TOPICS IN NUMBER THEORY

VOLUME I

To A. J. Kempner

TOPICS IN NUMBER THEORY

VOLUME I

---

by

WILLIAM JUDSON LEVEQUE

Department of Mathematics
University of Michigan

ADDISON-WESLEY PUBLISHING COMPANY
READING, MASSACHUSETTS · MENLO PARK, CALIFORNIA
LONDON · AMSTERDAM · DON MILLS, ONTARIO · SYDNEY

ADDISON-WESLEY PUBLISHING COMPANY, INC

Printed in the United States of America

Library of Congress Catalog Card No 56–10138

Third printing—December 1965

ISBN 0-201-04225-8
DEFGH JKLM-CO 7987654

PREFACE

The theory of numbers, one of the oldest branches of mathematics, has engaged the attention of many gifted mathematicians during the past 2300 years. The Greeks, Indians, and Chinese made significant contributions prior to 1000 A.D., and in more modern times the subject has developed steadily since Fermat, one of the fathers of Western mathematics. It is therefore rather surprising that there has never been a strong tradition in number theory in America, although a few men of the stature of L. E. Dickson have emerged to keep the flame alive. But in most American universities the theory of numbers is regarded as a slightly peripheral subject, which has an unusual flavor and unquestioned historical importance, but probably merits no more than a one-term course on the senior or first-year graduate level. It seems to me that this is an inappropriate attitude to maintain toward a subject which is flourishing in European hands, and which has contributed so much to the mathematics of the past and which promises exciting developments in the future. Changing its status is complicated, however, by the paucity of advanced works suitable for use as textbooks in American institutions. There are several excellent elementary texts available, and an ever-increasing number of monographs, mostly European, but to the best of my knowledge no general book designed for a second course in the theory of numbers has appeared since Dickson ceased writing. In Volume II of the present work I have attempted partially to fill this gap.

When I began to write Volume II, the number of introductory texts was very small, and no one of them contained all the information I found occasion to refer to. Since I had already written lecture notes for a first course, there seemed to be some advantage in expanding them into a more complete exposition of the standard elementary topics. Volume I is the result; it is designed to serve either as a self-contained textbook for a one-term course in number theory, or as a preliminary to the second volume. The two volumes together are intended to provide an introduction to some of the important techniques and results of classical and modern number theory; I hope they will prove useful as a first step in the training of students who are or might become seriously interested in the subject.

In view of the diversity of problems and methods grouped together under the name of number theory, it is clearly impossible to write even an introductory treatment which in any sense covers the field completely My choice of topics was made partly on the basis of my own taste and knowledge, of course, but also more objectively on the grounds of the technical importance of the methods developed or of the results obtained It was this consideration which led me, for example to give a standard function-theoretic proof of the Prime Number Theorem in the second volume the analytic method has proved to be extremely powerful and is applicable to a large variety of problems, so that it must be considered an essential tool in the subject, while the elementary Erdos Selberg method has found only limited applications, and so for the time being must be regarded as an isolated device, of great interest to the specialist but of secondary importance to the beginner

In a similar vein, I have on several occasions given proofs which are neither the shortest nor most elegant known but which seem to me to be the most natural, or to lead to the deepest understanding of the phenomena under consideration For example, the proof given in Chapter 8, Volume I, of Hurwitz' theorem on the approximation of an irrational number by rationals is perhaps not as elegant as some others known but of those which make no use of continued fractions, it is the only one I am familiar with which does not require prior knowledge of the special role played by the number $\sqrt{5}$ To my mind these other proofs are inferior pedagogically, in that they give no hint as to how the student might attack a similar problem

Most of the material in the first volume is regularly included in various elementary courses, although it would probably be impossible to cover the entire volume in one semester This allows the instructor to choose topics to suit his taste and what is even more important for my general purpose, it presents the student with an opportunity for further reading in the subject

I consider the first volume suitable for presentation to advanced undergraduate and beginning graduate students insofar as the difficulty of the subject matter is concerned No technical knowledge is assumed except in Section 3–5 and in Chapter 6 where calculus is used On the other hand elementary number theory is by no means easy, and that vaguely defined quality called mathematical maturity is of great value in developing a sound feeling for the subject. I

doubt, though, that it should be considered a prerequisite, even if it could be measured; studying number theory is perhaps as good a way as any of acquiring it.

Rather few of the problems occurring at the ends of sections are of the routine computational type; I assume that the student can devise such problems as well as I. It has been my experience that many of those included offer some difficulty to most students. For this reason I have appended hints in profusion, and have indicated by asterisks a few problems that remain more difficult than the average.

The development of continued fractions in the final chapter may be sufficiently novel to warrant a word of explanation. I have chosen to regard as the basic problem that of finding the "good" rational approximations to a real number, and have derived the regular continued fraction as the solution. This procedure seems to me to be pedagogically better than the classical treatment, in which one simply defines a continued fraction and verifies that the convergents have the requisite property. Moreover, this same approach looks promising for the corresponding problem of approximating complex numbers by the elements of a fixed quadratic field, while earlier attempts to define a useful complex continued fraction algorithm have been conspicuously unsuccessful. The idea of associating an interval with each Farey point is derived from work by K. Mahler, who, with J. W. S. Cassels and W. Ledermann, investigated the much more complicated Gaussian case [*Philosophical Transactions of the Royal Society, A* (London) **243**, 585–628 (1951)].

I am grateful to Professors T. Apostol, A. Brauer, B. W. Jones, and K. Mahler for their many constructive criticisms, to Mrs. Edith Fisher for her help in typing the manuscript, and to Mr. Earl Lazerson for his invaluable aid in proofreading.

W. J. L.

Ann Arbor, Michigan
November, 1955

CONTENTS

CHAPTER 1

INTRODUCTION

1–1 What is number theory? In number theory we are concerned with properties of certain of the integers

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

or sometimes with those properties of the real or complex numbers which depend rather directly on the integers. As in most branches of abstract thought, it is easier to characterize the theory of numbers extensively, by giving a large number of examples of problems which are usually considered parts of number theory, than to define it intensively, by saying that exactly those problems having certain characteristics will be included in the subject. Before considering such a list of types of problems, however, it might be worth while to make an exclusion.

In the opinion of the author, the theory of numbers does not include the axiomatic construction or characterization either of systems of numbers (integers, rational numbers, real numbers, or complex numbers) or of the fundamental operations and relations in these sets. Toward the end of this chapter, a few properties of the integers are mentioned which the student may not have considered explicitly before; aside from these, no properties will be assumed beyond what any high-school pupil knows. It is, of course, quite possible that the student will not have read a logical treatment of elementary arithmetic; if he wishes to do so, he might examine E. Landau's elegant *Foundations of Analysis* (New York: Chelsea Publishing Company, 1951), but he should not expect to find a treatment of this kind here. The contents of such a book are, in a sense, assumed to be known to the reader, but as far as understanding number theory is concerned, this assumption is of little consequence.

The problems treated in number theory can be divided into groups according to a more or less rough classification. First, there are multiplicative problems, concerned with divisibility properties of the integers. It will be proved later that any positive integer n greater than 1 can be represented uniquely, except for the order of the factors,

as a product of primes, a prime being any integer greater than 1 having no exact divisors except itself and 1 This might almost be termed the Fundamental Theorem of number theory, so manifold and varied are its applications *From the decomposition of n into primes, it is* easy to determine the number of divisors of n This number is called $\tau(n)$ by some writers and $d(n)$ by others, we shall use the former designation The behavior of $\tau(n)$ is very erratic, the first few values are as follows

n	$\tau(n)$	n	$\tau(n)$
1	1	13	2
2	2	14	4
3	2	15	4
4	3	16	5
5	2	17	2
6	4	18	6
7	2	19	2
8	4	20	6
9	3	21	4
10	4	22	4
11	2	23	2
12	6	24	8

If $n = 2^m$, the divisors of n are 1 2 2^2, 2^m, so that $\tau(2^m) = m + 1$ On the other hand, if n is a prime then $\tau(n) = 2$ Since, as we shall see, there are infinitely many primes, it appears that the τ-function has arbitrarily large values, and yet has the value 2 for infinitely many n Many questions might occur to anyone who thinks about the subject for a few moments and studies the above table For example,

(a) Is it true that $\tau(n)$ is odd if and only if n is a square?

(b) Is it always true that if m and n have no common factor, then $\tau(m)\tau(n) = \tau(mn)$?

(c) Do the arguments of the form 2^m give the relatively largest values of the τ-function? That is is the inequality

$$\tau(n) \leq \frac{\log n}{\log 2} + 1$$

correct for all n? If not, is there any better upper bound than the trivial one, $\tau(n) \leq n$?

(d) How large is $\tau(n)$ on the average? That is, what can be said about the quantity

$$\frac{1}{N}\sum_{n=1}^{N}\tau(n),$$

as N increases indefinitely?

(e) For large N, approximately how many solutions $n \leq N$ are there of the equation $\tau(n) = 2$? In other words, about how many primes are there among the integers $1, 2, \ldots, N$?

Of the above questions, which are fairly typical problems in multiplicative number theory, the first is very easy to answer in the affirmative. The next three are more difficult; they will be considered in Chapter 6. The last is very difficult indeed. It was conjectured by C. F. Gauss and A. Legendre, two of the greatest of number theorists, that the number $\pi(N)$ of primes not exceeding N is approximately $N/\log N$, in the sense that the relative error

$$\frac{|\pi(N) - (N/\log N)|}{N/\log N} = \left|\frac{\pi(N)}{N/\log N} - 1\right|$$

is very small when N is sufficiently large. Many years later (1852–54), P. L. Chebyshev showed that if this relative error has any limiting value, it must be zero, but it was not until 1896 that J. Hadamard and C. de la Vallée Poussin finally proved what is now called the Prime Number Theorem, that

$$\lim_{N\to\infty}\frac{\pi(N)}{N/\log N} = 1.$$

In another direction, we have the problems of additive number theory: questions concerning the representability, and the number of representations, of a positive integer as a *sum* of integers of a specified kind. For instance, upon examination it appears that some integers, like $5 = 1^2 + 2^2$ and $13 = 2^2 + 3^2$, are representable as a sum of two squares, while others, like 3 or 12, are not. Which integers are so representable, and how many such representations are there?

A third category might include what are known as Diophantine equations, named after the Greek mathematician Diophantos, who first studied them. These are equations in one or more variables whose solutions are required to be integers, or at any rate rational

numbers For example, it is a familiar fact that $3^2 + 4^2 = 5^2$, which gives us a solution of the Diophantine equation $x^2 + y^2 = z^2$ *Giving a particular solution is hardly of interest, what is desired is an* explicit formula for *all* solutions A very famous Diophantine equation is that known as Fermat's equation $x^n + y^n = z^n$ P Fermat asserted that this equation has no solution (in nonzero integers, of course) if $n \geq 3$, the assertion has never been proved or disproved for general n There is at present practically no general theory of Diophantine equations although there are many special methods, most of which were devised for the solution of particular equations

Finally, there are problems in Diophantine approximations For example, given a real number ξ and a positive integer N, find that *rational number* p/q for which $q \leq N$ and $|\xi - (p/q)|$ is minimal The proofs that e and π are transcendental also fall in this category This branch of number theory probably borrows the most from, and contributes the most to other branches of mathematics

The theorems of number theory can also be subdivided along entirely different lines—for example according to the methods used in their proofs Thus the dichotomies of elementary and nonelementary, analytic and synthetic A proof is said to be elementary (although not necessarily simple[1]) if it makes no use of the theory of functions of a complex variable and synthetic if it does not involve the usual concepts of analysis—limits, continuity etc Sometimes, but not always the nature of the theorem shows that the proof will be in one or another of these categories For example the above-mentioned theorem about $\pi(x)$ is clearly a theorem of analytic number theory but it was not until 1948 that an elementary proof was found *On the other hand, the following theorem* first proved by D Hilbert, involves in its statement none of the concepts of analysis yet the only proofs known prior to 1942 were analytic Given any positive integer k there is another integer s depending only on k such that every positive integer is representable as the sum of at most s kth powers, i e such that the equation

$$n = x_1^k + \quad + x_s^k$$

is solvable in non-negative integers x_1 , x_s for every n

It may seem strange at first that the theory of functions of a com-

plex variable is useful in treating arithmetic problems, since there is, *prima facie*, nothing common to the two disciplines. Even after we understand how function theory can be used, we must still reconcile ourselves to the rather disquieting thought that it apparently *must* be used in some problems—that there is, at present, simply no other way to deal with them. What is perhaps not a familiar fact to the general reader is that function theory is only one of many branches of mathematics which are at best only slightly related to number theory, but which enter in an essential way into number-theoretic considerations. This is true, for example, of abstract algebra, probability, Euclidean and projective geometry, topology, the theory of Fourier series, differential equations, and elliptic and other automorphic functions. In particular, it would appear that the rather common subsumption of number theory under algebra involves a certain distortion of the facts.

1-2 Proofs. It is a well-known phenomenon in mathematics that an excessively simple theorem frequently is difficult to prove (although the proof, in retrospect, may be short and elegant) just because of its simplicity. This is probably due in part to the lack of any hint in the statement of the theorem as to the machinery to be used in proving it, and in part to the lack of available machinery. Since many theorems of elementary number theory are of this kind, and since there is considerable diversity in the types of arguments used in their proofs, it might not be amiss to discuss the subject briefly.

First a psychological remark. If we are presented with a rather large number of theorems bearing on the same subject but proved by quite diverse means, the natural tendency is to regard the techniques used in the various proofs as special tricks, each applicable only to the theorem with which it is associated. A technique ceases to be a trick and becomes a method only when it has been encountered enough times to seem natural; correspondingly, a subject may be regarded as a "bag of tricks" if the relative number of techniques to results is too high. Unfortunately, elementary number theory has sometimes been regarded as such a subject. On working longer in the field, however, we find that many of the tricks become methods, and that there is more uniformity than is at first apparent. By making a

conscious effort to abstract and retain the germs of the proofs that follow, the reader will begin to see patterns emerging sooner than he otherwise might

Consider, for example, the assertion that $\tau(n)$ is even unless n is a square, i e , the square of another integer A proof of this is as follows If d is a divisor of n, then so is the integer n/d If n is not a square, then $d \neq n/d$, since otherwise $n = d^2$ Hence if n is not a square, its divisors can be paired off into couples d, n/d, so that each divisor of n occurs just once as an element of some one of these couples The number of divisors is therefore twice the number of couples and being twice an integer, is even

The principle here is that when we want to count the integers having a certain property (here "count may also be replaced by "add"), it may be helpful first to group them in judicious fashion There are several problems in the present book whose solutions depend on this idea

In addition to the special methods appropriate to number theory, we shall have many occasions to use two quite general types of proof with which the student may not have had much experience proofs by contradiction and proofs by induction

An assertion P is said to have been proved by contradiction if it has been shown that by assuming P false we can deduce an assertion Q which is known to be incorrect or which contradicts the assumption that P is false As an example consider the theorem (known as early as the time of Euclid) that there are infinitely many prime numbers To prove this by contradiction we assume the opposite namely that there are only finitely many primes Let these be $p_1, p_2, \ldots, p_k$, let N be the integer $p_1 p_2 \cdots p_k + 1$, and let Q be the assertion that N is divisible by some prime different from any of $p_1, p_2, \ldots, p_k$ Now N is divisible by some prime p (if N is itself prime, then $p = N$), and N is not divisible by any of the $p_1, p_2, \ldots, p_k$, since each of these primes leaves a remainder of 1 when divided into N Hence Q is true Since Q is not compatible with the falsity of the theorem the theorem is true.

As for proofs by induction let $P(n)$ be a statement involving an integral variable n, we wish to prove that $P(n)$ is valid for every integer n not less than a particular one say n_0 The induction principle says that if $P(n_0)$ is valid and if for every $n \geq n_0$ one can

deduce $P(n+1)$ by assuming that $P(n)$ is valid, then $P(n)$ is valid for every integer $n \geq n_0$. The statement $P(n_0)$ must, of course, be proved independently; usually, though not always, by direct verification. The difficulty, if any, normally lies in showing that $P(n)$ implies $P(n+1)$.

As an example, let us undertake to prove that the formula

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

is correct, whatever the positive integer n may be. Here $n_0 = 1$. There are three steps:

(a) Show that the formula is correct when $n = 1$. This is trivial here.

(b) Show that if n is an integer for which (1) holds, the same is true of $n + 1$. But if (1) holds, then adding $n + 1$ to both sides gives

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

and this is simply (1) with n replaced by $n + 1$, so that $P(n)$ implies $P(n+1)$.

(c) Use the principle of induction to deduce that (1) holds for every positive integer n.

As a second example, consider the Fibonacci sequence

$$1, 2, 3, 5, 8, 13, 21, \ldots,$$

in which every element after the second is the sum of the two numbers immediately preceding it. If we denote by u_n the nth element of this sequence, then the sequence is recursively defined by the conditions

$$\begin{aligned} u_1 &= 1, \\ u_2 &= 2, \\ u_n &= u_{n-1} + u_{n-2}, \qquad n \geq 3. \end{aligned} \tag{2}$$

We may verify, for as large n as we like, that

$$u_n < (\tfrac{7}{4})^n.$$

To prove this for all positive integral n, we take for $P(n)$ the following statement: the inequalities

$$u_n < (\tfrac{7}{4})^n \qquad \text{and} \qquad u_{n+1} < (\tfrac{7}{4})^{n+1} \tag{3}$$

hold This is clearly equivalent to the earlier statement We repeat the three steps

(a) For $n = 1$, $P(n)$ reduces to the assertion that $1 < \frac{7}{4}$ and $2 < (\frac{7}{4})^2$

(b) The induction hypothesis now is that $u_n < (\frac{7}{4})^n$ and $u_{n+1} < (\frac{7}{4})^{n+1}$, where n is a positive integer Since $n + 2 \geq 3$, we have that

$$u_{n+2} = u_{n+1} + u_n,$$

by (2) Hence

$$u_{n+2} < (\tfrac{7}{4})^{n+1} + (\tfrac{7}{4})^n = (\tfrac{7}{4})^n(1 + \tfrac{7}{4}) < (\tfrac{7}{4})^n \quad (\tfrac{7}{4})^2 = (\tfrac{7}{4})^{n+2},$$

and this inequality, together with the induction hypothesis, shows that

$$u_{n+1} < (\tfrac{7}{4})^{n+1} \qquad \text{and} \qquad u_{n+2} < (\tfrac{7}{4})^{n+2},$$

so that $P(n)$ implies $P(n + 1)$

(c) By the induction principle, it follows that the inequality (3) holds for all positive integral n

To avoid the artificial procedure of the last proof, it is frequently convenient to use the following formulation of the principle of induction, which can be shown to be equivalent to the first if $P(n_0)$ is valid, and if for every $n \geq n_0$ the propositions $P(n_0)$ $P(n_0 + 1)$, , $P(n)$ together imply $P(n + 1)$, then $P(n)$ is valid for every $n \geq n_0$

Using this formulation, we could have taken $P(n)$ to be the assertion $u_n < (\frac{7}{4})^n$ in the second example

Besides the principle of induction, we shall have occasion to use three other properties of integers which the reader may not have encountered explicitly before

(a) Every nonempty set of positive integers (or of non-negative integers) has a smallest element

(b) If a and b are positive integers there exists a positive integer n such that $na > b$

(c) Let n be a positive integer If a set of $n + 1$ elements is subdivided into n or fewer subsets, in such a way that each element belongs to precisely one subset, then some subset contains more than one element

These assertions, which are consequences of the underlying axioms, will be assumed without proof

PROBLEMS

1. Show that $\tau(n)$ is odd if n is a square.

2. Prove that

$$\text{(a) } \sum_{m=1}^{n} m = \frac{n(n+1)}{2}, \qquad \text{(b) } \sum_{m=1}^{n} m^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\text{(c) } \sum_{m=1}^{n} m^3 = \frac{n^2(n+1)^2}{4},$$

first by induction on n, and second by writing

$$\sum_{m=1}^{n+1} m^k = \sum_{m=1}^{n+1} ((m-1)+1)^k$$

and applying the binomial theorem to the summands on the right. Since the terms $\sum_1^n m^k$ drop out, this method can be used with $k = 2$ to prove (a), then with $k = 3$ to prove (b), etc.

3. Prove by induction that no two consecutive elements of the Fibonacci sequence $u_1, u_2, \ldots$ have a common divisor greater than 1.

4. Carry out the second proof of the inequality

$$u_n < (\tfrac{7}{4})^n$$

as indicated in the text.

5. Prove by induction that every integer greater than 1 can be represented as a product of primes.

6. Anticipating Theorem 1–1, suppose that every integer can be written in the form $6k + r$, where k is an integer and r is one of the numbers 0, 1, 2, 3, 4, 5.

(a) Show that if $p = 6k + r$ is a prime different from 2 and 3, then $r = 1$ or 5.

(b) Show that the product of numbers of the form $6k + 1$ is of the same form.

(c) Show that there exists a prime of the form $6k - 1 = 6(k-1) + 5$.

(d) Show that there are infinitely many primes of the form $6k - 1$.

1–3 Radix representation. Although we have assumed a knowledge of the structure of the system of integers, we have said nothing about the method by which we will assign names to the integers. There are, of course, various ways of doing this, of which the Roman and decimal systems are probably the best known. While the decimal system has obvious advantages over Roman numerals, and the advantage of familiarity over any other method, it is not always the best system for theoretical purposes. A rather more general scheme

is sometimes convenient, and it is the object of the following two theorems to show that this kind of representation is possible, i e , that each integer can be given a unique name Here, and until Chapter 6, lower-case Latin letters will denote integers.

THEOREM 1–1 *If a is positive and b is arbitrary, there is exactly one pair of integers q, r such that the conditions*

$$b = qa + r, \qquad 0 \leq r < a, \tag{4}$$

hold

Proof First, we show that (4) has at least one solution

Consider the set D of integers of the form $b - ua$, where u runs over all integers, positive and nonpositive For the particular choice

$$u = \begin{cases} -1, & \text{if } b \geq 0, \\ b, & \text{if } b < 0, \end{cases}$$

the number $b - ua$ is non-negative, so that D contains non-negative elements The subset consisting of the non negative elements of D has a smallest element Take r to be this number, and q the value of u which corresponds to it Then

$$r = b - qa \geq 0, \qquad r - a = b - (q + 1)a < 0,$$

so that (4) is satisfied

To show the uniqueness, assume that also

$$b = q'a + r', \qquad 0 \leq r' < a$$

Then if $q' < q$,

$$b - q'a = r' \geq b - (q - 1)a = r + a \geq a,$$

while if $q' > q$

$$b - q'a = r' \leq b - (q + 1)a = r - a < 0$$

Hence $$q' = q, r' = r$$

THEOREM 1–2 *Let g be greater than 1 Then each a greater than 0 can be represented uniquely in the form*

$$a = c_0 + c_1 g + \qquad + c_n g^n,$$

where c_n is positive and $0 \leq c_m < g$ for $0 \leq m \leq n$

Proof: We prove the representability by induction on a. For $a = 1$ we have $n = 0$, $c_0 = 1$.

Take a greater than 1 and assume that the theorem is true for 1, 2, $\ldots$, $a - 1$. Since g is larger than 1, the numbers $g^0, g^1, g^2, \ldots$ form an increasing sequence, and any positive integer lies between some pair of successive powers of g. More precisely, there is a unique $n \geq 0$ such that $g^n \leq a < g^{n+1}$. By Theorem 1-1,

$$a = c_n g^n + r, \qquad 0 \leq r < g^n.$$

Here $c_n > 0$, since $c_n g^n = a - r > g^n - g^n = 0$; moreover, $c_n < g$ because $c_n g^n \leq a < g^{n+1}$. If $r = 0$,

$$a = 0 + 0 \cdot g + \cdots + 0 \cdot g^{n-1} + c_n g^n;$$

while if r is positive, the induction hypothesis shows that r has a representation of the form

$$r = b_0 + b_1 g + \cdots + b_t g^t,$$

where b_t is positive and $0 \leq b_m < g$ for $0 \leq m \leq t$. Moreover, $t < n$. Thus

$$a = b_0 + b_1 g + \cdots + b_t g^t + 0 \cdot g^{t+1} + \cdots + 0 \cdot g^{n-1} + c_n g^n.$$

Now use the induction principle.

To prove uniqueness, assume that

$$a = c_0 + c_1 g + \cdots + c_n g^n = d_0 + d_1 g + \cdots + d_r g^r,$$

with $n \geq 0$, $c_n > 0$, and $0 \leq c_m < g$ for $0 \leq m \leq n$, and also $r \geq 0$, $d_r > 0$, and $0 \leq d_m < g$ for $0 \leq m \leq r$. Then, by subtraction, we have

$$0 = e_0 + e_1 g + \cdots + e_s g^s,$$

where $e_m = c_m - d_m$ and where s is the largest value of m for which $c_m \neq d_m$, so that $e_s \neq 0$. If $s = 0$, we have the contradiction $e_0 = e_s = 0$. If $s > 0$ we have, since

$$|e_m| = |c_m - d_m| \leq g - 1$$

and

$$e_s g^s = -(e_0 + \cdots + e_{s-1} g^{s-1}),$$

$$\begin{aligned} g^s \leq |e_s g^s| &= |e_0 + \cdots + e_{s-1} g^{s-1}| \leq |e_0| + \cdots + |e_{s-1}| g^{s-1} \\ &\leq (g - 1)(1 + g + \cdots + g^{s-1}) = g^s - 1, \end{aligned}$$

which is also a contradiction We conclude that $n = r$ and $c_m = d_m$ for $0 \leq m \leq n$, and the representation is unique

By means of Theorem 1–2 we can construct a system of names or symbols for the positive integers in the following way We choose arbitrary symbols to stand for the *digits* (i e , the non-negative integers less than g) and replace the number

$$c_0 + c_1 g + \cdots + c_n g^n$$

by the simpler symbol $c_n c_{n-1} \cdots c_1 c_0$ For example, choosing g to be ten, and giving the smaller integers their customary symbols, we have the ordinary decimal system, in which, for example, 2743 is an abbreviation for the value of the polynomial $2x^3 + 7x^2 + 4x + 3$ when x is ten But there is no reason why we must use ten as the *base*, or *radix*, if we used seven instead, we would write the integer whose decimal representation is 2743 as 10666, since

$$2743 = 6 + 6 \cdot 7 + 6 \cdot 7^2 + 0 \cdot 7^3 + 1 \cdot 7^4$$

To indicate the base that is being used, we might write a subscript (in the decimal system), so that

$$(2743)_{10} = (10666)_7$$

Of course, if the radix is larger than $(10)_{10}$, it will be necessary to invent symbols to replace $(10)_{10}$, $(11)_{10}$, , $g - 1$ For example, taking $g = (12)_{10}$ and putting $(10)_{10} = \alpha$, $(11)_{10} = \beta$, we have

$$(14)_{12} + (7)_{12} = (1\beta)_{12}$$

and

$$(31)_{12} \cdot (\alpha)_{12} = (37)_{10} \cdot (10)_{10} = (370)_{10} = (26\alpha)_{12}$$

PROBLEMS

1 (a) Show that any integral weight less than 2^{n+1} can be weighed using only the standard weights 1, 2, 2^2, , 2^n, by putting the unknown weight on one pan of the balance and a suitable combination of standard weights on the other pan

(b) Prove that no other set of $n + 1$ weights will do this [*Hint* Name the weights so that $w_0 \leq w_1 \leq \cdots \leq w_n$ Let k be the smallest index such that $w_k \neq 2^k$ and obtain a contradiction, using the fact that the number of nonempty subsets of a set of $n + 1$ elements is $2^{n+1} - 1$]

2 Construct the addition and multiplication tables for the duodecimal digits i e , the digits in base twelve Using these tables evaluate

$$(21\alpha 9)_{12} \cdot (\beta 370)_{12}$$

3. Let $u_1, u_2, \ldots$ be the Fibonacci sequence defined in the preceding section.

(a) Prove by induction (or otherwise) that for $n > 0$,

$$u_{n-1} + u_{n-3} + u_{n-5} + \cdots < u_n,$$

the sum on the left continuing so long as the subscripts are positive.

(b) Show that every positive integer can be represented in a unique way in the form $u_{n_1} + u_{n_2} + \cdots + u_{n_k}$, where $k \geq 1$ and $n_{j-1} \geq n_j + 2$ for $j = 2, 3, \ldots, k$.

CHAPTER 2

THE EUCLIDEAN ALGORITHM AND ITS CONSEQUENCES

2–1 Divisibility. Let a be different from zero, and let b be arbitrary. Then, if there is a c such that $b = ac$, we say that *a divides b*, and write $a|b$ (negation $a\nmid b$). As usual, the letters involved represent integers.

The following statements are immediate consequences of this definition.

(a) For every $a \neq 0$, $a|0$ and $a|a$. For every b, $\pm 1|b$.

(b) If $a|b$ and $b|c$, then $a|c$.

(c) If $a|b$ and $a|c$, then $a|(bx + cy)$ for each x, y. (If $a|b$ and $a|c$, then a is said to be a *common divisor* of b and c.)

2–2 The Euclidean algorithm and greatest common divisor

THEOREM 2-1. *Given any two integers a, b not both zero, there is a unique integer d such that*

(a) $d > 0$,

(b) *$d|a$ and $d|b$,*

(c) *if $d_1|a$ and $d_1|b$, then $d_1|d$.*

Since $x|y$ implies that $|x| \leq |y|$, we call the d of Theorem 2–1 the *greatest common divisor* (abbreviated GCD) of a and b, and write $d = (a, b)$.

Proof. First let a and b be positive and assume that $a \geq b$. Then, by Theorem 1-1, there are unique integers q_1, r_1 such that

$$a = bq_1 + r_1, \qquad 0 \leq r_1 < a$$

Repeated application of this theorem shows the existence of unique pairs q_2, r_2, q_3, r_3, , such that

$$b = r_1q_2 + r_2, \qquad 0 \leq r_2 < r_1,$$

$$r_1 = r_2q_3 + r_3, \qquad 0 \leq r_3 < r_2$$

and this may be continued until we reach a remainder, say r_{k+1}, which is zero; the existence of such a k is assured because $r_1, r_2, \ldots$ is a decreasing sequence of non-negative integers. Thus the process terminates:

$$\begin{aligned} r_{k-3} &= r_{k-2}q_{k-1} + r_{k-1}, & 0 \leq r_{k-1} < r_{k-2}, \\ r_{k-2} &= r_{k-1}q_k + r_k, & 0 \leq r_k < r_{k-1}, \\ r_{k-1} &= r_k q_{k+1}. \end{aligned}$$

From the last equation we see that $r_k | r_{k-1}$; from the preceding equation, using statement (b) of Section 2–1, we see that $r_k | r_{k-2}$, etc. Finally, from the second and first equations, respectively, we have that $r_k | b$ and $r_k | a$. Thus r_k is a common divisor of a and b. Now let d_1 be any common divisor of a and b. From the first equation, $d_1 | r_1$; from the second, $d_1 | r_2$; etc.; from the equation before the last, $d_1 | r_k$. Thus we can take the d of the theorem to be r_k.

If $a < b$, interchange the names of a and b. If either a or b is negative, find the d corresponding to $|a|, |b|$. If a is zero, $(a, b) = |b|$.

If both d_1 and d_2 have the properties of the theorem, then d_1, being a common divisor of a and b, divides d_2. Similarly, $d_2 | d_1$. This clearly implies that $d_1 = d_2$, and the GCD is unique.

The chain of operations indicated by the above equations is known as the Euclidean algorithm; as will be seen, it is the cornerstone of multiplicative number theory. (In general, an algorithm is a systematic procedure which is applied repeatedly, each step depending on the results of the earlier steps. Other examples are the long division algorithm and the square root algorithm.) The Euclidean algorithm is actually quite practicable in numerical cases; for example, if we wish to find the GCD of 4147 and 10672, we have

$$\begin{aligned} 10672 &= 4147 \cdot 2 + 2378, \\ 4147 &= 2378 \cdot 1 + 1769, \\ 2378 &= 1769 \cdot 1 + 609, \\ 1769 &= 609 \cdot 2 + 551, \\ 609 &= 551 \cdot 1 + 58, \\ 551 &= 58 \cdot 9 + 29, \\ 58 &= 29 \cdot 2. \end{aligned}$$

Hence $(4147, 10672) = 29$.

It is frequently important to know whether two integers a and b have a common factor larger than 1 If they have not, so that $(a, b) = 1$, we say that they are *relatively prime*, or *prime to each other*

The following properties of the GCD are easily derived either from the definition or from the Euclidean algorithm

(a) The GCD of more than two numbers, defined as that positive common divisor which is divisible by every common divisor, exists and can be found in the following way Let there be n numbers $a_1, a_2 \quad , a_n$, and define

$$D_1 = (a_1, a_2), \qquad D_2 = (D_1, a_3), \qquad , \qquad D_{n-1} = (D_{n-2} \; a_n)$$

Then $(a_1, a_2, \quad , a_n) = D_{n-1}$

(b) $(ma, mb) = m(a, b)$, if $m \neq 0$

(c) If $m|a$ and $m|b$, then $(a/m, b/m) = (a, b)/m$

(d) If $(a \; b) = d$, there exist integers x, y such that $ax + by = d$ (An important consequence of this is that if a and b are relatively prime, there exist x, y such that $ax + by = 1$ Conversely, if there is such a representation of 1, then clearly $(a, b) = 1$)

(e) If a given integer is relatively prime to each of several others, it is relatively prime to their product For if $(a \; b) = 1$ and $(a, c) = 1$, there are x, y, t, and u such that $ax + by = 1$ and $at + cu = 1$, whence $ax + by(at + cu) = a(x + byt) + bc(yu) = 1$, and therefore $(a, bc) = 1$

The Euclidean algorithm can be used to find the x and y of property (d) Thus, using the numerical example above, we have

$$\begin{aligned}
29 &= 551 - 58 \;\; 9 && (58 \;\; = 609 - 551 \;\; 1)\\
&= 551 - 9(609 - 551 \;\; 1)\\
&= 10 \;\; 551 - 9 \;\; 609 && (551 \;\; = 1769 - 2 \;\; 609)\\
&= 10(1769 - 2 \;\; 609) - 9 \;\; 609\\
&= 10 \;\; 1769 - 29 \;\; 609 && (609 \;\; - 2378 - 1 \;\; 1769)\\
&= 10 \;\; 1769 - 29(2378 - 1 \;\; 1769)\\
&= 39 \;\; 1769 - 29 \;\; 2378 && (1769 = 4147 - 2378)\\
&= 39(4147 - 2378) - 29 \;\; 2378\\
&= 39 \;\; 4147 - 68 \;\; 2378 && (2378 = 10672 - 2 \;\; 4147)\\
&= 175 \;\; 4147 - 68 \;\; 10672,
\end{aligned}$$

so that $x = 175 \;\; y = -68$

PROBLEMS

1. Evaluate (4655, 12075), and express the result as a linear combination of 4655 and 12075, that is, in the form $4655x + 12075y$.

2. Show that if $(a, b) = 1$, then $(a - b, a + b) = 1$ or 2.

3. Show that if $ax + by = m$, then $(a, b) | m$.

4. Show that no cancellation is possible in the fraction

$$\frac{a_1 + a_2}{b_1 + b_2}$$

if $a_1b_2 - a_2b_1 = \pm 1$.

5. Show that if $b|a$ and $c|a$, and $(b, c) = 1$, then $bc|a$.

6. Show that if $(b, c) = 1$, then $(a, bc) = (a, b)(a, c)$. [*Hint:* Prove that each member of the last equation divides the other. Use property (d) above, and the preceding problem.]

*7. Show that if $a + b \neq 0$, $(a, b) = 1$, and p is an odd prime, then

$$\left(a + b, \frac{a^p + b^p}{a + b}\right) = 1 \text{ or } p.$$

[*Hint:* If this GCD is d, then $a + b = kd$ and $(a^p + b^p)/(a + b) = ld$. Replace b and $a + b$ in the second equation by their values from the first, apply the binomial theorem, and show that $d|p$.]

*8. In the notation introduced in the proof of Theorem 2–1, show that each nonzero remainder r_m $(m \geq 2)$ is less than $r_{m-2}/2$. (Consider separately the cases in which r_{m-1} is less than, equal to, and greater than $r_{m-2}/2$.) Deduce that the number of divisions in the Euclidean algorithm is less than

$$\frac{2 \log b}{\log 2} = 2.88 \ldots \log b,$$

where b is the larger of the two numbers whose GCD is being found. (Here and elsewhere, "log" means the natural logarithm.)

2–3 The Unique Factorization Theorem

THEOREM 2–2. *Every integer $a > 1$ can be represented as a product of one or more primes.*

Proof: The theorem is true for $a = 2$. Assume it true for 2, 3, 4, $\ldots$, $a - 1$. If a is prime, we are through. Otherwise a has a divisor different from 1 and a, and we have $a = bc$, with $1 < b < a$, $1 < c < a$.

*Here and in all problems throughout the book, an asterisk is used to indicate a particularly difficult problem.

The induction hypothesis then implies that

$$b = \prod_{i=1}^{s} p_i', \qquad c = \prod_{i=1}^{t} p_i''$$

with p_i', p_i'' primes, and hence $a = p_1'p_2' \ldots p_s'p_1'' \ldots p_t''$

Any positive integer which is not prime and which is different from unity is said to be *composite* Hereafter p will be used to denote a prime number, unless otherwise specified

THEOREM 2–3 *If* $a|bc$ *and* $(a, b) = 1$, *then* $a|c$

Proof If $(a, b) = 1$, there are integers x and y such that $ax + by = 1$, or $acx + bcy = c$ But a divides both ac and bc, and therefore divides c

THEOREM 2–4 *If*

$$p \Big| \prod_{m=1}^{n} p_m,$$

then for at least one m, $p = p_m$

Proof Suppose that $p|p_1p_2 \quad p_n$ but that p is different from any of the $p_1, p_2, \quad , p_{n-1}$ Then p is relatively prime to each of the $p_1, \quad , p_{n-1}$, and so is relatively prime to their product By Theorem 2–3, $p|p_n$, whence $p = p_n$

THEOREM 2–5 (*Unique Factorization Theorem*) *The representation of* $a > 1$ *as a product of primes is unique up to the order of the factors*

Proof We must show exactly the following From

$$a = \prod_{m=1}^{n_1} p_m = \prod_{m=1}^{n_2} p_m', \qquad (p_1 \le p_2 \le \quad \le p_{n_1}, p_1' \le p_2' \le \quad \le p_{n_2}'),$$

it follows that $n_1 = n_2$ and $p_m = p_m'$ for $1 \le m \le n_1$

For $a = 2$ the assertion is true, since $n_1 = n_2 = 1$, $p_1 = p_1' = 2$

Take $a > 2$ and assume the assertion correct for $2, 3, \quad , a - 1$

(a) If a is prime, $n_1 = n_2 = 1$, $p_1 = p_1' = a$

(b) Otherwise $n_1 > 1$, $n_2 > 1$ From

$$p_1' \Big| \prod_{m=1}^{n_1} p_m, \qquad p_1 \Big| \prod_{m=1}^{n_2} p_m'$$

it follows by Theorem 2–4 that for at least one r and at least one s,

$$p_1' = p_r, \qquad p_1 = p_s'.$$

Since

$$p_1 \leq p_r = p_1' \leq p_s' = p_1,$$

we have $p_1 = p_1'$. Moreover, since $1 < p_1 < a$ and $p_1 | a$, we have

$$1 < \frac{a}{p_1} = \prod_{m=2}^{n_1} p_m = \prod_{m=2}^{n_2} p_m' < a,$$

and hence by the induction hypothesis,

$$n_1 - 1 = n_2 - 1 \qquad \text{and} \qquad p_m = p_m' \qquad \text{for } 2 \leq m \leq n_1.$$

Theorem 2–5, which appears natural enough when one is accustomed to working only with the ordinary integers, assumes greater significance when we encounter more general types of "integers" for which it is not true.

PROBLEMS

1. Show that if the reduced fraction a/b is a root of the equation

$$c_0 x^n + c_1 x^{n-1} + \cdots + c_n = 0,$$

where x is a real variable and $c_0, c_1, \ldots, c_n$ are integers with $c_0 \neq 0$, then $a | c_n$ and $b | c_0$. In particular, show that if k is an integer then $\sqrt[n]{k}$ is rational if and only if it is an integer.

2. The Unique Factorization Theorem shows that each integer $a > 1$ can be written uniquely as a product of powers of distinct primes. If the primes that do not divide a are included in this product with exponents 0, we can write

$$a = \prod_{i=1}^{\infty} p_i^{\alpha_i},$$

where p_i is the ith prime, $\alpha_i \geq 0$ for each i and $\alpha_i = 0$ for sufficiently large i, and the α_i's are uniquely determined by a. Show that if also

$$b = \prod_{i=1}^{\infty} p_i^{\beta_i},$$

then

$$(a, b) = \prod_{i=1}^{\infty} p_i^{\min(\alpha_i, \beta_i)},$$

where $\min(\alpha, \beta)$ is the smaller of α and β. Use this to give a different solution of Problem 6, Section 2–2.

3 Show that the Diophantine equation

$$x^2 - y^2 = N$$

is solvable in non negative integers x and y if and only if N is odd or divisible by 4 Show further that the solution is unique if and only if N or $N/4$, respectively, is unity or a prime [*Hint* Factor the left side]

4 Show that the following identity is formally correct

$$\sum_{k=0}^{\infty} \frac{1}{2^{2k}} \quad \sum_{k=0}^{\infty} \frac{1}{3^{2k}} \quad \sum_{k=0}^{\infty} \frac{1}{5^{2k}} \quad = \sum_{n=1}^{\infty} \frac{1}{n^2}$$

The denominators occurring on the left are the even powers of the primes

2-4 The linear Diophantine equation. For simplicity, we consider only the equation in two variables

$$ax + by = c \tag{1}$$

It is easy to devise a scheme for finding an infinite number of solutions of this equation in case any exist, it can best be explained by means of a numerical example, say $5x + 22y = 18$ Since x is to be an integer, $\frac{1}{5}(18 - 22y)$ must also be integral Writing

$$x = \frac{18 - 22y}{5} = 3 - 4y + \frac{3 - 2y}{5},$$

we see that $\frac{1}{5}(3 - 2y)$ must also be an integer, say z This gives

$$z = \frac{3 - 2y}{5}, \qquad 2y + 5z = 3$$

We now repeat the argument, solving as before for the unknown which has the smaller coefficient

$$y = \frac{3 - 5z}{2} = 1 - 2z + \frac{1 - z}{2},$$

$$\frac{1 - z}{2} = t, \qquad z = 1 - 2t$$

Clearly, z will be an integer for any integral t and we have

$$y = \frac{3 - 5(1 - 2t)}{2} = -1 + 5t,$$

$$x = \frac{18 - 22(-1 + 5t)}{5} = 8 - 22t$$

Moreover, it is easily seen that any solution x, y of the original equation must be of this form, so that we have a *general solution* of the equation.

The same idea could be applied in the general case, but it is somewhat simpler to adopt a different approach. First of all, it should be noticed that (1) has no solution unless $d|c$, where $d = (a, b)$, and that if this requirement is satisfied, we can divide through in (1) by d to get a new equation

$$a'x + b'y = c', \tag{2}$$

where now $(a', b') = 1$. We now use property (d) of Section 2–2 to assert the existence of numbers x_0', y_0' such that

$$a'x_0' + b'y_0' = 1,$$

so that $c'x_0'$, $c'y_0'$ is a solution of (2). Put $c'x_0' = x_0$, $c'y_0' = y_0$. If t is any integer, we have

$$a'(x_0 + b't) + b'(y_0 - a't) = a'x_0 + b'y_0 = c',$$

so that $x_0 + b't$, $y_0 - a't$ is a solution of (2) for each t. Finally, if x_1, y_1 is *any* solution of (2), we have

$$a'x_0 + b'y_0 = c', \qquad a'x_1 + b'y_1 = c',$$

and, by subtraction,

$$a'(x_0 - x_1) + b'(y_0 - y_1) = 0.$$

Thus $a'|(y_0 - y_1)$, $y_0 - y_1 = a't_1$, and $b'|(x_0 - x_1)$, $x_0 - x_1 = b't_2$. This gives $x_1 = x_0 - b't_2$, $y_1 = y_0 - a't_1$, and, requiring that these numbers satisfy (2), we have $t_2 = -t_1$. Hence every solution of (2) is of the form $x_0 + b't$, $y_0 - a't$, and every such pair constitutes a solution. Since every solution of (1) is a solution of (2) and conversely, we have the following theorem.

THEOREM 2–6. *A necessary and sufficient condition that the equation*

$$ax + by = c$$

have a solution x, y in integers is that $d|c$, where $d = (a, b)$. If there is one solution, there are infinitely many; they are exactly the numbers of the form

$$x = x_0 + \frac{b}{d}t, \qquad y = y_0 - \frac{a}{d}t,$$

where t is an arbitrary integer.

There are various ways of getting a particular solution Sometimes one can be found by inspection, if not, the method explained at the beginning of the section may be used or, what is almost the same thing, the Euclidean algorithm may be applied to find a solution of the equation which results from dividing the original equation through by (a, b) The latter process of successively eliminating the remainders in the Euclidean algorithm can be systematized, but this we shall not do at present (See Section 9–2)

PROBLEMS

1 Find a general solution of the linear Diophantine equation

$$2072x + 1813y = 2849$$

2 *Find all solutions of* $19x + 20y = 1909$ *with* $x > 0$, $y > 0$

3 Let m and n be positive integers, with $m \leq n$, and let $x_0, x_1, \quad , x_k$ be all the distinct numbers among the two sequences

$$\frac{0}{m}, \frac{1}{m}, \quad , \frac{m}{m} \quad \text{and} \quad \frac{0}{n}, \frac{1}{n}, \quad , \frac{n}{n},$$

arranged so that $x_0 < x_1 < \quad < x_k$ Describe k as a function of m and n What is the shortest distance between successive x's?

*4 Let a and b be positive relatively prime integers Then for certain non negative integers n (which we shall refer to briefly as the *representable* integers), the equation $ax + by = n$ has a solution with $x \geq 0$, $y \geq 0$, while for other n it may not have For example if $n = 0, 3$ 5, or 6, or if $n \geq 8$, then $3x + 5y = n$ has such a solution Show that this example is typical, in the following sense

(a) There is always a number $N(a, b)$ such that for all $n \geq N(a, b)$, n is representable (It may be helpful to combine the theory of the present section with the elementary analytic geometry of the line $ax + by = c$, interpreting x and y in the latter case as real variables Note that so far it is only the existence of $N(a, b)$ that is in question and not its size)

(b) The minimal value of $N(a\ b)$ is always $(a - 1)(b - 1)$

(c) Exactly half the integers up to $(a - 1)(b - 1)$ are representable

2–5 The least common multiple

THEOREM 2–7 *The number* $\langle a, b\rangle = \dfrac{|ab|}{(a, b)}$ *has the following properties* (a) $\langle a, b\rangle \geq 0$, (b) $a|\langle a, b\rangle$, $b|\langle a, b\rangle$, (c) *If* $a|m$ *and* $b|m$, *then* $\langle a, b\rangle|m$

Proof: (a) Obvious.

(b) Since $(a, b)|b$, we can write

$$\langle a, b\rangle = |a| \cdot \frac{|b|}{(a, b)},$$

and hence $a|\langle a, b\rangle$. Similarly,

$$\langle a, b\rangle = |b| \cdot \frac{|a|}{(a, b)},$$

and so $b|\langle a, b\rangle$.

(c) Let $$m = ra = sb,$$

and put $d = (a, b)$, $a = a_1 d$, $b = b_1 d$. Then

$$m = ra_1 d = sb_1 d;$$

thus $a_1|sb_1$, and since $(a_1, b_1) = 1$, it must be that $a_1|s$. Thus $s = a_1 t$, and

$$m = ta_1 b_1 d = t\frac{ab}{d}.$$

Because of the properties listed in Theorem 2–7, the number $\langle a, b\rangle$ is called the *least common multiple* (LCM) of a and b. The definition is easily extended to the case of more than two numbers, just as for the GCD. It is useful to remember that

$$ab = \pm(a, b)\langle a, b\rangle.$$

PROBLEMS

1. In the notation of Problem 2, Section 2–3, show that

$$\langle a, b\rangle = \prod_{i=1}^{\infty} p_i^{\max(\alpha_i, \beta_i)},$$

where max (α, β) is the larger of α and β.

*2. Show that

$$\min(\alpha, \max(\beta, \gamma)) = \max(\min(\alpha, \beta), \min(\alpha, \gamma)).$$

(By symmetry, one may suppose $\beta \geq \gamma$.) Deduce that

$$(a, \langle b, c\rangle) = \langle (a, b), (a, c)\rangle.$$

CHAPTER 3

CONGRUENCES

3-1 Introduction The problem of solving the Diophantine equation $ax + by = c$ is just that of finding an x such that ax and c leave the same remainder when divided by b, since then $b|(c - ax)$ and we can take $y = (c - ax)/b$ As we shall see, there are many other instances also in which a comparison must be made of the remainders after dividing each of two numbers a and b by a third, say m Of course, if the remainders are the same, then $m|(a - b)$, and conversely, and this might seem to be an adequate notation But, as Gauss noticed, the following, for most purposes, is more suggestive if $m|(a - b)$, then we write $a \equiv b \pmod{m}$, and say that a is *congruent to b modulo m*

The use of the symbol "$\equiv$" is suggested by the similarity of the relation we are discussing to ordinary equality Each of these two relations is an example of an *equivalence relation*, i e , a relation $\mathcal{R}$ between elements of a set, such that if a and b are arbitrary elements, either a stands in the relation $\mathcal{R}$ to b (more briefly, $a \mathcal{R} b$) or not, and having the following properties

(a) $a \mathcal{R} a$

(b) If $a \mathcal{R} b$, then $b \mathcal{R} a$.

(c) If $a \mathcal{R} b$ and $b \mathcal{R} c$, then $a \mathcal{R} c$

These are called the *reflexive*, *symmetric* and *transitive* properties, respectively That ordinary equality between numbers is an equivalence relation is obvious (or it may be taken as an axiom) either $a = b$ or $a \neq b$, $a = a$, if $a = b$, then $b = a$, if $a = b$ and $b = c$, then $a = c$

THEOREM 3-1 *Congruence modulo a fixed number m is an equivalence relation*

Proof (a) $m|(a - a)$, so that $a \equiv a \pmod{m}$

(b) If $m|(a - b)$, then $m|(b - a)$, if $a \equiv b \pmod{m}$, then $b \equiv a \pmod{m}$

(c) If $m|(a - b)$ and $m|(b - c)$, then $a - b = km$, $b - c = lm$, say, so that $a - c = (k + l)m$, if $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$

Since we shall have occasion later to use several other equivalence relations, we pause to show a simple but important property enjoyed by all such relations. If $\mathcal{R}$ is an equivalence relation with respect to a set S, then corresponding to each element a of S there is a subset S_a of S which consists of exactly those elements of S which are equivalent to a, so that b is in S_a if and only if $a \mathcal{R} b$. Now if $a \mathcal{R} b$, then the sets S_a and S_b are identical: if c is in S_b, then $c \mathcal{R} b$, and since $a \mathcal{R} b$, also $c \mathcal{R} a$, so that c is in S_a. If, on the other hand, a is not equivalent to b, then S_a and S_b are disjoint; that is, they have no element in common. For if c is in S_a and in S_b, then $c \mathcal{R} a$ and $c \mathcal{R} b$, which entails $a \mathcal{R} b$. These disjoint sets, which jointly exhaust S, are called *equivalence classes;* an element of an equivalence class is sometimes called a *representative* of the class, and a *complete system of representatives* is any subset of S which contains exactly one element from each equivalence class.

Section 3-3 provides examples of all these notions, with somewhat different terminology.

PROBLEM

Decide whether each of the following is an equivalence relation. If it is, describe the equivalence classes.

(a) Congruence of triangles.
(b) Similarity of triangles.
(c) The relations "$\neq$", "$>$", and "$\geq$", relating real numbers.
(d) Parallelism of lines.
(e) Having the same mother.
(f) Having a parent in common.

3-2 Elementary properties of congruences. One reason for the superiority of the congruence notation is that congruences can be combined in much the same way as can equations.

THEOREM 3-2. *If* $a \equiv b \pmod{m}$ *and* $c \equiv d \pmod{m}$, *then* $a + c \equiv b + d \pmod{m}$, $ac \equiv bd \pmod{m}$, *and* $ka \equiv kb \pmod{m}$ *for every integer* k.

Proof: These statements follow immediately from the definition. For if

$$m|(a - b) \quad \text{and} \quad m|(c - d),$$

then

$$m|(a - b + c - d) \quad \text{and} \quad m|((a + c) - (b + d)).$$

If $m|(a-b)$, then $m|k(a-b)$ Finally, if $m|(a-b)$ and $m|(c-d)$, then $m|(a-b)(c-d)$ But

$$(a-b)(c-d) = ac - bd + b(d-c) + d(b-a),$$

so that also $m|(ac-bd)$

THEOREM 3–3 *If $f(x)$ is a polynomial with integral coefficients, and $a \equiv b \pmod m$, then $f(a) \equiv f(b) \pmod m$*

Proof Let $f(x) = c_0 + c_1x +$ $+ c_nx^n$

If $a \equiv b \pmod m$, then for every non-negative integer j,

$$a^j \equiv b^j \pmod m,$$

and

$$c_ja^j \equiv c_jb^j \pmod m,$$

by Theorem 3–2 Adding these last congruences for $j = 0, 1,$ $, n,$ we have the theorem

The situation is a little more complicated when we consider dividing both sides of a congruence by an integer We cannot deduce from $ka \equiv kb \pmod m$ that $a \equiv b \pmod m$, for it may be that part of the divisibility of $ka - kb = k(a-b)$ by m is accounted for by the presence of the factor k What is clearly necessary is that the part of m which does not divide k should divide $a - b$

THEOREM 3–4 *If $ka \equiv kb \pmod m$ and $(k, m) = d$, then*

$$a \equiv b \left(\mod \frac{m}{d}\right)$$

Proof Theorem 2–3

PROBLEMS

1 Let $f(x) = a_0x^n + a_1x^{n-1} +$ $+ a_n$

where $a_0,$ $, a_n$ are integers Show that if d consecutive values of f (i e , values for consecutive integers) are all divisible by the integer d then $d|f(x)$ for all integral x Show by an example that this sometimes happens with $d > 1$ even when $(a_0$ $a_n) = 1$

2 In Theorem 3–3 take $a = 10$ $b = 1$ $m = 9$ to deduce the rule that an integer is divisible by 9 if and only if this is true of the sum of its digits What is the corresponding rule for divisibility by 11? Use the fact that $7 \cdot 11 \cdot 13 = 1001$ to obtain a test for divisibility by any of the integers 7, 11, or 13

3–3 Residue classes and Euler's φ-function. When dealing with congruences modulo a fixed integer m, the set of all integers breaks down into m classes, such that any two elements of the same class are congruent and two elements from two different classes are incongruent. For many purposes it is completely immaterial which element of one of these *residue classes* is used; for example, Theorem 3–3 shows this to be the case when one considers the values modulo m of a polynomial with integral coefficients. In these cases it suffices to consider an arbitrary set of representatives of the various residue classes; that is, a set consisting of one element of each residue class. Such a set $a_1, a_2, \ldots, a_m$, called a *complete residue system modulo* m, is characterized by the following properties.

(a) If $i \neq j$, then $a_i \not\equiv a_j \pmod{m}$.

(b) If a is any integer, there is an index i with $1 \leq i \leq m$ for which $a \equiv a_i \pmod{m}$.

Examples of complete residue systems $(\text{mod } m)$ are the set of integers $0, 1, 2, \ldots, m-1$, and the set $1, 2, \ldots, m$. The elements of a complete residue system need not be consecutive integers, however; for $m = 5$ we could take $1, 22, 13, -6, 2500$ as such a set.

THEOREM 3–5. *If $a_1, a_2, \ldots, a_m$ is a complete residue system* $(\text{mod } m)$ *and $(k, m) = 1$, then also $ka_1, ka_2, \ldots, ka_m$ is a complete residue system* $(\text{mod } m)$.

Proof: We show directly that properties (a) and (b) above hold for this new set.

(a) If $ka_i \equiv ka_j \pmod{m}$, then by Theorem 3–4, $a_i \equiv a_j \pmod{m}$, whence $i = j$.

(b) Theorem 2–6 shows that if $(k, m) = 1$, the congruence $kx \equiv a \pmod{m}$ has a solution for any fixed a. Let a solution be x_0. Since $a_1, \ldots, a_m$ is a complete residue system, there is an index i such that $x_0 \equiv a_i \pmod{m}$. Hence $kx_0 \equiv ka_i \equiv a \pmod{m}$.

The reason that we use the adjective "complete" when speaking of a residue system is that there is another kind which is frequently useful, called a *reduced residue system*. This is a set of integers $a_1, \ldots, a_h$, incongruent $(\text{mod } m)$, such that if a is any integer prime to m, there is an index i, $1 \leq i \leq h$, for which $a \equiv a_i \pmod{m}$. In other words, a reduced residue system is a set of representatives, one from each of the residue classes containing integers prime to m. (Clearly,

$(a, m) = (b, m)$ if $a \equiv b \pmod m$, since then $m|(a - b)$, so that $(a, m)|(a - b)$, and hence $(a, m)|b$, this implies that $(a, m)|(b, m)$, and also, by symmetry, that $(b, m)|(a, m)$) The number h is the number of positive integers not exceeding m and prime to m This function of m is customarily designated by $\varphi(m)$, and is called *Euler's φ-function* or the *totient* of m

THEOREM 3–6 *If $a_1, \ldots, a_{\varphi(m)}$ is a reduced residue system* (mod m) *and* $(k, m) = 1$, *then also $ka_1, \ldots, ka_{\varphi(m)}$ is a reduced residue system* (mod m)

The proof is exactly parallel to that of Theorem 3–5

Euler's φ-function has many interesting properties and, as we shall see, it occurs repeatedly in number-theoretic investigations

THEOREM 3–7 *If* $(m, n) = 1$, *then* $\varphi(mn) = \varphi(m)\varphi(n)$

(A function with this property is called a *multiplicative* function For another example, see Problem 6, Section 2–2)

Proof Take integers m, n with $(m, n) = 1$, and consider the numbers of the form $mx + ny$ If we can so restrict the values which x and y assume that these numbers form a reduced residue system (mod mn), there must be $\varphi(mn)$ of them But also their number is then the product of the number of values which x assumes and the number of values which y assumes Clearly, in order for $mx + ny$ to be prime to m, it is necessary that $(m, y) = 1$, and likewise we must have $(n, x) = 1$ Conversely, if these last two conditions are satisfied, then $(mx + ny, mn) = 1$ Hence let x range over a reduced residue system (mod n), say $x_1, \ldots, x_{\varphi(n)}$, and let y run over a reduced residue system (mod m), say $y_1, \ldots, y_{\varphi(m)}$ If for some indices i, j, k, l we have

$$mx_i + ny_j \equiv mx_k + ny_l \pmod{mn},$$

then

$$m(x_i - x_k) + n(y_j - y_l) \equiv 0 \pmod{mn}$$

Since divisibility by mn implies divisibility by m, we have

$$m(x_i - x_k) + n(y_j - y_l) \equiv 0 \pmod m,$$

$$n(y_j - y_l) \equiv 0 \pmod m,$$

$$y_j \equiv y_l \pmod m,$$

whence $j = l$. Similarly, $i = k$. Thus the numbers $mx + ny$ so formed are incongruent (mod mn). Now let a be any integer prime to mn; in particular, $(a, m) = 1$ and $(a, n) = 1$. Then Theorem 2–6 shows that there are integers X, Y (not necessarily in the chosen reduced residue systems) such that $mX + nY = a$, whence also $mX + nY \equiv a \pmod{mn}$. But there is an x_i such that $X \equiv x_i \pmod{n}$, and there is a y_j such that $Y \equiv y_j \pmod{m}$. This means that there are integers k, l such that $X = x_i + kn$, $Y = y_j + lm$. Hence

$$mX + nY = m(x_i + kn) + n(y_j + lm) \equiv mx_i + ny_j \equiv a \pmod{mn}.$$

Hence as x and y run over fixed reduced residue systems (mod n) and (mod m) respectively, $mx + ny$ runs over a reduced residue system (mod mn), and the proof is complete.

THEOREM 3–8. $$\varphi(m) = m \prod_{p \mid m} \left(1 - \frac{1}{p}\right),$$

where the notation indicates a product over all the distinct primes which divide m.

Proof: By Theorem 3–7, if $m = \prod_{i=1}^{r} p_i^{\alpha_i}$,

then

$$\varphi(m) = \prod_{i=1}^{r} \varphi(p_i^{\alpha_i}).$$

But we can easily evaluate $\varphi(p^\alpha)$ directly; all the positive integers not exceeding p^α are prime to p^α except the multiples of p, and there are just $p^{\alpha-1}$ of these. Hence

$$\varphi(p_i^{\alpha_i}) = p_i^{\alpha_i} - p_i^{\alpha_i - 1} = p_i^{\alpha_i}\left(1 - \frac{1}{p_i}\right),$$

and so

$$\varphi(m) = \prod_{i=1}^{r} p_i^{\alpha_i}\left(1 - \frac{1}{p_i}\right) = \prod_{i=1}^{r} p_i^{\alpha_i} \cdot \prod_{i=1}^{r}\left(1 - \frac{1}{p_i}\right)$$
$$= m \prod_{p \mid m}\left(1 - \frac{1}{p}\right).$$

For example, the integers 1, 5, 7, 11 are all those which do not exceed 12 and are prime to 12, and

$$\varphi(12) = 12(1 - \tfrac{1}{2})(1 - \tfrac{1}{3}) = 4.$$

THEOREM 3–9 $\sum_{d|n} \varphi(d) = n$

Proof Let $d_1, \quad , d_k$ be the positive divisors of n We separate the integers between 1 and n inclusive into classes $C(d_1), \quad , C(d_k)$, putting an integer into the class $C(d_i)$ if its GCD with n is d_i The number of elements in $C(d_i)$ is then

$$\sum_{\substack{a \le n \\ (a,n)=d}} 1,$$

and since every integer up to n is in exactly one of the classes,

$$\sum_{d_i|n} \sum_{\substack{a \le n \\ (a,n)=d_i}} 1 = n$$

The number of integers a such that $a \le n$ and $(a, n) = d_i$ is exactly equal to the number of integers b such that $b \le n/d_i$ and $(b, n/d_i) = 1$, in fact, multiplying the b's by d_i, we get the a's But from the definition of the Euler function, the number of b's is clearly $\varphi(n/d_i)$ Thus

$$\sum_{d|n} \varphi\left(\frac{n}{d_i}\right) = n,$$

which is equivalent to the theorem, since, as d_i runs over the divisors of n, n/d_i also runs over these divisors, but in reverse order

To illustrate the theorem and its proof, take $n = 12$ Then

$$\varphi(1) + \varphi(2) + \varphi(3) + \varphi(4) + \varphi(6) + \varphi(12)$$
$$= 1 + 1 + 2 + 2 + 2 + 4 = 12,$$

$C(1) = \{1, 5\ 7\ 11\}, \quad C(2) = \{2, 10\} \quad C(3) = \{3, 9\},$

$C(4) = \{4, 8\}, \quad C(6) = \{6\} \quad C(12) = \{12\}$

PROBLEMS

*1 *Prove that if* $(a\ b) = d$ *then*

$$\varphi(ab) = \frac{d\varphi(a)\varphi(b)}{\varphi(d)}$$

2 Show that if $n > 1$, then the sum of the positive integers less than n and prime to it is

$$\frac{n\varphi(n)}{2}$$

[*Hint* If m satisfies the conditions, so does $n - m$]

3. Show that if $d|n$, then $\varphi(d)|\varphi(n)$.

4. Let n be positive. Show that any solution of the equation

$$\varphi(x) = 4n + 2$$

is of one of the forms p^α or $2p^\alpha$, where p is a prime of the form $4s - 1$. [*Hint:* Use the factorization of $\varphi(x)$ as given in Theorem 3-8.]

*5. Let $f(x)$ be a polynomial with integral coefficients, and let $\psi(n)$ denote the number of values

$$f(0), f(1), \ldots, f(n-1)$$

which are prime to n.

(a) Show that ψ is multiplicative:

$$\psi(mn) = \psi(m) \cdot \psi(n) \qquad \text{if } (m, n) = 1.$$

(b) Show that

$$\psi(p^\alpha) = p^{\alpha-1}(p - b_p),$$

where b_p is the number of integers $f(0), f(1), \ldots, f(p-1)$ which are divisible by the prime p.

6. How many fractions r/s are there satisfying the conditions

$$(r, s) = 1, \qquad 0 \leq r < s \leq n?$$

3-4 Linear congruences. Because of the analogy between congruences and equations, it is natural to ask about the solution of congruences involving one or more (integral) unknowns. In the case of an algebraic congruence $f(x) \equiv 0 \pmod{m}$, where $f(x)$ is a polynomial in x with integral coefficients, we see by Theorem 3-3 that if $x = a$ is a solution, so is every element of the residue class containing a. For this reason it is customary, for such congruences, to list only the solutions between 0 and $m - 1$, inclusive, with the understanding that any x congruent to one of those listed is also a solution. Similarly, when mention is made of the number of roots of a certain congruence, it is actually the number of residue classes that is meant.

The simplest case to treat is the linear congruence in one unknown; that is, the congruence

$$ax \equiv b \pmod{m}.$$

As we have already noticed, this is equivalent to the linear Diophantine equation

$$ax - my = b,$$

and by Theorem 2-6 this equation is solvable if and only if $(a, m)|b$.

If it is solvable, and if x_0, y_0 is a solution, then a general solution is

$$x \equiv x_0 \left(\mathrm{mod}\, \frac{m}{d}\right), \qquad y \equiv y_0 \left(\mathrm{mod}\, \frac{a}{d}\right),$$

where $d = (a, m)$ Among the numbers x satisfying the first of these congruences, the numbers

$$x_0, x_0 + \frac{m}{d}, x_0 + \frac{2m}{d}, \quad \ldots, x_0 + \frac{(d-1)m}{d}$$

are incongruent (mod m), while every other such x is congruent (mod m) to one of these Hence we have the following theorem

THEOREM 3–10 *A necessary and sufficient condition that the congruence*

$$ax \equiv b \pmod{m}$$

be solvable is that $(a, m) | b$ *If this is the case, there are exactly* (a, m) *solutions* (mod m)

While Theorem 3–10 gives assurance of the existence of a solution under appropriate circumstances and predicts the number of such solutions, it says nothing about finding them For this purpose the simplest procedure, if no solution can be found by inspection, is to convert the congruence to an equation and solve by the method given at the beginning of Section 2–4

Consider, for example the congruence

$$34x \equiv 60 \pmod{98}$$

Since $(34, 98) = 2$ and $2|60$, there are just two solutions, to be found from

$$17x \equiv 30 \pmod{49}$$

This is equivalent to $17x - 49y = 30$ and we get

$$x = \frac{49y + 30}{17} = 3y + 2 - \frac{2y + 4}{17}, \qquad t = \frac{2y + 4}{17},$$

$$y = \frac{17t - 4}{2} = 8t - 2 + \frac{t}{2}, \qquad z = \frac{t}{2},$$

$$t = 2z$$

Take $z = 0$; then $t = 0$, $y = -2$, $x = -4$. Hence

$$x \equiv -4 \pmod{49},$$

and the two solutions of the original congruence are

$$x \equiv -4, 45 \pmod{98}.$$

The solution of a linear congruence in more than one unknown can be effected by the successive solution of a (usually large) number of congruences in a single unknown. Consider the congruence

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n \equiv c \pmod{m}.$$

The obviously necessary condition for solvability, that $(a_1, \ldots, a_n, m)$ should divide c, is also sufficient, just as in the former case. For, assuming it satisfied, we can divide through by $(a_1, \ldots, a_n, m)$ to get

$$a_1'x_1 + \cdots + a_n'x_n \equiv c' \pmod{m'},$$

where now $(a_1', \ldots, a_n', m') = 1$. If $(a_1', \ldots, a_{n-1}', m') = d'$, we must have

$$a_n'x_n \equiv c' \pmod{d'};$$

since $(a_n', d') = 1$, this has just one solution (mod d'). Thus there are m'/d' numbers x_n with $0 \leq x_n < m'$ satisfying this congruence. Substituting these into the preceding congruence, we get m'/d' congruences in $n - 1$ unknowns, and the process can be repeated.

As an example, consider the congruence

$$2x + 7y \equiv 5 \pmod{12}.$$

Here $(2, 7, 12) = 1$. Since $(2, 12) = 2$, we must have

$$7y \equiv 5 \pmod{2},$$

which clearly gives $y \equiv 1 \pmod{2}$, or $y \equiv 1, 3, 5, 7, 9, 11 \pmod{12}$. These give

$$2x \equiv 10, 8, 6, 4, 2, 0 \pmod{12}$$

respectively, or

$$x \equiv 5, 4, 3, 2, 1, 0 \pmod{6}.$$

Thus the solutions (mod 12) are

$$x, y = 5, 1;\ 11, 1;\ 4, 3;\ 10, 3;\ 3, 5;\ 9, 5;$$
$$2, 7;\ 8, 7;\ 1, 9;\ 7, 9;\ 0, 11;\ 6, 11.$$

The general situation is given in the following theorem, which is easily proved by induction on the number of unknowns.

THEOREM 3–11 *The congruence*

$$a_1x_1 + \quad + a_nx_n \equiv c \pmod{m}$$

has just dm^{n-1} *or no solutions* $\pmod{m}$ *according as* $d|c$ *or* $d\nmid c$, *where* $d = (a_1, \quad , a_n, m)$

Turning now to the simultaneous solution of a system of linear congruences we consider the system

$$\alpha_1x \equiv \beta_1 \pmod{m_1}, \quad , \quad \alpha_nx \equiv \beta_n \pmod{m_n},$$
$$\alpha_i \text{ and } \beta_i \text{ integers}$$

Clearly, no x satisfies all these congruences unless each is solvable separately Assuming that this is so, we can restrict our attention to systems of the form

$$x \equiv c_1 \pmod{m_1}, \quad , \quad x \equiv c_n \pmod{m_n}$$

It is clear that this system will have no solution unless every pair has From the first of the congruences

$$x \equiv c_i \pmod{m_i}, \qquad x \equiv c_j \pmod{m_j},$$

we get $x = c_i + m_iy$, substituting in the second yields

$$m_iy \equiv c_j - c_i \pmod{m_j},$$

and consequently it must be true that

$$(m_i, m_j)|(c_i - c_j)$$

If this is the case, then y is unique $(\text{mod } m_j/(m_i, m_j))$, and x is unique $(\text{mod } m_im_j/(m_i, m_j))$, that is modulo the LCM of m_i and m_j We have thus proved part of the following theorem

THEOREM 3–12 *A necessary and sufficient condition that the system of congruences* $x \equiv c_i \pmod{m_i}$ $(i = 1, 2 \quad , n)$ *be solvable is that for every pair of indices* i j *between* 1 *and* n *inclusive,*

$$(m_i, m_j)|(c_i - c_j)$$

The solution, if it exists is unique modulo the LCM *of* $m_1, \quad , m_n$

Proof To prove the sufficiency we must show the following If every pair from among the n congruences is solvable, and if any two of them are solved to give a single new congruence, then the $n - 1$ congruences consisting of this new one and the remaining $n - 2$ original congruences also have the property that every pair from among them

is solvable. That is, assume that for all i and j with $1 \leq i \leq n$, $1 \leq j \leq n$, it is true that $(m_i, m_j) | (c_i - c_j)$, and let the solution of

$$x \equiv c_1 \pmod{m_1}, \qquad x \equiv c_2 \pmod{m_2}$$

be

$$x \equiv f \pmod{\langle m_1, m_2 \rangle}.$$

Then we must show that for $3 \leq i \leq n$,

$$(m_i, \langle m_1, m_2 \rangle) | (c_i - f).$$

This can easily be seen by considering the exponent α of any prime p which occurs in the prime-power factorization of $(m_i, \langle m_1, m_2 \rangle)$. Let the exponent of p in the factorization of m_j be β_j, for $j = 1, 2, \ldots, i$. Then p occurs in $\langle m_1, m_2 \rangle$ with exponent $\max(\beta_1, \beta_2)$, so that

$$\alpha = \min(\beta_i, \max(\beta_1, \beta_2)) = \max(\min(\beta_1, \beta_i), \min(\beta_2, \beta_i)).$$

But our assumption is that

$$p^{\min(\beta_1, \beta_i)} | (c_1 - c_i) \qquad \text{and} \qquad p^{\min(\beta_2, \beta_i)} | (c_2 - c_i),$$

and since $p^{\beta_1} | (c_1 - f)$ and $p^{\beta_2} | (c_2 - f)$ we see, by writing

$$c_1 - c_i = (c_1 - f) + (f - c_i),$$

$$c_2 - c_i = (c_2 - f) + (f - c_i),$$

that

$$p^{\min(\beta_1, \beta_i)} | (c_i - f) \qquad \text{and} \qquad p^{\min(\beta_2, \beta_i)} | (c_i - f),$$

so that also $p^{\alpha} | (c_i - f)$. Since p^{α} was an arbitrary prime-power factor of $(m_i, \langle m_1, m_2 \rangle)$, it follows that

$$(m_i, \langle m_1, m_2 \rangle) | (c_i - f),$$

and the sufficiency of the condition is proved.

Finally, solving the first two congruences simultaneously, we get a solution which is unique (mod $\langle m_1, m_2 \rangle$); solving this with the third, we get a solution unique (mod $\langle m_3, \langle m_1, m_2 \rangle \rangle$), that is, unique (mod $\langle m_1, m_2, m_3 \rangle$), etc.

As a consequence of Theorem 3–12, we have the following important result.

THEOREM 3–13 (*Chinese Remainder Theorem*). *Every system of linear congruences in which the moduli are relatively prime in pairs is solvable, the solution being unique modulo the product of the moduli.*

PROBLEMS

1 Solve the congruence $6x + 15y \equiv 9 \pmod{18}$

2 Solve simultaneously

$$x \equiv 1 \pmod 2,$$

$$x \equiv 1 \pmod 3,$$

$$x \equiv 3 \pmod 4,$$

$$x \equiv 4 \pmod 5$$

3 Suppose that the system of congruences

$$x \equiv a_i \pmod{m_i}, \qquad i = 1, 2, \quad , n,$$

is to be solved, where $(m_i, m_j) = 1$ for all i, j with $i \neq j$ Put

$$M = m_1 \quad m_n,$$

and for $i = 1, \quad , n$, let $y = b_i$ be a solution of the congruence

$$\frac{M}{m_i} y \equiv 1 \pmod{m_i}$$

Then show that the solution x of the original system is given by

$$x \equiv \sum_{i=1}^{n} a_i b_i \frac{M}{m_i} \pmod M$$

4 *Show that given a, b, and n, with $(a\ b) = 1$, there is an x such that*

$$(ax + b, n) = 1$$

[*Hint* If $p|a$ and $p|n$ then $p\nmid(ax + b)$ for any x If $p|n$ and $p\nmid a$, there is a solution of

$$ax + b \equiv 1 \pmod p$$

Use the Chinese Remainder Theorem]

3-5 Congruences of higher degree. *We consider now the congruence*

$$f(x) = a_0x^n + a_1x^{n-1} + \quad + a_n \equiv 0 \pmod m,$$

where the a_i are not all congruent to zero (mod m) If $m = \prod_{i=1}^{r} p_i^{\alpha_i}$, then clearly the given congruence is equivalent to the system of congruences

$$f(x) \equiv 0 \pmod{p_1^{\alpha_1}}, \qquad , \qquad f(x) \equiv 0 \pmod{p_r^{\alpha_r}}$$

If for each i with $1 \leq i \leq r$, c_i is a root of $f(x) \equiv 0 \pmod{p_i^{\alpha_i}}$, then by the Chinese Remainder Theorem there is a solution x_0 of the system

$$x \equiv c_1 \pmod{p_1^{\alpha_1}}, \quad \ldots, \quad x \equiv c_r \pmod{p_r^{\alpha_r}},$$

and this x_0, which is unique modulo m, is a solution of the original congruence. Consequently, the number of solutions of the original congruence is the product of the numbers of roots of the congruences modulo the prime-power divisors of m. Hence we can restrict our attention to the case where the modulus is a power of a prime.

The reduction can easily be carried a step further, so that we have only to consider the higher degree congruence with prime modulus, together with a number of linear congruences with prime moduli. The idea is that the solutions of

$$f(x) \equiv 0 \pmod{p^\alpha} \tag{1}$$

are to be found among those of

$$f(x) \equiv 0 \pmod{p^\beta} \tag{2}$$

with $\beta < \alpha$. Suppose that for some $\beta < \alpha$ a solution of (2) is known, say a. (There may be others, of course.) Then every number $a + tp^\beta$ is a solution of (2); it is desired to determine t so that $a + tp^\beta$ is also a solution of

$$f(x) \equiv 0 \pmod{p^{\beta+1}}. \tag{3}$$

By Taylor's theorem,

$$f(a + tp^\beta) = f(a) + tp^\beta f'(a) + \frac{(tp^\beta)^2 f''(a)}{2!} + \cdots + \frac{(tp^\beta)^n f^{(n)}(a)}{n!}.$$

A term $c_j x^j$ in $f(x)$ leads to the term

$$\frac{j(j-1)\ldots(j-k+1)}{k!} c_j x^{j-k} = \binom{j}{k} c_j x^{j-k}$$

in $f^{(k)}(x)/k!$, so that the numbers $f^{(k)}(a)/k!$ are integers. Hence

$$f(a + tp^\beta) \equiv f(a) + tp^\beta f'(a) \pmod{p^{\beta+1}},$$

and (3) becomes

$$f(a) + tp^\beta f'(a) \equiv 0 \pmod{p^{\beta+1}}.$$

Now $p^\beta | f(a)$, so that this reduces to the linear congruence

$$f'(a) \cdot t \equiv -\frac{f(a)}{p^\beta} \pmod{p},$$

of which the number of solutions is

$$\begin{cases} 0, & \text{if } p | f'(a) \text{ but } p \nmid \dfrac{f(a)}{p^\beta}, \\ p, & \text{if } p | f'(a) \text{ and } p | \dfrac{f(a)}{p^\beta}, \\ 1, & \text{if } p \nmid f'(a) \end{cases}$$

The general procedure should now be clear, if all solutions of (2) with $\beta = 1$ are known Choose one of them, say a_1 Corresponding to it there are 0, 1, or p solutions of (2) with $\beta = 2$, to be found by solving a linear congruence If there are no solutions, start over with a different a_1 If there are solutions, choose one and find the corresponding solutions of (2) with $\beta = 3$ If all possibilities are explored in this way all solutions of (1) can eventually be found

Consider for example the congruence

$$f(x) = x^3 - 4x^2 + 5x - 6 \equiv 0 \pmod{27}$$

We first search for roots of

$$x^3 - 4x^2 + 5x - 6 \equiv x^3 + 2x^2 + 2x \equiv 0 \pmod{3}$$

Trying successively 0, 1, 2, we find the only solution of this congruence to be $x \equiv 0 \pmod 3$ Putting $x = 0 + 3t$, we now wish to find t's for which

$$f(0 + 3t) \equiv 0 \pmod 9$$

As above, this reduces to

$$3f'(0)t \equiv -f(0) \pmod 9,$$

or

$$15t \equiv 6 \pmod 9,$$

or

$$5t \equiv 2 \pmod 3,$$

so that $t \equiv 1 \pmod 3$ Putting $t = 1 + 3t_1$, we get $x = 3 + 9t_1$, and we ask that

$$f(3 + 9t_1) \equiv 0 \pmod{27}$$

This gives

$$f(3) + 9t_1 f'(3) \equiv 0 \pmod{27},$$

or

$$9 \cdot 8 \cdot t_1 \equiv 0 \pmod{27},$$

$$t_1 \equiv 0 \pmod{3}.$$

Thus $t_1 = 3t_2$ and $x = 3 + 27t_2$, so that the only solution of the original congruence is

$$x \equiv 3 \pmod{27}.$$

If at any stage in the above argument there had been more than one possibility, each of them would have had to be followed through to obtain corresponding solutions.

PROBLEMS

1. Find all solutions of the congruence

$$x^3 - 3x^2 + 27 \equiv 0 \pmod{1125}.$$

[*Answer:* $x \equiv 51, 426, 801 \pmod{1125}$.]

2. If $f(x)$ is a nonconstant polynomial with integral coefficients, show that it assumes composite values for arbitrarily large x. [*Hint:* Apply Taylor's theorem to $f(m + k \cdot f(m))$.]

3. Suppose that the congruence $f(x) \equiv 0 \pmod{p}$ has as roots the s numbers $x_1, \ldots, x_s$, which are distinct $(\bmod p)$. Show that if $p \nmid f'(x_k)$ for $k = 1, \ldots, s$, then the congruence $f(x) \equiv 0 \pmod{p^\alpha}$ also has exactly s roots, for every $\alpha \geq 1$.

3–6 Congruences with prime moduli. If $f(x)$ and $f_1(x)$ are two polynomials whose corresponding (integral) coefficients are congruent modulo m, then we say that $f(x)$ and $f_1(x)$ are *identically congruent modulo m*, and write

$$f(x) \equiv f_1(x) \pmod{m}. \tag{4}$$

When there is no reference made to the numerical values of x in such a relation, it will always mean identical congruence. It should be noted that (4) is not equivalent to the assertion

$$f(x) \equiv f_1(x) \pmod{m} \qquad \text{for all } x,$$

since, for example, $x^3 \equiv x \pmod{3}$ for all x, but x^3 and x are not identically congruent modulo 3.

If $g(x)$ is also a polynomial with integral coefficients, and $g(x)$ has leading coefficient 1, then $f(x)$ can be divided by $g(x)$ in the usual fashion to obtain a quotient $q_1(x)$ and a remainder $r_1(x)$ Both q_1 and r_1 are polynomials with integral coefficients, and the degree of r_1 is less than that of g If now

$$q_1(x) \equiv q(x) \pmod{m} \qquad \text{and} \qquad r_1(x) \equiv r(x) \pmod{m},$$

then

$$f(x) \equiv g(x)q(x) + r(x) \pmod{m} \tag{5}$$

Such *division modulo m* is not always possible if the leading coefficient of $g(x)$ is not 1, since fractional coefficients may then be encountered In the case of a prime modulus, however, it is possible to find an integer c such that $cg(x)$ has leading coefficient congruent to 1, and so to carry out the division

If in (5),

$$r(x) \equiv 0 \pmod{m},$$

then $g(x)$ is said to *divide* $f(x)$ *modulo* m, or to *be a factor of* $f(x)$ *modulo* m, and we write

$$g(x) | f(x) \pmod{m}$$

If $f(x)$ has no nonconstant factor $\pmod{m}$ of lower degree than itself, it is said to be *irreducible* $\pmod{m}$ If $(x - a) | f(x) \pmod{m}$, then a is said to be a *zero* of $f(x) \pmod{m}$, or a *root* of the congruence $f(x) \equiv 0 \pmod{m}$

In the case of prime modulus, the Euclidean algorithm can easily be generalized, so that we can find the GCD $\pmod{p}$ of any two polynomials For example, if

$$f(x) = x^3 + 2x^2 - x + 1, \qquad g(x) = x^2 - x + 1,$$

then

$$f(x) \equiv x \quad g(x) + (x + 1) \pmod{3},$$

$$g(x) \equiv (x + 1)(x + 1) \pmod{3},$$

and so the GCD $\pmod{3}$ of $f(x)$ and $g(x)$ is the last nonvanishing remainder, namely $x + 1$ But

$$f(x) \equiv (x + 3)g(x) + (x - 2) \pmod{5},$$

$$g(x) \equiv (x + 1)(x - 2) + 3 \pmod{5},$$

$$x - 2 \equiv 3(2x + 1) \pmod{5},$$

so that $f(x)$ and $g(x)$ are relatively prime (i.e., have no common nonconstant divisor) modulo 5.

If the leading coefficient of $g(x)$ is not 1, it may be made so by multiplication by a suitable constant, and then one can find $(f(x), cg(x))$.

It is now possible to prove theorems analogous to Theorems 2–1 through 2–5, and so to show that every polynomial is congruent to a product of polynomials which are irreducible (mod p), and that this representation is unique except for the order of factors and the presence of a set of constant factors whose product is 1 (mod p). Notice that this result is not valid when the modulus is composite, for example,

$$(x - 1)x \equiv (x - 3)(x + 2) \pmod 6,$$

and each of the linear polynomials is of course irreducible.

Another assertion which holds only for prime modulus is that if

$$f(x)g(x) \equiv 0 \pmod p,$$

then either

$$f(x) \equiv 0 \qquad \text{or} \qquad g(x) \equiv 0 \pmod p.$$

For otherwise we may suppose, with no loss in generality, that the leading coefficients of $f(x)$ and $g(x)$ are 1. But then the leading coefficient of $f(x) \cdot g(x)$ is also 1, and therefore not 0.

THEOREM 3–14 (*Factor theorem*). *If a is a root of the congruence*

$$f(x) \equiv 0 \pmod m,$$

then

$$(x - a) | f(x) \pmod m,$$

and conversely.

Proof: Take $g(x) = x - a$ in equation (5). Then $r(x) = r$ is constant, and

$$f(x) \equiv (x - a)q(x) + r \pmod m.$$

Putting $x = a$, we see that $r \equiv 0 \pmod m$ if $f(a) \equiv 0 \pmod m$. Conversely, if

$$f(x) \equiv (x - a)g(x) \pmod m,$$

then

$$f(a) \equiv 0 \pmod m.$$

THEOREM 3–15 (*Lagrange's theorem*) *The congruence*

$$f(x) \equiv 0 \pmod{p}$$

in which

$$f(x) = a_0x^n + \cdots + a_n, \qquad a_0 \not\equiv 0 \pmod{p},$$

has at most n roots

Proof For $n = 1$ this follows from Theorem 3–10 Assume that every congruence of degree $n - 1$ has at most $n - 1$ solutions, and that a is a root of the original congruence Then

$$f(x) \equiv (x - a)q(x) \pmod{p},$$

where $q(x)$ is not identically zero (mod p) and is of degree $n - 1$ It therefore has at most $n - 1$ zeros, say $c_1, \quad , c_r$, where $r \leq n - 1$ Then if c is any number such that $f(c) \equiv 0 \pmod{p}$, then

$$(c - a)q(c) \equiv 0 \pmod{p},$$

so that either

$$c \equiv a \pmod{p}$$

or

$$q(c) \equiv 0 \pmod{p}, \qquad \text{that is, } c = c_i \text{ for some } i,\ 1 \leq i \leq r$$

In other words, the original congruence has at most $r + 1 \leq n$ roots The theorem now follows by the induction principle

Again, this theorem is not valid for composite modulus

PROBLEM

Let $f(x)$ be a polynomial of degree n, with integral coefficients Show that if $n + 1$ consecutive values of $f(x)$ are divisible by a fixed prime p, then $p \mid f(x)$ for every integral x Cf Problem 1, Section 3–2

3–7 The theorems of Fermat, Euler, and Wilson

THEOREM 3–16 (*Fermat's theorem*) *If* $p \nmid a$, *then*

$$a^{p-1} \equiv 1 \pmod{p}$$

Since $\varphi(p) = p - 1$, this is a special case of

THEOREM 3 17 (*Euler's theorem*) *If* $(a, m) = 1$, *then*

$$a^{\varphi(m)} \equiv 1 \pmod{m}$$

Proof: Let $c_1, \ldots, c_{\varphi(m)}$ be a reduced residue system (mod m), and let a be prime to m. Then $ac_1, \ldots, ac_{\varphi(m)}$ is also a reduced residue system (mod m), and

$$\prod_{i=1}^{\varphi(m)} ac_i = a^{\varphi(m)} \prod_{i=1}^{\varphi(m)} c_i \equiv \prod_{i=1}^{\varphi(m)} c_i \pmod{m}.$$

Since $(m, \prod c_i) = 1$, this implies that

$$a^{\varphi(m)} \equiv 1 \pmod{m}.$$

We see from Euler's theorem that if we take the least positive remainders (mod m) of the sequence of powers $a, a^2, a^3, \ldots$ of a number a which is prime to m, we will have a periodic sequence, of period less than or equal to $\varphi(m)$. The period of this sequence—that is, the least positive exponent t such that $a^t \equiv 1 \pmod{m}$—is called the *order of* a (mod m), or the *exponent to which* a *belongs modulo* m, and we write $\operatorname{ord}_m a = t$.

THEOREM 3-18. *If* $a^u \equiv 1 \pmod{m}$, *then* $\operatorname{ord}_m a \mid u$.

Proof: Put $\operatorname{ord}_m a = t$, and let $u = qt + r$, $0 \leq r < t$. Then

$$a^u = a^{qt+r} = (a^t)^q \cdot a^r \equiv a^r \equiv 1 \pmod{m},$$

and if r were different from zero, there would be a contradiction with the definition of t.

THEOREM 3-19. *For every* a *prime to* m, $\operatorname{ord}_m a \mid \varphi(m)$.

Proof: Follows immediately from Theorems 3-16 and 3-18.

As we shall see in the next chapter, the numbers a of order $\varphi(m)$ are of great importance.

The direct converse of Fermat's theorem does not hold; that is, it is not true that if for some a, $a^{m-1} \equiv 1 \pmod{m}$, then m is prime. For example, the powers of 3, reduced modulo 91, are 3, 9, 27, 81, 61, 1, so that $\operatorname{ord}_{91} 3 = 6$. Since $6 \mid 90$, $3^{90} \equiv 1 \pmod{91}$. But 91 is not prime. The clue to the proper converse lies in the observation that $\varphi(m) \leq m - 1$ always, and $\varphi(m) = m - 1$ if and only if m is prime, so that m will certainly be prime if there is an a such that $\operatorname{ord}_m a = m - 1$.

THEOREM 3-20. *If there is an* a *for which* $a^{m-1} \equiv 1 \pmod{m}$, *while none of the congruences* $a^{(m-1)/p} \equiv 1 \pmod{m}$ *holds, where* p *runs over the prime divisors of* $m - 1$, *then* m *is prime.*

Proof By the first hypothesis and Theorem 3–18, the exponent t to which a belongs (mod m) divides $m - 1$ On the other hand, since every proper divisor of $m - 1$ is a divisor of at least one of the numbers $(m - 1)/p$, the second hypothesis and Theorem 3–18 imply that t is not a proper divisor of $m - 1$ Consequently $t = m - 1$ By Theorem 3–19, $m - 1 | \varphi(m)$, and so $m - 1 = \varphi(m)$ and m is prime

In a way, Theorem 3–20 is simply a restatement of the fact that $\varphi(m) = m - 1$ if and only if m is prime But in distinction to this statement, it can actually be used to investigate the primality of large numbers

Fermat's theorem exhibits congruences which have the maximum number of roots allowable by Lagrange's theorem The following theorem gives another important example of such a situation

THEOREM 3–21 *If p is prime and d divides $p - 1$, then there are exactly d roots of the congruence*

$$x^d \equiv 1 \pmod{p}$$

Proof Since $d | p - 1$,

$$x^{p-1} - 1 \equiv (x^d - 1)q(x) \pmod{p},$$

where $q(x)$ is a polynomial of degree $p - 1 - d$ in x By Lagrange's theorem, the congruence

$$q(x) \equiv 0 \pmod{p}$$

has at most $p - 1 - d$ solutions Since $x^{p-1} \equiv 1 \pmod{p}$ has exactly $p-1$ solutions $x^d \equiv 1 \pmod{p}$ must have at least $p-1-(p-1-d)=d$ solutions Since it can have no more than this, it must have exactly d solutions

As another consequence of Fermat's theorem, we have

THEOREM 3 22 (*Wilson s theorem*) *If p is prime, then*

$$(p - 1)! \equiv -1 \pmod{p}$$

Proof Fermat's theorem and Theorem 3–14 show that

$$x^{p-1} - 1 \equiv (x - 1)(x - 2) \quad (x - p + 1) \pmod{p}$$

identically so that the constant terms must be congruent

$$-1 \equiv (-1)^{p-1}(p - 1)! \pmod{p}$$

If p is odd, this gives the theorem. If $p = 2$, then we have

$$-1 \equiv 1 \equiv 1! \pmod 2.$$

The converse of Wilson's theorem does hold.

THEOREM 3-23. *If $m > 1$ and $(m-1)! \equiv -1 \pmod m$, then m is prime.*

Proof: If m is composite, it has a proper divisor $d > 1$. But then

$$(m-1)! \equiv 0 \not\equiv -1 \pmod d,$$

and *a fortiori,*

$$(m-1)! \not\equiv -1 \pmod m.$$

There is another way of obtaining Wilson's theorem which also throws some light on a subject to be considered in much more detail in Chapter 5. Let a be any integer not divisible by the odd prime p, and let b be one of the numbers $1, \ldots, p-1$. Then we know that there is a unique solution (mod p) of the congruence $bx \equiv a \pmod p$. Let b', called the *associate* of b, be that positive solution which is less than p. We must distinguish two cases, according as some b is associated with itself or not. If $b = b'$, then $b^2 \equiv a \pmod p$, so that the congruence $x^2 \equiv a \pmod p$ has a solution; in this case a is said to be a *quadratic residue* of p. If the congruence $x^2 \equiv a \pmod p$ has no solution, a is called a *quadratic nonresidue* of p. (Similar definitions hold for nth power residues and nonresidues.)

If a is a quadratic residue of p, and if ${b_1}^2 \equiv a \pmod p$, then clearly also $(p - b_1)^2 \equiv a \pmod p$; by Lagrange's theorem there are no other solutions. Thus in this case the numbers $1, \ldots, p-1$ can be grouped into $(p-3)/2$ pairs of associates, the product of each pair being congruent to $a \pmod p$, together with the two numbers b_1 and $p - b_1$. Thus

$$(p-1)! = \prod_{b=1}^{p-1} b \equiv a^{(p-3)/2} \cdot b_1(p - b_1) \equiv -a^{(p-1)/2} \pmod p. \quad (6)$$

On the other hand, if a is a quadratic nonresidue of p, the numbers $1, 2, \ldots, p-1$ can be grouped into $(p-1)/2$ pairs of associates, and

$$(p-1)! = \prod_{b=1}^{p-1} b \equiv a^{(p-1)/2} \pmod p. \quad (7)$$

In order to give a uniform statement of (6) and (7), we define the *Legendre symbol* (a/p) (also frequently written $\left(\frac{a}{p}\right)$ or $(a|p)$) to

mean 1 if a is a quadratic residue of p, and -1 if a is a quadratic nonresidue of p Here a is called the "first entry," and p the "second entry" (Note that (a/p) is not yet defined if $p|a$) Then (6) and (7) become

$$(p-1)! \equiv -(a/p)a^{(p-1)/2} \pmod{p} \tag{8}$$

Taking $a = 1$, and noting that the congruence $x^2 \equiv 1 \pmod{p}$ has the solution $x=1$, so that $(1/p)=1$, we have $(p-1)!\equiv -1 \pmod{p}$, which is Wilson's theorem again Substituting in (8), this gives

$$(a/p)a^{(p-1)/2} \equiv 1 \pmod{p},$$

or since $(a/p) = \pm 1$,

$$(a/p) \equiv a^{(p-1)/2} \pmod{p}$$

Thus we have proved

THEOREM 3-24 (*Euler's criterion*) *A necessary and sufficient condition that a be a quadratic residue of an odd prime p is that the congruence*

$$a^{(p-1)/2} \equiv 1 \pmod{p}$$

hold

PROBLEMS

1 Show that if $ab \equiv 1 \pmod{m}$ then

$$\operatorname{ord}_m a = \operatorname{ord}_m b$$

2 Show that if p is an odd prime and $\operatorname{ord}_{p^\alpha} a = 2t$ then

$$a^t \equiv -1 \pmod{p^\alpha}$$

Show that this need not be true if $p = 2$

3 Show that if p is an odd prime and $a^t \equiv -1 \pmod{p}$ then a belongs to an even exponent $2u \pmod{p}$ and t is an odd multiple of u

*4 Show that if p is an odd prime and $p|(x^{2^r}+1)$ then $p \equiv 1 \pmod{2^{r+1}}$ Deduce that there are infinitely many primes congruent to 1 modulo any fixed power of 2

5 Show that for $a > 1$ and $n > 0$ $n|\varphi(a^n-1)$

6 Use Theorem 3-20 with $a = 2$ to show that 389 is prime

*7 Show that if $(a\ b) = 1$, p is an odd prime not dividing $a+b$, and

$$d \left| \frac{a^p + b^p}{a+b}, \right.$$

then $d \equiv 1 \pmod{p}$ Cf Problem 7 Section 2 2 [*Hint* Let q be a prime

divisor of $(a^p + b^p)/(a + b)$, so that $a^p \equiv -b^p \pmod{q}$. Show that a k exists such that $b|(kq + a)$, and put $r = (kq + a)/b$; then $r^p \equiv -1 \pmod{q}$, so that $\text{ord}_q(-r) = 1$ or p. If the first alternative is eliminated, then $p|(q - 1)$.]

8. Show that the congruence $f(x) \equiv 0 \pmod{p}$, of degree $m < p$, has m roots if and only if $f(x)|(x^p - x) \pmod{p}$.

9. Use Theorem 3-21 and the method of Section 3-5 to show that if p is prime and $d|p - 1$, then there are exactly d roots $(\text{mod } p^n)$ of the congruence

$$x^d \equiv 1 \pmod{p^n},$$

where $n \geq 1$.

*10. Show that the Diophantine equation

$$(n - 1)! = n^k - 1$$

has only the solutions $n, k = 2, 1; 3, 1;$ and $5, 2$. [*Hint:* Prove and use the following statements:

(a) There is no solution with n even and larger than 2.

(b) $n - 1|(n - 2)!$ if n is odd and larger than 5.

(c) $(n - 1)^2|(n^k - 1)$ only if $(n - 1)|k$. It is useful to write $n^k - 1 = ((n - 1) + 1)^k - 1$.]

CHAPTER 4

PRIMITIVE ROOTS AND INDICES

4–1 Integers belonging to a given exponent (mod p)

THEOREM 4–1 *If* $\operatorname{ord}_m a = t$, *then* $\operatorname{ord}_m a^n = t/(n, t)$

Proof. Let $(n, t) = d$ Then, since $a^t \equiv 1 \pmod{m}$, we have

$$(a^t)^{n/d} = (a^n)^{t/d} \equiv 1 \pmod{m},$$

so that if $\operatorname{ord}_m a^n = t'$, then

$$t' \left| \frac{t}{d} \right. \tag{1}$$

But from the congruence

$$(a^n)^t \equiv 1 \pmod{m},$$

we have that $t|nt'$, or

$$\frac{t}{d} \left| \frac{n}{d} t' \right.$$

Since

$$\left(\frac{t}{d}, \frac{n}{d}\right) = 1,$$

this gives

$$\frac{t}{d} \left| t' \right. \tag{2}$$

Combining (1) and (2), we have

$$t' = \frac{t}{d}$$

THEOREM 4–2 *If any integer belongs to* $t \pmod{p}$, *then exactly* $\varphi(t)$ *incongruent numbers belong to* $t \pmod{p}$

Proof Assume that $\operatorname{ord}_p a = t$ Then by Theorem 3–19, $t|(p - 1)$, so that by Theorem 3–21 there are exactly t roots of the congruence

$x^t \equiv 1 \pmod{p}$. But all the numbers $a, a^2, \ldots, a^t$ are roots of this congruence and they are incongruent (mod p), so that they are the only roots. By Theorem 4–1, the powers of a which belong to $t \pmod{p}$ are the numbers a^n with $(n, t) = 1$, $1 \leq n \leq t$, and there are just $\varphi(t)$ of these numbers.

THEOREM 4–3. *If $t \mid (p-1)$, there are $\varphi(t)$ incongruent numbers (mod p) which belong to $t \pmod{p}$.*

Proof: Let d run over the divisors of $p-1$, and for each such d let $\psi(d)$ be the number of integers among $1, 2, \ldots, p-1$ of order $d \pmod{p}$. By Theorem 3–19 and Fermat's theorem, each of the integers $1, 2, \ldots, p-1$ belongs to exactly one of the d. Hence

$$\sum_{d \mid p-1} \psi(d) = p - 1.$$

But also

$$\sum_{d \mid p-1} \varphi(d) = p - 1,$$

by Theorem 3–9, so that

$$\sum_{d \mid p-1} \psi(d) = \sum_{d \mid p-1} \varphi(d).$$

By Theorem 4–2, the value of $\psi(d)$ is either zero or $\varphi(d)$ for each d, and we deduce from the last equation that $\psi(d) = \varphi(d)$ for each d dividing $p-1$.

If $\operatorname{ord}_m a = \varphi(m)$, then a is said to be a *primitive root* of m. The importance of this notion lies in the fact that if g is such a primitive root, then its powers

$$g, g^2, \ldots, g^{\varphi(m)}$$

are distinct (mod m) and are all relatively prime to m; they therefore constitute a reduced residue system modulo m. Thus we have a convenient way of representing all the elements of a reduced residue system, some of the implications of which are to be found later in this chapter and in the problems.

It follows immediately from Theorem 4–1 that the other primitive roots of m are those powers g^k for which $(k, \varphi(m)) = 1$. Either from this remark or from Theorem 4–3 we have

THEOREM 4–4. *There are exactly $\varphi(\varphi(p))$ primitive roots of a prime p.*

PROBLEMS

1 Show that if $\operatorname{ord}_p a = t$ $\operatorname{ord}_p b = u$ and $(t, u) = 1$, then $\operatorname{ord}_p (ab) = tu$

2 Show that if $p \equiv 1 \pmod 4$ and g is a primitive root of p, then so is $-g$ Show by a numerical example that this need not be the case if $p \equiv 3 \pmod 4$

3 Show that if p is of the form $2^m + 1$ and $(a/p) = -1$, then a is a primitive root of p

4. Show that if p is an odd prime and $\operatorname{ord}_p a = t > 1$, then

$$\sum_{k=1}^{t-1} a^k \equiv -1 \pmod p$$

4–2 Primitive roots of composite moduli Theorem 4–4 immediately brings the following questions to mind Do all numbers have primitive roots? If not, which do and how many are there? The first question is easily answered in the negative since 8 has none $\varphi(8) = 4$, but

$$\operatorname{ord}_8 1 = 1, \qquad \operatorname{ord}_8 3 = 2, \qquad \operatorname{ord}_8 5 = 2, \qquad \operatorname{ord}_8 7 = 2$$

On the other hand, since 5 is a primitive root of 6 there are composite numbers having primitive roots The answer to the second question is, therefore, not just the set of primes, as one might think

After the primes themselves, the simplest moduli to treat are the prime powers We need a preliminary result

THEOREM 4 5 (a) *If p is prime, then*

$$a \equiv b \pmod{p^n} \quad \textit{implies} \quad a^{p^s} \equiv b^{p^s} \pmod{p^{n+s}} \tag{3}$$

for every pair of positive integers n, s

(b) *If p is an odd prime and $p \nmid b$, then*

$$a^{p^s} \equiv b^{p^s} \pmod{p^{n+s}} \quad \textit{implies} \quad a \equiv b \pmod{p^n} \tag{4}$$

for every pair of positive integers n s

Proof (a) We use induction on s Assume that $a \equiv b \pmod{p^n}$ Then

$$a = hp^n + b,$$

and

$$a^p = (hp^n)^p + \binom{p}{1}(hp^n)^{p-1}b + \quad + \binom{p}{p-1} hp^n b^{p-1} + b^p$$

Now p occurs in the numerator of the binomial coefficient

$$\binom{p}{k} = \frac{p!}{k!(p-k)!},$$

but it is not present in the denominator if $0 < k < p$; hence $p \mid \binom{p}{k}$ for $0 < k < p$, and for such k,

$$p^{n+1} \mid \binom{p}{k} p^{n(p-k)}.$$

But also $p^{n+1} \mid p^{np}$, so that $a^p \equiv b^p \pmod{p^{n+1}}$. Hence (3) is correct for $s = 1$ and every n.

Suppose that (3) is valid for $s = 1, 2, \ldots, s'$, for every n, and suppose that $a \equiv b \pmod{p^n}$. (This congruence is now to be regarded as the premise of (3) with $s = s' + 1$.) Then the induction hypothesis with $s = 1$ gives

$$a^p \equiv b^p \pmod{p^{n+1}}. \tag{5}$$

Using (5) as the premise of (3) with $s = s'$ gives

$$(a^p)^{p^{s'}} \equiv (b^p)^{p^{s'}} \pmod{p^{n+1+s'}},$$

or

$$a^{p^{s'+1}} \equiv b^{p^{s'+1}} \pmod{p^{n+(s'+1)}},$$

which is the conclusion of (3) with $s = s' + 1$. Hence (3) holds for every pair of positive integers n, s.

(b) We first prove (4) for $s = 1$, by induction on n; we suppose throughout that $p \neq 2$ and $p \nmid b$. Thus we wish to show that

$$a^p \equiv b^p \pmod{p^{n+1}} \qquad \text{implies} \qquad a \equiv b \pmod{p^n}.$$

If $a^p \equiv b^p \pmod{p^2}$, then also $a^p \equiv b^p \pmod{p}$, and, by Fermat's theorem, $a \equiv b \pmod{p}$. Now assume that

$$a^p \equiv b^p \pmod{p^{n'}} \qquad \text{implies} \qquad a \equiv b \pmod{p^{n'-1}}$$

and that

$$a^p \equiv b^p \pmod{p^{n'+1}}.$$

Then

$$a^p \equiv b^p \pmod{p^{n'}},$$

so that

$$a \equiv b \pmod{p^{n'-1}}.$$

But if $a = up^{n'-1} + b$, then

$$a^p \equiv b^p + up^{n'}b^{p-1} \pmod{p^{n'+1}}$$

if $p > 2$, and so $p|u$, whence $a = u_1p^{n'} + b$ and

$$a \equiv b \pmod{p^{n'}},$$

and the implication follows by induction on n.

To complete the proof of (4), we use induction on s. Assume that

$$a^{p^{s-1}} \equiv b^{p^{s'-1}} \pmod{p^{n+s'-1}} \quad \text{implies} \quad a \equiv b \pmod{p^n}$$

for every n, and assume that

$$a^{p^{s'}} \equiv b^{p^{s'}} \pmod{p^{n+s'}}$$

Then

$$(a^p)^{p^{s-1}} \equiv (b^p)^{p^{s-1}} \pmod{p^{n+s'}},$$

whence

$$a^p \equiv b^p \pmod{p^{n+1}},$$

so that, by what we have just proved,

$$a \equiv b \pmod{p^n}$$

The result follows by induction on s.

Let p be a prime. Then if $p^n|a$ and $p^{n+1}\nmid a$, we will write for brevity $p^n\|a$.

THEOREM 4–6. *If p is an odd prime, $\operatorname{ord}_p a = t$, and $p^z\|(a^t - 1)$, then*

$$\operatorname{ord}_{p^n} a = t\ p^{\max(0,n-z)}$$

Proof. Assume the hypotheses of the theorem are satisfied. If $n \leq z$, then $p^n|(a^t - 1)$. This is not true for any exponent $t' < t$, since if $p^n|(a^{t'} - 1)$, then $p|(a^{t'} - 1)$, so that $t|t'$. Hence in this case $\operatorname{ord}_{p^n} a = t$, which proves the theorem for $n \leq z$.

If $n > z$, we get from Theorem 4–5 and the last hypothesis of the present theorem that

$$a^{tp^{n-z}} \equiv 1 \pmod{p^n}$$

We must show that $a^d \not\equiv 1 \pmod{p^n}$ if d is a proper divisor of tp^{n-z}. Let $d = t_1p^r$, where $r \leq n - z$ and $t_1|t$, and assume that $a^{t_1p^r} \equiv 1 \pmod{p^n}$.

By Theorem 4–5 again, $a^{t_1} \equiv 1 \pmod{p^{n-r}}$,

whence $$a^{t_1} \equiv 1 \pmod{p},$$

so that $t | t_1$, and $t = t_1$. Since $p^z \| (a^t - 1)$ and $p^{n-r} | (a^t - 1)$, we have $n - r \leq z$, whence $n - z = r$.

We can use Theorem 4-6 to construct primitive roots of p^n, where p is an odd prime; that is, numbers which belong to $p^{n-1}(p-1)$ modulo p^n. Let g be a primitive root of p. Then if $p^2 \nmid (g^{p-1} - 1)$, Theorem 4-6 shows that

$$\operatorname{ord}_{p^n} g = (p-1)p^{n-1},$$

and g is also a primitive root of p^n for all positive n. If $p^2 | (g^{p-1} - 1)$, then $g + p$ is also a primitive root of p, and

$$\begin{aligned}(g+p)^{p-1} - 1 &\equiv g^{p-1} + (p-1)g^{p-2}p - 1 \\ &\equiv (p-1)pg^{p-2} \not\equiv 0 \pmod{p^2},\end{aligned}$$

so that by Theorem 4-6,

$$\operatorname{ord}_{p^n}(g+p) = (p-1)p^{n-1},$$

and $g + p$ is a primitive root of p^n for all positive n. We have thus proved

THEOREM 4-7. *Any power of an odd prime has a primitive root.*

Turning now to other composite numbers, it is convenient to define a function $\lambda(m)$, called the *universal exponent* of m:

$$\lambda(1) = 1,$$

$$\lambda(2^\alpha) = \begin{cases} \varphi(2^\alpha) = 2^{\alpha-1} \text{ if } \alpha = 1, 2, \\ \frac{1}{2}\varphi(2^\alpha) = 2^{\alpha-2} \text{ if } \alpha > 2, \end{cases}$$

$$\lambda(p^\alpha) = \varphi(p^\alpha), \qquad p \text{ an odd prime},$$

$$\lambda(2^\alpha \cdot p_1^{\alpha_1} \dots p_r^{\alpha_r}) = \langle \lambda(2^\alpha), \lambda(p_1^{\alpha_1}), \dots, \lambda(p_r^{\alpha_r}) \rangle,$$

$$p_1, \dots, p_r \text{ distinct odd primes.}$$

Euler's theorem can now be strengthened somewhat.

THEOREM 4-8. *If* $(a, m) = 1$, *then*

$$a^{\lambda(m)} \equiv 1 \pmod{m}.$$

Proof: (a) If $m = 2^\alpha$ with $\alpha \leq 2$, this is Euler's theorem.
(b) If $m = 2^\alpha$ with $\alpha > 2$, a must be odd, so that $a^2 \equiv 1 \pmod{2^3}$.

By Theorem 4–5, $(a^2)^{2^{\alpha-3}} = a^{2^{\alpha-2}} \equiv 1 \pmod{2^\alpha}$

(c) If $m = p^\alpha$, where p is odd, we have Euler's theorem again

(d) Finally, suppose that $m = 2^\alpha \quad p_1^{\alpha_1} \quad p_r^{\alpha_r}$ By (a), (b) and (c), each of the congruences

$$a^{\lambda(2^\alpha)} \equiv 1 \ (\mathrm{mod}\ 2^\alpha),$$

$$a^{\lambda(p_i^{\alpha_i})} \equiv 1 \pmod{p_i^{\alpha_i}} \qquad i = 1, 2, \quad , r,$$

holds Since all the exponents of a divide $\lambda(m)$, it follows that

$$a^{\lambda(m)} \equiv 1 \pmod{2^\alpha},$$

$$a^{\lambda(m)} \equiv 1 \pmod{p_i^{\alpha_i}}, \qquad i = 1, 2, \quad , r,$$

and hence

$$a^{\lambda(m)} \equiv 1 \pmod{m}$$

As a complement to Theorem 4–8, we have

THEOREM 4–9 *$\lambda(m)$ is the smallest positive value of x such that $a^x \equiv 1 \pmod{m}$ for every a prime to m That is, there is always an integer which belongs to $\lambda(m) \pmod{m}$*

Proof (a) If $m = 1 \quad \lambda(1) = 1$ and $\mathrm{ord}_1 1 = 1$

(b) If $m = 2 \quad \lambda(2) = 1$ and $\mathrm{ord}_2 1 = 1$

(c) If $m = 4 \quad \lambda(4) = 2$ and $\mathrm{ord}_4 3 = 2$

(d) If $m = 2^\alpha$, $\alpha > 2 \quad \lambda(2^\alpha) = 2^{\alpha-2}$ and $\mathrm{ord}_{2^\alpha} 5 = 2^{\alpha-2}$ For if $\mathrm{ord}\ 5 = d$, then $d | 2^{\alpha-2}$, so that $d = 2^\beta$, where $\beta \leq \alpha - 2$ But it is easily proved by induction on α that for $\alpha \geq 3$

$$5^{2^{\alpha-3}} = 1 + 2^{\alpha-1} h_\alpha,$$

where h_α is an odd number Hence $5^{2^{\alpha-3}} \not\equiv 1 \pmod{2^\alpha}$ and $\beta = \alpha - 2$

(e) If $m = p^\alpha$ with p odd $\lambda(p^\alpha) = \varphi(p^\alpha)$, and by Theorem 4–7, p^α has a primitive root

(f) If m is arbitrary Let $m = p_1^{\alpha_1} \quad p_r^{\alpha_r}$, with $2 \leq p_1 < \quad < p_r$ By the first five steps of the proof, there are numbers $a_1, \quad , a_r$ such that $\mathrm{ord}_{p_i^{\alpha_i}} a_i = \lambda(p_i^{\alpha_i})$ for $i = 1, \quad , r$ By the Chinese Remainder Theorem, there is a single integer a such that $a \equiv a_i \pmod{p_i^{\alpha_i}}$ for $i = 1 \quad , r$, and the order of $a \pmod{p_i^{\alpha_i}}$ is the same as that of a_i, for each i Hence if $a^x \equiv 1 \pmod{m}$, then $\lambda(p_i^{\alpha_i}) | x$ for each i, and so $\lambda(m)$, since it is the LCM of the numbers $\lambda(p_i^{\alpha_i})$, also divides x By Theorem 4 8, $\mathrm{ord}_m a = \lambda(m)$

An integer whose order (mod m) is $\lambda(m)$ is called a *primitive λ-root* of m. Theorem 4–9 says in effect that every modulus has a primitive λ-root.

As a combination of Theorems 4–2 and 4–9, we have

THEOREM 4–10. *There are $\varphi(\lambda(m))$ primitive λ-roots of m congruent to powers of any given primitive λ-root.*

Notice that in general it is not the case that all the primitive λ-roots are congruent to powers of a single one. For example, if $m = 2^4$, $g = 5$, then the only other primitive λ-root congruent to a power of 5 is 13, while 3 and 11 are also primitive λ-roots.

Moreover, we can now deduce

THEOREM 4–11. *The numbers having primitive roots are*

$$1, 2, 4, p^\alpha, 2p^\alpha,$$

where p is any odd prime.

Proof: We already know that 1, 2, 4, and p^α have primitive roots. Since

$$\lambda(2p^\alpha) = \langle \lambda(2), \lambda(p^\alpha) \rangle = \lambda(p^\alpha) = \varphi(p^\alpha) = \varphi(2p^\alpha),$$

every number $2p^\alpha$ has primitive roots. On the other hand, if $m = 2^\alpha \cdot {p_1}^{\alpha_1} \cdots {p_r}^{\alpha_r}$ with $\alpha > 2$, p_i odd, $r \geq 1$, then

$$\lambda(m) \leq \tfrac{1}{2}\varphi(2^\alpha)\varphi({p_1}^{\alpha_1}) \cdots \varphi({p_r}^{\alpha_r}) \leq \tfrac{1}{2}\varphi(m),$$

and if

$$m = {p_1}^{\alpha_1} \cdots {p_r}^{\alpha_r} \qquad \text{with } r > 1$$

or if

$$m = 4{p_1}^{\alpha_1} \cdots {p_r}^{\alpha_r} \qquad \text{with } r \geq 1,$$

then each of the numbers $\lambda(4)$, $\lambda({p_i}^{\alpha_i})$ is even, so that again

$$\lambda(m) \leq \tfrac{1}{2}\varphi(m).$$

This completes the proof.

The problem of efficiently finding a primitive root of a given large modulus q is not simple. It is, of course, a finite problem, and for specific modulus can be solved by successively testing the elements of a reduced residue system. A slightly more rapid method is indicated in Problem 4 at the end of the next section, but it also is laborious for large q, particularly if $\varphi(q)$ has many distinct prime divisors.

PROBLEMS

1 Show that if q has primitive roots, there are $\varphi(\varphi(q))$ of them, and their product is congruent to 1 (mod q) if $q > 6$ [*Hint* Represent all the primitive roots in terms of a single one]

2 Find all the primitive roots of 25

*3 It is an unproved conjecture that no two consecutive integers, except 8 and 9, are perfect powers Show that at any rate the only pair x, y satisfying the conditions

$$3^x - 2^y = 1, \qquad x > 1, y > 1$$

is 2, 3 [*Hint* Use Theorem 4–6 and Problem 3, Section 3–7 to show that $3^{x-1} \mid y$]

4 Show that if g is a primitive root of p^2, then the roots of the congruence

$$x^{p-1} \equiv 1 \pmod{p^2}$$

are g^{np}, $n = 1, 2, \quad , p-1$, that is, that these numbers are distinct roots, and there are no others [*Hint* Show that the congruence has only $p-1$ roots Cf Problem 9, Section 3–7]

4–3 Indices Let q be a number having primitive roots and let g be one of them Then the numbers $g, g^2, \quad , g^{\varphi(q)}$ are distinct (mod q), and they are all prime to q, therefore they constitute a reduced residue system (mod q) The relation between a number a and the exponent of a power of g which is congruent to a (mod q) is very similar to the relation between an ordinary positive real number x and its logarithm This exponent is called an *index* of a to the base g, and written "$\operatorname{ind}_g a$" That is, $\operatorname{ind}_g a$ will stand for any number t such that $g^t \equiv a \pmod{q}$, it is defined only if $(a, q) = 1$, and is unique modulo $\varphi(q)$ The following facts are immediate consequences of the definition

THEOREM 4–12 *If g is a primitive root of q and $a \equiv b \pmod{q}$, then*

$$\operatorname{ind}_g a \equiv \operatorname{ind}_g b \pmod{\varphi(q)},$$

$$\operatorname{ind}_g (ab) \equiv \operatorname{ind}_g a + \operatorname{ind}_g b \pmod{\varphi(q)},$$

and

$$\operatorname{ind}_g a^n \equiv n \operatorname{ind}_g a \pmod{\varphi(q)}$$

The procedure for finding the indices of the elements of a reduced residue system is quite simple if a primitive root is known If g is a primitive root of q, construct a table of two rows and $\varphi(q)$ columns, of

which the second row consists of the integers $1, 2, \ldots, \varphi(q)$ in order. In the first row enter g in the first column. Multiply this by g and reduce modulo q for the element in the second column, multiply this result by g and reduce modulo q for the element in the third column, etc. (When the table is complete, the last element in the first row should be 1.) Then the index of any element of the first row appears directly below that element.

If, for example, $q = 17$ and $g = 3$, we have the table

a:	3	9	10	13	5	15	11	16	14	8	7	4	12	2	6	1
ind a:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

,

while if $q = 18$ and $g = 5$, we have

a:	5	7	17	13	11	1
ind a:	1	2	3	4	5	6

.

By Theorem 4–1, if $\operatorname{ord}_m g = \varphi(m)$, then

$$\operatorname{ord}_m g^n = \frac{\varphi(m)}{(n, \varphi(m))},$$

so that a is a primitive root of m if and only if $(\operatorname{ind} a, \varphi(m)) = 1$. Thus in the above table we see that the primitive roots of 18 are 5 and 11, since the only numbers less than $\varphi(18) = 6$ and prime to it are 1 and 5.

Indices are quite useful in solving binomial congruences. For example, the congruence

$$10x \equiv 8 \pmod{18}$$

implies

$$5x \equiv 4 \pmod{9},$$

which implies

$$\operatorname{ind} 5 + \operatorname{ind} x \equiv \operatorname{ind} 4 \pmod{6},$$

$$\operatorname{ind} x \equiv \operatorname{ind} 4 - \operatorname{ind} 5 \pmod{6}.$$

Since 2 is a primitive root of 9, we construct the table as before:

n:	2	4	8	7	5	1
ind n:	1	2	3	4	5	6

.

Thus $$\text{ind } x \equiv 2 - 5 \equiv 3 \pmod 6,$$

whence $$x \equiv 8 \pmod 9,$$

so that $$x \equiv 8 \text{ or } 17 \pmod{18}$$

The investigation of the congruence $x^n \equiv c \pmod m$, where $(m, c) = 1$, can be reduced to the study of the solutions of

$$x^n \equiv c \pmod p$$

by previously explained methods But the latter is entirely equivalent to

$$n \text{ ind } x \equiv \text{ind } c \pmod{p-1},$$

which has solutions if and only if $(n, p-1) | \text{ind } c$, if this condition is satisfied there are $d = (n, p-1)$ roots This criterion has the disadvantage that it requires knowledge of the value of ind c, the *following is more useful*

THEOREM 4–13 *Let $(c\ q) = 1$, q being any number which has primitive roots Then a necessary and sufficient condition that the congruence*

$$x^n \equiv c \pmod q \tag{6}$$

be solvable is that

$$c^{\varphi(q)/d} \equiv 1 \pmod q,$$

where $d = (n, \varphi(q))$

Proof By an argument similar to that just given for prime modulus, a necessary and sufficient condition for the solvability of (6) is that $\text{ind } c \equiv 0 \pmod d$ This is equivalent to

$$\frac{\varphi(q)}{d} \text{ind } c \equiv 0 \pmod{\varphi(q)},$$

or, what is the same thing,

$$c^{\varphi(q)/d} \equiv 1 \pmod q$$

If $x^n \equiv c \pmod m$ is solvable, and $(m, c) = 1$, c is said to be an *nth power residue* of m, otherwise a nonresidue

THEOREM 4–14 *The number of incongruent nth power residues of q is $\varphi(q)/d$, and these residues are the roots of the congruence*

$$x^{\varphi(q)/d} \equiv 1 \pmod q$$

Proof: The second statement is a paraphrase of Theorem 4–13. Since q has a primitive root g, the roots of the congruence $x^{\varphi(q)/d} \equiv 1 \pmod{q}$ are the numbers g^t for which

$$g^{t\varphi(q)/d} \equiv 1 \pmod{q},$$

and this requires that $d|t$. But the number of multiples t of d with $1 \leq t \leq \varphi(q)$ is exactly $\varphi(q)/d$. (Note that this is a generalization of Theorem 3–24.)

PROBLEMS

1. Show that if g and h are primitive roots of p, then

$$\text{ind}_h a \equiv \text{ind}_g a \cdot \text{ind}_h g \pmod{p-1}.$$

2. Given that 2 is a primitive root of 29, construct a table of indices, and use it to solve the following congruences:

(a) $17x \equiv 10 \pmod{29}$ (b) $17x^2 \equiv 10 \pmod{29}$.

3. Develop a method for solving the congruence

$$Ax^2 + Bx + C \equiv 0 \pmod{p}$$

by use of indices, when p is an odd prime which does not divide A. (First show that the given congruence can be replaced by one in which the coefficient of x^2 is 1; then, after suitable modifications, complete the square.) Apply your method to

(a) $17x^2 - 3x + 10 \equiv 0 \pmod{29}$ (b) $17x^2 - 4x + 10 \equiv 0 \pmod{29}$.

4. Let q be a number having primitive roots. Show that h is a primitive root of q if and only if h is an rth power nonresidue of q for every prime r dividing $\varphi(q)$. [*Hint:* Write $h = g^k$, where g is a primitive root of q, and show that each of the allegedly equivalent statements is equivalent to the equation $(k, \varphi(q)) = 1$.] By eliminating all the appropriate powers of the elements of a reduced residue system, find all the primitive roots of (a) 13, (b) 29. (Cf. Problem 3, Section 4–1.)

*5. Show that for $x > 1$ the quantity

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

where q is prime, $q^n > 2$, and $y = x^{q^{n-1}}$, has the following properties:

(a) For $x \equiv 1 \pmod{q}$, $q||f(x)$, and for $x \not\equiv 1 \pmod{q}$, $q \nmid f(x)$.

(b) $f(x) > q$.

(c) $(f(x), x) = 1$

(d) If $p \neq q$ and $p|f(x)$, then $p \equiv 1 \pmod{q^n}$

Deduce that there exists a prime $p \equiv 1 \pmod{q^n}$, and then by taking $x = p_1 \cdots p_r$, where each $p_i \equiv 1 \pmod{q^n}$, that there are infinitely many primes $p \equiv 1 \pmod{q^n}$ (Cf Problem 4, Section 3-7, where $q = 2$)

4-4 An application to Fermat's conjecture A simple way of attempting to show that the equation

$$x^n + y^n = z^n \tag{7}$$

has no nonzero solutions for $n \geq 3$ is to show that the infinitely many congruences

$$x^n + y^n \equiv z^n \pmod{p}, \qquad p = 2, 3, 5, \ldots,$$

impose absurd conditions on the variables For example, in the case $n = 3$ the congruence

$$x^3 + y^3 \equiv z^3 \pmod{7}$$

implies that $7|xyz$ For if $7 \nmid u$, then $u^6 \equiv 1 \pmod{7}$, so that $u^3 \equiv \pm 1 \pmod{7}$, and for no choice of signs is $\pm 1 \pm 1 \equiv \pm 1 \pmod{7}$ If we could find infinitely many primes p such that

$$x^3 + y^3 \equiv z^3 \pmod{p}$$

implies $p|xyz$, then clearly equation (7) could have no nonzero solution for $n = 3$ We shall show that this cannot be done, either for $n = 3$ or for larger n The proof depends on the following combinatorial lemma

THEOREM 4-15 *If the numbers $1, 2, \ldots, N$ are distributed into m disjoint classes, and if* $N > me$, then at least one class contains the difference of two of its elements*

Proof Suppose that the numbers $1, 2, \ldots, N$ have been put into m disjoint classes so that no class contains the difference of any two of its elements. Let a class having the largest number of elements be called Z_1, then if Z_1 is composed of $x_1, \ldots, x_{n_1}$, we have $N \leq n_1 m$ If the names are so chosen that $x_1 < x_2 < \cdots < x_{n_1}$, the $n_1 - 1$ differences

$$x_2 - x_1, x_3 - x_1, \ldots, x_{n_1} - x_1 \tag{8}$$

are also integers between 1 and N, inclusive, and by assumption they

*Here contrary to our convention, the number $e = 2.718 \ldots$ is not an integer

lie in the remaining $m - 1$ classes. Let Z_2 be a class in which the largest number of differences (8) lie. If Z_2 contains the n_2 differences

$$x_\alpha - x_1, x_\beta - x_1, \ldots, \tag{9}$$

then clearly $n_1 - 1 \leq n_2 (m - 1)$. Now the $n_2 - 1$ differences

$$x_\beta - x_\alpha, x_\gamma - x_\alpha, \ldots \tag{10}$$

do not lie in either Z_1 or Z_2, so they must be distributed among the remaining $m - 2$ classes. If n_3 is the largest number of differences (10) in any single class, then $n_2 - 1 \leq n_3(m - 2)$. Continuing in this way, we have

$$n_\mu - 1 \leq n_{\mu+1}(m - \mu), \tag{11}$$

for $\mu = 1, 2, \ldots, m_1$, where m_1 is such that $n_{m_1} = 1$. From (11), we have

$$\frac{n_\mu}{(m-\mu)!} \leq \frac{1}{(m-\mu)!} + \frac{n_{\mu+1}}{(m-\mu-1)!}, \qquad \mu = 1, 2, \ldots, n_1$$

and adding all these inequalities gives

$$\frac{n_1}{(m-1)!} \leq \frac{1}{(m-1)!} + \frac{1}{(m-2)!} + \cdots + \frac{1}{(m-m_1)!} < e.$$

Hence

$$N \leq n_1 m < m!e,$$

and the proof is complete.

THEOREM 4–16. *There are only finitely many primes p for which every solution of the congruence*

$$x^n + y^n \equiv z^n \pmod{p} \tag{12}$$

is such that $p|xyz$. More precisely, if $p > n!e + 1$, then (12) has solutions such that $p\nmid xyz$.

Proof: First suppose that $n|(p - 1)$, so that $p - 1 = nr$ for suitable r. Let g be a primitive root of p, and let s_m be the smallest positive residue (mod p) of g^m. Then the numbers $s_1, \ldots, s_{p-1}$ are the integers $1, 2, \ldots, p - 1$ in some order. We now classify the numbers s_m according to the residue classes of their subscripts (mod n), so that for each t with $0 \leq t \leq n - 1$, the numbers

$$s_t, s_{t+n}, \ldots, s_{t+(r-1)n}$$

form a single class, there being n classes altogether. By Theorem

4–15, if $p - 1 > n^2e$, then some class contains three elements, say s_{t+jn}, s_{t+kn}, s_{t+ln}, such that

$$s_{t+jn} - s_{t+kn} = s_{t+ln}$$

But then

$$g^{t+jn} \equiv g^{t+kn} + g^{t+ln} \pmod{p},$$

whence

$$g^{jn} \equiv g^{kn} + g^{ln} \pmod{p},$$

and the numbers $x = g^k$, $y = g^l$, $z = g^j$ give the desired solution of (12)

If $n \nmid (p-1)$, let $d = (n, p-1)$ Then by what we have just proved, the congruence

$$x^d + y^d \equiv z^d \pmod{p}$$

is solvable with $p \nmid xyz$ if $p - 1 > d^2e$ But by Theorem 4–13, any dth power is an nth power residue of p, since

$$(u^d)^{(p-1)/(p-1,n)} = (u^d)^{(p-1)/d} = u^{p-1} \equiv 1 \pmod{p}$$

Thus there exist an x_1, y_1, and z_1 such that

$$x_1^n \equiv x^d, \qquad y_1^n \equiv y^d, \qquad z_1^n \equiv z^d \pmod{p},$$

and hence

$$x_1^n + y_1^n \equiv z_1^n \pmod{p}$$

PROBLEM

Show that if $x^3 + y^3 \equiv z^3 \pmod 9$, then $3 \mid xyz$ Use the result of this section, together with the method of Section 3–5 to show that this is an atypical phenomenon that for fixed n the congruence

$$x^n + y^n \equiv z^n \pmod{p^\alpha}$$

has a solution such that $p \nmid xyz$ if p is sufficiently large and $\alpha \geq 1$

REFERENCES

Section 4–4

The main theorem was first proved by L E Dickson, *Journal fur die Reine und Angewandte Mathematik* (Berlin) **135**, 134–141 181 188 (1909) The proof given here is due to I Schur, *Jahresbericht der Deutschen Matematikervereinigung* (Leipzig) **25**, 114–117 (1917) Dickson's proof is more difficult, but shows that equation (12) is solvable if $p > cn^4$ for suitable c.

CHAPTER 5

QUADRATIC RESIDUES

5–1 Introduction. The subject of nth power residues is a large and difficult one. It happens, however, that for the case $n = 2$ many elegant and important results can be obtained by elementary considerations, and it is to these that we now turn our attention. A fundamental tool in the investigation of quadratic residues is Euler's criterion, Theorem 3–24, which was generalized somewhat in Theorem 4–13, namely that a necessary and sufficient condition for a number a prime to q to be a quadratic residue of q is that $a^{\varphi(q)/2} \equiv 1 \pmod{q}$. (Here q is a number greater than 2 having primitive roots.) The two problems with which we shall deal are, first, to extend this criterion to general composite moduli (in so doing we shall find that it suffices to restrict our attention to odd prime moduli); and, second, to find an efficient method for determining all the primes of which a given integer a is a quadratic residue.

The prime 2 plays a rather special role in the theory of quadratic residues, not so much because of an intrinsic difference between it and the odd primes (which does exist, as we saw in the discussion of primitive roots of composite moduli) as because the congruences are quadratic; in a similar fashion, 3 must be treated separately when considering cubic congruences. On account of this, we shall use the symbol p to represent an *odd* prime throughout this chapter.

5–2 Composite moduli

THEOREM 5–1. *A number a prime to m is a quadratic residue of m if and only if it is a quadratic residue of all odd prime divisors of m and is congruent to* 1 (mod 4) *if* $m \equiv 4 \pmod{8}$, *and congruent to* 1 (mod 8) *if* $8|m$.

Proof: Let

$$m = 2^{\alpha} \prod_{i=1}^{r} p_i^{\alpha_i}.$$

Then the congruence

$$x^2 \equiv a \pmod{m}$$

is equivalent to the system of congruences

$$x^2 \equiv a \pmod{2^\alpha}$$

$$x^2 \equiv a \pmod{p_i^{\alpha_i}}, \qquad i = 1, \quad , r,$$

so that a is a quadratic residue of m if and only if it is a quadratic residue of every prime-power divisor of m

(a) *If a is a quadratic residue of p, it is a quadratic residue of p^α,* and conversely (The converse is trivial) For *if a is a residue of p* it follows from Euler's criterion that

$$a^{(p-1)/2} \equiv 1 \pmod{p}$$

By Theorem 4–5,

$$a^{p^{\alpha-1}(p-1)/2} \equiv 1 \pmod{p^\alpha},$$

and so a is a quadratic residue of p^α by Euler's criterion again If a is a quadratic residue of p and $x_1^2 \equiv a \pmod{p}$, then also $(-x_1)^2 \equiv a \pmod{p}$, and these are the only solutions, by Lagrange's theorem Using the method of Section 3–5, it is easily seen that if $p \nmid f'(x_1)$ for each root x_1 of $f(x) \equiv 0 \pmod{p}$, then the congruence $f(x) \equiv 0 \pmod{p^\alpha}$ has exactly as many roots as the congruence with prime modulus In this case $f(x) = x^2 - a$, $f'(x) = 2x$, and since p is odd and $p \nmid x_1$, it follows that $x^2 \equiv a \pmod{p^\alpha}$ has exactly two solutions $\pmod{p^\alpha}$ if *a is a quadratic residue of p*

(b) *For modulus 2^α the situation is more complicated* If *a is odd,*

(i) $x^2 \equiv a \pmod{2}$ is always uniquely solvable,

(ii) $x^2 \equiv a \pmod{4}$ is solvable if and only if $a \equiv 1 \pmod{4}$, it then has two roots,

(iii) $x^2 \equiv a \pmod{2^\alpha}$, for $\alpha \geq 3$, is solvable if and only if $a \equiv 1 \pmod{8}$, it then has four roots

The first statement is obvious, and the second follows immediately upon noting that any odd square is congruent to 1 (mod 4) (The two solutions are, of course, ± 1) For the case $\alpha \geq 3$, recall that it was shown in the proof of Theorem 4–9 that 5 is a primitive λ-root of 2^α, and so the numbers 5 5^2, 5^3, , $5^{2^{\alpha-2}}$ are distinct $\pmod{2^\alpha}$ Since by the binomial theorem $(1 + 2^2)^n \equiv 1 + 4n \pmod{8}$, $5^n \equiv 1 \pmod{8}$ *if and only if n is even* Thus 5^2, 5^4, 5^6, , $5^{2^{\alpha-2}}$ are $2^{\alpha-3}$ numbers which are distinct $\pmod{2^\alpha}$ and which are all con-

gruent to 1 (mod 8). But since there are exactly $2^{\alpha-3}$ numbers in a complete residue system (mod 2^α) which are congruent to 1 (mod 8), it follows that every number congruent to 1 (mod 8) is congruent to 5^{2n} for some n, so that every $a \equiv 1 \pmod 8$ is a quadratic residue of 2^α. (If $5^{2n} \equiv a \pmod{2^\alpha}$, the congruence $x^2 \equiv a \pmod{2^\alpha}$ has the solution $x = 5^n$.) On the other hand, every odd square is of the form $8k + 1$, so that $x^2 \equiv a \pmod{2^\alpha}$ is certainly not solvable unless $a \equiv 1 \pmod 8$.

Assume that $b^2 \equiv a \pmod{2^\alpha}$, and let x be any other solution of this congruence, so that also $x^2 \equiv a \pmod{2^\alpha}$. Then $x^2 - b^2 = (x - b)(x + b) \equiv 0 \pmod{2^\alpha}$. Both x and b are odd, so $x - b$ and $x + b$ are even; since $(x - b, x + b) = 2(x, b)$, one of them has 2 as a simple factor. Since

$$\frac{x - b}{2} \cdot \frac{x + b}{2} \equiv 0 \pmod{2^{\alpha-2}},$$

one factor must be divisible by $2^{\alpha-2}$, that is,

$$\frac{x \pm b}{2} \equiv 0 \pmod{2^{\alpha-2}}.$$

Hence $x \equiv \pm b \pmod{2^{\alpha-1}}$, and x must be congruent to one of $\pm b$, $\pm b + 2^{\alpha-1} \pmod{2^\alpha}$. It is immediately verified that each of these four numbers is a solution.

Combining the results of (a) and (b) gives Theorem 5-1. By the Chinese Remainder Theorem, the number of roots of $x^2 \equiv a \pmod m$ is the product of the numbers of roots of the congruences with prime-power moduli. As shown above, if a is a quadratic residue of m, this number is 2 for each odd prime-power factor, and 1, 2 or 4 according as α is 0 or 1, 2, or more than two, where $2^\alpha \| m$. Hence we have

THEOREM 5-2. *If* $(a, m) = 1$ *and the congruence* $x^2 \equiv a \pmod m$ *is solvable, it has exactly* $2^{\sigma+\tau}$ *solutions, where* σ *is the number of distinct odd prime divisors of* m *and* τ *is* 0, 1, *or* 2 *according as* $4 \nmid m$, $2^2 \| m$, *or* $8 \mid m$.

PROBLEMS

1. Decide whether 5 is a quadratic residue of 44.

2. Show that the product of the quadratic residues of a prime p is congruent to 1 or $-1 \pmod p$ according as $p \equiv -1$ or $1 \pmod 4$. [*Hint:* Write the residues of p in terms of a primitive root.]

*3 Prove the following generalization of Wilson's theorem The product of the positive integers less than m and prime to m is congruent to $-1 \pmod m$ if $m = 4$, p^a, or $2p^a$, and to 1 otherwise. [*Hint* Proceed as in the second proof of Wilson's theorem, associating a and a' if $aa' \equiv 1 \pmod m$ Use Theorem 5–2 to count the elements associated with themselves]

5–3 *Quadratic residues of primes, and the Legendre symbol* As was seen in Section 5–2, the quadratic residues of powers of 2 can be given explicitly, and the quadratic residues of powers of an odd prime are *identical with those of the prime itself* Consequently, there remains only the investigation of quadratic residues of odd primes Hereafter we shall make use of the simplifying notation of the Legendre symbol (a/p), introduced at the end of Chapter 3 It will be recalled that for $(a, p) = 1$, we put

$$(a/p) = \begin{cases} 1, & \text{if } a \text{ is a quadratic residue of } p, \\ -1, & \text{if } a \text{ is a quadratic nonresidue of } p \end{cases}$$

For completeness we put $(a/p) = 0$ if $p|a$, so that (a/p) is now defined for every odd prime p

THEOREM 5–3 *The Legendre symbol* (a/p) *has the following properties*

(a) $(ab/p) = (a/p)(b/p)$ *Thus the product of two residues or two nonresidues is a residue, the product of a residue and a non residue is a nonresidue*

(b) *If* $a \equiv b \pmod p$, *then* $(a/p) = (b/p)$

(c) $(a^2/p) = 1$ *if* $p \nmid a$

(d) $(-1/p) = (-1)^{(p-1)/2}$

Proof The first two parts are obvious if $p|ab$, so suppose that $p \nmid ab$ In the proof of Theorem 3 24 it was shown that $(a/p) \equiv a^{(p-1)/2} \pmod p$ Hence

$$(ab/p) \equiv (ab)^{(p-1)/2} = a^{(p-1)/2} b^{(p-1)/2} \equiv (a/p)(b/p) \pmod p,$$

and since (a/p) assumes only the values ± 1, it follows that $(ab/p) = (a/p)(b/p)$ Property (d) also follows immediately from this congruence Properties (b) and (c) are obvious

It follows from Theorem 5–3 that in investigating the Legendre symbol (a/p), there will be no loss in generality in assuming that a is

a positive prime. For example, Theorem 5-3 shows that

$$\begin{aligned}(-48/31) &= (-1/31)(48/31) = (-1/31)(3/31)(16/31)\\ &= (-1/31)(3/31)\\ &= (30/31)(3/31) = (2/31)(3/31)(5/31)(3/31)\\ &= (2/31)(5/31),\end{aligned}$$

so that $(-48/31)$ can be evaluated either from

$$(-48/31) = (-1)^{\frac{1}{2}(31-1)}(3/31) = -(3/31)$$

or from

$$(-48/31) = (2/31)(5/31).$$

In general, (a/p) can be written as the product of Legendre symbols, in which the first entries are the distinct prime divisors of a which divide a to an odd power.

Although it will be used only in the case where a is prime, the following theorem is valid for all a's for which $p\nmid a$.

THEOREM 5-4 (*Gauss's lemma*). *If μ is the number of elements of the set a, $2a$, ..., $\frac{1}{2}(p-1)a$ whose numerically least residues* (mod p) *are negative, then*

$$(a/p) = (-1)^{\mu}.$$

Example: If $a = 3, p = 31$, the numerically least residues (mod 31) of $3 \cdot 1, 3 \cdot 2, \ldots, 3 \cdot 15$ are 3, 6, 9, 12, 15, -13, -10, -7, -4, -1, 2, 5, 8, 11, 14; thus $\mu = 5$, $(3/31) = -1$, and from the above numerical example, $(-48/31) = 1$.

Proof: Replace the numbers of the set $a, 2a, \ldots, \frac{1}{2}(p-1)a$ by their numerically smallest residues (mod p); denote the positive ones by $r_1, r_2, \ldots$ and the negative ones by $-r_1', -r_2', \ldots$. Clearly no two r_i's are equal, and no two r_i''s are equal. If $m_1 a \equiv r_i$ and $m_2 a \equiv -r_j'$ (mod p), then $r_i = r_j'$ would imply $a(m_1 + m_2) \equiv 0$ (mod p), which implies $m_1 + m_2 \equiv 0$ (mod p), and this is impossible because the m's are strictly between 0 and $p/2$. Hence the $(p-1)/2$ numbers r_i, r_i' are distinct integers between 1 and $(p-1)/2$ inclusive, and are therefore exactly the numbers $1, 2, \ldots, (p-1)/2$ in some order. Hence,

$$a \cdot 2a \cdots \frac{p-1}{2}a \equiv (-1)^{\mu}\frac{p-1}{2}! \pmod{p},$$

$$a^{(p-1)/2} \equiv (-1)^{\mu} \pmod{p}.$$

Since also $a^{(p-1)/2} \equiv (a/p) \pmod{p}$, it follows that

$$(a/p) \equiv (-1)^{\mu} \pmod{p},$$

and finally,

$$(a/p) = (-1)^{\mu}$$

In distinction to Euler's criterion, Gauss's lemma can be used to characterize the primes of which a given integer a is a quadratic residue For example, if $a = 2$, then μ is the number of numbers $2m$, with $1 \leq m \leq (p-1)/2$, which are greater than $p/2$, this is clearly true if and only if $m > p/4$ Thus if we write $[x]$ to stand for the largest integer not exceeding x, it follows that

$$\mu = \frac{p-1}{2} - \left[\frac{p}{4}\right]$$

If now

$p=8k+1$, then $\mu=4k-[2k+\frac{1}{4}]=4k-2k\equiv 0 \pmod 2$,

$p=8k+3$, then $\mu=4k+1-[2k+\frac{3}{4}]=4k+1-2k\equiv 1 \pmod 2$,

$p=8k+5$, then $\mu=4k+2-[2k+1+\frac{1}{4}]=2k+1\equiv 1 \pmod 2$,

$p=8k+7$, then $\mu=4k+3-[2k+1+\frac{3}{4}]=2k+2\equiv 0 \pmod 2$,

and we deduce that 2 is a quadratic residue of primes of the form $8k \pm 1$ and a nonresidue of primes $8k \pm 3$ Since it happens that the quantity $(p^2-1)/8$ satisfies exactly the same congruences as μ above, this result can be stated in the following form

THEOREM 5–5 $(2/p) = (-1)^{(p^2-1)/8}$

As an application of Theorem 5–5, we have

THEOREM 5–6 (a) *2 is a primitive root of the prime $p = 4q + 1$ if q is an odd prime*

(b) *2 is a primitive root of $p = 2q + 1$ if q is a prime of the form $4k + 1$*

(c) *-2 is a primitive root of $p = 2q + 1$ if q is a prime of the form $4k - 1$*

Proof (a) If $\operatorname{ord}_p 2 = t$, then $t|p-1$, which is equivalent to saying that $t|4q$ Aside from 4, every proper divisor of $4q$ is also a divisor of $2q$, and if $2^4 \equiv 1 \pmod p$, then p is 5 and q is not prime Hence it suffices to show that $2^{2q} \not\equiv 1 \pmod p$ But

$$2^{2q} = 2^{(p-1)/2} \equiv (2/p) \equiv (-1)^{(p^2-1)/8} \equiv (-1)^{2q^2+q} \equiv -1 \pmod p$$

Parts (b) and (c) can be proved in a similar fashion. Part (a) shows that 2 is a primitive root of 13, 29, 53, . . . ; part (b) shows that 2 is a primitive root of 11, 59, 83, . . . , and part (c) that -2 is a primitive root of 7, 23, 47, ... It is an unproved conjecture that 2 is a primitive root of infinitely many primes, which would follow from Theorem 5–5 if it could be shown that there are infinitely many primes p of the kinds described in (a) and (b).

Referring to (a), this requires a proof that the function $4x + 1$ assumes prime values for infinitely many prime arguments. Unfortunately, there is no nonconstant rational function known to have this property. If one could prove that the function $x + 2$ has it, one would have proved a conjecture which is one of the outstanding problems in additive number theory: that there are infinitely many "twin primes," such as 17 and 19, or 101 and 103.

PROBLEMS

1. Apply Gauss's lemma to determine the primes of which -2 is a quadratic residue, and show that your result is consistent with Theorem 5–3, parts (a) and (d), and Theorem 5–5.

2. Complete the proof of Theorem 5–6.

*3. Show that 7 is a primitive root of any prime of the form $2^{4n} + 1$ with $n > 0$. [*Hint:* Show first that it suffices to prove that $(7/p) = -1$, and then show that any *prime* of the specified form is congruent to 3 or 5 (mod 7). Note that $2^4 \equiv 2 \pmod 7$.]

4. Show that the numbers $6k - 1$ and $6k + 1$ are twin primes if and only if the equation $k = 6xy \pm x \pm y$ has no solution in positive integers x and y for any of the four choices of sign. [Note that if $6k + 1 = mn$, then $m \equiv n \equiv \pm 1 \pmod 6$.] Show that this characterizes all the twin primes except 3 and 5.

5–4 The law of quadratic reciprocity. Gauss's lemma can be used to establish a deep property of the Legendre symbol which is an essential tool both in determining the quadratic character of a prime $q \pmod p$ and in finding the primes p of which q is a quadratic residue.

THEOREM 5–7 (*Quadratic reciprocity law*). *If p and q are distinct odd primes, then*

$$(p/q)(q/p) = (-1)^{\frac{1}{2}(p-1)\cdot\frac{1}{2}(q-1)}.$$

In other words, $(p/q) = (q/p)$ unless both p and q are of the form $4k - 1$, in which case $(p/q) = -(q/p)$

Proof By Gauss's lemma, the numbers μ and ν in the equations

$$(q/p) = (-1)^{\mu}, \qquad (p/q) = (-1)^{\nu}$$

are the numbers of the multiples

$$q, 2q, \quad , \frac{p-1}{2} q,$$

and

$$p, 2p, \quad , \frac{q-1}{2} p$$

whose absolutely smallest residues (mod p) and (mod q) respectively are negative, and we need only show that

$$\mu + \nu \equiv \frac{p-1}{2} \frac{q-1}{2} \pmod 2$$

If y is chosen so that

$$-\frac{p}{2} < qx - py < \frac{p}{2},$$

then clearly $qx - py$ is the numerically smallest residue of qx (mod p) From this inequality we get

$$\frac{qx}{p} - \frac{1}{2} < y < \frac{qx}{p} + \frac{1}{2}$$

Thus y is unique and non-negative, if $y = 0$ then $qx - py = qx > 0$, *and there is no contribution to μ in this case* Moreover, we see that for $x \leq (p-1)/2$,

$$\frac{qx}{p} - \frac{1}{2} < \frac{q-1}{2},$$

so that also $y \leq (q-1)/2$ The number μ denotes therefore the number of combinations of x and y from the sequences

$$(p) \qquad 1, 2, \quad , \frac{p-1}{2}$$

and

$$(q) \qquad 1, 2, \quad , \frac{q-1}{2},$$

respectively, for which

$$0 > qx - py > -\frac{p}{2}.$$

Similarly, ν is the number of pairs x and y from the sequences (p) and (q) respectively, for which

$$0 > py - qx > -\frac{q}{2}.$$

For any other pair x and y from (p) and (q) respectively, either

$$py - qx > \frac{p}{2}$$

or

$$py - qx < -\frac{q}{2};$$

let there be λ of the former and ρ of the latter. Then clearly

$$\frac{p-1}{2} \cdot \frac{q-1}{2} = \mu + \nu + \lambda + \rho.$$

Finally, as x and y run through (p) and (q) respectively, the numbers

$$x' = \frac{p+1}{2} - x \qquad \text{and} \qquad y' = \frac{q+1}{2} - y$$

run through the same sequences, but in the opposite order. And if $py - qx > p/2$, then

$$\begin{aligned} py' - qx' &= p\left(\frac{q+1}{2} - y\right) - q\left(\frac{p+1}{2} - x\right) \\ &= \frac{p-q}{2} - (py - qx) < \frac{p-q}{2} - \frac{p}{2} = -\frac{q}{2}. \end{aligned}$$

Hence $\lambda = \rho$, and

$$\frac{p-1}{2} \cdot \frac{q-1}{2} = \mu + \nu + 2\lambda \equiv \mu + \nu \pmod 2.$$

By combining the law of quadratic reciprocity with the properties of the Legendre symbol mentioned in Theorem 5-3, it is easy to evaluate (q/p) if p and q do not lie beyond the extent of the available

tables of factorizations of integers For example, 2819 and 4177 are both primes and $4177 \equiv 1 \pmod 4$, *so that*

$$\begin{aligned}(2819/4177) &= (4177/2819) = (1358/2819) = (2\ \ 7\ \ 97/2819)\\ &= (2/2819)(7/2819)(97/2819)\\ &= -1 \cdot -(2819/7)(2819/97) = (5/7)(6/97)\\ &= (7/5)(2/97)(97/3)\\ &= (2/5)(1/3) = -1,\end{aligned}$$

and so 2819 is not a quadratic residue of 4177

Moreover, the quadratic reciprocity law can be used to determine the primes p of which a given prime q is a quadratic residue This result, which is contained in the next theorem, has sometimes been taken as the quadratic reciprocity law, rather than Theorem 5–7

THEOREM 5–8 *Every $p \neq q$ can be uniquely represented in the form $4qk \pm a$, where $0 < a < 4q$ and $a \equiv 1 \pmod 4$ For a fixed odd prime q, the solutions of the equation $(q/p) = 1$ are exactly the primes $p \neq q$ such that the corresponding a is a quadratic residue of q, that is, $(q/p) = (a/q)$ The numbers a such that*

$$0 < a < 4q, \qquad a \equiv 1 \pmod 4 \qquad \text{and} \qquad (a/q) = 1, \qquad (1)$$

are given by the least positive residues (mod $4q$) of the numbers $1^2, 3^2, 5^2, \ldots, (q-2)^2$

Proof Clearly every odd number can be written in the form $4qk' + a'$ where $1 \leq a' < 4q$ and a' is odd If $a' \equiv 1 \pmod 4$, take $a = a'$ and $k = k'$, while if $a' \equiv -1 \pmod 4$, take $a = 4q - a'$ and $k = k' + 1$ Thus every odd number, and therefore every p, has a representation either as $4qk + a$ (if $a' \equiv 1 \pmod 4$) or as $4qk - a$ (if $a' \equiv -1 \pmod 4$) This proves the first sentence

If $p \equiv a \pmod{4q}$, then $p \equiv 1 \pmod 4$ so that

$$(q/p) = (p/q) = (a/q)$$

If, on the other hand, $p \equiv -a \pmod{4q}$, then $p \equiv -1 \pmod 4$, and

$$\begin{aligned}(q/p) &= (-1)^{\frac{1}{2}(p-1)\frac{1}{2}(q-1)}(p/q) = (-1)^{\frac{1}{2}(p-1)\frac{1}{2}(q-1)}(-a/q)\\ &= (-1)^{\frac{1}{2}(p-1)\ \frac{1}{2}(q-1)}(-1)^{\frac{1}{2}(q-1)}(a/q)\\ &= (-1)^{\frac{1}{2}(p+1)\ \frac{1}{2}(q-1)}(a/q) = (a/q)\end{aligned}$$

Thus always $(q/p) = (a/q)$, which proves the second sentence

Finally, if $(a/q) = 1$, there is an x such that

$$x^2 \equiv a \pmod{q} \qquad \text{and} \qquad 1 \leq x \leq q-1,$$

whence also

$$(q-x)^2 \equiv a \pmod{q} \qquad \text{and} \qquad 1 \leq q-x \leq q-1.$$

Since either x or $q-x$ is odd—say x'—we have

$$x'^2 \equiv a \pmod{q}, \qquad 1 \leq x' \leq q-2, \qquad x' \equiv 1 \pmod{2}.$$

But then

$$x'^2 \equiv 1 \equiv a \pmod{4},$$

so that

$$x'^2 \equiv a \pmod{4q},$$

and the proof is complete.

To illustrate, take $q = 3$. Then the only integer satisfying the conditions (1) is 1, so that 3 is a quadratic residue of primes $12k \pm 1$. Every other odd number is of one of the forms $12k \pm 3$ or $12k \pm 5$, and no prime except 3 occurs in the progressions $12k \pm 3$. Hence $(3/p)$ is completely determined by the equations

$$(3/p) = \begin{cases} 1, & \text{if } p \equiv \pm 1 \pmod{12}, \\ -1, & \text{if } p \equiv \pm 5 \pmod{12}. \end{cases}$$

Similarly, taking $q = 17$ we consider the squares

$$1^2, 3^2, 5^2, 7^2, 9^2, 11^2, 13^2, 15^2,$$

which reduce (mod 68) to

$$1, 9, 25, 49, 13, 53, 33, 21.$$

We have that 17 is a quadratic residue of primes of the forms

$$68k \pm 1, 9, 13, 21, 25, 33, 49, \text{ and } 53,$$

and a nonresidue of primes of the forms

$$68k \pm 5, 29, 37, 41, 45, 57, 61, \text{ and } 65;$$

17 itself is the only prime of the forms $68k \pm 17$.

In general, out of the $2q$ progressions $4qk \pm a$, $q-1$ contain only primes of which q is a residue, $q-1$ contain only primes of which q is a nonresidue, and two (either $4qk \pm q$ or $4qk \pm 3q$, according as $q \equiv 1$ or $3 \pmod{4}$) contain no primes besides q itself.

Determining the primes of which a composite number is a quadratic residue is somewhat more complicated To illustrate, consider the problem of finding the primes p for which $(10/p) = 1$ This requires that either $(2/p) = (5/p) = 1$ or $(2/p) = (5/p) = -1$, so that either

$$p \equiv \pm 1 \pmod 8 \qquad \text{and} \qquad p \equiv \pm 1 \pmod{10}$$

or

$$p \equiv \pm 3 \pmod 8 \qquad \text{and} \qquad p \equiv \pm 3 \pmod{10},$$

all combinations of signs being allowed Thus we have the following pairs of congruences, each pair to be solved simultaneously

$p \equiv 1 \pmod 8$, $p \equiv 1 \pmod{10}$
$p \equiv -1 \pmod 8$, $p \equiv -1 \pmod{10}$
$p \equiv 1 \pmod 8$, $p \equiv -1 \pmod{10}$
$p \equiv -1 \pmod 8$, $p \equiv 1 \pmod{10}$
$p \equiv 3 \pmod 8$, $p \equiv 3 \pmod{10}$
$p \equiv -3 \pmod 8$, $p \equiv -3 \pmod{10}$
$p \equiv 3 \pmod 8$, $p \equiv -3 \pmod{10}$
$p \equiv -3 \pmod 8$, $p \equiv 3 \pmod{10}$

Solving (by the method of Problem 3, Section 3–4, for example), we obtain

$$p \equiv 1, -1, 9, 31, 3, -3, 27, 13 \pmod{40},$$

that is, 10 is a quadratic residue of the primes $40k \pm 1, 3, 9, 13$, and a nonresidue of the others

PROBLEMS

1 Evaluate the Legendre symbols (503/773) and (501/773)

2 Characterize the primes of which 5 is a quadratic residue, those of which 6 is a quadratic residue

3 Show that if $p = 4m + 1$ and $d|m$, then $(d/p) = 1$ [*Hint* Let q be a prime divisor of m, and consider separately the cases $q = 2$ and $q > 2$]

4 Deduce from the representation $N = 6119 = 82^2 - 5 \cdot 11^2$ that if $p|N$, then $(5/p) = 1$ Use this to find the factorization of N (It suffices to consider $p < 80$) Use similar ideas to factor $43993 = 211^2 - 2^4 \cdot 33$

5 Prove that 4751 is prime

5–5 An application. It is clear that if a given integer a is congruent to 1 (mod p) for every prime p, then $a = 1$, since $p|(a - 1)$ implies $p \leq |a| + 1$ unless $a - 1 = 0$ Here we have an instance of the following principle if an assertion involving a congruence holds

for every prime modulus p, then the statement with the congruence replaced by the corresponding equation may be implied. With this in mind, it is natural to ask whether it is true that if, for fixed integers a and n, a is an nth power modulo p for every p, then a must be an nth power. (Saying that a is an nth power (mod p) means, of course, that a is congruent to the nth power of some integer; in other words, that a is an nth power residue of p.) Unfortunately, this is not quite the case: if the congruence $x^n \equiv a \pmod{p}$ is solvable for every p, then $a = b^n$ for some b if $8 \nmid n$, but if $8 \mid n$, either $a = b^n$ or $a = 2^{n/2}b^n$. Powers of 2 higher than the second cause difficulty here, just as they did in the study of primitive roots. (Cf. Problem 1 at the end of this section.)

At the present time, the theorem just stated cannot be proved in a simple way. Even in the special case $n = 2$ which we now treat, it is necessary to use a rather deep result about the existence of primes in certain arithmetic progressions.

THEOREM 5–9. *A fixed integer is a quadratic residue of every prime if and only if it is a square.*

Proof: If $a = b^2$, the congruence $x^2 \equiv a \pmod{p}$ has the solution $x \equiv b \pmod{p}$ for every p.

Suppose, on the other hand, that a is not a square. Then it can be written as $\pm m^2 p_1 p_2 \cdots p_r$, where $r \geq 1$ and $p_i \neq p_j$ if $i \neq j$. Suppose first that a is positive; then we wish to show the existence of a prime p such that

$$(a/p) = (m^2 p_1 \cdots p_r/p) = (p_1/p) \cdots (p_r/p) = -1.$$

We attempt to find a p such that $(p_i/p) = 1$ if $1 \leq i < r$, while $(p_r/p) = -1$. Here, of course, one of the primes $p_1, \ldots, p_r$ may be 2. But since 2 is a quadratic residue of primes $8k \pm 1$, and a nonresidue of primes $8k \pm 5$, the following statement is true for *every* prime q:

If $p \equiv 1 \pmod{4q}$, then $(q/p) = 1$. On the other hand, for each q there is a u such that $q \nmid u$, $u \equiv 1 \pmod{4}$, and if $p \equiv u \pmod{4q}$, then $(q/p) = -1$.

The first part is obvious. When $q = 2$, u may be taken to be 5 in the second part, while if $q > 2$, u may be taken as any of the N numbers remaining out of the q integers between 1 and $4q$ which are congruent to 1 (mod 4), after the removal of (a) the least positive

residues (mod $4q$) of the $(q-1)/2$ squares $1^2, 3^2, \quad , (q-2)^2$, and (b) that one of $q, 3q$ which is congruent to 1 (mod 4) Since

$$N = q - \frac{q-1}{2} - 1 = \frac{q-1}{2} \geq 1,$$

such an integer u exists

Now consider the system of congruences

$$x \equiv 1 \pmod{4p_1}$$

$$x \equiv 1 \pmod{4p_{r-1}}$$

$$x \equiv u \pmod{4p_r},$$

when $r > 1$, or the single congruence

$$x \equiv u \pmod{4p_1}$$

when $r = 1$, where u is the number characterized above, with $q = p_r$ or p_1 For $r > 1$, the necessary and sufficient condition that the system be solvable is, by Theorem 3–12, that for all i and j,

$$(4p_i, 4p_j) \mid (c_i - c_j),$$

where $c_i = 1$ if $i < r$ and $c_i = u$ if $i = r$ Since $c_i \equiv 1 \pmod 4$ for every i, this requirement is clearly satisfied, so that the system can be replaced by a single congruence

$$x \equiv u' \pmod{4p_1 \quad p_r},$$

where $(u', 4p_1 \quad p_r) = 1$ If now $p = 4p_1 \quad p_r k + u'$ or $p = 4p_1 k + u$, in the cases $r > 1$ and $r = 1$, respectively, then

$$(a/p) = 1 \quad 1 \quad (-1) = -1,$$

and it is seen that in the case $a > 0$, the theorem is a consequence of the famous

DIRICHLET'S THEOREM *If s and t are relatively prime, there are infinitely many primes of the form $sk + t$*

Proofs of special cases of Dirichlet's theorem have been indicated in Problem 4 of Section 3–7 and Problem 5 of Section 4–3 The general theorem is proved in Volume II of this work

If $a = -m^2$, then $(a/p) = -1$ if $p \nmid a$ and $p \equiv -1 \pmod 4$ If $p_1, \quad , p_k$ is any set of primes of the form $4k - 1$, then the number

$4p_1 \dots p_k - 1$ has a prime divisor of this same form distinct from $p_1, \dots, p_k$, so there are infinitely many primes of this form, and in particular there is one which does not divide a. For this p, $(a/p) = -1$.

If $a = -m^2 p_1 \cdots p_r$, where $r \geq 1$ and $p_i \neq p_j$ if $i \neq j$, then we must find a p such that $p \nmid a$ and

$$(-1/p)(p_1/p) \dots (p_r/p) = -1.$$

But if p is a prime for which $(-a/p) = -1$, as determined above, then $p \equiv 1 \pmod 4$, so that $(a/p) = (-a/p)$. The proof is complete.

PROBLEMS

1. Show that the congruence

$$x^{2^\alpha} \equiv 2^{2^{\alpha-1}} \pmod p$$

has a solution for every prime p, if $\alpha \geq 3$. [*Hint:* Consider the factorization

$$x^{2^\alpha} - 2^{2^{\alpha-1}} = (x^2-2)(x^2+2)((x-1)^2+1)((x+1)^2+1)(x^{2^2}+2^{2^2}) \cdots (x^{2^{\alpha-1}}+2^{2^{\alpha-1}}),$$

and show that every p divides one of the first three factors for suitable x.]

2. Show that if the congruence $x^n \equiv a \pmod m$ is solvable for every m, then a is an nth power. [*Hint:* Consider the moduli $p^{\alpha+1}$, where $p^\alpha \| a$ and α is positive.]

5-6 The Jacobi symbol. As was pointed out at the end of the proof of the law of quadratic reciprocity, it is necessary to have available rather extensive factorization tables if one is to evaluate Legendre symbols with large entries. Partly to obviate such a list, and partly for theoretical purposes, it has been found convenient to extend the definition of the Legendre symbol (a/p) so as to give meaning to (a/b) when b is not a prime. This is done in the following way: put $(a/1) = 1$, and if b is greater than 1 and odd, put

$$(a/b) = (a/p_1)(a/p_2) \cdots (a/p_r), \tag{2}$$

where $p_1 p_2 \cdots p_r$ is the prime factorization of b, and the symbols on the right in (2) are Legendre symbols. Then the symbol on the left in (2) is called a *Jacobi symbol;* like the Legendre symbol, it is undefined for even second entry. As we shall see, many more of its properties are similar to those of the Legendre symbol, but there is one

crucial point at which the analogy breaks down It may happen that $(a/b) = 1$ even when a is not a quadratic residue of b For it is clearly necessary that each of the Legendre symbols (a/p_i) have the value 1 in order for a to be a residue of b, while $(a/b) = 1$ if an even number of the factors in (2) are -1 while the remainder are $+1$ On the other hand, a is certainly not a quadratic residue of b if $(a/b) = -1$

The following theorem lists properties of the Jacobi symbol which were proved for the Legendre symbol in Theorems 5–3, 5–5, and 5–7, together with one (the second) which is peculiar to the extended function

THEOREM 5–10 *The Jacobi symbol has these properties*

(a) $(a_1a_2/b) = (a_1/b)(a_2/b)$

(b) $(a/b_1b_2) = (a/b_1)(a/b_2)$

(c) *If* $a_1 \equiv a_2 \pmod{b}$, *then* $(a_1/b) = (a_2/b)$

(d) $(-1/b) = (-1)^{(b-1)/2}$

(e) $(2/b) = (-1)^{\frac{1}{8}(b^2-1)}$

(f) *If* $(a, b) = 1$, *then* $(a/b)(b/a) = (-1)^{\frac{1}{4}(a-1)\,\frac{1}{4}(b-1)}$

Here the second entry in each symbol is a positive odd number

Proof (a) Put $b = p_1 \quad p_r$ Then

$$(a_1a_2/b) = (a_1a_2/p_1) \quad (a_1a_2/p_r),$$

and since these are Legendre symbols,

$$(a_1a_2/b) = (a_1/p_1) \quad (a_1/p_r)(a_2/p_1) \quad (a_2/p_r) = (a_1/b)(a_2/b)$$

(b) Put $b_1 = p_1 \quad p_r$ and $b_2 = p_1' \quad p_s'$ Then

$$\begin{aligned}(a/b_1b_2) &= (a/p_1 \quad p_rp_1' \quad p_s') \\ &= ((a/p_1) \quad (a/p_r))((a/p_1') \quad (a/p_s')) \\ &= (a/b_1)(a/b_2)\end{aligned}$$

(c) If $a_1 \equiv a_2 \pmod{b}$ and $b = p_1 \quad p_r$, then $a_1 \equiv a_2 \pmod{p_i}$ *for* $i = 1, \quad , r$ Hence $(a_1/p_i) = (a_2/p_i)$ and

$$(a_1/b) = (a_1/p_1) \quad (a_1/p_r) = (a_2/p_1) \quad (a_2/p_r) = (a_2/b)$$

(d) Put $b = p_1 \quad p_r$ Then

$$(-1/b) = \prod_{i=1}^{r} (-1/p_i) = \prod_{i=1}^{r} (-1)^{(p_i-1)/2}$$

or

$$(-1/b) = (-1)^{\sum_{i=1}^{r} \frac{p_i - 1}{2}} \tag{3}$$

But if m and n are odd, then

$$(m-1)(n-1) \equiv 0 \pmod 4,$$

$$mn - 1 \equiv m + n - 2 \pmod 4,$$

$$\frac{mn-1}{2} \equiv \frac{m-1}{2} + \frac{n-1}{2} \pmod 2.$$

Repeated application of this fact shows that

$$\sum_{i=1}^{r} \frac{p_i - 1}{2} \equiv \frac{p_1 \cdots p_r - 1}{2} \pmod 2,$$

so that $(-1/b) = (-1)^{(b-1)/2}$, by (3).

(e) The proof of this is the same as that just given, except that, using the fact that $m^2 \equiv 1 \pmod 8$ if m is odd, we deduce from the congruence

$$(m^2-1)(n^2-1) \equiv 0 \pmod{64}$$

that

$$\frac{m^2-1}{8} + \frac{n^2-1}{8} \equiv \frac{(mn)^2-1}{8} \pmod 2.$$

(f) Put $a = p_1 \cdots p_r$, $b = p_1' \cdots p_s'$. Then

$$\begin{aligned}
(a/b)(b/a) &= \prod_{i=1}^{s} (a/p_i') \prod_{j=1}^{r} (b/p_j) \\
&= \prod_{j=1}^{r} \prod_{i=1}^{s} (p_j/p_i') \cdot \prod_{j=1}^{r} \prod_{i=1}^{s} (p_i'/p_j) \\
&= \prod_{j=1}^{r} \prod_{i=1}^{s} (p_j/p_i')(p_i'/p_j) \\
&= (-1)^{\sum_{j=1}^{r} \sum_{i=1}^{s} \frac{p_j-1}{2} \cdot \frac{p_i'-1}{2}} = (-1)^{\sum_{j=1}^{r} \frac{p_j-1}{2} \cdot \sum_{i=1}^{s} \frac{p_i'-1}{2}} \\
&= (-1)^{\frac{1}{2}(a-1) \cdot \frac{1}{2}(b-1)}.
\end{aligned}$$

Because the laws of operation and combination are the same for the two types, Jacobi symbols can be used (and according to the same rules) in evaluating Legendre symbols, even though they do not give complete information about the quadratic character of a modulo b; all that is required is that one begin with a Legendre symbol. This means that the first entry in each symbol does not have to be factored,

except that powers of 2 must be removed Thus, using the numerical example considered earlier, we have

$$\begin{aligned}(2819/4177) &= (4177/2819) = (1358/2819) = (2/2819)(679/2819)\\ &= -(679/2819) = (2819/679) = (103/679)\\ &= -(679/103) = -(61/103) = -(103/61)\\ &= -(42/61) = -(2/61)(21/61) = (61/21)\\ &= (19/21) = (21/19) = (2/19) = -1,\end{aligned}$$

and we can again conclude that 2819 is a nonresidue of 4177

PROBLEM

Evaluate (751/919), both with and without the use of Jacobi symbols The entries are primes

REFERENCES

Section 5–5

The general theorem stated in the first paragraph is due to E Trost, *Nieuw Archief voor Wiskunde* (Amsterdam) **18**, 58–61 (1934) It has been generalized by H Flanders *Annals of Mathematics* **57**, 392–400 (1953)

CHAPTER 6

NUMBER-THEORETIC FUNCTIONS AND THE DISTRIBUTION OF PRIMES

6–1 Introduction. A number-theoretic function is any function which is defined for positive integral argument or arguments. Euler's φ-function is such, as are $n!$, n^2, e^n, etc. The functions which are interesting from the point of view of number theory are, of course, those like φ whose value depends in some way on the arithmetic nature of the argument, and not simply on its size. Two of the most interesting of such functions are $\tau(n)$, the number of positive divisors of n, and $\sigma(n)$, the sum of these divisors. These functions have been treated extensively in the literature, partly because of their simplicity and partly because they occur in a natural way in the investigation of many other problems. For this reason we shall pause briefly to demonstrate some of their fundamental properties. Recall that, as noted in Chapter 3, a number-theoretic function which is not identically zero is said to be multiplicative if $f(mn) = f(m)f(n)$ whenever $(m, n) = 1$.

THEOREM 6–1. *The functions σ and τ are multiplicative.*

Proof: Assume that $(m, n) = 1$. Then by the Unique Factorization Theorem, every divisor of mn can be represented uniquely as the product of a divisor of m and a divisor of n, and conversely, every such product is a divisor of mn. Clearly this implies that τ is multiplicative, and that

$$\sum_{d|m} d \cdot \sum_{d'|n} d' = \sum_{d''|mn} d'',$$

so that also $\sigma(m)\sigma(n) = \sigma(mn)$.

If f is any multiplicative function and the prime-power factorization of n is

$$n = \prod_{i=1}^{r} p_i^{\alpha_i},$$

then clearly

$$f(n) = f\left(\prod_{i=1}^{r} p_i^{\alpha_i}\right) = \prod_{i=1}^{r} f(p_i^{\alpha_i}),$$

and so the function is completely determined when its value is known for every prime-power argument In the cases at hand, we have

$$\tau(p^\alpha) = \alpha + 1$$

and $$\sigma(p^\alpha) = 1 + p + \quad + p^\alpha = \frac{p^{\alpha+1} - 1}{p - 1}$$

Thus we have proved

THEOREM 6–2 *If* $n = p_1^{\alpha_1} \quad p_r^{\alpha_r}$, *then*

$$\tau(n) = \prod_{i=1}^{r} (\alpha_i + 1) \qquad \text{and} \qquad \sigma(n) = \prod_{i=1}^{r} \frac{p_i^{\alpha_i+1} - 1}{p_i - 1}$$

There is another way of proving the multiplicativity of σ and τ which uses a basic property of all multiplicative functions

THEOREM 6–3 *If f is multiplicative and F is the function defined by the equation*

$$F(n) = \sum_{d|n} f(d),$$

then F is also multiplicative

Remark The multiplicativity of σ and τ follows immediately from the relations

$$\sigma(n) = \sum_{d\,n} d, \qquad \tau(n) = \sum_{d|n} 1,$$

since the functions f_1 and f_2 defined by the equations

$$f_1(n) = n \qquad \text{and} \qquad f_2(n) = 1 \qquad \text{for all } n$$

are obviously multiplicative

Proof Let $(m, n) = 1$ Then every divisor d of mn can be written uniquely as the product of a divisor d_1 of m and a divisor d_2 of n, and $(d_1, d_2) = 1$ Hence

$$F(rin) = \sum_{d|mn} f(d) = \sum_{\substack{d_1|m \\ d_2|n}} f(d_1 d_2) = \sum_{\substack{d_1\ m \\ d_2|n}} f(d_1)f(d_2)$$

$$= \sum_{d_1|m} f(d_1) \sum_{d_2|n} f(d_2) = F(m)F(n)$$

We shall see in the next section that the converse of Theorem 6–3 also holds

A problem that was of great interest to the Greeks was that of determining all the *perfect numbers*, that is, numbers such as 6 which

are equal to the sum of their proper divisors. In our notation this amounts to asking for all solutions of the equation

$$\sigma(n) = 2n.$$

It was known as early as Euclid's time that every number of the form

$$n = 2^{p-1}(2^p - 1),$$

in which both p and $2^p - 1$ are primes, is perfect. This is easy to verify:

$$\sigma(n) = \frac{2^p - 1}{2 - 1} \cdot \frac{(2^p - 1)^2 - 1}{(2^p - 1) - 1} = (2^p - 1) \cdot 2^p = 2n.$$

It happens that a partial converse also holds: every *even* perfect number is of the Euclid type. To see this we put $n = 2^{k-1} \cdot n'$, where $k \geq 2$. Then

$$\sigma(n) = \sigma(2^{k-1})\sigma(n') = (2^k - 1)\sigma(n'),$$

so that if n is perfect, it must be that

$$(2^k - 1)\sigma(n') = 2n = 2^k n'.$$

This implies that $(2^k - 1) | n'$, so we put $n' = (2^k - 1)n''$ and obtain

$$\sigma(n') = 2^k n''.$$

Since n' and n'' are divisors of n' whose sum is

$$n'' + (2^k - 1)n'' = 2^k n'' = \sigma(n'),$$

it must be that they are the only divisors of n', so that n' must be prime, and so $n'' = 1$, $n' = 2^k - 1$. Thus $n = 2^{k-1}(2^k - 1)$, where $2^k - 1$ is prime; this can happen only if k itself is prime.

There are two problems connected with perfect numbers which have not yet been solved. One is whether there are any odd perfect numbers; various necessary conditions are known for an odd number to be perfect, which show that any such number must be extremely large, but no conclusive results have been obtained. The other question is about the primes p for which $2^p - 1$ is prime. These *Mersenne primes* $2^p - 1$ are completely known for $p < 2300$ (the corresponding p's are 2, 3, 5, 7, 13, 17, 19, 31, 61, 89, 107, 127, 521, 607, 1279, 2203, 2281), but it is not known whether there are infinitely many such primes.

Aside from φ, σ, and τ, the function with which we shall be most concerned in this chapter is $\pi(x)$, already defined in Chapter 1 as the number of primes not exceeding x (We now drop the restriction that all variables are integer-valued) It was shown there that $\pi(x)$ increases indefinitely with x, that is, that *there are infinitely many primes* We now give another proof, which depends on the Unique Factorization Theorem

Assume that there are only k primes, say $p_1, \ldots, p_k$ By the Unique Factorization Theorem, every integer larger than 1 can be written uniquely as the product of a square-free number (that is, an integer which is the product of distinct primes) and a square But with only k primes at our disposal, there are only

$$\binom{k}{1} + \quad + \binom{k}{k-1} + 1 = 2^k - 1 < 2^k$$

square-free numbers, and there are not more than $\sqrt{n}$ perfect squares less than or equal to n This means that there are fewer than $2^k \sqrt{n}$ positive integers not exceeding n, which is obviously false if $n \geq 2^k\sqrt{n}$, that is if $\sqrt{n} \geq 2^k$ Actually, this argument proves a little more, namely that

$$2^{\tau(n)} > \sqrt{n}, \qquad \text{or} \qquad \pi(n) > \frac{\log n}{2 \log 2}$$

For later use in this chapter we now prove a general combinatorial theorem of very wide applicability (The product representation for the φ-function, for example, is a special case) The result is sometimes called the principle of cross-classification

THEOREM 6–4 *Let S be a set of N distinct elements, and let $S_1, \quad , S_r$ be arbitrary subsets of S containing $N_1 \quad , N_r$ elements, respectively For $1 \leq i < j < \quad < l \leq r$, let $S_{i,j,\ \ l}$ be the intersection of $S_i, S_j, \quad , S_l$, that is, the set of all elements of S common to $S_i\ S_j, \quad , S_l$, and let $N_{i,j,\ \ l}$ be the number of elements of $S_{i,j,\ \ l}$ Then the number of elements of S not in any of $S_1, \quad , S_r$ is*

$$K = N - \sum_{1 \leq i \leq r} N_i + \sum_{1 \leq i < j \leq r} N_{i,j} - \sum_{1 < i < j < k \leq r} N_{i,j,k} + \\ + (-1)^r N_{12 \quad r}$$

Remark To obtain the product formula for the φ-function, take S to be the set of integers $1, \quad , n$, and for $1 \leq k \leq r$, take S_k to be the set of elements of S which are divisible by p_k, where $n = p_1^{a_1} \quad p_r^{a_r}$

If $d|n$, the number of integers $s \leq n$ such that $d|s$ is n/d; hence

$$\varphi(n) = n - \sum_{1\leq i\leq r} \frac{n}{p_i} + \sum_{1\leq i<j\leq r} \frac{n}{p_i p_j} - \cdots = n \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

Proof: Let a certain element s of S belong to exactly m of the sets $S_1, \ldots, S_r$. If $m = 0$, s is counted just once, in N itself. If $0 < m \leq r$, then s is counted once, or $\binom{m}{0}$ times, in N, $\binom{m}{1}$ times in the terms N_i, $\binom{m}{2}$ times in the terms N_{ij}, etc. Hence the total contribution to K arising from the element s is

$$\binom{m}{0} - \binom{m}{1} + \binom{m}{2} - \cdots + (-1)^m \binom{m}{m} = (1-1)^m = 0.$$

PROBLEMS

1. Find an expression for $\sigma_k(n)$, the sum of the kth powers of the divisors of n.

2. Prove that $$\sum_{d|n} \tau^3(d) = \left(\sum_{d|n} \tau(d)\right)^2.$$

[*Hint:* Both sides are multiplicative functions, so it suffices to consider the case $n = p^\alpha$. Cf. Problem 2, Section 1-2.]

3. Show that, if $\sigma(n)$ is odd, then n is a square or the double of a square.

4. Show that the number of representations of an integer n as a sum of one or more consecutive positive integers is $\tau(n_1)$, where n_1 is the largest odd divisor of n. [*Hint:* If

$$n = (r+1) + (r+2) + \cdots + (r+s) = \sum_{k=1}^{r+s} k - \sum_{k=1}^{r} k = \tfrac{1}{2}s(s+2r+1),$$

then either s or $s + 2r + 1$ divides n_1.]

*5. Show that the number of ordered pairs of integers whose LCM is n is $\tau(n^2)$.

6. Show that

$$\sum_{n=1}^{\infty} \frac{\tau(n)}{n^s} = \left(\sum_{n=1}^{\infty} \frac{1}{n^s}\right)^2$$

if $s > 1$. [The series involved converge absolutely, and therefore can be rearranged in any order.]

*7. (a) Show that the sum of the odd divisors of n is

$$-\sum_{d|n} (-1)^{n/d} d.$$

[*Hint:* Let d_1 be an odd divisor of n, and find the total contribution to this sum from all divisors of n of the form $2^k d_1$.]

(b) Show that if n is even, then

$$\sum_{d|n} (-1)^{n/d} d = 2\sigma(n/2) - \sigma(n)$$

*8 Show that, if $d|n$ and $(n, r) = 1$, then the number of solutions (mod n) of

$$x \equiv r \pmod{d}, \qquad (x, n) = 1$$

is

$$\frac{\varphi(n)}{\varphi(d)} = \frac{n}{d} \prod_{\substack{p|n \\ p\nmid d}} \left(1 - \frac{1}{p}\right)$$

[*Hint* Take S of Theorem 6–4 to be the n/d numbers

$$x = r + td \qquad 1 \leq t \leq n/d$$

If $p|d$, then $p\nmid x$ Let the subsets consist of those elements of S divisible by the various primes which divide n but not d]

6–2 The Mobius function As we saw in Theorem 6–3, if f is any multiplicative function and F is its sum function, so that

$$F(n) = \sum_{d|n} f(d),$$

then F is also multiplicative We now ask whether the converse is true—whether the multiplicativity of F implies the multiplicativity of f To this end we attempt to express $f(n)$ as a sum, over the divisors of n of terms involving $F(d)$ Assuming that F is multiplicative, it is enough to consider $F(p^n)$, and if the converse in question is valid we can also restrict attention to $f(p^n)$ Since

$$f(p^n) = F(p^n) - F(p^{n-1}),$$

we can write

$$f(p^n) = \sum_{\alpha=0}^{n} \mu(p^{n-\alpha}) F(p^\alpha) = \sum_{d|p^n} \mu\left(\frac{p^n}{d}\right) F(d),$$

if we define the function μ in the following way

$$\mu(1) = 1,$$

$$\mu(p) = -1,$$

$$\mu(p^n) = 0 \qquad \text{for } n > 1$$

If we now require in addition that μ be multiplicative, then $\mu(n)$ is defined for all positive integral n, and it is easily seen that

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{if } n \text{ is divisible by a square larger than 1,} \\ (-1)^r & \text{if } n = p_1 \cdots p_r, \text{ where the } p_i \text{ are distinct primes} \end{cases}$$

This function μ is commonly called the *Möbius function*; it plays an important role in the theory of numbers. On the basis of the heuristic argument above, it is reasonable to conjecture that, for any n,

$$f(n) = \sum_{d|n} \mu\left(\frac{n}{d}\right) F(d),$$

and that from this formula one might be able to deduce the multiplicativity of f from that of F. We now substantiate these conjectures.

THEOREM 6–5. $$\sum_{d|n} \mu(d) = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{if } n > 1. \end{cases}$$

Proof: By Theorem 6–3, the function

$$M(n) = \sum_{d|n} \mu(d)$$

is multiplicative, and since

$$M(p^\alpha) = \begin{cases} 1, & \text{if } \alpha = 0, \\ 1 - 1 + 0 + \cdots + 0, & \text{if } \alpha \geq 1, \end{cases}$$

we see that $M(n) = 0$ if n is divisible by any prime, that is, if $n > 1$.

THEOREM 6–6 (*Möbius inversion formula*). *If f is any number-theoretic function (not necessarily multiplicative) and*

$$F(n) = \sum_{d|n} f(d),$$

then

$$f(n) = \sum_{d|n} F(d)\mu\left(\frac{n}{d}\right) = \sum_{d|n} F\left(\frac{n}{d}\right)\mu(d).$$

Proof: We have

$$\begin{aligned} \sum_{d|n} \mu(d) F\left(\frac{n}{d}\right) &= \sum_{d_1 d_2 = n} \mu(d_1) F(d_2) = \sum_{d_1 d_2 = n} \mu(d_1) \sum_{d|d_2} f(d) \\ &= \sum_{d_1 d|n} \mu(d_1) f(d) = \sum_{d|n} f(d) \sum_{d_1 \left| \frac{n}{d}\right.} \mu(d_1), \end{aligned}$$

and, by Theorem 6–5, the coefficient of $f(d)$ is zero unless $n/d = 1$ (that is, unless $d = n$), when it is 1, so that this last sum is equal to $f(n)$.

As an example of Theorem 6–6, we have

THEOREM 6–7. $$\varphi(n) = n \sum_{d|n} \frac{\mu(d)}{d}.$$

This follows immediately from Theorem 3–9

$$\sum_{d|n} \varphi(d) = n$$

It can also be obtained directly from the product representation of $\varphi(n)$

$$\varphi(n) = n \prod_{p|n} \left(1 - \frac{1}{p}\right) = n \left(1 + \sum_{\substack{p_1 \\ p_1 <}} \sum_{\substack{p_r | n \\ < p_r}} \frac{(-1)^r}{p_1 \quad p_r}\right) = n \sum_{d|n} \frac{\mu(d)}{d}$$

THEOREM 6–8 *If*

$$F(n) = \sum_{d\, n} f(d)$$

and F is multiplicative, so is f

Proof If $(m, n) = 1$, then

$$\begin{aligned} f(mn) &= \sum_{\substack{d_1|m \\ d_2|n}} F(d_1 d_2) \mu\left(\frac{mn}{d_1 d_2}\right) \\ &= \sum_{\substack{d_1|m \\ d_2\, n}} F(d_1) F(d_2) \mu\left(\frac{m}{d_1}\right) \mu\left(\frac{n}{d_2}\right) \\ &= \sum_{d\,|m} F(d_1) \mu\left(\frac{m}{d_1}\right) \sum_{d_2|n} F(d_2) \mu\left(\frac{n}{d_2}\right) = f(m) f(n) \end{aligned}$$

PROBLEMS

*1 Show that

$$\frac{1}{\varphi(n)} = \frac{1}{n} \sum_{d|n} \frac{\mu^2(d)}{\varphi(d)}$$

2 Show that

$$\sum_{d^2|n} \mu(d) = |\mu(n)|$$

3 Let f be any number theoretic function of two variables Show that if F is defined by the equation

$$F(m\ n) = \sum_{\substack{d\,|m \\ d_2|n}} f(d_1\ d_2),$$

then

$$f(m, n) = \sum_{\substack{d\,|m \\ d_2|n}} \mu(d_1) \mu(d_2) F\left(\frac{m}{d_1}, \frac{n}{d_2}\right)$$

*4. Let $J_k(n)$ be the number of ordered sets of k equal or distinct positive integers, none of which exceeds n and whose GCD is prime to n. Show, in the order indicated, that

(a) $\sum_{d|n} J_k(d) = n^k$,

(b) J_k is multiplicative,

(c) $J_k(n) = n^k \prod_{p|n} \left(1 - \frac{1}{p^k}\right)$.

5. Let

$$\Lambda(n) = \begin{cases} \log p & \text{if } n \text{ is a power of any prime } p, \\ 0 & \text{otherwise.} \end{cases}$$

Show that

$$\log n = \sum_{d|n} \Lambda(d),$$

and deduce that

$$\sum_{d|n} \mu(d) \log d = -\Lambda(n).$$

6. If ϑ is any multiplicative function, then the function ϑ' defined by the equation

$$\sum_{d|n} \vartheta(d)\vartheta'\left(\frac{n}{d}\right) = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{if } n > 1, \end{cases}$$

is also multiplicative. In this notation, find μ' and τ'.

7. If ϑ and ϑ' have the relation specified in Problem 6 and if

$$F(n) = \sum_{d|n} f(d)\vartheta\left(\frac{n}{d}\right),$$

then

$$f(n) = \sum_{d|n} F(d)\vartheta'\left(\frac{n}{d}\right).$$

8. Show that if f is multiplicative, then

$$\sum_{d|n} \mu(d)f(d) = \prod_{p|n} (1 - f(p)).$$

[*Hint:* Show that the function on the left is multiplicative.]

6-3 The function [x]. Another function which is of importance in number theory is the function $[x]$, introduced in the last chapter to represent the largest integer not exceeding x. In other words, for each real x, $[x]$ is the unique integer such that

$$x - 1 < [x] \leq x < [x] + 1.$$

For later purposes we list some of the properties of $[x]$

(a) $x = [x] + \vartheta$, where $0 \leq \vartheta < 1$ ϑ is called the *fractional part* of x

(b) $[x + n] = [x] + n$, if n is an integer

(c) $[x] + [-x] = \begin{cases} 0 & \text{if } x \text{ is an integer,} \\ -1 & \text{otherwise} \end{cases}$

(d) $[x_1] + [x_2] \leq [x_1 + x_2]$

(e) $[x/n] = [[x]/n]$ if n is a positive integer

(f) $0 \leq [x] - 2[x/2] \leq 1$ (Equivalently, $[x] - 2[x/2]$ assumes only the values 0 and 1)

(g) The number of integers m for which $x_1 < m \leq x_2$ is $[x_2] - [x_1]$

(h) The number of multiples of m which do not exceed x is $[x/m]$

(i) The least non negative residue of a, modulo m, is the number a' defined by the equation

$$a = m\left[\frac{a}{m}\right] + a'$$

These properties may easily be proved using the definition of $[x]$ and the first property above

Another quantity closely related to $[x]$ is the nearest integer to x, which is $[x + \frac{1}{2}]$ Sometimes the quantity $-[-x]$ is also useful, it is the smallest integer not less than x

In order to simplify the notation summation signs will sometimes be used with the real variable x as upper limit In these cases, it is understood that the summation variable takes values up to $[x]$, in other words,

$$\sum_{k=a}^{x} f(k) = \sum_{k=a}^{[x]} f(k)$$

The following relation between the greatest integer function and *the factorial function will be of importance later*

THEOREM 6-9 *If n is a positive integer, the exponent of the highest power of a prime p which divides $n!$ is*

$$\left[\frac{n}{p}\right] + \left[\frac{n}{p^2}\right] + \left[\frac{n}{p^3}\right] +$$

That is, if we set $$\sum_{k=1}^{\infty} \left[\frac{n}{p^k}\right] = E(p, n),$$

then $$p^{E(p,n)} || n!.$$

Remark: The sum has, of course, only finitely many nonzero terms.

Proof: The multiples of p from among the numbers $1, 2, \ldots, n$ are counted once each in $[n/p]$, those which are also multiples of p^2 are counted again in $[n/p^2]$, etc. Thus if $p^r || m$, the total contribution to the sum

$$\left[\frac{n}{p}\right] + \left[\frac{n}{p^2}\right] + \cdots$$

from the number m is exactly r, as it should be.

PROBLEMS

1. Carry out the proofs of the properties of $[x]$ listed in the text.

2. Prove that $[2x] + [2y] \geq [x] + [y] + [x + y]$, where x and y are arbitrary real numbers. [*Hint:* Consider separately the cases that neither, one, or both of $x - [x]$, $y - [y]$ are greater than $\frac{1}{2}$.]

3. Let $f(x, n)$ be the number of integers less than or equal to x and prime to n. Show that

(a) $\sum_{d|n} f\left(\frac{x}{d}, \frac{n}{d}\right) = [x]$. [Parallel the proof of Theorem 3–9.]

(b) $f(x, n) = \sum_{d|n} \mu(d) \cdot \left[\frac{x}{d}\right]$.

4. Let x be a number between 0 and 1. Let a_1 be the smallest positive integer such that

$$x_1 = x - \frac{1}{a_1} \geq 0,$$

let a_2 be the smallest positive integer such that

$$x_2 = x_1 - \frac{1}{a_2} \geq 0,$$

etc. Show that this leads to a finite expansion

$$x = \frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}$$

(that is, that $x_{n+1} = 0$ for some n) if and only if x is rational.

5 (a) Show that

$$\frac{(ab)!}{a!(b!)^a}$$

is an integer if a and b are positive integers [*Hint* Use induction on a]

(b) Show that

$$\frac{(2a)!(2b)!}{a!b!(a+b)!}$$

is an integer [*Hint* Use Problem 2]

6–4 The symbols "O", "o", and "$\sim$". If we construct tables of values of the common number-theoretic functions, we are immediately struck by how erratically they behave Thus $\tau(n)$ can be arbitrarily large, since for example $\tau(2^m) = m + 1$, and yet $\tau(n) = 2$ whenever n is prime Neither φ nor σ varies quite so wildly, in the sense that each of them definitely grows with n, but they are still far from monotonic It is one of the objects of this chapter to see what can be said about the size of these and other functional values simply in terms of the size of their arguments.

A very convenient notation has been introduced by Landau for use in this connection Let $g(x)$ be defined and positive for all positive x Then if $f(x)$ is any function defined on some unbounded set S of positive numbers (which in all applications here will be either the set of positive integers or the set of positive real numbers), and if there is a number M such that

$$\frac{|f(x)|}{g(x)} < M$$

for all sufficiently large $x \in S$, then we write $f(x) = O(g(x))$ (The symbol $\in$ means "is an element of ") If

$$\lim_{\substack{x\to\infty \\ x\in S}} \frac{f(x)}{g(x)} = 0,$$

we write $f(x) = o(g(x))$, and if

$$\lim_{\substack{x\to\infty \\ x\in S}} \frac{f(x)}{g(x)} = 1,$$

we write $f(x) \sim g(x)$, and say that $f(x)$ is asymptotically equal to $g(x)$. For example,

$$\sin x = O(x),$$
$$\sin x = o(x),$$
$$\sin x = O(1),$$
$$\varphi(n) = O(n),$$
$$\sqrt{x} = o(x),$$
$$x^k = o(e^x) \quad \text{for every constant } k,$$
$$\log^k x = o(x^\alpha) \quad \text{for every pair of constants } \alpha > 0 \text{ and } k,$$
$$[x] \sim x.$$

Here each of the second and third equations gives more information than the one preceding it; the first says that $\sin x$ does not grow any faster than x itself, the second that it does not grow as fast, and the third that $\sin x$ remains bounded as x increases. In the fourth equation, $O(n)$ could not be replaced by $o(n)$, since $\varphi(p) = p - 1 \sim p$.

The purpose of introducing these symbols is that, by their use, a complicated expression can be replaced by its principal or largest term, plus a remainder or error term whose possible size is indicated. Retaining an estimate for the error term is necessary because if several such expressions are combined, one has eventually to show that the sum of the error terms is still of smaller order of magnitude than the principal term. This in turn makes it necessary to combine terms involving "O" and "o". The following abbreviated rules apply:

(a) $O(O(g(x))) = O(g(x))$,
(b) $O(o(g(x))) = o(O(g(x))) = o(o(g(x))) = o(g(x))$,
(c) $O(g(x)) \pm O(g(x)) = O(g(x)) \pm o(g(x)) = O(g(x))$,
(d) $o(g(x)) \pm o(g(x)) = o(g(x))$,
(e) $\{O(g(x))\}^2 = O(g^2(x))$,
(f) $O(g(x)) \cdot o(g(x)) = \{o(g(x))\}^2 = o(g^2(x))$.

The meaning of the first statement, for example, is that if $f(x) = O(g(x))$ and $h(x) = O(f(x))$, then $h(x) = O(g(x))$; this follows from the fact that, if $0 < f(x) < M_1 g(x)$ and $|h(x)| < M_2 f(x)$, then $|h(x)| < M_1 M_2 g(x)$. The other assertions are equally straightforward; they need not be remembered explicitly, but are listed here to help orient the student, who should analyze all of them. Notice that suit-

able combinations of these rules give more general ones, for example, rules (a) and (c) show that

$$O(f(x)) \pm O(g(x)) = O(\max(f(x), g(x)))$$

A useful fact to remember is that the implication

$$f(x) = O(g(x)) \qquad \text{implies} \qquad h(f(x)) = O(h(g(x)))$$

does not hold in general, a sufficient condition is that $h(kx) = O(h(x))$ for every positive constant k, if $h(x), f(x) \to \infty$ as $x \to \infty$ Thus if $f(x)$ is larger than some positive constant for every $x > 0$, then $f(x) = O(g(x))$ implies that $\log f(x) = O(\log g(x))$, but it does not imply that

$$e^{f(x)} = O(e^{g(x)}),$$

since, for example, $\log x = O(\log \sqrt{x})$ but $x \neq O(\sqrt{x})$

The situation is quite different for the "o" symbol If $f(x) = o(g(x))$, then

$$e^{f(x)} = o(e^{g(x)})$$

if $f(x)$ increases indefinitely with x, but the relation $\log f(x) = o(\log g(x))$ may be false, e g , if $f(x) = \sqrt{x}$, $g(x) = x$

Another important point arises when we want to add together a set of error terms, the number $a(x)$ of such terms being an increasing function of x It is not true without restriction that

$$\sum_{k=1}^{a(x)} O(g_k(x)) = O\left(\sum_{k=1}^{a(x)} g_k(x)\right),$$

since, for example,

$$x = O(x), \qquad 2x = O(x), \qquad ,$$

but

$$\sum_{k=1}^{[x]} kx \neq O\left(\sum_{k=1}^{[x]} x\right)$$

What is needed here, of course, is that the constants implied in the symbols $O(g_k(x))$ all be bounded above by some number independent of k The corresponding principle for the "o" symbol is this if $f_k(x) = o(g_k(x))$, then we can write $f_k(x) = \epsilon_k(x) g_k(x)$, where $\epsilon_k(x) \to 0$ as $x \to \infty$, for fixed k, and if $\max(|\epsilon_1(x)|, \quad , |\epsilon_{a(x)}(x)|) \to 0$ as $x \to \infty$, then

$$\sum_{k=1}^{a(x)} f_k(x) = o\left(\sum_{k=1}^{a(x)} g_k(x)\right)$$

Turning now to the relation $f(x) \sim g(x)$, notice first that it is equivalent to the equation $f(x) = g(x) + o(g(x))$. Hence if $g(x) \to \infty$ as x increases indefinitely, the difference $f(x) - g(x)$ need not remain bounded; all that is asserted is that it is of smaller order of magnitude than $g(x)$ itself.

To give more precise information about $f(x)$, we must consider not $f(x)$ but $f(x) - g(x)$. As an example of this, consider the following theorem, which is not strictly a number-theoretic result, but which will be useful in what follows:

THEOREM 6-10. *There is a constant* $\gamma = 0.57721\ldots$ (*called Euler's constant*) *such that*

$$\sum_{k=1}^{n} \frac{1}{k} = \log n + \gamma + O\left(\frac{1}{n}\right). \tag{1}$$

Remark: The relation

$$\sum_{k=1}^{n} \frac{1}{k} - \log n \sim \gamma, \qquad \text{or} \qquad \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \log n\right) = \gamma, \tag{2}$$

is weaker than (1), since it says nothing about the error except that it approaches zero. Notice that (2) is not equivalent to

$$\sum_{k=1}^{n} \frac{1}{k} \sim \log n + \gamma \tag{3}$$

(that is, terms may not be "transposed" in an asymptotic relation), for (3) has no more content than the simpler relation

$$\sum_{k=1}^{n} \frac{1}{k} \sim \log n.$$

Proof: Put

$$\alpha_k = \log k - \log (k-1) - \frac{1}{k}, \qquad k = 2, 3, \ldots,$$

and put

$$\gamma_n = \sum_{k=1}^{n} \frac{1}{k} - \log n, \qquad n = 1, 2, \ldots,$$

so that

$$1 - \gamma_n = \sum_{k=2}^{n} \alpha_k, \qquad n = 2, 3, \ldots.$$

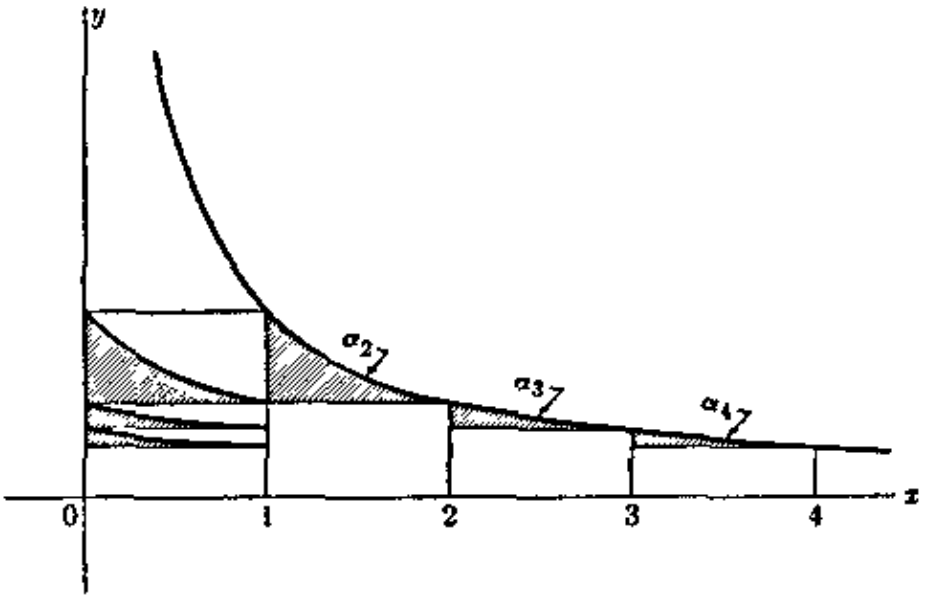

FIGURE 6–1

Geometrically, the number α_k represents the difference between the area of the region between the x-axis and the curve $y = 1/x$ in the interval $k - 1 \leq x \leq k$, and the area of the rectangle inscribed in this region, it is therefore positive The regions having areas α_2, α_3, and α_4 are shaded in Fig 6–1 If the regions having areas α_2 , α_n are translated parallel to the x-axis into the interval $0 \leq x \leq 1$, it becomes obvious that $0 < 1 - \gamma_n < 1$ and that $1 - \gamma_{n+1} > 1 - \gamma_n$, for $n = 1, 2,$ Since every bounded increasing sequence is convergent, we have that $\lim_{n\to\infty} (1 - \gamma_n)$ exists, we call the limit $1 - \gamma$ Referring again to the square $0 \leq x \leq 1$, $0 \leq y \leq 1$, we see that the region whose area is

$$\gamma_n - \gamma = (1 - \gamma) - (1 - \gamma_n) = \sum_{k=n+1}^{\infty} \alpha_k$$

is contained in the rectangle $0 \leq x \leq 1$, $0 \leq y \leq 1/n$, of area $1/n$, *so that*

$$\gamma - \gamma_n = O\left(\frac{1}{n}\right)$$

The proof is complete

If $f(x) \sim g(x)$ and $g(x) \to \infty$ as $x \to \infty$, then $\log f(x) \sim \log g(x)$ The relation $e^{f(x)} \sim e^{g(x)}$ is usually false, however, it is true only when $f(x) - g(x) = o(1)$ Finally, under the above suppositions,

together with that of the continuity of f and g, one may deduce that

$$\int_a^x f(t)dt \sim \int_a^x g(t)dt$$

for sufficiently large fixed a, by applying L'Hôpital's rule, but the corresponding relation $f'(x) \sim g'(x)$ is not always valid.

PROBLEMS

1. Carry out the proofs of all the unproved statements in this section.
2. Show that

$$\sum_{\substack{i,j=1 \\ i \neq j}}^{\infty} \left[\frac{x}{p_i p_j}\right] = x \sum_{\substack{p_i p_j \leq x \\ i \neq j}} \frac{1}{p_i p_j} + O(x),$$

where p_i is the ith prime.

3. Show that if $f(x)$ tends to zero monotonically as x increases without limit, and is continuous for $x > 0$, and if the series

$$\sum_{k=1}^{\infty} f(k)$$

diverges, then

$$\sum_{k=1}^{n} f(k) \sim \int_1^n f(x)dx.$$

What can be said if the infinite series converges?

4. It is known that for every n, the nth prime p_n is greater than $n \log n$. Use this to show that if B_n is defined by the equation

$$\sum_{i=1}^{n} \frac{1}{p_i} - \log \log n = B_n, \qquad n = 3, 4, \ldots,$$

then $B_3, B_4, \ldots$ is a decreasing sequence.

6-5 The sieve of Eratosthenes. We now turn to the study of $\pi(x)$, and shall obtain many of the classical elementary results concerning the distribution of primes. None of these estimates is the best of its kind that is known, but to obtain more accurate results would require either too long a discussion to be worth while or the use of tools not available here, as, for example, the theory of functions of a complex variable. For many purposes our results are quite as useful as the better estimates.

One method of estimating $\pi(x)$ is based upon the observation that if n is less than or equal to x and is not divisible by any prime less than or equal to $\sqrt{x}$, then it is prime Thus if we eliminate from the integers between 1 and x first all multiples of 2, then all multiples of 3, then all multiples of 5, etc , until all multiples of all primes less than or equal to $\sqrt{x}$ have been eliminated then the numbers remaining are prime This method of eliminating the composite numbers is known as the *sieve of Eratosthenes*, it has been adapted by Viggo Brun and others into a powerful method of estimating the number of integers in a certain interval having specified divisibility properties with respect to a certain set of primes.

We can modify the process just described by striking out the multiples of the first r primes $p_1, \quad , p_r$, retaining r as an independent variable until the best choice for it can be clearly seen If p_r is not the largest prime less than or equal to $\sqrt{x}$, but is some smaller prime, then of course it is no longer the case that all the integers remaining are primes, but certainly none of the primes except $p_1, \quad , p_r$ have been removed Thus if $A(x, r)$ is the number of integers remaining after all multiples of $p_1, \quad , p_r$ (including $p_1, \quad , p_r$ themselves, of course) have been removed from the integers less than or equal to x, then

$$\pi(x) \leq r + A(x, r)$$

In order to estimate $A(x, r)$ we use Theorem 6–4 If we take the $N = [x]$ objects to be the positive integers $\leq x$ and take S_k, for $1 \leq k \leq r$, to be the set of elements of S divisible by p_k then

$$N_1 = \left[\frac{x}{2}\right], \qquad N_2 = \left[\frac{x}{3}\right], \qquad , \qquad N_{12} = \left[\frac{x}{2\ 3}\right], \qquad ,$$

and so

$$A(x, r) = [x] - \sum_{i=1}^{r} \left[\frac{x}{p_i}\right] + \sum_{1 \leq i < j \leq r} \left[\frac{x}{p_i p_j}\right] - \sum_{1 \leq i < j < k \leq r} \left[\frac{x}{p_i p_j p_k}\right]$$
$$+ \quad + (-1)^r \left[\frac{x}{p_1 p_2 \quad p_r}\right]$$

The difference between this expression and

$$x - \sum_{1 \leq i \leq r} \frac{x}{p_i} + \sum_{1 \leq i < j \leq r} \frac{x}{p_i \ p_j} - \quad + (-1)^r \frac{x}{p_1 p_2 \quad p_r}$$

does not exceed

$$1 + \binom{r}{1} + \binom{r}{2} + \cdots + \binom{r}{r} = 2^r,$$

and consequently

$$\pi(x) \leq r + x \prod_{i=1}^{r} \left(1 - \frac{1}{p_i}\right) + 2^r.$$

We need an estimate for the product occurring here.

THEOREM 6–11. *If* $x \geq 2$, *then*

$$\prod_{p \leq x} \left(1 - \frac{1}{p}\right) < \frac{1}{\log x}.$$

Proof: We have

$$\prod_{p \leq x} \frac{1}{1 - 1/p} = \prod_{p \leq x} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots\right),$$

and, by the Unique Factorization Theorem, this is the sum of the reciprocals of all integers having only the primes not exceeding x as prime divisors. In particular, all integers less than or equal to x are of this form, and so

$$\prod_{p \leq x} \frac{1}{1 - 1/p} > \sum_{k=1}^{x} \frac{1}{k} > \int_1^{[x]+1} \frac{du}{u} > \log x,$$

and the theorem follows.

We can now prove

THEOREM 6–12.

$$\pi(x) = O\left(\frac{x}{\log \log x}\right).$$

Proof: As above,

$$\pi(x) \leq r + 2^r + x \cdot \prod_{i=1}^{r} \left(1 - \frac{1}{p_i}\right) \leq 2^{r+1} + x \prod_{i=1}^{r} \left(1 - \frac{1}{p_i}\right),$$

and by Theorem 6–11,

$$\pi(x) \leq 2^{r+1} + \frac{x}{\log p_r}.$$

But $p_r \geq r$, and so

$$\pi(x) < \frac{x}{\log r} + 2^{r+1}$$

Taking $r = [\log x]$, this becomes

$$\pi(x) < \frac{x}{\log \log x} + 2 \quad 2^{\log x} = \frac{x}{\log \log x} + O(x^{\log 2})$$

The last term is $O(x^{1-\epsilon})$ for some $\epsilon > 0$, and this is $o\left(\frac{x}{\log \log x}\right)$

Hence

$$\pi(x) = O\left(\frac{x}{\log \log x}\right) + o\left(\frac{x}{\log \log x}\right) = O\left(\frac{x}{\log \log x}\right).$$

*PROBLEM

Show by a sieve argument that the number of square-free integers not exceeding x is less than

$$x \prod \left(1 - \frac{1}{p^2}\right) + o(x)$$

6–6 Sums involving primes. Theorems 6–11 and 6–12 bear a rather peculiar relation to each other Theorem 6–11 was used in the proof of Theorem 6–12, yet the import of Theorem 6–11 is that the primes are not too infrequent, while that of Theorem 6–12 is that they are not too frequent For if, for example, the primes were so scarce that $p_n > cn^2$ for some positive number c, and for all n, the product

$$\prod_{p \leq x} \left(1 - \frac{1}{p}\right)$$

would be bounded away from zero as $x \to \infty$, which it is not It follows from Theorem 6–12, however, that there is no constant c' such that $p_n < c'n$ for all n The following theorem is, in its implications, analogous to Theorem 6–11

THEOREM 6–13 *The series* $\sum_p \frac{1}{p}$ *diverges*

Proof By Theorem 6–11,

$$\log \prod_{p \leq x} \left(1 - \frac{1}{p}\right) = \sum_{p \leq x} \log \left(1 - \frac{1}{p}\right) < - \log \log x$$

But since the curve $y = \log(1 + x)$ lies entirely above the curve $y = 2x$ in the interval $-\frac{1}{2} \leq x < 0$ (see Fig. 6-2), and since $p \geq 2$, we have

$$-\frac{2}{p} < \log\left(1 - \frac{1}{p}\right)$$

for all primes p, and so

$$\sum_{p \leq x} \frac{2}{p} > \log\log x.$$

In order to get more precise information about the behavior of the sum

$$\sum_{p \leq x} \frac{1}{p},$$

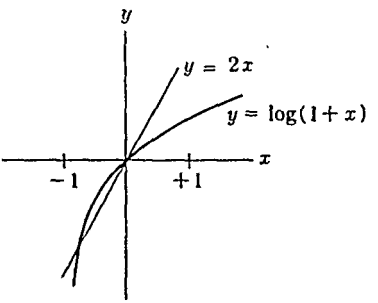

FIGURE 6-2

we proceed in a rather roundabout way, making use of the connection established in Theorem 6-9 between $n!$ and the primes not exceeding n.

THEOREM 6-14. $$\sum_{p \leq x} \frac{\log p}{p} \sim \log x.$$

Proof: By Theorem 6-9,

$$n! = \prod_{p \leq n} p^{[n/p] + [n/p^2] + \cdots},$$

and so

$$\log n! = \sum_{p \leq n} \left[\frac{n}{p}\right] \log p + \sum_{p \leq n} \left(\left[\frac{n}{p^2}\right] + \left[\frac{n}{p^3}\right] + \cdots\right) \log p.$$

Now

$$\sum_{p \leq n} \left[\frac{n}{p}\right] \log p \leq \sum_{p \leq n} \frac{n}{p} \log p,$$

and

$$\sum_{p \leq n} \left[\frac{n}{p}\right] \log p \geq \sum_{p \leq n} \left(\frac{n}{p} - 1\right) \log p = \sum_{p \leq n} \frac{n}{p} \log p - \sum_{p \leq n} \log p$$

$$\geq \sum_{p \leq n} \frac{n}{p} \log p - \log n \sum_{p \leq n} 1.$$

Moreover,

$$0 \leq \sum_{p \leq n} \left(\left[\frac{n}{p^2}\right] + \left[\frac{n}{p^3}\right] + \cdots\right) \log p \leq \sum_{p \leq n} \left(\frac{n}{p^2} + \frac{n}{p^3} + \cdots\right) \log p.$$

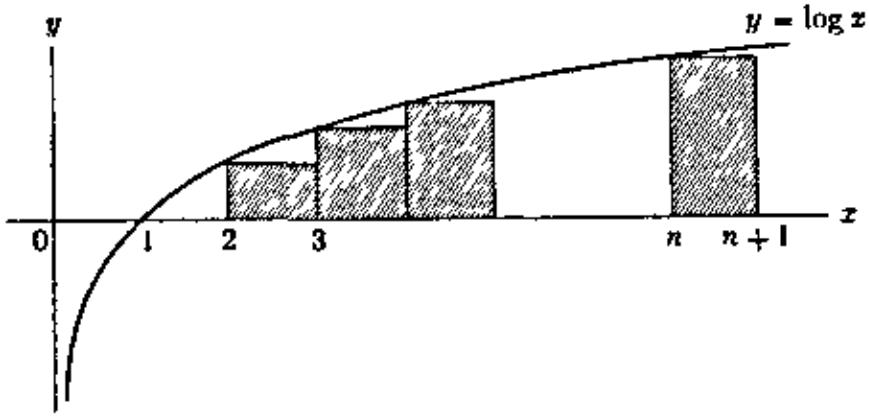

FIGURE 6–3

Thus

$$\log n! = \sum_{p \le n} \frac{n}{p} \log p + O(\pi(n) \log n)$$
$$+ O\left(n \sum_{p \le n} \left(\frac{1}{p^2} + \frac{1}{p^3} + \quad\right) \log p\right)$$
$$= n \sum_{p \le n} \frac{\log p}{p} + O(\pi(n) \log n) + O\left(n \sum_{p} \frac{\log p}{p(p-1)}\right),$$

and by Theorem 6–12, and the fact that the series

$$\sum_{k} \frac{\log k}{k(k-1)}$$

converges, this gives

$$\log n! = n \sum_{p \le n} \frac{\log p}{p} + O\left(\frac{n \log n}{\log \log n}\right) + O(n)$$
$$= n \sum_{p \le n} \frac{\log p}{p} + O\left(\frac{n \log n}{\log \log n}\right)$$

On the other hand, by comparing the area under the curve $y = \log x$ with the total area of the inscribed rectangles (see Fig 6–3), we see easily that

$$\log n! = \sum_{m=1}^{n} \log m = \int_{1}^{n} \log t \, dt + O(\log n) = n \log n + O(n)$$

Combining the two estimates obtained for $\log n!$, we have

$$n \sum_{p \le n} \frac{\log p}{p} + O\left(\frac{n \log n}{\log \log n}\right) = n \log n + O(n),$$

so that

$$\sum_{p\le n}\frac{\log p}{p}=\log n+O\left(\frac{\log n}{\log\log n}\right).$$

This proves the theorem when x is an integer n. But if $n\le x<n+1$, then

$$\sum_{p\le x}\frac{\log p}{p}=\sum_{p\le n}\frac{\log p}{p}=\log n+O\left(\frac{\log n}{\log\log n}\right)$$
$$=\log x+O\left(\frac{\log x}{\log\log x}\right). \tag{4}$$

THEOREM 6-15. *Suppose that $\lambda_1, \lambda_2, \ldots$ is a nondecreasing sequence with limit infinity, that $c_1, c_2, \ldots$ is an arbitrary sequence of real or complex numbers, and that $f(x)$ has a continuous derivative for $x\ge\lambda_1$. Put*

$$C(x)=\sum_{\lambda_n\le x}c_n,$$

where the summation is over all n for which $\lambda_n\le x$. Then for $x\ge\lambda_1$,

$$\sum_{\lambda_n\le x}c_nf(\lambda_n)=C(x)f(x)-\int_{\lambda_1}^{x}C(t)f'(t)\,dt.$$

Proof: We have

$$\sum_{\lambda_n\le x}c_nf(\lambda_n)=C(\lambda_1)f(\lambda_1)+(C(\lambda_2)-C(\lambda_1))f(\lambda_2)+\cdots$$
$$+(C(\lambda_\nu)-C(\lambda_{\nu-1}))f(\lambda_\nu),$$

where λ_ν is the greatest λ_n which does not exceed x. Regrouping the terms, we have

$$\sum_{\lambda_n\le x}c_nf(\lambda_n)=C(\lambda_1)(f(\lambda_1)-f(\lambda_2))+\cdots$$
$$+C(\lambda_{\nu-1})(f(\lambda_{\nu-1})-f(\lambda_\nu))$$
$$+C(\lambda_\nu)(f(\lambda_\nu)-f(x))+C(\lambda_\nu)f(x)$$
$$=-\int_{\lambda_1}^{x}C(t)f'(t)\,dt+C(x)f(x),$$

since $C(t)$ is a step function, constant over each of the intervals $(\lambda_{i-1}, \lambda_i)$ and over the interval (λ_ν, x).

THEOREM 6-16. $$\sum_{p\le x}\frac{1}{p}\sim\log\log x.$$

Proof. Take $\lambda_n = p_n$, $c_n = \log p_n/p_n$, $f(t) = 1/\log t$ in Theorem 6–15 By (4),

$$\begin{aligned}
\sum_{p\le x}\frac{1}{p} &= \frac{1}{2} + \sum_{2<p\le x}\left(\frac{\log p}{p}\cdot\frac{1}{\log p}\right)\\
&= \frac{1}{\log x}\sum_{2<p<x}\frac{\log p}{p} - \int_3^x\left(\sum_{p\le t}\frac{\log p}{p}\right)\frac{-dt}{t\log^2 t} + \frac{1}{2}\\
&= \frac{1}{\log x}\left\{\log x + O\left(\frac{\log x}{\log\log x}\right)\right\}\\
&\qquad + \int_3^x\left\{\log t + O\left(\frac{\log t}{\log\log t}\right)\right\}\frac{dt}{t\log^2 t} + \frac{1}{2}\\
&= O(1) + \int_3^x\frac{dt}{t\log t} + \int_3^x O\left(\frac{1}{t\log t\log\log t}\right)dt\\
&= \log\log x + O(1) + \int_3^x O\left(\frac{1}{t\log t\log\log t}\right)dt
\end{aligned}$$

Now for some constant M,

$$\begin{aligned}
\left|\int_3^x O\left(\frac{1}{t\log t\log\log t}\right)dt\right| &< M\int_3^x\frac{dt}{t\log t\log\log t}\\
&= O(\log\log\log x),
\end{aligned}$$

so that $$\sum_{p<x}\frac{1}{p} = \log\log x + O(\log\log\log x),$$

which proves the theorem

*PROBLEM

Use Theorems 6–15 and 6–16 to show that

$$\int_2^x\frac{\pi(t)}{t^2}\,dt = \sum_{p\le x}\frac{1}{p} + o(1) \sim \log\log x,$$

and deduce that for no positive constant δ is there a $T = T(\delta)$ such that for all $t > T$,

$$\pi(t) > (1+\delta)\frac{t}{\log t},$$

and that for no $\delta > 0$ is there a $T = T(\delta)$ such that for all $t > T$,

$$\pi(t) < (1-\delta)\frac{t}{\log t}$$

This implies that, if

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\log x}$$

exists, it must be 1.

6-7 The order of $\pi(x)$. We now show that the actual order of $\pi(x)$ is $x/\log x$.

THEOREM 6-17. *There are positive finite constants c_1 and c_2 such that for $x \geq 2$,*

$$c_1 \frac{x}{\log x} < \pi(x) < c_2 \frac{x}{\log x}.$$

Proof: Take $n \geq 2$. Corresponding to each $p \leq 2n$ there is a unique integer r_p such that $p^{r_p} \leq 2n < p^{r_p+1}$. We first prove that

$$\prod_{n<p\leq 2n} p \,\Big|\, \frac{(2n)!}{n!n!} \quad \text{and} \quad \frac{(2n)!}{n!n!} \,\Big|\, \prod_{p\leq 2n} p^{r_p}. \tag{5}$$

The first part is obvious, since any prime between n and $2n$ occurs as a factor of $(2n)!$ but does not occur in the denominator $(n!)^2$. For the second part, we have that the highest power of p which divides the numerator $(2n)!$, by Theorem 6-9, has exponent

$$\sum_{m=1}^{r_p} \left[\frac{2n}{p^m}\right],$$

while the highest power of p which divides the denominator has exponent

$$2\sum_{m=1}^{r_p} \left[\frac{n}{p^m}\right],$$

so that the highest power of p dividing $\binom{2n}{n}$ has exponent

$$\sum_{m=1}^{r_p} \left\{\left[\frac{2n}{p^m}\right] - 2\left[\frac{n}{p^m}\right]\right\} \leq \sum_{m=1}^{r_p} 1 = r_p.$$

Here we have used property (f) of $[x]$, from the list in Section 6-3. From (5) we get

$$n^{\pi(2n)-\pi(n)} \leq \prod_{n<p\leq 2n} p \leq \binom{2n}{n} \leq \prod_{p\leq 2n} p^{r_p} \leq (2n)^{\pi(2n)},$$

whence $(\pi(2n) - \pi(n)) \log n \leq \log \binom{2n}{n} \leq \pi(2n) \log 2n$

Clearly $\binom{2n}{n} \leq 2^{2n}$, and also

$$\binom{2n}{n} = \frac{(n+1)\cdots(2n)}{1\cdots n} = \prod_{a=1}^{n} \frac{n+a}{a} \geq \prod_{a=1}^{n} 2 = 2^n,$$

thus $$(\pi(2n) - \pi(n)) \log n \leq 2n \log 2,$$

or* $$\pi(2n) - \pi(n) \leq c_3 \frac{n}{\log n}, \tag{6}$$

and $$\pi(2n) \log 2n \geq n \log 2,$$

or $$\pi(2n) > c_4 \frac{n}{\log n} \tag{7}$$

If $x \geq 4$, we get from (7) that

$$\pi(x) \geq \pi\left(2\left[\frac{x}{2}\right]\right) > c_4 \frac{[x/2]}{\log [x/2]} > c_5 \frac{x}{\log x},$$

and since $\pi(x) \geq 1$ for $2 \leq x < 4$, $\pi(x) > c_1(x/\log x)$ for $x \geq 2$

If $y \geq 4$, we get from (6) that

$$\pi(y) - \pi\left(\frac{y}{2}\right) = \pi(y) - \pi\left(\left[\frac{y}{2}\right]\right) \leq 1 + \pi\left(2\left[\frac{y}{2}\right]\right) - \pi\left(\left[\frac{y}{2}\right]\right)$$

$$< 1 + c_3 \frac{[y/2]}{\log [y/2]} < c_6 \frac{y}{\log y},$$

and so for $y \geq 2$, $$\pi(y) - \pi\left(\frac{y}{2}\right) < c_7 \frac{y}{\log y}$$

Using the trivial bound $\pi(y/2) \leq y/2$, we get

$$\pi(y) \log y - \pi\left(\frac{y}{2}\right) \log \frac{y}{2} = \left\{\pi(y) - \pi\left(\frac{y}{2}\right)\right\} \log y + \pi\left(\frac{y}{2}\right) \log 2$$

$$< \log y \quad c_7 \frac{y}{\log y} + \frac{y}{2} < c_8 y$$

*Here c_3, c_4, will denote certain positive constants, whose exact values will be of no concern

If we put $y = x/2^m$ with $2^m \leq x/2$ and $m \geq 0$, this becomes

$$\pi\left(\frac{x}{2^m}\right)\log\frac{x}{2^m} - \pi\left(\frac{x}{2^{m+1}}\right)\log\frac{x}{2^{m+1}} < c_8\frac{x}{2^m},$$

and summing over all such m's, we have

$$\pi(x)\log x - \pi\left(\frac{x}{2^{\mu+1}}\right)\log\frac{x}{2^{\mu+1}} < c_2 x,$$

where $2^\mu \leq x/2 < 2^{\mu+1}$. But $x/2^{\mu+1} < 2$, so that $\pi(x/2^{\mu+1}) = 0$, and we have

$$\pi(x)\log x < c_2 x,$$

which completes the proof.

THEOREM 6–18. *There are positive constants c_9, c_{10} such that for $r > 1$,*

$$c_9 r\log r < p_r < c_{10} r\log r.$$

Proof: Taking x to be p_r in Theorem 6–17, we get

$$c_1\frac{p_r}{\log p_r} < r < c_2\frac{p_r}{\log p_r}.$$

The right-hand inequality gives immediately

$$p_r > c_9 r\log p_r > c_9 r\log r.$$

Using the other inequality and the fact that $\log u = o(\sqrt{u})$, we have that for $r > c_{11}$,

$$\frac{\log p_r}{\sqrt{p_r}} < c_1 < \frac{r\log p_r}{p_r},$$

$$p_r < r^2,$$

$$\log p_r < 2\log r,$$

and so for $r > c_{11}$,

$$p_r < \frac{1}{c_1} r\cdot 2\log r,$$

whence finally $p_r < c_{10} r\log r$ for all $r > 1$.

We can use Theorem 6–17 to improve Theorems 6–14 and 6–16. Examining the proof of Theorem 6–14, we see that the error term can

now be reduced to $O(n)$, since by Theorem 6–17, $\pi(n) \log n = O(n)$ Thus we have

THEOREM 6–19 $$\sum_{p \le x} \frac{\log p}{p} = \log x + O(1)$$

Following the argument used for Theorem 6–14, we now have

$$\sum_{p \le x} \frac{1}{p} = \frac{1}{\log x}(\log x + O(1)) + \int_2^x \log t \, \frac{dt}{t \log^2 t} + \int_2^x \left(\sum_{p \le t} \frac{\log p}{p} - \log t \right) \frac{dt}{t \log^2 t}$$

$$= 1 + O\left(\frac{1}{\log x}\right) + \log\log x - \log\log 2 + \int_2^\infty \left(\sum_{p \le t} \frac{\log p}{p} - \log t \right) \frac{dt}{t \log^2 t} - \int_x^\infty \frac{O(1)dt}{t \log^2 t}.$$

Here the first integral is convergent, and the second is clearly $O(1/\log x)$ This proves

THEOREM 6–20 *There is a constant C such that*

$$\sum_{p \le x} \frac{1}{p} = \log\log x + C + O\left(\frac{1}{\log x}\right).$$

PROBLEM

Apply Theorem 6–20 to show that for some constant B,

$$\sum_{p \le x} \log\left(1 - \frac{1}{p}\right) = -\log\log x - B + O\left(\frac{1}{\log x}\right)$$

Deduce that

$$\prod_{p \le x} \left(1 - \frac{1}{p}\right) = \frac{e^{-B}}{\log x} + O\left(\frac{1}{\log^2 x}\right).$$

By Theorem 6–11 $B \ge 0$, although we do not prove it, B is Euler's constant [Use the law of the mean to show that if $f(x) \to 0$ as $x \to \infty$, then $e^{f(x)} = 1 + O(f(x))$]

6–8 Bertrand's conjecture. In 1845 J Bertrand showed empirically that there is a prime between n and $2n$ for all n greater than 1 and less than six million, and predicted that this is true for all positive integral n Chebyshev proved this in 1850, and indeed that for every

$\epsilon > \frac{1}{5}$ there is a ξ such that for every $x > \xi$ there is a prime between x and $(1 + \epsilon)x$. Since that time, analytic methods have been used to show that this last theorem is true for every $\epsilon > 0$. We shall content ourselves here with a proof of Bertrand's original conjecture. The proof given is due to P. Erdös.

It is worth noting that Theorem 6–19 implies a weak form of the theorem.

THEOREM 6–21. *There exists a positive constant c_{12} such that there is a prime between n and $c_{12}n$ for all n.*

Proof: By Theorem 6–19, there is a constant A such that

$$\log n - A < \sum_{p \le n} \frac{\log p}{p} < \log n + A$$

for all n. Suppose that there is no prime between n and ne^{2A}. Then

$$\sum_{p \le n} \frac{\log p}{p} = \sum_{p \le ne^{2A}} \frac{\log p}{p},$$

and so by Theorem 6–19 again,

$$\sum_{p \le n} \frac{\log p}{p} > \log (ne^{2A}) - A = \log n + A.$$

With this contradiction, the theorem is proved with $c_{12} = e^A$.

For the proof of the more exact theorem we need two lemmas.

THEOREM 6–22. $\displaystyle\prod_{p \le n} p < 4^n.$

Proof: We use induction on n. The theorem is obvious if $n = 1$ or 2. Suppose it is true for $1, 2, \ldots, n - 1$, where $n \ge 3$. Then we can restrict attention to odd n, since otherwise

$$\prod_{p \le n} p = \prod_{p \le n-1} p < 4^{n-1} < 4^n,$$

so we can put $n = 2m + 1$. From its definition, the binomial coefficient

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

is divisible by every prime p with $m + 2 \le p \le 2m + 1$. Hence

$$\prod_{p \le 2m+1} p \le \binom{2m+1}{m} \cdot \prod_{p \le m+1} p < \binom{2m+1}{m} 4^{m+1}.$$

But the numbers

$$\binom{2m+1}{m} \quad \text{and} \quad \binom{2m+1}{m+1}$$

are equal, and both occur in the expansion of $(1+1)^{2m+1}$, so that

$$\binom{2m+1}{m} \leq \frac{1}{2}\ 2^{2m+1} = 4^m,$$

and so

$$\prod_{p \leq 2m+1} p < 4^m\ \ 4^{m+1} = 4^{2m+1}$$

The theorem follows by induction on n

THEOREM 6–23 *If $n \geq 3$ and $\frac{2}{3}n < p \leq n$, then $p \nmid \binom{2n}{n}$*

Proof The restrictions on n and p are such that

(a) p is greater than 2,

(b) p and $2p$ are the only multiples of p which are less than or equal to $2n$ since $3p$ is greater than $2n$,

(c) p itself is the only multiple of p which is less than or equal to n

From (a) and (b), $p^2 \| (2n)!$, and from (c), $p^2 \| (n!)^2$, so that $p \nmid (2n)!/(n!)^2$

THEOREM 6–24 *For any positive integer n there is a prime p such that $n < p \leq 2n$*

Proof This is true for $n = 1, 2, 3$ Assume the theorem false for a certain integer $n \geq 4$ Then by Theorem 6–23, every prime which divides $\binom{2n}{n}$ must be less than or equal to $2n/3$ Let p be such a prime, and suppose that $p^\alpha \| \binom{2n}{n}$ Then by the proof of Theorem 6–17, since

$$\binom{2n}{n} \Bigg| \prod_{p \leq 2n} p^{r_p}, \qquad (p^{r_p} \leq 2n < p^{r_p+1})$$

it follows that $p^\alpha \leq 2n$ Thus if $\alpha \geq 2$ then $p \leq \sqrt{2n}$, and so there are at most $[\sqrt{2n}]$ primes appearing in the prime power factorization of $\binom{2n}{n}$ with exponent larger than 1 Hence

$$\binom{2n}{n} \leq (2n)^{[\sqrt{2n}]} \prod_{p \leq 2n/3} p$$

But $\binom{2n}{n}$ is the largest of the $2n+1$ terms in the expansion of $(1+1)^{2n}$, so that

$$4^n < (2n+1)\binom{2n}{n},$$

and so

$$\frac{4^n}{2n+1} < (2n)^{\sqrt{2n}} \cdot \prod_{p \le 2n/3} p.$$

By Theorem 6–22, this implies that

$$\frac{4^n}{2n+1} < (2n)^{\sqrt{2n}} \cdot 4^{2n/3},$$

and since $2n+1 < 4n^2$, this gives

$$4^n < (2n)^{\sqrt{2n}+2} \cdot 4^{2n/3}, \qquad \text{or} \qquad 4^{n/3} < (2n)^{\sqrt{2n}+2}.$$

Taking logarithms, we have

$$\frac{n \log 4}{3} < (\sqrt{2n}+2) \log 2n.$$

This inequality is false for $n > 512$, so the theorem is true for $n > 512$. But in the sequence of primes

$$2, 3, 5, 7, 13, 23, 43, 83, 163, 317, 557,$$

each number is smaller than twice the one preceding it, and the theorem is also true for all $n \le 512$. It is therefore true for all n.

PROBLEMS

1. It follows from the Problem of Section 6–6 that in Theorem 6–17, $c_1 < 1 < c_2$. If estimates had been made of c_1 and c_2 in the proof of Theorem 6–17 (which would be simple to do), we would know, as a consequence, two particular constants c_1 and c_2 for which the inequality of Theorem 6–13 holds. Suppose that this is the case, and that $c_2/c_1 = \beta > 1$. Show that if $\epsilon > 0$, then

$$\frac{\pi((1+\epsilon)x) - \pi(x)}{x/\log x} > c_1(1+\epsilon-\beta) + O\left(\frac{1}{\log x}\right).$$

Deduce that if $\epsilon > \beta - 1$, the number of primes between x and $(1+\epsilon)x$ tends to infinity with x.

2 Show that there is a constant $A > 0$ such that

$$\sum_{x<p\leq x^2} \frac{1}{p} > A$$

for all sufficiently large x Deduce that for each $\epsilon > 0$ there are infinitely many pairs p_n and p_{n+1} of consecutive primes such that

$$p_{n+1} < (1 + \epsilon)p_n$$

6–9 The order of magnitude of φ, σ, and τ. The quantity $\pi(x)$ is reasonably well-behaved, and so one can make fairly precise statements about its size as a function of x This is not true of the other functions we have considered, which vary much too wildly to permit asymptotic approximations There are, however, various weaker statements which can be made about their size which still yield considerable information

Consider, for example, the quantity $\tau(n)$ A moment's thought shows that the number of divisors of n is much smaller than n itself, for large n, it is to be expected that $\tau(n) = o(n)$ And while $\tau(n) = 2$ infinitely many times it is also possible to make $\tau(n)$ arbitrarily large for suitable n Thus if the points $(n, \tau(n))$ are plotted in a coordinate system, as in Fig 6–4, there is a unique "lowest" polygonal path extending upward and to the right from (1, 1) which is concave downward and is such that every point $(n, \tau(n))$ lies on or below it Suppose that this path is described by the equation $y = T(x)$ While we shall not obtain an asymptotic estimate for $T(x)$, the following theorem shows that it increases more rapidly than any power of $\log x$, and less rapidly than any positive power of x

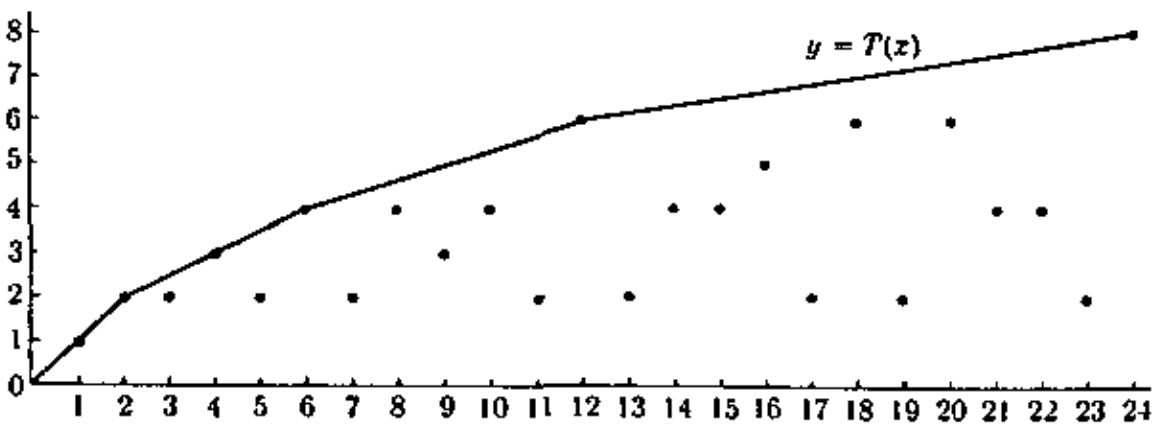

FIGURE 6–4

THEOREM 6–25. (a) *The relation* $\tau(n) = O(\log^h n)$ *is false for every constant* h.

(b) *The relation* $\tau(n) = O(n^\delta)$ *is true for every fixed* $\delta > 0$.

Proof: (a) Let n be any of the numbers $(2 \cdot 3 \cdots p_r)^m$, $m = 1, 2, \ldots$; here r is arbitrary but fixed. Then

$$\tau(n) = \prod_1^r (m+1) = (m+1)^r > m^r.$$

But $m = \dfrac{\log n}{\log (2 \cdot 3 \cdots p_r)}$, so that

$$\tau(n) > \frac{\log^r n}{(\log (2 \cdot 3 \cdots p_r))^r} > c_{13} \log^r n,$$

where $c_{13} > 0$ is a constant depending only on r, and not on n.

(b) Let $$f(n) = \frac{\tau(n)}{n^\delta};$$

then f is multiplicative. But

$$f(p^m) = \frac{m+1}{p^{m\delta}} \leq \frac{2m}{p^{m\delta}} = \frac{2}{\log p} \frac{\log p^m}{p^{m\delta}} \leq \frac{2}{\log 2} \cdot \frac{\log p^m}{p^{m\delta}},$$

so that $f(p^m) \to 0$ as $p^m \to \infty$, i.e., as either p or m, or both, increases. This clearly implies that $f(n) \to 0$ as $n \to \infty$, which is the assertion.

Alternative proof of (b): Let δ be positive, and let

$$n = \prod_1^r p_i^{\alpha_i}.$$

Then $$\frac{\tau(n)}{n^\delta} = \frac{\alpha_1 + 1}{p_1^{\alpha_1 \delta}} \cdots \frac{\alpha_r + 1}{p_r^{\alpha_r \delta}} \leq \prod_{p_i | n} \max_{x \geq 0} \left(\frac{x+1}{p_i^{\delta x}} \right).$$

For fixed δ, the quantity

$$\max_{x \geq 0} \frac{x+1}{p^{\delta x}}$$

is equal to 1 for sufficiently large p, and is never smaller than 1. Hence

$$\frac{\tau(n)}{n^\delta} \leq \prod_p \max_{x \geq 0} \left(\frac{x+1}{p^{\delta x}} \right) = c_\delta,$$

and c_δ is a finite constant independent of n. Hence $\tau(n) = O(n^\delta)$.

By actually evaluating the constant c_3, one obtains inequalities such as

$$\tau(n) \leq \sqrt{3n}, \qquad \tau(n) < 4\sqrt[3]{n},$$

which of course are very poor estimates for large n, but which are sometimes more useful than the statement of Theorem 6-25, where nothing is said about the behavior of $\tau(n)$ for all n, but only for large n.

As regards the φ-function, we have the trivial upper bound $\varphi(n) \leq n - 1$ for $n > 1$, equality being attained whenever n is prime. The corresponding lower bound is less obvious.

THEOREM 6-26 *There is a positive constant c_{14} such that for all $n > 3$,*

$$\varphi(n) > \frac{c_{14} n}{\log \log n}$$

Proof We have

$$\frac{\varphi(n)}{n} = \prod_{p|n} \left(1 - \frac{1}{p}\right),$$

so that

$$\log \frac{\varphi(n)}{n} = \sum_{p|n} \log\left(1 - \frac{1}{p}\right) = -\sum_{p|n} \frac{1}{p} + \sum_{p|n} \left\{\log\left(1 - \frac{1}{p}\right) + \frac{1}{p}\right\}$$

$$> -\sum_{p|n} \frac{1}{p} - c_{15},$$

since

$$\sum_{p} \left\{\log\left(1 - \frac{1}{p}\right) + \frac{1}{p}\right\} > \sum_{p} \left(\frac{1}{p} - \frac{1}{p-1}\right)$$

$$> -\sum_{n=2}^{\infty} \left(\frac{1}{n-1} - \frac{1}{n}\right) = -1$$

Now let $p_1, \quad , p_{r-\rho}$ be the primes less than $\log n$ which divide n so that

$$\sum_{p|n} \frac{1}{p} = \sum_{k=1}^{r-\rho} \frac{1}{p_k} + \sum_{k=r-\rho+1}^{r} \frac{1}{p_k} = S_1 + S_2$$

Then

$$\log^{\rho} n \leq p^{\rho}_{r-\rho+1} \leq \prod_{k=r-\rho+1}^{r} p_k \leq n,$$

so that
$$\rho \leq \frac{\log n}{\log \log n},$$

and
$$S_2 \leq \frac{1}{\log n} \cdot \frac{\log n}{\log \log n} < c_{16}.$$

By Theorem 6–20,

$$S_1 < \log \log p_{r-\rho} + c_{17} < \log \log \log n + c_{17}.$$

Combining these results, we get

$$\log \frac{\varphi(n)}{n} > -\log \log \log n - c_{18},$$

and so

$$\frac{\varphi(n)}{n} > \frac{c_{14}}{\log \log n}.$$

We can use Theorem 6–26 to obtain a corresponding upper bound for $\sigma(n)$, with the help of the following theorem.

THEOREM 6–27. *There is a positive constant* c_{19} *such that*

$$c_{19} < \frac{\sigma(n)\varphi(n)}{n^2} < 1.$$

Proof: If $n = \prod p^{\alpha}$, then

$$\begin{aligned}\sigma(n)\varphi(n) &= \prod_{p|n} \left(\frac{p^{\alpha+1} - 1}{p - 1}\right) n \prod_{p|n} \left(1 - \frac{1}{p}\right) \\ &= n \prod_{p|n} \frac{1 - p^{-(\alpha+1)}}{1 - 1/p} \cdot n \prod_{p|n} \left(1 - \frac{1}{p}\right) \\ &= n^2 \prod_{p|n} (1 - p^{-(\alpha+1)}).\end{aligned}$$

Here the coefficient of n^2 is clearly less than 1 and greater than or equal to

$$\prod_{p|n} \left(1 - \frac{1}{p^2}\right) > \prod_{k=2}^{n} \left(1 - \frac{1}{k^2}\right).$$

Now

$$\log \prod_{k=2}^{n} \left(1 - \frac{1}{k^2}\right) = \sum_{k=2}^{n} \log \left(1 - \frac{1}{k^2}\right),$$

and for $x > 0$, $$\log (1 - x) > \frac{-x}{1 - x},$$

so that $$-\sum_{k=2}^{n} \frac{1}{k^2 - 1} < \sum_{k=2}^{n} \log\left(1 - \frac{1}{k^2}\right).$$

Since the first sum in this inequality tends to a limit as $n \to \infty$, it follows that the above coefficient of n^2 is bounded away from zero, and the theorem follows

THEOREM 6-28 $\sigma(n) = O(n \log \log n)$

Proof By Theorem 6–26,

$$\frac{\varphi(n)}{n} > \frac{c_{14}}{\log \log n},$$

and by Theorem 6–27,

$$\frac{\sigma(n)}{n} < \frac{n}{\varphi(n)} < \frac{\log \log n}{c_{14}},$$

$$\frac{\sigma(n)}{n} = O(\log \log n),$$

$$\sigma(n) = O(n \log \log n)$$

PROBLEM

Show that there is an infinite sequence of positive integers $n_1, n_2,$ such that

$$\varphi(n_k) < \frac{cn_k}{\log \log n_k}, \qquad k = 1, 2, \quad ,$$

for some constant c

6–10 Average order of magnitude. Another way of describing the behavior of a number-theoretic function is in terms of its average order, that is, in terms of the quantity

$$\frac{1}{n} \sum_{m=1}^{n} f(m)$$

Summing the values of a function has the effect of smoothing out its irregularities, so that it is frequently possible to make quite precise statements about the size of the sum

THEOREM 6-29. *If*

$$F(n) = \sum_{d|n} f(d),$$

then

$$\sum_{m=1}^{n} F(m) = \sum_{m=1}^{n} \left[\frac{n}{m}\right] f(m).$$

Proof:

$$\sum_{m=1}^{n} F(m) = \sum_{m=1}^{n} \sum_{d|m} f(d) = \sum_{d=1}^{n} \sum_{k=1}^{[n/d]} f(d) = \sum_{d=1}^{n} \left[\frac{n}{d}\right] f(d).$$

THEOREM 6-30. *If*

$$F(n) = \sum_{d|n} f(d),$$

then

$$\sum_{m=1}^{n} F(m) = \sum_{m=1}^{t} \left[\frac{n}{m}\right] f(m) + \sum_{m=1}^{n/t} G\left(\frac{n}{m}\right) - \left[\frac{n}{t}\right] G(t),$$

where t is any positive integer not exceeding n, and

$$G(\xi) = G([\xi]) = \sum_{m=1}^{\xi} f(m).$$

Proof: By the definition of G, $f(m) = G(m) - G(m-1)$, and so by partial summation (cf. the proof of Theorem 6-15),

$$\begin{aligned}
\sum_{m=1}^{n} F(m) &= \sum_{m=1}^{t} \left[\frac{n}{m}\right] f(m) + \sum_{m=t+1}^{n} \left[\frac{n}{m}\right] f(m) \\
&= \sum_{m=1}^{t} \left[\frac{n}{m}\right] f(m) + \sum_{m=t+1}^{n} \left[\frac{n}{m}\right] (G(m) - G(m-1)) \\
&= \sum_{m=1}^{t} \left[\frac{n}{m}\right] f(m) + \sum_{m=t}^{n} \left\{G(m) \cdot \left(\left[\frac{n}{m}\right] - \left[\frac{n}{m+1}\right]\right)\right\} \\
&\qquad - \left[\frac{n}{t}\right] G(t).
\end{aligned}$$

As was noted earlier, $[n/m] - [n/(m+1)]$ is the number of integers u such that

$$\frac{n}{m+1} < u \leq \frac{n}{m}.$$

For each such u, $n/u - 1 < m \leq n/u$, so that $m = [n/u]$ Hence

$$G(m)\left(\left[\frac{n}{m}\right] - \left[\frac{n}{m+1}\right]\right) = \sum_{n/(m+1)<u\leq n/m} G\left(\frac{n}{u}\right),$$

$$\sum_{m=t}^{n} G(m)\left(\left[\frac{n}{m}\right] - \left[\frac{n}{m+1}\right]\right) = \sum_{m=t}^{n} \sum_{n/(m+1)<u\leq n/m} G\left(\frac{n}{u}\right)$$

$$= \sum_{u=1}^{n/t} G\left(\frac{n}{u}\right),$$

and the proof is complete

THEOREM 6–31 $\displaystyle\sum_{m=1}^{n} \tau(m) = n \log n + (2\gamma - 1)n + O(n^{\frac{1}{2}}),$

where γ is Euler's constant

Proof Take $F = \tau$, $f = 1$, $t = [\sqrt{n}]$ in Theorem 6–30, then $G(\xi) = [\xi]$ and

$$\sum_{m=1}^{n} \tau(m) = \sum_{m=1}^{[\sqrt{n}]} \left[\frac{n}{m}\right] + \sum_{m=1}^{n/[\sqrt{n}]} \left[\frac{n}{m}\right] - \left[\frac{n}{[\sqrt{n}]}\right][\sqrt{n}]$$

$$= 2\sum_{m=1}^{\sqrt{n}} \left[\frac{n}{m}\right] - n + O(\sqrt{n})$$

$$= 2\sum_{m=1}^{\sqrt{n}} \frac{n}{m} + O(\sqrt{n}) - n + O(\sqrt{n})$$

$$= 2n\sum_{m=1}^{\sqrt{n}} \frac{1}{m} - n + O(\sqrt{n})$$

$$= 2n\,(\log \sqrt{n} + \gamma + O(1/\sqrt{n})) - n + O(\sqrt{n})$$

$$= n \log n + n(2\gamma - 1) + O(\sqrt{n})$$

The term $O(\sqrt{n})$ in Theorem 6–31 is not the best possible estimate of the error The problem of increasing the accuracy of the estimate, usually called Dirichlet's divisor problem has received a large amount of study It is known that $O(n^{\frac{1}{2}})$ can be replaced by $O(n^{\frac{1}{3}})$, but not by $O(n^{\frac{1}{4}})$ The exact exponent, if such exists is still unknown

For the purpose of illustrating the methods available for estimating averages, we give a second proof of Theorem 6–31 By Theorem 6–29,

$$\sum_{m=1}^{n} \tau(m) = \sum_{m=1}^{n} \left[\frac{n}{m}\right]$$

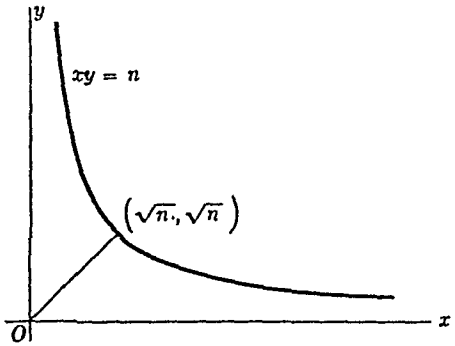

FIGURE 6-5

Geometrically, this last sum represents the number of *lattice points* (x, y) (that is, points such that x and y are integers) with positive coordinates, on or below the hyperbola $xy = n$, since for fixed x the number of integers y such that $1 \leq y \leq n/x$ is exactly $[n/x]$.

By symmetry, the number of lattice points (x, y) with $0 < xy \leq n$, $y > x$, is equal to the number with $0 < xy \leq n$, $y < x$ (see Fig. 6-5). Hence the number of points (x, y) with $0 < xy \leq n$ is twice the number of those with $y > x$, plus the number with $y = x$:

$$\begin{aligned}\sum_{m=1}^{n} \tau(m) &= 2\sum_{x=1}^{\sqrt{n}} \left(\left[\frac{n}{x}\right] - x\right) + [\sqrt{n}] \\ &= 2n\sum_{1}^{n} \frac{1}{x} + O(\sqrt{n}) - 2\frac{[\sqrt{n}]([\sqrt{n}]+1)}{2} + O(\sqrt{n}) \\ &= 2n\left(\log\sqrt{n} + \gamma + O(1/\sqrt{n})\right) - n + O(\sqrt{n}) \\ &= n\log n + (2\gamma - 1)n + O(\sqrt{n}).\end{aligned}$$

To get an asymptotic estimate for the sum of the first n values of the φ-function, we need a preliminary result concerning the famous *Riemann ζ-function*, which is defined for $s > 1$ by the equation

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

For $s \leq 1$ the function is, for the time being, undefined, since the series fails to converge for such s. It is a well-known result, which we shall use without proof (see Problem 6 below), that

$$\zeta(2) = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

THEOREM 6–32 *For* $s > 1$,

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}.$$

Proof The series of the theorem and that for $\zeta(s)$ converge absolutely for $s > 1$, so that they may be multiplied together by adding all possible products of a term from one series and a term from the other, and the resulting terms may be arranged in any convenient order Hence

$$\sum_{m=1}^{\infty} \frac{1}{m^s} \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s} = \sum_{m,n=1}^{\infty} \frac{\mu(n)}{(mn)^s} = \sum_{t=1}^{\infty} \frac{1}{t^s} \sum_{d|t} \mu(d) = 1$$

THEOREM 6–33 $\displaystyle\sum_{m=1}^{n} \varphi(m) = \frac{3n^2}{\pi^2} + O(n \log n)$

Proof Since

$$\varphi(m) = m \sum_{d|m} \frac{\mu(d)}{d},$$

we have

$$\begin{aligned}
\sum_{m=1}^{n} \varphi(m) &= \sum_{m=1}^{n} m \sum_{d|m} \frac{\mu(d)}{d} = \sum_{dd' \le n} d'\mu(d) = \sum_{d=1}^{n} \mu(d) \sum_{d'=1}^{n/d} d' \\
&= \sum_{d=1}^{n} \mu(d) \frac{[n/d]^2 + [n/d]}{2} = \frac{1}{2} \sum_{d=1}^{n} \mu(d) \left[\frac{n}{d}\right]^2 \\
&\qquad + O\left(\sum_{d=1}^{n} \left[\frac{n}{d}\right]\right) \\
&= \frac{1}{2} \sum_{d=1}^{n} \mu(d) \frac{n^2}{d^2} + O\left(\sum_{d=1}^{n} \frac{n}{d}\right) + O(n \log n) \\
&= \frac{n^2}{2} \left(\sum_{d=1}^{\infty} \frac{\mu(d)}{d^2} - \sum_{d=n+1}^{\infty} \frac{\mu(d)}{d^2}\right) + O(n \log n) \\
&= \frac{n^2}{2} \frac{1}{\zeta(2)} + O\left(n^2 \sum_{n+1}^{\infty} \frac{1}{d^2}\right) + O(n \log n) \\
&= \frac{3n^2}{\pi^2} + O(n) + O(n \log n) \\
&= \frac{3n^2}{\pi^2} + O(n \log n)
\end{aligned}$$

THEOREM 6–34. $\sum_{m=1}^{n} \sigma(m) = \frac{\pi^2 n^2}{12} + O(n \log n)$.

Proof:

$$\sum_{m=1}^{n} \sigma(m) = \sum_{m=1}^{n} \sum_{d|m} d = \sum_{m=1}^{n} \sum_{d=1}^{n/m} d = \frac{1}{2} \sum_{m=1}^{n} \left(\left[\frac{n}{m}\right]^2 + \left[\frac{n}{m}\right] \right)$$

$$= \frac{1}{2} n^2 \sum_{1}^{n} \frac{1}{m^2} + O(n \log n)$$

$$= \frac{n^2}{2} \left(\zeta(2) - O(1/n)\right) + O(n \log n)$$

$$= \frac{n^2 \zeta(2)}{2} + O(n \log n).$$

PROBLEMS

1. Show that $\sum_{n \le x} \frac{\tau(n)}{n} = \frac{\log^2 x}{2} + 2\gamma \log x + O(1)$. [*Hint:* Use Theorems 6–15 and 6–31.]

2. Let $\delta(n)$ be the largest odd divisor of n. Show that

$$\sum_{n \le x} \delta(n) = \frac{x^2}{3} + O(x) \quad \text{and} \quad \sum_{n \le x} \frac{\delta(n)}{n} = \frac{2x}{3} + O(1).$$

[*Hint:* Classify the numbers less than or equal to x according to the exponents of the powers of 2 dividing them, and show that

$$\sum_{n \le x} \delta(n) = \sum_{n=0}^{(x-1)/2} (2n+1) + \sum_{n=0}^{(x-2)/4} \frac{4n+2}{2} + \sum_{n=0}^{(x-4)/8} \frac{8n+4}{4} + \cdots.]$$

*3. Show that $\sum_{n \le x} \frac{\varphi(n)}{n} \sim \frac{6x}{\pi^2}$.

Deduce that the numbers $\varphi(n)/n$ are not uniformly distributed in the interval (0, 1). [A sequence $\{a_n\}$ of numbers in (0, 1) is said to be uniformly distributed if, for every α and β with $0 \le \alpha < \beta < 1$,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{\substack{n \le N \\ \alpha \le a_n < \beta}} 1 = \beta - \alpha.]$$

4. Prove that $\sum_{d=1}^{n} \varphi(d) \left[\frac{n}{d}\right] = \frac{n(n+1)}{2}$.

(Use Theorem 6–29.)

*5 Using the result of Problem 1 Section 6–2 show that

$$\sum_{n\le x}\frac{1}{\varphi(n)}\sim \log x \sum{}'\frac{1}{n\varphi(n)},$$

where the accent designates summation over the square free integers

*6 Prove that $\zeta(2)=\pi^2/6$ by evaluating the double integral

$$I=\int_0^1\int_0^1\frac{dx\,dy}{1-xy}$$

in two ways [*Hint* Obtain $I=\zeta(2)$ from the expansion

$$\frac{1}{1-xy}=1+xy+x^2y^2+x^3y^3+\quad ,$$

which is valid for $|xy|<1$ Then evaluate the integral directly by rotating the coordinate system about the origin through 45° to obtain

$$I=4\int_0^{1/\sqrt{2}}du\int_0^u\frac{dv}{2-u^2+v^2}+4\int_{1/\sqrt{2}}^{\sqrt{2}}du\int_0^{\sqrt{2}-u}\frac{dv}{2-u^2+v^2},$$

integrating with respect to v and making the substitution $u=\sqrt{2}\cos\vartheta$]

6–11 An application As was pointed out earlier, it is not known whether 2 is a primitive root of infinitely many primes, nor has the same question been settled for any other fixed integer It is therefore natural to ask, what can be said about the size of the smallest primitive root g_p, as a function of p? Unfortunately the little that is known (such as the theorem that g_p is less than $\sqrt{p}\log^{17}p$ for large p) cannot be developed here, we content ourselves with an estimate for the smallest quadratic nonresidue n_p of p Since g_p is certainly a nonresidue of p any upper bound for g_p would imply the same bound for n_p, but not conversely

THEOREM 6–35 *There is a quadratic nonresidue of p between* 1 *and* $\sqrt{p}$, *for all sufficiently large primes p*

Proof Corresponding to each pair of integers x, y with $(x,y)=1$, $0<x<\sqrt{p}$, $0<y<\sqrt{p}$, there corresponds an integer z, unique modulo p, such that

$$x\equiv yz \pmod p \tag{8}$$

Different pairs x, y yield different z's. For if

$$x_1 \equiv y_1 z \pmod{p} \qquad \text{and} \qquad x_2 \equiv y_2 z \pmod{p},$$

then

$$x_1 y_2 \equiv x_2 y_1 \pmod{p},$$

whence $x_1 y_2 = x_2 y_1$. But this, together with the hypothesis that $(x_1, y_1) = (x_2, y_2) = 1$, clearly implies that $x_1 = x_2$ and $y_1 = y_2$.

Now there are $2\varphi(m)$ ordered pairs x, y of relatively prime positive integers whose largest element is m, if $1 < m < \sqrt{p}$, and one pair both of whose elements are 1, so that the total number of pairs is

$$1 + 2 \sum_{m=2}^{\sqrt{p}} \varphi(m) = 2 \sum_{m=1}^{\sqrt{p}} \varphi(m) - 1.$$

If this number is larger than $p/2$, then there are more than $(p-1)/2$ different residue classes z, and since there are only $(p-1)/2$ quadratic residues of p, *at least one z must be a nonresidue. But then it* follows from (8) that one of x and y (each of which is smaller than $\sqrt{p}$) is also a quadratic nonresidue of p. Thus the proof will be complete when it is shown that, for all large p,

$$2 \sum_{m=1}^{\sqrt{p}} \varphi(m) - 1 > \frac{p}{2}. \tag{9}$$

Using the estimate of Theorem 6-33, we have that

$$\begin{aligned} 2 \sum_{m=1}^{\sqrt{p}} \varphi(m) - 1 &= \frac{6}{\pi^2} [\sqrt{p}]^2 + O(\sqrt{p} \log p) \\ &= \frac{6}{\pi^2} p + O(\sqrt{p} \log p) \sim \frac{6p}{\pi^2}. \end{aligned}$$

Since $6/\pi^2 > \frac{1}{2}$, it is clear that (9) holds for sufficiently large p, where the lower bound of validity depends upon the implied constant in the term $O(\sqrt{p} \log p)$.

By refining the argument slightly, it can be shown that the error term in Theorem 6-33 is numerically less than $1 \cdot n \log n$. Using this, the phrase "sufficiently large p" of Theorem 6-35 can be replaced by "$p > 10^4$", and finally, by reference to tables of the smallest primitive roots of primes less than 10^4, it can be shown that $n_p < \sqrt{p}$ for every $p \neq 2, 3, 7, 23$.

REFERENCES

Section 6–1

J J Sylvester (*Mathematical Papers*, vol 4, New York Cambridge University Press, 1912, pp 588, 625–629) proved that an odd perfect number must have at least five distinct prime factors, and conjectured that it must have at least six His conjecture was verified by U Kühnel, *Mathematische Zeitschrift* (Berlin) **52**, 202–211 (1949) It follows that if there is an odd perfect number, it cannot be smaller than $2\,2\,10^{12}$

For a complete account of the Mersenne numbers, see R C Archibald, *Scripta Mathematica* **3**, 112–119 (1935), and H S Uhler, *Scripta Mathematica* **18**, 122–131, (1952)

Section 6–8

The proof of Bertrand's conjecture given here is a modification of that given by P Erdos, *Acta Universitatis Szegediensis* (Szeged, Hungary) **5**, 194–198 (1932) The proof of Theorem 6–22 is simpler than that originally given, it was found independently by Erdos and L Kalmár in 1939, but was not published

Section 6–11

The inequality $g_p < \sqrt{p}\log^{17}p$ is due to Erdos, *Bulletin of the American Mathematical Society* **51**, 131–132 (1945) The inequality

$$\left|\sum_{m=1}^{n}\varphi(m) - \frac{3n^2}{\pi^2}\right| < n\log n$$

is due to R Tambs-Lyche, *Kongelige Norske Videnskabers Selskabs Forhandlinger* (Trondhjem, Norway) **9**, 58–61 (1936) For further results on this error term, see Erdős and H N Shapiro, *Canadian Journal of Mathematics* **3**, 375–385 (1951) and A Z Valfish, *Akademiya Nauk Gruzinskoi SSR Trudy Tbilisskogo Matematicheskogo Instituta imeni A N Razmadze* **19**, 1–31 (1953) [American Mathematical Society Translations, Series 2, 4, 1–30]

CHAPTER 7

SUMS OF SQUARES

7–1 An approximation theorem. In this chapter we consider the following questions: Given k, what integers can be represented as a sum of k squares? If an integer is so representable, how many representations are there? Both problems will be completely solved for $k = 2$, a partial answer to the first will be given for $k = 3$, and it will be shown that every integer is a sum of four squares (and hence of k squares, if $k \geq 4$).

We shall need the following approximation theorem.

THEOREM 7–1. *If ξ is a real number and t is a positive integer, there are integers x and y such that*

$$\left|\xi - \frac{x}{y}\right| \leq \frac{1}{y(t+1)}, \qquad 1 \leq y \leq t.$$

Proof: The $t + 1$ numbers

$$0 \cdot \xi - [0 \cdot \xi], \qquad 1 \cdot \xi - [1 \cdot \xi], \qquad \ldots, \qquad t\xi - [t\xi]$$

all lie in the interval $0 \leq u < 1$. Call them, in increasing order of magnitude, $\alpha_0, \alpha_1, \ldots, \alpha_t$. Mark the numbers $\alpha_0, \ldots, \alpha_t$ on a circle of unit circumference, that is, a unit interval on which 0 and 1 are identified. Then the $t + 1$ differences

$$\alpha_1 - \alpha_0, \qquad \alpha_2 - \alpha_1, \qquad \ldots, \qquad \alpha_t - \alpha_{t-1}, \qquad \alpha_0 - \alpha_t + 1$$

are the lengths of the arcs of the circle between successive α's, and so they are non-negative and

$$(\alpha_1 - 0) + (\alpha_2 - \alpha_1) + \cdots + (1 - \alpha_t) = 1.$$

It follows that at least one of these $t + 1$ differences does not exceed $(t + 1)^{-1}$. But each difference is of the form

$$g_1\xi - g_2\xi - N,$$

where N is an integer, and we can take $y = |g_1 - g_2|$, $x = \pm N$.

7-2 Sums of two squares A representation of the positive integer n as a sum of two squares, say $n = x^2 + y^2$, will be termed *proper* or *improper* according as $(x, y) = 1$ or $(x, y) > 1$ Throughout this section "representable" will mean "representable as a sum of two squares," with an analogous meaning for "properly representable"

THEOREM 7-2 *If* $p \equiv 3 \pmod 4$ *and* $p|n$, *then* n *has no proper representation*

Proof If $p|n$, $n = x^2 + y^2$, and $(x, y) = 1$, then $p\nmid x$ and $p\nmid y$ Hence there is an integer u such that $y \equiv ux \pmod p$, and

$$x^2 + y^2 \equiv x^2 + u^2x^2 \equiv x^2(1 + u^2) \equiv 0 \pmod p,$$

so that

$$u^2 \equiv -1 \pmod p$$

It follows that -1 is a quadratic residue of p, and so either $p = 2$ or $p \equiv 1 \pmod 4$, by Theorem 5-3

THEOREM 7-3 *An integer* $n = \Pi p_i^{\alpha_i}$ *is representable if and only if* α_i *is even for every* i *for which* $p_i \equiv 3 \pmod 4$

Proof Suppose first that $p^{2k+1}||n$, where $p \equiv 3 \pmod 4$, and suppose that $n = x^2 + y^2$, where $(x, y) = d$ and $p^j||d$ Then $x = dx_1$ and $y = dy_1$, where $(x_1\ y_1) = 1$ But if $x_1^2 + y_1^2 = N$, then $p^{2k-2j+1}|N$, and $2k - 2j + 1 > 0$, this contradicts Theorem 7-2

It remains only to show that if $n = n_1n_2^2$, where n_1 is square-free and without divisors congruent to 3 (mod 4), then n is representable It suffices to prove n_1 representable Since

$$(x_1^2 + y_1^2)(x_2^2 + y_2^2) = (x_1x_2 + y_1y_2)^2 + (x_1y_2 - x_2y_1)^2, \quad (1)$$

the product of representable numbers is representable, this, together with the fact that $1 = 1^2 + 0^2$ is representable, shows that we need only consider the various prime factors of n_1 Since $2 = 1^2 + 1^2$ is representable, it suffices to show that if $p \equiv 1 \pmod 4$, then p is representable For later purposes, however, we prove a more precise result

THEOREM 7-4 *If* $n > 1$ *and* $u^2 \equiv -1 \pmod n$, *there are unique integers* x *and* y *such that*

$$n = x^2 + y^2, \qquad x > 0, \qquad y > 0, \qquad (x, y) = 1,$$
$$y \equiv ux \pmod n$$

Remark: In case n is a prime $p \equiv 1 \pmod 4$, the congruence $u^2 \equiv -1 \pmod p$ is solvable, and so p is representable.

Proof: The idea of the proof is to replace the equation $x^2 + y^2 = n$ by the equivalent conditions

$$x^2 + y^2 \equiv 0 \pmod n, \qquad 0 < x^2 + y^2 < 2n.$$

To satisfy these conditions, we require that x be one of the numbers $1, 2, \ldots, [\sqrt{n}]$, and then seek a y such that $y \equiv ux \pmod n$ (so that $x^2 + y^2 \equiv 0 \pmod n$) and $1 \leq y < \sqrt{n}$. But if $y \equiv ux \pmod n$, then $y = ux + an$, and so we want a linear combination of u and n to be small.

We apply Theorem 7–1, with

$$\xi = -\frac{u}{n}, \qquad t = [\sqrt{n}],$$

and see that there are integers a and x_1 such that $1 \leq x_1 \leq [\sqrt{n}]$ and

$$\left| -\frac{u}{n} - \frac{a}{x_1} \right| \leq \frac{1}{x_1(1 + [\sqrt{n}])} < \frac{1}{x_1\sqrt{n}},$$

so that

$$|ux_1 + na| < \sqrt{n}.$$

Put $y_1 = ux_1 + na$. If $y_1 > 0$, put $y' = y_1, x' = x_1$. If $y_1 < 0$, then $-y_1 \equiv -ux_1 \pmod n$, and since $u^2 \equiv -1 \pmod n$ we have

$$u^2 y_1 \equiv -ux_1 \pmod n,$$

$$u(-y_1) \equiv x_1 \pmod n,$$

and we take $x' = -y_1$, $y' = x_1$. In either case,

$$y' \equiv ux' \pmod n, \qquad x'^2 + y'^2 = n, \qquad x' > 0, \qquad y' > 0.$$

From the relation

$$\begin{aligned} n = x_1^2 + y_1^2 &= x_1^2 + u^2x_1^2 + 2ux_1na + n^2a^2 \\ &= x_1^2(1 + u^2) + ux_1an + na(ux_1 + an), \end{aligned}$$

we obtain

$$1 = \left(\frac{1 + u^2}{n} x_1 + ua\right) x_1 + ay_1,$$

so that $(x_1, y_1) = 1$, whence $(x', y') = 1$.

Finally, to prove the uniqueness, suppose that besides x', y', there is a pair x'', y'' satisfying all conditions of the theorem Then by equation (1),

$$n^2 = (x'^2 + y'^2)(x''^2 + y''^2) = (x'x'' + y'y'')^2 + (x'y'' - x''y')^2$$

But

$$x'x'' + y'y'' \equiv x'x'' + u^2x'x'' = x'x''(1 + u^2) \equiv 0 \pmod{n},$$

and since $x'x'' + y'y'' > 0$, it follows that $x'x'' + y'y'' = n$, $x'y'' - x''y' = 0$ Hence $x'' = kx'$, $y'' = ky'$, and it is clear that $k = 1$

THEOREM 7–5 *The number $P_2(n)$ of proper representations of n is four times the number of solutions of the congruence $u^2 \equiv -1 \pmod{n}$ Hence (by Theorems 5–1 and 5–2),*

$$P_2(n) = \begin{cases} 0 & \text{if } 4 \nmid n \text{ or if some } p \equiv 3 \pmod 4 \text{ divides } n, \\ 2^{s+2} & \text{if } 4 \nmid n, \text{ no } p \equiv 3 \pmod 4 \text{ divides } n, \text{ and } s \text{ is the} \\ & \quad \text{number of distinct odd prime divisors of } n \end{cases}$$

Proof The theorem is trivial if $n = 1$ If $n > 1$, then $xy \neq 0$, and the number of representations is four times the number of positive representations To each u such that $u^2 \equiv -1 \pmod{n}$, there corresponds exactly one proper representation with $x > 0$, $y > 0$, $y \equiv ux \pmod{n}$ Conversely, if $x^2 + y^2 = n$ and $(x, y) = 1$, then $(x, n) = 1$, so that the congruence $y \equiv ux \pmod{n}$ has a unique solution (mod n), and

$$x^2 + y^2 \equiv x^2(1 + u^2) \equiv 0 \pmod{n},$$

which implies that $u^2 \equiv -1 \pmod{n}$

COROLLARY *A prime $p \equiv 1 \pmod 4$ can be represented uniquely (up to order and sign) as a sum of two squares*

This follows immediately from the theorem, for in this case $P_2(p)$ is 8, so that p has essentially only one proper representation It clearly has no improper representation

PROBLEM

Show that if n is a positive odd number of which -2 is a quadratic residue, then there are integers x and y such that $2x^2 + y^2 = n$ and $(x, y) = 1$

7-3 The Gaussian integers. In order to obtain an expression for the *total* number of representations of an integer as a sum of two squares, we turn our attention momentarily to the arithmetic of the so-called Gaussian integers: the complex numbers $a + bi$, where a and b are ordinary (or rational) integers. In this section, Greek letters will be used exclusively to designate Gaussian integers, and the set of all such integers will be denoted by $R[i]$.

It is clear that if α and β are in $R[i]$, then so also are $\alpha \pm \beta$ and $\alpha\beta$. If $\alpha = a + bi$, then $(a + bi)(a - bi) = a^2 + b^2$ is called the *norm* of α, and designated by $\mathrm{N}\alpha$. It is easily verified that $\mathrm{N}\alpha\mathrm{N}\beta = \mathrm{N}\alpha\beta$.

An integer α whose reciprocal is also an integer is called a *unit*; since

$$\frac{1}{\alpha} = \frac{1}{a + bi} = \frac{a - bi}{\mathrm{N}\alpha},$$

α is a unit if and only if

$$(a^2 + b^2) | a \qquad \text{and} \qquad (a^2 + b^2) | b.$$

If $a \neq 0$ and $b \neq 0$, then $a^2 + b^2 > \max(|a|, |b|)$; hence either a or b must be zero. If $a = 0$, then $b^2 | b$, whence $b = \pm 1$. If $b = 0$, then $a = \pm 1$. Thus the units are ± 1, $\pm i$. The numbers $\pm\alpha$ and $\pm\alpha i$ are called the *associates* of α. An integer is a unit if and only if its norm is 1.

We say that α *divides* β, and write $\alpha | \beta$, if there is an integer γ such that $\beta = \alpha\gamma$. If $\alpha | \beta$, then $\mathrm{N}\alpha | \mathrm{N}\beta$. A unit divides any integer; if an integer has no divisors other than its associates and the units, it is said to be *prime*. Thus $1 + i$ is prime, since the equation

$$1 + i = (a + bi)(c + di)$$

implies

$$\mathrm{N}(1 + i) = 2 = \mathrm{N}(a + bi)\mathrm{N}(c + di),$$

which shows that either $\mathrm{N}(a + bi)$ or $\mathrm{N}(c + di)$ is 1, so that either $a + bi$ or $c + di$ is a unit. More generally, this argument shows that if $\mathrm{N}\alpha$ is a rational prime, then α is a prime of $R[i]$. Thus, corresponding to the representation $p = x^2 + y^2$ of a rational prime $p \equiv 1 \pmod 4$, we have the decomposition

$$p = (x + iy)(x - iy)$$

into primes of $R[i]$. In this case the factors are not associated: x and

y are relatively prime and numerically larger than 1, and are therefore distinct, so that the supposition that a relation of the form

$$i^n(x + iy) = x - iy$$

holds, yields

$$i^n = 1 \qquad \text{and} \qquad i^{n+1} = -i,$$

which is impossible

The primes $p \equiv 3 \pmod 4$ do not split further in $R[i]$, that is, they are also prime in the larger set For if

$$p = (a + bi)(c + di),$$

then

$$p^2 = (a^2 + b^2)(c^2 + d^2)$$

But the only factorizations of p^2 are $p \quad p$ and $1 \quad p^2$, and it is impossible that $a^2 + b^2 = c^2 + d^2 = p$, by Theorem 7-3, hence one of the numbers $a + bi$, $c + di$ is a unit

If α is not prime, it can be represented as a product of primes For then $\alpha = \beta\gamma$, where $\mathrm{N}\beta > 1$ and $\mathrm{N}\gamma > 1$, and consequently $\mathrm{N}\beta < \mathrm{N}\alpha$, $\mathrm{N}\gamma < \mathrm{N}\alpha$ If β and γ are primes we are through, if one is not, it can be factored with the factors having still smaller norms The process cannot continue indefinitely, since the norms are strictly decreasing positive rational integers, and so we come eventually to a *prime factorization*

To show the uniqueness of this factorization, we use the following analog of Theorem 1-1

THEOREM 7-6 *If α and β are integers of $R[i]$, and $\beta \neq 0$ then there are integers ρ, κ such that*

$$\alpha = \beta\kappa + \rho, \qquad \mathrm{N}\rho < \mathrm{N}\beta$$

Proof Since $\beta \neq 0$, we can write

$$\frac{\alpha}{\beta} = \frac{a + bi}{c + di} = \frac{(a + bi)(c - di)}{c^2 + d^2} = A + Bi,$$

where A and B are rational numbers Let x be the nearest integer to A, and y the nearest integer to B, so that

$$|A - x| \leq \tfrac{1}{2},$$

$$|B - y| \leq \tfrac{1}{2}$$

Then

$$\left|\frac{\alpha}{\beta} - (x + iy)\right| = |(A - x) + (B - y)i|$$
$$= ((A - x)^2 + (B - y)^2)^{\frac{1}{2}} \leq (\tfrac{1}{4} + \tfrac{1}{4})^{\frac{1}{2}} < 1.$$

Hence if we put

$$x + iy = \varkappa, \qquad \alpha - \beta(x + iy) = \rho,$$

then

$$\mathbf{N}\rho = \mathbf{N}(\alpha - \varkappa\beta) = \mathbf{N}\beta \cdot \mathbf{N}\left(\frac{\alpha}{\beta} - \varkappa\right) < \mathbf{N}\beta,$$

and $\varkappa, \rho \in R[i]$. The proof is complete.

Starting from Theorem 7–6, the development in Chapter 2 leading to the Unique Factorization Theorem for rational integers can now be paralleled, and we obtain a Unique Factorization Theorem for $R[i]$.

THEOREM 7–7. *Every integer of $R[i]$ can be represented as a product of primes. This representation is unique, aside from the order of the factors and the presence of a set of units whose product is* 1.

It can be shown that the primes of $R[i]$ are exactly the ones we have already found, i.e., the associates of the following numbers:

(a) $1 + i$,

(b) $a \pm bi$, where $a^2 + b^2 = p \equiv 1 \pmod 4$,

(c) q, where $q \equiv 3 \pmod 4$.

PROBLEMS

1. Use the ideas of this section to give a new proof of the Corollary to Theorem 7–5.

2. Find the GCD (in $R[i]$) of $21 + 49i$ and $78 + 8i$.

3. Show that if $\alpha = a + bi$ is prime in $R[i]$, then either $(a, b) = 1$ or $ab = 0$. Use this to deduce that the only primes in $R[i]$ are those listed above. [*Hint:* If $(a, b) = 1$, show that $\mathbf{N}\alpha = 2^t p_1 \ldots p_r$, where $t = 0$ or 1 and $p_i \equiv 1 \pmod 4$ for $i = 1, 2, \ldots, r$. Note also that $\mathbf{N}\alpha = \mathbf{N}\bar{\alpha}$.]

7–4 The total number of representations. Suppose that n has the factorization

$$n = 2^u \cdot \prod_{p_j \equiv 1 \pmod 4} p_j^{t_j} \cdot \prod_{q_j \equiv 3 \pmod 4} q_j^{s_j}.$$

Put

$$\prod_{p_j \equiv 1 \pmod 4} p_j^{t_j} = n', \qquad \prod_{q_j \equiv 3 \pmod 4} q_j^{s_j} = m.$$

THEOREM 7–8. *If $n \geq 1$, then the number $r_2(n)$ of representations of n as a sum of two squares is zero if m is not a square, and is $4\tau(n')$ if m is a square*

Proof The case in which m is not a square is covered by Theorem 7–3 If m is a square, each s_j is even, and we can put $s_j = 2r_j$ In this case we shall prove the theorem by establishing, by means of the identity $x^2 + y^2 = (x + iy)(x - iy)$, a one-to-one correspondence between the various representations of n on the one hand, and the factorizations of n as a product of two conjugate Gaussian integers, on the other We must count these factorizations Since $1 + i = i(1 - i)$, and $2 = i(1 - i)^2$, we can write the prime decomposition of n in $R[i]$ in the form

$$n = i^u(1 - i)^{2u}\prod((a + bi)(a - bi))^t\prod q^{2r},$$

where the subscripts in the products have been omitted for clarity, and where

$$a > 0, \qquad b > 0, \qquad p = a^2 + b^2$$

Then every divisor of n in $R[i]$ is of the form

$$x + iy = i^v(1 - i)^{u_1}\prod((a + bi)^{t_1}(a - bi)^{t_2})\prod q^{r_1},$$

where

$$0 \leq v \leq u, \qquad 0 \leq u_1 \leq 2u, \qquad 0 \leq t_1 \leq t, \qquad 0 \leq t_2 \leq t,$$
$$0 \leq r_1 \leq 2r$$

Not every such divisor leads to a representation, we must also require that the complex conjugate,

$$\begin{aligned} x - iy &= (-i)^v(1 + i)^{u_1}\prod((a + bi)^{t_2}(a - bi)^{t_1})\prod q^{r_1} \\ &= i^{u_1 - v}(1 - i)^{u_1}\prod((a + bi)^{t_2}(a - bi)^{t_1})\prod q^{r_1}, \end{aligned}$$

be such that $(x + iy)(x - iy) = n$ It is clear that this is the case if and only if $u_1 = u$, $t_1 + t_2 = t$, $r_1 = r$ Since the powers of i are periodic, with period 4, we obtain all the distinct factorizations of n into conjugate factors by listing the numbers

$$i^v(1 - i)^u\prod((a + bi)^{t_1}(a - bi)^{t - t_1})\prod q^r,$$

where u, t, and r are fixed, v is one of the integers 0, 1 2 3, and t_1 is one of 0, 1, , t Their total number is $4\prod(t + 1) = 4\tau(n')$

PROBLEM

Show that $$\sum_{m=1}^{n} r_2(m) = \pi n + O(\sqrt{n}).$$

[*Hint:* The sum on the left is the number of lattice points inside or on the circle $x^2 + y^2 = n$. Associate each such point with the unit square of which it is the lower left corner. The resulting region has a polygonal boundary, no point of which is at distance greater than $\sqrt{2}$ from the circle.]

7-5 Sums of three squares. The problem of the solvability of the equation

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

is much more difficult than the corresponding question for the sum of either two or four squares. The result is this: (2) is solvable if and only if n is not of the form $4^t(8k + 7)$. We prove here only the trivial half of this theorem, that if n is of the specified form then (2) has no integral solutions.

Since a square can have only the values 0, 1, or 4 (mod 8), the sum of three squares is congruent to 0, 1, 2, 3, 4, 5, or 6 (mod 8), so that no $n \equiv 7 \pmod 8$ is so representable. If $4|n$ and (2) holds, then x, y, and z must all be even, so that $n/4$ must also be a sum of three squares. Therefore n cannot be a power of 4 times a nonrepresentable number.

It might be mentioned that one reason that problems concerning three squares are more difficult than those concerning either two or four is that there is no composition identity in this case analogous to (1) or to that given below for four squares. Indeed, the fact that 3 and 5 are sums of three squares, while 15 is not, shows that no such identity is possible.

7-6 Sums of four squares

THEOREM 7-9. *Every positive integer can be represented as a sum of four squares.*

Since

$$\begin{aligned}(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) \\ = (x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4)^2 + (x_1y_2 - x_2y_1 + x_3y_4 - x_4y_3)^2 \\ + (x_1y_3 - x_3y_1 + x_4y_2 - x_2y_4)^2 + (x_1y_4 - x_4y_1 + x_2y_3 - x_3y_2)^2,\end{aligned}$$

the product of representable numbers is representable Since 1 is also representable, *it suffices to prove that every prime p is representable* The proof, which uses the same idea as the proof of Theorem 7–4 depends on the following theorem

THEOREM 7–10 *Let r, s, and m be positive integers with $r < s$, and let λ_σ ($\sigma = 1, \quad , s$) be positive numbers (not necessarily integers) smaller than m, such that*

$$\lambda_1 \quad \lambda_s > m^r$$

Then the system of r linear congruences

$$\sum_{\sigma=1}^{s} a_{\rho\sigma} x_\sigma \equiv 0 \pmod{m}, \qquad \rho = 1, \quad , r,$$

where the a's are integers, has a solution in integers $x_1, \quad , x_s$ not all zero such that $|x_\sigma| \leq \lambda_\sigma$ for $\sigma = 1, \quad , s$

Proof Put

$$y_\rho = \sum_{\sigma=1}^{s} a_{\rho\sigma} x_\sigma, \qquad \text{for } \rho = 1, \quad , r$$

For each σ, let x_σ range over the integers $0, 1, \quad , [\lambda_\sigma]$, this gives $1 + [\lambda_\sigma]$ choices for x which are distinct (mod m) since $\lambda_\sigma < m$, and there are

$$\prod_{\sigma=1}^{s} (1 + [\lambda_\sigma])$$

different s tuples $x_1, \quad , x_s$ Corresponding to each s tuple $x_1, \quad , x_s$ there is an r tuple $y_1, \quad , y_r$, and so we have found

$$\prod_{\sigma=1}^{s} (1 + [\lambda_\sigma]) > \prod_{\sigma=1}^{s} \lambda_\sigma > m^r$$

r tuples $y_1, \quad , y_r$ But there are only m^r integral r tuples which are distinct (mod m), *so that there must be sets* $y_1', \quad , y_r'$ and $y_1'', \quad , y_r''$ such that $y_\rho' \equiv y_\rho''$ (mod m) for $\rho = 1, \quad r$ If these r-tuples correspond to $x_1' \quad x_s'$ and $x_1'', \quad , x_s''$ respectively, then

$$y_\rho' - y_\rho'' = \sum_{\sigma=1}^{s} a_{\rho\sigma}(x_\sigma' - x_\sigma'') \equiv 0 \pmod{m}, \qquad \rho = 1, \quad r$$

and not all of $x_\sigma' - x_\sigma''$ are zero, while $|x_\sigma' - x_\sigma''| \leq \lambda_\sigma$ for $\sigma = 1, \quad , s$

Proof of Theorem 7–9: If p is a prime, then the congruence

$$x^2 + y^2 + 1 \equiv 0 \pmod{p}$$

has a solution. For if x and y range independently over the numbers $0, 1, \ldots, (p-1)/2$ (this for odd p; the assertion is clearly correct for $p = 2$), then all the numbers x^2 are distinct (mod p), and the same is true of the numbers $-(1 + y^2)$. For if ${x_i}^2 \equiv {x_j}^2 \pmod{p}$, then $p \mid (x_i - x_j)(x_i + x_j)$. But $0 < x_i + x_j < p$, unless $x_i = x_j = 0$, so $p \mid (x_i - x_j)$, $x_i \equiv x_j \pmod{p}$, and so $x_i = x_j$. But we have altogether

$$\frac{p+1}{2} + \frac{p+1}{2} = p + 1$$

numbers x^2 and $-1 - y^2$, so some x^2 is congruent to some $-1 - y^2$, modulo p, which is the assertion.

Suppose that $a^2 + b^2 + 1 \equiv 0 \pmod{p}$. By Theorem 7–10, the congruences

$$x \equiv az + bt \pmod{p},$$

$$y \equiv bz - at \pmod{p}$$

have a nontrivial solution x, y, z, t with

$$\max(|x|, |y|, |z|, |t|) \leq \sqrt{p} + \epsilon;$$

here $r = 2$, $s = 4$, $m = p$, and we have chosen $\lambda_\sigma = \sqrt{p} + \epsilon$, where $\epsilon > 0$ is so small that $\sqrt{p} + \epsilon < p$. Now x, y, z, and t are integers, while $\sqrt{p}$ is not; if ϵ is chosen so small that $\sqrt{p} + \epsilon < 1 + [\sqrt{p}]$, it follows that

$$\max(|x|, |y|, |z|, |t|) < \sqrt{p}.$$

We have

$$x^2 + y^2 \equiv (a^2 + b^2)(z^2 + t^2) \equiv -(z^2 + t^2) \pmod{p},$$

while

$$0 < x^2 + y^2 + z^2 + t^2 < p + p + p + p = 4p,$$

so that

$$x^2 + y^2 + z^2 + t^2 = Ap,$$

where $A = 1$, 2, or 3. If $A = 1$, we are finished. If $A = 2$, then x is congruent to y, z, or t (mod 2). If $x \equiv y \pmod{2}$, then $z \equiv t \pmod{2}$,

and $$p = \left(\frac{x+y}{2}\right)^2 + \left(\frac{x-y}{2}\right)^2 + \left(\frac{z+t}{2}\right)^2 + \left(\frac{z-t}{2}\right)^2,$$

where the quantities in parentheses are integers

In the case $A = 3$, we note first that $p = 3$ has a representation $3 = 1^2 + 1^2 + 1^2$, so that we need only consider $p \neq 3$ The square of an integer is congruent to 0 or 1 (mod 3), and the equation

$$x^2 + y^2 + z^2 + t^2 = 3p$$

implies that

$$x^2 + y^2 + z^2 + t^2 \equiv 0 \pmod 3,$$

while

$$x^2 + y^2 + z^2 + t^2 \not\equiv 0 \pmod 9$$

By the congruence, one of the quantities—say x—is divisible by 3, and either all the others are, or all are not, divisible by 3 Because of the incongruence, $3 \nmid yzt$, so that y, z, and t are all congruent to $\pm 1 \pmod 3$ Let z' be that one of $\pm z$ such that $z' \equiv y \pmod 3$, and let t' be that one of $\pm t$ such that $t' \equiv y \pmod 3$ Then

$$p = \left(\frac{y+z'+t'}{3}\right)^2 + \left(\frac{x+z'-t'}{3}\right)^2 + \left(\frac{x-y+t'}{3}\right)^2 + \left(\frac{x+y-z'}{3}\right)^2,$$

where the quantities in parentheses are integers The proof is complete

REFERENCES

Section 7–5

A brief proof of the theorem that every number not of the form $4^k(8l + 7)$ can be represented as the sum of three squares is to be found in Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, Leipzig Teubner Gesellschaft, 1909, vol 1, pp 550–555

Section 7–6

Theorem 7 10, and its application to Theorem 7–9 are due to A Brauer and T L Reynolds, *Canadian Journal of Mathematics* 3, 367–373 (1951)

CHAPTER 8

PELL'S EQUATION AND SOME APPLICATIONS

8-1 Introduction. The Diophantine equation $x^2 - dy^2 = N$ (where N and d are integers), commonly known as Pell's equation, was actually never considered by Pell; it was because of a mistake on Euler's part that his name has been attached to it. The early Greek and Indian mathematicians had considered special cases, but Fermat was the first to deal systematically with it. He said that he had shown, in the special case where $N = 1$ and $d > 0$ is not a perfect square, that there are infinitely many integral solutions x, y; as usual, he did not give a proof. The first published proof was given by Lagrange, using the theory of continued fractions. Prior to this, Euler had shown that there are infinitely many solutions if there is one.

Before beginning a systematic investigation, it might be worth while to indicate some of the ways in which the equation arises and some of the reasons, therefore, for its importance. On the one hand, knowledge of the solutions of Pell's equation is essential in finding integral solutions of the general quadratic equation

$$ax^2 + bxy + cy^2 + dx + ey + f = 0,$$

in which $a, b, \ldots, f$ are integers. For, writing the left side as a polynomial in x,

$$ax^2 + (by + d)x + cy^2 + ey + f = 0,$$

it is clear that, if the equation is solvable for a certain y, the discriminant

$$(by + d)^2 - 4a(cy^2 + ey + f),$$

or, what is the same thing,

$$(b^2 - 4ac)y^2 + (2bd - 4ae)y + d^2 - 4af,$$

must be a perfect square, say z^2. Putting

$$b^2 - 4ac = p, \qquad 2bd - 4ae = q, \qquad d^2 - 4af = r,$$

we have

$$py^2 + qy + r - z^2 = 0.$$

Again, the discriminant of this quadratic in y must be a perfect square, say

$$q^2 - 4p(r - z^2) = w^2.$$

Thus we are led to consider the Pell equation

$$w^2 - 4pz^2 = q^2 - 4pr,$$

knowing solutions of it, we can, at any rate, obtain rational solutions of the original equation

As a second example, consider the *real quadratic field* $R(\sqrt{d})$, consisting of all the numbers of the form

$$a + b\sqrt{d}, \qquad d > 1, \ d \text{ square-free}$$

where a and b are rational numbers—positive, negative or zero Each of these numbers with $b \neq 0$ is a zero of a unique quadratic polynomial with relatively prime integral coefficients, that of x^2 being positive *If the leading coefficient of the polynomial is 1, the corresponding number is said to be an integer* of the field (Notice the close analogy between the present discussion and that in the first portion of Section 7–3 There, of course, we were working with the integers of the nonreal quadratic field $R(\sqrt{-1})$) Starting with this notion of integer, it is possible to construct an arithmetic very similar to that developed in Chapter 2 for the ordinary, or rational, integers Denote by $R[\sqrt{d}]$ the set of all integers of $R(\sqrt{d})$ If α, β, and α/β are in $R[\sqrt{d}]$, then we say that β *divides* α, and write $\beta|\alpha$ If $\alpha|1$, then α is a *unit* of $R[\sqrt{d}]$ If every factorization $\alpha = \beta\gamma$ into the product of integers of $R[\sqrt{d}]$ is such that either β or γ is a unit, then α is *prime* in $R[\sqrt{d}]$ Finally, the *norm* $\mathrm{N}\alpha$ of $\alpha = a + b\sqrt{d}$ is the product of α and its algebraic conjugate $\bar{\alpha} = a - b\sqrt{d}$, namely $a^2 - db^2$. *It is a rational integer if α is an integer of the field, and* $\mathrm{N}\alpha \cdot \mathrm{N}\beta = \mathrm{N}(\alpha\beta)$ always

Two complications now arise, however, which must be dealt with The more serious, with which we shall not be concerned for the time being, is that the analog of the Unique Factorization Theorem does not hold for every d, and it is necessary to introduce a rather sophisticated mechanism to deal with this problem The other complication is that, in distinction to the set of rational integers where there are only the two units ± 1, a real quadratic field has infinitely many, as will follow from the theorems of this chapter For it is easily seen

that α is a unit of $R(\sqrt{d})$ if and only if $\mathrm{N}\alpha = \pm 1$, that is, if and only if

$$a^2 - db^2 = \pm 1.$$

Since it can be shown that $a + b\sqrt{d}$ is a quadratic integer if and only if a and b are both rational integers (or in case $d \equiv 1 \pmod 4$, a and b may also be halves of odd integers), the infinitude of units follows from Lagrange's theorem concerning Pell's equation. Clearly, knowledge of the structure of this set of units will depend on a thorough analysis of that equation for $N = \pm 1$ and ± 4.

We shall give a third application of Pell's equation, this time to the minimum of an indefinite quadratic form, later in this chapter. There will be others in Volume II.

PROBLEMS

*1. Let d be greater than 1 and square-free, and let α be in $R[\sqrt{d}]$. Show that if $d \not\equiv 1 \pmod 4$, then

$$\alpha = a + b\sqrt{d}, \qquad a, b \text{ integers},$$

while if $d \equiv 1 \pmod 4$, then

$$\alpha = \frac{a + b\sqrt{d}}{2}, \qquad a, b \text{ integers such that } a \equiv b \pmod 2.$$

[*Hint:* First show that if $\bar{\alpha}$ is the conjugate of α in $R(\sqrt{d})$, then $\alpha + \bar{\alpha}$ and $\alpha\bar{\alpha}$, and therefore also $4\alpha\bar{\alpha} - (\alpha + \bar{\alpha})^2$, must be in $R[\sqrt{d}]$.]

2. Show that α is a unit of $R[\sqrt{d}]$ if and only if $\mathrm{N}\alpha = \pm 1$.

3. Find some solutions of the Diophantine equation

$$x^2 + 6xy - 4y^2 - 4x - 12y - 19 = 0.$$

8-2 The case $N = \pm 1$. For the present we shall concern ourselves with the equation

$$x^2 - dy^2 = 1. \tag{1}$$

The case in which d is a negative integer is easily dealt with: if $d = -1$, then the only solutions are $\pm 1, 0$ and $0, \pm 1$, while if $d < -1$, the only solutions are $\pm 1, 0$. So from now on we may restrict attention to equations of the form (1) with $d > 0$. If d is a square, then (1) can be written as

$$x^2 - (d'y)^2 = 1,$$

and since the only two squares which differ by 1 are 0 and 1, the only solutions in this case are $\pm 1, 0$. Suppose then that d is not a square.

THEOREM 8–1 *For any irrational number ξ, the inequality*

$$|x - \xi y| < \frac{1}{y} \tag{2}$$

has infinitely many solutions

Proof According to Theorem 7–1, if ξ is irrational the inequalities

$$0 < |x - \xi y| < \frac{1}{t}, \qquad 1 \leq y \leq t, \tag{3}$$

have a solution for each positive integer t It is clear that each solution of (3) is also a solution of (2) Taking $t = 1$ in (3) gives a solution x_1, y_1 of (2) Then for suitable $t_1 > 1$,

$$|x_1 - \xi y_1| > \frac{1}{t_1},$$

and taking $t = t_1$ in (3) gives a solution x_2, y_2 of (2) Since

$$|x_2 - \xi y_2| < |x_1 - \xi y_1|,$$

the two solutions so far found are distinct Now choose $t_2 > t_1$ so that

$$|x_2 - \xi y_2| > \frac{1}{t_2},$$

and for $t = t_2$ find x_3, y_3 Clearly this procedure can be continued indefinitely, yielding infinitely many solutions of (2)

THEOREM 8–2 *There are infinitely many solutions of the equation*

$$x^2 - dy^2 = k \tag{4}$$

in positive integers x, y for some k with $|k| < 1 + 2\sqrt{d}$

Proof If x, y is a solution of (2), then

$$|x + y\sqrt{d}| = |x - y\sqrt{d} + 2y\sqrt{d}| < \frac{1}{y} + 2y\sqrt{d} \leq (1 + 2\sqrt{d})y,$$

and so $$|x^2 - dy^2| < \frac{1}{y}(1 + 2\sqrt{d})y = 1 + 2\sqrt{d}$$

Since there are infinitely many distinct pairs x, y available, but only finitely many integers numerically smaller than $1 + 2\sqrt{d}$, infinitely many of the numbers $x^2 - dy^2$ must have a common value, which is the theorem

THEOREM 8-3. *Equation* (1) *has at least one solution with* $y \neq 0$.

Proof: Separate the infinitely many solutions of (4) into k^2 classes, putting two solutions x_1, y_1 and x_2, y_2 in the same class if and only if $x_1 \equiv x_2 \pmod{k}$ and $y_1 \equiv y_2 \pmod{k}$. Then some class contains at least two different solutions, say x_1, y_1 and x_2, y_2, with $x_1x_2 > 0$. Put

$$x = \frac{x_1x_2 - dy_1y_2}{k}, \qquad y = \frac{x_1y_2 - x_2y_1}{k};$$

we shall show that x and y are integers with $y \neq 0$ for which $x^2 - dy^2 = 1$. It follows immediately from the congruences

$$x_1 \equiv x_2 \pmod{k}, \qquad y_1 \equiv y_2 \pmod{k},$$

that

$$x_1y_2 \equiv x_2y_1 \pmod{k},$$

and so y is an integer. Also, from these congruences and from (4),

$$x_1x_2 - dy_1y_2 \equiv x_1^2 - dy_1^2 = k \equiv 0 \pmod{k},$$

and so x is an integer. Furthermore

$$\begin{aligned} x^2 - dy^2 &= \frac{1}{k^2}\left((x_1x_2 - dy_1y_2)^2 - d(x_1y_2 - x_2y_1)^2\right) \\ &= \frac{1}{k^2}\left(x_1^2x_2^2 - dx_1^2y_2^2 + d^2y_1^2y_2^2 - dx_2^2y_1^2\right) \\ &= \frac{1}{k^2}\left(x_1^2 - dy_1^2\right)\left(x_2^2 - dy_2^2\right) = 1. \end{aligned}$$

Finally, if $y = 0$, then

$$x_1y_2 = x_2y_1,$$

so that for some a, $x_1 = ax_2$ and $y_1 = ay_2$. But since x_1, y_1 and x_2, y_2 are both solutions of (4), it must be that $a = 1$, contrary to the assumption that x_1, y_1 and x_2, y_2 are different solutions.

THEOREM 8-4. *If* x_1, y_1 *and* x_2, y_2 *are solutions of the Pell equation* (1), *then so also are the integers* x, y *defined by the equation*

$$(x_1 + y_1\sqrt{d})(x_2 + y_2\sqrt{d}) = x + y\sqrt{d}. \tag{5}$$

Proof: It follows from (5) that also

$$(x_1 - y_1\sqrt{d})(x_2 - y_2\sqrt{d}) = x - y\sqrt{d},$$

and multiplying corresponding sides of these equations gives

$$x^2 - dy^2 = (x_1^2 - dy_1^2)(x_2^2 - dy_2^2) = 1$$

In particular, it follows from Theorems 8–3 and 8–4, taking $x_1 = x_2$, $y_1 = y_2$, that the numbers x, y defined by

$$(x_1 + y_1\sqrt{d})^n = x + y\sqrt{d}$$

form a solution for every positive value of n; that this is also true for negative values of n follows from the fact that

$$\frac{1}{x_1 + y_1\sqrt{d}} = x_1 - y_1\sqrt{d}$$

We shall now show that a general solution can be obtained in this fashion For brevity, we shall refer to $x + y\sqrt{d}$, as well as x, y, as a solution of equation (1) It will be called *positive* if $x > 0$ and $y > 0$ The positive solutions will be ordered by the size of x, or what is the same thing, by the size of $x + y\sqrt{d}$, since if $x_1 + y_1\sqrt{d}$ and $x_2 + y_2\sqrt{d}$ are positive solutions, and $x_1 > x_2$ then

$$x_1 + y_1\sqrt{d} > x_2 + y_2\sqrt{d},$$

and conversely

THEOREM 8 5 *If x_1, y_1 is the minimal positive solution of equation* (1), *then a general solution is given by the equation*

$$x + y\sqrt{d} = \pm(x_1 + y_1\sqrt{d})^n, \tag{6}$$

where n can assume any integral value, positive, negative or zero

Remark Because of this theorem, the minimal positive solution of (1) is sometimes called the *fundamental solution*

Proof That (6) actually furnishes a solution for each $n > 0$, we have just seen Since

$$(x + y\sqrt{d})^{-n} = (x - y\sqrt{d})^n,$$

(6) also gives a solution for each $n < 0$ Since the solutions with $y = 0$ correspond to $n = 0$ it suffices to show that (6) gives every solution with $y \neq 0$ Furthermore, if x_1, y_1 and n are positive and

$$x + y\sqrt{d} = (x_1 + y_1\sqrt{d})^n > 1,$$

then
$$-x + (-y)\sqrt{d} = -(x_1 + y_1\sqrt{d})^n < 1,$$
$$x + (-y)\sqrt{d} = (x_1 + y_1\sqrt{d})^{-n} < 1,$$
$$-x + y\sqrt{d} = -(x_1 + y_1\sqrt{d})^{-n} < 1,$$

so that it suffices to show that every solution of (1) with both x and y positive (so that $x + y\sqrt{d} > 1$) satisfies the equation

$$x + y\sqrt{d} = (x_1 + y_1\sqrt{d})^n, \qquad n > 0.$$

Put $x_1 + y_1\sqrt{d} = \alpha$; then if x, y is any positive solution of (1), $x + y\sqrt{d} \geq \alpha$, since α is minimal. Hence there is an $n > 0$ such that

$$\alpha^n \leq x + y\sqrt{d} < \alpha^{n+1}.$$

But then

$$1 \leq (x + y\sqrt{d})\alpha^{-n} = (x + y\sqrt{d})(x_1 - y_1\sqrt{d})^n < \alpha,$$

and this, by Theorem 8–4, contradicts the minimality of α unless

$$(x + y\sqrt{d})(x_1 - y_1\sqrt{d})^n = 1,$$

whence

$$x + y\sqrt{d} = (x_1 + y_1\sqrt{d})^n.$$

Turning now to the case $N = -1$, we find a somewhat similar situation, with the essential difference that the equation is not always solvable. This is the case, for example, when $d = 3$, for the expression $x^2 - 3y^2$ assumes only the values 0, 1, and 2 (mod 4). However, it is again true that all solutions can be expressed in terms of a single one, when such exists.

THEOREM 8–6. *Let d be a positive nonsquare integer. Then if the equation*

$$z^2 - dt^2 = -1 \tag{7}$$

is solvable, and if $z_1 + t_1\sqrt{d}$ is the minimal positive solution, a general solution is given by

$$z + t\sqrt{d} = \pm(z_1 + t_1\sqrt{d})^{2n+1}, \qquad n = 0, \pm 1, \ldots.$$

With the earlier notation,

$$\alpha = x_1 + y_1\sqrt{d} = (z_1 + t_1\sqrt{d})^2.$$

Proof: We prove the second assertion first. It is clear that

$(z_1 + t_1\sqrt{d})^2$ is a solution of (1), so that

$$1 < x_1 + y_1\sqrt{d} \leq (z_1 + t_1\sqrt{d})^2 \tag{8}$$

This gives

$$-z_1 + t_1\sqrt{d} < -z_1x_1 + dy_1t_1 + (-z_1y_1 + t_1x_1)\sqrt{d} \leq z_1 + t_1\sqrt{d},$$

where the number in the center of this inequality (which we will call $z + t\sqrt{d}$, for the moment) is again a solution of (7), so that in particular $t \neq 0$ But if a number lies between the minimal positive solution of (7) and its reciprocal, the same is true of the reciprocal of that number, so that either

$$1 < z + t\sqrt{d} \leq z_1 + t_1\sqrt{d}$$

or

$$1 < -z + t\sqrt{d} < z_1 + t_1\sqrt{d}$$

Using the minimality of $z_1 + t_1\sqrt{d}$, we conclude that

$$z + t\sqrt{d} = z_1 + t_1\sqrt{d}$$

Now suppose that $z + t\sqrt{d}$ is any solution of (7), where we can again restrict attention to the case $z, t > 0$ Then as in the proof of Theorem 8–5, we can find an n such that

$$1 \leq (z + t\sqrt{d})\alpha^{-n} < \alpha = (z_1 + t_1\sqrt{d})^2,$$

or, dividing through by $z_1 + t_1\sqrt{d}$,

$$-z_1 + t_1\sqrt{d} \leq x' + y'\sqrt{d} < z_1 + t_1\sqrt{d}$$

where $x' + y'\sqrt{d}$ satisfies (1) This inequality implies that

$$\alpha^{-1} < x' + y'\sqrt{d} < \alpha,$$

so that $x' + y'\sqrt{d} = 1$, and

$$z + t\sqrt{d} = (z_1 + t_1\sqrt{d})\alpha^n = (z_1 + t_1\sqrt{d})^{2n+1}$$

PROBLEMS

1 Find a general solution of the equation $x^2 - 2y^2 = 1$

2 Describe all the integral solutions of the equation

$$x^2 + 6xy + 7y^2 + 8x + 24y + 15 = 0$$

3. Show that

$$\liminf_{n\to\infty} (n(n\sqrt{2} - [n\sqrt{2}])) = \frac{1}{2\sqrt{2}}.$$

[The assertion means simply that if a_n stands for the quantity in parentheses, and if $\epsilon > 0$, then

$$a_n < \frac{1+\epsilon}{2\sqrt{2}}$$

for infinitely many n, while

$$a_n > \frac{1-\epsilon}{2\sqrt{2}}$$

for all sufficiently large n.]

4. Show that a necessary condition that the equation $x^2 - dy^2 = -1$ be solvable is that d have a proper representation as a sum of two squares.

8-3 The case $|N| > 1$. Because of its special interest in connection with the units of real quadratic fields, we consider separately the case $|N| = 4$.

THEOREM 8-7. *Let d be positive and not a square. If $r_1 + s_1\sqrt{d}$ is the minimal positive solution of the equation*

$$r^2 - ds^2 = 4, \tag{9}$$

then a general solution is given by

$$r + s\sqrt{d} = \pm 2\left(\frac{r_1 + s_1\sqrt{d}}{2}\right)^n, \qquad n = 0, \pm 1, \ldots. \tag{10}$$

If the equation

$$r'^2 - ds'^2 = -4 \tag{11}$$

is solvable, and its minimal positive solution is $r_1' + s_1'\sqrt{d}$, then a general solution is given by

$$r' + s'\sqrt{d} = \pm 2\left(\frac{r_1' + s_1'\sqrt{d}}{2}\right)^{2n+1}, \qquad n = 0, \pm 1, \ldots.$$

Proof: Clearly, the double of any solution of (1) is a solution of (9). While this remark shows that (9) is always solvable, not all the solutions can necessarily be found in this way, since, for example, $3^2 - 5 \cdot 1^2 = 4$, and 3 and 1 are odd.

If $r_2 + s_2\sqrt{d}$ and $r_3 + s_3\sqrt{d}$ are any solutions of (9), then

$$r + s\sqrt{d} = 2\,\frac{r_2 + s_2\sqrt{d}}{2}\,\frac{r_3 + s_3\sqrt{d}}{2}$$

is another integral solution For, from (9), $r_i^2 \equiv ds_i^2 \pmod 2$, so that $r_i \equiv ds_i \pmod 2$ Hence

$$2r = r_2r_3 + ds_2s_3 \equiv d^2s_2s_3 + ds_2s_3 \equiv d(d+1)s_2s_3 \equiv 0 \pmod 2$$

and

$$2s = r_2s_3 + r_3s_2 \equiv ds_2s_3 + ds_2s_3 \equiv 2ds_2s_3 \equiv 0 \pmod 2,$$

so that r and s are integers Also,

$$(r + s\sqrt{d})(r - s\sqrt{d}) = r^2 - ds^2 = 4\,\frac{r_2^2 - ds_2^2}{4}\,\frac{r_3^2 - ds_3^2}{4} = 4$$

It follows that the numbers $r + s\sqrt{d}$ defined in (10) are solutions of (9), for every n The remainder of the proof for the case $N = 4$ is an easy modification of the proof of Theorem 8–5 The proof for $N = -4$ is a straightforward combination of the above considerations *and the proof of Theorem 8–6*

For general N, the situation is rather complicated The following theorem gives a partial result

THEOREM 8–8 *If $d > 0$ is not a square, and if the Pell equation*

$$u^2 - dv^2 = N \tag{12}$$

has one solution, it has infinitely many In particular, if x_1, y_1 is a solution of equation (1) *and u_1, v_1 is a solution of equation* (12), *then the integers u, v determined by*

$$u + v\sqrt{d} = (x_1 + y_1\sqrt{d})(u_1 + v_1\sqrt{d}), \tag{13}$$

form a solution of equation (12)

Proof The second statement is proved in exactly the same way as was Theorem 8–4 The first statement follows immediately from the second, making use of Theorem 8–5

Notice that it may not be possible to obtain *all* solutions of (12) from one solution and the set of all solutions of (1) For example, the equation $u^2 - 2v^2 = 49$ has the solutions 7 and $9 + 4\sqrt{2}$, and neither can be obtained from the other by multiplying by a solution of $x^2 - 2y^2 = 1$

Theorem 8–8 can be used to obtain a finite criterion for the solvability of (12). If two solutions of (12) are related as in (13), we say that they belong to the same *class*. We now find bounds on the smallest element of each class, where the solutions are ordered by the size of u. (We can require that u be positive, since $u + v\sqrt{d}$ and $-u - v\sqrt{d}$ are in the same class.) The investigation is carried out only for $N > 0$; the case $N < 0$ is similar.

To do this we ask, given a solution $u_1 + v_1\sqrt{d}$ of (12), with $u_1 > 0$, when is it possible to find a smaller solution $u + v\sqrt{d}$, with $u > 0$, in the same class? That is, we want to find u and v such that

$$u + v\sqrt{d} = (x + y\sqrt{d})(u_1 + v_1\sqrt{d}), \qquad 0 < u < u_1,$$
$$x^2 - dy^2 = 1.$$

Let $\alpha = x_1 + y_1\sqrt{d}$ be the minimal positive solution of (1). If $v_1 > 0$, take $x + y\sqrt{d} = \alpha^{-1} = x_1 - y_1\sqrt{d}$; while if $v_1 < 0$, take $x + y\sqrt{d} = \alpha$; in either case, we get

$$u = u_1x_1 - y_1|v_1|d = u_1\left(x_1 - y_1\sqrt{d}\,\frac{|v_1|\sqrt{d}}{u_1}\right)$$
$$= u_1\left\{x_1 - y_1\sqrt{d} + y_1\sqrt{d}\left(1 - \sqrt{1 - \frac{N}{u_1^2}}\right)\right\}.$$

Here $0 < N/u_1^2 < 1$. Since

$$0 < 1 - \sqrt{1 - t} = \frac{t}{1 + \sqrt{1 - t}} < \frac{t}{2 - t}$$

for $0 < t < 1$, we have

$$0 < u < u_1\left(\alpha^{-1} + \frac{y_1\sqrt{d}\,N}{2u_1^2 - N}\right).$$

A little manipulation shows that the coefficient of u_1 is smaller than 1, so that $u < u_1$, if

$$u_1 > \sqrt{\frac{\beta y_1\sqrt{d} + 1}{2}\,N}, \qquad \text{where } \beta = \frac{\alpha}{\alpha - 1}.$$

Since $y_1\sqrt{d} = \sqrt{x_1^2 - 1} < x_1$, we have proved

THEOREM 8–9. *If equation* (12) *is solvable, it has a solution with*

$$0 < u < \sqrt{\frac{\beta x_1 + 1}{2} \cdot N}, \tag{14}$$

where $\alpha = x_1 + y_1\sqrt{d}$ *is the minimal positive solution of equation* (1) *and* $\beta = \alpha/(\alpha - 1)$ *If there are two or more classes of solutions of equation* (12), *each contains an element for which equation* (14) *holds*

This reduces the question of the solvability of (12) to a finite problem, once the minimal positive solution of (1) is known, it suffices to decide whether any of the numbers $(u^2 - N)/d$ is a square, for u in the interval (14) If there are two or more such values of u, it is a simple matter to decide whether the corresponding solutions are in the same class

For example, when $d = 2$ we have the minimal positive solution $3^2 - 2\ 2^2 = 1$, and it is easily seen that (14) holds if $0 < u < \frac{3}{2}\sqrt{N}$ Since also $N = u^2 - 2v^2 \leq u^2$, we need only examine the integers u between $\sqrt{N}$ and $\frac{3}{2}\sqrt{N}$, for each N

PROBLEMS

1 Complete the proof of Theorem 8 7

2 The statement obtained from Theorem 8 7 by replacing 2 and 4 by 7 and 49, respectively, is false as is seen by considering the numerical example immediately following Theorem 8-8 Where would the analogous proof break down?

3 Show that if $N < 0$, Theorem 8 9 remains correct if the inequality (14) is replaced by

$$0 < u < \sqrt{\frac{\alpha x_1 |N|}{2(\alpha + 1)}}$$

[*Hint* Prove and use the fact that for $t > 0$ $\sqrt{1 + t} - 1 < t/2$]

4 Describe all the units of $R(\sqrt{2})$, of $R(\sqrt{5})$ Cf Problem 1 Section 8-1

8–4 An application We showed in Theorem 8–1 that if ξ is irrational the inequality

$$\left|\xi - \frac{x}{y}\right| < \frac{1}{y^2}$$

has infinitely many solutions in integers x, y It is the object of the present section to make a more detailed examination of the approximability of a quadratic irrationality (that is, an irrational root of a quadratic equation with integral coefficients) by rationals, making use of the preceding results concerning Pell's equation

It is easy to see that when ξ is a quadratic irrationality, there is a constant $g_0 = g_0(\xi)$ such that the inequality

$$\left|\xi - \frac{x}{y}\right| < \frac{1}{gy^2}$$

does not hold for any x, y if $g > g_0$. For if ξ is defined by the equation

$$f(\xi) = a\xi^2 + b\xi + c = 0, \qquad a, b, c \text{ integers},$$

and if $f(x)$ factors as

$$f(x) = a(x - \xi)(x - \xi'),$$

then

$$\left|\xi - \frac{x}{y}\right| = \frac{|ax^2 + bxy + cy^2|}{y^2|a|\cdot|\xi' - x/y|} \geq \frac{1}{y^2|a|\cdot|\xi' - x/y|},$$

since $ax^2 + bxy + cy^2$ is an integer different from zero. Since ξ is irrational, $\xi \neq \xi'$. Hence from the above inequality, either

$$\left|\xi - \frac{x}{y}\right| > \frac{|\xi - \xi'|}{2} \qquad \text{or} \qquad \left|\xi - \frac{x}{y}\right| \geq \frac{1}{3y^2|a|\cdot|(\xi - \xi')/2|},$$

and we can take

$$g(\xi) = \min\left(\frac{2}{|\xi - \xi'|}, \quad \frac{2}{3|a|\cdot|(\xi - \xi')/2|}\right)$$

for any $\epsilon > 0$.

We are thus led to consider the quantity $M(\xi)$, which is the upper limit of those numbers λ for which the inequality

$$\left|\xi - \frac{x}{y}\right| < \frac{1}{\lambda y^2}$$

has infinitely many solutions. It was first treated by A. Markov, who made an extensive investigation of $M(\xi)$ in connection with the problem of determining an upper bound for the minimum value assumed by an *indefinite quadratic form*, i.e., an expression

$$Ax^2 + Bxy + Cy^2$$

in which $D = B^2 - 4AC > 0$, D is not a square, and x, y are integral variables. Markov made use of the theory of continued fractions, but we shall derive certain of his results using only the theorems just proved concerning Pell's equation.

In order to avoid interrupting the argument later, we first prove a lemma

THEOREM 8–10 *Let*

$$f(x, y) = ax^2 + bxy + cy^2$$

have integral coefficients such that $d = b^2 - 4ac > 0$ *and* d *is not a square Then if the equation*

$$f(x, y) = k \tag{15}$$

has one solution in integers, it has infinitely many, and each of the two quantities

$$|2ax + by + y\sqrt{d}| \qquad \text{and} \qquad |2ax + by - y\sqrt{d}| \tag{16}$$

is less than any prescribed positive number for infinitely many such solutions

Proof (a) In the case $k = a$, let X, Y be integers such that $X^2 - dY^2 = 1$, so that

$$4a^2X^2 - 4a^2dY^2 = 4a^2$$

If we put

$$2aX = 2ax + by, \qquad 2aY = y,$$

in which case x and y, given by

$$x = X - bY, \qquad y = 2aY, \tag{17}$$

are integers, then

$$(2ax + by)^2 - dy^2 = 4af(x, y) = 4a^2,$$

or

$$f(x, y) = a \tag{18}$$

Since by Theorem 8–5 there are infinitely many pairs X, Y, there are infinitely many solutions of (18) Since

$$\lim_{y \to \infty} \left\{ \left(\frac{2ax + by}{y} \right)^2 - d \right\} = \lim_{y \to \infty} \frac{4a^2}{y^2} = 0,$$

it is clear that one of the quantities in (16) is smaller than any prescribed $\epsilon > 0$ *for sufficiently large* y *Moreover, if the integers* x, y determined by X, Y in (17) are such that one of the quantities in (16) is small, then the numbers x', y' determined by X, $-Y$ are such that the other quantity is small

(b) In the case $k \neq a$, let x_1, y_1 be a solution of (15), and x_2, y_2 a solution of (18). Then the integers x, y defined by the equation

$$2ax + by + y\sqrt{d} = \frac{(2ax_1 + by_1 + y_1\sqrt{d})(2ax_2 + by_2 - y_2\sqrt{d})}{2a}$$

again satisfy (15). Since there are infinitely many solutions of (18), the same is true of (15). Furthermore, for fixed x_1, y_1, the first quantity in (16) will be small if x_2, y_2 ranges over those solutions of (18) for which $|2ax_2 + by_2 - y_2\sqrt{d}|$ is small, while the second will be small for the remaining solutions of (18).

THEOREM 8-11. *Let ξ be a real quadratic irrationality of discriminant d:*

$$a\xi^2 + b\xi + c = 0, \qquad d = b^2 - 4ac > 0, \qquad d \text{ not a square},$$
$$(a, b, c) = 1, \qquad a > 0.$$

Then if k is the smallest positive integer for which the equation

$$|ax^2 + bxy + cy^2| = k$$

has an integral solution,

$$M(\xi) = \frac{\sqrt{d}}{k}.$$

Proof: (a) $M(\xi)$ must be less than or equal to $\sqrt{d}/k$. For assume on the contrary that

$$M(\xi) = \frac{\sqrt{d}}{(1 - \delta)k},$$

where $0 < \delta < 1$. Then the inequality

$$\left|\frac{\sqrt{d} - b}{2a} - \frac{x}{y}\right| < \frac{(1 - \delta)k}{\sqrt{d}y^2}$$

holds for an infinite sequence S of distinct fractions x/y. (The case that $\xi = (-b - \sqrt{d})/2a$ is treated similarly.) Then, multiplying through by $\sqrt{d}$, we have

$$\left|\left(\frac{b}{2a} + \frac{x}{y}\right)\sqrt{d} - \frac{d}{2a}\right| < \frac{(1 - \delta)k}{y^2},$$

or

$$\left|\sqrt{d} - \frac{dy}{2ax + by}\right| < \frac{2a(1 - \delta)k}{y|2ax + by|}.$$

Hence

$$\frac{2a(1-\delta)k}{y|2ax+by|}\left|\sqrt{d}+\frac{dy}{2ax+by}\right| > \left|d-\frac{d^2y^2}{(2ax+by)^2}\right|$$

$$= \frac{4ad}{(2ax+by)^2}|ax^2+bxy+cy^2| \geq \frac{4akd}{(2ax+by)^2},$$

and, therefore,

$$1-\delta > \frac{2dy}{|2ax+by|\ \ |\sqrt{d}+dy/(2ax+by)|}$$

$$= \frac{2d}{|b+2ax/y|\ \ |\sqrt{d}+dy/(2ax+by)|} \tag{19}$$

But as x/y runs through the sequence S, y increases without limit and

$$b+2a\frac{x}{y} \to \sqrt{d},$$

$$\sqrt{d}+\frac{dy}{2ax+by} \to 2\sqrt{d},$$

so that

$$\lim_{\substack{y\to\infty \\ x/y\in S}} \frac{2d}{|b+2ax/y|\ \ |\sqrt{d}+dy/(2ax+by)|} = 1,$$

which contradicts (19)

(b) $M(\xi)$ must be greater than or equal to $\sqrt{d}/k$ For from the definition of k, and Theorem 8 10, we have that the equation

$$|(2ax+by)^2-dy^2| = 4ak$$

has infinitely many integral solutions x, y The left side factors into

$$|2ax+by-y\sqrt{d}|\ \ |2ax+by+y\sqrt{d}| = 4ak \tag{20}$$

By Theorem 8 10, each of these factors is small for infinitely many pairs x, y Henceforth we restrict x, y to the set T of solutions of (20) for which the first factor is smaller than the second (The proof in the alternate case proceeds similarly) Then

$$\frac{x}{y} \to \frac{-b+\sqrt{d}}{2a}$$

as $|y|$ tends to infinity Furthermore,

$$\left|\frac{x}{y} - \frac{-b+\sqrt{d}}{2a}\right| = \frac{4ak}{4a^2|x/y + (b+\sqrt{d})/2a|y^2}$$

$$= \frac{k}{a|x/y + (b+\sqrt{d})/2a|y^2}$$

and

$$\frac{x}{y} + \frac{b+\sqrt{d}}{2a} \to \frac{\sqrt{d}}{a}$$

as $|y|$ tends to infinity. Hence, given $\epsilon > 0$, the inequality

$$\left|\frac{x}{y} - \frac{-b+\sqrt{d}}{2a}\right| < \frac{k(1+\epsilon)}{y^2\sqrt{d}}$$

holds for all $(x, y) \in T$ with $|y| > y_0(\epsilon)$. Hence $M(\xi) \geq \sqrt{d}/k$.

The proof is now complete, since if $M(\xi) \geq \sqrt{d}/k$ and also $M(\xi) \leq \sqrt{d}/k$, it must be true that $M(\xi) = \sqrt{d}/k$.

COROLLARY. *If ξ is defined as in Theorem* 8–11, *then*

$$\frac{\sqrt{d}}{a} \leq M(\xi) \leq \sqrt{d}.$$

For clearly $k \geq 1$, and $k \leq a$ since $a \cdot 1^2 + b \cdot 1 \cdot 0 + c \cdot 0^2 = a$.

PROBLEM

Generalize the result of Problem 3, Section 8–2, evaluating

$$\liminf_{n\to\infty} n(n\sqrt{m} - [n\sqrt{m}]),$$

where m is a positive square-free integer.

8–5 The minima of indefinite quadratic forms. So far we have used Theorem 8–11 to obtain information concerning the quantity $M(\xi)$; it can also be used, in conjunction with the following well-known theorem of A. Hurwitz, to obtain information about the numerically smallest value assumed by an indefinite quadratic form.

THEOREM 8–12 (*Hurwitz' theorem*). *If ξ is any irrational number, then there are infinitely many integral solutions x, y of the inequality*

$$\left|\xi - \frac{x}{y}\right| < \frac{1}{\sqrt{5}\,y^2}.$$

Consequently, $M(\xi) \geq \sqrt{5}$ for every irrational ξ.

We defer the proof for a moment Assuming the theorem to be correct, a comparison of it and Theorem 8–11 yields the following result

THEOREM 8–13 *If $f(x, y)$ is an indefinite binary quadratic form of nonsquare discriminant d, then*

$$0 < |f(x, y)| \leq \frac{\sqrt{d}}{\sqrt{5}}$$

for suitable integers x, y

The coefficient $1/\sqrt{5}$ occurring here is best possible, in the sense that the theorem becomes false (for some quadratic forms) if $1/\sqrt{5}$ is replaced by a smaller constant For the form $k(x^2 + xy - y^2)$ has discriminant $5k^2$, and it is clear that this form assumes no nonzero value numerically smaller than $|k(1^2 + 1 \ \ 0 - 0^2)| = |k|$

8–6 Farey sequences, and a proof of Hurwitz' theorem A very simple proof of Hurwitz' theorem can be deduced from the well-known properties of the so-called *Farey sequences* F_n, which are the sequences of rational numbers a/b with $0 < b \leq n$, $(a, b) = 1$, arranged in increasing order of magnitude Thus for the first few values of n we have

$$F_1 \quad , \frac{-1}{1}, \frac{0}{1}, \frac{1}{1}, \frac{2}{1},$$

$$F_2 \quad , \frac{-1}{2}, \frac{0}{1}, \frac{1}{2}, \frac{1}{1}, \frac{3}{2},$$

$$F_3 \quad , \frac{-1}{3}, \frac{0}{1}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{1}{1}, \frac{4}{3},$$

$$F_4 \quad , \frac{-1}{4}, \frac{0}{1}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}, \frac{5}{4},$$

$$F_5 \quad , \frac{-1}{5} \ \frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1}, \frac{6}{5},$$

Clearly the number of elements of F_n which lie between 0 and 1 inclusive is $1 + \varphi(1) + \varphi(2) + \quad + \varphi(n)$

The rational numbers p/q and r/s are said to be *adjacent in* F_n if they are successive elements of F_n.

THEOREM 8–14. (a) *If* p/q *and* r/s *are adjacent in* F_n, *then* $|ps - qr| = 1$.

(b) *If* $|ps - qr| = 1$, *then* p/q *and* r/s *are adjacent in* F_n *for*

$$\max(q, s) \leq n < q + s,$$

and they are separated by the single element $(p + r)/(q + s)$ *in* F_{q+s}.

Remark: This theorem on the one hand gives necessary and sufficient conditions that p/q and r/s be adjacent in F_n, and on the other hand gives the law of formation of the new elements that appear in going from F_n to F_{n+1}. The number $(p + r)/(q + s)$ is called the *mediant* of p/q and r/s.

Proof: Suppose that p/q and r/s are elements of F_n such that $qr - ps = 1$, so that $r/s > p/q$. As t varies continuously from zero to infinity, the number

$$f(t) = \frac{p + tr}{q + ts}$$

increases steadily from p/q to r/s, so that there is a one-to-one correspondence between the positive real numbers t and the points of the interval

$$\frac{p}{q} < x < \frac{r}{s}. \tag{21}$$

Moreover, it is clear that $f(t)$ is rational if and only if t is rational; since we are interested only in the rational numbers in the interval, we put $t = u/v$, where $(u, v) = 1$ and $u > 0$, $v > 0$. This gives

$$f\left(\frac{u}{v}\right) = \frac{vp + ur}{vq + us}.$$

Since

$$q(vp + ur) - p(vq + us) = u(qr - ps) = u,$$
$$s(vp + ur) - r(vq + us) = v(ps - qr) = -v,$$

we have $(vp + ur, vq + us) = 1$. Thus we have shown that as u and v run over all pairs of relatively prime positive integers, the reduced fraction $(vp + ur)/(vq + us)$ runs over all rational numbers between p/q and r/s.

Among these fractions, the one with $u = v = 1$ is clearly the unique one of smallest denominator, it is the mediant of p/q and r/s, and

$$|(p + r)q - (q + s)p| = 1, \qquad |r(q + s) - s(p + r)| = 1$$

Since $q + s > \max(q\ s)$, part (b) of the theorem follows To prove (a), we proceed inductively F_1 consists of the integers , $-1/1, 0/1, 1/1,$, and $|a\ \ 1 - (a + 1)\ \ 1| = 1$, so that (a) is true for $n = 1$ If it is true for $n = m$, it is also true for $n = m + 1$, since the only elements of F_{m+1} not in F_m are certain mediants of adjacent elements of F_m The assertion follows by the induction principle

Proof of Hurwitz' theorem If a/b is a reduced fraction and c is a positive real number, designate by $I_c(a/b)$ the closed interval

$$\left[\frac{a}{b} - \frac{1}{cb^2}, \frac{a}{b} + \frac{1}{cb^2}\right]$$

Hurwitz' theorem says that if ξ is irrational, there are infinitely many fractions x/y such that $\xi \in I_{\sqrt{5}}(x/y)$

For each n, ξ lies between some two adjacent elements of F_n, say

$$\frac{p}{q} < \xi < \frac{r}{s}$$

We divide the interval $[p/q, r/s]$ into left and right halves

$$J_L = \left[\frac{p}{q}, \frac{p + r}{q + s}\right], \qquad J_R = \left[\frac{p + r}{q + s}\ \ \frac{r}{s}\right]$$

We now ask, how large may c be if it is required that the three intervals $I_c(p/q)$, $I_c\big((p + r)/(q + s)\big)$, $I_c(r/s)$ together completely cover the interval J_L? If this is the case, and $\xi \in J_L$, then ξ must be an interior point of one of these intervals I_c and we have a solution of the inequality $|\xi - x/y| < 1/cy^2$

Clearly $I_c(p/q)$ and $I_c(r/s)$ overlap (or abut) if and only if

$$\frac{p}{q} + \frac{1}{cq^2} \geq \frac{r}{s} - \frac{1}{cs^2},$$

and this reduces, with the aid of the relation $rq - ps = 1$, to

$$c \leq qs\left(\frac{1}{q^2} + \frac{1}{s^2}\right) = \frac{s}{q} + \frac{q}{s},$$

or, putting $f(x) = x + 1/x$, to

$$c \leq f\left(\frac{s}{q}\right).$$

Similarly, $I_c(p/q)$ and $I_c((p+r)/(q+s))$ overlap if and only if

$$c \leq f\left(\frac{q+s}{q}\right) = f\left(1 + \frac{s}{q}\right),$$

and so J_L is certainly covered by the intervals I_c if

$$c \leq \max\,(f(s/q), f(1 + s/q)),$$

and *a fortiori* if

$$c \leq \min_{x>0} \{\max\,(f(x), f(1+x))\} = c_0.$$

But a glance at the curves $y = f(x)$ and $y = f(1+x)$ shows that the curve $y = \max\,(f(x), f(1+x))$ is concave upward for $x > 0$, and has its minimum c_0 at $x = x_0$, where $f(x_0) = f(1 + x_0)$. (See Fig. 8–1.) A simple calculation gives

$$x_0 = \frac{\sqrt{5}-1}{2}, \qquad c_0 = \sqrt{5}.$$

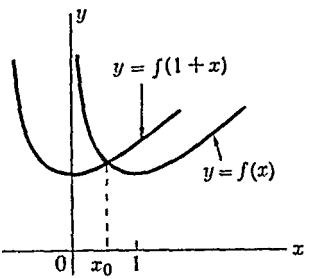

FIGURE 8–1

The proof can now be completed in either of two ways. The simpler is to note that ξ, being irrational, must lie, for infinitely many n, in the left half J_L of the interval between its surrounding Farey points; for if not, it would have to lie in all the intervals

$$\left[\frac{p}{q}, \frac{r}{s}\right], \qquad \left[\frac{p+r}{q+s}, \frac{r}{s}\right], \qquad \left[\frac{p+2r}{q+2s}, \frac{r}{s}\right], \qquad \ldots,$$

and the only point common to all these intervals is r/s itself. And whenever $\xi \in J_L$, the above argument shows that at least one of the numbers

$$\frac{p}{q}, \qquad \frac{p+r}{q+s}, \qquad \frac{r}{s}$$

affords a solution of Hurwitz' inequality. Finally, this gives infinitely

many solutions, because ξ lies in infinitely many J_L's and infinitely many J_R's, so that only finitely many of these intervals have a common end point

Alternatively, one can also examine the conditions under which the intervals $I_c(p/q)$, $I_c\big((p+r)/(q+s)\big)$, and $I_c(r/s)$ completely cover the interval J_R, an argument similar to that given above shows that this is the case if

$$c \leq \min_{x>0}\left(\max\left\{f(x), f\left(\frac{x}{1+x}\right)\right\}\right) = \sqrt{5}$$

It is then not necessary to distinguish the cases $\xi \in J_L$, $\xi \in J_R$

Again using the fact that ξ lies in infinitely many left half intervals and infinitely many right half-intervals, we deduce the following *stronger form of Hurwitz' theorem*

THEOREM 8-15 *If ξ is irrational there are infinitely many solutions of the inequality*

$$\left|\xi - \frac{x}{y}\right| < \frac{1}{\sqrt{5}\,y^2}. \tag{22}$$

If, for arbitrary n, ξ lies between the adjacent elements p/q and r/s of F_n, then at least one of the three numbers p/q, $(p+r)/(q+s)$, r/s is a solution of the inequality (22)

CHAPTER 9

RATIONAL APPROXIMATIONS TO REAL NUMBERS

9–1 Introduction. In the investigation of the solvability of the equation $n = x^2 + y^2$ in Chapter 7, it was convenient to use the fact that if x is real and t is a positive integer, there are integers p and q such that

$$|qx - p| \leq \frac{1}{t+1}, \qquad 1 \leq q \leq t.$$

In connection with Pell's equation we used an easy consequence of this theorem, that if x is irrational, the inequality

$$|qx - p| < \frac{1}{q}$$

has infinitely many integral solutions q and p with $q > 0$. Finally, the investigation, in Chapter 8, of the numerically smallest nonzero value assumed by an indefinite binary quadratic form hinged on Hurwitz' theorem, which states that the inequality

$$|qx - p| < \frac{1}{\sqrt{5}q}$$

has infinitely many integral solutions q and p with $q > 0$, if x is irrational. These theorems, while quantitatively different, all tell something about how small the absolute value of the linear form $qx - p$ can be made if the integers q and p are not both zero. Several generalizations of this problem come to mind at once, involving either a larger number of variables, or more than one such form, or both. The investigation of the behavior of such sets of forms is a central problem in the theory of Diophantine approximations; while many results have been obtained, few of them have the quantitative precision of Hurwitz' theorem, which becomes false if $\sqrt{5}$ is replaced by any larger constant. (This statement has not yet been proved; it is a consequence of Theorem 9–9.) One reason for this is that it is only in the simple case of one linear form in two variables, that a simple

algorithm can be constructed which yields all the pairs p, q for which $|qx - p|$ is "small," in a sense which will be made explicit below. Naturally, it is much easier to investigate the small values of a function if one knows what values to use for the arguments One of the objects of this chapter is to develop this algorithm

Rewriting Hurwitz' inequality in the form

$$\left|x - \frac{p}{q}\right| < \frac{1}{\sqrt{5}q^2},$$

we see that we are here concerned with a notion of "good" rational approximations to an irrational number which differs essentially from that generally understood in analysis There, we say that p/q is a better approximation to x than is r/s if

$$\left|x - \frac{p}{q}\right| < \left|x - \frac{r}{s}\right|$$

The question of finding this kind of good approximation is rather uninteresting arithmetically, although of course it may be necessary to use approximate values of irrational numbers in arithmetic investigations *What is involved in the theorems we are now discussing* is a comparison of the exactness of the approximation with the size of the denominator of the fraction used, the comparison being effected by taking the product

$$q\left|x - \frac{p}{q}\right|.$$

At least if x is irrational, the first factor in this product gets large as the second approaches zero, for out of all the elements of an arbitrary Farey sequence F_N there is one which is nearest to x, so to find a nearer rational number it is necessary to consider fractions whose denominators exceed N To require the above product to be small is therefore a much stronger condition than that imposed in analysis

Instead of searching immediately for appropriate fractions corresponding to a given x, it is fruitful (and indeed necessary, to make precise the meaning of "appropriate") to make x, instead of p/q, the unknown quantity That is, we fix a rational number p/q, and ask what numbers should be considered as having p/q as a good approximation Put so crudely, the question is unanswerable, we must decide what other rational numbers are competing with p/q It

seems natural to consider just the elements of some Farey sequence F_N which contains p/q, and to say that p/q is a "good" approximation to x if, for some $N \geq q$, $|qx - p| \leq |sx - r|$ for all r/s in F_N. This is perhaps most easily thought of this way: we measure distance from p/q in F_N, not by the usual expression $|x - p/q|$, but by $q|x - p/q|$, so that there is an individual measuring rod (or, more briefly, a *metric*) associated with each element of F_N. It is clear that "distances" increase more rapidly (in comparison with ordinary length) when measured from an element of F_N with large denominator than from an element with small denominator. We now associate with p/q all those points x such that the "distance" $|qx - p|$ from p/q is less than or equal to the "distance" $|sx - r|$ from an arbitrary element r/s of F_N. Call this set of points $R_N(p, q)$; formally, $R_N(p, q)$ is the set of x such that

$$\min_{r/s \in F_N} (|sx - r|) = |qx - p|.$$

Clearly, p/q itself is in $R_N(p, q)$.

Each of these sets $R_N(p, q)$ consists of a single interval. To see this, we first prove that if p/q and r/s are adjacent in F_N, then no point x between them belongs to any $R_N(t, u)$, if t/u is neither p/q nor r/s. This is obvious if x is either p/q or r/s. Suppose that

$$\frac{t}{u} < \frac{p}{q} < x < \frac{r}{s};$$

the other possible order, in which $t/u > r/s$, is treated similarly. If $q \leq u$, then

$$0 < qx - p = q\left(x - \frac{p}{q}\right) < q\left(x - \frac{t}{u}\right) \leq u\left(x - \frac{t}{u}\right) = ux - t,$$

so if the assertion is false, it must be that $q > u$. But then if

$$qx - p \geq ux - t,$$

so that

$$x \geq \frac{p - t}{q - u},$$

we have

$$0 < r - sx \leq r - s\,\frac{p - t}{q - u} = \frac{(qr - sp) - (ur - st)}{q - u} \leq 0,$$

since $qr - sp = 1$ while $ur - st \geq 1$. This contradiction shows that

$R_N(p, q)$ does not extend past the two elements to which p/q is adjacent in F_N But the condition

$$|qx - p| \leq |sx - r|$$

gives

$$qx - p \leq r - sx,$$

or

$$x \leq \frac{p + r}{q + s},$$

so that $R_N(p, q)$ consists of all x between the two points which are the mediants of p/q and its immediate neighbors in F_N

In particular, it follows that the new points which appear in going from F_N to F_{N+1} always appear at end points of intervals R_N

We now adopt the following convention if for some N, the number x is a point of $R_N(p, q)$ then p/q will be called a *best approximation* to x For this N, $|qx - p|$ is less than or equal to any expression $|sx - r|$ with $s \leq N$, *a fortiori*, $|x - p/q|$ is less than or equal to $|x - r/s|$ if $s \leq q$ so that if p/q is a best approximation to x in our present sense, it is also the rational number closest to x (in the ordinary sense) out of all those with denominators not exceeding q

There is thus associated with a fixed x a unique sequence of best approximations, the sequence will be infinite unless x is rational, in which case x lies inside its own interval R_N, for N greater than or equal to the denominator of x (For rational x, the sequence is not quite unique, since x is a common end point of two intervals R_N for some N) If $N \geq q$, $R_{N+1}(p, q)$ is contained in $R_N(p, q)$, so that if p/q is a best approximation to x, then certainly x is in $R_q(p, q)$ If h is the largest non negative integer for which $x \in R_{q+h}(p, q)$, then $x \in R_N(p, q)$ for $q \leq N \leq q + h$ Thus, if for fixed x we define a_N/b_N for $N = 1, 2,$ as that rational number such that $x \in R_N(a_N, b_N)$, then for suitable $N_0, N_1,$ we have $1 = N_0 < N_1 < N_2 <$ and

$$\frac{a_{N_0}}{b_{N_0}} = \frac{a_1}{b_1} = \frac{a_2}{b_2} = \quad = \frac{a_{N_1-1}}{b_{N_1-1}} \neq \frac{a_{N_1}}{b_{N_1}},$$

$$\frac{a_{N_1}}{b_{N_1}} = \frac{a_{N_1+1}}{b_{N_1+1}} = \quad = \frac{a_{N_2-1}}{b_{N_2-1}} \neq \frac{a_{N_2}}{b_{N_2}},$$

$$\frac{a_{N_2}}{b_{N_2}} = \frac{a_{N_2+1}}{b_{N_2+1}} = \quad = \frac{a_{N_3-1}}{b_{N_3-1}} \neq \frac{a_{N_3}}{b_{N_3}},$$

Since $a_{N_{k-1}}/b_{N_{k-1}}$ and a_{N_k}/b_{N_k} are adjacent elements of F_{N_k}, we have

$$|b_{N_k}a_{N_{k-1}} - b_{N_{k-1}}a_{N_k}| = 1, \qquad k = 1, 2, \ldots. \tag{1}$$

Now consider the following problem: *given a real number x, to find a systematic method for determining the sequence of best approximations to x.* We begin by reducing x by its greatest integer $[x] = \lambda_0$; the new number $x' = x - [x]$ is then in the interval $(0, 1)$. Put $P_0 = \lambda_0$, $Q_0 = 1$, so that P_0/Q_0 is or is not a_1/b_1 according as the fractional part x' of x is less than or greater than $\frac{1}{2}$. (In what follows, we shall assume that if x is rational, its denominator is sufficiently large that equality does not occur in statements such as the preceding one. This point will be considered in detail later.) If $P_0/Q_0 = a_1/b_1$, we put $P_k/Q_k = a_{N_k}/b_{N_k}$, while if $P_0/Q_0 \neq a_1/b_1$ we put $P_k/Q_k = a_{N_{k-1}}/b_{N_{k-1}}$, for $k = 1, 2, \ldots$. Thus the sequence $\{P_k/Q_k\}$ coincides with $\{a_{N_k}/b_{N_k}\}$, except that P_0/Q_0 may not be a best approximation. If we also put $P_{-1} = 1$, $Q_{-1} = 0$, then

$$Q_0P_{-1} - Q_{-1}P_0 = 1. \tag{2}$$

The numbers $P_1/Q_1, P_2/Q_2, \ldots$ are now to be determined. It turns out that this can be done using an algorithm, closely related to the Euclidean algorithm, of considerable importance in many branches of mathematics. Unfortunately, the deduction of this algorithm is necessarily somewhat complicated, since one must obtain the sequences $\{P_k\}$ and $\{Q_k\}$ from three others yet to be defined: $\{\alpha_k\}$, $\{x_k\}$, and $\{\lambda_k\}$. The final result, however, is quite simple.

If $P_0/Q_0 = a_1/b_1$, the relation

$$|Q_kP_{k-1} - Q_{k-1}P_k| = 1 \qquad k = 0, 1, \ldots, \tag{3}$$

holds, by (1) and (2). If $P_0/Q_0 \neq a_1/b_1$, then

$$\frac{P_1}{Q_1} - \frac{P_0}{Q_0} = 1, \qquad Q_0 = Q_1 = 1,$$

and

$$|Q_1P_0 - Q_0P_1| = 1, \tag{4}$$

so that (3) again holds, by (1), (2), and (4). The relation (3) is therefore always valid.

The numbers P_k and Q_k are now defined recursively, as follows: $P_{-1} = 1$, $Q_{-1} = 0$, $P_0 = [x]$, $Q_0 = 1$, and, for $k \geq 1$, P_k and Q_k constitute that solution p, q of the inequality

$$|qx - p| < |Q_{k-1}x - P_{k-1}|$$

for which q is positive and minimal If we put $\alpha_k = Q_k x - P_k$ for $k = 0, 1,$, we must find the minimal solution of the inequality

$$|qx - p| < |\alpha_{k-1}|$$

Fortunately, we need not consider all pairs p, q, but only those for which

$$|P_{k-1}q - Q_{k-1}p| = 1,$$

on account of (3) Since we know that one solution of this equation is $q = Q_{k-2}$, $p = P_{k-2}$, it follows that every solution is of the form

$$q = \epsilon(Q_{k-2} + \lambda Q_{k-1}), \qquad p = \epsilon(P_{k-2} + \lambda P_{k-1}), \tag{5}$$

where $\epsilon = \pm 1$ and λ is an integer, so that

$$|qx - p| = |\lambda(Q_{k-1}x - P_{k-1}) + (Q_{k-2}x - P_{k-2})| = |\lambda\alpha_{k-1} + \alpha_{k-2}|$$

Thus we can rephrase the definition of P_k and Q_k if $k \geq 1$ and *P_l and Q_l are known for $l < k$, then P_k, Q_k are the p and q of equations* (5) if λ and ϵ are so determined that

$|\lambda\alpha_{k-1} + \alpha_{k-2}| < |\alpha_{k-1}|$, $\epsilon(\lambda Q_{k-1} + Q_{k-2})$ is positive and minimal

Since $Q_{k-1} > 0$, these conditions are equivalent to the following

$$\left|\lambda - \left(-\frac{\alpha_{k-2}}{\alpha_{k-1}}\right)\right| < 1, \quad \epsilon\left\{\lambda - \left(-\frac{Q_{k-2}}{Q_{k-1}}\right)\right\} \text{ positive and minimal} \tag{6_k}$$

To see how to solve (6_k) let us consider the case $k = 1$ We have

$$\alpha_{-1} = 0 \quad x - 1 = -1 \qquad \alpha_0 = 1 \quad x - [x],$$

so that (6_k) becomes

$$\left|\lambda - \frac{1}{x - [x]}\right| < 1, \qquad \epsilon\left(\lambda - \frac{0}{1}\right) \text{ positive and minimal} \tag{6_1}$$

The number $$x_1 = \frac{1}{x - [x]}$$

is larger than 1, so that the two integral solutions λ of the inequality of (6_1) are positive, the solution of (6_1) is clearly

$$\lambda = \lambda_1 = [x_1], \qquad \epsilon = +1$$

This gives $P_1 = \lambda_1 P_0 + P_{-1}, \qquad Q_1 = \lambda_1 Q_0 + Q_{-1},$

$$\begin{aligned} Q_1P_0 - Q_0P_1 &= (\lambda_1 Q_0 + Q_{-1})P_0 - Q_0(\lambda_1 P_0 + P_{-1}) \\ &= -(Q_0P_{-1} - Q_{-1}P_0), \end{aligned}$$

and since $Q_0P_{-1} - Q_{-1}P_0 = 1 \cdot 1 - 0 \cdot \lambda_0 = 1,$

we have $Q_1P_0 - Q_0P_1 = -1.$

These calculations provide a basis for an inductive proof of the following theorem.

THEOREM 9–1. *Put* $x_0 = x$, *and define*

$$x_k = \frac{1}{x_{k-1} - [x_{k-1}]} \qquad \text{for } 1 \leq k < n + 1,$$

where n *is the smallest index for which* $x_n - [x_n] = 0$, *if there is such, and is infinity otherwise. Then for* $1 \leq k < n + 1$,

$$x_k = -\frac{\alpha_{k-2}}{\alpha_{k-1}}, \tag{7}$$

and the solution λ, ϵ *of* (6_k) *is* $\lambda = \lambda_k = [x_k]$, $\epsilon = +1$. *Hence* $\{P_k/Q_k\}$ *is recursively defined by the equations*

$$\begin{aligned} &P_{-1} = 1, \quad P_0 = \lambda_0, \quad P_k = P_{k-1}\lambda_k + P_{k-2} \\ &Q_{-1} = 0, \quad Q_0 = 1, \quad Q_k = Q_{k-1}\lambda_k + Q_{k-2} \end{aligned} \qquad \text{for } 1 \leq k < n + 1, \tag{8}$$

$$\text{and} \quad Q_kP_{k-1} - Q_{k-1}P_k = (-1)^k \qquad \text{for } 0 \leq k < n + 1. \tag{9}$$

Proof: As we have just seen, all the assertions of the theorem are true when $k = 1$, and (9) holds for $k = 0$. Suppose the assertions true for some $k < n + 1$, and for all indices smaller than this k. We wish to determine P_{k+1} and Q_{k+1} by solving (6_{k+1}).

If n is finite and $k + 1 = n + 1$, then $x_k - [x_k] = x_k - \lambda_k = 0$. Reversing the argument that led to (6_k), this means that $Q_kx - P_k = 0$, or that $x = P_n/Q_n$, and the sequence $\{P_k/Q_k\}$ terminates with P_n/Q_n. Thus the entire sequence of best approximations has already been determined when $k = n$, so that we need only consider the case $k < n$.

If $k < n$, we must solve

$$\left|\lambda - \left(-\frac{\alpha_{k-1}}{\alpha_k}\right)\right| < 1; \quad \epsilon\left\{\lambda - \left(-\frac{Q_{k-1}}{Q_k}\right)\right\} \text{ positive and minimal.} \tag{6_{k+1}}$$

From the induction hypothesis and the definition of x_{k+1}, we have

$$-\frac{\alpha_{k-1}}{\alpha_k} = -\frac{\alpha_{k-1}}{Q_k x - P_k} = -\frac{\alpha_{k-1}}{(Q_{k-1}\lambda_k + Q_{k-2})x - (P_{k-1}\lambda_k + P_{k-2})}$$

$$= -\frac{\alpha_{k-1}}{\alpha_{k-1}\lambda_k + \alpha_{k-2}} = \frac{1}{-\alpha_{k-2}/\alpha_{k-1} - \lambda_k} = \frac{1}{x_k - [x_k]} = x_{k+1} > 1$$

Since $-Q_{k-1}/Q_k < 0$, the solution of (6_{k+1}) is clearly

$$\lambda = \lambda_{k+1} = [x_{k+1}], \qquad \epsilon = +1,$$

whence $Q_{k+1} = Q_k\lambda_{k+1} + Q_{k-1}, \qquad P_{k+1} = P_k\lambda_{k+1} + P_{k-1}$

Moreover,

$$\begin{aligned} Q_{k+1}P_k - Q_kP_{k+1} &= (Q_k\lambda_{k+1} + Q_{k-1})P_k - Q_k(P_k\lambda_{k+1} + P_{k-1}) \\ &= -(Q_kP_{k-1} - Q_{k-1}P_k) = (-1)^{k+1}, \end{aligned}$$

by the induction hypothesis, and the proof is complete

To see how Theorem 9-1 solves the problem of finding the best approximations to x, take $x = \sqrt{2}$ Then $\lambda_0 = [\sqrt{2}] = 1$ and

$$x_1 = \frac{1}{x - [x]} = \frac{1}{\sqrt{2} - 1} = \sqrt{2} + 1, \qquad \lambda_1 = [\sqrt{2} + 1] = 2,$$

$$x_2 = \frac{1}{x_1 - [x_1]} = \frac{1}{\sqrt{2} - 1} = \sqrt{2} + 1, \qquad \lambda_2 = [\sqrt{2} + 1] = 2,$$

$$\vdots$$

and in general, $x_k = \sqrt{2} + 1$ and $\lambda_k = 2$, for $k \geq 1$ Hence

$$P_{-1} = 1, \qquad P_0 = 1, \qquad P_k = 2P_{k-1} + P_{k-2},$$

so that $\{P_k\} = 1, 1, 3, 7, 17, \quad$, and

$$Q_{-1} = 0, \qquad Q_0 = 1, \qquad Q_k = 2Q_{k-1} + Q_{k-2},$$

so that $\{Q_k\} = 0, 1, 2, 5, 12,$

Thus the best approximations to $\sqrt{2}$ are

$$1, \frac{3}{2}, \frac{7}{5}, \frac{17}{12}, \ldots$$

Of course, not every x will give a constant sequence of λ's, as happens with $\sqrt{2}$ In general, while arbitrarily many λ's can be determined, no explicit (i e , nonrecursive) formula for the entire sequence can be given

THEOREM 9-2. *If* $\{x_k\}$, $\{\lambda_k\}$, $\{P_k\}$, *and* $\{Q_k\}$ *are determined as in Theorem* 9-1, *then the following relations hold:*

$$x = \frac{P_{k-1}x_k + P_{k-2}}{Q_{k-1}x_k + Q_{k-2}}, \qquad 1 \leq k < n+1, \tag{10}$$

$$x = \lambda_0 + \cfrac{1}{\lambda_1 + \cfrac{1}{\lambda_2 + \cfrac{}{\ddots + \cfrac{1}{\lambda_{k-1} + \cfrac{1}{x_k}}}}}, \qquad 1 \leq k < n+1, \tag{11}$$

$$\frac{P_k}{Q_k} = \lambda_0 + \cfrac{1}{\lambda_1 + \cfrac{1}{\lambda_2 + \cfrac{}{\ddots + \cfrac{1}{\lambda_{k-1} + \cfrac{1}{\lambda_k}}}}}, \qquad 0 \leq k < n+1. \tag{12}$$

Proof: From (7) and the definition of the α's, we obtain

$$x_k = -\frac{Q_{k-2}x - P_{k-2}}{Q_{k-1}x - P_{k-1}},$$

which yields (10). The definition of $\{x_k\}$ and $\{\lambda_k\}$ gives

$$\begin{aligned} x &= \lambda_0 + \frac{1}{x_1}, \\ x_1 &= \lambda_1 + \frac{1}{x_2}, \\ &\vdots \\ x_{k-1} &= \lambda_{k-1} + \frac{1}{x_k}, \end{aligned} \tag{13}$$

and if we successively eliminate $x_1, x_2 \ldots, x_{k-1}$, equation (11) results. To obtain (12), consider the equations (13) with x_k as an

independent variable which assumes values greater than 1, with fixed λ_0 λ_1, , λ_{k-1} Then x is a function of x_k, given more briefly by (11) Since P_{k-1}, P_{k-2} Q_{k-1}, and Q_{k-2} depend only on λ_0 , λ_{k-1}, (10) and (11) are different expressions for the same functional relation If in (10) x_k is given the value λ_k then $x = P_k/Q_k$, and substituting these values in (11) gives (12)

PROBLEMS

1 Carry through the procedure described in this section to find the first few best approximations to $\sqrt{3}$

2 Find all the best approximations to 339/62

3 Show that if $x = \frac{1}{2}(1 + \sqrt{5})$, then each λ_i is 1

9-2 The rational case We now suppose that x is rational If x is an interior point of $R_{Q_k}(P_k, Q_k)$, we have the strict inequality $|Q_k x - P_k| < |Q_{k-1}x - P_{k-1}|$, or $|\lambda_k - x_k| < 1$ It may happen, however, that for some r/s with $s > Q_{k-1}$, x is the common end point of the abutting intervals $R_s(P_{k-1}, Q_{k-1})$ and $R_s(r, s)$, while x is an interior point of $R_{s-1}(P_{k-1}, Q_{k-1})$ In this case, it is a matter of choice whether r/s is to be included among the best approximations to x, it has not been included up to now, since we have required the strict inequality $|\lambda_k - x_k| < 1$ Fortunately, this ambiguity occurs only once for each x, for we know from earlier calculations that x is the mediant of P_{k-1}/Q_{k-1} and r/s

$$x = \frac{P_{k-1} + r}{Q_{k-1} + s}, \tag{14}$$

so that x is the first rational number to appear between P_{k-1}/Q_{k-1} and r/s in the sequences F_s F_{s+1}, Hence $k = n$, and if r/s is to be included among the best approximations, and if we put $r/s = P_n/Q_n$ then $P_{n+1}/Q_{n+1} = x$ Comparing equations (8) and (14), we see that $\lambda_{n+1} = 1$, and by (12),

$$x = \lambda_0 + \cfrac{1}{\lambda_1 + \cfrac{}{\ddots + \cfrac{1}{\lambda_n + \cfrac{1}{1}}}}$$

If r/s is not included, then $x = P_n/Q_n$ and

$$x = \lambda_0 + \cfrac{1}{\lambda_1 + \cfrac{\ddots}{\ddots + \cfrac{1}{(\lambda_n + 1)}}}.$$

To illustrate, take $x = \frac{2}{3}$. Then

$$x_1 = \frac{1}{\frac{2}{3} - [\frac{2}{3}]} = \frac{3}{2}, \qquad x_2 = \frac{1}{\frac{3}{2} - [\frac{3}{2}]} = 2,$$

$$\lambda_0 = [\tfrac{2}{3}] = 0, \qquad \lambda_1 = [x_1] = 1, \qquad \lambda_2 = [x_2] = 2,$$

and we have

$$\frac{2}{3} = 0 + \frac{1}{1 + \frac{1}{2}}, \qquad \frac{P_0}{Q_0} = \frac{0}{1}, \qquad \frac{P_1}{Q_1} = \frac{1}{1}, \qquad \frac{P_2}{Q_2} = \frac{2}{3}.$$

But we could also write

$$\frac{2}{3} = 0 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1}}},$$

with $\qquad \frac{P_0}{Q_0} = \frac{0}{1}, \qquad \frac{P_1}{Q_1} = \frac{1}{1}, \qquad \frac{P_2}{Q_2} = \frac{1}{2}, \qquad \frac{P_3}{Q_3} = \frac{2}{3}.$

In this case, $\frac{2}{3}$ is the right end point of $R_2(1, 2)$ and the left end point of $R_2(1, 1)$; with the normal procedure, $\frac{1}{2}$ would not be included among the best approximations to $\frac{2}{3}$.

An expression

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{\ddots}{\ddots + \cfrac{1}{a_n}}}} \tag{15}$$

is called a *finite regular continued fraction*, it is finite because there are only finitely many a's, and regular because the a's are integers $a_1, \quad , a_n$ are positive, and the numerators are all $+1$ We shall deal only with regular continued fractions in this book For typographical simplicity, we write

$$a_0 + \frac{1}{a_1 +} \quad \frac{1}{a_2 +} \quad \frac{1}{a_n} \,. \tag{16}$$

in place of (15) The numbers

$$\frac{p_0}{q_0} = a_0, \qquad \frac{p_1}{q_1} = a_0 + \frac{1}{a_1}, \qquad \frac{p_2}{q_2} = a_0 + \frac{1}{a_1 +} \frac{1}{a_2},$$

are called the *convergents* of the continued fraction (16) If (16) has the value x, we can put

$$x = a_0 + \frac{1}{a_1 +} \quad \frac{1}{a_{k-1} +} \frac{1}{x_k'},$$

where

$$x_k' = a_k + \frac{1}{a_{k+1} +} \quad \frac{1}{a_n}, \qquad \text{or} \qquad x_k' = a_k + \frac{1}{x_{k+1}'},$$

for $k = 1, 2, \quad , n$ Since $x_k' \geq 1$, we have

$$x_1' = \frac{1}{x - a_0} = \frac{1}{x - [x]}, \qquad a_1 = [x_1'],$$

$$x_2' = \frac{1}{x_1' - a_1} = \frac{1}{x_1' - [x_1']}, \qquad a_2 = [x_2'],$$

Thus the sequence $\{x_k'\}$ is identical with the sequence $\{x_k\}$ defined in Theorem 9–1 and $\{a_k\}$ is therefore identical with $\{\lambda_k\}$ Hence we have the following theorem

THEOREM 9–3 *The convergents (possibly excepting p_0/q_0) of any finite regular continued fraction are the best approximations to the value of the continued fraction Every rational number can be expanded into a finite regular continued fraction and this expansion is unique, except for the variation indicated by the identity*

$$a_0 + \frac{1}{a_1 +} \quad \frac{1}{a_n} = a_0 + \frac{1}{a_1 +} \quad \frac{1}{(a_n - 1) + 1}, \qquad a_n > 1$$

Moreover, the identities

(a) $p_0 = a_0,\ p_k = p_{k-1}a_k + p_{k-2},$

(b) $q_0 = 1,\ q_k = q_{k-1}a_k + q_{k-2},$

(c) $q_k p_{k-1} - q_{k-1}p_k = (-1)^k,$

(d) $$x = \frac{p_{k-1}x_k + p_{k-2}}{q_{k-1}x_k + q_{k-2}}$$

hold for $k = 1, 2, \ldots n$, *if we define* $p_{-1} = 1$, $q_{-1} = 0$, *and*

$$x = a_0 + \frac{1}{a_1 +} \cdots \frac{1}{a_n}, \qquad a_n \geq 1.$$

It might be worth mentioning that the continued fraction expansion of $x = p/q$ can be deduced immediately from the Euclidean algorithm as applied to a and b. For if

$$\begin{aligned} a &= ba_0 + r_0, \\ b &= r_0a_1 + r_1, \\ r_0 &= r_1a_2 + r_2, \\ &\vdots \\ r_{n-3} &= r_{n-2}a_{n-1} + r_{n-1}, \\ r_{n-2} &= r_{n-1}a_n, \end{aligned}$$

then

$$\begin{aligned} \frac{a}{b} &= a_0 + \frac{r_0}{b} = a_0 + \frac{1}{b/r_0}, \\ \frac{b}{r_0} &= a_1 + \frac{r_1}{r_0} = a_1 + \frac{1}{r_0/r_1}, \\ \frac{r_0}{r_1} &= a_2 + \frac{r_2}{r_1} = a_2 + \frac{1}{r_1/r_2}, \\ &\vdots \\ \frac{r_{n-3}}{r_{n-2}} &= a_{n-1} + \frac{r_{n-1}}{r_{n-2}} = a_{n-1} + \frac{1}{r_{n-2}/r_{n-1}}, \\ \frac{r_{n-2}}{r_{n-1}} &= a_n, \end{aligned}$$

so that

$$\frac{a}{b} = a_0 + \frac{1}{a_1 +} \frac{1}{a_2 +} \cdots \frac{1}{a_n}.$$

PROBLEMS

1 Prove that the convergents p_k/q_k are in reduced form, i e , that $(p_k, q_k) = 1$

2 If $x = a/b$ where $(a\ b) = 1$, and

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \quad \cfrac{1}{a_n}}},$$

then $p_n = a\ \ q_n = b$ Use this and an identity of Theorem 9–3 to find a solution of the linear Diophantine equation $ax + by = c$ In particular find a general solution of $247x + 77y = 31$

3 Show that for $k \leq n$,

$$\frac{q_k}{q_{k-1}} = a_k + \cfrac{1}{a_{k-1} + \quad \cfrac{1}{a_2 + \cfrac{1}{a_1}}}$$

9–3 The irrational case. Now consider the case that x is an irrational number ξ The sequences $\{x_k\}$, $\{\lambda_k\}$, and $\{P_k/Q_k\}$ are now infinite, and we write

$$\xi = \lambda_0 + \cfrac{1}{\lambda_1 + \cfrac{1}{\lambda_2 + }} \tag{17}$$

This equation must be understood as an abbreviation for the equation

$$\xi = \lim_{n\to\infty} \left(\lambda_0 + \cfrac{1}{\lambda_1 + \quad \cfrac{1}{\lambda_n}}\right) = \lim_{n\to\infty} \frac{P_n}{Q_n}, \tag{18}$$

the convergents P_n/Q_n play a role here analogous to that of the partial sums of an infinite series

Conversely, if we start with an arbitrary infinite regular continued fraction

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + }} \quad , \tag{19}$$

we can show that the convergents

$$\frac{p_0}{q_0} = a_0, \qquad \frac{p_1}{q_1} = a_0 + \frac{1}{a_1}, \qquad \frac{p_2}{q_2} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2}},$$

always converge to an irrational number ξ For take $n > 2$ and put

$$\frac{p}{q} = \frac{p_n}{q_n} = a_0 + \cfrac{1}{a_1 + \quad \cfrac{1}{a_n}}$$

Then by Theorem 9–1, the numbers $a_0, a_1, \ldots, a_{n-2}$ are uniquely determined by p/q, and the convergents of p/q, which are also convergents of (19), satisfy the usual recursion relations:

$$\begin{aligned} p_{-1} &= 1, \qquad p_0 = a_0, \qquad p_k = p_{k-1}a_k + p_{k-2}, \\ q_{-1} &= 0, \qquad q_0 = 1, \qquad q_k = q_{k-1}a_k + q_{k-2}, \end{aligned} \tag{20}$$

for $k = 1, 2, \ldots, n-2$. Moreover,

$$\begin{aligned} q_k p_{k-1} - q_{k-1}p_k &= (q_{k-1}a_k + q_{k-2})p_{k-1} - q_{k-1}(p_{k-1}a_k + p_{k-2}) \\ &= -(q_{k-1}p_{k-2} - q_{k-2}p_{k-1}) = \cdots \\ &= (-1)^k(q_0p_{-1} - q_{-1}p_0), \end{aligned}$$

so that

$$q_k p_{k-1} - q_{k-1}p_k = (-1)^k \tag{21}$$

for $k = 0, 1, \ldots, n-2$. Since n is arbitrary, the relations (20) and (21) hold for all $k \geq 1$. By (21),

$$\begin{aligned} q_k p_{k-2} - q_{k-2}p_k &= (q_{k-1}a_k + q_{k-2})p_{k-2} - q_{k-2}(p_{k-1}a_k + p_{k-2}) \\ &= a_k(q_{k-1}p_{k-2} - q_{k-2}p_{k-1}) \\ &= (-1)^{k-1}a_k. \end{aligned} \tag{22}$$

From (22), we see that

$$\frac{p_{2k-2}}{q_{2k-2}} < \frac{p_{2k}}{q_{2k}}, \qquad \frac{p_{2k-1}}{q_{2k-1}} > \frac{p_{2k+1}}{q_{2k+1}},$$

so that

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \frac{p_4}{q_4} < \cdots, \qquad \frac{p_1}{q_1} > \frac{p_3}{q_3} > \frac{p_5}{q_5} > \cdots.$$

By (21),

$$\frac{p_{2k}}{q_{2k}} < \frac{p_{2k+1}}{q_{2k+1}},$$

so that

$$\frac{p_{2k}}{q_{2k}} < \frac{p_{2l+1}}{q_{2l+1}}$$

for every $l \geq k$. Hence

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \frac{p_4}{q_4} < \cdots < \frac{p_5}{q_5} < \frac{p_3}{q_3} < \frac{p_1}{q_1},$$

so that the sequences $\{p_{2k}/q_{2k}\}$ and $\{p_{2k+1}/q_{2k+1}\}$, being monotonic and bounded, are convergent But (21) can be rewritten in the form

$$\frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k} = \frac{(-1)^k}{q_{k-1}q_k},$$

and since $q_k \to \infty$ as $k \to \infty$ we see that

$$\lim_{k\to\infty}\left(\frac{p_{2k}}{q_{2k}} - \frac{p_{2k+1}}{q_{2k+1}}\right) = 0,$$

and consequently $\lim p_k/q_k$ exists Call this limit ξ, and put

$$\xi = a_0 + \frac{1}{a_1 +} \quad \frac{1}{a_{n-1} +} \frac{1}{\xi_n}.$$

It follows, just as in the rational case, that

$$\xi_1 = \frac{1}{\xi - [\xi]}, \qquad \xi_2 = \frac{1}{\xi_1 - [\xi_1]}, \qquad \cdot \quad ,$$

and $$a_0 = [\xi], \qquad a_1 = [\xi_1], \qquad ,$$

so that the convergents p_k/q_k of (19) are the best approximations P_k/Q_k to ξ From this we deduce the following assertions

THEOREM 9–4 *Every infinite regular continued fraction converges to an irrational number, the best approximations to which are afforded by the convergents of the continued fraction Every irrational number can be expanded into an infinite regular continued fraction, and this expansion is unique Moreover, the following identities hold, if*

$$\xi = a_0 + \frac{1}{a_1 +}$$

$$\begin{aligned} p_{-1} &= 1, \quad p_1 = a_0, \quad p_k = p_{k-1}a_k + p_{k-2}, \\ q_{-1} &= 0, \quad q_1 = 1, \quad q_k = q_{k-1}a_k + q_{k-2}, \end{aligned} \qquad k = 1, 2, 3, \quad , \quad (23)$$

$$q_k p_{k-1} - q_{k-1}p_k = (-1)^k, \qquad k = 0, 1, 2, \quad , \quad (24)$$

$$q_k p_{k-2} - q_{k-2}p_k = (-1)^{k-1}a_k, \qquad k = 1, 2, 3, \quad , \quad (25)$$

$$\xi = \frac{p_{k-1}\xi_k + p_{k-2}}{q_{k-1}\xi_k + q_{k-2}}, \qquad \text{where} \qquad \xi = a_0 + \frac{1}{a_1 +} \quad \frac{1}{a_{k-1} + \frac{1}{\xi_k}},$$

$$k = 1, 2, 3, \qquad (26)$$

The numbers a_k are called the *partial quotients*, and the ξ_k the *complete quotients*, in the expansion.

Once the continued fraction expansion of ξ is known, the successive convergents can be computed very simply. For example, let $\xi = \sqrt{7}$. Then

$$\sqrt{7} = 2 + (\sqrt{7} - 2), \qquad a_0 = 2, \quad \xi_1 = (\sqrt{7} - 2)^{-1},$$

$$\frac{1}{\sqrt{7} - 2} = \frac{\sqrt{7} + 2}{3} = 1 + \frac{\sqrt{7} - 1}{3}, \quad a_1 = 1, \quad \xi_2 = \left(\frac{\sqrt{7} - 1}{3}\right)^{-1},$$

$$\frac{3}{\sqrt{7} - 1} = \frac{\sqrt{7} + 1}{2} = 1 + \frac{\sqrt{7} - 1}{2}, \quad a_2 = 1, \quad \xi_3 = \left(\frac{\sqrt{7} - 1}{2}\right)^{-1},$$

$$\frac{2}{\sqrt{7} - 1} = \frac{\sqrt{7} + 1}{3} = 1 + \frac{\sqrt{7} - 2}{3}, \quad a_3 = 1, \quad \xi_4 = \left(\frac{\sqrt{7} - 2}{3}\right)^{-1},$$

$$\frac{3}{\sqrt{7} - 2} = \sqrt{7} + 2 = 4 + (\sqrt{7} - 2), \quad a_4 = 4, \quad \xi_5 = (\sqrt{7} - 2)^{-1}.$$

Since $\xi_5 = \xi_1$, also $\xi_6 = \xi_2$, $\xi_7 = \xi_3, \ldots$, so $\{\xi_k\}$ (and therefore also $\{a_k\}$) is periodic. Thus we have the periodic expansion

$$\sqrt{7} = 2 + \frac{1}{1+} \frac{1}{1+} \frac{1}{1+} \frac{1}{4+} \frac{1}{1+} \frac{1}{1+} \frac{1}{1+} \frac{1}{4+} \cdots.$$

Using the relations (23), we construct the following table:

k	-1	0	1	2	3	4	5	6	$\cdots$
a_k		2	1	1	1	4	1	1	$\cdots$
p_k	1	2	3	5	8	37	45	82	$\cdots$
q_k	0	1	1	2	3	14	17	31	$\cdots$

Here the element $37 = p_4$, for example, is determined by multiplying $a_4 = 4$ by $p_3 = 8$ and adding $p_2 = 5$. Thus the best approximations to $\sqrt{7}$ are $3, \frac{5}{2}, \frac{8}{3}, \frac{37}{14}, \frac{45}{17}, \ldots$.

PROBLEMS

1 It can be verified, using a sufficiently good decimal approximation to π, that the beginning of the continued fraction expansion for π is

$$\pi = 3 + \frac{1}{7+}\,\frac{1}{15+}\,\frac{1}{1+}$$

Find the first four best approximations to π

2 Show that $\sqrt{13}$ has an expansion which is eventually periodic, and find the first few convergents

9–4 *Quadratic irrationalities.* *The problem of finding the best* approximations to a real number x has thus been completely solved in terms of the regular continued fraction expansion of x Of course, unless x is of a very special form, it may be impossible to give the complete expansion of x, just as one cannot give the rule of formation for the digits occurring in the decimal expansion of π But if a decimal approximation of x is known, a corresponding part of the continued fraction expansion of x can be determined quite easily For example, from the series expansion for e, one can easily show that

$$2\,7182 < e < 2\,7183$$

By a simple computation, we find that

$$2\,7182 = 2 + \frac{1}{1+}\,\frac{1}{2+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{4+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{3+}\,\frac{1}{1+}\,\frac{1}{9},$$

$$2\,7183 = 2 + \frac{1}{1+}\,\frac{1}{2+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{4+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{19+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{3},$$

so that $e = 2 + \dfrac{1}{1+}\,\dfrac{1}{2+}\,\dfrac{1}{1+}\,\dfrac{1}{1+}\,\dfrac{1}{4+}\,\dfrac{1}{1+}\,\dfrac{1}{1+}$

(Actually, it is known that the sequence of partial quotients is

$$2,\quad 1, 2, 1,\quad 1, 4, 1,\quad 1, 6, 1,\quad \ldots,\quad 1, 2n, 1,\quad \ldots)$$

There is, however, one simple case in which the complete expansion can be determined that in which the partial quotients $a_0, a_1, a_2, \ldots$ constitute a sequence which is eventually periodic Consider for example the continued fraction

$$\xi = 1 + \frac{1}{3+}\,\frac{1}{1+}\,\frac{1}{2+}\,\frac{1}{1+}\,\frac{1}{2+}\quad \ldots,$$

where $a_{2n} = 1$ and $a_{2n+1} = 2$ for $n \geq 1$. If ξ_2 is related to ξ as in (26), we have

$$\xi_2 = 1 + \frac{1}{2+}\,\frac{1}{1+}\,\frac{1}{2+}\cdots = 1 + \frac{1}{2+}\,\frac{1}{\xi_2},$$

so that

$$\xi_2 = 1 + \frac{1}{2 + \dfrac{1}{\xi_2}} = 1 + \frac{\xi_2}{2\xi_2 + 1} = \frac{3\xi_2 + 1}{2\xi_2 + 1},$$

$$2\xi_2^2 - 2\xi_2 - 1 = 0,$$

$$\xi_2 = \frac{-1 + \sqrt{3}}{2}.$$

(The plus sign is chosen since $\xi_2 > 0$.) Hence

$$\xi = 1 + \cfrac{1}{3 + \cfrac{1}{\cfrac{\sqrt{3}-1}{2}}} = \frac{4\sqrt{3} - 2}{3\sqrt{3} - 1} = \frac{17 - \sqrt{3}}{13}.$$

Conversely, if we start with a quadratic irrationality—say $\xi = 1 + \sqrt{6}$—we get

$$a_0 = [1 + \sqrt{6}] = 3,$$

$$\xi_1 = \frac{1}{\sqrt{6} - 2} = \frac{\sqrt{6} + 2}{2}, \qquad a_1 = \left[\frac{\sqrt{6} + 2}{2}\right] = 2,$$

$$\xi_2 = \cfrac{1}{\cfrac{\sqrt{6}+2}{2} - 2} = \frac{2}{\sqrt{6} - 2} = \sqrt{6} + 2, \qquad a_2 = 4,$$

$$\xi_3 = \frac{1}{\sqrt{6} - 2} = \xi_1,$$

so that

$$\sqrt{6} + 1 = 3 + \frac{1}{2+}\,\frac{1}{4+}\,\frac{1}{2+}\,\frac{1}{4+}\cdots.$$

We can now show that these are not isolated phenomena.

THEOREM 9–5 *Every eventually periodic regular continued fraction converges to a quadratic irrationality, and every quadratic irrationality has a regular continued fraction expansion which is eventually periodic*

Proof The first part is quite simple Suppose that the first period begins with a_n, and let the length of the period be h, so that $a_{k+h} = a_k$ for $k > n$ Put

$$\xi = a_0 + \frac{1}{a_1 +} \cdots, \qquad \text{and} \qquad \xi_k = a_k + \frac{1}{a_{k+1} +} \cdots,$$

so that $\xi_{k+h} = \xi_k$ for $k \geq n$ By this and equation (26),

$$\xi = \frac{p_{n-1}\xi_n + p_{n-2}}{q_{n-1}\xi_n + q_{n-2}} = \frac{p_{n+h-1}\xi_n + p_{n+h-2}}{q_{n+h-1}\xi_n + q_{n+h-2}},$$

so that ξ_n satisfies a quadratic equation with integral coefficients Since ξ_n is obviously not rational, it is a quadratic irrationality By (26) again, the same is true of ξ itself, since if

$$A\xi_n^2 + B\xi_n + C = 0,$$

then

$$\begin{aligned} A(-q_{n-2}\xi + p_{n-2})^2 + B(-q_{n-2}\xi + p_{n-2})(q_{n-1}\xi - p_{n-1}) \\ + C(q_{n-1}\xi - p_{n-1})^2 = 0 \end{aligned}$$

and this is a quadratic equation in ξ

The proof of the converse involves a little more computation Suppose that

$$A\xi^2 + B\xi + C = 0,$$

where A, B, and C are integers and ξ is irrational Then equation (26) gives

$$\begin{aligned} A(p_{k-1}\xi_k + p_{k-2})^2 + B(p_{k-1}\xi_k + p_{k-2})(q_{k-1}\xi_k + q_{k-2}) \\ + C(q_{k-1}\xi_k + q_{k-2})^2 = 0, \end{aligned}$$

or

$$A_k\xi_k^2 + B_k\xi_k + C_k = 0,$$

where the integers A_k, B_k, and C_k are given by the equations

$$\begin{aligned} A_k &= Ap_{k-1}^2 + Bp_{k-1}q_{k-1} + Cq_{k-1}^2, \\ B_k &= 2Ap_{k-1}p_{k-2} + B(p_{k-1}q_{k-2} + p_{k-2}q_{k-1}) + 2Cq_{k-1}q_{k-2}, \\ C_k &= Ap_{k-2}^2 + Bp_{k-2}q_{k-2} + Cq_{k-2}^2 \end{aligned}$$

If $f(x) = Ax^2 + Bx + C$, then

$$A_k = q_{k-1}^2 f\left(\frac{p_{k-1}}{q_{k-1}}\right), \qquad C_k = q_{k-2}^2 f\left(\frac{p_{k-2}}{q_{k-2}}\right).$$

Using Taylor's theorem, we have

$$A_k = q_{k-1}^2 \left\{ f(\xi) + f'(\xi)\left(\frac{p_{k-1}}{q_{k-1}} - \xi\right) + \frac{1}{2} f''(\xi)\left(\frac{p_{k-1}}{q_{k-1}} - \xi\right)^2 \right\},$$

since $f'''(x)$ is identically zero. Now $f(\xi) = 0$, and

$$\begin{aligned} \xi - \frac{p_{k-1}}{q_{k-1}} &= \frac{p_{k-1}\xi_k + p_{k-2}}{q_{k-1}\xi_k + q_{k-2}} - \frac{p_{k-1}}{q_{k-1}} = \frac{q_{k-1}p_{k-2} - q_{k-2}p_{k-1}}{q_{k-1}(q_{k-1}\xi_k + q_{k-2})} \\ &= \frac{(-1)^{k-1}}{q_{k-1}(q_{k-1}\xi_k + q_{k-2})}; \end{aligned} \tag{27}$$

since $\xi_k > 1$, it is certainly true that

$$\left| \xi - \frac{p_{k-1}}{q_{k-1}} \right| < \frac{1}{q_{k-1}^2}.$$

Hence

$$|A_k| < |f'(\xi)| + \frac{|f''(\xi)|}{2q_{k-1}^2},$$

and similarly

$$|C_k| < |f'(\xi)| + \frac{|f''(\xi)|}{2q_{k-2}^2},$$

so that $|A_k|$ and $|C_k|$ remain bounded as $k \to \infty$.

To see that $|B_k|$ is also bounded, we use the fact that all the quantities $B_k^2 - 4A_kC_k$ have the common value $B^2 - 4AC = D$. (This can be proved by a straightforward but tedious computation or, if one is acquainted with the theory of linear transformations, by noting that the expression $A_kx'^2 + B_kx'y' + C_ky'^2$ is obtained from $Ax^2 + Bxy + Cy^2$ by the unimodular substitution

$$\begin{aligned} x &= p_{k-1}x' + p_{k-2}y', \\ y &= q_{k-1}x' + q_{k-2}y', \end{aligned}$$

and that two such forms have the same discriminant.) Since A_k and C_k are bounded and D is fixed,

$$B_k^2 = D + 4A_kC_k$$

must be bounded also.

Thus, there is a constant M such that

$$|A_k| < M, \qquad |B_k| < M, \qquad |C_k| < M$$

for all k Since there are fewer than $(2M + 1)^3$ triples of integers numerically smaller than M, there must be three indices, say n_1, n_2, and n_3, which give the same triple

$$A_{n_1} = A_{n_2} = A_{n_3}, \qquad B_{n_1} = B_{n_2} = B_{n_3}, \qquad C_{n_1} = C_{n_2} = C_{n_3}$$

Since the equation $A_{n_1}x^2 + B_{n_1}x + C_{n_1} = 0$ has only two roots, two of the numbers ξ_{n_1}, ξ_{n_2}, ξ_{n_3} must be equal; with proper naming, they can be taken to be ξ_{n_1} and ξ_{n_2}, where $n_1 < n_2$ If $n_2 - n_1 = h$, then $\xi_{n_1+h} = \xi_{n_1}$, and

$$\xi_{n_1+h+1} = \frac{1}{\xi_{n_1+h} - [\xi_{n_1+h}]} = \frac{1}{\xi_{n_1} - [\xi_{n_1}]} = \xi_{n_1+1},$$

$$\xi_{n_1+h+2} = \frac{1}{\xi_{n_1+h+1} - [\xi_{n_1+h+1}]} = \frac{1}{\xi_{n_1+1} - [\xi_{n_1+1}]} = \xi_{n_1+2},$$

and in general, $\xi_{k+h} = \xi_k$ for $k \geq n_1$ Thus the ξ_k's are eventually periodic Since each a_k is determined exclusively by the corresponding ξ_k, the same is true of the a_k's, and the proof is complete

The relation (27) is of course valid for general ξ, although it was used above only when ξ is a quadratic irrationality It provides a proof of the following assertions

THEOREM 9–6 *If p_k/q_k is a convergent of the continued fraction expansion of ξ, then*

$$\xi - \frac{p_k}{q_k} = \frac{(-1)^k}{q_k(q_k\xi_{k+1} + q_{k-1})} \tag{28}$$

A fortiori, $$\frac{1}{q_k(q_k + q_{k+1})} < \left|\xi - \frac{p_k}{q_k}\right| < \frac{1}{q_kq_{k+1}}$$

and $$\left|\xi - \frac{p_k}{q_k}\right| < \frac{1}{q_k^2}. \tag{29}$$

As a partial converse, we have

THEOREM 9–7 *If* $$\left|\xi - \frac{p}{q}\right| < \frac{1}{2q^2}, \tag{30}$$

then p/q is a convergent of the continued fraction expansion of ξ

Proof: If p/q is adjacent to r/s in F_q, then the end point of $R_q(p, q)$ lying between p/q and r/s is the mediant

$$\frac{p+r}{q+s},$$

and
$$\left|\frac{p+r}{q+s} - \frac{p}{q}\right| = \left|\frac{qr-ps}{q(q+s)}\right| = \frac{1}{q(q+s)} \geq \frac{1}{2q^2}.$$

Hence, if (30) holds, either

$$\frac{p}{q} < \xi < \frac{p+r}{q+s} < \frac{r}{s} \quad \text{or} \quad \frac{r}{s} < \frac{p+r}{q+s} < \xi < \frac{p}{q},$$

and $\xi \in R_q(p, q)$, so that p/q is a best approximation, and therefore a convergent, to ξ.

PROBLEMS

1. Below is an outline of a proof that the expansion of $\sqrt{d}$ (d a positive nonsquare integer) is periodic after a_0. Fill in all details. (If $\alpha = r + s\sqrt{d}$, where r and s are rational, then $\bar{\alpha} = r - s\sqrt{d}$.)

Put $\xi = \sqrt{d} + [\sqrt{d}]$. Then $-1 < \bar{\xi} < 0$, and from the equation

$$\xi_k = a_k + \frac{1}{\xi_{k+1}}$$

it follows that $-1 < \bar{\xi}_k < 0$ for $k \geq 1$. This in turn shows that $a_k = [-1/\bar{\xi}_{k+1}]$. Now suppose that the periodicity of $\{\xi_k\}$ begins when $k = n$, and that the period is of length h, so that $\xi_n = \xi_{n+h}$. It follows that $a_{n-1} = a_{n+h-1}$, and hence that $\xi_{n-1} = \xi_{n+h-1}$, so that $\{\xi_k\}$ is periodic from the beginning.

2. It is a consequence of Theorem 8-1 that if ξ is irrational, then to each positive integer t there corresponds at least one pair of integers x, y such that

$$\left|\xi - \frac{y}{x}\right| < \frac{1}{tx}, \qquad 1 \leq x \leq t.$$

Show that this becomes false, for any irrational ξ and infinitely many t, if the second inequality above is replaced by $1 \leq x \leq t/2$. [*Hint:* Take $t = q_k + q_{k+1}$, and use Theorems 9-7 and 9-6.]

9-5 Application to Pell's equation

THEOREM 9-8. *If N and d are integers with $d > 0$ and $|N| < \sqrt{d}$, and d is not a square, then all positive solutions of the Pell equation*

$$x^2 - dy^2 = N \tag{31}$$

are such that x/y is a convergent of the continued fraction expansion of $\sqrt{d}$

Proof Suppose that $x + y\sqrt{d}$ is a positive solution of (31) Then, if N is positive,

$$0 < x - y\sqrt{d} = \frac{N}{x + y\sqrt{d}} < \frac{\sqrt{d}}{x + y\sqrt{d}} = \frac{1}{\frac{x}{\sqrt{d}} + y} = \frac{1}{y\left(\frac{x}{y\sqrt{d}} + 1\right)}$$

Since $x/y > \sqrt{d}$, we have

$$\left| \sqrt{d} - \frac{x}{y} \right| < \frac{1}{2y^2} \tag{32}$$

If N is negative, we deduce from the equation

$$y^2 - \frac{x^2}{d} = \frac{-N}{d}$$

the relations

$$0 < y - \frac{x}{\sqrt{d}} = \frac{-N/d}{y + \frac{x}{\sqrt{d}}} < \frac{1}{y\sqrt{d} + x} = \frac{1}{x\left(1 + \frac{y\sqrt{d}}{x}\right)},$$

and

$$\left| \frac{1}{\sqrt{d}} - \frac{y}{x} \right| < \frac{1}{2x^2} \tag{33}$$

Now if

$$\sqrt{d} = a_0 + \frac{1}{a_1 + \quad},$$

then

$$\frac{1}{\sqrt{d}} = 0 + \frac{1}{a_0 + a_1 + \quad \frac{1}{\quad}},$$

so that the convergents of the continued fraction expansion of $1/\sqrt{d}$ are $0/1$ and the reciprocals of the convergents of the continued fraction expansion of $\sqrt{d}$ Using this, the inequalities (32) and (33), and Theorem 9–7, we have the result

This theorem provides an effective method of determining all integers N, numerically smaller than $\sqrt{d}$, for which equation (31)

is solvable, for it happens that the sequence $\{p_k^2 - dq_k^2\}$ is eventually periodic, and consequently only finitely many values of k need to be examined. To see this, put $\xi = \sqrt{d}$ and

$$\sqrt{d} = \frac{p_{k-1}\xi_k + p_{k-2}}{q_{k-1}\xi_k + q_{k-2}}. \tag{34}$$

Solving for ξ_k and rationalizing the denominator, we can write

$$\xi_k = \frac{\sqrt{d} + r_k}{s_k},$$

where r_k and s_k are rational numbers. Substituting this back into (34), and replacing k by $k + 1$ throughout, we have

$$\sqrt{d} = \frac{p_k(\sqrt{d} + r_{k+1}) + p_{k-1}s_{k+1}}{q_k(\sqrt{d} + r_{k+1}) + q_{k-1}s_{k+1}},$$

or

$$(q_k r_{k+1} + q_{k-1}s_{k+1} - p_k)\sqrt{d} - (p_{k-1}s_{k+1} + p_k r_{k+1} - q_k d) = 0.$$

The rational and irrational parts must separately be zero, so

$$q_k r_{k+1} + q_{k-1}s_{k+1} = p_k, \qquad p_k r_{k+1} + p_{k-1}s_{k+1} = q_k d.$$

The determinant of this system is $q_k p_{k-1} - q_{k-1}p_k = (-1)^k$, so that

$$\begin{aligned} r_{k+1} &= (-1)^k(p_k p_{k-1} - q_k q_{k-1} d), \\ s_{k+1} &= (-1)^k(q_k^2 d - p_k^2). \end{aligned} \tag{35}$$

Now the numbers r_k and s_k are uniquely determined by ξ_k; since $\{\xi_k\}$ is eventually periodic, the same is true of $\{s_k\}$, and the eventual periodicity of $\{p_k^2 - dq_k^2\}$ follows from the second equation of (35).

The discussion of Pell's equation with $N = \pm 1$, in Chapter 8, had the serious drawback that no effective method was given for finding the fundamental solution, nor even of deciding when one exists for $N = -1$. The results obtained above entirely clarify these points: the first solution encountered, being the smallest, is the fundamental solution, and the equation $x^2 - dy^2 = -1$ is solvable if and only if a solution exists among the convergents to $\sqrt{d}$ up to the end of the second period. (For $\{s_k\}$ becomes periodic at the same point as $\{\xi_k\}$, and $\{(-1)^{k-1}s_k\}$ has period at most twice that of $\{s_k\}$.) It can be shown by the method sketched in Problem 4, below, that $s_k = 1$ for the first time at the end of the first period, so that the preceding

convergent is the fundamental solution of one of the equations, if for this convergent $p_k^2 - dq_k^2 = 1$, then the equation $x^2 - dy^2 = -1$ is unsolvable, while if $p_k^2 - dq_k^2 = -1$, the convergents preceding ends of periods are alternately solutions for $N = -1$ and $N = 1$

PROBLEMS

1 For what N with $|N| < \sqrt{7}$ is the equation $x^2 - 7y^2 = N$ solvable?

2 Show that the numbers r_k and s_k defined in this section are positive integers [*Hint* Use equation (28)]

3 Find the fundamental solution of $x^2 - 95y^2 = 1$, of $x^2 - 74y^2 = 1$

4 (a) Using Problem 1, Section 9–4, show that if the length of the period in the expansion of $\sqrt{d}$ is h, then $s_h = 1$, and hence that

$$p_{h-1}^2 - dq_{h-1}^2 = (-1)^h$$

Thus $x^2 - dy^2 = -1$ is solvable if h is odd

(b) Using the fact that the numbers $\xi, \xi_1, \quad , \xi_{h-1}$ are distinct, show that $s_k > 1$ if $1 \leq k \leq h - 1$ Deduce that the equation $x^2 - dy^2 = -1$ is solvable only if h is odd

9–6 Equivalence of numbers. Because each element of the sequence $\{\xi_k\}$ depends only on the preceding one, and because the defining rule

$$\xi_k = [\xi_k] + \frac{1}{\xi_{k+1}}$$

is the same for all $k \geq 1$, it is clear that if

$$\xi = a_0 + \cfrac{1}{a_1 + \quad \cfrac{1}{a_{n-1} + \cfrac{1}{\xi_n}}} = a_0 + \cfrac{1}{a_1 + \quad},$$

then

$$\xi_n = a_n + \cfrac{1}{a_{n+1} + \quad}$$

If we are interested in the possibility of finding infinitely many solutions of the inequality

$$\left|\xi - \frac{p}{q}\right| < \frac{1}{cq^2},$$

equation (28) shows that we need only examine the numbers ξ_k with large k *For this reason, we shall term two irrational numbers* ξ

and ξ' *equivalent* if, for some j and k, $\xi'_j = \xi_k$. Then $\xi'_{j+m} = \xi_{k+m}$ for $m \geq 0$, and by the above remark, this means that if

$$\xi = a_0 + \frac{1}{a_1 +} \cdots \frac{1}{a_{k-1} +} \frac{1}{a_k +} \frac{1}{a_{k+1} +} \cdots$$

then

$$\xi' = b_0 + \frac{1}{b_1 +} \cdots \frac{1}{b_{j-1} +} \frac{1}{a_k +} \frac{1}{a_{k+1} +} \cdots,$$

so that

$$\xi = \frac{p_{k-1}\xi_k + p_{k-2}}{q_{k-1}\xi_k + q_{k-2}}, \qquad \xi' = \frac{p'_{j-1}\xi_k + p'_{j-2}}{q'_{j-1}\xi_k + q'_{j-2}}.$$

THEOREM 9-9. *Two irrational numbers ξ and ξ' are equivalent, in the sense that their continued fraction expansions are identical from some points on, if and only if there are integers A, B, C, and D such that*

$$\xi' = \frac{A\xi + B}{C\xi + D}, \qquad \text{where } AD - BC = \pm 1. \tag{36}$$

Proof: Eliminating ξ_k from the equations preceding the theorem gives

$$\frac{-q_{k-2}\xi + p_{k-2}}{q_{k-1}\xi - p_{k-1}} = \frac{-q'_{j-2}\xi' + p'_{j-2}}{q'_{j-1}\xi' - p'_{j-1}},$$

or

$$\xi = \frac{A\xi' + B}{C\xi' + D},$$

where

$$A = p_{k-1}q'_{j-2} - p_{k-2}q'_{j-1}, \qquad B = p_{k-2}p'_{j-1} - p_{k-1}p'_{j-2},$$
$$C = q_{k-1}q'_{j-2} - q_{k-2}q'_{j-1}, \qquad D = q_{k-2}p'_{j-1} - q_{k-1}p'_{j-2}.$$

A simple calculation shows that

$$AD - BC = (p'_{j-1}q'_{j-2} - p'_{j-2}q'_{j-1})(p_{k-1}q_{k-2} - p_{k-2}q_{k-1}) = \pm 1.$$

To complete the proof, suppose that equation (36) holds. By replacing A, B, C, and D by their negatives if necessary, we may suppose also that $C\xi + D > 0$. Substituting the value of ξ from equation (26) into (36) gives

$$\xi' = \frac{a\xi_k + b}{c\xi_k + d}, \tag{37}$$

where

$$a = Ap_{k-1} + Bq_{k-1}, \qquad b = Ap_{k-2} + Bq_{k-2},$$

$$c = Cp_{k-1} + Dq_{k-1}, \qquad d = Cp_{k-2} + Dq_{k-2},$$

and

$$ad - bc = (AD - BC)(p_{k-1}q_{k-2} - p_{k-2}q_{k-1}) = \pm 1$$

By the inequality (29),

$$p_{k-1} = q_{k-1}\xi + \frac{\delta_{k-1}}{q_{k-1}}, \qquad p_{k-2} = q_{k-2}\xi + \frac{\delta_{k-2}}{q_{k-2}},$$

where

$$|\delta_{k-1}| < 1, \qquad |\delta_{k-2}| < 1$$

Hence

$$c = (C\xi + D)q_{k-1} + \frac{C\delta_{k-1}}{q_{k-1}}, \qquad d = (C\xi + D)q_{k-2} + \frac{C\delta_{k-2}}{q_{k-2}}$$

Since $C\xi + D$, q_{k-1}, and q_{k-2} are positive, and since $q_{k-1} > q_{k-2}$ and $q_k \to \infty$ with k, we have $c > d > 0$ for k sufficiently large But by Theorem 8–14, this means that a/c and b/d are adjacent in F_c, and from (37) and the fact that $\xi_k > 1$, it is seen that ξ' lies between a/c and b/d, and is closer to a/c than is the mediant $(a + b)/(c + d)$ It follows that $\xi' \in R_{c-1}(b, d)$ and $\xi' \in R_c(a, c)$, so that b/d and a/c are successive convergents of the continued fraction expansion of ξ'

$$a = p'_{j-1}, \qquad b = p'_{j-2}, \qquad c = q'_{j-1}, \qquad d = q'_{j-2}$$

But from

$$\xi' = \frac{p'_{j-1}\xi_k + p'_{j-2}}{q'_{j-1}\xi_k + q'_{j-2}} = \frac{p'_{j-1}\xi_j' + p'_{j-2}}{q'_{j-1}\xi_j' + q'_{j-2}}$$

it follows that $\xi_k = \xi_j'$, as was to be proved

In the course of the proof, the following useful fact emerged

THEOREM 9–10 *If a, b, c, and d are integers, and*

$$\xi = \frac{a\xi' + b}{c\xi' + d}, \qquad ad - bc = \pm 1, \qquad \xi' > 1, \quad c > d > 0,$$

then b/d and a/c are successive convergents of the continued fraction expansion of ξ, and ξ' is the corresponding complete quotient for suitable k,

$$a = p_{k-1}, \qquad b = p_{k-2}, \qquad c = q_{k-1}, \qquad d = q_{k-2}, \qquad \xi' = \xi_k$$

We shall use the symbol "$\cong$" to designate equivalence in the regular continued fraction sense. The notion of equivalence, together with equation (28), can be used to gain new insight concerning the Markov constant $M(\xi)$, which was defined in Section 8-4 as the upper limit of those numbers λ such that the inequality

$$\left|\xi - \frac{p}{q}\right| < \frac{1}{\lambda q^2}$$

has infinitely many solutions p, q. From (28), it is clear that $|q_k{}^2(\xi - p_k/q_k)|$ is approximately inversely proportional to ξ_k, so that $M(\xi)$ will probably have its smallest value for those ξ for which $a_k = 1$ for all large k. Now if

$$\xi = 1 + \frac{1}{1+}\,\frac{1}{1+}\cdots,$$

then

$$\xi = 1 + \frac{1}{\xi}, \qquad \xi = \frac{1+\sqrt{5}}{2}.$$

These remarks lead one to expect that the first part of the following theorem might be true.

THEOREM 9-11. *If* $\xi \cong \xi'$, *then* $M(\xi) = M(\xi')$. *If* $\xi \cong (1+\sqrt{5})/2$, *then* $M(\xi) = \sqrt{5}$. *If* ξ *is irrational and not equivalent to* $(1+\sqrt{5})/2$, *then* $M(\xi) \geq \sqrt{8}$. *If* $\xi \cong \sqrt{2}$, *then* $M(\xi) = \sqrt{8}$. *If* ξ *is not equivalent to either* $(1+\sqrt{5})/2$ *or* $\sqrt{2}$, *and is irrational, then* $M(\xi) \geq 17/6$.

Proof: By (28),

$$M(\xi) = \limsup_{k\to\infty}\left(\xi_{k+1} + \frac{q_{k-1}}{q_k}\right).$$

Now

$$\xi_{k+1} = a_{k+1} + \frac{1}{a_{k+2}+}\cdots,$$

and

$$\frac{q_{k-1}}{q_k} = \frac{q_{k-1}}{q_{k-1}a_k + q_{k-2}} = \frac{1}{a_k + \dfrac{q_{k-2}}{q_{k-1}}} = \frac{1}{a_k +}\,\frac{1}{a_{k-1} + \dfrac{q_{k-3}}{q_{k-2}}} = \cdots$$

$$= \frac{1}{a_k +}\,\frac{1}{a_{k-1} +}\cdots\frac{1}{a_2 + \dfrac{q_0}{q_1}} = \frac{1}{a_k +}\cdots\frac{1}{a_1},$$

so that

$$M(\xi)=\limsup_{k\to\infty}\left\{\left(\frac{1}{a_k+}\,\frac{1}{a_{k-1}+}\quad\frac{1}{a_1}\right)+\left(a_{k+1}+\frac{1}{a_{k+2}+}\quad\right)\right\} \tag{38}$$

If $\xi'\cong\xi$, then $\xi_j'=\xi_k$ and $a_j'=a_k$ for all sufficiently large j and k for which $j-k$ has a suitable fixed value h If the convergents of ξ' are p_j'/q_j', then for such j and k the continued fraction expansions of q_{k-1}/q_k and q_{j-1}'/q_j' have the same partial quotients at the beginning, and the interval of agreement can be made arbitrarily long by choosing j and k sufficiently large Suppose that they agree in the first $l+1$ partial quotients, that $r_l/s_l(l=0,1,\quad,l)$ are the common convergents, and that

$$\frac{q_{k-1}}{q_k}=\frac{r_{l-1}\alpha_l+r_{l-2}}{s_{l-1}\alpha_l+s_{l-2}},\qquad\frac{q_{j-1}'}{q_j'}=\frac{r_{l-1}\alpha_l'+r_{l-2}}{s_{l-1}\alpha_l'+s_{l-2}}$$

Then using the fact that $[\alpha_l]=[\alpha_l']\geq 1$, we have

$$\left|\frac{q_{k-1}}{q_k}-\frac{q_{j-1}'}{q_j'}\right|=\frac{|\alpha_l-\alpha_l'|}{(s_{l-1}\alpha_l+s_{l-2})(s_{l-1}\alpha_l'+s_{l-2})}\leq\frac{1}{s_{l-1}^2},$$

so that

$$\lim_{\substack{k\to\infty\\ j-k=h}}\left\{\left(\xi_k+\frac{q_{k-1}}{q_k}\right)-\left(\xi_j'+\frac{q_{j-1}'}{q_j'}\right)\right\}=\lim_{\substack{k\to\infty\\ j-k=h}}\left(\frac{q_{k-1}}{q_k}-\frac{q_{j-1}'}{q_j'}\right)=0,$$

and so $M(\xi)=M(\xi')$

To prove the second assertion of Theorem 9–11, we need only notice that

$$M\left(\frac{1+\sqrt{5}}{2}\right)=\lim_{k\to\infty}\left\{\left(1+\frac{1}{1+}\quad\right)+\underbrace{\left(\frac{1}{1+}\,\frac{1}{1+}\quad\frac{1}{1}\right)}_{k\text{ terms}}\right\}$$

$$=\frac{1+\sqrt{5}}{2}+\frac{1}{(1+\sqrt{5})/2}=\sqrt{5},$$

by (38)

To prove the third part, we may suppose that $a_{k+1}\geq 2$ for infinitely many indices k If $a_{k+1}\geq 3$ for infinitely many k, it is clear that $M(\xi)\geq 3$ Since $\sqrt{8}<3$, we need only consider those ξ s for which a_k is either 1 or 2 for all large k If there are infinitely many 1's and 2 s, there are infinitely many values of k such that $a_k=1\ \ a_{k+1}=2$

But then, since the value of a continued fraction is always at least equal to its convergent with index 2,

$$a_{k+1} + \frac{1}{a_{k+2} + \cdots} \geq 2 + \frac{1}{a_{k+2} + \dfrac{1}{a_{k+3}}} \geq 2 + \frac{1}{2 + \dfrac{1}{1}} = \frac{7}{3}$$

and

$$\frac{1}{a_k + \dfrac{1}{a_{k-1} + \cdots \dfrac{1}{a_1}}} \geq \frac{1}{1 + \dfrac{1}{a_{k-1}}} \geq \frac{1}{1 + \dfrac{1}{1}} = \frac{1}{2},$$

so that

$$M(\xi) \geq \tfrac{7}{3} + \tfrac{1}{2} = \tfrac{17}{6} = 2.833\ldots > \sqrt{8}.$$

On the other hand, if $a_k = 2$ for all large k, then

$$\xi \cong 1 + \frac{1}{2 + \dfrac{1}{2 + \cdots}} = \sqrt{2},$$

and

$$M(\xi) = \lim_{k \to \infty} \left\{ \left(2 + \frac{1}{2 + \cdots}\right) + \underbrace{\left(\frac{1}{2 + \dfrac{1}{2 + \cdots \dfrac{1}{2}}}\right)}_{k \text{ terms}} \right\}$$

$$= (\sqrt{2} + 1) + (\sqrt{2} - 1) = \sqrt{8}.$$

To clarify the significance of Theorem 9–11, we make use of the concept of countability, introduced by G. Cantor. Let S be an arbitrary infinite set. If it is possible to establish a one-to-one correspondence between the elements of S and the set of positive integers, then S is said to be *countable*. Another formulation of this requirement is that it should be possible to arrange the elements of S in a sequence having a first element, second element, and so on, in such a way that each element of S occurs only finitely far out in the sequence. The integers are countable, since every integer occurs in the sequence

$$0, 1, -1, 2, -2, \ldots, n, -n, \ldots.$$

The rational numbers between 0 and 1 are also countable, although they cannot be arranged by size. A suitable sequence is given by

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \ldots,$$

in which the reduced fractions with denominators 2 are listed first, then those with denominators 3, etc On the other hand, the real numbers between 0 and 1 are not countable (i e , the set of such numbers is *uncountable*) For each such number is uniquely represented by an infinite decimal (which may consist exclusively of 0's from some point on, but not of 9's), and conversely Suppose that the set could be arranged in a sequence, say $a_1, a_2,$, and let the decimal expansions be

$$a_1 = 0\, a_{11}a_{12}a_{13} \quad ,$$

$$a_2 = 0\, a_{21}a_{22}a_{23} \quad ,$$

$$a_3 = 0\, a_{31}a_{32}a_{33} \quad ,$$

where the a_{ij} are digits Let $b = 0\, b_1b_2b_3$ be the real number determined according to the following rule for $j = 1, 2,$,

$$b_j = \begin{cases} 0 & \text{if } a_{jj} \neq 0 \\ 1 & \text{if } a_{jj} = 0 \end{cases}$$

Then since $b_j \neq a_{jj}$, it is clear that $b \neq a_j$, and since this is true for every j, b is not in the sequence $a_1, a_2,$, so that the sequence does not contain every real number in the interval $0 < x < 1$

If it can be shown that one set is countable, while another is not, then there must be some element of the second set which is not in the first Moreover, every subset of a countable set is countable

It is relevant to our present purpose to note that the quadruples of integers (A, B, C, D) such that $|AD - BC| = 1$ are countable For without the restriction, we have the larger set of all quadruples of integers, and these can be arranged in a sequence by first writing (0, 0, 0, 0), then all quadruples whose elements are 0 or ± 1, then those whose elements are 0, ± 1 or ± 2, etc It follows from Theorem 9–9 that the set of numbers equivalent to a fixed number is countable, and it follows easily from this that the set of numbers equivalent to any of a fixed countable set of numbers is itself countable

Theorem 9–11 contains the first two of an infinite sequence of assertions about the values less than 3 assumed by $M(\xi)$ Markov showed that there are only countably many such values, that their sole limit point is 3, and that each such value corresponds precisely to the set of numbers equivalent to a certain quadratic irrationality

There is no really simple proof known. For later purposes, we show that these results cannot be extended to values $M(\xi) \geq 3$.

THEOREM 9-12. *There are uncountably many numbers ξ such that $M(\xi) = 3$.*

Proof: Let $r_1, r_2, \ldots$ be a strictly increasing sequence of positive integers, and let

$$\xi = \underbrace{1 + \frac{1}{1+} \cdots \frac{1}{1+}}_{r_1} \frac{1}{2+} \frac{1}{2+} \underbrace{\frac{1}{1+} \cdots \frac{1}{1+}}_{r_2} \frac{1}{2+} \frac{1}{2+} \frac{1}{1+} \cdots, \tag{39}$$

where there are r_1 partial quotients 1, then two 2's, then r_2 1's, then two 2's, then r_3 1's, etc. Thus two blocks consisting entirely of 1's are always separated by two 2's, and the blocks of 1's become longer as we move out in the sequence. Let

$$\beta_k = \xi_{k+1} + \frac{q_{k-1}}{q_k} = \left(a_{k+1} + \frac{1}{a_{k+2}+} \cdots\right) + \left(\frac{1}{a_k+} \frac{1}{a_{k-1}+} \cdots \frac{1}{a_1}\right).$$

If we choose k so that $a_{k+1} = 1$, then clearly $\xi_{k+1} < 2$, $q_{k-1}/q_k < 1$, and $\beta_k < 3$. If k runs through a sequence of indices such that $a_{k+1} = a_{k+2} = 2$, then

$$\beta_k = \left(2 + \frac{1}{2+} \frac{1}{1+} \frac{1}{1+} \cdots\right) + \left(\frac{1}{1+} \frac{1}{1+} \cdots \frac{1}{1}\right)$$

$$\to 2 + \frac{1}{2 + (\sqrt{5} - 1)/2} + \frac{\sqrt{5} - 1}{2} = 3,$$

while if k runs through a sequence of indices for which $a_k = a_{k+1} = 2$, then

$$\beta_k = \left(2 + \frac{1}{1+} \frac{1}{1+} \cdots\right) + \left(\frac{1}{2+} \frac{1}{1+} \frac{1}{1+} \cdots \frac{1}{1}\right)$$

$$\to 2 + \frac{1}{(1 + \sqrt{5})/2} + \frac{1}{2 + (\sqrt{5} - 1)/2} = 3.$$

Hence $M(\xi) = \limsup \beta_k = 3$.

To complete the proof, it is required to show that the set of inequivalent ξ's defined as in (39) is not countable. Now ξ and ξ' are

equivalent if and only if the sequences $r_1, r_2, \ldots$ and $r_1', r_2', \ldots$ associated with them are identical from some points on, so that we can *transfer the notion of equivalence from the numbers* ξ and ξ' *to* the sequences $\{r_k\}$ and $\{r_k'\}$ Suppose that the inequivalent sequences among all the increasing sequences of positive integers can themselves be arranged in a sequence, say $R_1, R_2, \ldots$, where R_i stands for the sequence $r_{i1}, r_{i2}, \ldots$, with $r_{i1} < r_{i2} < \ldots$. With proper naming we can suppose that R_1 is the sequence $1, 2, 3, \ldots$ of all positive integers in order If $i > 1$, R_i is not equivalent to R_1, and *there are therefore infinitely many positive integers not included in it.*

For $i > 1$ let $S_i = \{s_{ik}\}$ be the sequence complementary to R_i, that is, the positive integers, ordered by size, which do not occur in R_i Each S_i is an infinite sequence Now define a sequence T as follows Pick t_1 in S_2, and then successively choose $t_2, t_3, \ldots$ so that

$$t_1 \in S_2, \qquad t_1 < t_2 \in S_3,$$

$$1 + t_2 < t_3 \in S_2, \qquad t_3 < t_4 \in S_3, \qquad t_4 < t_5 \in S_4,$$

$$1 + t_5 < t_6 \in S_2 \qquad t_6 < t_7 \in S_3, \qquad t_7 < t_8 \in S_4, \qquad t_8 < t_9 \in S_5, \ldots$$

From this scheme it is apparent that T is an increasing sequence of integers, infinitely many of which are contained in S_k, and therefore not contained in R_k, for arbitrary $k \geq 2$ Hence T is certainly not equivalent to any of $R_2, R_3, \ldots$ Since each element $t_3, t_6, t_{10}, \ldots$ of T which lies in S_2 exceeds its predecessor in T by more than one, *T is also not equivalent to R_1* Hence *T is not equivalent to any R_k,* contrary to the hypothesis that the sequence $\{R_k\}$ contains an element equivalent to any increasing sequence of positive integers

PROBLEMS

1 Are the numbers $\sqrt{5}$ and $(1 + \sqrt{5})/2$ equivalent? What about $\sqrt{3}$ and $(1 + \sqrt{3})/2$?

2 Show that if ξ is irrational, then at least one of any two consecutive convergents to ξ satisfies the inequality

$$\left|\xi - \frac{p}{q}\right| < \frac{1}{2q^2}$$

REFERENCES

Section 9–6

Markov's work appears in *Mathematische Annalen* (Leipzig) **15**, 381–407 (1879) and **17**, 379–400 (1880). A quite different treatment was given by J. W. S. Cassels, *Annals of Mathematics* **50**, 676–685 (1949).

SUPPLEMENTARY READING

Chapters 1–5

DAVENPORT, H, *The Higher Arithmetic*, London Hutchinson & Co (Publishers), Ltd, 1952

DICKSON, L E, *History of the Theory of Numbers*, Washington Carnegie Institution of Washington, 1919 Reprinted, Chelsea Publishing Company, New York, 1950

GRIFFIN, H, *Elementary Theory of Numbers*, New York McGraw Hill Book Company, Inc, 1954

HARDY, G H, AND E M WRIGHT, *An Introduction to the Theory of Numbers*, 3rd edition, New York Oxford University Press, 1954

JONES, B W, *The Theory of Numbers*, New York Rinehart & Company, Inc, 1955

NAGELL, T, *Introduction to Number Theory*, New York John Wiley & Sons, Inc, 1951

ORE, Ø, *Number Theory and Its History*, New York McGraw Hill Book Company, Inc, 1948

STEWART, B M, *Theory of Numbers*, New York The Macmillan Company, 1952

VINOGRADOV, I M, *Elements of Number Theory*, translation of 5th Russian edition, New York Dover Publications, 1954

WRIGHT, H N, *First Course in Theory of Numbers*, John Wiley and Sons, Inc, New York, 1939

Chapter 6

HARDY AND WRIGHT, *op cit*

LANDAU, E, *Handbuch der Lehre von der Verteilung der Primzahlen*, vol 1, Leipzig Teubner Verlagsgesellschaft, 1909 Reprinted Chelsea Publishing Company, New York, 1953

LANDAU, E, *Vorlesungen uber Zahlentheorie*, vol 1 part 1 Leipzig S Hirzel Verlag 1927 Reprinted as *Elementare Zahlentheorie*, Chelsea Publishing Company, New York, 1950

Chapter 7

HARDY AND WRIGHT, *op cit*

LANDAU, E, *Vorlesungen uber Zahlentheorie*

Chapter 8

CAHEN, E, *Théorie des Nombres*, Paris Hermann & Cie, 1914 1924

LANDAU, E, *Vorlesungen uber Zahlentheorie*

NAGELL, T, *op cit*

Chapter 9

HARDY AND WRIGHT, *op. cit.*

KOKSMA, J. F., *Diophantische Approximationen*, Berlin: Springer-Verlag OHG, 1936. (Ergebnisse der Mathematik, vol. 4, no. 4.) Reprinted, Chelsea Publishing Company, New York, 1951.

PERRON, O., *Die Lehre von den Kettenbrüchen*, 3rd edition, Stuttgart: Teubner Verlagsgesellschaft, 1954. Second edition reprinted, Chelsea Publishing Company, New York, 1950.

ZUELLIG, J., *Geometrische Deutung Unendlicher Kettenbrüche*, Zürich: Orell Füssli Verlag, 1928.

LIST OF SYMBOLS

INDEX