equiv 1$ mod p with the equation $x^{p-1} = 1$ over the complex number system, for a root of the latter equation which is not a root of $x^d = 1$ for $1 \leq d < p - 1$ has long been called a primitive $(p - 1)$ root of unity. The object of the next corollary to **G.17** is to show the *existence* of primitive roots mod p.

G.17.1: For every prime p there are $\phi(p - 1)$ primitive roots.

Proof: Since p is a prime and $p - 1$ is a divisor of $p - 1$, it follows from **G.17** that there are $\phi(p - 1)$ residue classes belonging to $p - 1$ modulo p and, according to our definition, each of these classes is a primitive root mod p.

The important feature of **G.17.1** is that it guarantees for every prime the existence of at least one primitive root. The unfortunate feature of the proof is that it is an existence proof, not a constructive proof, and there seems to be no really simple way of finding a primitive root for large values of p. For small values of p a primitive root may be found by trial, and once it has been found, it can be used, together with EX. *21.1*, to determine rather rapidly to what exponent each residue class of p belongs. For if a is a primitive root mod p, then EX. *21.1* shows that a^s belongs to $(p - 1)/(p - 1,s)$.

For example, when $p = 13$, we find by trial that 2 is a primitive root. The table, mod 13, is as follows:

s:	1	2	3	4	5	6	7	8	9	10	11	12
2^s:	2	4	8	3	6	12	11	9	5	10	7	1

We can find $\phi(12) = 4$ values of $s = 1,5,7,11$ such that $(s,12) = 1$; hence there are $\phi(12)$ primitive roots mod 13, namely, the corresponding values of $2^s \equiv 2, 6, 11, 7$, respectively. By choosing $s = 2, 10$ so that $(s,12) = 2$, we find $\phi(6) = 2$ classes belonging to 6 mod 13, namely, $2^s \equiv 4, 10$, respectively. Choosing $s = 3,9$ so that $(s,12) = 3$, we find $\phi(4) = 2$ classes belonging to 4 mod 13, namely, $2^s \equiv 8,5$, respectively. With $s = 4,8$ so that $(s,12) = 4$, we find $\phi(3) = 2$ classes belonging to 3 mod 13, namely, $2^s \equiv 3,9$, respectively. With

$s = 6$, so that $(s,12) = 6$, we find $\phi(2) = 1$ class belonging to 2 mod 13, namely, $2^s \equiv 12$. With $s = 12$, so that $(s,12) = 12$, we find $\phi(1) = 1$ class belonging to 1 mod 13, namely, $2^s \equiv 1$.

21.3. The theory of indices. If p is a prime and if a is a primitive root mod p, then the powers

$$a, a^2, a^3, \ldots, a^{p-2}, a^{p-1} \equiv 1 \bmod p$$

are $p - 1$ in number and are incongruent in pairs, for otherwise a contradiction of **G.15.2** and the fact that a is a primitive root would appear. Hence it follows that these powers represent, in some order, the non-zero residue classes mod p. In other words, if $b \not\equiv 0 \bmod p$, there exists an integer x such that $a^x \equiv b \bmod p$. We now agree, for convenience of reference, to give this x a new name; we shall write $x = \text{ind}_a\, b$, to be read "x is the index of b to the base a mod p"

In the very definition of the index the reader will no doubt recognize the close analogy with the usual definition in analysis of the logarithm of b to the base a; and in studying the following rules of indices, the reader can easily anticipate the results by thinking of the usual rules of logarithms. Just as in the study of logarithms where the base a is usually kept constant in a given discussion, so that the a is not written in the logarithms, so here, too, we will agree to dispense with the subscript a on each index, simply adopting the understanding that in a given problem or in a given set of rules it is a fixed primitive root a which is being used as the base of the system of indices; but there is a further agreement here, not of concern in logarithms, that the modulus p is also being held constant.

G.18: Rules of indices with the base a modulo p:

G.18.1: If $b \equiv c \not\equiv 0 \bmod p$, then $\text{ind}\, b \equiv \text{ind}\, c \bmod p - 1$; and conversely.

G.18.2: If $d \equiv bc \not\equiv 0 \bmod p$, then

$$\text{ind}\, d \equiv \text{ind}\, b + \text{ind}\, c \bmod p - 1;$$

and conversely.

G.18.3: If $d \equiv b^k \not\equiv 0 \bmod p$, then $\text{ind}\, d \equiv k\, \text{ind}\, b \bmod p - 1$; and conversely.

Proof: Every one of these rules is a direct consequence of the usual rules of exponents and of **G.15.2** with $e = p - 1$, and with

$s = \text{ind } b$, $t = \text{ind } c$, $u = \text{ind } d$. For example, in **G.18.1** by hypothesis and definition we have

$$a^{\text{ind } b} \equiv b \equiv c \equiv a^{\text{ind } c} \bmod p;$$

then since a belongs to $p - 1$ we apply **G.15.2** to conclude that

$$\text{ind } b \equiv \text{ind } c \bmod p - 1.$$

Conversely, from ind $b \equiv$ ind c mod $p - 1$ we may write ind $b =$ ind $c + K(p - 1)$ where K is an integer. Then

$$b \equiv a^{\text{ind } b} \equiv a^{\text{ind } c + K(p-1)} \equiv a^{\text{ind } c}(a^{p-1})^K \equiv a^{\text{ind } c} \equiv c \bmod p.$$

The details in proving **G.18.2** and **G.18.3** will be left as exercises for the reader.

In the following example with $p = 29$ we use the primitive root $a = 2$ and construct a complete table of indices, comparable to the usual table of logarithms. However, with logarithms it is not thought necessary usually to give a companion table of anti-logarithms, because if the numbers for which logarithms are given are arranged in increasing order, then the logarithms themselves automatically appear in increasing order. But when the non-zero residue classes mod p for which indices are given are arranged in the natural order, then the indices do not appear, in general, in the natural order; hence a separate table of anti-indices is a very great convenience. In fact in constructing such tables, it is the latter table which it is most natural to form at the outset, so we give it first in the following example:

ind b	1	2	3	4	5	6	7	8	9	10	11	12	13	14	ANTI-INDICES Given ind b, mod 28; to find b, mod 29.
b	2	4	8	16	3	6	12	24	19	9	18	7	14	28	
ind b	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
b	27	25	21	13	26	23	17	5	10	20	11	22	15	1	
b	1	2	3	4	5	6	7	8	9	10	11	12	13	14	INDICES Given b, mod 29; to find ind b, mod 28.
ind b	28	1	5	2	22	6	12	3	10	23	25	7	18	13	
b	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
ind b	27	4	21	11	9	24	17	26	20	8	16	19	15	14	

As an example of the use of the tables of indices and the rules of indices, let us solve $21x \equiv 36 \bmod 29$.

First we note that $36 \equiv 7 \bmod 29$ and then using **G.18.1** and **G.18.2**, we have ind $21 +$ ind $x \equiv$ ind 7 mod 28. From the table of indices we have $17 +$ ind $x \equiv 12 \bmod 28$, whence ind $x \equiv -5 \equiv 23$

mod 28. Finally, by the table of anti-indices and by an application of the converse part of **G.18.1**, we find the unique solution $x \equiv 10$ mod 29.

As another example, let us solve $x^{36} \equiv 36 \bmod 29$.

By Euler's theorem $x^{28} \equiv 1 \bmod 29$, and $36 \equiv 7 \bmod 29$, so the problem reduces to the simpler form $x^8 \equiv 7 \bmod 29$. By **G.18.1** and **G.18.3** we find that an equivalent problem is $8 \text{ ind } x \equiv \text{ind } 7 \bmod 28$; from the tables this last congruence may be written $8 \text{ ind } x \equiv 12$ mod 28. Since $(8,28) = 4$ is a divisor of **12**, this congruence may be solved as in **G.8.1**. First we consider $2 \text{ ind } x \equiv 3 \bmod 7$ and multiplying by 4, we discover the solution $\text{ind } x \equiv 12 \equiv 5 \bmod 7$. Therefore $\text{ind } x \equiv 5, 12, 19, 26 \bmod 28$ are the only possibilities. By the table of anti-indices and **G.18.1** it follows that $x \equiv 3, 7, 26, 22 \bmod 29$ are the respective solutions of the given problem and form the complete set of solutions.

Using the same attack, we find the congruence $x^8 \equiv 8 \bmod 29$ has *no* solution; for the equivalent linear congruence $8 \text{ ind } x \equiv \text{ind } 8 \equiv 3$ mod 28 has no solution by **G.8.1** inasmuch as $(8,28) = 4$ does not divide 3.

For extensive problem-solving of this type it may be useful to know that tables of indices and anti-indices for all primes <100 are given in the Uspensky and Heaslet text, listed in **1.3**.

21.4. A slide rule for problems mod 29. Since the theory behind the ordinary slide rule is the theory of logarithms, it is reasonably clear that with the aid of the theory of indices we may construct a slide rule for the solution of all problems of the type suggested by **G.18**. As an example we shall show here how to construct a circular slide rule to be used in solving problems mod 29.

By way of preliminary discussion we need to digress for a moment and explain the possibility of defining, for real numbers, congruence modulo m where m is a fixed real number. For real numbers a and b, we shall define $a \equiv b \bmod m$ if and only if $a - b = Km$ where K is an *integer*. This notion is an equivalence relation dividing all the real numbers into mutually exclusive classes of congruent numbers.

For example, a very useful device in some problems is the notion of congruence mod 1; of course, in strict number theory this concept may not be of much use, because all the integers fall into one class, say the 0-class, mod 1; but for all fractions, say, or for all real numbers, the concept is useful, every real number a being congruent mod 1 to one and only one real number b in the interval $0 \leqq b < 1$.

In particular, here we want to use the notation $\theta_1 \equiv \theta_2 \bmod 2\pi$ as a convenient way of saying $\theta_1 = \theta_2 + 2\pi K$ where K is an integer. For if θ_1 and θ_2 are central angles measured in radians with the same initial sides, then their terminal sides will be coincident, inasmuch as 2π radians is one revolution. In other words, to write $\theta_1 \equiv \theta_2 \bmod 2\pi$ is equivalent to the usual "equals relation" for angles.

In making a circular slide rule mod 29 we shall use five concentric circular scales, each of which we shall describe in terms of polar coordinates, the radius vector of each scale being constant, while the polar angle is in each case a function of an integral parameter, with all five functions involving the same constant $k = 2\pi/28$. The exact description of the scales is as follows:

A-scale: $r = r_1, \quad \theta = k\,A;$

c-scale: $r = r_2, \quad \theta = k \text{ ind } c;$

d-scale: $r = r_3, \quad \theta = k \text{ ind } d;$

R-scale: $r = r_4, \quad \theta = -k \text{ ind } R;$

Q-scale: $r = r_5, \quad \theta = k \text{ (ind } Q)/2$, if ind Q is *even.*

We shall make a rule* in which $r_1 < r_2 < r_3 < r_4 < r_5$ and in which the A- and c-scales are constructed upon one circular disk, while the other scales are constructed upon another sheet, the disk being pinned to the sheet, and free to rotate, at the common center of the scales. The first four scales have the parameters A, c, d, R running from 1 to 28; and as usual in constructing a slide rule, we use the formula to locate the correct θ-position corresponding to a given value of the parameter, but we label that position *not* with the value of θ, but with the value of the parameter.

We know from **G.18.1** that if $C \equiv c \bmod 29$ then ind $C \equiv$ ind c mod 28, or ind $C =$ ind $c + 28K$ where K is an integer. Hence we discover the relation

$$k \text{ ind } C = k \text{ ind } c + k28K = k \text{ ind } c + 2\pi K$$

so that with reference to the c-scale we have

$$\theta(C) \equiv \theta(c) \bmod 2\pi.$$

But this is exactly the type of relation discussed earlier and shows that $\theta(C)$ and $\theta(c)$ are "equal" angles. Hence this type of slide rule

*See the tailpiece to Chapter 21 and construct a working model from the sheet facing page 154.

automatically takes care of our need of staying within the same residue classes, mod 29 (or mod 28 in case of the A-scale), for the same position on our circular scales. It will be unnecessary, therefore, to add any labels, different from those of the non-zero residue classes mod 29 already marked.

If the disk carrying the A- and c-scales is rotated so that c-ray marked 1 falls upon the d-ray marked x, then if the d-ray marked z falls upon the c-ray marked y, it will follow that $xy \equiv z$ mod 29. Hence if any two of the three quantities x,y,z are given, the third can be found.

The proof resides in the fact that in the rotated position of the central disk we have in terms of angles:

$$\theta(z) \equiv \theta(x) + \theta(y) \bmod 2\pi;$$

but in terms of the c- and d-scales this congruence implies

$$k \text{ ind } z = k \text{ ind } x + k \text{ ind } y + 2\pi K$$

where K is an integer. Multiplying by $1/k = 28/2\pi$, we arrive at the relation

$$\text{ind } z = \text{ind } x + \text{ind } y + 28K;$$

thus ind $z \equiv$ ind x + ind y mod 28 and by **G.18.2** we know that this implies $xy \equiv z$ mod 29.

A ray extending from the A-scale to the c-scale obviously solves $A =$ ind c, so here in graphic form is a table of anti-indices, and with just a bit of looking (because of the c's not appearing in the natural order) it may also be considered a table of indices.

In the d- and R-scales is provided a direct solution of the congruence $dR \equiv 1$ mod 29, obtained by merely extending the ray from d on the d-scale to R on the R-scale.

The proof is simple since the proposed construction gives $\theta(d) \equiv \theta(R)$ mod 2π, or k ind $d \equiv -k$ ind R mod 2π, whence

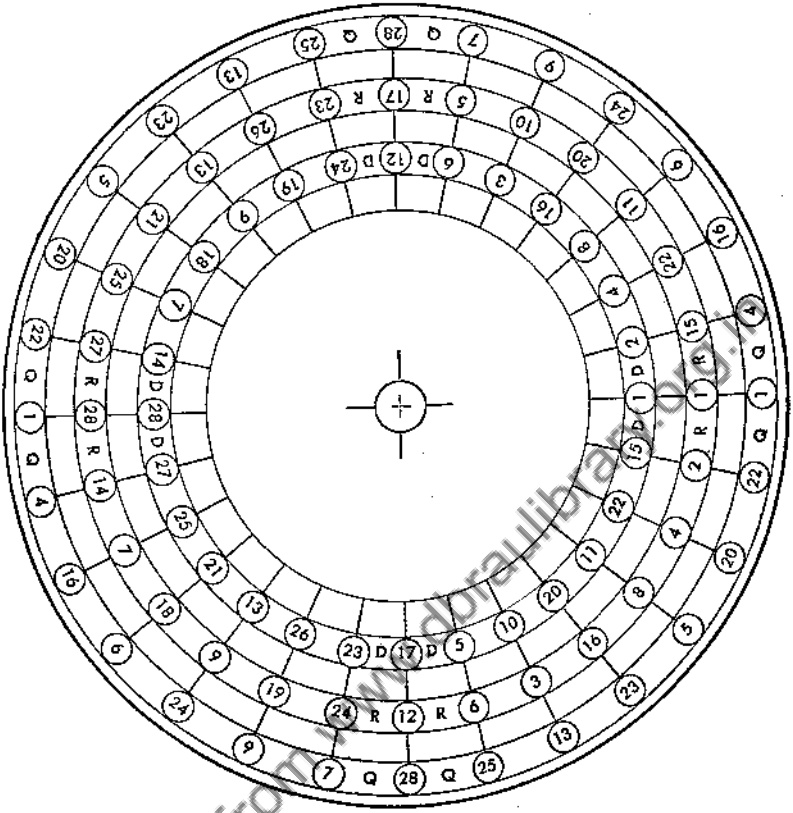
$$\text{ind } d + \text{ind } R \equiv 0 \equiv 28 \equiv \text{ind } 1 \bmod 28;$$

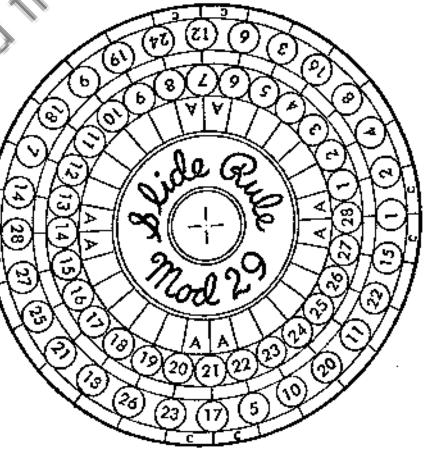
but this last congruence by **G.18.2** implies $dR \equiv 1$ mod 29.

Similarly, the d- and Q-scales provide a direct solution of the problem $d^2 \equiv Q$ mod 29 in the *fourteen* cases where there are solutions. For from **G.18.3** we know that the given congruence is equivalent to

$$2 \text{ ind } d \equiv \text{ind } Q \bmod 28;$$

but by **G.8.1** we know that there are solutions, in fact just *two* solutions, of this latter congruence if and only if ind Q is *even*. Since this last restriction is exactly that placed on the function defining the Q-scale, the connection is fairly obvious. However, it is to be





Cut along this line

If desired, a working model of the slide rule may be constructed by mounting the above components on cardboard and combining the separate parts on a common axis.

noted that $Q \equiv q$ mod 29 implies ind $Q \equiv$ ind q mod 28 and $\theta(Q) \equiv \theta(q)$ mod π, *not* 2π. Hence for a given value of Q there are found two entries on the Q-scale differing by π. If then a ray is drawn from either position of Q on the Q-scale to d_1 and d_2, respectively, on the d-scale there will be found the two solutions of $d^2 \equiv Q$ mod 29. For the construction gives $\theta(d) \equiv \theta(Q)$ mod π,

$$k \text{ ind } d \equiv k\,(\text{ind}\,Q)/2 \bmod \pi, \text{ or } 2k \text{ ind } d \equiv k \text{ ind}\,Q \bmod 2\pi,$$

$$2 \text{ ind } d \equiv \text{ind } Q \bmod 28, \text{ or } d^2 \equiv Q \bmod 29.$$

The fact that $d^2 \equiv Q$ mod 29 has solutions for fourteen values of Q of even index and fails to have a solution for fourteen values of Q of odd index, will provide us with a good introduction to the next chapter.

EXERCISES

EX. *21.1.* If $(a,m) = 1$ and a belongs to e mod m, prove that a^s belongs to E where $d = (e,s)$ and $e = Ed$.

EX. *21.2.* Prove that 1,5,7,11 all belong to 1 or 2 mod 12 and that there are no residue classes belonging to $\phi(12)$ mod 12.

EX. *21.3.* If there is one class a belonging to $\phi(m)$ mod m, prove that there are $\phi(\phi(m))$ classes belonging to $\phi(m)$ mod m.

EX. *21.4.* Find 8 residue classes belonging to 20 mod 25.

EX. *21.5.* Prove that 2 is *not* a primitive root mod 17.

EX. *21.6.* Prove that 3 *is* a primitive root mod 17 and then use EX. *21.1* to find *all* the primitive roots mod 17.

EX. *21.7.* Find a primitive root mod 19 and then use EX. *21.1* to classify all the residue classes mod 19 according to the exponents to which they belong.

EX. *21.8.* If a belongs to e mod m, prove that e divides $\lambda(m)$ as defined in EX. *18.7*.

EX. *21.9.* If p is an odd prime and a is a primitive root mod p, prove that $p - a$ is a primitive root if and only if $p \equiv 1$ mod 4.

EX. *21.10.* Complete the proof of **G.18.2**.

EX. *21.11.* Complete the proof of **G.18.3**.

EX. *21.12.* Use the tables given in **21.3** to solve the following congruences mod 29:

$$x \equiv (12)(18), \quad x \equiv (21)(25), \quad 3x \equiv 7, \quad 17x \equiv 23, \quad 8x \equiv 1,$$
$$x \equiv (15)^{10}, \quad x \equiv (33)^{33} \equiv 4^5, \quad x^2 \equiv 16, \quad x^2 \equiv 18,$$
$$x^3 \equiv 7, \quad x^3 \equiv 18, \quad x^5 \equiv 22, \quad x^6 \equiv 5, \quad x^6 \equiv 6.$$

EX. *21.13.* Solve the above congruences mod 29 by the use of the slide rule mod 29.

EX. *21.14.* Construct a table of indices mod 29 using the primitive root 3.

EX. *21.15.* If p is a prime, let $N(k,p,y)$ be the number of solutions x of the congruence $x^k \equiv y \bmod p$. If k is fixed, show that N has one of two values for various values of $y \not\equiv 0 \bmod p$ and find the number Q of y's leading to each of these values of N.

$$N_1 = (k,p-1),\ N_2 = 0;\ Q_1 = (p-1)/N_1,\ Q_2 = p - 1 - Q_1.$$

EX. *21.16.* Why do the diametrically opposite entries on the c-scale of the circular slide rule mod 29 add to 29?

EX. *21.17.* Construct a circular slide rule mod 31.

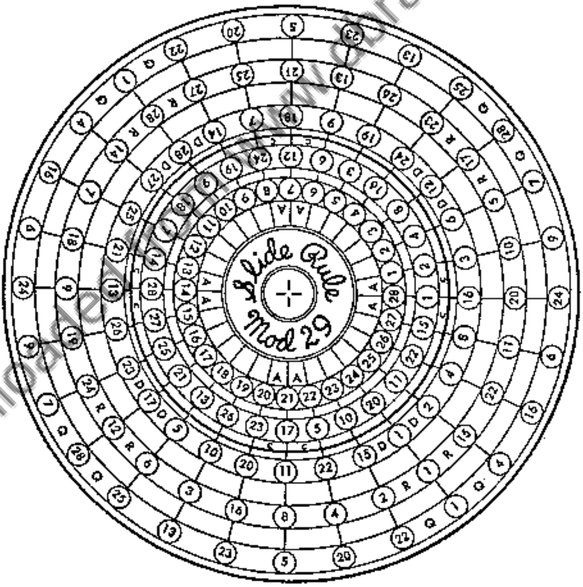
EX. *21.18.* Establish the following theorems:

(a) if $(a,m) = 1$ and $a^{m-1} \not\equiv 1 \bmod m$, then m is composite;

(b) if a belongs to $m - 1 \bmod m$, then m is prime;

(c) if $m - 1$ has just k distinct prime factors and $a^{m-1} \equiv 1 \bmod m$, then k tests will suffice to decide whether a belongs to $m - 1 \bmod m$.

EX. *21.19.* With $a = 2$ use the ideas of EX. *21.18* to test the primality of (a) 600,001; (b) 700,001.



An example of the use of the slide rule mod 29. Turned to this position, the rule solves $D \equiv 16C \bmod 29$ with D and C on the same ray.

> ▶*What Gauss put into print is as true and important today as when first published; his publications are statutes, superior to other statutes in this, that nowhere and never has a single error been detected in them.*
>
> —M. CANTOR

CHAPTER 22*

QUADRATIC RESIDUES AND LEGENDRE'S SYMBOL

22.1. Quadratic congruences. The purpose of this and the following chapter is to provide a complete test for the *existence* of solutions of the general quadratic congruence

$$ax^2 + bx + c \equiv 0 \bmod m, \qquad a \not\equiv 0 \bmod m.$$

In a sense the discussion of Chapter 20 completely solves this problem; however, the reader will recall that by the method of that chapter a given congruence problem mod m was reduced to a series of problems mod p, a prime, and that the solutions mod p were to be found, presumably, by trial of all the residue classes. This is feasible for small primes p, but impractical for large primes. It is that defect which we propose to remedy in the case of quadratic congruences, in the sense of providing another way of deciding whether solutions exist.

If in the general congruence displayed above we have $b \equiv 0 \bmod m$, then the congruence is called a pure quadratic congruence. The purpose of the next theorem is to show that for a prime modulus the general quadratic congruence is equivalent to a chain of two congruences: the first of which is a pure quadratic congruence, which may

*Chapter 22 is a basic chapter.

or may not have a solution, and the second of which is a linear congruence, which is always solvable provided that the first congruence of the chain has a solution. The formulas are remarkably like those of the quadratic formula, so familiar in the theory of equations for complex numbers.

Since the case $p = 2$ is trivially solved by trial, we shall limit ourselves in this and the next lesson to having p represent an odd prime.

G.19: If p is an odd prime and if $a \not\equiv 0 \bmod p$, then

$$ax^2 + bx + c \equiv 0 \bmod p$$

is equivalent to the chain of congruences which follows:

$$u^2 \equiv b^2 - 4ac, \quad 2ax \equiv u - b \bmod p.$$

Proof: By the hypothesis that p is an odd prime and that $a \not\equiv 0 \bmod p$, it follows that $(4a,p) = 1$; hence **G.6** may be employed to show that the given congruence is equivalent to the following congruence:

$$4a^2x^2 + 4abx + 4ac \equiv 0 \bmod p.$$

By subtracting $4ac$ and adding b^2 to each side of the latter congruence we succeed in completing the square and arriving at the equivalent congruence which follows:

$$(2ax + b)^2 \equiv b^2 - 4ac \bmod p.$$

By setting $u \equiv 2ax + b \bmod p$, we complete the proof; for if there is no u satisfying $u^2 \equiv b^2 - 4ac \bmod p$, then there is no x satisfying the given congruence; and if there is a u satisfying $u^2 \equiv b^2 - 4ac \bmod p$, then because $(2a,p) = 1$ there is an x satisfying $2ax \equiv u - b \bmod p$ and satisfying the given congruence.

22.2. Quadratic residues. As shown in the preceding section, the solution of the general quadratic congruence mod m reduces finally to the solution of pure quadratic congruences modulo primes. With due notice, we now change notation and consider as our typical problem the following pure quadratic congruence:

$$x^2 \equiv a \bmod p, \quad (a,p) = 1, \quad p \text{ an odd prime.}$$

The case $a \equiv 0 \bmod p$ is trivial, having the unique solution $x \equiv 0 \bmod p$, and is not included in the discussion.

We begin with the following useful definitions.

If the congruence $x^2 \equiv a \bmod m$ has a solution, then a is called a *quadratic residue mod m*; if there is no solution, then a is called a *quadratic non-residue mod m*.

For example, 1,4,9,3,12,10 are quadratic residues mod 13, for they are the remainders mod 13 of the squares of $\pm1,\pm2,\pm3,\pm4,\pm5,\pm6$, respectively. Since this exhausts the list of possible solutions (except for the case $x \equiv 0$ leading to $a \equiv 0$ which we have agreed to exclude from the discussion, although by definition 0 is certainly a quadratic residue), it follows that the other non-zero residue classes, 2, 5, 6, 7, 8, 11 are quadratic non-residues mod 13.

G.20: Exactly half the non-zero residues mod p are quadratic residues mod p.

Proof: By **G.18**, $x^2 \equiv a \bmod p$ implies $2 \operatorname{ind} x \equiv \operatorname{ind} a \bmod p - 1$. Since $(2,p-1) = 2$, it follows from **G.8** that there is a solution for ind x, and hence for x, if and only if ind a is *even*. Since the indices mod p run $1,2,\ldots,p-1$, exactly half the indices are even, which completes the proof.

G.21: The integer $a \not\equiv 0 \bmod p$ is a quadratic residue, or a quadratic non-residue mod p, according as

$$a^s \equiv 1 \bmod p, \quad \text{or} \quad a^s \equiv -1 \bmod p,$$

where $s = (p-1)/2$.

Proof: By Euler's theorem, $x^{p-1} - 1 \equiv 0 \bmod p$ has $p - 1$ solutions made up of the non-zero residue classes mod p. In factored form $x^{p-1} - 1 = (x^s - 1)(x^s + 1)$. By **G.16**, the congruence $x^s - 1 \equiv 0 \bmod p$ has exactly s solutions. But it is easy to see that every quadratic residue mod p is a solution of $x^s - 1 \equiv 0 \bmod p$; for if there exists an integer x so that $x^2 \equiv a \bmod p$, then $a^s \equiv x^{2s} \equiv x^{p-1} \equiv 1 \bmod p$, the last congrucnce being justified by Euler's theorem. However, by **G.20** the quadratic residues are s in number, so all the solutions of $x^s - 1 \equiv 0 \bmod p$ are quadratic residues. Hence the remaining s solutions of $x^{p-1} - 1 \equiv 0 \bmod p$ must be quadratic non-residues all of which solve $x^s + 1 \equiv 0 \bmod p$, because by **G.16** and **G.6** the factor $x^s - 1 \not\equiv 0 \bmod p$ may be cancelled, when x is a non-residue.

G.21.1: The product ab is a quadratic non-residue mod p if and only if exactly one of a or b is a quadratic non-residue mod p.

Proof: Since $(ab)^s \equiv a^s b^s \bmod p$, it follows from **G.21** that a^s and b^s are either congruent to $+1$ or $-1 \bmod p$ so that $(ab)^s \equiv -1$ if and only if *just one* of a^s or b^s is congruent to $-1 \bmod p$. By **G.21** this result may be rephrased in terms of residues and non-residues as stated in **G.21.1.**

22.3. Legendre's symbol. In explaining and carrying out tests to decide whether a given integer a is a quadratic residue or non-residue mod p, we shall find it extremely convenient to use a special number-theoretic function known as Legendre's symbol, written (a/p), with its values defined as follows:

$(a/p) = +1$ if $a \not\equiv 0 \bmod p$ and if a is a quadratic residue mod p;
$(a/p) = -1$ if a is a quadratic non-residue mod p.

Of course, it is essential that the user of this Legendre symbol be a bit cautious, and not interpret the symbol as a mere fraction in parentheses, for as we know already from the definition and see again in the next theorems, the properties of the symbol are quite different from those of fractions.

G.22.1: If $a \equiv b \bmod p$, then $(a/p) = (b/p)$.

G.22.2: $(a/p) \equiv a^s \bmod p$.

G.22.3: $(ab/p) = (a/p)(b/p)$.

G.22.4: $(c^2b/p) = (b/p)$.

Proofs: If $a \equiv b \bmod p$ then $x^2 \equiv a \bmod p$ has exactly the same solutions, if any, as has $x^2 \equiv b \bmod p$, which establishes **G.22.1.**

The result in **G.22.2** is a direct consequence of the definition of the Legendre symbol and of **G.21.**

The result in **G.22.3** is a mere restatement of **G.21.1** in terms of the Legendre symbol.

G.22.4 is a special case of **G.22.3** making use of the fact that $a = c^2$ is obviously a quadratic residue so that $(c^2/p) = 1$.

For the present in evaluating Legendre's symbol, we may be content with the above theorems, but in the next chapter a much more elegant method of evaluation, avoiding computations of $a^s \bmod p$, will be explained.

For example, the question "Is 113 a quadratic residue mod 101?" should first be mentally compared with the equivalent question "Does $x^2 \equiv 113 \bmod 101$ have a solution?" and then rephrased "Find the value of (113/101)."

According to **G.22.1** and **G.22.4**, since $113 \equiv 12 \equiv 2^23 \bmod 101$.

the problem reduces to finding $(3/101)$. Then, at this stage of our work, we must have recourse to finding the value of 3^{50} mod 101, as in **G.22.2.** Since $3^5 = 243 \equiv 41$ mod 101, we find in turn that $3^{10} \equiv (41)^2 = 1681 \equiv 65$, $3^{20} \equiv (65)^2 = 4225 \equiv 84$, $3^{25} \equiv (84)(41) = 3444 \equiv 10$, $3^{50} \equiv (10)^2 = 100 \equiv -1$, mod 101. Therefore $(3/101) = -1$; and 3 and 12 and 113 are *not* quadratic residues mod 101.

EXERCISES

EX. *22.1.* As in **G.19**, find the chain of congruences equivalent to $2x^2 + 3x - k \equiv 0$ mod 5 and determine for what values of k there will be solutions.

EX. *22.2.* If $x^2 \equiv a$ mod p has a solution x_1, show that $x_2 = p - x_1$ is also a solution. If $a \not\equiv 0$ and if p is odd show that $x_2 \not\equiv x_1$.

EX. *22.3.* If p is an odd prime and if $(a/p) = 1$, prove that a is a quadratic residue mod p^n and $x^2 \equiv a$ mod p^n has *exactly two* solutions. Use **G.12** and induction.

EX. *22.4.* If a is odd and $m \geqq 3$, then $x^2 \equiv a$ mod 2^m is impossible unless $a \equiv 1$ mod 8.

EX. *22.5.* If $a \equiv 1$ mod 8 and if $m \geqq 3$, then $x^2 \equiv a$ mod 2^m has *exactly four* solutions. Use **G.12** and induction, noting that the four solutions are related: x_1, $x_2 = -x_1$, $x_3 = x_1 + 2^{m-1}$, $x_4 = -x_1 + 2^{m-1}$.

EX. *22.6.* Prove that $(-1/p) = +1$ if and only if p has the form $p = 4K + 1$.

EX. *22.7.* As in **G.22**, find the value of $(791/101)$.

EX. *22.8.* Solve $x^2 \equiv 140$ mod 221.

EX. *22.9.* Solve $x^2 \equiv 65$ mod 280.

EX. *22.10.* Solve $x^2 \equiv 11$ mod 101 (no solutions!).

EX. *22.11.* Solve $x^2 \equiv 33$ mod 101.

EX. *22.12.* Show that for every prime $p > 3$ the *sum* of the quadratic residues mod p is divisible by p. *Hint:* Use EX. *3.3.*

> ▶ *The beautiful has its place in mathematics for here are triumphs of the creative imagination, beautiful theorems, proofs and processes whose perfection of form has made them classic. He must be a "practical" man who can see no poetry in mathematics.*
>
> —W. F. WHITE

CHAPTER 23[*]

THE QUADRATIC RECIPROCITY LAW

23.1. Results leading to the quadratic reciprocity law. We now begin a chain of theorems and corollaries which culminate in the justly revered quadratic reciprocity law with whose aid the Legendre symbol can always be evaluated, and hence the solvability of every pure quadratic congruence and, indeed, of every quadratic congruence decided.

It is this reciprocity law which the master, Gauss, declared to be "the jewel of arithmetic."

It will greatly simplify the statement of all the theorems if it is always understood, as in the previous lesson, that p is an odd prime, that $(a,p) = 1$, and that q is an odd prime distinct from p; furthermore, that $s = (p - 1)/2$ and that $t = (q - 1)/2$.

G.23: If K is the number of least positive residues of the set a, $2a$, $3a, \ldots, sa$ which exceed $p/2$, then $(a/p) = (-1)^K$.

For example, to evaluate $(3/101)$ we may note that $3, 2\cdot 3, \ldots, 16\cdot 3 = 48$ are $<101/2$; then $17\cdot 3, \ldots, 33\cdot 3 = 99$ are $>101/2$; next, $34\cdot 3 = 102 \equiv 1 < 101/2$, $35\cdot 3 \equiv 1 + 3, \ldots,$ $50\cdot 3 \equiv 1 + 16\cdot 3 < 101/2$; hence $K = 33 - 16 = 17$; therefore $(3/101) = (-1)^{17} = -1$.

[*] Chapter 23 is a basic lesson. Some knowledge of the bracket function defined in 9.1 is a prerequisite.

Proof: Suppose $R_i a \equiv r_i$, $S_j a \equiv s_j \bmod p$, $0 < r_i < p/2$, $p/2 < s_j < p$, with $j = 1,2,\ldots,K$; $i = 1,2,\ldots,H = s - K$. Then we claim that $r_1,r_2,\ldots,r_K,p - s_1, p - s_2,\ldots,p - s_H$ represent $1,2,3,\ldots,s$ in some order. Certainly the numbers of this list are all positive, all less than $p/2$, and are s in number; that the numbers are distinct, which will complete the claim, may be seen as follows:

(*1*) If $r_i = r_m$, then $R_i a \equiv R_m a \bmod p$; since $(a,p) = 1$, it follows that $R_i \equiv R_m \bmod p$; hence $R_i = R_m$ for $0 < R_i,R_m \leqq s < p$.

(2) Similarly, $p - s_j = p - s_m$ implies $S_j = S_m$.

(3) If $r_i = p - s_j$, then $r_i + s_j \equiv 0 \bmod p$ so that $(R_i + S_j)a \equiv 0 \bmod p$; since $(a,p) = 1$, it follows that $R_i + S_j \equiv 0 \bmod p$; but this is impossible since $0 < R_i,S_j \leqq s$ so that $0 < R_i + S_j < 2s = p - 1 < p$.

Hence if we multiply together the numbers of the set $a,2a,3a,\ldots,sa$ we may write

$$s!\,a^s = (R_1a)\ldots(R_Ha)(S_1a)\ldots(S_Ka) \equiv r_1\ldots r_H s_1\ldots s_K$$
$$= (-1)^K r_1\ldots r_H(p - s_1)\ldots(p - s_K) = (-1)^K s! \bmod p$$

Since p is a prime, it follows that $(s!,p) = 1$, so that we may apply **G.6** and arrive at $a^s \equiv (-1)^K \bmod p$.

Finally, because of **G.22.2** it follows that $(a/p) = (-1)^K$, the congruence replaced by equality because of the limited range of values of the two symbols appearing and because $p > 2$.

Before continuing with the next theorem let us make use of the notions introduced in the preceding proof to define

$$A = r_1 + r_2 + \ldots + r_H, \quad B = s_1 + s_2 + \ldots + s_K,$$
$$M = [a/p] + [2a/p] + \ldots + [sa/p].$$

where the brackets indicate the bracket function of Chapter 9.

G.24: $(a - 1)(p^2 - 1)/8 = (M - K)p + 2B.$

Proof: By the division algorithm we find

$$R_i a = p[R_i a/p] + r_i, \quad S_j a = p[S_j a/p] + s_j.$$

By EX. *3.1* we know that $1 + 2 + 3 + \ldots + s = s(s + 1)/2 = (p^2 - 1)/8$. Hence we may write

(*24.1*) $a(p^2 - 1)/8 = a + 2a + \ldots + sa = Mp + A + B.$

Then using the preliminary claim in the *proof* of **G.23**, we may write

(*24.2*) $(p^2 - 1)/8 = 1 + 2 + \ldots + s = r_1 + \ldots + r_H + (p - s_1) + \ldots + (p - s_K) = A + Kp - B.$

By subtracting (*24.2*) from (*24.1*) we eliminate A and arrive at **G.24**.

G.24.1: $(2/p) = (-1)^{(p^2-1)/8}.$

Proof: When we take $a = 2$ in **G.24**, we must take $M = 0$, because $M = [2/p] + [4/p] + \dots + [(p-1)/p]$ contains only summands for which the bracket function is zero. Hence **G.24** shows that $(p^2 - 1)/8 = 2B - Kp \equiv -Kp \equiv -K \equiv +K \bmod 2$. Then from **G.23**, we obtain **G.24.1**.

G.24.2: If $M = [q/p] + [2q/p] + \dots + [sq/p]$, then $(q/p) = (-1)^M$.

Proof: Since q is odd, $(q - 1)$ is even and if we take $a = q$ in **G.24** we find since p also is odd that $M \equiv K \bmod 2$. By **G.23** it follows that $(q/p) = (-1)^M$.

G.24.3: If $N = [p/q] + [2p/q] + \dots + [tp/q]$, then $(p/q) = (-1)^N$.

Proof: For this corollary we need but change the roles of p and q in the preceding **G.24.2**.

G.25: In the notation of **G.24.2** and **G.24.3**, $M + N = st$.

Proof: The proof is a geometric one, originated by Eisenstein, a pupil of Gauss. Consider a Cartesian coordinate system and (as in **11.3**) define a lattice point to be a point (x,y) both of whose coordinates are integers.

On the one hand, since p and q are odd, the number of lattice points *inside* the rectangle whose vertices are O:(0,0), A: $(p/2,0)$, B: $(p/2,q/2)$, C: $(0,q/2)$ is given by st.

On the other hand, there are no lattice points within the rectangle on the diagonal OB; and the numbers of lattice points *inside* triangles OAB and OBC are given by M and N, respectively.

The first of these assertions follows from the fact that the equation of OB is $py = qx$; then inasmuch as $(p,q) = 1$, it follows that a lattice point (x,y) satisfying this equation would have to have x a multiple of p(and y a multiple of q), but the x's under consideration range only from 1 to s.

The second assertion follows from the fact that $[kq/p]$ is the number of lattice points on the vertical line $x = k$ between OA and OB, because these lattice points must have y-coordinates satisfying $0 < y \leqq [kq/p]$. Summing from $k = 1$ to $k = s$, we find M lattice points inside triangle OAB. In a similar manner $[up/q]$ is the number of lattice points on the horizontal line $y = u$ between OC and OB.

Summing from $u = 1$ to $u = t$, we find N lattice points inside triangle OBC.

Equating the results of the two ways of counting the number of lattice points inside the rectangle, we have the desired relation $M + N = st$.

G.26: The *quadratic reciprocity law:* $(q/p) = (p/q)(-1)^{st}$.

Proof: From **G.24.2**, **G.24.3**, and **G.25**, we find

$$(p/q)(q/p) = (-1)^{M}(-1)^{N} = (-1)^{M+N} = (-1)^{st}.$$

Finally, whether (p/q) is $+1$ or -1, we have $(p/q)^2 = +1$; hence if we multiply the last displayed equation by (p/q) we arrive at the law stated in **G.26**.

This law receives its name for obvious reasons. On the one hand it deals with symbols which concern "quadratic" residues or non-residues. On the other hand, the symbols (q/p) and (p/q) which appear in the law are in a sense "reciprocal." The implications of this last statement are well used in the next section.

23.2. The evaluation of Legendre's symbol.

Given any A not a multiple of p, we may decide whether the congruence $x^2 \equiv A \bmod p$ has a solution, or not, by finding whether (A/p) is $+1$, or -1, respectively.

To evaluate (A/p) we may proceed as follows:

(*1*) If a is the absolutely least residue of A mod p, we may write $(A/p) = (a/p)$ by **G.22.1**.

(2) If a contains any perfect squares, say $a = m^2b$, where b is "square-free," we may write $(a/p) = (b/p)$ by **G.22.4**.

(3) The most complicated form which b can have is $b = (-1)2q_1q_2\ldots q_k$ where the q's are distinct odd primes; by **G.22.3** we may write $(b/p) = (-1/p)(2/p)(q_1/p)\ldots(q_k/p)$.

(4) To evaluate $(-1/p)$ we use $(-1)^s$ as in **G.21**.

(5) To evaluate $(2/p)$ we use -1 with the exponent $(p^2 - 1)/8$ as in **G.24.1**.

(6) To evaluate each (q/p) we use the quadratic reciprocity law **G.26** for this leads us to a new problem with a smaller "denominator," since in (p/q) we have $q \leq |b| \leqq |a| \leqq s < p$; we may begin the above routine again for (p/q) and eventually arrive at Legendre symbols that can be evaluated directly.

(7) Collecting the results in (4), (5), and (6) and substituting carefully in (3) we find the value of (A/p).

As an example we consider the evaluation of $(231/997)$. Here the prime 997 is so large that a direct consideration of the congruence $x^2 \equiv 231 \bmod 997$ is not practical. After factoring $231 = 3 \cdot 7 \cdot 11$, we write $(231/997) = (3/997)(7/997)(11/997)$.

To find $(3/997)$ we use **G.26** to write

$$(3/997) = (997/3)(-1)^{498 \cdot 1} = (1/3) = +1.$$

Here we have used $s = (997 - 1)/2$, $t = (3 - 1)/2$, and $997 \equiv 1 \bmod 3$.

To find $(7/997)$ we use **G.26** twice to write

$$(7/997) = (997/7)(-1)^{498 \cdot 3} = (3/7) = (7/3)(-1)^{3 \cdot 1} = -(1/3) = -1.$$

To find $(11/997)$ we use **G.26** twice to write

$$(11/997) = (997/11)(-1)^{498 \cdot 5} = (7/11) = (11/7)(-1)^{5 \cdot 3} = -(4/7) = -1.$$

Combining these results we conclude that

$$(231/997) = (+1)(-1)(-1) = +1.$$

Hence 231 *is* a quadratic residue of 997.

As a further example let us consider the problem of finding all odd primes p for which 11 is a quadratic residue. Evidently we must determine p so that $(11/p) = +1$ and by **G.26** and **G.22.1** we may suppose $p \equiv p' \bmod 11$ and write

$$(11/p) = +1 = (p'/11)(-1)^{5(p-1)/2}, \qquad 0 < p' < 11.$$

When $(p'/11) = +1$, i.e., when $p' = p_1$ is a quadratic residue mod 11, we must have $(p - 1)/2$ *even*, or $p \equiv 1 \bmod 4$. When $(p'/11) = -1$, i.e., when $p' = p_2$ is a quadratic non-residue mod 11, we must have $(p - 1)/2$ *odd*, or $p \equiv 3 \bmod 4$. By the Chinese remainder theorem, we must have in the first case $p \equiv p_1 \bmod 11$, $p \equiv 1 \bmod 4$, or $p \equiv 33 + 12p_1 \bmod 44$; and in the second case, with $p \equiv p_2 \bmod 11$, $p \equiv 3 \bmod 4$, we must have $p \equiv 11 + 12p_2 \bmod 44$. Since $11 \equiv 3 \bmod 4$, it follows from **G.21** that $(-1/11) = -1$, hence we may pair off the numbers p_2 and p_1 by the relation $p_2 = 11 - p_1$. But also $11 \equiv -33 \bmod 44$, so the two cases in the above argument may be combined into one formula: $p \equiv \pm(33+12p_1) \bmod 44$. Specifically, since 1,4,9,5,3 are the quadratic residues mod 11, we find that 11 is a quadratic residue of an odd prime p if and only if p has the form

$$p = 44T \pm u$$

where $u = 1,5,7,9,19$, and where T is an arbitrary integer such that p is prime and $p > 0$.

Some examples are as follows:

$$p = 5,\ 7,\ 19,\ 37,\ 43,\ 53,\ 79,\ 83,\ 89,\ 97,\ 107,\ 113,\ 127,\ 131.$$

23.3. Concluding remarks. Legendre's symbol and the quadratic reciprocity law afford an elegant solution of the problem of determining the *existence* of solutions of $x^2 \equiv a \bmod p$; but in those cases where solutions exist, and it is required that they be found, there remains considerable labor, especially in case of a large prime. Some labor-saving suggestions are given in the Uspensky and Heaslet text cited in **1.3**.

If we consider $x^2 \equiv a \bmod m$, where the modulus is composite, we may use the methods of Chapter 22 to solve the problem. In particular, by virtue of EX. *22.3*, EX. *22.4*, EX. *22.5*, and the results of this lesson, we can decide whether the congruence has a solution without actually solving it. Some reduction in this last problem can be effected by the use of the Jacobi symbol, which is an interesting generalization of the Legendre symbol. The properties of the Jacobi symbol are discussed in Uspensky and Heaslet, and some of the properties are discussed in the following exercises.

EXERCISES

EX. *23.1* Evaluate (783/997) and (127/997).

EX. *23.2.* Evaluate $(2/p)$ for $p = 8K + 1$, $8K + 3$, $8K + 5$, $8K + 7$.

EX. *23.3.* Determine whether $x^2 \equiv 239 \bmod 2431$ has solutions. Note that $2431 = 11 \cdot 13 \cdot 17$.

EX. *23.4.* Find all primes p for which 7 is a quadratic residue.

EX. *23.5.* Find all primes p for which 13 is a quadratic residue.

EX. *23.6.* If p and q are distinct odd primes with $p \equiv 1 \bmod 4$, show that $(p/q) = +1$, if and only if q has the form

$$q \equiv p + a(p + 1) \bmod 2p$$

where $(a/p) = +1$.

EX. *23.7.* If p and q are distinct odd primes with $p \equiv 3 \bmod 4$, show that $(p/q) = +1$, if and only if q has the form

$$q \equiv \pm \{3p + a(p + 1)\} \bmod 4p$$

where $(a/p) = +1$.

EX. *23.8.* If $P = p_1p_2 \ldots p_k$ where the p's are odd primes, use induction to prove that
$(P - 1)/2 \equiv (p_1 - 1)/2 + (p_2 - 1)/2 + \ldots + (p_k - 1)/2 \bmod 2.$

EX. *23.9.* Using EX. *23.8* and assuming $(P,q) = 1$ where q is an odd prime, show that $(P/q) = (q/p_1) \ldots (q/p_k)(-1)^{(P-1)(q-1)/4}$.

EX. *23.10.* If $P = p_1p_2 \ldots p_k$ where the p's are odd primes and if $(Q,P) = 1$ define the Jacobi symbol (Q/P) as follows

$$(Q/P) = (Q/p_1)(Q/p_2) \ldots (Q/p_k).$$

If $(Q/P) = -1$, show that $x^2 \equiv Q \bmod P$ has no solution.

If $(Q/P) = +1$, show that $x^2 \equiv Q \bmod P$ may *or* may not have solutions.

EX. *23.11.* Use the preceding exercises to show if Q and P are odd with $(Q,P) = 1$, then

$$(P/Q) = (Q/P)(-1)^{(P-1)(Q-1)/4}.$$

▶*Arithmetical symbols are written diagrams and geometrical figures are graphic formulas.* —D. HILBERT

CHAPTER 24*

ADDITIVE ARITHMETIC

24.1. Introduction to additive arithmetic. The chief purpose of the present lesson is that of background for the following lesson. For more detailed discussion a very good reference is the work by Hardy and Wright.

The principal problem of the additive theory of numbers may be phrased as follows: given a set of A of integers $a_1,a_2,a_3,\ldots$, consider the representation of an integer n in the form $n = a_{i_1} + a_{i_2} + \ldots + a_{i_s}$, where s may or may not be fixed, where the a's may or may not be different, and where the order of the a's may or may not be relevant. Let $A(n)$ be the number of representations of n in this form. The simpler problem is to determine whether $A(n)$ is positive; the harder problem is to find the exact value of $A(n)$; a related problem would be to find the greatest restrictions on s under which $A(n)$ will remain positive.

As an example, let P be the set of positive integers, let s be unrestricted, let repetitions be allowed, and let order be disregarded. Let $P(n)$ be the number of representations of n, as explained above, for this particular problem. To explain Euler's description of $P(n)$

*Chapter 24 is a basic chapter in the sense of providing background for the following chapter.

we need to digress and explain an abbreviation often used in discussion of this subject matter.

Adapting EX. *3.2* we may write

$$(1 - x^i)(1 + x^i + x^{2i} + \ldots + x^{qi}) = 1 - x^{(q+1)i} \quad .$$

If x is a rational number, say, such that $0 \leqq |x| < 1$, then for any assigned rational number $\epsilon > 0$, there can be found an integer Q so that $|x^{(q+1)i}| < \epsilon$ whenever $q > Q$. For this reason let us agree that $1/(1 - x^i)$ is a suggestive abbreviation for the *infinite* polynomial $(1 + x^i + x^{2i} + \ldots + x^{qi} + \ldots)$.

Theorem: $P(n)$ is the coefficient of x^n in the product

$$1/(1 - x)(1 - x^2)\ldots(1 - x^n) \quad .$$

Proof: The proof consists in applying the above definition to each of the factors $1/(1 - x^i)$ for $i = 1,2,\ldots,n$ and then forming the product. To a term x^n of the product each $1/(1 - x^i)$ must contribute one and only one factor; and by the rules of exponents if x^{ji} contributes as a factor to some x^n, it is because j of the i's are summands in a representation of n; conversely, each representation of n corresponds to a unique set of factors, one from each $1/(1 - x^i)$, whose product is x^n.

For example, to find $P(5)$ we may compute the coefficient of x^5 in the product $1/(1 - x)(1 - x^2)(1 - x^3)(1 - x^4)(1 - x^5)$ or $(1 + x + x^2 + x^3 + x^4 + x^5 + \ldots)(1 + x^2 + x^4 + \ldots)(1 + x^3 + \ldots)$ $(1 + x^4 + \ldots)(1 + x^5 + \ldots)$, where we have purposely relegated to the ellipsis those powers of x which are not relevant. We find $P(5) = 7$ and if we analyze the contributing terms and corresponding representations, as an illustration of the argument intended above, we find

$1 \cdot 1 \cdot 1 \cdot 1 \cdot x^5$ or 5, $x \cdot 1 \cdot 1 \cdot x^4 \cdot 1$ or $1 + 4$, $1 \cdot x^2 \cdot x^3 \cdot 1 \cdot 1$ or $2 + 3$,
$x^2 \cdot 1 \cdot x^3 \cdot 1 \cdot 1$ or $1 + 1 + 3$, $x \cdot x^4 \cdot 1 \cdot 1 \cdot 1$ or $1 + 2 + 2$,
$x^3 \cdot x^2 \cdot 1 \cdot 1 \cdot 1$ or $1 + 1 + 1 + 2$, $x^5 \cdot 1 \cdot 1 \cdot 1 \cdot 1$ or $1 + 1 + 1 + 1 + 1$.

The product function of the theorem is described as a "generating function" for $P(n)$ and the theorem itself is described as part of the "theory of partitions."

At first we feel pleased with the simplicity of the theorem, but if we try the theorem as a means of finding $P(n)$ for a large value of n, we are liable to be disappointed; for if we try directly to find the coefficient of x^n, our work is equivalent to writing down all the representations. However, by clever manipulation of the generating functions we might be able to find recursion formulas which would make computation much easier. But, in general, only the more modern methods of analytic number theory have provided ways of computing enumerative functions like $P(n)$ for large values of n.

We may specialize the above example by saying: let A contain only 1,2,3; let s be unrestricted, repetitions allowed, and order disregarded. Then the generating function appropriate for $A(n)$ is $1/(1 - x)(1 - x^2)(1 - x^3)$ with $A(n)$ as the coefficient of x^n in this product. The argument is almost word for word like that in the preceding proof, except that i is restricted to the values 1,2,3. The expanded form of the generating function begins as follows:

$$1 + x + 2x^2 + 3x^3 + 4x^4 + 5x^5 + 7x^6 + 8x^7 + 10x^8 + \cdots,$$

hence the values of $A(n)$ for $n = 1,2,\ldots,8$ are in evidence.

A problem amusingly related to the preceding one is provided by letting A^* contain all non-negative integers, requiring s to be 3, allowing repetition, and disregarding order. For we may show $A^*(n) = A(n)$.

On the one hand, if $n = u + 2v + 3w$, with u,v,w non-negative, then we may write $n = a + b + c$, where $a = u + v + w$, $b = v + w$, $c = w$ are non-negative, and such that $a \geqq b \geqq c$. Conversely, if $n = a + b + c$, with a,b,c non-negative and arranged in the order $a \geqq b \geqq c$, then $n = u + 2v + 3w$, where $u = a - b$, $v = b - c$, $w = c$ are non-negative.

Thus when $n = 6$, we have $A(6) = 7$ cases, as follows:

$$6=0\cdot1+0\cdot2+2\cdot3=1\cdot1+1\cdot2+1\cdot3=3\cdot1+0\cdot2+1\cdot3=0\cdot1+3\cdot2+0\cdot3$$
$$=2\cdot1+2\cdot2+0\cdot3=4\cdot1+1\cdot2+0\cdot3=6\cdot1+0\cdot2+0\cdot3;$$

and the corresponding cases, as described above, showing $A^*(6) = 7$, are as follows:

$$6 = 2 + 2 + 2 = 3 + 2 + 1 = 4 + 1 + 1 = 3 + 3 + 0$$
$$= 4 + 2 + 0 = 5 + 1 + 0 = 6 + 0 + 0.$$

24.2. Waring's problem. In terms of additive arithmetic we may describe one of the most famous problems of the theory of numbers, usually known as "Waring's problem" although it seems that there was probably no case of his problem for which Waring could give a demonstration.

Let k be a fixed integer, $k \geqq 2$, and let A be the set of kth powers of non-negative integers: $0^k,1^k,2^k,\ldots$; then Waring's problem is to determine whether there exists an integer $s = s(k)$, depending on k but *independent* of n, such that if we allow repetitions, we have $A(n) > 0$ for all n.

In other words, for a given k we seek an $s = s(k)$, such that every n can be written in at least one way as

$$n = a_1{}^k + a_2{}^k + \ldots + a_s{}^k,$$

where $a_1,a_2,\ldots,a_s$ are non-negative integers, not necessarily distinct.

It was a triumph, more for analysis than number theory, that Waring's problem was answered in the affirmative, for all k, by Hilbert, one hundred years after Waring. But the proof is an existential one, and the attempt to give an explicit value for s, for all k, is not yet quite successful. Knowing that s exists, we can see that any greater integer has the same property. Hence it is natural to define $g(k)$ to be the *least* value of s, such that every n is representable as the sum of g kth powers, but there is *at least one* n which cannot be represented by fewer than g kth powers. For example, it has been shown that $g(3) = 9$, meaning that every integer may be represented as the sum of 9 cubes of non-negative integers, and that there is at least one integer which actually requires 9 cubes in this kind of representation. However, it turns out in this case that there are *only two* integers n requiring the full complement of 9 cubes; for this, and other reasons, it is natural to define $G(k)$ to be such that all but a finite number of integers n can be represented as the sum of G kth powers, and *infinitely many* integers n cannot be represented as the sum of fewer than G kth powers. For example, the value of $G(3)$ is still in doubt, but is restricted to the range $4 \leqq G(3) < 9$, by the facts given above and the additional fact that there are known to be infinitely many integers requiring 4 cubes in their representation.

The only case of Waring's problem sufficiently simple for these lessons is the case when $k = 2$; and in our next chapter we will show that $g(2) = G(2) = 4$. In other words, every integer n can be written in at least one way as the sum of 4 squares of integers, and there are infinitely many integers which cannot be written as the sum of fewer than 4 squares. For a discussion of the recent status of the $g(k)$ and $G(k)$ problems, for other values of k, the reader may refer to Hardy and Wright.

24.3. Polygonal numbers. Let t be a fixed integer, $t \geqq 3$. Let A_t be the set of all *polygonal numbers of order* t, defined for $i = 0,1,\ldots,$

by $a(i,t) = i\{2 + (t - 2)(i - 1)\}/2$. Let $s(t) = t$, let repetitions be allowed and order disregarded. Then the Cauchy-Fermat result is that $A_t(n) > 0$ for every positive integer n. Thus every n is the sum of three triangular numbers, four square numbers, five pentagonal numbers, etc.

The numbers receive their geometric description because, with the exception of 0 and 1, which occur in every set, they can be described, for a given t, as a nest of regular polygons, each of t sides, homothetic with respect to a common vertex, and having, successively, $i = 2,3,\ldots$ points on a side. For if we count the number of points in the polygon at the stage where there are i points on a side, we obtain the polygonal number $a(i,t)$, inasmuch as the sum of the terms of the arithmetic progression

$$1,\ 1 + (t - 2),\ 1 + 2(t - 2),\ \ldots,\ 1 + (i - 1)(t - 2)$$

is precisely $a(i,t)$.

For example, the triangular numbers $a(i,3) = i(i + 1)/2$ are $0,1,3,6,10,15,21,\ldots$; and examples of the Cauchy-Fermat theorem are as follows:

$18=15+3+0=6+6+6,\ 19=15+3+1=10+6+3,\ 20=10+10+0.$

Since it turns out that the square numbers $a(i,4) = i^2$ are, indeed, the squares of integers, this particular case of the Cauchy-Fermat theorem is the same as the Waring problem for $k = 2$, so the proofs of the next chapter are applicable. A modern discussion of the Cauchy-Fermat theorem, for all values of t, can be found in Dickson's "*Modern Elementary Theory of Numbers.*"

EXERCISES

EX. *24.1.* If A contains 1 and k, if s is unrestricted, if repetitions are allowed and order disregarded, show that $A(n) = [n/k] + 1$ in three ways:

(*1*) by direct enumeration;

(*2*) by an appropriate generating function;

(*3*) by considering $A^*(n)$, where A^* consists of all non-negative integers, $s^* = 2$, repetitions allowed and order disregarded.

EX. *24.2.* If A contains 2 and 3 with s unrestricted, repetitions allowed and order disregarded, show that $A(n) = [n/6]$, or $[n/6] + 1$ according as $n \equiv 1$ or $n \not\equiv 1$, mod 6.

EX. *24.3.* If A contains $a_1,a_2,\ldots,a_k$ where $0 < a_1 < a_2 < \ldots < a_k$, if s is unrestricted, if repetitions are allowed and if order *is* considered, show that by defining $A(0) = 1$ and noting that $A(n) = 0$ for $n = 1,2,\ldots,$

$a_1 - 1$, then all other $A(n)$ may be computed by the following recursion formulas:

$$A(n) = A(n - a_1) + A(n - a_2) + \ldots + A(n - a_i), \qquad a_i \leqq n < a_{i+1},$$
$$i = 1,2,\ldots,k-1;$$
$$A(n) = A(n - a_1) + A(n - a_2) + \ldots + A(n - a_k), \qquad a_k \leqq n.$$

(*Hint:* Arrange the representations of n in lexicographic order.)

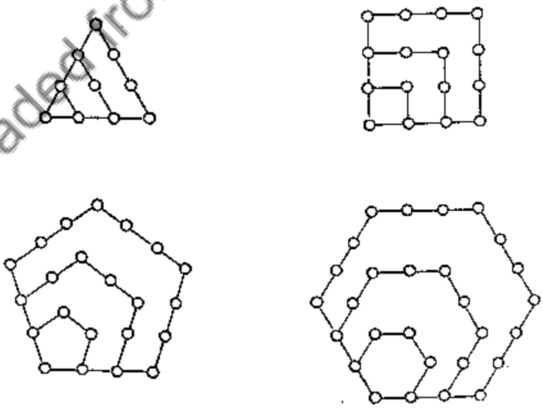
EX. *24.4.* In the special case of EX. *24.3* where A contains only 1 and 2, compute some of the $A(n)$—these are the *Fibonacci numbers.*

EX. *24.5.* If $(a_1,a_2,\ldots,a_k) = d$ and $a_i = A_i d$, let A^* contain $A_1,A_2,\ldots,A_k$; then under the conditions of EX. *24.3* compare $A(n)$ and $A^*(n)$, making appropriate use of **12.1.**

EX. *24.6.* Show (a) geometrically and (b) algebraically:

(a) $4a(i,3) + (i+1) = a(i+1,6)$,
(b) $(t-2)a(i,3) + (i+1) = a(i+1,t)$;
(a) $a(i,5) + a(i-1,3) = a(i,6)$,
(b) $a(i,t) + a(i-1,3) = a(i,t+1)$;
(a) $a(i,6) + 1 = 2a(i,3) + a(i-1,4)$,
(b) $a(i,2t) + 1 = 2a(i,t) + a(i-1,4)$.

EX. *24.7.* For $1 \leqq k \leqq n$, define $P(n,k)$ to be the number of representations of n in the form $n = a_1 + a_2 + \ldots + a_t$ where $k \leqq a_1 \leqq a_2 \leqq \ldots \leqq a_t$ and $t \geqq 1$. Show that $P(n) = P(n,1)$. If $[n/2] < k \leqq n$, prove that $P(n,k) = 1$. If $1 \leqq k \leqq [n/2]$, prove that $P(n,k) = 1 + \Sigma P(n-i,i)$, summed over $k \leqq i \leqq [n/2]$. Also prove that $P(n,k+1) = P(n,k) - P(n-k,k)$.



▶ *The majority of ideas we deal with were conceived by others, often centuries ago. In a great measure, it is really the intelligence of other people that confronts us in science.*
—D. MACH

CHAPTER 25[*]

SUM OF FOUR SQUARES

25.1. Four lemmas. In this lesson we will present a proof, essentially due to Euler, that every positive integer is the sum of four squares of integers; i.e., in the language of the preceding lesson, we will solve Waring's problem, when $k = 2$, with the very precise result that $g(2) = G(2) = 4$. To this end the following lemmas will be useful.

L.1: It is true that $g(2) \geqq G(2) \geqq 4$.

Proof: Consider the following table:

If $\qquad x \equiv 0, 1, 2, 3, 4, 5, 6, 7 \bmod 8,$

then $\qquad x^2 \equiv 0, 1, 4, 1, 0, 1, 4, 1 \bmod 8$, respectively.

Consequently, a study of the various cases shows that if x,y,z are any three given integers, then

$$x^2 + y^2 + z^2 \equiv 0, 1, 2, 3, 4, 5, \text{ or } 6 \bmod 8.$$

Therefore there are infinitely many positive integers of the form $8m + 7$ which are not representable as the sum of three squares of integers. In the language of Chapter 24, this is **L.1**.

L.2: If every prime is the sum of four squares, then every composite integer is the sum of four squares.

[*] Chapter 25 is a basic chapter.

Proof: It is a matter of patience to verify the following remarkable identity found by Euler:

$$(25.1)\begin{cases}(a^2+b^2+c^2+d^2)(a_1^2+b_1^2+c_1^2+d_1^2) = A^2+B^2+C^2+D^2\\ A = aa_1+bb_1+cc_1+dd_1, \quad B = ab_1-ba_1+cd_1-dc_1,\\ C = ac_1-bd_1-ca_1+db_1, \quad D = ad_1+bc_1-cb_1-da_1.\end{cases}$$

From this identity **L.2** is an immediate consequence, for every composite integer n is the product of primes and by application of (25.1), an appropriate number of times, a representation of n as the sum of four squares can be obtained if representations are known for each of the prime factors of n.

L.3: If p is an odd prime, there exists a solution in integers x,y,z,m of $x^2+y^2+z^2 = mp$ with $0 < m < p$.

Proof: First we show by contradiction that there is a solution x,y,z of the congruence $x^2+y^2+z^2 \equiv 0 \bmod p$, other than the trivial solution $x \equiv y \equiv z \equiv 0 \bmod p$. For if we suppose there is no solution of the given congruence, except the trivial solution, then (using the Legendre symbol of Chapter 22) we must have $(-1/p) = -1$. Otherwise, we would be able to find y so that $y^2 \equiv -1 \bmod p$, and we would have $1^2+y^2+0^2 \equiv 0 \bmod p$ and a non-trivial solution. Again, if for any integer $a \not\equiv 0$ or -1, mod p, we assume $(a/p) = 1$, we must have $(-(a+1)/p) = -1$. Otherwise we would be able to find x,y,z with $x^2 \equiv 1$, $y^2 \equiv a$, $z^2 \equiv -(a+1)$ mod p and would have a non-trivial solution of $x^2+y^2+z^2 \equiv 0$ mod p. Combining these observations and using **G.22.3** in Chapter 22, we find that if $(a/p) = 1$, for an $a \not\equiv 0$ or -1, mod p, then $((a+1)/p) = (-1/p)(-(a+1)/p) = (-1)(-1) = +1$. Beginning with the case $a = 1$, where $(1/p) = 1$, this would imply by induction that *every* non-zero residue class mod p is a quadratic residue mod p which would be an obvious contradiction of **G.20.**

Having shown the existence of a non-trivial solution x,y,z of the congruence, we may suppose this solution replaced by X,Y,Z where $X \equiv \pm x$, $Y \equiv \pm y$, $Z \equiv \pm z$ mod p with the signs so chosen that $|X| < p/2$, $|Y| < p/2$, $|Z| < p/2$. Then X,Y,Z is also a non-trivial solution of the congruence so that

$$0 < mp = X^2+Y^2+Z^2 < 3p^2/4 < p^2, \quad \text{hence } 0 < m < p.$$

L.4: If p is an odd prime and if $x^2+y^2+z^2+w^2 = mp$ with

$1 < m < p$, then there exist integers x_1,y_1,z_1,w_1, and M such that $x_1^2 + y_1^2 + z_1^2 + w_1^2 = Mp$ with $1 \leqq M < m$.

Proof: The proof is divided into two cases according as m is *even* or *odd*.

When m is *even*, then x,y,z,w are all even; or all are odd; or two are even and two are odd, say x and y. With this agreement we may use the hypothesis to write
$((x + y)/2)^2 + ((x - y)/2)^2 + ((z + w)/2)^2 + ((z - w)/2)^2 = (m/2)p.$
Hence $x_1 = (x + y)/2$, $y_1 = (x - y)/2$, $z_1 = (z + w)/2$, $w_1 = (z - w)/2$, and $M = m/2$ are integers satisfying the conclusions of L.4.

When m is odd, we may use the modified division algorithm for least absolute value remainder to write

$$x = am + a_1, \quad y = bm + b_1, \quad z = cm + c_1, \quad w = dm + d_1,$$

where $|a_1| < m/2, \quad |b_1| < m/2, \quad |c_1| < m/2, \quad |d_1| < m/2.$

Substituting these expressions into the given equation and making use of the symbols introduced in (*25.1*) we find

$$(25.2) \quad a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2Am + (a^2 + b^2 + c^2 + d^2)m^2 = mp.$$

Hence there exists a non-negative integer M such that

$$(25.3) \qquad a_1^2 + b_1^2 + c_1^2 + d_1^2 = Mm.$$

Furthermore we cannot have $M = 0$ for this would imply that $a_1 = b_1 = c_1 = d_1 = 0$; then m^2 would divide $x^2 + y^2 + z^2 + w^2 = mp$, and m would divide p; but $1 < m < p$ and p is a prime, so this case cannot occur. Since we know $a_1^2 + b_1^2 + c_1^2 + d_1^2 < 4(m^2/4) = m^2$, it follows that M in (*25.3*) satisfies $M < m$. Putting these results together we have $1 \leqq M < m$. Finally, from the relations (25.2) and (*25.3*) we find on dividing by m that

$$M + 2A + (a^2 + b^2 + c^2 + d^2)m = p.$$

If we multiply this last equation by M and employ (*25.3*) and (*25.1*) we obtain

$$M^2 + 2AM + A^2 + B^2 + C^2 + D^2 = Mp,$$
$$(M + A)^2 + B^2 + C^2 + D^2 = Mp.$$

Thus $x_1 = M + A$, $y_1 = B$, $z_1 = C$, $w_1 = D$, and M are integers satisfying the conclusions of **L.4**.

25.2. Representation by four squares. The preceding lemmas allow a precise disposition of Waring's problem when $k = 2$.

Theorem: For representation as a sum of squares $G(2) = g(2) = 4$.

Proof: For every odd prime p, **L.3** guarantees the existence of a solution of $x^2 + y^2 + z^2 + w^2 = mp$ with $1 \leqq m < p$. If $m > 1$, then **L.4** allows a descent in a finite number of steps (say with $p > m > M = M_1 > M_2 > \ldots > M_k = 1$) to the situation

$$x_k{}^2 + y_k{}^2 + z_k{}^2 + w_k{}^2 = p.$$

In other words, this shows that every odd prime may be represented as the sum of four squares. The only even prime 2 may be represented in the form $2 = 1^2 + 1^2 + 0^2 + 0^2$. Since every prime can be represented as the sum of four squares, **L.2** guarantees that every composite number may be represented as the sum of four squares. For the unit 1 we have $1 = 1^2 + 0^2 + 0^2 + 0^2$. Thus we have shown that every positive integer may be represented as the sum of four squares. In other words, $g(2) \leqq 4$.

If we now apply **L.1**, we have $4 \leqq G(2) \leqq g(2) \leqq 4$. Therefore, we conclude that $G(2) = g(2) = 4$.

According to W. W. R. Ball, the mental calculator Jacques Inaudi could express numbers less than $(10)^5$ as a sum of four squares in a minute or two. Such ability is certainly unusual and either depended on unusual memory or on the application of some trick process, certainly not on following through the processes indicated by the above proof, for in the case of large primes it is not so easy to produce a solution of the type whose existence is guaranteed by **L.3**.

If $n = u^2N$, it is clear that a representation for N, say

$$N = x^2 + y^2 + z^2 + w^2,$$

leads to a representation for n, say

$$n = X^2 + Y^2 + Z^2 + W^2,$$

with $X = ux$, $Y = uy$, $Z = uz$, $W = uw$. Thus if the problem is merely that of finding one representation, it suffices to deal with an N that is square-free.

For example, if $n = 351 = 9 \cdot 39$, we can write

$$N = 39 = 6^2 + 1^2 + 1^2 + 1^2$$

and then we can easily obtain $351 = (18)^2 + 3^2 + 3^2 + 3^2$. But we can also write $27 = 5^2 + 1^2 + 1^2 + 0^2$ and $13 = 3^2 + 2^2 + 0^2 + 0^2$, and apply *(25.1)* to obtain

$$A = 15 + 2 + 0 + 0 = 17, B = 10 - 3 + 0 - 0 = 7,$$
$$C = 0 - 0 - 3 + 0 = -3, D = 0 + 0 - 2 - 0 = -2,$$

so that $351 = (17)^2 + 7^2 + 3^2 + 2^2$.

The question of the total number of representations has received its neatest answer in the case in which the set A is described as including the squares of all integers, counting $(-x)^2$ as different from $(+x)^2$ except, of course, when $x = 0$, requiring $s = 4$, allowing repetitions and considering order. In this form of the problem Jacobi has shown that

$A(n) = 8\sigma(n)$, when n is odd;

$A(n) = 24\sigma(m)$, when m is odd and $n = 2^a m$ with $a \geqq 1$.

Here $\sigma(n)$ is the number-theoretic function of Chapter 8 denoting the sum of the positive divisors of n.

For example, to explain $A(1) = 8$, we need to realize that the representations involving (1,0,0,0), (−1,0,0,0), (0,1,0,0), (0,−1,0,0), (0,0,1,0), (0,0,−1,0), (0,0,0,1), (0,0,0,−1) as (x,y,z,w) are being counted as distinct. From this point of view, we find that 351 has $A(351) = 8 \cdot 40 \cdot 14 = 4480$ representations. From the point of view where negative solutions are not used and order is disregarded, there are just 14 representations of 351, as follows:

(18,3,3,3), (18,5,1,1), (17,7,3,2), (17,6,5,1), (15,11,2,1),
(15,10,5,1), (15,9,6,3), (14,11,5,3), (14,9,7,5), (13,13,3,2),
(13,11,6,5), (13,10,9,1), (11,11,10,3), (11,10,9,7).

It is easily found that ten of these solutions, under choice of sign and permutations, each lead to $16 \cdot 24 = 384$ of the solutions considered by Jacobi; there are three of these solutions which each lead to $16 \cdot 12 = 192$ of Jacobi's; and one solution which gives $16 \cdot 4 = 64$ of Jacobi's. The grand total of $3840 + 576 + 64 = 4480$ is in agreement with Jacobi's formula.

For discussion of Jacobi's enumeration formula, the reader may refer to Dickson's books, or Uspensky and Heaslet.

EXERCISES

EX. *25.1.* Verify Euler's identity *(25.1)*.

EX. *25.2.* Show that no integer of the form $4^k(8m + 7)$, $k \geqq 0$, can be a sum of three squares. (*Hint*: If $k > 0$, make an argument by descent to the case $k - 1$; for $k = 0$, use **L.1**.)

EX. *25.3.* Note $p = 151 \equiv 7 \bmod 8$. Illustrate the proof of **L.4**, starting with $(20)^2 + 7^2 + 2^2 = 3 \cdot 151$.

EX. *25.4.* Show $(a^2 + b^2 + c^2)^2 = (a^2 + b^2 - c^2)^2 + (2ac)^2 + (2bc)^2$ (Catalan).

EX. *25.5.* Find all representations of 408 as the sum of four squares.

EX. *25.6.* If $x_1 \geqq x_2 \geqq x_3 \geqq x_4$, show that

$$6(x_1^2 + x_2^2 + x_3^2 + x_4^2)^2 = \Sigma((x_i + x_j)^4 + (x_i - x_j)^4)$$

summed over the six cases where $i < j$.

EX. *25.7.* Write $n = 6m + r$, $0 \leqq r < 6$; $m = a_1^2 + a_2^2 + a_3^2 + a_4^2$; $a_k = x_{1k}^2 + x_{2k}^2 + x_{3k}^2 + x_{4k}^2$, $k = 1,2,3,4$; $r = r \cdot 1^4$. Then apply $g(2) = 4$ and EX *25.6* to show $15 < g(4) \leqq 53$.

▶*I do not know what I may appear to the world, but to myself I seem to have been only a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.*

—I. NEWTON

CHAPTER 26*

SUM OF TWO SQUARES

26.1. Four lemmas and a theorem. From the preceding chapter it is clear that not every integer may be represented as the sum of two squares, so the object of the present lesson is to establish just which integers may be so represented. The lemmas which follow and their proofs are almost parallel to those of Chapter 25.

L.1: No integer of the form $4m + 3$ is a sum of two squares.

Proof: If we consider a table in which $x \equiv 0,1,2,3 \bmod 4$ implies $x^2 \equiv 0,1,0,1 \bmod 4$, respectively, it is clear that for given integers x and y, we must have $x^2 + y^2 \equiv 0,1,$ or $2 \bmod 4$; whence **L.1** is an immediate consequence.

L.2: If the prime factors of a composite number n may each be written as the sum of two squares, then n is the sum of two squares.

Proof: It is easy to verify the following identity:

(26.1) $(a^2 + b^2)(a_1^2 + b_1^2) = A^2 + B^2,\ A = aa_1 + bb_1,\ B = ab_1 - ba_1.$

From (26.1), applied several times if necessary, it follows that **L.2** is correct.

L.3: If p is a prime of the form $4K + 1$, there exists a solution in integers x,y,m of $x^2 + y^2 = mp$ with $0 < m < p$.

*Chapter 26 is a supplementary chapter.

Proof: In EX. *22.6* we have shown that $(-1/p) = +1$ if $p = 4K + 1$, hence there exists an integer y such that $1 + y^2 \equiv 0 \bmod p$. We may find $Y \equiv \pm y \bmod p$ and such that $|Y| < p/2$, then $0 < mp = 1 + Y^2 < 1 + p^2/4 < p^2$, so $0 < m < p$. Thus the integers $1, Y$, and m satisfy the conclusions of **L.3**.

L.4: If p is a prime of the form $4K + 1$ and if $x^2 + y^2 = mp$ with $1 < m < p$, then there exist integers x_1, y_1 and M such that $x_1^2 + y_1^2 = Mp$ with $1 \leqq M < m$.

Proof: If m is *even*, we must have $x \equiv y \bmod 2$ and we may rewrite the equation of the hypothesis in the form

$$((x + y)/2)^2 + ((x - y)/2)^2 = (m/2)p$$

to see that $x_1 = (x + y)/2$, $y_1 = (x - y)/2$, $M = m/2$ satisfy the conclusions of **L.4**.

If m is *odd*, we can use a modified division algorithm to write

$$x = am + a_1,\ |a_1| < m/2; \quad y = bm + b_1,\ |b_1| < m/2.$$

If these expressions are substituted in the given equation, we find, using the symbols of (*26.1*), that

$$a_1^2 + b_1^2 + 2Am + (a^2 + b^2)m^2 = mp.$$

Hence it follows that there is a non-negative integer M such that $a_1^2 + b_1^2 = Mm$, and we may write

$$M + 2A + (a^2 + b^2)m = p,$$

$$M^2 + 2AM + (a^2 + b^2)(a_1^2 + b_1^2) = (M + A)^2 + B^2 = Mp.$$

If $M = 0$, we would have $a_1 = b_1 = 0$, so that m^2 would divide $x^2 + y^2 = mp$ and m would divide p. Since p is a prime and $1 < m < p$, this is a contradiction. Hence we have $1 \leqq M$. But also $Mm = a_1^2 + b_1^2 < m^2/2 < m^2$, so $M < m$. Thus $x_1 = M + A$, $y_1 = B$, and M are integers satisfying the conclusions of **L.4**.

Theorem: Every prime of the form $4K + 1$ can be represented as the sum of two squares.

Proof: By **L.3** we may find integers x,y so that $x^2 + y^2 = mp$, $1 \leqq m < p$. In case $m > 1$, we may apply **L.4** a finite number of times (say with $m > M = M_1 > M_2 > \ldots > M_k = 1$) to "descend" to the situation where $x_k^2 + y_k^2 = p$.

26.2. Representation as a sum of two squares. In the preceding section **L.1** shows that no prime of the form $4K + 3$ is the sum of two squares. But since, for example, the product n of two such primes, say $n = (4K + 3)(4K_1 + 3)$, is of the form $4T + 1$, further investigation is required to see if such an n is representable as the sum of two squares. The answer, for this example, turns out to be yes, if $K = K_1$; and no, if $K \neq K_1$. The general case is discussed in what follows.

Let us say that n has a *proper* representation as the sum of two squares if and only if there exist relatively prime integers x and y such that $n = x^2 + y^2$.

Theorem 1: If n is divisible by a prime p of the form $4K + 3$, then n has no proper representation as the sum of two squares.

Proof: The proof is by contradiction. Suppose there is a proper representation $n = x^2 + y^2$, $(x,y) = 1$. Then we must have $(x,p) = 1$. Otherwise, we would have p dividing x and n, and hence y, thus denying $(x,y) = 1$. But with $(x,p) = 1$ we can solve $xu \equiv y \bmod p$ for u as in **G.8** of **19.2**. Then $n \equiv 0 \bmod p$ shows $x^2 + y^2 \equiv x^2(1 + u^2) \equiv 0 \bmod p$. Since $(x,p) = 1$, the cancellation law applies to show $1 + u^2 \equiv 0 \bmod p$. But this is a contradiction, for in EX. *22.6*, we have shown $(-1/p) = -1$ for primes of the form $p = 4K + 3$.

Theorem 2: If $n = p^c m$, where p is a prime of the form $4K + 3$, where c is *odd* and $(p,m) = 1$, then n has no representation as the sum of two squares.

Proof: The proof is by contradiction. Suppose there is a representation $n = x^2 + y^2$. Let $(x,y) = d$, $x = Xd$, $y = Yd$. Then $(X,Y) = 1$ and $n = Nd^2$. Since p^c divides n and c is *odd*, it follows that p divides N. But $N = X^2 + Y^2$ with $(X,Y) = 1$, and to have N with such a proper representation, yet divisible by a prime p of the form $4K + 3$, is a contradiction of the preceding *Theorem 1*.

Theorem 3: A positive integer is representable as the sum of two squares if and only if each of its prime factors of the form $4K + 3$ appears to an even power.

Proof: (A) For the unit 1, we have $1 = 1^2 + 0^2$. For the only even prime 2, we have $2 = 1^2 + 1^2$. For every prime of the form $4K + 1$ a representation as the sum of two squares exists by the theorem of **26.1**. An even power p^{2s} of a prime of the form $p = 4K + 3$ is a sum of two squares since $p^{2s} = (p^s)^2 + 0^2$. Then by **L.2** of **26.1**, every composite number n in which prime factors of the form $4K + 3$ appear only to even powers is representable as a sum of two squares. This includes the case where such prime factors are *absent*, if we interpret $p^0 = 1$ with the zero exponent as an even power.

(B) If even one prime of the form $4K + 3$ appears to an odd power, and not to a higher power, as a factor of n, then n is not representable as the sum of two squares; for this is the content of the preceding *Theorem* 2.

For example, in our previous examination of $n = 351$ no representation as the sum of two squares was found. This could have been predicted since $351 = 3^3 13$ with the prime 3 appearing to an odd power. On the other hand, we have examples like $117 = 3^2 13 = 9^2 + 6^2$ (where only improper representations are available, see *Theorem 1*) and $65 = 5 \cdot 13 = (1^2 + 2^2)(2^2 + 3^2) = 8^2 + 1^2$ (to illustrate **L.2**).

An elegant result, due to Jacobi and discussed in Uspensky and Heaslet, shows how to enumerate the representations of n as the sum of two squares, distinguishing $(-x)^2$ from $(+x)^2$ and considering order. Jacobi considered the positive divisors of n separated into four classes according to their residues 1,2,3,4 mod 4 and indicated the number of divisors in each of these classes by $\tau_1(n)$, $\tau_2(n)$, $\tau_3(n)$, $\tau_4(n)$, respectively. (In this notation the $\tau(n)$ of Chapter 8 would be given by $\tau(n) = \tau_1(n) + \tau_2(n) + \tau_3(n) + \tau_4(n)$.) Jacobi showed that there are $A(n) = 4(\tau_1(n) - \tau_3(n))$ representations of n as a sum of two squares.

For example, if $n = 351 = 3^3 13$, we find $\tau_1(n) = 4$ and $\tau_3(n) = 4$, corresponding to the sets of divisors 1,9,13,117 and 3,27,39,351, respectively; so $A(n) = 0$ and there are no representations. If $n = 72 = 2^3 3^2$, we have $\tau_1(n) = 2$ and $\tau_3(n) = 1$, corresponding to the sets of divisors 1,9 and 3, respectively; so $A(n) = 4$, the appropriate representations being $(\pm 6)^2 + (\pm 6)^2$. If $n = 65 = 5 \cdot 13$, then $\tau_1(n) = \tau(n) = 4$; so $A(n) = 16$, the appropriate representations being $(\pm 8)^2 + (\pm 1)^2$, $(\pm 1)^2 + (\pm 8)^2$, $(\pm 7)^2 + (\pm 4)^2$, $(\pm 4)^2 + (\pm)7^2$.

It is more difficult to discuss in entirety the result concerning representation as the sum of three squares, although EX. 25.2 establishes the easier part of the proof. The correct theorem is that a positive integer n is the sum of three squares of integers if and only if n is not of the form $4^k(8m + 7)$, $k \geqq 0$. Expositions of this result can be found in the books of Dickson or in Uspensky and Heaslet.

26.3. Representation as the difference of two squares. We shall let $Q(n)$ indicate the number of solutions of the Diophantine equation $x^2 - y^2 = n$, where n is a given positive integer and we require x and y to be positive integers.

Theorem: (a) If $n \equiv 2 \bmod 4$, then $Q(n) = 0$.
(b) If $n \equiv 1$ or $n \equiv 3 \bmod 4$, then $Q(n) = [\tau(n)/2]$.
(c) If $n \equiv 0 \bmod 4$, then $Q(n) = [\tau(n/4)/2]$.

Proof: (a) We note that according as $x \equiv 0,1,2,3 \bmod 4$, we have $x^2 \equiv 0,1,0,1 \bmod 4$. Hence for any given integers x and y, we must have $x^2 - y^2 \equiv 0,1,$or $3 \bmod 4$; but we cannot have $x^2 - y^2 \equiv 2 \bmod 4$. Therefore $Q(n) = 0$ when $n \equiv 2 \bmod 4$.

A solution $x > 0$, $y > 0$ of $n = x^2 - y^2 = (x + y)(x - y)$ implies a factorization of n in the form $n = dd'$ where $d = x + y$ and $d' = x - y$ so that $d + d' = 2x$ and $d - d' = 2y$. It follows that $d > d' > 0$ and $d \equiv d' \bmod 2$. Conversely, for every factorization $n = dd'$ with $d > d' > 0$ and $d \equiv d' \bmod 2$, there is a solution $x = (d + d')/2$ and $y = (d - d')/2$ of the Diophantine equation with $x > 0$ and $y > 0$.

(b) If $n \equiv 1$ or if $n \equiv 3 \bmod 4$, then n is odd and both d and d' must be odd so $d \equiv d' \bmod 2$ is satisfied. If n is not a square, every factorization $n = dd'$ has $d \neq d'$. There are $\tau(n)$ choices for d, where $\tau(n)$ is even (Ex. *8.3*); and exactly $\tau(n)/2$ choices of d with $d > d' > 0$. Hence as explained above, there are the same number of solutions in positive integers of the Diophantine equation; so $Q(n) = \tau(n)/2$. If n is a square, there is one, but only one, factorizarion $n = dd'$ in which $d = d'$, which would *not* lead to a suitable solution with $y > 0$. In this case $\tau(n)$ is odd, so the number of suitable factorizations of $n = dd'$ with $d > d' > 0$, each leading to a solution of the Diophantine equations and all solutions so obtainable, is given by $Q(n) = (\tau(n) - 1)/2$. The two cases are readily combined by writing $Q(n) = [\tau(n)/2]$.

(c) If $n \equiv 0 \bmod 4$, then n is even and if $n = dd'$, then one, at least, of d and d' must be even; then in order to satisfy $d \equiv d' \bmod 2$, both d and d' must be even, say $d = 2D$, $d' = 2D'$. Then the number of solutions of the Diophantine equation depends exactly on the number of factorizations $n = 4K = (2D)(2D')$, or $n/4 = K = DD'$ $D > D' > 0$. As in part (*b*), if K is not a square, we find $Q(n) =$

$\tau(K)/2$; but if K is a square, $Q(n) = (\tau(K) - 1)/2$. Both cases are correctly described by $Q(n) = [\tau(n/4)/2]$.

For example, with $n = 351 = 3^3 13 \equiv 3 \bmod 4$, since $\tau(n) = 8$, we find $Q(n) = 4$ solutions. Corresponding to the factorizations $351 \cdot 1$, $117 \cdot 3$, $39 \cdot 9$, $27 \cdot 13$, the solutions x,y are 176,175; 60,57; 24,15; 20,7.

EXERCISES

EX. *26.1.* Show that (*26.1*) is a special case of (*25.1*).

EX. *26.2.* Find all representations as a sum of two squares for (a) 209, (b) 221, (c) 1225.

EX. *26.3.* Find all representation as a difference of two squares for (a) 426, (b) 427, (c) 428, (d) 429.

EX. *26.4.* If a is a given positive integer and positive integers b,c are required so that

$$b^2 - a^2 = c(c + 1),$$

prove that the number $N(a)$ of solutions b,c is given by

$$N(a) = \frac{\tau(4a^2 - 1)}{2} - 1.$$

EX. *26.5.* Find the number of Pythagorean triplets of a given *side*.

▶ *A scientist worthy of the name, above all a mathematician, experiences in his work the same impressions as an artist; his pleasure is as great and of the same nature.*

—H. POINCARE

CHAPTER 27[*]

INTRODUCTION TO QUADRATIC FORMS

27.1. Equivalent functions. It will be evident that this chapter presents generalizations of the material in Chapter 26, but it may be necessary for the reader to review carefully the ideas of Chapter 11 before proceeding, in particular **11.3** and **11.4**.

In the present section "function" will be used to mean a polynomial $f(x,y)$ in two variables with integers as coefficients; in other words, $f(x,y)$ is the sum of a finite number of terms of the type rx^sy^t, where r is an integer and s and t are non-negative integers.

A function $f(x,y)$ will be said to *represent* an integer n if and only if there exists a pair of integers x,y (i.e., a lattice point of S_2 as in **11.3**) such that $f(x,y) = n$. An integer n will be said to be *properly* represented if and only if there is a representation $f(x,y) = n$ in which x and y are relatively prime. If a function is such that it represents every integer, the function will be called *universal*.

A function $F(X,Y)$ will be said to be *equivalent* to a function $f(x,y)$ if and only if there exists a linear transformation T of the lattice group G of S_2, say,

$$T: \quad x = aX + bY, \quad y = cX + dY, \quad ad - bc = \pm 1,$$

such that $f(x,y) \underset{T}{=} F(X,Y)$.

[*] Chapter 27 is a basic chapter.

One motivation for this terminology is provided by the following theorem.

F.1: Equivalent functions represent the same integers.

Proof: We have shown in **M.5** and **M.6** of **11.3**, and in the discussion of **11.4**, how the linear transformations of S_2 which are completely reversible in integers are precisely those of the lattice group of S_2, namely, those of unit determinant. The definition of equivalent functions is phrased to take advantage of this property. For if $f(x,y) \underset{T}{=} F(X,Y)$ and if X, Y are integers such that $F(X,Y) = n$, then the integers x,y defined by $(x,y) = (X,Y)T$ are such that $f(x,y) = n$. Conversely, since T is of unit determinant, there exists an inverse transformation:

$$T^{-1}: \quad X = dx - by, \quad Y = -cx + ay, \quad da - cb = \pm 1$$

such that $F(X,Y) \underset{T^{-1}}{=} f(x,y)$. Hence, if x,y are integers such that $f(x,y) = m$, then the integers X,Y defined by $(X,Y) = (x,y)T^{-1}$ are such that $F(X,Y) = m$. Combining these observations, we find that the totalities of integers represented by equivalent functions $f(x,y)$ and $F(X,Y)$ are exactly the same.

For example, if we extend the discussion of **26.3** to all integers x and y, we find that $f(x,y) = x^2 - y^2$ represents all integers n, except those for which $n \equiv 2 \bmod 4$. Using T: $x = 2X - 3Y$, $y = X - Y$ which has determinant $+1$, we find

$$f(x,y) \underset{T}{=} (2X - 3Y)^2 - (X - Y)^2 = 3X^2 - 10XY + 8Y^2 = F(X,Y).$$

By **F.1** we can assert that $F(X,Y)$ also represents all integers n, except those for which $n \equiv 2 \bmod 4$. Thus from $f(13,7) = 120$, we can compute $(13,7)T^{-1} = (8,1)$ and assert that $F(8,1) = 120$.

F.2: Equivalence of functions is an equivalence relation.

Proof: The proof follows closely the known properties of the lattice group G of S_2 as given in **M.6** of **11.3**. With these group properties proved it is easy to establish that equivalence of functions has the four properties of an equivalence relation.

(1) *Determinative:* given $f(x,y)$ and $F(X,Y)$, either there is or is not a T of G such that $f(x,y) \underset{T}{=} F(X,Y)$.

(2) *Reflexive:* the group G contains I and $f(x,y) \underset{I}{=} f(x,y)$.

(3) *Symmetric:* given that $f(x,y) \underset{T}{=} F(X,Y)$ for a T in G, then there is T^{-1} in G such that $F(X,Y) \underset{T^{-1}}{=} f(x,y)$.

(4) *Transitive:* given that $f(x,y) \underset{T}{=} F(X,Y)$ and that $F(X,Y) \underset{U}{=} F_1(X_1,Y_1)$ with T and U in G, then TU is in G and is such that $f(x,y) \underset{TU}{=} F_1(X_1,Y_1)$.

From **F.2** it follows that equivalence of functions divides all functions into mutually exclusive classes of equivalent functions. From **F.1** it follows that all the functions in such an equivalence class represent exactly the same integers. Now to follow out the program outlined in **11.4**, we should seek for each equivalence class some representative, characterized by its simplicity and, if possible, so described that it is canonical, i.e., so that in an equivalence class there is one and only one function of this description.

In the next section we shall consider certain equivalence classes for which this program can be achieved.

27.2. Positive definite binary quadratic forms. A *binary quadratic form* is a special function of the type

$$f(x,y) = ax^2 + bxy + cy^2$$

where, of course, a,b,c, are given integers, and since they completely determine the form, the abbreviation $f = [a,b,c]$ is convenient.

F.3: A function equivalent to a binary quadratic form is a binary quadratic form.

Proof: Let T be a linear transformation of the lattice group G of S_2 defined by

$$T\colon\quad x = a_1X + b_1Y, \quad y = c_1X + d_1Y, \quad a_1d_1 - b_1c_1 = \pm 1.$$

Then by substitution and expansion we find $[a,b,c] \underset{T}{=} [A,B,C]$,

$$(27.1)\begin{cases} A = aa_1^2 + ba_1c_1 + cc_1^2 = f(a_1,c_1), \\ B = 2aa_1b_1 + b(a_1d_1 + b_1c_1) + 2cc_1d_1 = f(a_1 + b_1, c_1 + d_1) - A - C, \\ C = ab_1^2 + bb_1d_1 + cd_1^2 = f(b_1,d_1). \end{cases}$$

Since A,B,C are integers, $[A,B,C]$ is a binary quadratic form.

The *discriminant* of a binary quadratic form a,b,c is defined to be the integer $b^2 - 4ac$.

F.4: Equivalent binary quadratic forms have the same discriminant.

Proof: A straightforward, but tedious simplification, starting from (*27.1*), will show that $B^2 - 4AC = b^2 - 4ac$. However, a simpler proof is obtained by writing (*27.1*) in matric form, as follows:

$$(27.2) \qquad \begin{pmatrix} a_1 & c_1 \\ b_1 & d_1 \end{pmatrix}\begin{pmatrix} b & 2a \\ 2c & b \end{pmatrix}\begin{pmatrix} d_1 & c_1 \\ b_1 & a_1 \end{pmatrix} = \begin{pmatrix} B & 2A \\ 2C & B \end{pmatrix} .$$

If to the matric equation (*27.2*) we apply **M.7** and note that even though matric multiplication is not, in general, commutative, the determinants are commutative in their multiplication, then since $(a_1d_1 - b_1c_1)^2 = +1$, we find, rather elegantly, that $b^2 - 4ac = B^2 - 4AC$.

For example, we showed above that $[1,0,-1]$ and $[3,-10,8]$ are equivalent. To illustrate **F.4** we can now check that $0^2 - 4(1)(-1) = 4 = (-10)^2 - 4(3)(8)$.

Let us describe a form as *positive definite* if it represents, in addition to zero, positive and only positive integers. If there are such forms, then it follows by **F.1** that all forms equivalent to a positive definite form are also positive definite.

F.5: A form $f = [a,b,c]$ is positive definite if and only if
$$a \geqq 0, \quad c \geqq 0, \quad a^2 + c^2 > 0, \quad b^2 - 4ac \leqq 0.$$

Proof: (A) It is clear that a positive definite form must not have $a < 0$, for then $f(1,0) = a$ would be a negative integer represented by the form; similarly, since $f(0,1) = c$, it follows that a positive definite form must not have $c < 0$. If $a = b = c = 0$, the form represents only zero and is not positive definite; if $a = c = 0$ and $b \neq 0$, then $f(x,y) = bxy$ so that $f(1,1) = b$ and $f(1,-1) = -b$, so such a form is not positive definite; therefore a positive definite form must have at least one of a and c positive, i.e., $a^2 + c^2 > 0$. From $f(b,-2a) = -a(b^2 - 4ac)$ it follows that even with $a > 0$, it is necessary to have $b^2 - 4ac \leqq 0$ to make f positive definite. If $a = 0$, but $c > 0$, then $f(-2c,b) = -cb^2$, hence it is necessary to have $b^2 - 4ac = b^2 = 0$ to make f positive definite. Both of these cases are covered by the requirement $b^2 - 4ac \leqq 0$.

(B) Conversely, if the conditions mentioned in **F.5** are satisfied, then $f(x,y) = [a,b,c]$, in addition to zero, represents positive and only positive integers. First assume $a > 0$, then $f(1,0) = a$ shows that f represents at least one positive integer. Secondly, from the following identity:

(27.3) $4af(x,y) = 4a(ax^2 + bxy + cy^2) = (2ax + by)^2 - (b^2 - 4ac)y^2$,

we see, since by hypothesis $b^2 - 4ac \leqq 0$, that the right-hand member of the identity is non-negative; since $a > 0$, it follows that f is non-negative. Finally, if $a = 0$, then $b = 0$ and $c > 0$, so that $f(x,y) = cy^2$ is obviously positive definite.

We will deal henceforth with positive definite forms and for each equivalence class of these forms we can establish a canonical form as described in the following two theorems and EX. *27.9*.

F.6: Any given positive definite form $f = [a,b,c]$ with $b^2 - 4ac < 0$ is equivalent to a form $F = [A,B,C]$ in which

(27.4) $$0 \leqq B \leqq A \quad \text{and} \quad 0 < A \leqq C.$$

Proof: By the hypothesis $b^2 - 4ac < 0$ and by **F.5** it follows that both a and c are positive. Both are represented by f. Hence there is a least positive integer A represented by f and an upper limit on its value is already available. By (*27.3*), or its analogue for $4cf$, the value of A can be found in a finite number of steps. Such an A has a proper representation by f. For if we suppose $f(x_0,y_0) = A$ with $(x_0,y_0) = d$ and $x_0 = X_0d$, $y_0 = Y_0d$, then $f(X_0,Y_0) = A/d^2$. Hence if $d > 1$, then A/d^2 would be a positive integer less than A represented by f. Therefore we must have $(x_0,y_0) = 1$. By the Euclid algorithm we know there exist integers r and s so that $rx_0 - sy_0 = 1$. Then T: $x = x_0X + sY$, $y = y_0X + rY$, is a linear transformation belonging to the lattice group G of S_2, for its coefficients are integers and its determinant $+1$. T transforms $[a,b,c]$ into the equivalent form $[A,B',C']$.

The transformation U_q: $X = X_1 + qY_1$, $Y = Y_1$, where q is an integer, is also in the lattice group of S_2, and U_q transforms $[A,B',C']$ into the equivalent form $[A,B^*,C]$ where $B^* = 2qA + B'$. Thus by a suitable choice of q we can make B^* a least absolute value residue of B' mod $2A$; i.e., $-A < B^* \leqq A$.

If $B^* < 0$, then the transformation V: $X_1 = X_2$, $Y_1 = -Y_2$, is in the lattice group of S_2, and V transforms $[A,B^*,C]$ into $[A,B,C]$ where $B = -B^*$. If $B^* \geqq 0$, we set $B = B^*$. Hence in both cases we have $0 \leqq B \leqq A$.

By the transitive property in **F.2** it follows that $[a,b,c]$ is equivalent to $[A,B,C]$. By **F.4** it follows that these forms have the same discriminant. By the hypothesis $B^2 - 4AC = b^2 - 4ac < 0$. Then since $A > 0$, it follows that $C > 0$. Furthermore by **F.1**, A is the least positive integer represented by $[a,b,c]$ and by $[A,B,C]$; hence $A \leqq C$.

This completes the proof of the theorem. A form with the properties *(27.4)* will be called a *reduced* form.

For example, if given $f = [4,-27,48]$, we check that $b^2 - 4ac = -39$, so the form is positive definite and of negative discriminant. It is clear that $A \leqq 4$. From *(27.3)* we write

$$16A = (8x - 27y)^2 + 39y^2 \leqq 64$$

which requires $y = 0$ or $y = 1$. The first case requires $x = 0$ and $A = 0$; or $x = 1$ and $A = 4$. The second case requires both $(8x - 27)^2 \leqq 25$ and $(8x - 27)^2 \equiv 9 \bmod 16$ so that only two solutions are found: either $x = 3$ and $A = 3$; or $x = 4$ and $A = 4$. Thus the correct value of A is 3 with the proper representation $f(3,1) = 3$. The transformation T: $x = 3X + 2Y$, $y = X + Y$, takes $[4,-27,48]$ into $[3,-9,10]$. Since $-9 = 6(-2) + 3$, we may use U_2: $X = X_1 + 2Y_1$, $Y = Y_1$, to transform $[3,-9,10]$ into $[3,3,4]$ which is a reduced form. The next theorem guarantees that we cannot, by some other sequence of transformations, arrive at any other reduced form.

F.7: In each class of equivalent positive definite forms of negative discriminant there is one and only one reduced form.

Proof: By **F.6** there is at least one reduced form in every class of equivalent positive definite forms of negative discriminant. Let us suppose that $[a,b,c]$ and $[A,B,C]$ are two equivalent reduced forms, each satisfying the conditions *(27.4)* so $0 \leqq b \leqq a$, $0 < a \leqq c$; $0 \leqq B \leqq A$, $0 < A \leqq C$. Let us suppose that these forms are equivalent under a transformation T such that the relations *(27.1)* hold.

It is no restriction to assume $a \geqq A$. From $(a_1 \pm c_1)^2 \geqq 0$ it follows that $a_1^2 + c_1^2 \geqq 2|a_1c_1|$. From $0 \leqq b \leqq a$, whether $a_1c_1 \geqq 0$ or $a_1c_1 < 0$, we have $ba_1c_1 \geqq -a|a_1c_1|$. Then since $c \geqq a$, we may use *(27.1)* to see that

$$A = aa_1^2 + ba_1c_1 + cc_1^2 \geqq 2a|a_1c_1| - a|a_1c_1| = a|a_1c_1|.$$

Hence $a \geqq A \geqq a|a_1c_1|$, so $1 \geqq |a_1c_1|$.

If $|a_1c_1| = 0$, and $a_1 = 0$, then $c_1 \neq 0$, for $a_1d_1 - b_1c_1 = \pm 1$; then

$a \geqq A = cc_1^2 \geqq c \geqq a$ shows $A = a$. A similar argument holds when $|a_1c_1| = 0$ and $c_1 = 0$. If $|a_1c_1| = 1$, the concluding line of the last paragraph shows $a \geqq A \geqq a$, so that $A = a$. Thus in every case we have $A = a$.

If $c = a = A = C$, we may use **F.4** and $b^2 - 4ac = B^2 - 4AC$ to conclude that $b^2 = B^2$. Since b and B are non-negative, it follows that $b = B$, hence in this case the reduced forms are identical.

In the remaining case it is no restriction to consider $c > a$, rather than $C > A$, inasmuch as we have already shown $a = A$. Then the inequality established above is more restrictive and instead of $A \geqq a|a_1c_1|$, we may say $A > a|a_1c_1| = A|a_1c_1|$, so $|a_1c_1| = 0$.

If $a_1 = 0$, then $c_1 \neq 0$, for $a_1d_1 - b_1c_1 = \pm 1$; furthermore, $|b_1c_1| = 1$, so $c_1^2 = 1$. Then $a = A = cc_1^2 = c$, a contradiction of the assumption $c > a$, so this case cannot arise.

If $a_1 \neq 0$, then $c_1 = 0$, and as above $a_1d_1 = +1$ or $a_1d_1 = -1$. From (*27.1*) we may write

$$B = 2aa_1b_1 + b(a_1d_1 + b_1c_1) + 2cc_1d_1 = 2aa_1b_1 + ba_1d_1.$$

If $a_1d_1 = 1$, then $B - b = 2aa_1b_1$ is a multiple of $2a$; but with $0 \leqq b \leqq a$, $0 \leqq B \leqq A = a$, we have $-a \leqq B - b \leqq a$. Hence it follows that $a_1b_1 = 0$; since $a_1 \neq 0$, we have $b_1 = 0$. Either $a_1 = 1$, $d_1 = 1$; or $a_1 = -1, d_1 = -1$. In these cases the corresponding transformations $T = I$ or $T = -I$ are such that $B = b$ and $C = c$.

If $a_1d_1 = -1$, then $B + b = 2aa_1b_1$. Since $0 \leqq B + b \leqq 2a$, we have two cases to consider. If $B + b = 0$, then $B = b = 0$, and also $a_1b_1 = 0$ so that $b_1 = 0$; then either $a_1 = 1, d_1 = -1$; or $a_1 = -1$, $d_1 = 1$; in either case the transformation T thus determined is such that $C = c$. If $B + b = 2a$, then $B = b = a$, and $a_1b_1 = 1$; then either $a_1 = 1, b_1 = 1, d_1 = -1$; or $a_1 = -1, b_1 = -1, d_1 = 1$; in either case the transformation T thus determined is such that $C = ab_1^2 + bb_1d_1 + cd_1^2 = a - b + c = c$.

Thus in every case reduced and equivalent forms have been shown to be identical, which completes the proof of **F.7**.

The final note of clarification is contained in the following theorem, which carries the warning that there may be more than one reduced form of a given discriminant but tempers the warning with words of finiteness.

F.8: There are only a finite number of reduced forms with the same discriminant.

Proof: From the condition (*27.4*) we have $B^2 \leqq A^2 \leqq AC < 4AC$ or $B^2 - 4AC < 0$, so a reduced form is automatically positive definite. Let us set $K = 4AC - B^2 > 0$. Then $4A^2 \leqq 4AC = B^2 - (B^2 - 4AC) \leqq A^2 + K$, so that $3A^2 \leqq K$. Hence if K is fixed, there are only a finite number of choices of A. The conditions $0 \leqq B \leqq A$ and $B^2 \equiv -K \bmod 4A$ show that there are only a finite number of B's to go with a choice of A. As soon as A and B are selected, C is already fixed by $4AC = B^2 + K$. In short, there are only a finite number of reduced forms of a given discriminant, $-K$.

An immediate corollary, of course, is that there are only a finite number of equivalence classes of positive definite forms of a given negative discriminant. For by **F.7** each such class contains one and just one reduced form. Since by **F.8** there are only a finite number of these reduced forms, there are just the same number of classes.

For example, if $K = 6$, then $3A^2 \leq 6$, shows $A = 1$; but since neither $B = 0$ or $B = 1$ solves $B^2 \equiv -6 \bmod 4$, there are *no* reduced forms, and *no* positive definite forms, of discriminant -6. But if $K = 7$, then $3A^2 \leqq 7$, shows $A = 1$, and $B = 1$ (but not $B = 0$) solves $B^2 \equiv -7 \bmod 4$; hence there is one, and just one, class of positive definite forms of discriminant -7, and this class is represented by its only reduced form [1,1,2].

With these ideas as background Hermite gave a simple proof of the theorem of **27.1**, which we restate as follows:

F.9: Every prime of the form $4K + 1$ can be represented as the sum of two squares.

Proof: Since -1 is a quadratic residue of the prime $p = 4K + 1$, there are integers s and t such that $s^2 + 1 = tp$. Hence the form $[t,2s,p]$ with $p > 0$, $t > 0$, and discriminant $(2s)^2 - 4tp = -4$ is positive definite. When $K = 4$, we have $3A^2 \leqq 4$, so $A = 1$; then from $B^2 \equiv -4 \bmod 4$, only the solution $B = 0$ satisfies $0 \leqq B \leqq A$; and for this solution $C = 1$. Thus there is one and only one reduced form of discriminant -4 and it is [1,0,1]. By **F.7** it follows that $[t,2s,p]$ and [1,0,1] are equivalent, for they have the same discriminant and there is only the one reduced form with this discriminant. By **F.1** these equivalent forms represent the same integers. But it is clear that $F(X,Y) = [t,2s,p]$ represents p, inasmuch as $F(0,1) = p$. Hence it follows that [1,0,1] represents p. However, $f(x,y) = [1,0,1] = x^2 + y^2$ is the familiar sum of two squares, now written as a binary quadratic form. Hence p may be written as the sum of two squares.

The literature about quadratic forms and universal functions is extensive. But among modern writers, few, except Dickson, include the topic because Dickson and his students have written so extensively on the subject; so it is to this author's books we refer the student who may wish to pursue the subject in greater detail.*

*A recent book: B. W. Jones, *The Arithmetic Theory of Quadratic Forms*, Carus Monograph No. 10, New York, Wiley, 1950.

EXERCISES

EX. *27.1.* Prove that $f(x,y) = x^3 + y^3$ is equivalent to $F(X,Y) = 37X^3 - 90X^2Y + 72XY^2 - 19Y^3$.

EX. *27.2.* (a) Using $T: x = 5X + 2Y, y = 7X + 3Y$, find the form which is T-equivalent to $f(x,y) = [3,5,1]$.

(b) Check that the equivalent forms of part (a) do have the same discriminant.

EX. *27.3.* Decide which of the following forms are positive definite: (a) $4xy$; (b) $x^2 + 3xy + 2y^2$; (c) $-x^2 + 3xy - 12y^2$; (d) $x^2 + 3xy + 3y^2$.

EX. *27.4.* Find all the reduced forms of discriminant -104.

EX. *27.5.* Find the reduced form equivalent to

$$37x^2 - 194xy + 255y^2.$$

EX. *27.6* Define $f(x,y)$ to be *strictly* equivalent to $F(X,Y)$ if and only if $f(x,y) \underset{T}{=} F(X,Y)$ where $T: x = aX + bY, y = cX + dY$ has $ad - bc = +1$. Show that strict equivalence of functions is an equivalence relation.

EX. *27.7.* Show that any given positive definite binary quadratic form $[a,b,c]$ of negative discriminant is strictly equivalent to a reduced form A,B,C where either (1) $0 < A < C$, $-A < B \leqq A$; or (2) $0 < A = C$, $0 \leqq B \leqq A$.

EX. *27.8.* Show that two reduced forms of the type described in EX. *27.7* which are strictly equivalent are identical.

EX. *27.9.* Show that in each class of equivalent positive definite forms of zero discriminant there is one and only one form $[A,0,0]$ with $0 < A$.

▶*Mathematics is the science which draws necessary conclusions.* —B. PIERCE

CHAPTER 28*

PEANO'S AXIOMS FOR THE NATURAL INTEGERS

28.1. Concerning mathematical systems. The reader has probably become acquainted with the postulational method in mathematics by a study of plane geometry, but if he has had only the traditional courses in algebra he may never have realized that algebra, too, is susceptible of such a postulational treatment. Historically, this is understandable, for geometry has been regarded abstractly for over twenty centuries, while algebra has been so viewed for scarcely one century. But there was a revolution in attitude toward the axiomatic basis of geometry at the beginning of the nineteenth century; and the revolution spread out to cause a study of the foundations of all branches of mathematics.

In order to describe how the postulational method touches the theory of numbers, it will be convenient at the outset to have a definition of a general mathematical system. All such definitions have their faults, being criticized as either too general or too restrictive, but the following one seems quite useful.

A mathematical system is the resultant of the application of a system of logic to a set of elements, relations, and operations whose

*Chapter 28 is a supplementary chapter.

properties are described by a consistent set of postulates. If the elements, relations, and operations are left *undefined*, except that they are *assumed* to satisfy the postulates, then the system may be described as a *pure* or *abstract* mathematical system. If the elements, relations, and operations are *defined* in terms of previously studied concepts and the postulates are *proved* to hold, then the system may be described as an *applied* mathematical system or as a *concrete example*.

A student with some mathematical experience will sense that in the development of a mathematical system both the abstract and the concrete approaches are worth while. From a concrete example one obtains suggestions about theorems that may hold in the abstract system; but if the theorem can be proved in the abstract form (and such a proof is sometimes easier, being free of distracting special details found in the example), then it holds for *all* the concrete examples without any further special investigations.

It may be helpful to discuss in detail some of the terms used in the above definition.

By the word *resultant* we imply that not only the elements, relations, operations, and postulates shall be thought of as part of the system, but also all propositions that can be derived as a formal logical consequence from the postulates. One natural way in which such a study of *all* propositions may be limited is suggested in a later paragraph. We should speak of derived propositions as being *valid*, rather than true, to remind ourselves that they can be no more "true" than the originally assumed postulates or the previously studied systems.

In ordinary mathematics we use the Aristotelian system of logic with the following basic laws:

(1) Law of the identity: A is A, a thing is itself.

(2) Law of the excluded middle: either A or not -A, a proposition is valid or is not-valid, there being no other value that can be assigned to it;

(3) Law of contradiction: not both A and not -A, a proposition is not both valid and not-valid.

But in our definition we have used the phrase, "a" system of logic, because today there is study of systems of logic in which there are more than two "truth values" to be considered for each proposition, and the mathematical systems developed with such systems of logic may be considered a part of mathematics.

In any definition, to avoid circular reasoning, certain basic ideas must be left undefined. For example, the ideas of an *element*, a *set* of elements, and of an element *belonging to* a set of elements are of this fundamental nature.

An important example of a relation in a mathematical system is an equivalence relation, a concept which we have already described at length in **17.2.** If we consider the properties required for an equivalence relation as postulates, then we already have at hand a simple type of mathematical system.

A possible synonym for the word operation is the word function. For example, a rule which determines for each element a of a set S a corresponding element $b = f(a)$ of S is an example of a *unary* operation. A rule which determines for each ordered pair of elements a,b of S a corresponding element $c = f(a,b)$ of S is an example of a *binary* operation. In similar manner, we may define functions of three or more variables. Operations are usually required to be closed and well defined: thus a binary operation $f(a,b)$ is said to be closed if $f(a,b)$ is in S for every a and b in S and is said to be well defined (with respect to a specified equivalence relation) if $a = a'$ and $b = b'$ imply $f(a,b) = f(a',b')$ for all a,b in S. Similar definitions apply when unary, ternary, and other operations are being considered. Some of these ideas have already been well illustrated in **17.3.**

The mathematical systems studied in modern algebra may be characterized, at least roughly, as those in which are present operations resembling some of the familiar operations of addition, multiplication, subtraction, and division.

The set of postulates in a mathematical system is, of course, man-made, sometimes suggested by external situations, sometimes pure invention; but in such a selection complete arbitrariness is not allowed, for the set of postulates is always required to be *consistent*. This term means that in the propositions derived from the postulates there must never appear a violation of the law of contradiction, i.e., it must never happen that a proposition is both valid and not-valid. The only known test for the consistency of a set of postulates is the exhibition of at least one example satisfying all the postulates.

It is desirable, although not absolutely necessary, that a set of postulates be such that no postulate can be derived as a theorem from the other postulates; a set of postulates with this property is said to *independent*. The test for independence of a set of postulates is the exhibition of as many examples as there are postulates, one for each postulate, with the property that the example does not satisfy that particular postulate but does satisfy all the other postulates, hence the postulate in question could not possibly be derived as a theorem from the other postulates.

To describe one further property of a set of postulates it is first necessary to explain just when two examples satisfying a set of postulates shall be considered distinct. It is reasonably clear that if the elements $a_1,b_1,\ldots$ of example S_1 in which there is, say, a binary operation $f_1(a_1,b_1)$, may be paired off in a "one-to-one" manner (see below) with the elements $a_2,b_2,\ldots$ of a second example S_2 in which there is a corresponding operation $f_2(a_2,b_2)$, and the pairing off is of such a nature that whenever a_1 and a_2, b_1 and b_2 are corresponding elements, it then follows that $f_1(a_1,b_1)$ and $f_2(a_2,b_2)$ are corresponding elements, then the examples S_1 and S_2, although distinct in the sense that their elements and operations have different names, are "abstractly the same" or, to use technical terms (see below), they are "*isomorphic* with respect to the operations f_1 and f_2."

An important goal in the study of a mathematical system is the characterization of all the non-isomorphic examples satisfying the abstractly defined system. If it happens that all the possible examples are isomorphic, then the set of postulates is described as *categorical*. This property is interesting, but not usually desirable, since it much restricts the realm of applications of the abstractly obtained results.

To make the notion of isomorphism more precise let us consider the general matter of *mapping* one system S_1 *into* a second system S_2 (which may be the system S_1 itself). Such a mapping is simply a correspondence, which we indicate with a letter F, by which *each* element x_1 of S_1 is made to "correspond" to a *unique* element $x_2 = x_1F$ of S_2. If every element of S_2 appears as the map of some element of S_1, then the mapping is from S_1 *onto* S_2. If there is no restriction on the number of elements of S_1 which map onto a given element of S_2, the mapping is *many-to-one*. A mapping of S_1 *onto* S_2 in which *only one* element of S_1 maps onto each element of S_2 is described as *one-to-one*.

If there is in the system S_1 an operation, say $f_1(x_1,y_1)$, and in the system S_2 an operation on the same number of elements, say $f_2(x_2,y_2)$, then a one-to-one mapping F from S_1 into S_2 is called an "isomorphism of S_1 and S_2 with respect to the operations f_1 and f_2" if and only if

$$f_1(x_1,y_1)F = f_2(x_1F,y_1F) \qquad \text{for all } x_1,y_1 \text{ in } S_1.$$

This critical property for the one-to-one mapping to qualify as an isomorphism may be called the "operation-preserving" property.

If there is a one-to-one mapping of S_1 onto S_2, then S_1 and S_2 are said to be *equivalent* sets. Equivalence of sets is an equivalence relation, and if we note the separation into mutually exclusive classes produced by any equivalence relation, then we may appreciate the following definition due to B. Russell: the set of all sets equivalent to a given set shall be called the *cardinal number* of the set.

A *subset* T of a set S is a set of elements all of which belong to S. A *proper subset* T of a set S is a subset such that there exists at least one element of S not belonging to T. A set S is said to have a *finite* cardinal number if S is not equivalent to any of its proper subsets; but if S is equivalent to a proper subset of itself, then S is said to have an *infinite* cardinal number.

28.2. Peano's axioms. Among the studies of mathematical systems, one of the most vital is an abstract description of the natural integers, i.e., the positive integers or counting numbers or finite cardinal numbers, for they are building blocks for almost all other more complicated mathematical systems. The following description, due essentially to G. Peano (1858-1932), is only one of several possibilities, but it seems like a natural one for us to pursue, at least briefly, for we shall find that an essential part of Peano's axioms is the very postulate of mathematical induction with which we have been concerned in previous chapters.

We do not, however, think that this is the place for an exhaustive study of the consequences that can be drawn from Peano's simple assumptions, in fact we shall not go far enough even to show just how all of the postulates come into play. Rather what we want is to give the reader an introduction to the subject, enough of the philosophy of assumptions, definitions, and proofs that he can perhaps carry on for himself—and at least appreciate the remarks of the first lesson to the effect that "... these [commutative, associative, distributive, and cancellation] laws may be proved as theorems on the basis of other still simpler postulates."

According to Peano, a system of natural integers is a set S of elements $a,b,\ldots,$ called "natural integers" or briefly "integers," with an equals relation written $a = b$ or $a \neq b$ (to be read "a equals b" or "a is not equal to b," respectively), and a "sequels" operation, indicated by a' and read "a' is the *sequel* of a" or "a is the *antecedent* of a'," subject to the following postulates:

S.1: There exists an integer called "one," written 1.

S.2: Every integer a has a unique sequel a'.

S.3: The integer 1 has no antecedent.

S.4: If $a' = b'$, then $a = b$.

S.5: If M is a set of integers such that:

(I) M contains 1;

(II) if M contains a, then M contains a';

then M contains all the integers of S.

By way of preliminary comment on the various axioms we note that **S.1** guarantees that the set S is not empty; **S.2** implies that if $a = b$, then $a' = b'$; **S.3** implies $a' \neq 1$; **S.4** says that antecedents are unique; and **S.5** is one form of the principle of mathematical induction.

In the proofs of the following theorems we will adopt the habit of writing under each equality sign the name of the definition, postulate, hypothesis, or previously proved theorem which justifies the equality.

28.3. The operation of addition. In terms of Peano's axioms it is possible to define for the integers of S an ordered binary operation, called addition and written $a + b$, having many familiar properties to justify this terminology. The *definition of addition* may be made as follows:

D.I: We define $1 + b = b'$ for every b in S.

D.II: For every a in S for which

A.I: $a + 1 = a'$, and

A.II: $a + b' = (a + b)'$ for every b in S,

we define $a' + b = (a + b)'$ for every b in S.

Theorem: The operation of addition is closed and well defined.

Proof: Before beginning the proof we should mention that the above definition is due to Grandjot, Landau, and Kalmar who first pointed out in Landau's *Foundation of Analysis* that the definition of Peano which employed only **A.I** and **A.II** is actually incomplete, defined only for a *fixed* a, not for *all* a. Attempts to prove well-definedness, for example, using **A.I** and **A.II** only, are unsuccessful.

(A) Let M be the set of all integers a for which the operations of addition defined by **D.I** and **D.II** has properties **A.I** and **A.II**.

(I) M contains 1: for

$$1 + 1 \underset{\textbf{D.I}}{=} 1'$$

which is **A.I** when $a = 1$; and then

$$1 + b' \underset{\textbf{D.I}}{=} (b')' \underset{\textbf{D.I, S.2}}{=} (1 + b)'$$

which is **A.II** when $a = 1$.

(II) If by hypothesis **H**: M contains a, then M contains a': for

$$a' + 1 \underset{\textbf{D.II}}{=} (a + 1)' \underset{\textbf{H, A.I}}{=} (a')'$$

which is **A.I** for a'; and then

$$a' + b' \underset{\textbf{D.II}}{=} (a + b')' \underset{\textbf{H, A.II}}{=} ((a + b)')' \underset{\textbf{D.II, S.2}}{=} (a' + b)'$$

which is **A.II** for a'.

By (I), (II), and **S.5** it follows that M contains all integers. But with **A.I** and **A.II** holding for *every* a, it follows that $a + b$ is a uniquely defined integer of S for every a and b in S, so the operation of addition is closed.

(B) To show that addition is well defined we suppose **G**: $b = B$ and let M be that set of all integers a for which $a + b = a + B$.

(I) M contains 1, for

$$1 + b \underset{\textbf{D.I}}{=} b' \underset{\textbf{G, S.2}}{=} B' \underset{\textbf{D.I}}{=} 1 + B.$$

(II) If M contains a, so that **H**: $a + b = a + B$, then M contains a', for

$$a' + b \underset{\textbf{D.II}}{=} (a + b)' \underset{\textbf{H, S.2}}{=} (a + B)' \underset{\textbf{D.II}}{=} a' + B.$$

By (I), (II), and **S.5** it follows that M contains all integers.

Similarly (except that we use **A.I** and **A.II** instead of **D.I** and **D.II**, respectively), we may show that if $a = A$, then $a + b = A + b$ for every b in S.

When $a = A$ and $b = B$, we combine the above results and have

$$a + b = a + B = A + B$$

so that addition is well defined.

Attention to this last property is by no means trivial, for in the next proof we find that we often need to make replacements, such as $a + 1$ by a'.

T.1: The *associative* law of addition: $(a + b) + c = a + (b + c)$.

Proof: Let a,b be fixed and let M be the set of all integers c for which **T.1** holds.

(I) M contains 1, for

$$(a + b) + 1 \underset{\text{A.I}}{=} (a + b)' \underset{\text{A.II}}{=} a + b' \underset{\text{A.I}}{=} a + (b + 1).$$

(II) If M contains c, so that **H**: $(a + b) + c = a + (b + c)$, then M contains c', for

$$(a+b) + c' \underset{\text{A.II}}{=} ((a+b) + c)' \underset{\text{H, S.2}}{=} (a + (b+c))' \underset{\text{A.II}}{=} a + (b+c)' \underset{\text{A.II}}{=} a + (b + c').$$

By (I), (II), and **S.5**, M contains all integers, hence **T.1** is always valid.

T.2: The *commutative* law for addition: $a + b = b + a$.

Proof: Let b be fixed and let M be the set of all integers a for which **T.2** holds.

(I) M contains 1. To prove this we use another induction argument, letting N be the set of all integers b for which $1 + b = b + 1$. Clearly, N contains 1, for $1 + 1 = 1 + 1$. Suppose that N contains b, so that **H**: $1 + b = b + 1$. Then N contains b', for

$$1 + b' \underset{\text{A.II}}{=} (1 + b)' \underset{\text{H, S.2}}{=} (b + 1)' \underset{\text{A.I}}{=} (b + 1) + 1 \underset{\text{A.I}}{=} b' + 1.$$

Hence by **S.5**, N contains all integers. Hence M contains 1.

(II) If M contains a, so that **H**: $a + b = b + a$, then M contains a', for

$$a' + b \underset{\text{A.I}}{=} (a + 1) + b \underset{\text{T.1}}{=} a + (1 + b) \underset{\text{(I)}}{=} a + (b + 1) \underset{\text{T.1}}{=}$$
$$(a + b) + 1 \underset{\text{A.I}}{=} (a + b)' \underset{\text{H}}{=} (b + a)' \underset{\text{A.II}}{=} b + a'.$$

By (I), (II), and **S.5**, M contains all integers and **T.2** is always valid.

T.3: The *cancellation* law for addition: if $a + c = b + c$, then $a = b$.

Proof: Let M be the set of all integers c for which **T.3** holds.

(I) M contains 1, for if $a + 1 = b + 1$, then by **A.I** we have $a' = b'$, whence by **S.4** it follows that $a = b$.

(II) If M contains c, so that **H**: $a + c = b + c$ implies $a = b$, then M contains c', because if $a + c' = b + c'$, then by **A.II** we have $(a + c)' = (b + c)'$, and by **S.4** we have $a + c = b + c$, whence by **H**, $a = b$.

By (I), (II), and **S.5**, M contains all integers, hence **T.3** is always valid.

Believing that the proofs of **T.1**, **T.2**, **T.3** are a sufficient indication of the methods of developing the properties of the natural integers from Peano's axioms, we shall merely indicate, without proof, but in proper sequence, a series of theorems, definitions, and lemmas that culminate in some comforting facts about the integers.

T.4: For all integers a,b, in S, $a \neq a + b$.

T.5: If $a \neq 1$, there exists just one integer u such that $a = u'$.

T.6: The *trichotomy* law: for every pair of integers a,b in S, one and only one of the following cases must hold:
(1) $a + u = b$; (2) $a = b$; (3) $a = b + v$.

Definition: We write $a < b$, read "a is less than b," if and only if there exists an integer u such that $a + u = b$; we write $a \leqq b$, read "a is less than or equal to b," if either $a < b$ or $a = b$.

L.1: $1 \leqq a$ for every a in S.

L.2: If $a < b$, then $a + 1 \leqq b$.

Definition: If in a given set N of integers there is an integer m such that $m \leqq x$ for all integers x in N, then m is called a "smallest integer in N."

T.7: In every non-empty set of integers, there is a smallest integer.

The proof of **T.7** is a little more subtle than proofs for the other theorems, perhaps because the conclusion of the theorem seems so obvious. How really fundamental this particular theorem **T.7** is, is shown by the fact that if **T.7** is assumed, then **S.5** may be proved as a theorem.

T.8: Peano's axioms are *categorical*.

In **28.1** we have already explained the significance of the term categorical. In this case it means that any two systems S and $\overline{S}$ satisfying Peano's axioms may be shown to be isomorphic, meaning that a one-to-one mapping can be established between the two systems, say that a in S corresponds to $\overline{a}$ in $\overline{S}$ and a' to $\overline{a'}$, such that this correspondence is "sequels-preserving," so that if $\overline{a}^*$ denotes the sequels operation in $\overline{S}$, then $\overline{a'} = \overline{a}^*$ for every a in S.

Since all the other usual operations on integers may be defined in terms of the sequels operation, it follows that **T.8** is of considerable logical importance in that although several different schemes of representing the integers may be proposed, they are abstractly the same. There is no theorem provable with one scheme of representation that is not true under all the other schemes. For example, of the various representations suggested in Chapter 4, some one may be more familiar or more convenient for a particular proof, but **T.8** assures us that the results, correctly translated, are true in every representation.

Ordinarily it is rather restricted and uninteresting to study a mathematical system that is categorical. But when we consider how the natural integers are building blocks for so many other systems, it is comforting to know that this basic system is essentially unique.

Perhaps the reader has noted that no attempt has been made to prove that Peano's axioms are consistent—this is another reflection of the basic nature of the system of integers. To avoid circular reasoning there must surely be some basic system whose consistency cannot be demonstrated and the integers seem a natural choice for this basic role.

28.4. The operation of multiplication. Just as in elementary arithmetic, where multiplication is introduced as a convenient shorthand for certain types of addition problems, so here in the abstract development, it proves convenient to define the operation of multiplication in terms of the previously studied addition. We wish to define for the integers of S an ordered binary operation, called multiplication, written ab. Following Landau, rather than Peano, we proceed as follows:

R.I: We define $1b = b$, for every b in S.

R.II: For every a in S for which

M.I: $a1 = a$, and

M.II: $ab' = ab + a$, for every b in S,
we define $a'b = ab + b$, for every b in S.

The following theorems will be left as exercises, it being understood that they had best be considered in sequence.

Theorem: The operation of multiplication is closed and well defined.

T.9: The *distributive* law: $(a + b)c = ac + bc$.

T.10: The *commutative* law for multiplication: $ab = ba$.

T.11: The *associative* law for multiplication: $(ab)c = a(bc)$.

T.12: The *cancellation* law for multiplication: if $ac = bc$, then $a = b$.

In terms of the sequels operation and the operation of multiplication it is possible (see Landau) to define for the integers of the system S a binary operation, called exponentiation, written a^b, that is closed and well defined and has the following properties:

E.I: $a^1 = a$; **E.II:** $a^{b'} = a^b a$.

The following theorems then represent exercises in the use of **S.5** and the previous theorems and definitions.

T.13: $a^b a^c = a^{b+c}$.

T.14: $(a^b)^c = a^{bc}$.

T.15: $(ab)^c = a^c b^c$.

EXERCISES

EX. *28.1.* Show that for a fixed integer a there is only one way of defining an operation of addition that will possess properties **A.I** and **A.II**.

EX. *28.2.* Prove that the operation of multiplication defined in 28.4 is closed and well defined.

EX. *28.3.* Prove **T.9**.

EX. *28.4.* Prove **T.10**.

Ex. *28.5.* Prove **T.11**.

EX. *28.6.* Prove **T.12**.

EX. *28.7.* Define $1' = 2, 2' = 3, 3' = 4$, and show by the use of the theorems that $(2)(2) = 4$.

EX. *28.8* Supposing that positive integers written with the base 10 form a system S satisfying Peano's axioms, prove that all positive multiples of 5 form a system $\bar{S}$ satisfying Peano's axioms. Find the isomorphism between S and $\bar{S}$ that illustrates **T.8.**

EX. *28.9.* Prove **T.13.**

EX. *28.10.* Prove **T.14.**

EX. *28.11.* Prove **T.15.**

EX. *28.12.* Prove that there exist no integers a and x, such that $a < x < a'$.

▶*Strictly speaking, the theory of numbers has nothing to do with negative, or fractional, or irrational quantities, as such.*

—G. B. MATHEWS

CHAPTER 29*

INTEGERS—POSITIVE, NEGATIVE, AND ZERO

29.1. Integers as pairs of natural integers. We wish to devote one more lesson to the foundations of our subject and to indicate one way in which the complete system of integers, positive, negative, and zero, which we indicated in the first lesson to be the elements of our study, may be developed in a logical manner from the natural integers as described, say, by Peano's axioms. Although all theorems about integers can be written in terms of the natural integers alone, there is a certain tediousness and vexation in doing so. Hence we are going to employ a device that is frequently useful in the study of algebraic systems, namely, the embedding of the system in question (in the sense of an isomorphism) within a larger system wherein it is hoped that the theorems in question may be more easily formulated and more easily proved valid.

In our case most of the difficulties, if we do not use this embedding idea, will be found to arise from the fact that in the system of natural integers the equation $a = b + x$ has no solution when $a = b$ and when $a < b$ as is seen by reference to **T.4** in **28.3**. Hence we may set ourselves the problem of inventing a number system, retaining as many features as possible of the system of natural integers and,

*Chapter 29 is a supplementary chapter.

in particular, containing a subset isomorphic to the natural integers, and within which the *type* equation "$a = b + x$" is always solvable. If the elements of the new system are called "integers," we must henceforth be rather strict in calling the elements previously discussed by their full title "natural integers."

Let us define the system of *integers* to be the set N of all ordered pairs (a,b) of natural integers a,b subject to the following definitions:

E: equality: $(a,b) = (c,d)$ if and only if $a + d = b + c$;

A: addition: $(a,b) + (c,d) = (a + c, b + d)$;

M: multiplication: $(a,b)(c,d) = (ac + bd, ad + bc)$;

O: ordering: $(a,b) < (c,d)$ if and only if $a + d < b + c$.

Note that each of these concepts is defined entirely in terms of elements, relations and operations in the system S of natural integers; hence all the following theorems may be proved by referring back to the previously assumed postulates or the previously established theorems about natural integers.

In the proofs of the next theorems we will use the notation "…" → "…" to indicate that the first statement *implies* the second statement, and we will write under the arrow the name of the definition, postulate or theorem which justifies the implication. Note that if $p \to q$ and $q \to r$, then $p \to r$. Similarly, the notation "…" ⟷ "…" will indicate that the first statement is valid *if and only if* the second statement is valid.

N.1: The equality **E** of integers of N is an equivalence relation.

Proof: **R.1:** **E** is determinative ty **T.6.**

R.2: **E** is reflexive, for $a + b \underset{\text{T.2}}{=} b + a \xrightarrow[\text{E}]{} (a,b) = (a,b)$.

R.3: **E** is symmetric, for

$$(a,b) \underset{\text{H}}{=} (c,d) \xrightarrow[\text{E}]{} a + d = b + c \xrightarrow[\text{T.2, R.3 in } S]{} c + b = a + d \xrightarrow[\text{E}]{} (c,d) = (a,b).$$

R.4: **E** is transitive, for

H.1: $(a,b) = (c,d) \xrightarrow[\text{E}]{} a + d = b + c$;

H.2: $(c,d) = (e,f) \xrightarrow[\text{E}]{} c + f = d + e$;

$$d + (a + f) \underset{\text{T.1}}{=} (d + a) + f \underset{\text{T.2}}{=} (a + d) + f \underset{\text{H.1}}{=} (b + c) + f$$
$$\underset{\text{T.1}}{=} b + (c + f) \underset{\text{H.2}}{=} b + (d + e) \underset{\text{T.1}}{=} (b + d) + e \underset{\text{T.2}}{=} (d + b) + e$$
$$\underset{\text{T.1}}{=} d + (b + e) \xrightarrow[\text{T.3}]{} a + f = b + e \xrightarrow[\text{E}]{} (a,b) = (e,f).$$

N.2: The addition A of integers of N is (*1*) closed, (*2*) well defined, (*3*) commutative, and (*4*) associative.

Proof: (*1*) by definition **A** and by **A.I** and **A.II** of **28.3,** the sum of two integers is an integer.

(*2*) If **H**: $(a,b) = (A,B) \underset{\mathbf{E}}{\rightarrow} a + B = b + A$, $(c,d) = (C,D) \underset{\mathbf{E}}{\rightarrow} c + D = d + C$; then

$$a + c + B + D \underset{\text{H, T.1, T.2}}{=} b + d + A + C \underset{\mathbf{E}}{\rightarrow} (a + c, b + d)$$

$$= (A + C, B + D) \underset{\mathbf{A}}{\rightarrow} (a,b) + (c,d) = (A,B) + (C,D).$$

(*3*) $(a,b) + (c,d) \underset{\mathbf{A}}{=} (a + c, b + d) \underset{\text{T.2, E}}{=} (c + a, d + b) \underset{\mathbf{A}}{=} (c,d) + (a,b).$

(*4*) $\{(a,b) + (c,d)\} + (e,f) \underset{\mathbf{A}}{=} (a + c, b + d) + (e,f)$

$$\underset{\mathbf{A}}{=} \{(a + c) + e, (b + d) + f\} \underset{\text{T.1, E}}{=} \{a + (c + e), b + (d + f)\}$$

$$\underset{\mathbf{A}}{=} (a,b) + (c + e, d + f) \underset{\mathbf{A}}{=} (a,b) + \{(c,d) + (e,f)\}.$$

N.3: The multiplication **M** of integers of N is (*1*) closed, (*2*) well defined, (*3*) commutative, (*4*) associative, and (*5*) distributive with respect to addition.

Proof: (*1*) by definition **M** and by **A.I, A.II, M.I,** and **M.II** of **28.3** and **28.4,** the product of two integers is an integer.

(*2*) If **H**: $(a,b) = (A,B) \underset{\mathbf{E}}{\rightarrow} a + B = b + A$, $(c,d) = (C,D) \underset{\mathbf{E}}{\rightarrow} c + D = d + C$, then

$$(aD + bC + bD + aC) + (ac + bd + AD + BC)$$

$$\underset{\text{T.1, T.2, T.9, T.10}}{=} a(c + D) + b(d + C) + (b + A)D + (a + B)C$$

$$\underset{\mathbf{H}}{=} a(d + C) + b(c + D) + (a + B)D + (b + A)C$$

$$\underset{\text{T.1, T.2, T.9, T.10}}{=} (aD + bC + bD + aC) + (ad + bc + AC + BD)$$

$$\underset{\text{T.3}}{\rightarrow} ac + bd + AD + BC = ad + bc + AC + BD$$

$$\underset{\mathbf{E}}{\rightarrow} (ac + bd, ad + bc) = (AC + BD, AD + BC)$$

$$\underset{\mathbf{M}}{\rightarrow} (a,b)(c,d) = (A,B)(C,D).$$

(*3*) $(a,b)(c,d) \underset{\mathbf{M}}{=} (ac + bd, ad + bc)$

$$\underset{\text{T.2, T.10, E}}{=} (ca + db, cb, da) \underset{\mathbf{M}}{=} (c,d)(a,b).$$

$$(4)\quad \{(a,b)(c,d)\}(e,f) \underset{\mathbf{M}}{=} (ac + bd,ad + bc)(e,f)$$

$$\underset{\text{M, T.9, T.11, T.1, E}}{=} (ace + bde + adf + bcf,acf + bdf + ade + bce)$$

$$\underset{\text{T.1, T.2, E}}{=} (ace + adf + bcf + bde,acf + ade + bce + bdf)$$

$$\underset{\text{E, T.1, T.11, T.9, M}}{=} (a,b)(ce + df,cf + de) \underset{\mathbf{M}}{=} (a,b)\{(c,d)(e,f)\}.$$

$$(5)\quad (a,b)\{(c,d) + (e,f)\} \underset{\mathbf{A}}{=} (a,b)(c + e,d + f)$$

$$\underset{\text{M, T.9, E}}{=} (ac + ae + bd + bf,ad + af + bc + be)$$

$$\underset{\text{T.1, T.2, E}}{=} (ac + bd + ae + bf,ad + bc + af + be)$$

$$\underset{\text{T.1, A, E}}{=} (ac + bd,ad + bc) + (ae + bf,af + be)$$

$$\underset{\mathbf{M}}{=} (a,b)(c,d) + (a,b)(e,f).$$

N.4: The ordering **O** of integers of N satisfies a trichotomy law.

Proof: By **T.6** of **28.3**, one and only one of the cases $a + d < b + c$, $a + d = b + c$, $b + c < a + d$ must hold, hence by **O** and **E** one and only one of the following cases must hold:

(1) $(a,b) < (c,d)$; (2) $(a,b) = (c,d)$; (3) $(c,d) < (a,b)$.

29.2. Integers classified as positive, negative, and zero. From **T.6** one and only one of the cases $a < b$, $a = b$, $b < a$ must hold, hence there are three and only three types of integers (a,b). This classification may be made even more explicit as follows:

Lemma: $(x,y) = (a + k,a)$ if and only if $x = y + k$;
$(x,y) = (a,a)$ if and only if $x = y$;
$(x,y) = (a,a + k)$ if and only if $y = x + k$.

Proof:

$$(x,y) = (a+k,a) \underset{\mathbf{E}}{\longleftrightarrow} x+a=y+a+k \underset{\text{T.1, T.2, T.3}}{\longleftrightarrow} x=y+k;$$

$$(x,y) = (a,a) \underset{\mathbf{E}}{\longleftrightarrow} x + a = y + a \underset{\text{T.2, T.3}}{\longleftrightarrow} x = y;$$

$$(x,y) = (a,a+k) \underset{\mathbf{E}}{\longleftrightarrow} x+a+k=y+a \underset{\text{T.1, T.2, T.3}}{\longleftrightarrow} y=x+k.$$

In the light of this lemma the following more convenient notations may be introduced and the new definitions may be described as "well defined," because, as the lemma shows, the definitions are independent of the natural integer a which appears in them.

Definitions: $(a + k,a) = k$, called the "positive integer, *plus* k";
$(a,a) = 0$, called the "zero integer";
$(a,a + k) = -k$, called the "negative integer, *minus* k."

N.5: The system of natural integers is isomorphic to the subsystem of the integers consisting of the positive integers, the correspondence between the two systems being preserved under *both* addition and multiplication.

Proof: The correspondence $F(k) = (a + k,a)$, $F(m) = (b + m,b)$ is a one-to-one mapping of the natural integers onto the positive integers. Since

$$F(k) + F(m) = (a + k,a) + (b + m,b) = (a + b + k + m,a + b) = F(k + m),$$

it follows that the correspondence is an isomorphism with respect to the operations of addition in the two systems. Since

$$F(k)F(m) = (a + k,a)(b + m,b) = (ab + am + kb + ab + km,ab + am + kb + ab) = F(km).$$

it follows that the correspondence is also an isomorphism with respect to the operations of multiplication in the two systems.

Hence we have succeeded in constructing a system of numbers with a subsystem isomorphic to the natural integers. That the new system is larger than this subsystem is shown by the presence of zero and negative integers and in another way by the fact that the *type* equation "$a = b + x$" is now always solvable, although this equation in N has a different appearance than in S with new meanings for the elements, the equals relation, and the operation of addition, namely:

N.6: The equation **H:** $(a,b) = (c,d) + (x,y)$ has the unique solution $(x,y) = (a + d,b + c)$.

Proof: $$(a,b) \underset{\text{H, A}}{=} (c + x,d + y) \underset{\text{E, T.1}}{\longleftrightarrow} a + d + y = b + c + x$$
$$\underset{\text{T.1, T.2}}{\longleftrightarrow} x + b + c = y + a + d \underset{\text{T.1, E}}{\longleftrightarrow} (x,y) = (a + d,b + c).$$

We shall leave as exercises the proofs of the following theorems:

N.7: Properties of the zero integer: *(1)* $0 + (a,b) = (a,b)$; *(2)* $0(a,b) = 0$; *(3)* $(a,b) + (b,a) = 0$; *(4)* If $(a,b) + (c,d) = (a,b) + (e,f)$, then $(c,d) = (e,f)$.

N.8: The *restricted* cancellation law for multiplication in N: If $(a,b)(c,d) = (a,b)(e,f)$ *and* if $(a,b) \neq 0$, then $(c,d) = (e,f)$.

N.9: Properties of negative integers:

$$(1) \quad (-k)m = -km = k(-m); \qquad (2) \quad (-k)(-m) = km.$$

Theorem **N.9** seems worthy of some philosophical remark, for too often because of the limited background of his first teachers the student will be left with the impression that the rules of signs, embodied in **N.9**, have some absolute or preordained source. But here the rules of signs are seen to be merely incidental theorems, the offshoots of a more deepseated search for a system with commutative and associative operations as in **N.2** and **N.3** and within which **N.6** will be valid.

If we define a new operation, called subtraction, as follows: $(a,b) - (c,d) = (a,b) + (d,c)$, then we may prove as exercises the following theorems:

N.10: Properties of subtraction:

$$(1) \quad k - m = k + (-m); \qquad (2) \quad (k,m) = k - m;$$
$$(3) \quad k - (-m) = k + m.$$

EXERCISES

EX, *29.1.* Prove **N.7.**

EX. *29.2.* Prove **N.8.**

EX. *29.3.* Prove **N.9.**

EX. *29.4.* Prove **N.10.**

EX. *29.5.* If a and b are any *positive* integers, show $-a < b$.

EX. *29.6.* If b is a *positive* integer, show that $(s,t)b < (u,v)b$ if and only if $(s,t) < (u,v)$.

EX. *29.7.* If $-1 < (s,t) < 1$, show that $s = t$.

EX. *29.8.* If b is a *positive* integer and $-b < (s,t)\,b < b$, show that $s = t$. (*Hint:* Use EX, *29.6*, EX. *29.7.*)

EX. *29.9.* Show that $a < b$ if and only if $a + x < b + x$ for all x (a, b, and x in N).

EX. *29.10.* Show that the quotient and remainder in the division algorithm are unique by assuming the possibility of two representations, say $a = qb + r = q_1b + r_1$, $0 \leqq r < b$, $0 \leqq r_1 < b$, and proving that $r = r_1$ and $q = q_1$. (*Hint:* Use EX. *29.9*, EX. *29.8*, and **N.8.**)

▶*The existence of calculating machines proves that computation is not concerned with the significance of numbers, but only with the formal laws of operation; for it is only these which the machine can be constructed to obey, having no perception of the meaning of the numbers.* —F. KLEIN

CHAPTER 30*

RATIONAL NUMBERS

30.1. Logical foundation for rational numbers. In the preceding chapters on those few occasions when we have used fractions, we have assumed that the rules for operating with these numbers were well known to our readers. However, it is perhaps within the province of this text to discuss fractions more critically and to approach the subject in the spirit of Chapters 28 and 29, defining each new symbol and each new operation in terms of previously known numbers and operations. We shall use the technical term "rational number" instead of the colloquial word "fraction," but as a symbol we shall employ the familiar a/b.

Definitions: (*1*) A *rational number*, indicated a/b, is an ordered pair of integers a and b, with $b > 0$ (however, see EX. *30.1*); the first integer a is called the *numerator*; and second integer b, the *denominator*.

(2) Equality: $a/b = c/d$ if and only if $ad = bc$.

(3) Multiplication: $(a/b)(c/d) = ac/bd$.

(4) Addition: $a/b + c/d = (ad + bc)/bd$.

(5) Order: $a/b < c/d$ if and only if $ad < bc$.

Note well how each definition employs only concepts which were

[*] Chapter 30 is a supplementary chapter.

developed earlier. This fact is brought out clearly in the theorems which follow.

U.1: Equality of rational numbers is an equivalence relation.

Proof: Since the elements of the rational numbers are integers for which the rules of multiplication and equality are already known, the relation defined by (2) is *determinative*. The relation in (2) is *reflexive*, for from the commutative property of multiplication of integers we have $ab = ba$, which by (2) implies $a/b = a/b$. The relation in (2) is *symmetric*, for $a/b = c/d$ implies $ad = bc$, which in turn implies, by the commutative property of multiplication of integers and by the symmetric property of equality of integers, that $cb = da$, whence by (2) we have $c/d = a/b$. The relation in (2) is *transitive*, for if $a/b = c/d$ and $c/d = e/f$, then by (2) we have $ad = bc$ and $cf = de$; so employing the associative property of multiplication of integers we may write

$$(ad)f = (bc)f = b(cf) = b(de);$$

since by (1) we know $d > 0$, we may use the commutative and cancellation laws for multiplication of integers to cancel d and arrive at $af = be$, which implies by (2) that $a/b = e/f$.

This completes the proof that the relation (2) for rational numbers is an equivalence relation. Recalling that an equivalence relation divides the set concerned into mutually exclusive classes of equivalent elements, we may investigate the nature of these classes.

In considering the rational number a/b with $b > 0$, suppose $(a,b) = D$, with $D > 0$, then $a = AD$, $b = BD$, with $(A,B) = 1$ and $B > 0$. By (2) it follows that $a/b = AD/BD = A/B$ because $(AD)B = (BD)A$. Thus any given rational number is equal to one in which numerator and denominator are relatively prime. On the other hand, if $c/d = A/B$ where $(A,B) = 1$, then by (2) we have $cB = dA$; but by the fundamental lemma (EX. 6.2), it follows that $c = kA$ and $d = kB$, where k is an integer; moreover, since $d > 0$ and $B > 0$, it follows that $k > 0$. Conversely, for any positive integer k, we find kA/kB is defined and $kA/kB = A/B$.

In summary, we find all rational numbers (1) divided by the relation (2) into classes of equal numbers such that in each class there is one and only one rational number A/B with numerator and denominator relatively prime; all numbers of the form kA/kB with $k > 0$

are in the class; and every number in the class is of this form. We say that A/B with $(A,B) = 1$ is a "canonical" or "reduced" rational number, or that such a rational number is in "lowest terms."

For example, $-30/12$, $-15/6$, $-10/4$ are all in the same class whose canonical representative is $-5/2$; and every number in this class is represented by $-5k/2k$ with $k > 0$.

U.2: Multiplication of rational numbers is closed, well defined, commutative, and associative.

Proof: In a/b and c/d we have $b > 0$ and $d > 0$ so that $bd > 0$, hence ac/bd, with an integer in the numerator position and a positive integer in the denominator, is a rational number, so the operation defined by *(3)* is *closed.* If $a/b = A/B$ and $c/d = C/D$, so that $aB = bA$ and $cD = dC$, then $acBD = bdAC$ which shows $ac/bd = AC/BD$ and proves the operation defined by *(3)* to be *well defined* with $(a/b)(c/d) = (A/B)(C/D)$. Since in integers we have $acbd = bdca$, it follows from (2) that $ac/bd = ca/db$; then by *(3)* and **U.1** we see that $(a/b)(c/d) = (c/d)(a/b)$ which shows that the new multiplication is *commutative.* Since in integers $\{(ac)e\}\{b(df)\} = \{(bd)f\}\{a(ce)\}$, it follows from (2) and *(3)* that $(ac/bd)(e/f) = (a/b)(ce/df)$ and then that $\{(a/b)(c/d)\}(e/f) = (a/b)\{(c/d)(e/f)\}$, hence the new multiplication defined by *(3)* is *associative.*

U.3: Addition of rational numbers is closed, well defined, commutative, associative, and distributive with respect to multiplication.

Proof: In a/b and c/d we have $b > 0$ and $d > 0$ so that $bd > 0$, hence the symbol $(ad + bc)/bd$ with an integer in the numerator and a positive integer in the denominator is a rational number, so the operation defined by *(4)* is *closed.* If $a/b = A/B$ and $c/d = C/D$, then $aB = bA$ and $cD = dC$ and in integers we have

$$(ad + bc)BD = (aB)dD + bB(cD) = (bA)dD + bB(dC) = bd(AD + BC),$$

so that $(ad + bc)/bd = (AD + BC)/BD$ and the operation defined by *(4)* is therefore *well defined* with $a/b + c/d = A/B + C/D$. Since in integers $(ad + bc)db = bd(cb + da)$ we find by (2),*(4)*, and **U.1** that $a/b + c/d = (ad + bc)/bd = (cb + da)/bd = c/d + a/b$, which is the required *commutative* property. Since in integers

$$\{(ad + bc)f + (bd)e\}b(df) = (bd)f\{a(df) + b(cf + de)\}$$

we find by (2) and (4) and **U.1** that

$$(a/b + c/d) + e/f = a/b + (c/d + e/f)$$

which shows that the new addition is *associative.*

Since in integers we have

$$(ad + bc)e(bf)(df) = (bd)f\{ae(df) + (bf)ce\},$$

we find by (2) that

$$(ad + bc)e/(bd)f = \{ae(df) + (bf)ce\}/(bf)(df);$$

then by (3), (4), and **U.1** we find

$$\{(ad + bc)/bd\}\{e/f\} = ae/bf + ce/df.$$

Finally, by (4), (3), and **U.1** we have

$$(a/b + c/d)(e/f) = (a/b)(e/f) + (c/d)(e/f)$$

which is the required *distributive* property.

30.2. The rational field. In previous chapters we have presented the notions of an equivalence relation (**17.2**), of a group of transformations (**11.2**), and of an abstract group (following EX. *18.8*). A somewhat more elaborate mathematical system, at least in the sense of the number of operations involved, is the "abstract field" described by the following definition.

An *abstract field* is a set F of elements $a,b,\ldots$, with a relation, called equality, usually written $a = b$, and two operations: one called addition, written $a + b$, and the other called multiplication, written ab, all subject to the following postulates:

F.1: The equality defined in F is an equivalence relation.

F.2: The elements of F form a commutative group under addition, with an identity of addition called *zero.*

F.3: The elements of F, other than *zero*, form a commutative group under multiplication.

F.4: For all elements of F, the operations of addition and multiplication are related by a distributive law

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

U.4: The set *Ra* of all rational numbers, with elements, equality, addition and multiplication defined as in **30.1**, is a field, known as the "rational field."

Proof: By **U.1** the equality defined in *Ra* is an equals relation so **F.1** is satisfied.

By **U.3** the addition defined in *Ra* has some of the properties required for a commutative group. Beyond this we must show the existence of an identity element for addition, the number which we will call "zero." For this purpose we find $0/1$ to be satisfactory, since $a/b + 0/1 = (a\cdot 1 + b\cdot 0)/b\cdot 1 = a/b$ for every a/b in *Ra*. We must also show for every a/b in *Ra* the existence of an inverse (or negative) with respect to addition, i.e., a rational number x/y such that $a/b + x/y = 0/1$. Without difficulty we see that $x/y = -a/b$ is a suitable choice, since by (3) and (2) in **30.1** we have

$$a/b + (-a)/b = (ba + b(-a))/b = 0/b = 0/1.$$

Thus the numbers of *Ra* form a commutative group under addition, so **F.2** is satisfied.

By **U.2** the multiplication defined in *Ra* has some of the properties required for a commutative group. Beyond this we must show the existence of an identity element for multiplication. For this purpose we find $1/1$ to be satisfactory, since $(a/b)(1/1) = a/b$ for every a/b in *Ra*. Then for every non-zero number a/b of *Ra* we must produce an inverse (or reciprocal) with respect to multiplication, i.e., a rational number x/y such that $(a/b)(x/y) = 1/1$; furthermore, this inverse must be non-zero. First we will use (2) of **30.1** to check that $a/b = 0/1$ if and only if $a = a\cdot 1 = b\cdot 0 = 0$; so a non-zero rational number a/b is characterized by having $a \neq 0$. If $a > 0$, then b/a is a non-zero rational number such that $(a/b)(b/a) = ab/ba = 1/1$; if $a < 0$, then $-b/-a$ is a non-zero rational number such that $(a/b)(-b/-a) = -ab/-ba = 1/1$; hence every non-zero rational number has a non-zero inverse with respect to multiplication. Thus the non-zero numbers of *Ra* form a commutative group with respect to multiplication, so **F.3** is satisfied.

The distributive relation between addition and multiplication for the system *Ra* is included in **U.3**, so **F.4** is satisfied.

Thus the system *Ra* has been shown to satisfy all the postulates required of a field, so the proof of **U.4** is complete.

30.3. The rational domain. Let us consider a mathematical system *D* having all the properties **F.1**, **F.2**, **F.3**, **F.4**, of a field, *except* that one of the postulates implied by **F.3** is replaced by a weaker postulate; namely, the requirement, included in **F.3**, that every non-zero element of D have an inverse with respect to multiplication is to be replaced by the weaker requirement that the cancellation law of multiplication be valid for all non-zero numbers of D. Such a system D is called an *abstract domain.*

As is implied by our use of the adjective "weaker," every field is a domain, but not every domain is a field. Consider in a field F the equation $ab = ac$ with a not-zero; then a has an inverse x such that $xa = e$, where e is the identity of multiplication; by the well-defined property of multiplication and by the associative law we find

$$b = eb = (xa)b = x(ab) = x(ac) = (xa)c = ec = c,$$

so the cancellation law of multiplication is valid for all non-zero numbers of a field; thus every field is a domain.

On the other hand, the integers form a good example of a domain which is not a field. For in Chapter 29 in theorems **N.1**, **N.2**, **N.3**, **N.6**, **N.7**, **N.8** we have the necessary properties to prove that the integers form a domain; but among the integers only the units $+1$ and -1 have inverses which are integers, so the integers fail to form a field. We shall now demonstrate why the integers are called *rational integers* and why the domain of integers is called the *rational domain.*

U.5: In the rational field Ra the set $[Ra]$ of all rational numbers of the form $a/1$, each of which is called a "rational integer," is a domain, called the "rational domain," which is isomorphic to the domain of all integers with respect to equality, addition, and multiplication.

Proof: We check readily from $a/1 + b/1 = (a + b)/1$ and $(a/1)(b/1) = ab/1$ that the set $[Ra]$ is closed under addition and multiplication. Since the set includes the zero $0/1$, the negative $-a/1$ of $a/1$, and the identity $1/1$, it follows readily from the fact that Ra is a field, that its subset $[Ra]$ is a domain (in particular, the cancellation law is valid for the non-zero numbers of a field).

Moreover the one-to-one correspondence T defined by $aT = a/1$ between integers and rational integers is preserved under the equality, addition, and multiplication rules of the two systems. Thus by definition of the correspondence T we have $aT = a/1$, $bT = b/1$, $(a + b)T = (a + b)/1$, and $(ab)T = ab/1$. Since $a/1 = b/1$ if and only if $a = b$, the correspondence is one-to-one and preserves equivalence relations; since $(a + b)T = (a + b)/1 = a/1 + b/1 = aT + bT$, the correspondence preserves the addition operations; and since $(ab)T = ab/1 = (a/1)(b/1) = (aT)(bT)$, the correspondence also preserves the multiplication operations.

Henceforth, making use of the isomorphism which we have just established, we shall refer to the integers—positive, negative, and zero—as *rational integers*; however, we shall not usually want to write $a/1$ for a rational integer, but shall employ the simpler notation a. Because of the isomorphism there is little danger of confusion. Moreover, since in later lessons we want to study other domains whose elements also are called "integers," it will be helpful to have the full title of "rational integers" to designate the elements of this most fundamental, prototype domain [Ra].

30.4. Order and absolute value for rational numbers. Definition (5) in **30.1** defines order among the rational numbers in terms of the previously studied order for (rational) integers.

U.6: Order in Ra is well defined, trichotomous, and transitive.

Proof: If $a/b = A/B$ and $c/d = C/D$, then $aB = bA$ and $cD = dC$. If $a/b < c/d$, then by (5) we must have $ad < bc$. Since $BD > 0$, we may write $(bA)dD = (aB)dD = (ad)BD < (bc)BD = bB(cD) = bB(dC)$. Since $bd > 0$, we conclude that $AD < BC$, whence $A/B < C/D$, so that the order relation in Ra is well defined.

In integers we know that one and only one of the cases $ad < bc$, $ad = bc$, or $bc < ad$ will hold; hence in Ra, one and only one of the tricases $a/b < c/d$, $a/b = c/d$, $c/d < a/b$ will hold, which is the trichotomy law for Ra.

If $a/b < c/d$ and $c/d < e/f$, then $ad < bc$ and $cf < de$. Since $f > 0$ and $b > 0$, we may write $(ad)f < (bc)f = b(cf) < b(de)$; then since $d > 0$, we conclude that $af < be$, or that $a/b < e/f$. Thus the order relation in Ra is transitive. This completes the proof of **U.6**.

In particular, if $0/1 < a/b$, we shall call a/b a "positive" rational number; if $a/b < 0/1$, we shall call a/b a "negative" rational number. By the trichotomy property in **U.6**, all the rational numbers fall into three classes: positive, zero, and negatives. The rule of signs for multiplication of these classes parallels that for integers given in **N.9** of **29.2**.

Let us define the notion of absolute value for rational numbers in terms of the absolute value of integers as follows:

(*6*) Absolute value: $|a/b| = |a|/b$.

The following facts may be readily established:

U.7.1: $|a/b|$ is non-negative.

U.7.2: $|a/b|$ is zero, if and only if $a/b = 0/1$.

U.7.3: $|(a/b)(c/d)| = |a/b|\,|c/d|$.

U.7.4: $|a/b + c/d| \leqq |a/b| + |c/d|$.

Proof: Properties **U.7.1** and **U.7.2** of absolute value in *Ra* follow readily from definition (*6*). Properties **U.7.3** and **U.7.4** may be established by considering the various cases that arise according as one of a and c is zero, according as a and c have like or unlike signs, and according as $|a/b|$ or $|c/d|$ is the greater, or that $|a/b| = c/d|$.

EXERCISES

EX. *30.1.* Show that **U.1**, **U.2**, **U.3** are still valid when the only restriction placed on the number a/b is that $b \neq 0$. But show that order defined by (*5*) would now fail to be well defined.

EX. *30.2.* If m is composite, show that residue classes of integers mod m do not form a domain. (See **18.1**.)

EX. *30.3.* If p is a prime, show that residue classes of integers mod p form a field (known as a Galois field, $GF(p)$).

EX. *30.4.* Establish the rule of signs for multiplication of positive and negative rational numbers.

EX. *30.5.* If $a/b < c/d$, show that $a/b + e/f < c/d + e/f$ for any e/f in *Ra*; but that $(a/b)(e/f) < (c/d)(e/f)$ if and only if $0/1 < e/f$.

EX. *30.6.* Show that in any domain the zero z has the property that $az = z$ for every element a in the domain.

EX. *30.7.* Give the details in the proof of **U.7.1, U.7.2, U.7.3, U.7.4.**

EX. *30.8.* Show that *Ra* contains no solutions of the equations $(x/y)(x/y) = (-1/1)$ and $(x/y)(x/y) = (2/1)$.

EX. *30.9.* Consider a system G made up of ordered pairs (A,B) of rational numbers $A = a_1/a_2$ and $B = b_1/b_2$. Define:

equality: $(A,B) = (C,D)$ if and only if $A = B$ and $C = D$;

addition: $(A,B) + (C,D) = (A + C, B + D)$;

multiplication: $(A,B)(C,D) = (AC - BD, AD + BC)$.

Prove that G is a field (known as the Gaussian field).

EX. *30.10.* Show that G contains a subfield G^* made up of all numbers of G of the form $(A,0)$. Show that the mapping T defined by $AT = (A,0)$ is an isomorphism between Ra and G^*.

EX. *30.11.* Show that G contains a domain $[G]$ made up of all numbers of G of the form (A,B) where A and B are rational integers, $A = a/1$, $B = b/1$.

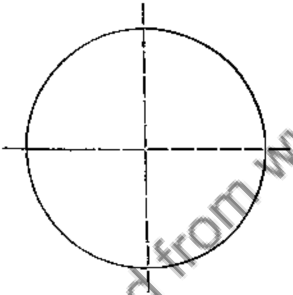
EX. *30.12.* Show that G contains numbers (X,Y) solving $(X,Y)(X,Y) = (-1/1,0)$.

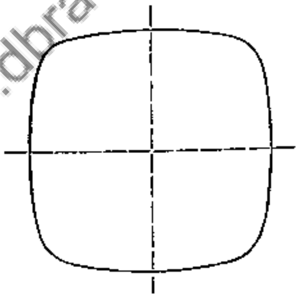
EX. *30.13.* Show that G does *not* contain any numbers (X,Y) solving $(X,Y)(X,Y) = (2/1,0)$.

EX. *30.14.* Define a point (x,y) of a rectangular coordinate system to be a "rational point" if and only if *both* x and y are rational numbers.

(a) Describe accurately all the *infinitely many rational points* on the locus of $x^2 + y^2 = 1$.

(b) Prove that there are *only four rational points* on the locus of $x^4 + y^4 = 1$.





▶ *The student who has thus far taken the system of real numbers for granted, and worked with them, may continue to do so to the end of his life without detriment to his mathematical thought. On the other hand, most mathematicians are curious, at one time or another in their lives, to see how the system of real numbers can be evolved from the natural numbers.* —W. F. OSGOOD

CHAPTER 31*

DECIMAL REPRESENTATION OF RATIONAL NUMBERS

31.1. Decimal fractions. If the base for representing rational integers is the usual base 10, then it is of particular interest to study decimal fractions a/b, where we limit the denominator b to be of the form $b = 10^k$ with the exponent k a non-negative integer, interpreting $10^0 = 1$. For such decimal fractions there is a convenient positional notation which we shall now describe in detail.

Since we may express $a > 0$ in the form

$$a = a_m 10^m + \ldots + a_1 10 + a_0; \quad 0 < a_m < 10; \quad 0 \leqq a_i < 10,\ 0 \leqq i < m;$$

it follows that when $a > 0$ we may write

$$a/10^k = a_m 10^m/10^k + \ldots + a_1 10/10^k + a_0/10^k$$

If $m \geqq k$, we have

$$a/10^k = a_m 10^{m-k} + \ldots + a_{k+1} 10 + a_k + a_{k-1}/10 + \ldots + a_1/10^{k-1} + a_0/10^k;$$

*Chapter 31 is, in general, a supplementary chapter, but sections 31.4 and 31.5 merit special attention since they present interesting applications of basic material from previous chapters.

and if we define $b_{i-k} = a_i$, for $i = 0,1,\ldots,m$, then we may write $a/10^k$ in "decimal notation" as follows:

$$a/10^k = b_{m-k}\ldots b_1b_0.b_{-1}b_{-2}\ldots b_{-k}.$$

In this notation, if $j \geqq 0$, we are to interpret b_j by its $j+1$ position to the *left* of the period or "decimal point" to represent b_j10^j; but if $j > 0$, we are to interpret b_{-j} by its jth position to the *right* of the decimal point to represent the decimal fraction $b_{-j}/10^j$; then if the whole symbol is understood to represent the sum of these components, it correctly represents $a/10^k$.

If $m < k$, we again set $b_{i-k} = a_i$, and find

$$\begin{aligned} a/10^k &= a_m/10^{k-m} + \ldots + a_1/10^{k-1} + a_0/10^k \\ &= b_{-(k-m)}/10^{k-m} + \ldots + b_{-(k-1)}/10^{k-1} + b_{-k}/10^k. \end{aligned}$$

But in this case to effect a suitable positional notation, if $k - m > 1$ we must define $b_{-1} = b_{-2} = \ldots = b_{-(k-m-1)} = 0$, so that in the symbol

$$a/10^k = .00\ldots 0b_{-(k-m)}\ldots b_{-k}$$

we will have b_{-j} occurring in the jth position to the *right* of the decimal point.

Thus, for examples, we have

$$30.302 = 30 + 3/10 + 2/1000 = 31302/1000;$$
$$.0071 = 7/1000 + 1/10000 = 71/10000.$$

In particular we want to observe that when $j \geqq i + 1$,

$$(31.1) \qquad .00\ldots 0b_{-(i+1)}\ldots b_{-j} < 1/10^i.$$

For the inequality to be checked is equivalent, by the definitions above, to the following inequality:

$$(b_{-(i+1)}10^{j-(i+1)} + \ldots + b_{-j})/10^j < 1/10^i,$$

which reduces to the following inequality in *integers*

$$b_{-(i+1)}10^{j-(i+1)} + \ldots + b_{-j} < 10^{j-i}.$$

But this last inequality is known to be valid because of the restriction $0 \leqq b_k < 10$ on all of the b's.

If $a < 0$ we may employ the above notation for $(-a)/10^k$ and write $a/10^k = -\{(-a)/10^k\}$, prefixing the negative sign to the decimal representation.

The set D of *all decimal fractions* is a domain, but *not* a field; for D is closed under addition and multiplication, contains $0/1$ and $1/1$, and has all the other properties for a domain because D is a subset of Ra; however, D is *not* a field because the inverse of a non-zero fraction $a/10^k$ of D is in D if and only if a is of the form 2^s5^t, $s \geqq 0$, $t \geqq 0$. For example, although $7/2 = 35/10$ is in D, the inverse rational number $2/7$ is not in D.

However, we can introduce a larger number system which contains numbers isomorphic to the rational numbers in such a way that we shall be able to represent *any rational number* (or, more precisely, its isomorphic image) using only decimal fractions. This enlarged number system is known as the "real number system" or (since it does have the required properties) as the "real field."

The device which we use in making this extension is already known to the reader, for he is accustomed to "approximating" a fraction such as 2/7 by an appropriately chosen decimal fraction, such as .285714285714. Here it will be worth while to examine the method of finding an appropriate decimal fraction, the exact meaning of the approximation, and the periodic character of the representation, for these matters are closely related to the division algorithm, to the properties of inequalities and absolute value, and to the theory of congruences, and these seem proper subjects for a lesson in the theory of numbers.

31.2. Regular sequences. We shall be interested in this section in infinite sequences of rational numbers. Each such sequence may be indicated by $a_1, a_2, \ldots, a_n, \ldots$ or more briefly by $\{a_i\}$.

A *regular* sequence $\{a_i\}$ is an infinite sequence of rational numbers, such that for any assigned positive rational number ϵ, however small, it is possible to find a corresponding rational integer $N = N(\epsilon)$ so that

$$|a_N - a_i| < \epsilon \quad \text{for all } i > N(\epsilon).$$

It is important to observe about this definition of a regular sequence, that it is *not* necessary to find one N which will serve for *all* choices of ϵ; rather, all that is required is that each time an ϵ is selected, that it shall be possible to find the corresponding $N(\epsilon)$—purposely written this way to emphasize that N depends upon ϵ. When a sequence is regular what will ordinarily happen is that as ϵ is chosen smaller and smaller, $N(\epsilon)$ must be selected larger and larger. On the other hand, it is *not* sufficient to guarantee that a sequence is regular to produce an $N(\epsilon)$ suitable for *one* assigned ϵ; we must be able to find an $N(\epsilon)$ for *any* assigned ϵ.

The sequence $\{a_i\}$ where $a_i = 2^i$ is *not* a regular sequence, for even with $\epsilon = 1$ it is impossible to make $|a_i - a_j| < \epsilon$ whenever $i \neq j$.

The sequence $\{a_i\}$ where $a_i = 1/3$ for every value of i is a trivial example of a regular sequence, for no matter how small the positive rational number ϵ is chosen, we may take $N(\epsilon) = 1$ and guarantee $|a_1 - a_i| < \epsilon$ for $i > 1$, inasmuch as $a_i - a_j = 0$ for all i and j.

A less trivial example is provided by the sequence $\{a_i\}$ where $a_n = 1 - 1/2^n$. Here $a_j - a_i = 1/2^i - 1/2^j$, hence by **U.7.4** of **30.4** we may write $|a_j - a_i| < 1/2^i + 1/2^j$. But when $j < i$, we have $2^j < 2^i$ and $1/2^j > 1/2^i$; hence by EX. *30.5* we have $|a_j - a_i| < 2/2^j$ when $i > j$. Given $\epsilon = a/b$, positive, however small, we can make $2/2^j < a/b = \epsilon$ or $2^j a > 2b$, by choosing j sufficiently large; for if, in the binary system, $2b = b_m 2^m + \ldots + b_1 2$, then $j = m + 1$ will suffice. Using the transitive property (see **U.6** in **30.4**), we may take $N(\epsilon) = m + 1$ and have

$$|a_N - a_i| < \epsilon \quad \text{when } i > N.$$

(For example, if $\epsilon = 1/5000$, then $2b = 10^4 = (10011100010000)_2$, $m = 13$, $N(\epsilon) = 14$, and $|a_{14} - a_i| < 1/5000$ for $i > 14$.) Thus $\{a_i\}$ is a regular sequence.

A fundamental example involving decimal fractions may be described as follows. Let $b_m, b_{m-1}, \ldots, b_0, b_{-1}, \ldots, b_{-i}, \ldots$ be any infinite sequence of integers b_i satisfying $0 \leqq b_i < 10$; define a corresponding sequence $\{a_i\}$ of decimal fractions as follows: $a_i = b_m \ldots b_0 . b_{-1} \ldots b_{-i}$. A sequence of this type will be called an "infinite decimal," designated by

$$\{a_i\} = b_m \ldots b_0 . b_{-1} \ldots b_{-i} \ldots .$$

Theorem: An infinite decimal is a regular sequence.

Proof: If $i > j$, then we may apply *(31.1)* to see that

$$a_i - a_j = .00\ldots 0 b_{-(j+1)} \ldots b_{-i} < 1/10^j.$$

Hence if we are given $\epsilon = a/b > 0$ with $b = q_t 10^t + \ldots + q_1 10 + q_0$ and $0 \leqq q_i < 10$, we may make $1/10^j < \epsilon$ if we can make $10^j a > b$; but this is easily arranged by taking $j = t + 1$. So we select $N(\epsilon) = t + 1$ and by **U.6** we have

$$|a_N - a_i| < \epsilon \quad \text{when} \quad i > N.$$

Thus we have shown $\{a_i\} = b_m \ldots b_0 . b_{-1} \ldots b_{-i} \ldots$ to be a regular sequence.

31.3. The real number system. The concepts of the preceding sections may be used to define the real number system.

(1) Real numbers: A real number is a regular sequence $\{a_i\}$ of rational numbers; and every regular sequence of rational numbers defines a real number.

(2) *Equality:* Two real numbers $\{a_i\}$ and $\{b_i\}$ are said to be equal, written $\{a_i\} = \{b_i\}$, if and only if for any given positive rational number ϵ, there exists an integer $N(\epsilon)$ such that $|a_i - b_i| < \epsilon$ when $i > N$.

(3) *Addition:* The sum of two real numbers $\{a_i\}$ and $\{b_i\}$ is defined to be the sequence $\{c_i\}$ in which $c_i = a_i + b_i$.

(4) *Multiplication:* The product of two real numbers $\{a_i\}$ and $\{b_i\}$ is defined to be the sequence $\{c_i\}$ in which $c_i = a_ib_i$.

Logically we should now proceed to prove the theorem: "The set *Re* of all real numbers forms a field, known as the real field." But all the details would take us too far afield from our main purpose; so we shall be content with suggesting a few pertinent exercises at the end of this lesson and with referring the reader to other texts, such as that of MacDuffee cited in **1.3**.

It is of particular interest that among the real numbers *Re* there is a subset isomorphic to *Ra*. The subset in question contains all sequences of the type $\{a_i\}$ where $a_i = a$ for all i; these sequences may be designated $\{a\}$ and are of a type where it is trivial to show that they are regular and hence represent real numbers. The suitable correspondence T to establish the isomorphism is defined by $aT = \{a\}$.

Knowing that every rational number a is represented, isomorphically speaking, by a certain real number $\{a\}$, we question whether $\{a\}$ can be written as an infinite decimal. If this proves possible, it will remedy in a sense the difficulties encountered in **31.1** in studying the domain D of finite decimal fractions. For example, it is seen that $1/3$ is not in D; but now we ask whether $\{1/3\}$ may be written as an infinite decimal, i.e., is there a real number of the type $\{a_i\} = b_m \dots b_0 . b_{-1} \dots b_{-1} \dots$ which is "equal" to $\{1/3\}$.

The reader is already familiar with the answer, although the question may never have been put to him in such a hard (precise) way. For we can show

$\{1/3\} = 0.333\dots$ with $b_i = 0$ for $i \geqq 0$ and $b_{-i} = 3$ for $i > 0$.

For if we set $\{c_i\} = \{1/3\}$, with $c_i = 1/3$ for all i; and if we set $\{a_i\} = 0.333\dots$, then we have

$$a_i = (3/10)\{1 + 1/10 + (1/10)^2 + \dots + (1/10)^{i-1}\}$$

and by applying EX. *3.2*, we find

$$a_i = (3/10)\{1 - (1/10)^i\}/(1 - 1/10) = (1/3)\{1 - (1/10)^i\}.$$

Hence $|c_i - a_i| = (1/3)(1/10)^i$ and can be made less than any given positive rational number ϵ by taking i sufficiently large. Therefore by definition (2) it follows that $\{c_i\} = \{a_i\}$ or that $\{1/3\} = 0.333\ldots$.

To answer the same question in the general case will lead us to a situation that is more obviously part of the theory of numbers. By these preliminary sections we hope to have placed the problem on a sound logical basis.

31.4. Periodic infinite decimals. An infinite decimal

$$b_m\ldots b_0.b_{-1}\ldots b_{-i}\ldots$$

will be said to be *periodic* or *repeating* if there exist two integers $s \geqq 0$ and $k > 0$ such that

$$b_{-t} = b_{-t'} \quad \text{whenever } t > s,\ t' > s,\ t \equiv t' \bmod k.$$

For example, it was shown above that $\{1/3\}$ is represented by $0.333\ldots$ which is a periodic infinite decimal having $s = 0$ and $k = 1$.

For a periodic infinite decimal we shall use the notation

$$\{a_i\} = b_m\ldots b_0.b_{-1}\ldots b_{-s}\dot{b}_{-(s+1)}\ldots\dot{b}_{-(s+k)}$$

with a dot above the number $b_{-(s+1)}$ and another dot above $b_{-(s+k)}$; if $k = 1$, only one dot will be required.

We shall call $Q = b_m\ldots b_0$ the "whole number part" of $\{a_i\}$; $S = .b_{-1}\ldots b_{-s}$, the "non-repeating part"; and $P = b_{-(s+1)}\ldots b_{-(s+k)}$, the "repeating part."

In this terminology, the example given above would appear as $0.\dot{3}$, indicating that $Q = 0, S = 0, P = 3, s = 0, k = 1$. In $31.04\dot{1}2\dot{3}$, we have $Q = 31, S = .04, P = 123, s = 2, k = 3$. In $3027.\dot{0}2\dot{7}$, we have $Q = 3027, S = 0, P = 27, s = 0, k = 3$. In 5.0125, a "finite" decimal fraction, we may interpret the notation to indicate that $Q = 5, S = .0125, P = 0, s = 4, k = 1$; sometimes we shall refer to this case as that of a "terminating" decimal; most of the time we shall prefer, for uniformity, to think of this case as a periodic infinite decimal with $P = 0, k = 1$.

Theorem: Every periodic infinite decimal represents a *rational* real number.

Proof: Using the terminology introduced above, if

$$\{a_i\} = b_m\ldots b_0.b_{-1}\ldots b_{-s}\dot{b}_{-(s+1)}\ldots\dot{b}_{-(s+k)},$$

then for $q \geqq 0$ we have, again using EX. 3.2,

$$\begin{aligned} a_{s+qk} &= Q + S + (P/10^{s+k})\{1 + (1/10^k) + \ldots + (1/10^k)^{q-1}\} \\ &= Q + S + (P/10^{s+k})\{1 - (1/10^k)^q\}/(1 - 1/10^k) \\ &= Q + S + \{P/10^s(10^k - 1)\}\{1 - (1/10^k)^q\}. \end{aligned}$$

For any $i \geqq s$, we may set $i - s = (q-1)k + r$, $0 \leqq r < k$, $q \geqq 1$, and have $a_{s+(q-1)k} \leqq a_i < a_{s+qk}$. Then if we set

$$X = Q + S + P/10^s(10^k - 1) \tag{31.2}$$

we have $a_i - X < a_{s+qk} - X = -P/10^{s+qk}(10^k - 1)$. Since $P \leqq 10^k - 1$ and $i < s + qk$, it follows that

$$|a_i - X| < 1/10^{s+qk} < 1/10^i, \qquad i \geqq s.$$

Since we may make $1/10^i < \epsilon$, for any assigned positive rational number ϵ by choosing i sufficiently large, it follows that $\{a_i\} = \{X\}$. Hence we conclude that the given periodic infinite decimal represents the rational number X of *(31.2)* in its "real disguise" of $\{X\}$.

For example, using this theorem we may show that:

$31.04\dot{1}2\dot{3} = 31 + 4/100 + 123/99900 = 31 + 1373/33300$;
$3027.\dot{0}2\dot{7} = 3027 + 27/999 = 3027 + 1/37$;
$5.0125 = 5 + 125/10000 = 5 + 1/80 = 401/80$;
$5.0124\dot{9} = 5 + 124/10000 + 9/90000 = 401/80$.

Converse theorem: Any positive rational real number $\{a/b\}$ may be represented by a periodic infinite decimal.

Proof: With $a > 0$, $b > 0$, we may use the division algorithm to write

$$a = Qb + r_0, \qquad 0 \leqq r_0 < b, \qquad Q \geqq 0. \tag{31.3}$$

Then as in Chapter 4 we may represent Q in the base 10 as $Q = b_m \ldots b_0$. We may use the division algorithm to find

$$10r_0 = q_1 b + r_1, \qquad 0 \leqq r_1 < b.$$

Since $0 \leqq r_0 < b$, it follows on the one hand that $0 \leqq 10r_0$, so that $-b < -r_1 \leqq q_1 b$, hence $0 \leqq q_1$; on the other hand, $10r_0 < 10b$, so that $q_1 b \leqq q_1 b + r_1 < 10b$, hence $0 \leqq q_1 < 10$. We continue the algorithm in this same manner with

$$10r_{i-1} = q_i b + r_i, \qquad 0 \leqq r_i < b, \qquad 0 \leqq q_i < 10, \qquad i \geqq 1, \tag{31.4}$$

until for minimal values of s and k we arrive at $r_{s+k} = r_s$, whereupon we conclude the algorithm. The conclusion will certainly be reached in at most b steps, for from the restriction $0 \leqq r_i < b$ there are only b different possible remainders. We now define

$$S = .q_1 \ldots q_s \quad \text{and} \quad P = q_{s+1} \ldots q_{s+k}$$

and assert that

$$\{a/b\} = b_m \ldots b_0 . q_1 \ldots q_s \dot{q}_{s+1} \ldots \dot{q}_{s+k} \;.$$

One way to establish this equality is to use the direct theorem above which asserts that the periodic infinite decimal which we have constructed is equal to $\{X\}$, where

$$X = Q + S + P/10^s(10^k - 1) \quad .$$

For from the relations (*31.3*) and (*31.4*) we have $a/b = Q + r_0/b$, $0 = q_1/10 - r_0/b + r_1/10b, \ldots, 0 = q_s/10^s - r_{s-1}/10^{s-1}b + r_s/10^sb$, and upon adding these equations and noting the telescoping cancellations, we find

$$a/b = Q + S + r_s/10^sb.$$

Again using (*31.4*) we may write

$$\begin{aligned} r_s/10^sb &= q_{s+1}/10^{s+1} + r_{s+1}/10^{s+1}b, \\ 0 &= q_{s+2}/10^{s+2} - r_{s+1}/10^{s+1}b + r_{s+2}/10^{s+2}b, \ldots, \\ 0 &= q_{s+k}/10^{s+k} - r_{s+k-1}/10^{s+k-1}b + r_{s+k}/10^{s+k}b. \end{aligned}$$

Upon adding these equations we find

$$r_s/10^sb = P/10^{s+k} + r_{s+k}/10^{s+k}b.$$

Recalling that $r_{s+k} = r_s$, we may solve this last equation to show

$$r_s/10^sb = P/10^s(10^k - 1).$$

Combining these results we see that $X = a/b$ which completes the proof.

The reader will recall from elementary arithmetic that the division process represented by (*31.4*) may be carried out very handily by mentally shifting the decimal point one place to the right at each step, corresponding to the multiplication of the preceding remainder by 10. For example, to find the periodic infinite decimal representing $\{2/7\}$ we may arrange our work as follows:

$$\begin{array}{r l}
.285714 & \\
7 \,\overline{)\,2.000000} & r_0 = 2, \quad Q = 0, \\
\underline{1\,4} & \\
60 & r_1 = 6, \quad q_1 = 2, \\
\underline{56} & \\
40 & r_2 = 4, \quad q_2 = 8, \\
\underline{35} & \\
50 & r_3 = 5, \quad q_3 = 5, \\
\underline{49} & \\
10 & r_4 = 1, \quad q_4 = 7, \\
\underline{7} & \\
30 & r_5 = 3, \quad q_5 = 1, \\
\underline{28} & \\
2 & r_6 = 2, \quad q_6 = 4.
\end{array}$$

Since $r_6 = r_0$ we have $s = 0$, $k = 6$, and $\{2/7\} = 0.\dot{2}8571\dot{4}$.

From the standpoint of number theory it is of interest that we can predict the minimal values of s and k in the above theorem without carrying through the complete division algorithm.

Theorem: If $b = 2^x 5^y A$ where $(A,10) = 1$ and if $(a,b) = 1$, then in the division algorithm for finding the periodic infinite decimal representing $\{a/b\}$ the minimal values of s and k for which $r_{s+k} = r_s$ are given as follows: s is the *maximum* of x and y and k is the *exponent to which* 10 *belongs modulo* A.

Proof: Equations (*31.3*) and (*31.4*) may be written as congruences mod b as follows:

$$a \equiv r_0, \quad 10r_{i-1} \equiv r_i \bmod b, \qquad i \geqq 1.$$

Since $10 \equiv 10 \bmod b$, these congruences are equivalent to

$$10^i a \equiv r_i \bmod b, \qquad i \geqq 0.$$

Then to have $r_{s+k} = r_s$, we must have

$$10^{s+k} a \equiv 10^s a \bmod b;$$

and since $(a,b) = 1$, we must have

$$10^{s+k} \equiv 10^s \bmod b$$

for minimal values of s and k. The last congruence requires the existence of an integer t such that

$$10^s(10^k - 1) = tb = t2^x 5^y A.$$

Since 2^x and 5^y are relatively prime to $10^k - 1$, it follows that they must divide 2^s and 5^s, respectively; hence s must be at least as large as the *maximum* of x and y. If we suppose s so chosen, we are able to find a suitable value of k, for the condition above reduces to

$$2^{s-x}5^{s-y}(10^k - 1) = tA.$$

Since $(A,10) = 1$, this implies that A must divide $10^k - 1$, or in terms of congruences that $10^k \equiv 1 \bmod A$. Since $(A,10) = 1$, we may use the language of **21.1** to assert that a positive integer k with this property exists and that the *minimal* value of k which we seek is the exponent to which 10 belongs modulo A. This completes the proof except for the comment that the choice made above guarantees that $q_s \neq q_{s+k}$, for if $q_s = q_{s+k}$, we would have from $r_s = r_{s+k}$ and (*31.4*) that $r_{s-1} = r_{s-1+k}$, which would contradict the argument given above; hence the first digit of the repeating part is definitely q_{s+1} where s is the maximum of x and y.

Thus in considering $\{17/520\}$, when we have found $520 = 2^3 5^1 13$ and $10^6 \equiv 1 \bmod 13$, we see that $s = 3$ and $k = 6$; hence we may predict $\{17/520\} = 0.q_1q_2q_3\dot{q}_4q_5q_6q_7q_8\dot{q}_9$. In fact, by actual computation we find $\{17/520\} = 0.032\dot{6}9230\dot{7}$. If $(a,520) = 1$, then the last theorem shows that $\{a/520\}$ will also have $s = 3$ and $k = 6$.

A few of the numerous corollaries to the above theorem are given in the exercises which follow this lesson. For example, the division algorithm leads to a terminating decimal representation ($k = 1$, $P = 0$) for $\{a/b\}$ if and only if b has the form $b = 2^x 5^y$, see EX. *31.10*.

From the preceding theorems we may be tempted to draw the conclusion that rational real numbers and periodic infinite decimals are in one-to-one correspondence. But this is not quite correct, the missing step in the argument being that we have not investigated the uniqueness of representation by means of infinite decimals. The example $5.0125\dot{0} = 5.0124\dot{9}$, given earlier, provides a partial clue.

If we define two infinite decimals

$$b_m \ldots b_0 . b_{-1} \ldots b_{-i} \ldots \quad \text{and} \quad b'_m \ldots b'_0 . b'_{-1} \ldots b'_{-i} \ldots$$

to be distinct if there exists an integer i such that $b_i \neq b_i'$, then the correct situation is as follows: two distinct infinite decimals represent distinct real numbers, except in the case of terminating infinite decimals ($k = 1$, $P = 0$) when two representations are possible, see EX. *31.11*.

It now follows that an infinite decimal represents a rational real number if and only if the infinite decimal is periodic. For our first theorem states that every periodic infinite decimal represents a rational real number. Our second theorem shows that every rational real number may be represented by at least one periodic infinite decimal. The theorem of the preceding paragraph shows that a rational real number may, in general, be represented by only one infinite decimal; even in the exceptional case there are only two corresponding infinite decimals, and these are both periodic.

Inasmuch as we can write infinite decimals that are not periodic, but which are regular sequences and hence real numbers, it follows that there exist real numbers other than rational real numbers and these are called *irrational* real numbers. It is correct to identify irrational real numbers with non-periodic infinite decimals, since it is possible to show that every real number may be represented by an infinite decimal (see EX. *31.13*), and an irrational real number, by only one infinite decimal, in view of the uniqueness theorem above.

A simple example of an irrational real number is $\sqrt{2}$. For in EX. *15.3* and again in EX. *30.8* we have seen that there is no rational number x/y satisfying $(x/y)^2 = 2$. But by seeking integers X_i such that $X_i^2 < 2(10)^{2i} < (X_i + 1)^2$ we find a real number $\{X_i/10^i\}$ such that $\{X_i/10^i\}^2 = \{2\}$. It follows readily that $10X_i \leqq X_{i+1} < X_{i+1} + 1 \leqq 10X_i + 10$; hence the digits of X_{i+1} differ at most in the units place from those in $10X_i$; thus the digits may be found recursively and $\{X_i/10^i\}$ appears as an infinite decimal, albeit a non-periodic one, obtainable to any desired number of decimal places. In fact by setting $X_{i+1} = 10X_i + q$, $0 \leqq q < 9$, we see that the algorithm proposed above is one of finding the maximum value of q such that $2(10)^{2(i+1)} > X_{i+1}^2 = 10^2X_i^2 + 20X_iq + q^2$ or such that

$$2(10)^{2(i+1)} - 10^2X_i^2 > (20X_i + q)q.$$

This algorithm may be conveniently condensed in the following manner which will be recognized as the "square-root process" given in many elementary arithmetics, usually without proof.

$$\begin{array}{l|rrrrl}
 & 1 & 4 & 1 & 4 & \\ \cline{2-5}
 & 2. & 00 & 00 & 00 & = m, \qquad\qquad X_0 = 1\\
 & 1 & & & & \\ \cline{2-5}
20X_0 + q_1 = 24, & 1 & 00 & & & = 10^2m - (10X_0)^2,\ q_1 = 4,\ X_1 = 14\\
 & & 96 & & & \\ \cline{2-5}
20X_1 + q_2 = 281, & & 4 & 00 & & = 10^4m - (10X_1)^2,\ q_2 = 1,\ X_2 = 141\\
 & & 2 & 81 & & \\ \cline{2-5}
20X_2 + q_3 = 2824, & & 1 & 19 & 00 & = 10^6m - (10X_2)^2,\ q_3 = 4,\ X_3 = 1414\\
 & & 1 & 12 & 96 & \\ \cline{2-5}
 & & & 6 & 04 & = 10^8m - (10X_3)^2.
\end{array}$$

In this manner we find $\sqrt{2} = 1.41421\ldots$.

In the lesson which follows we shall be particularly interested in such quadratic irrationalities for we shall discover a sense in which these are the most regular of irrational real numbers.

31.5. Basimal fractions. It is reasonably clear that the discussion of the preceding sections might well have been made more general by taking any desired fixed integer $B > 1$, not necessarily $B = 10$, as the base number in the representation.

Beginning as in Chapter 4, we know that a given positive integer a may be represented in the form

$$a = a_mB^m + \ldots + a_1B + a_0,\ 0 < a_m < B;\quad 0 \leqq a_i < B, 0 \leqq i < m.$$

We take the liberty of calling a/B^k a "*basimal* fraction," inasmuch as the adjective "decimal" is by its Latin original meaning suitable only when $B = 10$; then paralleling **31.1**, we use a positional notation with a "basimal point" to write

$$a/B^k = b_{m-k}\ldots b_0.b_{-1}\ldots b_{-k} \quad \text{or} \quad a/B^k = 0.0\ldots 0b_{-(k-m)}\ldots b_{-k}$$

according as $k \leqq m$ or $k > m$, with $b_{i-k} = a_i$. Where the context does not indicate the value of B, parentheses and a subscript may be used. For example,

$$(3.1052)_6 = 3 + 1/6 + 5/6^3 + 2/6^4.$$

If $b_m, b_{m-1}, \ldots, b_0, b_{-1}, \ldots, b_{-i}, \ldots$ is a given sequence of integers b_i satisfying $0 \leqq b_i < B$, we may define a corresponding sequence $\{a_i\}$ of rational numbers a_i as follows:

$$a_i = (b_m\ldots b_0.b_{-1}\ldots b_{-i})_B.$$

A sequence of this type will be called an "infinite basimal," designated by $\{a_i\} = (b_m\ldots b_0.b_{-1}\ldots b_{-i}\ldots)_B$. An infinite basimal is a regular sequence (see EX. *31.15*).

An infinite basimal will be said to be *periodic* if there exist two integers $s \geqq 0$ and $k > 0$ such that $b_{-t} = b_{-t'}$ whenever $t > s$, $t' > s$, and $t \equiv t' \bmod k$. A periodic infinite basimal will be denoted by

$$\{a_i\} = (b_m\ldots b_0.b_{-1}\ldots b_{-s}\dot{b}_{-(s+1)}\ldots \dot{b}_{-(s+k)})_B.$$

Every periodic infinite basimal represents a *rational* real number $\{X\}$ (see EX. *31.16*). Conversely, every positive rational real number may be represented by a periodic infinite basimal (see EX. *31.17*).

If in standard form $B = p_1^{s_1}p_2^{s_2}\ldots p_k^{s_k}$, $s_i > 0$, $p_i < p_{i+1}$, if $b = p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}A$, where $(A,B) = 1$, and if $(a,b) = 1$, then in the periodic infinite basimal representing $\{a/b\}$ to the base B the minimal value for s is the smallest *integer* greater than or equal to the maximum of $x_1/s_1, x_2/s_2, \ldots, x_k/s_k$ and the minimal value for k is the exponent to which B belongs modulo A (see EX. *31.18*).

For example, if $B = 2^2 3$, we may use $x = 9 + 1$, $L = x + 1$, $B = L + 1$ and may consider finding the infinite basimal to represent $\{15/260\}_B$. Since $(260)_B = (360)_{10} = 2^3 3^2 5$, we have $x_1/s_1 = 3/2$, $x_2/s_2 = 2$ so that $s = 2$. Since $B \equiv 2 \bmod 5$, we find that B belongs to 4 mod 5 so that $k = 4$. By the theorem we may predict that

$\{15/260\}_B = (0.q_1q_2\dot{q}_3q_4q_5\dot{q}_6)_B$. The division algorithm in the base B appears as follows:

$$\begin{array}{rl} & .069724 \\ \cline{2-2} 260 & \multicolumn{1}{|l}{15.00} \\ & \multicolumn{1}{|l}{13\,00} \\ \cline{2-2} & 2\,000 \\ & 1\,x60 \\ \cline{2-2} & 1600 \\ & 1560 \\ \cline{2-2} & 600 \\ & 500 \\ \cline{2-2} & 1000 \\ & x00 \\ \cline{2-2} & 200 \end{array} \qquad \begin{array}{l} \text{base } B = 2^2 3 \\ x = 9 + 1 \\ L = x + 1 \\ B = L + 1 \end{array}$$

Hence $\{15/260\}_B = (0.06\dot{9}72\dot{4})_B$, as predicted.
It is of some interest to see that the same rational number expanded in various bases may have different periodic character. Thus for the example just given $(15/260)_B = (17/360)_{10}$. Since $10 = 2^1 5^1$ we find $s = 3$; since $10 \equiv 1 \bmod 9$ we have $k = 1$; in fact $(17/360)_{10} = (0.047\dot{2})_{10}$.

A periodic infinite basimal is said to be terminating if $k = 1$, $P = 0$. A rational real number $\{a/b\}$ has a terminating basimal representation to the base $B = p_1^{s_1}p_2^{s_2}\ldots p_k^{s_k}$ only if $b = p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}$, $x_i \geqq 0$.

Two infinite basimals are said to be distinct if for some integer t, $b_t \neq b_t'$. Distinct infinite basimals represent unequal real numbers, except in the case of terminating infinite basimals for which two representations are possible, one with $k = 1$, $P = 0$, the other with $k = 1$, $P = B - 1$.

Finally, every real number may be represented by an infinite basimal: rational real numbers if and only if the infinite basimal is periodic; irrational real numbers if and only if the infinite basimal is non-periodic.

EXERCISES

EX. *31.1.* Prove that equality of real numbers is *transitive*.

EX. *31.2.* Prove that the *sum* of two regular sequences is a *regular* sequence, so that *addition* of real numbers is *closed*.

EX. *31.3.* Prove that *addition* of real numbers is *well defined.*

EX. *31.4.* Prove that *multiplication* of real numbers is *closed.*

EX. *31.5.* Find in lowest terms the rational real numbers represented by the following *infinite decimals:*

(a) $0.03\dot{0}027\dot{1}$; (b) $1.\dot{0}12\dot{1}$; (c) $0.\dot{1}76470588235294\dot{1}$.

EX. *31.6.* *Predict* the *form* of the periodic infinite decimals representing the following rational real numbers:

(a) $\{3/410\}$; (b) $\{25/11\}$; (c) $\{16/27\}$, (d) $\{355/4004\}$.

EX. *31.7.* If $PP' = 10^k - 1$, discuss the repeating parts of the periodic infinite decimals representing $\{1/P\}$ and $\{1/P'\}$.

EX. *31.8.* If $b = 2^x5^yA$ with $(A,10) = 1$, if $(a,b) = 1$, if $(a',b) = 1$, and if $a \equiv a' \bmod A$, show that $\{a/b\}$ and $\{a'/b\}$ have the same repeating part P, in their infinite decimals.

EX. *31.9.* If $\{a/b\}$ has remainders r_i and a repeating part $P = q_{s+1} \cdots q_{s+k}$, if $(a',b) = 1$ and if $10^sa' \equiv r_{s+i} \bmod b$, show that the infinite decimal representing $\{a'/b\}$ has a repeating part P' obtained from P by a cyclic advancement of digits.

EX. *31.10.* Show that the division algorithm leads to a terminating decimal representation for $\{a/b\}$ if and only if b has the form $b = 2^x5^y$.

EX. *31.11.* Show that distinct infinite decimals represent unequal real numbers, except in the case of terminating decimals when two representations are possible. (*Hint:* Consider four cases, each with t maximal such that $b_t \neq b_t'$: (*1*) $b_t > b_t' + 1$; (*2*) $b_t = b_t' + 1$, $b_u > 0$ for a maximal $u < t$; (*3*) $b_t = b_t' + 1$, $b_u = 0$ for $u < t$, $b_v' < 9$ for a maximal $v < t$; (*4*) $b_t = b_t' + 1$, $b_u = 0$ for $u < t$, $b_v' = 9$ for $v < t$.)

EX. *31.12.* If $\{a_i\}$ and $\{b_i\}$ are regular sequences, define $\{a_i\} > \{b_i\}$ if and only if there exists a rational number $\epsilon > 0$ and an integer N such that $a_i - b_i > \epsilon$ for $i > N$. Prove that this order relation for real numbers is *tricholomous.*

EX. *31.13.* Use EX. *31.12* to show that any real number $\{a_i\}$ is equal to an infinite decimal.

EX. *31.14.* Show that the algorithm represented by finding integers X_i such that $X_i^3 < 2(10)^{3i} < (X_i + 1)^3$ constructs an infinite decimal $\{X_i/10^i\}$ which is a real irrational cube root of 2. Explain why this algorithm cannot be so conveniently condensed as the one for square roots.

EX. *31.15.* Establish the analog of (*31.1*) and show that an infinite basimal is a regular sequence.

EX. *31.16.* Show that a periodic infinite basimal represents a rational real number $\{X\}$, $X = Q + S + P/B^s(B^k - 1)$, analogous to (*31.2*).

EX. *31.17.* Using the analogues of (*31.3*) and (*31.4*) show that every positive rational number may be represented by a periodic infinite basimal.

EX. *31.18.* For a periodic infinite *basimal* representing $\{a/b\}_B$ establish the theorem given in the text for the minimal values of s and k.

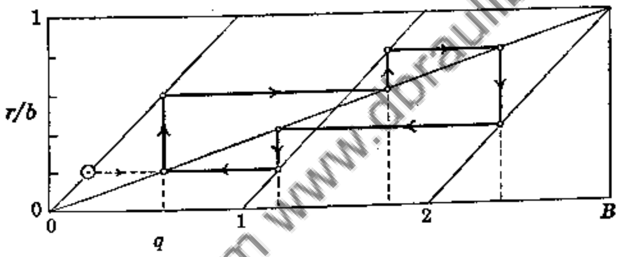
EX. *31.19.* In regard to periodic basimals show that k must divide $\lambda(A)$ and $\phi(A)$. (Recall **G.15.1** and EX. *18.7.*)

EX. *31.20.* If $b = (11)_{10}$, investigate $(1/b)_B$ for $B = 2,3,4,5,6,7,8,9,10,12$; compare with **G.17**.

EX. *31.21.* If $b = (13)_{10}$, investigate $(a/b)_5$ for $a = 1,2,\ldots,b-1$; compare with EX. *31.9*.

EX. *31.22.* Show that (a) $(1/(B-1))_B = 0.\dot{1}$; (b) $(1/(B+1))_B = 0.\dot{0}(B \dot{-} 1)$; (c) $(1/(B-1)^2)_B = 0.\dot{0}123\ldots(B-3)(B \dot{-} 1)$.

EX. *31.23.* Establish a graphic picture of the periodic character of $\{a/b\}_B$ by using $F(x) = x - [x]$ and $L(x) = x/B$ for $0 \leqq x < B$, starting with $F(r_0/b)$ on F and proceeding alternately, horizontally to L and vertically to F. Thus the tailpiece to this chapter illustrates $\{1/12\}_3 = (0.\dot{0}12\dot{1})_3$.



►*I shall set forth the method of forming fractions which is most pleasing to me today and it will rest in men's judgment to appraise what they see.* —R. BOMBELLI

CHAPTER 32[*]

CONTINUED FRACTIONS

32.1. Finite continued fractions. Given the non-negative integer b_0 and the positive integers $b_1,b_2,\ldots,b_n$ we may define, recursively, the following integers p_i and q_i:

$$(32.1)\qquad \begin{aligned} p_{-1} &= 1,\ p_0 = b_0,\ p_i = b_i p_{i-1} + p_{i-2},\ 1 \leqq i \leqq n; \\ q_{-1} &= 0,\ q_0 = 1,\ q_i = b_i q_{i-1} + q_{i-2},\ 1 \leqq i \leqq n. \end{aligned}$$

We shall call $a_n = p_n/q_n$ a *finite continued fraction* and denote its dependence upon $b_0,b_1,\ldots,b_n$ by the following symbol:

$$a_n = \{b_0,b_1,\ldots,b_n\}.$$

We shall call $a_i = p_i/q_i$, for $0 \leqq i \leqq n$, the ith *convergent* of the continued fraction. Evidently the notation is so chosen that each convergent is itself a continued fraction, i.e.,

$$a_i = \{b_0,b_1,\ldots,b_i\}, \qquad 0 \leqq i \leqq n.$$

The name, continued fraction, can best be explained by reviewing the notion of a *complex* fraction. We have seen that when $c \neq 0$, then $(c/d)(x/y) = (a/b)$ has the solution $x/y = ad/bc$. However, by analogy with the situation when $b \neq 0$ and $(b/1)(x/y) = (a/1)$ has

[*] Chapter 32 is a basic chapter, but one which will require some knowledge of fractions and real numbers such as is given in the preceding supplementary chapters.

the solution $x/y = a/b$, it is natural to write the solution to the first given equation in the form $(a/b)/(c/d)$ or $\dfrac{a/b}{c/d}$, with appropriate parentheses or a longer fraction bar or vinculum to explain the order of operation which is intended. Such a "fraction" with other fractions in its "numerator" or "denominator" is called a *complex fraction*. By repeated use of the rules for ordinary fractions and of the definitions above in which

$$(a/b)/(c/d) = ad/bc$$

we may reduce a complex fraction to an ordinary or *simple* fraction.

A continued fraction makes a good exercise in this reduction technique and the exercise reveals why the name *continued* fraction is relevant. For we soon discover that from

$$a_{i-1} = \frac{p_{i-1}}{q_{i-1}} = \frac{b_{i-1}p_{i-2} + p_{i-3}}{b_{i-1}q_{i-2} + q_{i-3}}, \qquad i \geqq 2$$

we can obtain $a_i = p_i/q_i$ by replacing b_{i-1} by $b_{i-1} + 1/b_i$. Thus

$$\frac{\left(b_{i-1} + \frac{1}{b_i}\right)p_{i-2} + p_{i-3}}{\left(b_{i-1} + \frac{1}{b_i}\right)q_{i-2} + q_{i-3}} = \frac{b_i(b_{i-1}p_{i-2} + p_{i-3}) + p_{i-2}}{b_i(b_{i-1}q_{i-2} + q_{i-3}) + q_{i-2}} =$$

$$\frac{b_ip_{i-1} + p_{i-2}}{b_iq_{i-1} + q_{i-2}} = \frac{p_i}{q_i}.$$

Hence, for example, starting with $p_0/q_0 = b_0/1 = b_0$ we find

$$\frac{p_1}{q_1} = b_0 + \frac{1}{b_1}, \quad \frac{p_2}{q_2} = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2}}, \quad \frac{p_3}{q_3} = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3}}}.$$

Thus p_n/q_n is really an "n-storied" complex fraction, meaning that fraction bars of n different lengths could be used to indicate its structure. It is readily appreciated that the notation $a_n = \{b_0, b_1, \ldots, b_n\}$ and the recursive relations *(32.1)* afford a much less cumbersome symbolism.

In the same manner if we set $R_{i,k} = \{b_i, \ldots, b_k\}$, $1 \leqq i \leqq k \leqq n$, and replace b_i by $R_{i,k}$ in the expressions for p_i and q_i, we find that

$$(32.2) \qquad \frac{p_k}{q_k} = \frac{R_{i,k}p_{i-1} + p_{i-2}}{R_{i,k}q_{i-1} + q_{i-2}}.$$

If the integers b_i are small, the successive computations required by *(32.1)* may be done mentally and entered in the following chart, working from left to right:

b		b_0	b_1	b_2	b_3	b_4	$\dots$	b_n
p	1	b_0	p_1	p_2	p_3	p_4	$\dots$	p_n
q	0	1	q_1	q_2	q_3	q_4	$\dots$	q_n

Comparison with *(12.5)* and the example which follows *(12.5)* will be interesting and suggestive.

For theoretical purposes it is worth while to note that the relations *(32.1)* may be written in matric form:

$$\begin{pmatrix} p_0 & q_0 \\ p_{-1} & q_{-1} \end{pmatrix} = \begin{pmatrix} b_0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} p_i & q_i \\ p_{i-1} & q_{i-1} \end{pmatrix} = \begin{pmatrix} b_i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_{i-1} & q_{i-1} \\ p_{i-2} & q_{i-2} \end{pmatrix}, \qquad i \geqq 1.$$

Then by an easy induction it follows that

$$\begin{pmatrix} p_i & q_i \\ p_{i-1} & q_{i-1} \end{pmatrix} = \begin{pmatrix} b_i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} b_{i-1} & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} b_0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad i \geqq 0.$$

Recalling **M.7** in **11.3** concerning determinants, we have the following useful result:

$$(32.3) \qquad p_i q_{i-1} - p_{i-1} q_i = (-1)^{i+1}, \qquad i \geqq 0.$$

It follows at once that the numerator and denominator of any convergent are relatively prime, i.e., $(p_i, q_i) = 1$, so the convergents are automatically in lowest terms.

C.1: A finite continued fraction represents a positive rational number; conversely, a positive rational number may be represented as a finite continued fraction.

Proof: The direct proposition is obvious from the definition of a finite continued fraction. For the converse if the given *positive* rational number is x/y we consider the Euclid algorithm for finding $d = (x/y)$ and rewrite the equations of the algorithm in the following manner:

$$\begin{array}{ll}
x = b_0 y + r_0,\ 0 < r_0 < y, & x/y = b_0 + r_0/y = b_0 + 1/(y/r_0); \\
y = b_1 r_0 + r_1,\ 0 < r_1 < r_0, & y/r_0 = b_1 + r_1/r_0 = b_1 + 1/(r_0/r_1); \\
\dots & \qquad \dots \qquad \dots \\
r_{k-2} = b_k r_{k-1} + r_k,\ 0 < r_k < r_{k-1}, & r_{k-2}/r_{k-1} = b_k + r_k/r_{k-1}; \\
r_{k-1} = b_{k+1} r_k, \qquad 0 = r_{k+1}, & r_k/r_{k-1} \quad = 1/b_{k+1}.
\end{array}$$

If $x \geqq y$, then $b_0 > 0$; if $x < y$, then $b_0 = 0$. In the other equations we have $b_i > 0$, $i = 1,\ldots,k+1$. Hence the finite continued fraction $\{b_0,b_1,\ldots,b_{k+1}\}$ exists and by the equations above we see that this continued fraction, considered as a $k+1$-storied complex fraction, is equal to x/y.

Corollary: If $(x,y) = d$, to find integers s and t such that $xs - ty = \pm d$ it suffices to expand x/y as a finite continued fraction, say $x/y = \{b_0,b_1,\ldots,b_n\}$ and take $t = p_{n-1}$, $s = q_{n-1}$.

Proof: By the theorem a finite continued fraction representing x/y exists. Since $(p_n,q_n) = 1$ and $x/y = p_n/q_n$, it follows that $x = p_n d$, $y = q_n d$. Since *(32.3)* holds, we may multiply *(32.3)* by d and obtain $xq_{n-1} - p_{n-1}y = (-1)^{n+1}d$. Thus with $s = q_{n-1}$ and $t = p_{n-1}$ we have $xs - ty = \pm d$.

For example, since $19/15 = 1 + 4/15$, $15/4 = 3 + 3/4$, and $4/3 = 1 + 1/3$, we have $19/15 = \{1,3,1,3\}$. From the table of convergents:

b		1	3	1	3
p	1	1	4	5	19
q	0	1	3	4	15

we check that $(19)(4) - (5)(15) = 1$ which illustrates the corollary with $x = 19$, $y = 15$, $s = 4$, $t = 5$, $d = 1$.

Using *(32.2)* and *(32.3)* for $1 \leqq i < k \leqq n$ we find

$$(32.4)\quad a_k - a_i = \frac{R_{i+1,k}p_i + p_{i-1}}{R_{i+1,k}q_i + q_{i-1}} - \frac{p_i}{q_i} = \frac{(-1)^i}{q_i(R_{i+1,k}q_i + q_{i-1})}.$$

In particular, when $k = i + 1$, we find $R_{i+1,i+1} = b_{i+1}$ so that

$$(32.5)\qquad a_{i+1} - a_i = (-1)^i/q_iq_{i+1}, \qquad 1 \leqq i < n.$$

Again, when $k = i + 2$, we find $R_{i+1,i+2} = b_{i+1} + 1/b_{i+2}$ so that

$$(32.6)\qquad a_{i+2} - a_i = (-1)^i b_{i+2}/q_iq_{i+2}, \qquad 1 \leqq i < n - 1.$$

C.2: For a finite continued fraction the successive convergents always have the following order: those of even subscript occur in increasing order; those of odd subscript occur in decreasing order; and every convergent of odd subscript is greater than every one of even subscript.

Proof: Since $b_i \geqq 1$ for $i \geqq 1$, it follows that $q_i \geqq 1$ for $i \geqq 1$ and that $R_{i,k} \geqq 1$ for $k \geqq i \geqq 1$. Hence when i is *even*, *(32.6)* shows $a_{i+2} > a_i$; but when i is *odd*, *(32.6)* shows $a_{i+2} < a_i$; and these inequalities prove the first two parts of **C.2.** Let i be any *even* integer; then for any odd integer k with $k > i$ we may use *(32.4)* to see that $a_k > a_i$; since we have previously shown for *odd* K and k that $a_K > a_k$ when $K < k$, it follows that for *any* odd k and any even i we have $a_k > a_i$, which completes the proof.

For example, with $a_6 = \{2,1,4,2,1,12,3\}$ we find

b		2	1	4	2	1	12	3
p	1	2	3	14	31	45	571	1758
q	0	1	1	5	11	16	203	625

Here $\quad a_0 = 2 < a_2 = 14/5 < a_4 = 45/16 < a_6 = 1758/625$
and $\quad a_6 = 1758/625 < a_5 = 571/203 < a_3 = 31/11 < a_1 = 3.$

An important property of a convergent to a finite continued fraction is that it is a closer approximation to the value of the continued fraction than any rational number of smaller denominator. It is this property which has given the study of continued fractions many practical applications.

C.3: If $a_n = \{b_0,b_1,\ldots,b_n\}$ and if $a_i = p_i/q_i = \{b_0,b_1,\ldots,b_i\}$ with $1 \leqq i < n$ and if q is an integer with $0 < q < q_i$, then

$$|a_n - a_i| < |a_n - p/q|$$

for all integers p.

Proof: The proof is by contradiction. If we set $T_{i+1} = R_{i+1,n}q_i + q_{i-1}$, then *(32.4)* shows $a_n - a_i = (-1)^i/q_iT_{i+1}$. Hence if we suppose p/q closer to a_n then a_i we must have

$$-1/q_iT_{i+1} < (-1)^{i+1}(a_n - p/q) < 1/q_iT_{i+1}$$

where the sign $(-1)^{i+1}$ has been introduced to use in what follows. Adding $1/q_iT_{i+1}$ to these inequalities, we find

$$0 < (-1)^{i+1}(a_n - (-1)^i/q_iT_{i+1} - p/q) = (-1)^{i+1}(a_i - p/q) < 2/q_iT_{i+1}.$$

Noting that $R_{i+1},n \geqq b_{i+1}$ we have $T_{i+1} \geqq q_{i+1}$ and may write

$$0 < (-1)^{i+1}(p_i/q_i - p/q) < 2/q_iq_{i+1}.$$

Multiplying these inequalities by the *positive* integer qq_i we find

$$0 < (-1)^{i+1}(p_iq - pq_i) < 2q/q_{i+1}.$$

Since $q_1 = b_1 \geqq 1$, it follows from *(32.1)* that $q_i < q_{i+1}$ for $i \geqq 1$. Since $q < q_i < q_{i+1}$, we have $2q/q_{i+1} < 2$, so the inequalities last displayed require the central *integer* to have the value $+1$; hence $p_iq - pq_i = (-1)^{i+1}$. But this is a Diophantine equation which by the theory in **12.1** can have *at most one* solution with $0 < q < q_i$. By *(32.3)* there *is* such a solution: namely, $p = p_{i-1}$, $q = q_{i-1}$. However with $q = q_{i-1}$ the inequality $1 < 2q/q_{i+1}$ or $q_{i+1} < 2q_{i-1}$ cannot hold; for since $q_{i+1} = b_{i+1}q_i + q_{i-1}$, it follows that $q_{i+1} \geqq 2q_{i-1}$ for $i \geqq 1$. This contradiction establishes the theorem.

For illustration we may take the previous example and assert that 45/16 is a better approximation to 1758/625 than any rational number p/q with $0 < q < 16$, such as 27/10 or 39/14.

32.2. Infinite continued fractions. Given any sequence of rational integers $b_0,b_1,\ldots,b_i,\ldots$ with $b_0 \geqq 0$ and $b_i \geqq 1$ for $i \geqq 1$, we may define a corresponding sequence $\{a_i\}$ of rational numbers by setting $a_i = \{b_0,b_1,\ldots,b_i\}$. Such a sequence $\{a_i\}$ is called an *infinite continued fraction* and may be designated by

$$\{a_i\} = \{b_0,b_1,\ldots,b_i,\ldots\}.$$

C.4: An infinite continued fraction is a regular sequence.

Proof: Since *(32.4)* is valid here for $1 \leqq i < k$ and since $R_{i+1,k} \geqq b_{i+1}$ so that $R_{i+1,k}q_i + q_{i-1} \geqq q_{i+1}$, we have

$$(32.7) \qquad |a_k - a_i| < 1/q_iq_{i+1}, \qquad 1 \leqq i < k.$$

Since the q_i form an increasing sequence of rational integers with $q_{i+1} > q_i > i$ for $i > 3$ (see EX. *32.5*), we may apply *(32.7)* with $i = N$ to see that

$$|a_N - a_k| < 1/q_Nq_{N+1} < 1/N^2 \qquad \text{for } 3 < N < k.$$

Since for any assigned rational number $\epsilon > 0$, we may find a positive integer $N > 3$ such that $N^2 > 1/\epsilon$ and thus $1/N^2 < \epsilon$, it follows that we can guarantee that $|a_N - a_k| < \epsilon$ for $k > N$. Therefore by the definition in **31.2** the sequence $\{a_i\}$ is regular.

To continue this development we need to use the notions of order, absolute value, and the bracket function for real numbers; the intuitive descriptions of Chapters 1 and 9 will suffice; a rigorous development, following the line of EX. *31.12*, may be found in the source cited in **31.3.**

C.5: Any positive irrational real number x may be represented by an infinite continued fraction.

Proof: The following process is, in general, not complete in a finite number of steps, so it cannot properly be called an algorithm; rather it is a recursive process that can be carried out to any desired number of steps:

$$\begin{aligned} &b_0 = [x], \quad r_0 = x - b_0; \\ &b_i = [1/r_{i-1}], \quad r_i = 1/r_{i-1} - b_i, \qquad i \geqq 1. \end{aligned}$$

By definition of the bracket function we would naturally have $0 \leqq r_i < 1$, $i \geqq 0$; but since x is irrational, each r_i must be irrational, so the 0 value is excluded, and we have $0 < r_i < 1$, $i \geqq 0$. Then $1/r_i$ is defined and $1/r_i > 1$, $i \geqq 0$, so that $b_i \geqq 1$, $i \geqq 1$. Since $x > 0$, we have $b_0 \geqq 0$. Thus the b_i are proper elements for an infinite continued fraction and we claim that $x = \{a_i\} = \{b_0, b_1, \ldots, b_i, \ldots\}$.

By induction we may see that the construction above makes

$$(32.8) \qquad x = \frac{(b_{i+1} + r_{i+1})p_i + p_{i-1}}{(b_{i+1} + r_{i+1})q_i + q_{i-1}}, \qquad i \geqq 0.$$

(I) When $i = 0$ we have, using *(32.1)*,

$$x = b_0 + r_0 = b_0 + 1/(b_1 + r_1) = \frac{(b_1 + r_1)b_0 + 1}{(b_1 + r_1)1 + 0} = \frac{(b_1 + r_1)p_0 + p_{-1}}{(b_1 + r_1)q_0 + q_{-1}}.$$

(II) If we assume *(32.8)* correct for i, we have only to use *(32.1)* and $r_{i+1} = 1/(b_{i+2} + r_{i+2})$ to be able to write

$$x = \frac{p_{i+1} + r_{i+1}p_i}{q_{i+1} + r_{i+1}q_i} = \frac{(b_{i+2} + r_{i+2})p_{i+1} + p_i}{(b_{i+2} + r_{i+2})q_{i+1} + q_i},$$

which is the form that *(32.8)* should take for the case $i + 1$.

From *(32.8)* and *(32.3)* it follows that

$$(32.9) \qquad x - a_i = (-1)^i/q_i(q_{i+1} + r_{i+1}q_i), \qquad i \geqq 0.$$

Since $0 < r_{i+1}$ and $i < q_i < q_{i+1}$ when $i > 3$, it follows that $|x - a_i| < 1/q_iq_{i+1} < 1/i^2$ for $i > 3$. Hence by taking i sufficiently large, say $i > N$ where $N^2 > 1/\epsilon$, and where ϵ is any assigned positive real number, we may make $|x - a_i| < \epsilon$, for $i > N$, which proves that $x = \{a_i\}$.

It is worth while to note that if x is a *rational* number the process described above becomes exactly the Euclid algorithm, but since the process will now terminate in a finite number of steps, the continued fraction that is obtained to represent x is finite, instead of infinite.

C.6: If $x = \{a_i\} = \{b_0, b_1, \ldots, b_i, \ldots\}$ is an infinite continued fraction, then the successive convergents always have the following order: those of even subscript occur in increasing order and all are less than x; those of odd subscript occur in decreasing order and all are greater than x. If q is an integer such that $0 < q < q_i$, then

$$|x - a_i| < |x - p/q|$$

for all integers p.

Proof: Concerning the order of the convergents the proofs are the same as in **C.2**. We may use *(32.9)* to see the order relation between x and a_i, according as i is even or odd. The result concerning closeness of approximation is proved exactly as in **C.3** *except* for replacing a_n by x and $R_{i+1,n}$ by $R_{i+1} = \{b_{i+1}, b_{i+2}, \ldots\} = b_{i+1} + r_{i+1}$ and using *(32.9)* instead of *(32.4)*.

To illustrate these theorems we may study x, where $x^2 = 21$, and carry forward the recursive process of **C.5** as follows: $x = 4 + r_0$,

$$\begin{aligned}
1/r_0 &= 1/(x - 4) = (x + 4)/5 = 1 + r_1,\\
1/r_1 &= 5/(x - 1) = (x + 1)/4 = 1 + r_2,\\
1/r_2 &= 4/(x - 3) = (x + 3)/3 = 2 + r_3,\\
1/r_3 &= 3/(x - 3) = (x + 3)/4 = 1 + r_4,\\
1/r_4 &= 4/(x - 1) = (x + 1)/5 = 1 + r_5,\\
1/r_5 &= 5/(x - 4) = (x + 4)/1 = 8 + r_6;
\end{aligned}$$

but at this point we find $r_6 = r_0$, so the process now repeats itself. For an adequate notation for this phenomenon let us digress from the illustration to make certain definitions.

An infinite continued fraction $x = \{b_0, b_1, \ldots, b_i, \ldots\}$ will be said to be *periodic* if there exist integers $s \geqq 0$ and $k > 0$ such that whenever $t > s$, $t' > s$, and $t \equiv t' \bmod k$, then $b_t = b_{t'}$. We may use the same convention as with periodic infinite decimals to write a periodic infinite continued fraction as follows:

$$x = \{b_0, \ldots, b_s, \dot{b}_{s+1}, \ldots, \dot{b}_{s+k}\}.$$

With this agreement the example worked above becomes $x =$

$\{4,\dot{1},1,2,1,1,\dot{8}\}$ with $s = 0$ and $k = 6$. The first few convergents for x appear in the following table:

b		4	1	1	2	1	1	8	1
p	1	4	5	9	23	32	55	472	527
q	0	1	1	2	5	7	12	103	115

In illustration of part of **C.6** we find the convergents are arranged as follows:

$$4 < 9/2 < 32/7 < 472/103 < x < 527/115 < 55/12 < 23/5 < 5.$$

By (*32.9*) we know 472/103 is an approximation to x correct to within $1/(103)(115) = 1/11845$ and by **C.6** that it is a better approximation than any other rational number of denominator less than 103.

A real number of the type $(A + \sqrt{M})/C$ where A, C, M are rational integers with $C \neq 0$, $M > 0$, and M *not* a perfect square is called a *real quadratic surd.* Such real numbers take on a peculiar interest from the standpoint of continued fractions because of the following theorem (and its converse).

C.7: A periodic infinite continued fraction represents a real quadratic surd.

Proof: Given $x = \{b_0,\ldots,b_s,\dot{b}_{s+1},\ldots,\dot{b}_{s+k}\}$ from (*32.8*) we may write

$$x = \frac{R_{s+1}p_s + p_{s-1}}{R_{s+1}q_s + q_{s-1}}$$

and make the proof depend upon evaluating $R_{s+1} = \{\dot{b}_{s+1},\ldots,\dot{b}_{s+k}\}$. This latter continued fraction by its periodic character has the property $R_{k+1}' = \{b_{s+k+1},b_{s+k+2},\ldots\} = R_{s+1}$. If we denote the convergents of R_{s+1} by $p_i'/q_i' = \{b_{s+1},\ldots,b_{s+i+1}\}$, then by (*32.8*) we have

$$R_{s+1} = \frac{R_{k+1}'p_k' + p_{k-1}'}{R_{k+1}'q_k' + q_{k-1}'} = \frac{R_{s+1}p_k' + p_{k-1}'}{R_{s+1}q_k' + q_{k-1}'}.$$

Hence R_{s+1} is a solution of the equation

$$q_k'R_{s+1}^2 + (q_{k-1}' - p_k')R_{s+1} - p_{k-1}' = 0.$$

Since this equation has rational integers as coefficients, its solutions, obtained by the quadratic formula, are of the form $(A + \sqrt{M})/C$ with $C = 2q_k' \neq 0$. Since the discriminant $M = (q_{k-1}' - p_k')^2 + 4q_k'p_{k-1}' > 0$, the roots are real; furthermore, only one root is positive, so R_{s-1} is uniquely determined as this positive root. Finally, M is not a perfect square and R_{s+1} is irrational, for if R_{s+1} were rational, its representation as a continued fraction would be finite (see EX. *32.7*). If we substitute the value of R_{s+1} in x and rationalize the denominator, we find that x is also a quadratic surd.

For example if $x = \{2,\dot{1},\dot{3}\}$, we set $R = \{\dot{1},\dot{3}\} = 1 + 1/(3 + 1/R)$ and find $3R^2 - 3R - 1 = 0$, whence $R = (3 + \sqrt{21})/6$. Then $x = 2 + 1/R = 2 + 6/(3 + \sqrt{21}) = 2 + (\sqrt{21} - 3)/2 = (1 + \sqrt{21})/2$.

It is more difficult to prove that every real quadratic surd may be represented by a periodic infinite continued fraction (see Hardy and Wright, *Theorem 177*). In particular, surds of the type $\sqrt{M}$ always have $s = 0$ and repeating parts that are rather symmetric

$$\sqrt{M} = \{b_0,\dot{b}_1,b_2,\ldots,b_2,b_1,2\dot{b}_0\}.$$

(See H. S. Hall and S. R. Knight, *Higher Algebra*, Fourth Edition, §363, New York, Macmillan, 1940.)

These results show in a fascinating way, from the standpoint of continued fractions, that the quadratic are better behaved than the other irrationalities.

EXERCISES

EX. *32.1.* Expand the decimal fraction 3.1415926535 as a continued fraction and show that $a_3 = 355/113$ is a correct approximation to within $1/(113)(33102)$.

EX. *32.2.* Expand the decimal fraction 2.71828 as a continued fraction.

EX. *32.3.* Use a continued fraction to help solve $53s - 17t = 1$.

EX. *32.4.* Establish formula (*32.2*) by induction on k.

EX. *32.5.* Prove that $q_i > i$ for $i > 3$.

EX. *32.6.* Show that $\{1,3,4\} = \{1,3,3,1\}$.

EX. *32.7.* Show that representation by a continued fraction is unique except for finite continued fractions where exactly two representations are possible (both finite). See EX. *32.6*.

EX. *32.8.* Prove that $\{a,2\dot{a}\} = \sqrt{a^2 + 1}$ and that $\{2a,\dot{a},4\dot{a}\} = 2\sqrt{a^2 + 1}$.

EX. *32.9.* Prove that $\{a,\dot{a},2\dot{a}\} = \sqrt{a^2 + 2}$.

EX. *32.10.* Show that $\sqrt{7}$ has $k = 4$, but $\sqrt{19}$ has $k = 6$.

EX. *32.11.* Show that $\{3a,2\dot{a},6\dot{a}\} = \sqrt{9a^2 + 3}$.

EX. *32.12.* Show that $\sqrt{41}$ has $k = 3$ and $\sqrt{13}$ has $k = 5$.

EX. *32.13.* Use (*32.9*) and (*32.5*) to show that $|x - a_{i+1}| < |x - a_i|$ and that $1/2q_iq_{i+1} < |x - a_i| < 1/q_iq_{i+1}$.

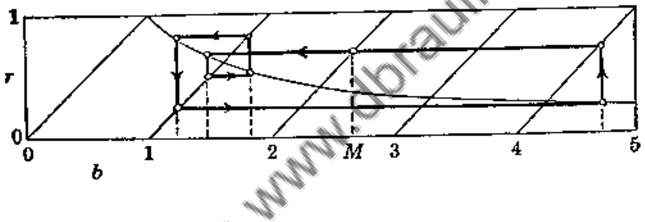
EX. *32.14.* Investigate Fibonacci numbers, which are the p_i (or q_i) in $x = \{\dot{1}\}$, independently or in reference books.

EX. *32.15.* Investigate Pell's Diophantine equation $x^2 - Ay^2 = N$, independently or in reference books.

EX. *32.16.* Investigate in reference books the use of continued fractions in the design of gears.

EX. *32.17.* On the usual coordinate system where *points* are designated by $P = (x,y)$ and $O = (0,0)$, plot the lattice points $P_i = (q_i,p_i)$. Use $a_i = p_i/q_i =$ slope OP_i to illustrate graphically *every* part of theorem **C.6** (F. Klein).

EX. *32.18.* Establish a graphic picture of the periodic character of the continued fraction representing $\sqrt{M}$ by using $F(x) = x - [x]$ for $x \geqq 0$ and $H(x) = 1/x$ for $x > 1$, starting with $F(\sqrt{M})$ on F and proceeding alternately, horizontally to H and vertically to F. Thus the tailpiece to this chapter illustrates $\sqrt{7} = \{2,\dot{1},1,1,\dot{4}\}$.



> ▶*When I use a word, it means just what I choose it to mean—neither more, nor less.*
> —H. DUMPTY; L. CARROLL; C. DODGSON

CHAPTER 33[*]

THE FUNDAMENTAL THEOREM RECONSIDERED

33.1. The domain $[Ra\sqrt{10}]$. In 29.3 we studied why the set of all ordinary integers might be described as the rational domain $[Ra]$, and it was suggested for explicitness that the ordinary integers be described as rational integers to distinguish them from the elements of other number systems which might with equal right also be called "integers."

For reasons which we will soon justify, the number system which we propose to study is called the "domain of $\sqrt{10}$ over the rational domain" and is designated $[Ra\sqrt{10}]$. The numbers A of this system will be indicated as ordered number pairs: $A = (a_1,a_2)$, so it will be necessary in this lesson not to interpret this notation as indicating a greatest common divisor.

By definition the number system $[Ra\sqrt{10}]$ consists of *all* ordered pairs $A = (a_1,a_2)$ of rational integers a_1,a_2 subject to the following rules:

equality: $(a_1,a_2) = (b_1,b_2)$ if and only if $a_1 = b_1$, $a_2 = b_2$;
addition: $(a_1,a_2) + (b_1,b_2) = (a_1 + b_1,a_2 + b_2)$;
multiplication: $(a_1,a_2)(b_1,b_2) = (a_1b_1 + 10a_2b_2,\ a_1b_2 + a_2b_1)$.

[*] Chapter 33 is a basic chapter; a few references to Chapter 29 are required.

Q.1: The system $[Ra\sqrt{10}]$ is a domain.

Proof: A review of the conditions set forth in **29.2** and **29.3** shows that the first requirement is to show that the equality of $[Ra\sqrt{10}]$ is an equals relation; but this follows at once, since the new equality is defined componentwise in terms of the equality of rational integers, which is known to be an equals relation. It is likewise easy to check that the elements of $[Ra\sqrt{10}]$ form a commutative group under addition, for addition is defined componentwise in terms of the addition of rational integers, and the rational integers are known to form a commutative group under their addition. In particular, $(0,0)$ is the zero of $[Ra\sqrt{10}]$.

It is obvious that multiplication is *closed* for $a_1b_1+10a_2b_2$ and $a_1b_2+a_2b_1$ are rational integers whenever a_1,b_1,a_2,b_2 are rational integers; furthermore, multiplication is *well defined* for the addition, and multiplication of rational integers, used in forming the new product, are well-defined operations. From the symmetry of the components of the product and the commutative laws for rational integers, it follows that multiplication in $[Ra\sqrt{10}]$ is *commutative*. From the associative and distributive laws for rational integers it follows that multiplication in $[Ra\sqrt{10}]$ is *associative*, for by either scheme of association we find $(a_1,a_2)(b_1,b_2)(c_1,c_2)$ given by
$\{a_1b_1c_1+10(a_1b_2c_2+a_2b_1c_2+a_2b_2c_1),\ 10a_2b_2c_2+(a_2b_1c_1+a_1b_2c_1+a_1b_1c_2)\}$. There is an identity of multiplication in $[Ra\sqrt{10}]$ provided by $(1,0)$ since $(1,0)(a_1,a_2)=(a_1,a_2)$.

The *distributive* law relating the addition and multiplication of $[Ra\sqrt{10}]$ is valid, for we have

$$\begin{aligned}(a_1,a_2)\{(b_1,b_2)+(c_1,c_2)\} &= (a_1,a_2)(b_1+c_1,b_2+c_2)\\ &= (a_1b_1+a_1c_1+10a_2b_2+10a_2c_2,\ a_1b_2+a_1c_2+a_2b_1+a_2c_1)\\ &= (a_1b_1+10a_2b_2,a_1b_2+a_2b_1)+(a_1c_1+10a_2c_2,a_1c_2+a_2c_1)\\ &= (a_1,a_2)(b_1,b_2)+(a_1,a_2)(c_1,c_2).\end{aligned}$$

Finally, we must establish the *cancellation* law for all non-zero numbers in $[Ra\sqrt{10}]$. If we consider $(a,b)\neq(0,0)$ and $(a,b)(x,y)=(0,0)$, we are led to the Diophantine equations:

$$ax+10by=0,\quad ay+bx=0.$$

This system of equations implies

$$axy+10by^2=b(10y^2-x^2)=0,\quad ax^2+10bxy=a(x^2-10y^2)=0.$$

Thus if $(a,b)\neq(0,0)$, then since at least one of a and b is not 0, we employ the cancellation law for rational integers to assert that $x^2 = 10y^2$. Since 10 is not a perfect square, the equation last given is impossible in *non-zero* integers x and y (see EX. *15.1*). Hence $(x,y) = (0,0)$, and this establishes the cancellation law for non-zero numbers of $[Ra\sqrt{10}]$. This completes the proof that $[Ra\sqrt{10}]$ is a domain.

Q.2: In the domain $[Ra\sqrt{10}]$ the subsystem K of all numbers of the form $(a,0)$ forms a domain isomorphic to the rational domain $[Ra]$.

Proof: We readily check that the one-to-one correspondence T defined by $(a,0)T = a$ from K to $[Ra]$ preserves both the operations of addition and multiplication in the respective systems, for

$$\begin{aligned}((a,0) + (b,0))T &= (a + b,0)T = a + b = (a,0)T + (b,0)T,\\ ((a,0)(b,0))T &= (ab,0)T \quad\;\; = ab \quad\;\; = (a,0)T(b,0)T.\end{aligned}$$

Henceforth, with **Q.2** in mind, we agree to identify K with $[Ra]$ and to write a for $(a,0)$ whenever convenient.

Q.3: The domain $[Ra\sqrt{10}]$ contains a solution of the equation $X^2 = 10$.

Proof: Written at greater length, the equation under consideration appears as $(x,y)^2 = (10,0)$. It is then easy to check that $X = (0,1)$ is such that $X^2 = (0,1)(0,1) = (10,0)$.

The theorems **Q.2** and **Q.3** show us that the domain $[Ra\sqrt{10}]$ may be thought of as an *extension* of the domain $[Ra]$; for not only does $[Ra\sqrt{10}]$ contain a *subsystem* K behaving just *like* $[Ra]$, but *other numbers* of a different character, for in the rational domain the equation $x^2 = 10$ has no solution. We can now see a motive for defining $(0,1) = \sqrt{10}$ and then using the symbol $[Ra\sqrt{10}]$ for this domain, meaning to indicate by the *brackets* that this is a domain (see **Q.1**) whose elements have properties making them deserve the title of "integers," by the Ra that rational integers are used in the description of the new domain and are themselves to be found in the domain under the isomorphic disguise K (see **Q.2**), and by the $\sqrt{10}$ that there is present in the new domain an "integer" whose square is "10" (see **Q.3**).

In fact, we may now write

$$(a_1,a_2) = (a_1,0) + (0,a_2) = (a_1,0) + (a_2,0)(0,1)$$

to see that under the agreements suggested above, a suitable notation for (a_1,a_2) would be as follows:

$$(a_1,a_2) = a_1 + a_2\sqrt{10}.$$

If one has already established the properties of the real number system, so that a real number $\sqrt{10}$ is available, this suggests a new approach to defining $[Ra\sqrt{10}]$. However, when it is recalled that the construction of the real number system involves a rather non-arithmetical study of *infinite* sequences, it is seen that there is considerable logical merit in insisting on this simple approach using *pairs* of rational integers. Nevertheless, the notation $a_1 + a_2\sqrt{10}$ has considerable mnemonic usefulness, since with its aid the rules of addition and multiplication for the number pairs are readily reconstructed.

33.2. Divisibility properties in $[Ra\sqrt{10}]$. With **Q.1** in mind it is natural to attempt a division of the integers of $[Ra\sqrt{10}]$ into classes according to divisibility properties as was done for the rational domain in **5.1.**

If $C = AB$, we will call C a *multiple* of B, and B, a *divisor* or *factor* of C. The *zero* (0,0) is exceptional from the point of view, being a multiple of every integer, so we put it in a class by itself.

If there are integers A and B such that $AB = 1 = (1,0)$, then A and B will be called *units*. (It is shown below that the domain $[Ra\sqrt{10}]$ has a *plentiful* supply of units.)

If an integer P is not a unit and is such that $P = AB$ implies that either A or B must be a unit, then P will be called a *prime*. (Some primes are exhibited later.)

Any integer that is not zero, not a unit, and not a prime, will be called *composite*.

Thus on the basis of rather simple divisibility properties the integers of $[Ra\sqrt{10}]$ fall into four distinct categories.

In discussing questions of divisibility a valuable device is the following concept. For every integer $A = (a_1,a_2)$ we define $N(A) = a_1^2 - 10a_2^2$ to be the "norm of A." It is clear that $N(A)$ is a rational integer. An alternative definition would be to associate with every integer $A = (a_1,a_2)$, its *conjugate* integer $\bar{A} = (a_1,-a_2)$, for then $A\bar{A} = (N(A),0) = N(A)$.

The most valuable property of the norm is derived from the following theorem which shows that every factorization of integers in $[Ra\sqrt{10}]$ must be accompanied by a factorization of rational integers.

Q.4: $N(AB) = N(A)N(B)$.

Proof: By direct substitution we may verify that

$$\begin{aligned} N(AB) &= (a_1b_1 + 10a_2b_2)^2 - 10(a_1b_2 + a_2b_1)^2 \\ &= (a_1^2 - 10a_2^2)(b_1^2 - 10b_2^2) = N(A)N(B). \end{aligned}$$

An alternative proof is suggested in EX. *33.3*.

Q.5: An integer A is a unit if and only if $N(A) = \pm 1$.

Proof: By definition A is a unit if and only if there is an integer B such that $AB = 1$. From **Q.4** it follows that $N(A)N(B) = N(AB) = N(1) = 1$; since $N(A)$ is a rational integer, it follows that a necessary condition for A to be a unit is that $N(A) = \pm 1$. Conversely, if $N(A) = \pm 1$, then since $A\bar{A} = N(A)$, we can take $B = \pm\bar{A}$ to show $AB = 1$, so that A is a unit.

From **Q.5** it follows that $A = (a_1,a_2)$ is a unit if and only if a_1,a_2 is a solution of one of the equations

$$a_1^2 - 10a_2^2 = \pm 1.$$

An unusual feature now presents itself for these equations have infinitely many solutions; hence $[Ra\sqrt{10}]$ has infinitely many units.

For example, it is readily checked that $N(3,1) = -1$; hence by **Q.4** it follows that $N(3,1)^k = (-1)^k$; hence $(3,1)^k$ is a unit for $k = 1,2,\ldots$. Moreover, an induction proof will show $(3,1)^k = (p_k,q_k)$ where $p_0 = 1$, $q_0 = 0$; $p_1 = 3$, $q_1 = 1$; and

$$p_{k+1} = 6p_k + p_{k-1},\ q_{k+1} = 6q_k + q_{k-1}, \text{ for } k > 1;$$

hence it follows readily that the units obtained in this manner are distinct (cf. Chapter 32). (It is more difficult to prove that *all* units are obtained from $\pm(3,1)^k$ where k is *any* integer, interpreting $A^{-k} = \bar{A}^k$ and $A^0 = 1$.)

For integers of $[Ra\sqrt{10}]$ let us define A to be an *associate* of B if and only if there exists a unit U such that $A = UB$.

Q.6: Being an associate is an equivalence relation.

Proof: By definition the concept of being an associate is *determinative*. Since $(1,0) = 1$ is a unit and $A = 1A$, we find that A is an associate of itself, so the concept is *reflexive*. If U is a unit, there is a companion unit V such that $UV = 1$; consequently $A = UB$

implies $VA = VUB = B$, so the concept is *symmetric*. If U and U_1 are units, so is UU_1, since by **Q.4** and **Q.5** we have $N(UU_1) = N(U)N(U_1) = (\pm 1)(\pm 1) = \pm 1$; therefore $A = UB, B = U_1C$ imply $A = (UU_1)C$ and show the concept of being an associate to be *transitive*. This completes the proof.

From **Q.6** it follows that the integers of $[Ra\sqrt{10}]$ are divided into mutually exclusive classes of associated integers. Moreover this division preserves the concepts of the zero, units, primes, and composites. The zero is in a class by itself, and all the units fall into one class; if one member of a class is a prime, so are all the others; if one member is composite, so are all its associates. The last remarks follow since, if P and Q are associates, we have $P = UQ, Q = VP$, where $UV = 1$; hence from $P = AB$ we can derive $Q = A_1B_1$ where $A_1 = VA$, $B_1 = B$; or conversely, from $Q = A_1B_1$ we can derive $P = AB$ where $A = UA_1$, $B = B_1$. Then by **Q.4** and **Q.5** we have $N(A) = \pm N(A_1)$, $N(B) = \pm N(B_1)$. Hence if P is composite, so that *both* of A and B are not units and hence by **Q.5** have norms in absolute value greater than 1, the same can be said of A_1 and B_1, so Q is composite; and conversely.

If we suppose $C = AB$, it is simple to take any unit U and its companion unit V such that $UV = 1$ and to write $C = AUVB = A_1B_1$ where $A_1 = AU$ and $B_1 = BV$. We will not consider two such factorizations, which differ only because factors have been replaced by their associates, as being distinct.

In our first discussion of the fundamental theorem of arithmetic for rational integers, we limited the argument to positive integers. Had we considered all rational integers, we would have had to state the theorem in such a way as to allow for the replacement of a prime p by $-p$, in other words for the replacement of a prime by one of its associates. With this in mind, we could restate the fundamental theorem in this form: every rational integer, not zero or a unit, can be factored into a product of primes, and this factorization is unique except for the order of the prime factors and the replacement of a prime by an associate.

We must make these same considerations when we discuss factorization in $[Ra\sqrt{10}]$, and the situation is even more critical because of the great multiplicity of units and hence of associates. When we compare the primes of one factorization with those of another factorization, which is claimed to be distinct, it will be necessary to guarantee that the primes used in one factorization are not associates of the primes used in the other factorization.

With these preliminary remarks in mind, we can anticipate how the following very special theorem may become of great interest.

Q.7: In $[Ra\sqrt{10}]$ the integers $(2,0),(3,0),(2,1),(-2,1)$ are primes, and no two of these are associates.

Proof: First we will show that there are no integers of $[Ra\sqrt{10}]$ with norm $\equiv 2$ or 3, mod 5. For if $A = (a_1,a_2)$ we have $N(A) = a_1^2 - 10a_2^2$, then $a_1^2 \equiv N(A)$ mod 5. But for every integer a_1 we have $a_1^2 \equiv 0,1$, or 4 mod 5, *not* 2 or 3.

Let C represent $(2,0)$ or $(3,0)$ or $(2,1)$ or $(-2,1)$. If $C = AB$, then by **Q.4** we have $N(A)N(B) = N(C)$, where $N(C) = 4,9$, or -6. Since $N(A)$ cannot be ± 2 or ± 3, it follows that one of $N(A)$ or $N(B)$ must be ± 1, hence A or B must be a unit, hence in every case C is a prime.

By **Q.4** and **Q.5** it follows that associated integers must have equal norms; hence $(2,0)$ and $(3,0)$ with norms 4 and 9, respectively, are not associates, nor are either of these associates of $(2,1)$ and $(-2,1)$ which both have the norm -6. Thus only the case of $(2,1)$ and $(-2,1)$ offers any delay; but when we examine the equation $(2,1)(x,y) = (-2,1)$, we obtain the equations $2x + 10y = -2$, $2y + x = 1$; since the solution $x = 7/3$, $y = -2/3$ is *not* a solution in rational *integers*, it follows that $(2,1)$ and $(-2,1)$ are *not* associates.

33.3. The fundamental theorem reconsidered.

Q.8: In the domain $[Ra\sqrt{10}]$ unique factorization of composite integers into primes does not hold.

Proof: The rather special results is **Q.7** were intended for use in this proof, since it is only necessary to produce *one counterexample* to show that unique factorization does *not* hold. For this counterexample we use

$$(6,0) = (2,0)(3,0) = (2,1)(-2,1).$$

By **Q.7**, all of $(2,0)$, $(3,0)$, $(2,1)$, $(-2,1)$ are primes of $[Ra\sqrt{10}]$, and the primes of the second factorization are not associates of those of the first factorization. Hence, by the standards proposed as a test, the integer $(6,0)$ has two essentially different factorizations into primes.

The fact that the domain $[Ra\sqrt{10}]$, so like the rational domain in many respects, fails to have a fundamental theorem demonstrates how essential it is that the fundamental theorem in $[Ra]$ be proved; it also opens up whole new fields of inquiry, such as the discovery of domains in which there is a fundamental theorem, or the study of a remedy in those cases, like $[Ra\sqrt{10}]$, where there fails to be the usual sort of fundamental theorem.

If we recall that the proof of the fundamental theorem in Chapter 6 stemmed from a study of the greatest common divisor, whereas the usual definition of a greatest common divisor fails in $[Ra\sqrt{10}]$ (see EX. *33.4*), it will not be too surprising to learn that a remedy for the anomalous cases, where a fundamental theorem is lacking, can be obtained by a suitable generalization of the notion of greatest common divisor by a scheme known as the theory of ideals. By this rather remarkable scheme unique factorization is restored. The theory involved is classical, but too long for inclusion here. The interested student may refer to treatises on algebraic numbers or, for example, to the MacDuffee text, cited in **1.3**.

EXERCISES

EX. *33.1*. Consider the set S of all matrices of the form

$$A = \begin{pmatrix} a_1 & a_2 \\ 10a_2 & a_1 \end{pmatrix}$$

where a_1 and a_2 are rational integers. Consider the mapping T from $[Ra\sqrt{10}]$ to S defined by $(a_1,a_2)T = A$. Show that T is an isomorphism between $[Ra\sqrt{10}]$ and S, the addition and multiplication in S being the usual matric operations.

EX. *33.2*. With reference to EX. *33.1* show that $N(a_1,a_2)$ in $[Ra\sqrt{10}]$ is the determinant of the corresponding matrix in S.

EX. *33.3*. Using EX. *33.1* and EX. *33.2* show $N(AB) = N(A)N(B)$.

EX. *33.4*. In $[Ra\sqrt{10}]$ consider $A = (6,0)$ and $B = (2,0)(2,1)$. Show that $D = (2,0)$ and $D_1 = (2,1)$ are common divisors of A and B, but that neither is a divisor of the other. Hence the usual definition of a greatest common divisor (see **5.2**) is not useful in $[Ra\sqrt{10}]$.

The student should review the concepts in EX. *30.9* through EX. *30.13* before beginning the next exercises.

EX. *33.5*. Explain why the Gaussian domain $[G]$ might be designated $[Ra\sqrt{-1}]$.

EX. *33.6.* For a number $A = (a_1,a_2)$ in $[G]$ define the norm by $N(A) = a_1^2 + a_2^2$ and the conjugate by $A = (a_1,-a_2)$. Prove that $N(A) = A\bar{A}$. Prove that $N(AB) = N(A)N(B)$ and compare with (*26.1*).

EX. *33.7.* Define A in $[G]$ to be a unit of $[G]$ if and only if there is a B in $[G]$ such that $AB = 1 = (1,0)$. Use EX. *33.6* to prove that $[G]$ has only *four* units.

EX. *33.8.* Define A and B of $[G]$ to be associates if and only if there is a unit U of $[G]$ such that $A = UB$. Prove that every $A \neq (0,0)$ in $[G]$ has a unique associate (x,y) such that $0 < x$, $0 \leqq y$.

EX. *33.9.* Define P of $[G]$ to be a prime of $[G]$ if P is not a unit and if $P = AB$ implies that one of A and B must be a unit. Use EX. *33.6* and **L.1** of **26.1** to show that $p = (p,0)$ is a prime of $[G]$ if and only if p is a rational prime of the form $4K + 3$.

EX. *33.10.* Show that for any $B \neq (0,0)$ of $[G]$ and any A of $[G]$ we can find Q and R in $[G]$ such that

$$A = QB + R, \quad 0 \leqq N(R) < N(B)$$

(*Hint:* Let $A\bar{B} = (x_1,x_2)$, $b = N(B)$, $x_1 = q_1b + s_1$, $x_2 = q_2b + s_2$, where $|s_1| \leqq b/2$, $|s_2| \leqq b/2$; take $Q = (q_1,q_2)$ and $R = A - QB$.)

EX. *33.11.* Use the "division algorithm" of EX. *33.10* to construct a "Euclid algorithm" for $[G]$ and show that any two numbers A,B of $[G]$, not both zero, have a greatest common divisor D, unique up to an associate, and that there exist numbers X,Y in $[G]$ such that $D = AX + BY$.

EX. *33.12.* Use EX. *33.11* to show that if a prime P divides a product AB in $[G]$, then P must divide at least one of A and B.

EX. *33.13.* Use EX. *33.6* and the "fundamental lemma" of EX. *33.12* to show that $[G]$ has a "fundamental theorem": Every number of $[G]$, not zero and not a unit, is either a prime, or can be written as a product of primes, this representation being unique except for the order of the primes and the possibility of replacing a prime by one of its associates.

EX. *33.14.* Use the fundamental theorem for $[G]$ to show that a rational prime p of the form $p = 4K + 1$ can be written in *only* one way as the sum of two squares.

INDEX