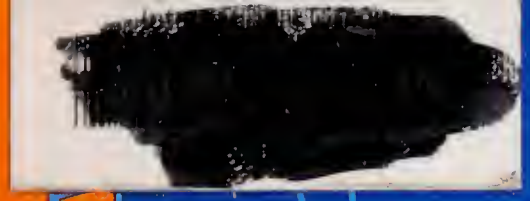


The History of **Mathematics**

A Brief Course

Roger **Cooke**

QA The history of mathematics
21
C649 Cooke, Roger
1997
Copy 2

The History of Mathematics

The History of Mathematics
A Brief Course

ROGER COOKE
University of Vermont

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This book is printed on acid-free paper. ♾

Library of Congress Cataloging in Publication Data:

Cooke, Roger, 1942–

The history of mathematics : a brief course / Roger Cooke.

p. cm.

"A Wiley-Interscience publication."

Includes bibliographical references and index.

ISBN 0-471-18082-3 (cloth : alk. paper)

1. Mathematics—History. I. Title.

QA21.C649 1997

510′.9—dc2 97-6046

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

"...all histories, to the extent that they contain a system, a drama, or a moral, are so much literary fiction... ." These wise words of George Santayana must give pause to anyone who hopes to write an interesting history on any subject. The operative word here is *interesting*; for without system, drama, or moral a book is certain to be both confusing and dull. Fortunately, the history of a subject with such a rich internal structure as mathematics must exhibit some kind of order, even though it may be only the order of a patchwork quilt. Although the past is, as C. S. Lewis said, a roaring cataract of hiccups and sneezes, it also contains, here and there, bits of comprehensible speech. My purpose in these pages is to exhibit those bits and draw connections between them for the reader to think about and perhaps argue with.

This book was begun after several years of teaching a general introduction to the history of mathematics aimed at mathematics and mathematics-education majors. As with every textbook I have used over three decades of university teaching, I could never be quite satisfied in my history courses with the approach the author had chosen. Each professor has a particular way of looking at a subject. That individual outlook makes the selection of material highly idiosyncratic. The following chapters are the result of my own reflection on the subject. Where pieces of original mathematical work can be described comprehensibly in a reasonable space, I prefer simply to present them, like photos in an album, with just a brief caption, so that the reader can appreciate the mathematics at first hand. I am not striving to give the reader a detailed chronological history of any part of mathematics. To do so in a first course (the only one most students will ever take, unfortunately) would require omitting the far more important element of *appreciation* that should (in my view) be at the heart of such a course.

It is easy to see that drastic principles of selection must be applied in order to provide a manageable amount of material for such a course. At one time the manuscript of this book was 50 percent larger than its present extent. One by one topics had to be eliminated—continued fractions, orthogonal expansions, Bourbaki, Plimpton 322, applications to thermodynamics, all sorts of gossip, all attempts at fine detail after the year 1700, nearly all biography. Each cut was painful. By what principles can a selection be made? If there were general agreement on a ranking of mathematical work or mathematicians by quality, one could simply start at the top and quit when the book was full. However, the importance of a mathematical topic is not always well defined, and there are other considerations to be kept in

mind when writing a book. The most important of these is the potential reading audience. What do the people in that audience most need to know? Of nearly equal importance is the fact that only the haziest idea of the real achievements of modern mathematics can be made comprehensible to undergraduates; without knowledge of the mathematical details, the student will not fully understand, but will soon fully forget, even the clearest summary description. These considerations have often dictated that the amount of space devoted to a mathematician or a mathematical topic is not proportional to his/her/its importance. Thus the reader will find more material on Pappus than on Euler in this book, even though there is no doubt that the latter was a much better mathematician than the former.

For the most part I have omitted biographies entirely and concentrated instead on the mathematics that was done and what it means in relation to other mathematics and other areas of human endeavor. This aim has led me to devote space to minor mathematicians, such as Boethius and Maria Gaetana Agnesi, and to some people who were not really mathematicians at all, such as Gerbert and Benjamin Banneker, while the work of some very good mathematicians, such as Simon Stevin, Hermann Amandus Schwarz and Norbert Wiener, is not mentioned. In taking this approach I am conscious of being influenced by the great enjoyment I once experienced in reading Bertrand Russell's *History of Western Philosophy*. Russell thought it worthwhile to discuss Lord Byron, who was not a philosopher, in order to explain the influence of the Romantic Movement on philosophy, and he explicitly stated that philosophical merit was not the basis on which text was allocated to a subject. The absence of biographies, though not a virtue in itself, does provide a wealth of topics for students to use for term papers. My own practice is to ask each student who wishes to write a biographical term paper to choose a notable mathematician and summarize that person's career, reading at least one original paper by the subject.

The principle of deciding what the reader needs to know, however, is still not a complete guide to the selection of material. What an author thinks the reader ought to know about such a vast amount of material is sure to be biased toward the familiar. I have learned by experience that it is very difficult to appreciate the importance of subject areas that lie beyond one's own complexity horizon (to use an elegant expression of John Allen Paulos). One can only make the effort to treat unfamiliar subject areas fairly and hope that the author's own interests will not be too apparent to the reader. Let me confess immediately that I find the following topics to be of stupefying dullness: (1) the bases for counting used by various peoples, (2) the evolution of the symbols for numbers from ancient India to modern Europe, and (3) pentagonal and hexagonal numbers. Having admitted my bias against these topics, I make a semi-apology for slighting them in the text that follows.

In the course of the writing I learned also that it is extremely difficult to give a concise summary of an area of work unless one is very familiar with that area. Two solutions to this problem naturally suggest themselves—leave the writing to the few mathematical giants who truly understand all these areas, or give a detailed presentation. The first alternative, besides leaving the author no book to write, has the further disadvantage that these mathematical giants never seem

to write for undergraduates; their intended audience always seems to be research mathematicians. The second, as I know from reading too many other books, produces long, tedious recitals of what was done, by whom, and when. The reader is presented with a huge agglomeration of facts without any unifying principle to assure that they will be remembered when the book is closed. It is, I believe, better to omit important topics than to produce a bleak landscape full of names and dates in a futile attempt to tell the whole story.

I have selected material that I consider interesting for its own sake, but nearly every piece of mathematics that is discussed has some larger significance than the mere fact that some clever person thought of it. I had originally thought of naming this book *Issues in the History of Mathematics*. At all stages I encourage the reader to ask why people were interested in the problems that mathematicians were solving and what consequences their solutions had for the further development of mathematics and its applications.

I wish to thank the reviewers whose comments have greatly improved the manuscript and express the earnest hope that its postpublication reviewers will be equally kind. The reviewers known to me are Joachim Lambek (McGill University), Millianne Lehmann (University of San Francisco), Richard L. Francis (Southeast Missouri State University), Frank J. Swetz (The Pennsylvania State University), G. G. Bilodeau (Boston College), and W. R. Wade (University of Tennessee). I owe a huge debt to the many historians of mathematics whom I know personally and through their works, especially my friend and coauthor V. Fred Rickey, who I believe also reviewed some of this book, although I never saw the review, and Milo Gardner, who has learned more about Egyptian arithmetic than I would have thought possible from the available documents. Inevitably some of these people will find that they disagree with some of my judgments and (what is worse) that I have made some errors. I welcome all corrections, and—provided they do not become too time-consuming—arguments over what is and is not a justified conclusion based on the facts. I am also grateful to the students at the University of Vermont who have patiently (I hope) learned (I hope) the history of mathematics in the courses I have taught over the past 15 years. I also wish to acknowledge the patience of a succession of editors, who have been extremely tolerant of my idiosyncrasies while guiding my writing with the interests of the reader in mind. The biggest debt of all is owed to my wife Cathie, who has patiently tolerated my mental absence as I sat before the home computers on which this book was written.

Roger Cooke

University of Vermont

June 1997

The History of Mathematics

PART I

Early Western Mathematics

In this first part of our study we shall look at the origins of mathematics and examine its progress in the world around the Mediterranean Sea from prehistoric times until the end of the Roman Empire. This study will provide a point of view and a basis for comparison with the independent development that took place in the Orient. Although we shall not develop this theme in any detail, you should notice the important role played by the Mediterranean Sea in fostering communication and commercial contacts among a large number of peoples speaking very distinct languages. The Mediterranean Sea borders Europe, Africa, and Asia, so that much of what we are calling "Western" mathematics has origins south and east of Europe. Even the Greeks, the quintessence of Western culture, were immigrants to the peninsula on which they lived when they made their great mark on the world. It may be this cosmopolitanism (a beautiful Greek word) that accounts for the uniqueness of the Greek contribution to mathematics. The organization of mathematics into a system of definitions, axioms, and theorems and the requirement of formal proof of results is without parallel anywhere else in human history. It is, however, only the summit of a large pyramid of knowledge whose lower levels were built by people who came before the Greeks. Those people also deserve to be remembered.

Our journey begins with the kind of mathematics that occurs spontaneously to human beings in the course of everyday life. We shall see how this mathematics becomes increasingly sophisticated as human society becomes more complex. Two facts stand out from the very beginning: (1) all societies need some mathematics; (2) those who create mathematics by solving mathematical problems nearly always develop the subject far beyond its practical value. We shall trace this development as one would follow the growth of a wave that ultimately breaks and crashes into the shore. The crest of this wave is the mathematics of the Hellenistic period and early Roman Empire. It is called "Greek mathematics" because it was written in Greek, even though most of it was discovered far from the mainland of Greece. After watching the wave break in the later Roman Empire, we shall leave the West and turn our attention to other "waves" in other parts of the world.

Chapter 1

Origins

We begin our study of the development of mathematics by trying to clarify the nature of the subject. What is mathematics and how does it arise? The first of these questions is a philosophical one, but must be addressed if we hope to make any sense out of the history of mathematics. The second question belongs really to the prehistory of mathematics, but our tentative answers to it provide the foundation for the historical study we are undertaking.

1.1 The Nature of Mathematics

In this section we shall survey some issues in the history of mathematics in order to free ourselves from the perspective of the school mathematics that we have all been taught. The pedagogical ordering of mathematical topics does not always follow the chronological order of their invention. In this discussion we are interested not in mathematical questions but in questions *about* mathematics.

Although twentieth-century mathematics encompasses hundreds of academic specialties, these specialties all began with two basic human activities. One of these activities is counting and measuring, which is at least as old as human government and commerce. The other is categorizing objects according to their shape, which is necessary wherever people manufacture tools and decorate their surroundings. Wherever people engage in agriculture, commerce, or industry instead of merely gathering or hunting the amount of food needed at the moment, they must do three things:

(1) Count separate units of things regarded as identical for the purpose (animals in a herd, pottery jars, coins, etc.)

(2) Measure continuous objects such as rope, land, wine, and bread, that is, find lengths, areas, volumes, and weights

(3) Make objects of simple geometric shape, such as houses (parallelepipeds), pottery jars (spheroids), and grain silos (cylinders), and lay out fields in the shape of rectangles or triangles.

The first and third of these activities lead to arithmetic and geometry, respectively. The second seems to involve both subjects, and it was precisely the attempt to mix the two that led to the first sophisticated research in mathematics, 2500 years ago.

Number and shape, the prototypes of formal mathematics, can be seen to share a common foundation based on the universal human tendency to compare things as like or different and rank them in order. At all times and in all places, wherever people take an interest in one another and in the world around them, they ask questions like, "How much bigger or smaller is my neighbor's land than mine?" "Who goes first in the ceremony?" "How can we apportion the harvest (or the representation in Congress) fairly?" "What shape land of a given size (area) can be enclosed with the smallest amount of fencing?" The solution of these problems requires counting and ordering, and arithmetic and geometry were developed partly in order to answer such practical questions. The theory of proportion, which is the result of reflecting on such questions of comparison, led to the first formal mathematics—the geometry and number theory of the early Greeks—and also to the first intellectual conflict between the continuous and the discrete (geometry and arithmetic) in the form of the problem of incommensurables and the paradoxes of Zeno. One of our main themes will be the development of the theory of proportion. This concept will be seen to occur in a large number of places in science, characterized mathematically by linear functions whose outputs are proportional to their inputs. It is the most important of the general concepts that have shaped the subsequent development of mathematics. We shall begin by looking at the relation between number and space as reflected in the problem of proportion. Other essential aspects of mathematics, such as the use of symbols and logical inference to state and prove propositions, will be taken up later.

1.1.1 Numbers, Space, Symbols, and Logic

The theory of proportion led to the first attempt to provide a unified foundation for mathematics. In the earliest written mathematical documents geometry hardly exists as a separate subject. The *shapes* of lines, surfaces, and solids are of interest only because they determine the method by which the *size* (length, area, or volume) of an object is to be found, and size is expressed by number. Thus arithmetic and geometry are not coequals at the origins of mathematics. Number is supreme. In early mathematical documents lengths, areas, volumes, and weights are measured using whole numbers and fractions. The use of fractions is the only thing that distinguishes measuring (say, the volume of a grain silo) from counting (say, a flock of sheep). It was probably taken for granted that the whole numbers, from which fractions are naturally derived, are adequate for the task of studying geometry; and indeed, from a practical point of view, they are: all the numbers ever recorded on measuring instruments have been rational numbers. Within human intuition, however, lay certain assumptions about the nature of the continuous entities that geometry deals with. What brought these assumptions to light was logic: the attempt to make geometry and arithmetic into deductive systems in Greece during

the fifth century B.C.E.[1] In the light of logic the numbers known at the time turned out to be inadequate for formulating the intuitive idea that a line is continuous. The fundamental tool used to compare lengths, the *common measure* of two objects, turned out to be nonexistent in the case of some common pairs of line segments, such as the sides and diagonals of regular polygons. The result was a logical difficulty known as the *problem of incommensurables*. The way in which this problem was solved is highly interesting, both for its own sake and because the success of the resulting geometric theory of proportion banished discrete concepts from "official" geometry for a long time.

The term "official" geometry is being used here to mean geometry as a deductive system based on the *Elements* of Euclid. This kind of geometry is difficult, and intuition suggests many problems whose resolution is nearly impossible using Euclidean principles. It is known, however, that a less formal kind of geometry, in which discrete concepts were freely used, existed side by side with the more restricted official version for centuries. Even Archimedes, who is acknowledged as the greatest of the ancient mathematicians, allowed himself the luxury of discovering his results by thinking of a plane region as a stack of line segments or a solid region as a stack of plane figures.

Numbers were used for more than just measuring space, however, as we learn by examining the practice of algebra by Hindu, Chinese, and Islamic mathematicians. In these early algebra problems the emphasis was on finding unknown numbers from certain given properties. The intrinsic properties of numbers were studied very little, and in many cases the solutions were illustrated geometrically. What we think of as the essence of algebra today—the use of equations to determine unknown numbers represented by symbols—was a rather late development in Europe, although it was present from early times in China. Three thousand years before algebra became prominent in Europe the mathematicians of the Akkadian period in what is now Iraq had rules for finding unknown numbers given certain information about those numbers. Nowadays this information would be coded as equations, and the rules that were followed to solve such problems are logically equivalent to our formulas for solving equations. The role of symbolism in this early "algebra" was very restricted compared to its present role, however. Although the words for *length* and *width* may have been used abstractly as symbols to represent unknown numbers in a problem, there were no such symbols to represent the *data* in a generic problem, such as a and b in the generic linear equation $ax + b = 0$. Consequently the general method of solution could not be expressed as a formula, but had to be explained by examples.

Some enlargement in the sphere in which symbols were used occurred in the writings of the third-century Greek mathematician Diophantus of Alexandria, but the same defect was present as in the case of the Akkadians. This nonsymbolic form of algebra was inherited by the Islamic world and further developed, again without extensive use of symbols. During the European Renaissance both arith-

[1] The traditional notations for eras, B.C. (before Christ) and A.D. (anno Domini) are gradually being replaced in scholarly work by B.C.E. (before the Christian era) and C.E. (Christian era). This change has come about because the phrase *anno Domini* (year of the Lord) assumes a specifically Christian doctrine.

metic and geometry were recast in algebraic terms using essentially the notation of today, and a third basic element of modern mathematics arose as a result: the use of symbols to represent unknown or variable quantities and the isolation of the equation as an important object of study (as opposed to a problem that can be stated as an equation). Thus it seems that the power of symbolism appeared surprisingly late in mathematics, considering that symbolism itself, in the form of written language, is very old.

1.2 The Origins of Mathematics

The prehistory of mathematics (mathematics invented before any written texts known to us) is an imaginative reconstruction based on information from many different sources. Each of the sources suggests, but does not conclusively demonstrate, something about the way in which mathematics arose as a human activity. Five such sources, what they tell us about the origins of mathematics, and their possible deficiencies are discussed below.

1.2.1 Animal Psychology

Certain ways of coping with the problems of life that may be called "mathematical" are shared by human beings and other mammals and birds, namely, distinguishing numbers and shape, the fundamental elements of arithmetic and geometry. Establishing exactly what animals are capable of in these areas is a subtle business. There have been claims of horses that "count" by scratching the ground, dogs that make sounds alleged to be human speech, and apes that use American Sign Language. It has been conclusively established, however, that horses do not really count in this sense; in every case that has been investigated the horse has been trained to scratch the ground and watch its trainer attentively for a signal to stop. As for the "talking" dog, it has been trained to make a certain sequence of sounds in response to a fixed cue; there is no question of its attaching any meaning to the phrases it produces. The question of ape use of sign language is still in dispute. The trainers of the apes are enthusiastic in their belief that apes really do communicate, while skeptics who have seen videotapes of the performance have pointed out subtle cues given by the response of the trainer to correct signs by the ape.

The pitfalls of such research were carefully avoided in a series of experiments with birds conducted in the 1930s and 1940s by Prof. O. Koehler (1889–1974) of the University of Freiburg. Koehler kept the trainer isolated from the bird. In the final tests (after the birds had been trained), the birds were filmed automatically without any human beings present. Koehler found that parrots and ravens could learn to compare the number of dots (up to 6) on the lid of a hopper with a "key" pattern in order to determine which hopper contained food. They could make the comparison no matter how the dots were arranged, so that the only clue they could have was the abstract number of dots.

By a variety of such experiments Koehler was able to establish that birds can learn to associate the abstract *number* of spots on the key pattern with the lid

having the same abstract number of spots. Thus it appears that the ability to *use* numerical aspects of the world is among the potential abilities of birds, even though no one knows of any example of birds using this ability in the wild.

If arithmetic is of value to birds in laboratories, it seems clear that the ability to perceive shape (geometry) might be of value to an animal even in less artificial settings than a laboratory, and indeed the ability of animals to perceive shape has been very well documented. In his famous experiments on conditioned reflexes using dogs as subjects the Russian scientist Pavlov (1849–1936) taught dogs to distinguish ellipses of very small eccentricity from circles. He began by projecting a circle of light on the wall every time he fed the dog. Eventually the dog came to expect food (as shown by salivation) every time it saw the circle. When the dog was completely conditioned, Pavlov began to show the dog an ellipse in which one axis was twice as long as the other. The dog soon learned not to expect food when shown the ellipse. At this point the malicious scientist began making the ellipse less eccentric, and found, with diabolical precision, that when the axes were nearly equal (in a ratio of 8 : 9, to be exact) the poor dog went berserk.

Our point in telling this story is a simple one: certain aspects of reality that we may call arithmetical or geometric must be dealt with by all living organisms. Organisms possessing at least rudimentary cognitive abilities are therefore capable of learning to use these properties of the world about them. In particular, the perceptual ability needed to create mathematical concepts is not uniquely human. Parrots can learn to identify two collections on the basis of the number of elements they contain; dogs can learn to distinguish similar but not identical shapes; and, as we shall now show, pigeons can make associations between one event and another based on causality or likelihood. In the language of Pavlov's conditioned reflexes, inferences can be made on the basis of incomplete or partial reinforcement. This area—the study of sporadic or random phenomena—was one of the last human concepts to be mathematized.

In dealing with reality a knowledge of the likely consequences of an event is probably even more valuable than the ability to perceive numbers and shape. The attempt to systematize such knowledge has led to the concepts of causality and randomness. It was a long time before randomness could be effectively handled by mathematical methods, yet the concept of causality itself, the idea of associating one event with another, seems to be innate, even in animals. In a fascinating article entitled " 'Superstition' in the pigeon," which describes research using Pavlov's methods in a nondeterministic manner, B. F. Skinner (1904–1990) gave a model for understanding how such associations form in the mind, whether justified by the facts or not. He put hungry pigeons in a cage and attached a food hopper to the cage with an automatic timer to permit access to the food at regular intervals. The pigeons at first engaged in aimless activity when not being fed, but tended to repeat whatever activity they happened to be doing when the food arrived, as if they made an association between the activity and the arrival of food. Naturally the more they repeated a given activity, the more likely that activity was to be reinforced by the arrival of food. Since they were always hungry, it was not long before they were engaged full-time in an activity that they apparently considered an infallible food producer. This activity varied from one bird to another. One pigeon thrust its head

into an upper corner of the cage; another made long sweeping movements with its head; another tossed its head back; yet another made pecking motions toward the floor of the cage.

Those with a playful imagination may wish to construct a conversation among pigeons conditioned to different behaviors: how would they settle among themselves the relative food-producing power of turning counterclockwise in the cage versus thrusting one's head into an upper corner? The many difficulties people (even mathematicians) have in understanding and applying probability can be seen in microcosm in this example. To take just one illustration, the human body has a certain power of healing itself. Yet sick people, like hungry pigeons, try various methods of alleviating their misery. Like the automatic timer that eventually provides food to the pigeon, the human immune system often overcomes the disease. The consequence is a wide variety of nostrums said to cure a cold or arthritis. One of the triumphs of modern mathematical statistics is the establishment of reliable systems of inference to replace the inferences Skinner called "superstitious."

This tendency to make associations of the form "a causes b" has entered mathematics in the form of a relation between propositions: "a implies b." This relation is the glue that holds mathematics together. The correspondence between implication and cause is a philosophical issue. As an illustration, consider the statement "if a is true, then b is true." This statement is logically equivalent to "if b is false, then a is false." However, absolute truth or falsehood is not available in relation to the observed world. As a result, science must actually deal with propositions of the form, "if a is true, then b is *highly probable*." One cannot infer from this statement that "if b is false, then a is highly *im*probable." For example, if X is an adult male, then X is very probably a law-abiding citizen. One cannot validly infer, however, that if X is not a law-abiding citizen, then X is probably not an adult male.

In summary, the ability to use arithmetic, geometric, and probabilistic/causal notions is a vital part of the human perceptual apparatus. This much can be inferred from the observation of animals. Animals, however, are clearly not capable of formal mathematical reasoning; parrots do not prove theorems about prime numbers; dogs do not seek methods of finding the asymptotes of hyperbolas; and pigeons do not contemplate the strong law of large numbers. In this respect human beings are unique. What were the manifestations of mathematics among the earliest human beings? Animal psychology cannot answer this question, and so we turn to other sciences for help.

1.2.2 Archaeology

Very ancient animal bones have been found in Africa and Europe containing notches, strongly suggesting that some sort of counting procedure was being carried on at a very early date, although what exactly was being counted remains unknown and perhaps forever unknowable. One thing is clear to everyone, however: the notches were made by human beings. No natural phenomenon and no animal would be capable of making them. One such bone, named after the fishing

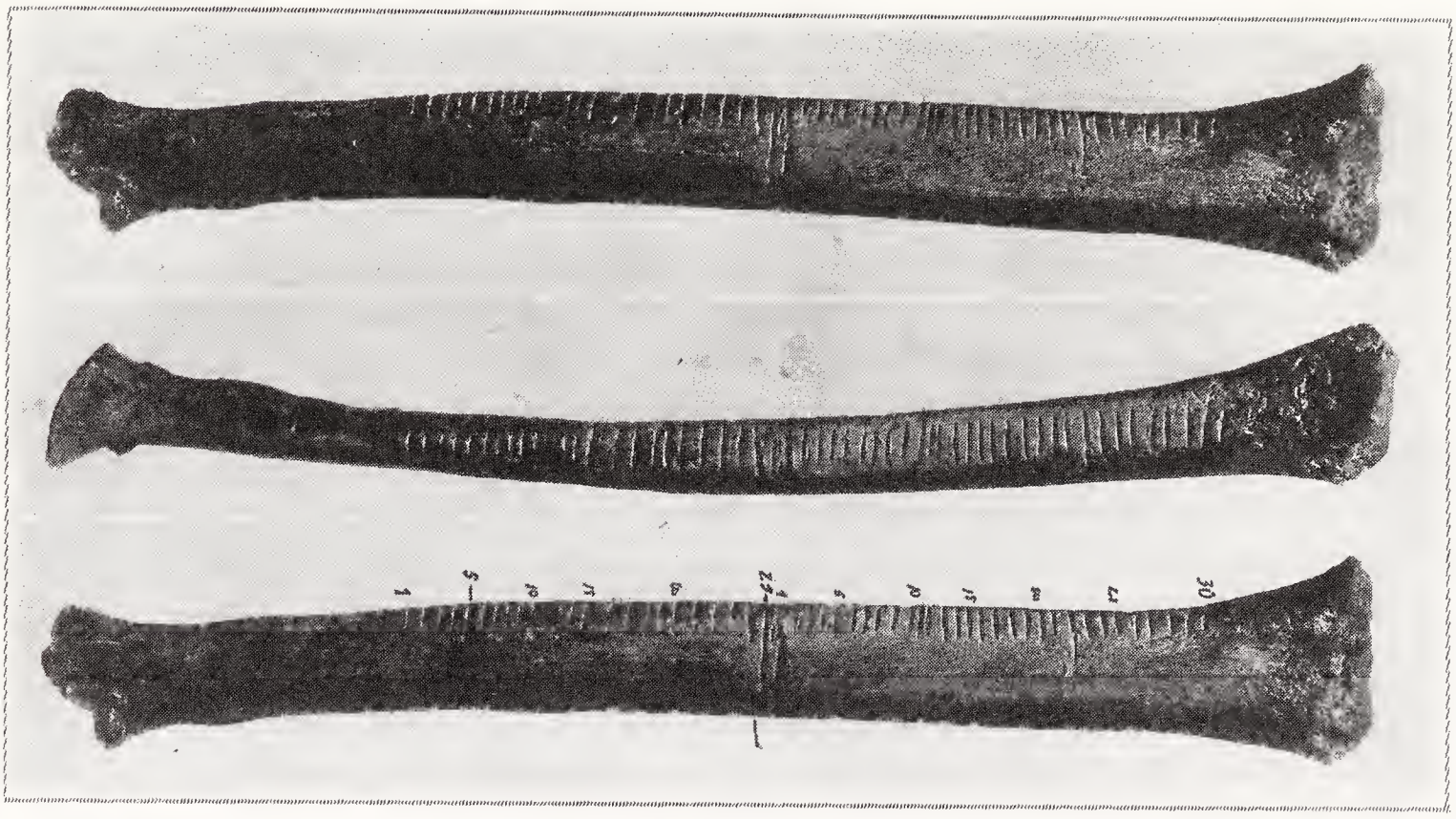


Figure 1.1: Paleolithic wolf bone, from the *Illustrated London News*, October 2, 1937. The Illustrated London News Picture Library.

village of Ishango on the shore of Lake Edward in Zaire where it was discovered, is believed to be between 8500 and 11,000 years old. The Ishango bone, which is now in the Musée d'Histoire Naturelle in Brussels, contains three columns of notches. One column consists of four series of notches containing 11, 21, 19, and 9 notches. Another consists of four series containing 11, 13, 17, and 19 notches. The third consists of eight series containing 3, 6, 4, 8, 10, 5, 5, and 7 notches, with larger gaps between the second and third series and between the fourth and fifth series. These columns present us with a mystery. Why were they put there? What activity was being engaged in by the person who carved them? Conjectures range from abstract experimentation with numbers to keeping score in a game. The bone could have been merely decorative, or it could have been a decorated tool.

Another such bone, the radius bone of a wolf (see Fig. 1.1), was discovered at Veronice (Czech Republic) in 1937. This bone was marked with two series of notches, grouped by fives, the first series containing five groups and the second six. Its discoverer, Dr. Karel Absolon (1887–1960), believed the bone to be about 30,000 years old, though other archaeologists thought it considerably younger. The people who produced this bone, though they hunted mammoth and rhinoceros in what is now Europe, were clearly a step above mere survival, since a human portrait carved in ivory was found in the same settlement along with a variety of sophisticated tools. Because of the grouping by fives, it seems clear that this bone was being used to count something. Even if the groupings are meant to be purely decorative, they point to a use of numbers and counting for a practical or artistic purpose.

Although the examples just given tell us nothing definite about the details of primitive mathematics, there are certain archaeological findings that can be woven

together with known facts about the physical world to support some highly interesting conjectures. To be specific, circular structures with various alignments that seem to be of astronomical significance are found all over the world. Stonehenge, near Salisbury, England, is the most famous example. Although we cannot be positive about the use of such circles (or even whether they had only one use, since different parts of Stonehenge were built at different times), it is clear that one possible use was in charting the motions of the sun and moon among the stars. If this currently fashionable interpretation of these circles is correct, it follows that one of the earliest stimuli to the use of both arithmetic and geometry comes from watching the sun and moon move among the stars. The properties of lines, circles, and angles come to be understood when stones must be arranged so as to point to significant places on the horizon (the sunrise and sunset points at the solstices, for example), and once the stones are in place, one can count the days between solstices and thereby compute a solar calendar.

Archaeologists have found significant astronomical alignments in most of these circles. One of the most interesting of these circles was discovered at Cahokia, Illinois in 1961. The Cahokia site contains more than one circle. Several arcs were discovered, but a portion of the site had been removed for construction before its archaeological significance was recognized. The archaeologist Warren L. Wittry found an arc, which he believes to be part of a full circle about 410 feet in diameter, containing post holes suitable for erecting tall poles evenly spaced and about 27.5 feet apart. Interestingly, there was no observation post at the exact center of the circle, but there was one about 5 feet east of the center, and among the post holes on the circle there are two so located that, viewed from the observation point, the sun rises exactly over them at the winter and summer solstice. Whatever poles may have been erected in these holes would have rotted away long ago (the construction is believed to be about 3000 years old).

As with all such circles, we shall probably never know with certainty whether the site was used as a calendar or merely for rituals (or even some purely mundane practical purpose—might it not have been the framework for a defensive fortress, for example?). If the site was used as a calendar, the scale is too small for a very precise determination of the solstices, but the precision may have been sufficient for the needs of a neolithic society.

Although we cannot be sure of our interpretations, it is plausible that the some of the first steps the human race took in the application of mathematics to understand the world were connected with astronomy. Artifacts such as the "sun circles" therefore do provide some confirmation that sunrises and full moons were among the earliest things that people counted. The important concept of time measurement, which depends on the fundamental relationship $distance = velocity \times time$, requires a standard motion that can be regarded as taking place at *constant* velocity to provide a definition of the quantity of time elapsed. The time elapsed is then directly proportional to the distance traversed; in fact, this property is the mathematical definition of constant velocity. Distance, of course, can be measured directly. Now of all the motions in the natural world that fit our intuitive notion of constant velocity, the regular progression of the stars from east to west in the sky every night is the most obvious standard. By comparing the

motions of the sun and moon with this star standard one can detect nonuniformities in their motions through the sky. This observation is the beginning of scientific astronomy, which came to be interwoven with geometry at an early date. This example shows that the concept of direct proportion, which we have mentioned above as a central problem of mathematics, is a key element in the measurement of time.

Archaeology can give us information not only on the origins of mathematics but also on the great variety of ways in which certain advanced cultures, now extinct, have dealt with mathematical concepts. The world's museums are full of mathematically related artifacts such as oracle bones and counting boards found in China, clay records of astronomical observations in Mesopotamia, and traveling "account books" (packets of knotted threads), known as *quipus*, used by the Incas of South America.

1.2.3 Anthropology

Studies of nontechnological societies can provide suggestions about the ways in which people create and use mathematics. Moreover, the information obtained from this source can be compared with the information available from archaeology to make plausible inferences about the earliest mathematics. Anthropological evidence, like archaeological evidence, comes to us filtered through the interpretations of its practitioners. It requires the further caveat that some of these societies may have learned their mathematics elsewhere rather than creating it from their own experience.

To give an example of the use of anthropological evidence of the origins of mathematics, we note that in rural areas in Africa calendars are more a matter of observation than a tool. As explained by the Rev. Dr. John S. Mbiti, the phases of the moon are so noticeable that the calendar adheres to them strictly, naming months after the prevailing weather conditions or the main activity taking place. (Contrast this strict lunarity with our "civil" months of 30 or 31 days.) Years tend to be counted according to cyclic human activities and weather conditions; they do not consist of a canonical number of days, as in Europe, Asia, and America.

From this description we can see that not every society *needs* precise mathematics. Only when the future must be planned in great detail is it necessary to perform a precise mathematical analysis of astronomical observations. Mbiti points out that the African concept of time is almost entirely concerned with the relations of earlier, later, and simultaneity in the important events of daily life. The mechanically measured time of a clock is needed only where it is necessary to coordinate the activities of large numbers of people who are not personally acquainted with one another. When people live in houses they have built themselves and manufacture the objects they use, there is no need to count days in order to compute rent, interest on debts, or the date of occupancy of a new dwelling. However, even the most uncomplicated life involves some counting, and it appears that the Africans discussed by Mbiti make a rough correlation between phases of the moon and their hunting activity and between seasons and agricultural activity.

As a general rule, people who engage in agriculture use mathematics only as a tool, not as an oracle. In making a particular judgment that involves mathematical questions they will place the mathematics in its proper perspective, weighing both numerical and nonnumerical considerations. This approach to problem-solving is taken in both low-technology and high-technology agriculture. For example, in modern America, when deciding when to harvest a crop of soybeans, the farmer knows that a penalty will be assessed at the grain elevator if the moisture content is above 17%. If the moisture content is above this level when harvest time arrives, the farmer will weigh the effects of delay on market prices, the possibility of future bad weather, and the fact that dry soybean pods tend to break open prematurely (causing the beans to drop onto the ground and be wasted) in order to decide the optimal time for harvest. These qualitative considerations reflect the same thought process described by Dr. Mbiti in connection with the calendar. Nevertheless, with the growth of technology the quantitative aspect of farming assumes an ever-larger role. Modern dairy farms provide measured nutrients individually adjusted to the needs of each cow and dispensed automatically by computer when the cow's collar sends a radio signal.

It must not be thought, however, that societies with small populations and little technology have no interest in mathematics beyond arithmetic. Prof. Marcia Ascher has assembled an impressive number of examples of rather arcane protomathematics among peoples who have very little technology. Mathematics, it seems, can be inspired by art as well as by science. The Bushoong people of Zaire make part of their living by supplying embroidered cloth, articles of clothing, and works of art to others in the economy of the Kuba chiefdom. In connection with this work (perhaps even as preparation for it) Bushoong children amuse themselves by tracing figures on the ground. The rule of the game is that a figure must be traced without repeating any strokes and without lifting the finger from the sand. In graph theory this problem is known as the *unicursal tracing problem.* It was thoroughly analyzed by the Swiss mathematician Leonhard Euler (1707–1783) in the eighteenth century in connection with the famous Königsberg bridge problem. According to Ascher, in 1905 some Bushoong children challenged the ethnologist Emil Torday (1875–1931) to trace a complicated figure without lifting his finger from the sand. Torday did not know how to do this, but he did collect several examples of such figures. The Bushoong children seem to learn intuitively what Euler proved mathematically: a unicursal tracing of a connected graph is possible if there are at most two vertices where an odd number of edges meet. The Bushoong children become very adept at finding such a tracing, even for figures as complicated as that shown in Fig. 1.2.

Any organized system of society puts everyone in a certain relation to other people. The codification of these relations requires a rudimentary analysis that may be called mathematical. The abstract study of relations is a part of set theory in modern mathematics, but we can see applications of it in nearly every society. Ascher gives the example of the Warlpiri of Australia, who assign each person to one of eight "sections" of the population. The sections are arranged in four ordered pairs, to which for convenience we shall assign numbers rather than names: $(1, 5)$, $(2, 6)$, $(3, 7)$, $(4, 8)$, and there is apparently strong social pressure to marry some-

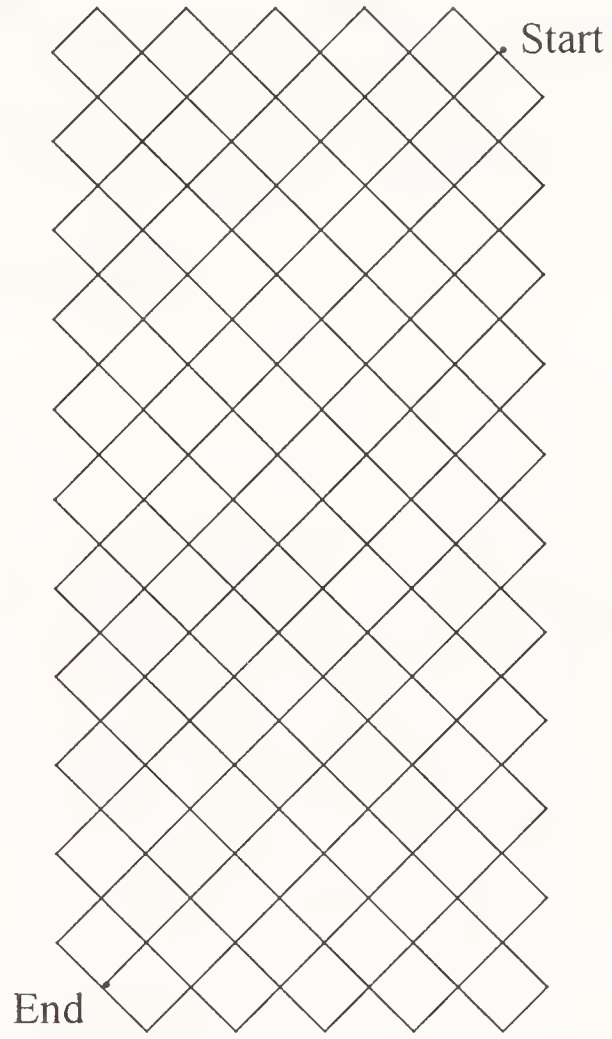


Figure 1.2: A graph for which a unicursal tracing is possible.

one from the group paired with one's own group. Children of these "preferred" marriages are assigned to a different section depending on the section to which their mother belongs, as depicted in the following two schemata:

$$1 \longrightarrow 3 \longrightarrow 2 \longrightarrow 4 \longrightarrow 1; \quad 5 \longrightarrow 8 \longrightarrow 6 \longrightarrow 7 \longrightarrow 5.$$

That is, children are assigned to the group following their mother's group in these schemata. One can thus deduce that, if only preferred marriages occur, a mother–daughter chain will stay among groups 1–4 if it begins there, and likewise that it will stay among groups 5–8 if it begins there. A father–son chain, on the other hand, will always alternate between groups 1–4 and groups 5–8, since men from groups 1–4 are expected to marry women from groups 5–8. The Warlpiri system is a striking example of the importance of the concept of relatedness in human society. As with the Warlpiri, such relations are usually family relations, and they were at one time of great importance in the politics of royal succession.

One obvious point of mathematical interest in the study of any culture is the way in which people count. In this connection, *words* are of primary importance. Every society, without exception, has a few proper names for small integers, from which the names of larger integers are formed by compounding. The integers with simple names indicate the *base* of the system. Among the bases that have been used by various peoples are 2, 3, 5, 10, 12, 20, and 60. Systems with a large base are probably a higher-order construction, superimposed on a smaller base. Studies

of the Andamans of Australia, for instance, revealed that these people would use their fingers to count to 10, saying, "ubatul, ikpor [one, two], anka [and this] ubatul ikpor, anka ubatul ikpor, anka ubatul ikpor, anka ubatul ikpor, ardura [that's all]." This last word was spoken while bringing the hands together. This example shows a system of base 2, with three and four essentially called "and [another] one two" and so forth up to 10. This example brings up a fourth source of knowledge about the origins of mathematics.

1.2.4 Language

Abstract mathematical concepts are usually given intuitive names having a vivid everyday meaning. We speak, for example, of *groups, rings*, and *fields* in algebra. As a more elementary example our word *line* is related to *linen* by more than a phonetic coincidence. Both come from the Latin word *linea*, which means *thread.* It is easy to see here the intuitive origin of the abstract concept of a line. As another example, take the words *angle* and *corner.* The former comes from the Latin word *angulus*, meaning *nook* or *corner*, while the latter is from the Latin word *cornus*, meaning *horn.* Again we can see what particular objects gave rise to the abstract mathematical concepts. This process occurs in any language that must express ideas that go beyond the ordinary functions of day-to-day living. The word *abstract* itself, for example, literally means *dragged away.* We don't really mean that abstract concepts are dragged away, of course; it is their concrete interpretation that has been removed when we use them.

This linguistic process can provide some clues to the origin of mathematics, for example, in the nursery rhyme that begins "Hickory, dickory, dock." This phrase preserves in a distorted form the words for "8, 9, 10" in a Celtic language spoken in Britain before there were any clocks for a mouse to run up. The most important ordinal numbers have recognizable relations to other words. Our word *first*, for example, is a shortened version of *foremost*, and *second* comes from a general root meaning *following.* As a method of finding out things about the origin of mathematics, linguistic considerations are somewhat limited, since number words are so old that they predate any language spoken today. The word *digit*, of course, means both *finger* and *positive integer less than ten*, but specific instances of digits (one, two, three, four, five, six, seven, eight, nine) have origins lost in dim antiquity. Were they perhaps names of specific fingers? Or was *two*, for example, the word for say *eyes* or *ears* in some language, and *five* the word for *hand*? The Russian word for *five* (*pyat'*) seems very close to the word for *metacarpus*[2] (*pyast'*); however, such phonetic coincidences are bound to occur in any large set of words. Although the names for numbers are similar across language groups, there is no consistent resemblance in English between the number names and names of objects that might plausibly be identified with them.

Despite the uncertainty of inference from linguistics, the study of the whole system of words used for counting, ordering, and dividing can be revealing. The

[2]For those who have not gone to medical school, the metacarpus is the portion of the hand excluding the fingers and thumb.

names people invent for abstract concepts provide a skeleton from which we can often discern the general shape of their thought processes. Partly for this reason, but mainly because we can conduct this kind of research without leaving our chairs, we shall follow up this idea in more detail in the exercises below.

1.2.5 Child Psychology

As our final example of a source of information about the origin of mathematics we note that our own thought processes and the learning experiences of children may provide a basis for conjectures about the creation of mathematics. Here again one must be very careful. We presume that mathematics has been created by adults, not children. Furthermore, *modern* children are not creating arithmetic and geometry; they are learning it, as did their teachers. Nevertheless, children clearly *learn* to use numbers and spatial concepts to the extent necessary to cope with ordinary life before they go to school. It is therefore of interest to see what situations require them to talk about numbers and space. The results of many studies have been summarized by Prof. Karen Fuson. A few of the results from observation of children at play and at lessons were as follows.

1. A group of nine children from 21 to 45 months was found to have used the word *two* 158 times, the word *three* 47 times, the word *four* 18 times, and the word *five* 4 times.
2. The children seldom had to count "one–two" in order to use the word *two* correctly; for the word *three* counting was necessary about half the time; for the word *four* it was necessary most of the time; for higher numbers it was necessary all the time.

One can thus observe in children the capacity to recognize groups of two or three without performing any conscious numerical process. This observation suggests that these numbers are primitive, while larger numbers are a conscious creation.

The most famous work on the development of mathematical concepts in children is due to Prof. Jean Piaget (1896–1980) of the University of Geneva, who wrote many books on the subject, some of which have been translated into English. Piaget divided the development of the child's ability to perceive space into three periods: a first period (up to about 4 months of age) consisting of pure reflexes and culminating in the development of primary habits; a second period (up to about one year) beginning with the manipulation of objects and culminating in purposeful manipulation; and a third period in which the child conducts experiments and becomes able to comprehend new situations. He categorized the primitive spatial properties of objects as proximity, separation, order, enclosure, and continuity. These elements are present in greater or less degree in any spatial perception. In the baby they come together at the age of about 2 months to provide recognition of faces. (The human brain seems to have some special "wiring" for recognizing faces.) The interesting thing about these concepts is that mathematicians recognize them as belonging to the subject of topology, an advanced branch of geometry that

developed in the late nineteenth and early twentieth centuries. It is an interesting paradox that the human ability to perceive shape depends on synthesizing various topological concepts; this progression reverses the pedagogical and historical ordering between geometry and topology.

Piaget pointed out that children can make topological distinctions (often by running their hands over models) before they can make geometric distinctions (3-year-old children may fail to distinguish between a square and an ellipse, for example). Discussing the perceptions of a group of 3–5-year-olds, Piaget claimed that the children had no trouble distinguishing between open and closed figures, surfaces with and without holes, intertwined rings and separate rings, and so forth, while the seemingly simpler relationships of geometry (distinguishing a square from an ellipse, for example) were not mastered until later.

1.3 Mathematics as a Human Endeavor

One fact revealed by a study of the history of mathematics is that all human cultures create and use mathematics to some degree, while in a few cultures it has risen to a very high degree of development. At present higher mathematics is the common heritage of the entire world, to which mathematicians of many nations have contributed for the last century. It is, along with music, a language understood by people from all parts of the world and the closest thing there is to a universal cultural phenomenon. Today there is a worldwide unity of mathematical standards expressed in mathematical journals published on every continent. To understand how this grand phenomenon came about is the ultimate object of our investigation. Having begun with the simplest protomathematics in this chapter, we shall advance to higher levels of mathematical creation in the later chapters.

There are several levels of mathematical sophistication. We shall distinguish three of these levels, forming a sort of pyramid. The use of numbers, shapes, and topological considerations to create tools and art and to engage in commerce is the bottom layer of the pyramid; all known societies exhibit this level of mathematical awareness. Just above it is the level at which mathematics is distinguished as a specialized subject of study. Within the group of societies whose mathematics is on this level there again seem to be certain intuitive universals. Rules for addition and subtraction and procedures equivalent to multiplication and division are found in all such societies. The Pythagorean theorem in its Euclidean form can be found in many cultures not known to have had contact with one another. This fact suggests that Euclidean geometry, at least, is a universal intuitive concept. Another intuitive approach is the attempt to "discretize" geometry, by regarding a plane figure as a stack of lines. This kind of intuitive formal mathematics existed and produced very similar results both in Europe and in Asia. At the very top of the pyramid is mathematics in the form of a logically organized *deductive system* in which conclusions are drawn according to rigorous logical rules rather than by intuition. This kind of mathematics was first developed by the ancient Greeks and grew alongside the intuitive variety, rather than displacing it. Although one would expect this approach to hinder rather than help the discovery of new results, it

was precisely the attempt to be clear about assumptions and logical in drawing conclusions that suggested a great many fertile fields for study. Noneuclidean geometry is the most obvious example of such a field. Although the intuitive ideas continued to exist into modern time, the ideal of logical precision persisted, and the continuing struggle to make mathematics clear and precise has led to new mathematical questions down to the present day. The advantages of this approach have been apparent to people throughout the world and have led to a broad consensus as to the proper way to state a mathematical proposition.

We cannot explore all of the history of mathematics in a one-semester course. Drastic cuts must be made in both the breadth and depth of the subject in order to produce a coherent and meaningful course. We shall first of all confine ourselves to the top two layers of the pyramid just described. We shall add nothing to what has already been said about the mathematics on the bottom layer. However, the amount of material in the top two layers is still prodigiously large. The material in the following chapters has been selected to show you the origins of the mathematics you learned in elementary and secondary school and the many different ways in which these topics have been studied by different societies. You may be surprised to find how much of it is a comparatively recent invention and how much has been transformed almost beyond recognition from its origins—mathematics does not spring full-grown from the earth, it develops like a living organism. An exclusive preoccupation with this "standard" mathematics, however, would not provide the perspective that comes from seeing other approaches to mathematics. For that reason we shall also spend some time studying the mathematics of ancient Egypt and the traditional mathematics of India, China, and Japan.

1.4 Problems and Questions

1.4.1 Questions on the Nature of Mathematics

Exercise 1.1 Describe the fundamental operations of grade-school arithmetic and geometry and tell how these operations relate to the theory of proportion. Find an application of the theory of proportion in the Fourteenth Amendment to the American Constitution.

Exercise 1.2 In what practical contexts of everyday life are the fundamental operations of arithmetic—addition, subtraction, multiplication, and division—needed? Give at least two examples of the use of each. How do these operations apply to the problems for which the theory of proportion was invented?

Exercise 1.3 It was stated above that measurement is performed on infinitely divisible, continuous objects. Are the two adjectives *infinitely divisible* and *continuous* synonyms, or is one stronger than the other?

Exercise 1.4 Why aren't there enough rational fractions so that we can assign one fraction to each point on a line? Consider in particular that one can write fractions as small as desired, and between any two fractions a and b there are other fractions,

such as $\frac{1}{2}(a+b)$. What more is needed if all points on a line are to be regarded as numbers? How can addition and multiplication be defined for all the points on a line?

Exercise 1.5 Why would the discovery of incommensurables be a challenge (or a nuisance, depending on your point of view)? What does it imply about the difference between infinite divisibility and continuity? What unfinished mathematical work does it point to?

1.4.2 Questions on the Origins of Mathematics

Exercise 1.6 What significance might there be in the fact that there are three columns of notches on the Ishango bone? What might be the significance of the numbers of notches in the three series?

Exercise 1.7 Can any firm conclusions be drawn from the Veronice wolf bone? Does it follow, for instance that, because the *bone* is 30,000 years old, the *notches* on it are also that old? If the notches were not used for counting things, what other purpose might they have served?

Exercise 1.8 The construction of a calendar is one of the fundamental problems that all large-scale societies have solved with varying degrees of precision. Why is a calendar needed by an organized society? Would a very small-scale society (consisting of, say, a few dozen families) require a calendar if it engaged mostly in hunting or fishing? What if the principal economic activity is agriculture?

Exercise 1.9 The basic blocks of time in our calendar are weeks, months, and years. What is the reason for attaching special significance to just these intervals of time and no others? Do these intervals give any clue as to the origin of our calendar? In the calendar described by Mbiti only months and years are mentioned; moreover months are astronomically defined, while a year is a meteorological term. What societal differences correspond to these calendrical differences?

Exercise 1.10 In constructing a calendar, we encounter the problem of measuring time. Measuring *space* is a comparatively straightforward task, based on the notion of congruent lengths. One can use a stick or a knotted rope stretched taut as a standard length and compare lengths or areas (rectangles) using it. Two lengths are congruent if each bears the same ratio to the standard length. In many cases one can move the objects around and bring them into coincidence. But what is meant by congruent *time intervals*? In what sense is the interval of time from 10:15 to 10:23 congruent to the time interval from 2:41 to 2:49?

Exercise 1.11 The preceding exercise brings up the general problem of measurement, which affects all of science. Except for mass (weight) and space, it seems that none of the quantities physics needs to measure has a straightforward definition, for example: time, velocity, acceleration, temperature, potential difference, and electric charge. Describe a way of measuring time and a way of measuring (constant) velocity. Can either of these concepts be described without the other?

Exercise 1.12 It appears from the preceding exercise that the scientific measurement of elapsed time requires three things: (1) the ability to measure distance, (2) a standard *motion* accepted as a standard constant velocity, and (3) the mathematical relation distance = velocity × time. Give an analogous description of a way of measuring temperature. Are there other physical quantities whose measurement must be approached in this way?

Exercise 1.13 Are there any other ways of measuring time besides the method already discussed, based on constant velocity?

Exercise 1.14 What conclusions can be drawn from the accounts of the Bushoong and Warlpiri given above? Are these people engaged in doing mathematics as we know it? Is there an "intuitive" preverbal knowledge of mathematics that guides the creation of mathematics? If so, what additional value is there in systematizing mathematics in treatises such as those of Euclid on geometry, in which everything is deduced logically from stated premises? Is anything lost in the transition from intuitive results to those that are rigorously proved?

Exercise 1.15 If only preferred marriages are taken into account among the Warlpiri, then a mother–daughter chain will have period 4, that is, each woman will belong to the same section as her great-great grandmother. Show that a father–son chain will have period two, that is, each man will belong to the same section as his grandfather. Also show that a female chain always stays within either the lower half of the numbers (1–4) or the upper half (5–8), while a male chain alternates between the two halves. What can be said about a gender-alternating sequence of generations, that is, mother–son–daughter–son, etc.?

Exercise 1.16 What may have happened in the history of the Bushoong and Warlpiri to motivate the particular mathematically oriented preoccupations discussed above?

Exercise 1.17 Find a unicursal tracing of the graph shown in Fig. 1.2.

Exercise 1.18 Consider the following three-column list of number names in English and Russian. The first column contains the cardinal numbers (those used for counting), the second column the ordinal numbers (those used for ordering), and the third the fractional parts. Study and compare the three columns carefully. The ordinal numbers and fractions and the numbers 1 and 2 are grammatically adjectives in Russian. They are given in the feminine form, since the fractions are always given that way in Russian, the noun *dolya*, meaning *part* or *share*, always being understood. If you know any other language, prepare a similar table for that language, then describe your observations and inferences. What does the table suggest about the origin of counting?

English			Russian		
one	first	whole	odna	pervaya	tselaya
two	second	half	dve	vtoraya	polovina
three	third	third	tri	tret'ya	tret'
four	fourth	fourth	chetyre	chetvyortaya	chetvert'
five	fifth	fifth	pyat'	pyataya	pyataya
six	sixth	sixth	shest'	shestaya	shestaya

Exercise 1.19 Does the development of personal knowledge of mathematics mirror the historical development of the subject? That is, do we learn mathematical concepts as individuals in the same order in which these concepts appeared historically? (When answering this question distinguish between formal mathematical knowledge and intuitive, nonverbal understanding of mathematical principles.)

Exercise 1.20 Topology, which may be unfamiliar to you, studies (among other things) the mathematical properties of knots, which have been familiar to the human race at least as long as most of the subject matter of geometry. Why was such a familiar object not studied mathematically until the twentieth century?

1.5 Endnotes

1. Koehler's work on counting birds was published in German in the *Bulletin of Animal Behavior*, No. 9. A translation can be found in Vol. 1 of *The World of Mathematics*, edited by James R. Newman and published by Simon and Schuster (New York, 1956) pp. 491–492.

2. The description of Pavlov's work is based on *Conditioned Reflexes* (1928), Dover reprint, New York (1960), p. 122, and his *Selected Works* (Foreign Languages Publishing House, Moscow, 1955).

3. Skinner's article was published in the *Journal of Experimental Psychology*, **38** (1), (Feb. 1948) pp. 168–172.

4. The discovery of the Moravian wolf bone and other artifacts is described in the *Illustrated London News* of Oct. 2, 1937.

5. The account of the Cahokia sun circle is summarized from E. C. Krupp's book *Echoes of the Ancient Skies* (Harper & Row, New York, 1983).

6. Dr. Mbiti's discussion of the calendar can found in his book *African Religions and Philosophy*, pp. 20–21. This passage is quoted at greater length by Prof. Claudia Zaslavsky in *Africa Counts*, pp. 62–63, which is the source of the discussion in this chapter.

7. Prof. Ascher's book is *Ethnomathematics*, Brooks/Cole, 1991. The discussion of the Bushoong is on pp. 70–72, and that of the Warlpiri on pp. 31 and 62. A full account of the work of Torday can be found in the book of

John Mack, *Emil Torday and the Art of the Congo. 1900–1909* (University of Washington Press, Seattle, 1990).

8. The Andaman method of counting was reported in *Numbers and Numerals* by David Eugene Smith and Jekuthiel Ginsburg, published by the National Council of Teachers of Mathematics, Washington, DC., 1937, pp. 2–3.

9. The book by Prof. Fuson is *Children's Counting and Concepts of Number* (Springer-Verlag, New York, 1988). The observations mentioned are on p. 15.

10. Among Piaget's published works are *The Child's Conception of Number* (The Humanities Press, Inc., New York, 1952) and (with Bärbel Inhelder) *The Child's Conception of Space* (Routledge & Kegan Paul, London, 1967). The conclusions described here are from the second of these.

Chapter 2

Ancient Egyptian Mathematics

2.1 Introduction and Historical Setting

In reading the preceding chapter you may have realized that the field of prehistory of mathematics is rich in unsubstantiated conjectures. When we come to the earliest written mathematical documents, we still find a wide field of allowable interpretation, but we also find a great many more hard facts to base our conjectures on. In the last chapter we saw some of the archaeological evidence for the most primitive forms of mathematics, which involve merely counting things or constructing physical objects in the shape of simple geometric figures. The invention of the more sophisticated mathematics involved in performing arithmetic operations on whole numbers (addition, subtraction, multiplication, and division), the handling of fractions, and the study of such geometric properties as area, volume, congruence, and proportion is harder to trace. In the earliest treatises the operations of arithmetic appear in finished form, without any indication as to how they were discovered. We do not know how or why addition and multiplication were invented. These operations appeared several thousand years ago, apparently independently, in China, India, Mesopotamia, and Egypt. The oldest mathematical records, dating back more than 4000 years, are found in Mesopotamia (Iraq). These records, however, are in the form of individual clay tablets devoted to particular mathematical problems; consequently they do not give a unified picture of the type of mathematics practiced.

The earliest systematic treatises on mathematics come from the Egyptian civilization, which was already 2000 years old before the mathematical treatises that survive today were written. After several thousand years during which the area now called Egypt was the home of isolated agricultural communities a process of consolidation began, and by 3100 B.C.E. there were two major kingdoms, Upper Egypt in the south and Lower Egypt in the north. Egypt became politically unified about this time when a ruler of Upper Egypt, variously said to be named Menes,

Narmer, or 'Scorpion,'[1] conquered Lower Egypt. In the four centuries following this conquest, a number of technological advances were made in Egypt making it possible to undertake large-scale engineering projects. Such projects required a certain amount of arithmetic and geometry. Shortly after the beginning of the Old Kingdom (2685 B.C.E.) the famous Step Pyramid of Djoser was built, the first structure made entirely of hewn stone. The Old Kingdom, which lasted just over five centuries, was a time of active building of temples and tombs. The collapse of central authority at the end of this period led to a century and a half during which the real power was held by provincial governors. The central authority recovered when the governors of Thebes extended their power northward, and over several generations brought about the Middle Kingdom (2040–1785 B.C.E.). When the central authority weakened again at the end of this period, foreign invaders known as the Hyksos conquered most of Egypt from the north. The Hyksos rule lasted for about a century, until some of their puppet governors became strong enough to usurp their authority; the Hyksos were driven out in 1570 B.C.E., which marked the beginning of the New Kingdom. It was during the Hyksos period that the earliest mathematical treatises still extant were written. We therefore begin with a discussion of mathematics as practiced in the Middle Kingdom.

2.2 Sources

The great architectural monuments of ancient Egypt are covered with hieroglyphics, some of which contain numbers. In fact, the ceremonial mace of the founder of the first dynasty contains records that mention oxen, goats, and prisoners and contain the hieroglyphic symbols for the numbers 10,000, 100,000, and 1,000,000. These hieroglyphs, while suitable for ceremonial recording of numbers, required some simplification for easy writing on papyrus or leather. The simplified cursive form of the hieroglyphics, known as the hieratic script, is the language of the earliest written documents that have come down to us.

The most detailed information about Egyptian mathematics comes from a single document written in the hieratic script on papyrus around 1650 B.C.E. and preserved in the dry Egyptian climate. This document is known properly as the Ahmose Papyrus, after its writer, but also as the Rhind Papyrus after the British lawyer Alexander Rhind (1833–1863), who went to Egypt for his health and became an Egyptologist. Rhind purchased the papyrus in Luxor, Egypt in 1857. Parts of the original document have been lost, but a section consisting of 14 sheets glued end to end to form a continuous roll $3\frac{1}{2}$ feet wide and 17 feet long remains. Part of it is on public display in the British Museum, where it has been since 1865. Some missing pieces of this document were later (1922) discovered in the Egyptian collection of the New York Historical Society; these are now on display at the Brooklyn Museum. A slightly earlier mathematical papyrus, now in the Moscow Museum of Fine Arts, consists of sheets about one-fourth the size of the Ahmose Papyrus. This papyrus was purchased by V. S. Golenishchev (1856–1947)

[1] It is not known with certainty whether these are one, two, or three persons.

in 1893 and donated to the museum in 1912. The Moscow Papyrus, however, is difficult reading, even for experts in the hieratic script in which it is written, and some of the parts that are completely clear duplicate parts of the Ahmose Papyrus. A third document, a leather roll purchased along with the Ahmose Papyrus, was not unrolled for 60 years after it reached the British Museum because the curators feared it would disintegrate if an attempt was made to unroll it. Because it was written on expensive leather, rather than cheap papyrus, the museum authorities assumed it contained material of great importance. At last suitable techniques were invented for softening the leather, and the document was unrolled in 1927. The results were disappointing, as the contents turned out to be a collection of 26 sums of unit fractions. A fourth set of documents, known as the Reisner Papyri after the American archaeologist George Andrew Reisner (1867–1942), who purchased them in 1904, consists of four rolls of records from dockyard workshops, apparently from the reign of Senusret I (1971–1926 B.C.E.). They are now in the Boston Museum of Fine Arts. These documents show the practical application of Egyptian mathematics in construction and commerce.

We are fortunate to be able to date the Ahmose Papyrus with such precision. The author himself gives us his name and tells us that he is writing in the fourth month of the flood season of the thirty-third year of the reign of Pharaoh Auserre (Apepi I). From this information Egyptologists arrived at a date of around 1650 B.C.E. for this papyrus. Ahmose tells us, however, that he is merely copying work written down in the reign of Pharaoh Nymaatre, also known as Amenemhet III (1842–1797 B.C.E.), the sixth pharaoh of the Twelfth Dynasty. It appears, therefore, that the mathematical knowledge contained in the papyrus is at least 4000 years old.

2.3 The Ahmose Papyrus

We shall give an overview of the contents of the papyrus before proceeding to a detailed study of Egyptian mathematics based on the information it contains.

The author begins by telling us that his work is a "correct method of reckoning, for grasping the meaning of things, and knowing everything that is, obscurities... and all secrets." Then follow tables resembling multiplication tables (more on this subject below), and then 87 problems involving various mathematical processes. Attempts have been made to discern a pattern in the arrangement of these problems, but the only suggestion that seems plausible is that the problems are grouped according to their application rather than their method of solution. The first six problems, for example, involve dividing loaves of bread among 10 people. Problems 7–23 are purely technical and show how to add fractional parts and, given a certain number of fractional parts, how to find complementary fractional parts to obtain a whole. Problems 24–38 are concerned with finding a quantity of which certain fractional parts will yield a given number. Area, volume, and general measurement problems are numbered from 40 to 60, and the remaining problems are concerned with various commercial applications to the distribution of goods.

2.4 Egyptian Arithmetic

2.4.1 General Features

What can we conclude about Egyptian mathematics from the problems considered in the Ahmose Papyrus and the methods used to solve them? To answer this question we must first clear our minds of modern modes of thought. We naturally think of arithmetic as consisting of the four operations of addition, subtraction, multiplication, and division performed on whole numbers and fractions. We learn the rules for carrying out these operations in childhood and do them automatically, without attempting to prove that they are correct. The situation was different for the Egyptian. To the Egyptian, it seems, the fundamental operations were addition and *doubling*, and these operations were performed on whole numbers and *parts*. Something needs to be said about these two fundamental differences from our way of thinking.

Let us consider first the absence of multiplication and division as we know them. The tables you looked at in Exercise 1.18 should have convinced you that there is something special about the number 2. We don't normally say "one-twoth" for the result of dividing something in two parts. This linguistic peculiarity suggests that *doubling* is psychologically different from applying the general concept of multiplying in the special case when the multiplier is 2. For the Egyptian the more general operation *did not exist*. It is important in trying to understand the thought processes of ancient mathematicians not to impose our own interpretation on the subject. Doubling can be done automatically in hieroglyphics. Having a hierarchical system of recording numbers in which a unit is represented by a vertical stroke (|), 10 by a hoop (∩), 100 by a coil resembling rope,[2] 1000 by a lotus flower, etc., the Egyptians could double any given number by simply drawing a copy of each symbol next to it, then "trading in" ten of any symbol for one of the next higher symbol. There is no need to think of multiplication in general, and the process can be carried out automatically with almost no conscious thought. It is therefore not surprising that this operation became a substitute for the more complicated arithmetic operations.

Next consider the absence of what we would call fractions.

The closest Egyptian equivalent to a fraction is what we have called a *part* above. For example, what we refer to nowadays as the fraction $\frac{1}{7}$ would be referred to as "the seventh part." This language conveys the image of a thing divided into seven equal parts arranged in a row and the seventh (and last) one being chosen. For that reason, according to B. L. Van der Waerden, (1903–1996) there can be only *one* seventh part (namely the last one), and so there would be no way of expressing what we call the fraction $\frac{3}{7}$. An exception was the fraction that we call $\frac{2}{3}$, which occurs constantly in the Ahmose Papyrus. There was a special symbol meaning "the two parts" (out of three). In general, however, the Egyptians used only *parts*, which in our way of thinking are *unit fractions*, that is, fractions whose

[2] We shall not burden the student's memory with the hieroglyphic symbols for larger numbers, since the Egyptian mathematical treatises were written in a cursive script rather than hieroglyphics. To get the flavor of Egyptian notation, the symbols for 1 and 10 will suffice.

numerator is 1. Again, thinking of them as unit fractions only makes the history of the subject harder to grasp; our familiarity with fractions in general makes it difficult to see what the fuss is about when the author asks what must be added to the two parts and the fifteenth part in order to make a whole (Problem 21). If this problem is stated in modern notation, it merely asks for the value of $1-\left(\frac{1}{15}+\frac{2}{3}\right)$, and of course, we get the answer immediately, expressing it as $\frac{4}{15}$. Both this process and the answer would have been foreign to the Egyptian, whose solution will be described below.

2.4.2 Notation

The Egyptian counting system was a decimal system, as already noted.

In order to understand the Egyptians we shall try to imitate their way of writing down a problem. On the other hand, we would be at a great disadvantage if our desire for authenticity led us to try to solve the whole problem using their notation. The best compromise seems to be to use our symbols for the whole numbers and express a *part* by the corresponding whole number with a bar over it. Thus *the fifth part* will be written $\overline{5}$ and *the thirteenth part* by $\overline{13}$, etc. For "the two parts," that is, $\frac{2}{3}$, we shall use a double bar, that is, $\overline{\overline{3}}$.

2.4.3 Proportion

Although no general theory of proportion is mentioned in the Ahmose Papyrus, the entire document is permeated with the implicit use of this concept. Take, for instance, the simplest problem of multiplying two integers, which occurs as a "subroutine" in many of the problems. Since the only operation other than addition and subtraction of integers (which are performed automatically without comment) is doubling, the problem that we would describe as "multiplying 11 by 19" would have been written out as follows:

	19	1	*
	38	2	*
	76	4	
	152	8	*
Result	209	11	

Inspection of this process shows its justification. The rows are kept strictly in proportion by doubling each time. The final result can be stated by comparing the first and last rows: 19 is to 1 as 209 is to 11. The rows in the right-hand column that must be added in order to obtain 11 are marked with an asterisk, and the corresponding entries in the left-hand column are then added to obtain 209. In this way any two positive integers can easily be multiplied. The only problem that arises is to decide how many rows to write down and which ones are to be marked with an asterisk. But that problem is easily solved. You stop creating rows when the next entry in the right-hand column would be bigger than the number you are multiplying by (in this case 11). You then mark your last row with an asterisk,

subtract the entry in its right-hand column (8) from 11 (getting a remainder of 3), then move up and mark the next row whose right-hand column contains an entry not larger than this remainder (in this case the second row), subtract the entry in its right-hand column (2), from the previous remainder to get a smaller remainder (in this case 1), and so forth.

If, for the sake of brevity, we refer to this general process of doubling and adding as "calculating," then what we call division is expressed in Egyptian terms by the same word. For example, what we would call the problem of dividing 873 by 97 would be expressed by the Egyptian as "calculate with 97 so as to obtain 873," and written out as follows:

*	97	1	
	194	2	
	388	4	
*	776	8	
	873	9	Result

The process, including the rules for creating the rows and deciding which ones to mark with an asterisk, is exactly the same as in the case of multiplication, except that now it is the left-hand column that is used rather than the right-hand column. We create rows until the next entry in the left-hand column would be larger than 873. We then mark the last row, subtract the entry in its left-hand column from 873 to obtain the remainder of 97, then look for the next row above whose left-hand entry contains a number not larger than 97, mark that row, and so on.

2.4.4 "Parts"

Obviously the second use of the two-column system can lead to complications. While in the first problem we can always express any positive integer as a sum of powers of two, the second problem is a different matter. We were just lucky that we happened to find multiples of 97 that add up to 873. If we hadn't found them, we would have had to deal with those *parts* that have already been discussed. For example, if the problem were "calculate with 12 so as to obtain 28," it might have been handled as follows:

	12	1	
*	24	2	
	8	$\overline{\overline{3}}$	
*	4	$\overline{3}$	
	28	$2\ \overline{3}$	Result

What is happening in this computation is the following. We stop creating rows after 24 because the next entry in the left-hand column (48) would be bigger than 28. Subtracting 24 from 28, we find that we still need 4, yet no 4 is to be found. We therefore go back to the first row and multiply by $\frac{2}{3}$, getting the row containing 8 and $\overline{\overline{3}}$. Dividing by 2 again gets a 4 in the left-hand column. We then have the numbers we need to get 28, and the answer is expressed as $2\ \overline{3}$.

There are two more complications that arise in doing arithmetic the Egyptian way. The first complication is obvious. Since the procedure is based on doubling, but the double of a *part* may not be expressible as a part, how does one "calculate" with parts? The answer to that question is contained in a table at the beginning of the document. It is easy to double, say, the twenty-sixth part: the double of the twenty-sixth part is the thirteenth part. If we try to double again, however, we are faced with the problem of doubling a part involving an odd number. The table gives the answer: the double of the thirteenth part is the eighth part plus the fifty-second part plus the one hundred fourth part. In our terms this tabular entry expresses the fact that

$$\frac{2}{13} = \frac{1}{8} + \frac{1}{52} + \frac{1}{104}.$$

With this table, which gives the doubles of all parts involving an odd number up to 99, straightforward multiplication involving parts is a feasible problem. There remains, however, one final complication before one can set out to solve any and all problems.

The calculation process described above requires subtraction at each stage in order to find out what sum is lacking in a given column. When the column already contains *parts*, this leads to the second complication: the problem of *subtracting parts*. (*Adding* parts is no problem. The author merely writes them one after another. The sum is condensed if, for example, the author knows that the sum of $\overline{3}$ and $\overline{6}$ is $\overline{2}$.) This technique, which is harder than the simple procedures discussed above, is explained in the papyrus itself in Problems 21–23. Problem 21, as mentioned above, asks for the parts that must be added to the sum of $\overline{\overline{3}}$ and $\overline{15}$ to obtain 1. The procedure used to solve this problem is as follows. Begin with the two parts in the first row:

$\overline{\overline{3}}$	$\overline{15}$	1

Now the problem is to see what must be added to the first column in order to obtain the second column. Preserving proportions, the author multiplies the row by 15, getting

10	1	15

It is now clear that, when the problem is "magnified" by a factor of 15, we need to add 4 units. Therefore the only remaining problem is, as we would put it, to divide 4 by 15, or, in language that may reflect better the thought process of the author, to "calculate with 15 so as to obtain 4." This operation is carried out in the usual way:

15	1	
1	$\overline{15}$	
2	$\overline{10}$ $\overline{30}$	[from the table]
4	$\overline{5}$ $\overline{15}$	Result

Thus the parts that must be added to the sum of $\overline{\overline{3}}$ and $\overline{15}$ in order to reach 1 are $\overline{5}$ and $\overline{15}$. It is of interest that this "subroutine," which is essential to make the

system of computation work, was always written in red ink in the manuscripts, as if the writers distinguished between computations made within the problem to find the answer and computations made in order to operate the system. Having learned how to complement (subtract) parts, what are called *hau* (or *aha*) computations by the author, one can confidently attack any arithmetic problem whatsoever. One point should be noted, however: there is no single way of doing these problems. In general, the author seems to proceed by getting close to the result that is needed in one column, then "shrinking" the whole problem by dividing by a large number, so that subsequent steps can be fine enough to hit the target. Specialists in this area have detected systematic procedures by which the table of doubles was generated and patterns in the solution of problems that indicate, if not an algorithmic procedure, at least a certain habitual approach to such problems.

We are now ready to attack a genuine problem from the papyrus. The one we pick is Problem 35, which, translated literally and misleadingly, reads as follows:

> Go down I times 3. My third part is added to me. It is filled. What is the quantity saying this?

Properly interpreted, this problem asks for a number that yields 1 when it is tripled and the result is then increased by the third part of the original number. In other words, "calculate with $3\ \overline{3}$ so as to obtain 1." The solution is as follows:

3	3	1
10		3 [multiplied by 3]
5		$1\ \overline{2}$
1		$\overline{5}\ \overline{10}$ Result

2.4.5 "Practical" Problems

The papyrus contains several problems of a superficially practical nature involving the slope of pyramids and the strength of beer. Both of these involve what we think of as a ratio. Thus they make good use of the format by which the Egyptians solved arithmetic problems. Several units of weight are mentioned in these problems, but the measurement we shall pay attention to is not a weight at all, but a measure of the dilution of bread or beer. It is called a *pesu* and defined as the number of loaves of bread or jugs of beer obtained from one *hekat* of grain. A hekat was slightly larger than a gallon, 4.8 liters to be precise. Unfortunately, this information by itself is useless, since we don't know the size of a standard loaf of bread or a standard jug of beer. What we do know is that the larger the *pesu*, the weaker the bread or beer.

Problem 71 tells of a jug of beer produced from half a hekat of grain (thus its *pesu* was 2). One-fourth ("the fourth part") of the beer is poured off, and the jug is then topped up with water. The problem asks for the new *pesu*. The author reasons that the eighth part of a hekat of grain was removed, leaving (in his terms) $\overline{4}\ \overline{8}$ (what we would call $\frac{3}{8}$) of a hekat of grain. Since this amount of grain goes into one jug, it follows that the *pesu* of that beer is $2\ \overline{\overline{3}}$. The author gives this

result immediately, apparently assuming that by now the reader will know how to "calculate with $\overline{4}\ \overline{8}$ until 1 is reached."

2.4.6 Algebra

Problems that require finding an unknown number without specifying which operations are to be performed on the data can be considered to be algebra, even though what we consider the distinguishing characteristic of algebra—the use of symbols for the unknown—is not present. Many such numerical problems occur in the Ahmose Papyrus, mostly involving the notion of proportion. The concept of proportion is the key to the problems based on the "rule of false position." Problem 24, for example, asks for the quantity that yields 19 when its seventh part is added to it. The author notes that if the quantity were 7 (the "false [sup]position"), it would yield 8 when its seventh part is added to it. Therefore the correct quantity will be obtained by performing the same operations on the number 7 that yield 19 when performed on the number 8. As we have already seen, the Egyptian format for such computations is well adapted for handling problems of this sort.

The scribes were also capable of performing operations more complicated than mere proportion. They could take the square root of a number, which they called a "corner." In a papyrus known as the Berlin Papyrus, nearly contemporaneous with the Ahmose Papyrus and kept in the State Museum in Berlin, one finds the following problem:

> ... the area of a square of 100 is equal to that of two smaller squares. The side of one is $\overline{2}\ \overline{4}$ the side of the other. Let me know the sides of the two unknown squares.

Here we are asking for two quantities given their ratio ($\frac{3}{4}$) and the sum of their squares (100). The scribe assumes that one of the squares has side 1 and the other has side $\frac{3}{4}$. Since the resulting total area is $1\ \overline{2}\ \overline{16}$, the square root of this quantity is taken ($1\ \overline{4}$), yielding the side of a square equal to the sum of these two given squares. This side is then multiplied by the correct proportionality factor so as to yield 10 (the square root of 100). That is, the number 10 is divided by $1\ \overline{4}$, giving 8 as the side of the larger square and hence 6 as the side of the smaller square. This example, incidentally, was cited by Van der Waerden as evidence of early knowledge of the Pythagorean theorem in Egypt.

Thus, despite having what appear to us to be rather crude computational methods, the Egyptians stretched their arithmetical techniques to the maximum and were able to handle problems involving two unknown quantities, provided the data allowed them to reduce the problem to one in which the two-column direct-proportion method applies. The operation of taking the square root makes this reduction possible in some cases. We can see that the Egyptians were doing mathematics in the full sense of the term—adapting known techniques to solve problems requiring a high degree of ingenuity. With the insight produced by our more advanced mathematical knowledge we can see that some of these problems would be more naturally handled by new techniques unknown to the Egyptians.

Mathematics sometimes develops in this way—existing techniques are applied in an ingenious way to solve a new and difficult problem. Then, after some insight has been gained through such a solution, the problem receives sufficient attention to allow the development of a systematic approach usable in a whole class of problems.

2.5 Egyptian Geometry

Like most ancient cultures, the Egyptians treated geometry as an area of application for arithmetic. The only geometric problems considered are those involving measurement. These few problems, however, show considerable insight into the properties of simple geometric figures such as the circle, the triangle, the rectangle, and of course the pyramid. What is frustrating or challenging historically is that the author never tells how the procedures for finding area and volume were arrived at. They are assumed to be known, and so we are left to conjecture their origin. Those involving polygons are correct from the point of view of Euclidean geometry, while those involving circles are merely approximations. The author makes no distinction between the two. For instance, Problems 48–52 calculate areas in the shape of circles, triangles, and rectangles, and in the case of the polygonal figures the areas are calculated in agreement with Euclidean geometry.

2.5.1 The Circle

Five of the problems (41–43, 48, and 50) involve calculating the area of a circle. In all these problems the author assumes the reader knows that the area of a circle is the area of the square whose side is obtained by removing the ninth part of the diameter.[3] There have been various conjectures as to how the Egyptians might have arrived at this result. One such conjecture involves a square of side 8. If a circle is drawn through the points 2 units from each corner, it is visually clear that the four fillets at the corners, at which the square is outside the circle, are nearly the same size as the four segments of the circle outside the square, hence this circle and this square may be considered equal in area. Now the diameter of this circle can be obtained by connecting one of the points of intersection to the opposite point, as in Fig. 2.1, and measurement will show that this line is very nearly 9 units in length (it is actually $\sqrt{80}$ in length).

It may be appropriate at this point to comment on the value of conjectures such as the one just described. Such conjectures are useful in showing us how people *may* have thought, so long as they are not taken as established fact. An example of such a metamorphosis of conjecture into "fact" may be known to you. It has been widely reported, in numerous textbooks and even in a recent film produced by the

[3] In our language the area is the square on eight-ninths of the diameter, that is, it is the square on $\frac{16}{9}$ of the radius. In our language, not that of Egypt, this gives a value of π equal to $\frac{256}{81}$. Please remember, however, that the Egyptians had no concept of the number π. The constant of proportionality that they always worked with represents what we would call $\pi/4$.

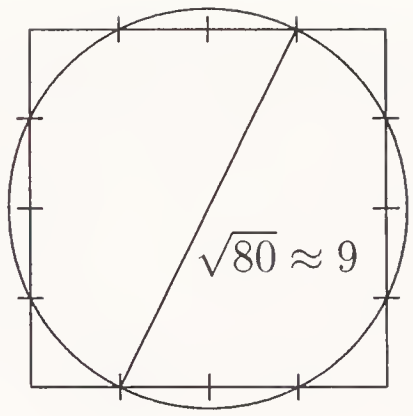


Figure 2.1: Conjectured Egyptian squaring of the circle.

Mathematical Association of America, that the Egyptians laid out right angles by stretching a rope with twelve equal intervals knotted on it so as to form a 3–4–5 right triangle. What is the evidence for this conjecture? Well, the Egyptians *did* lay out very accurate right angles, and it is known that their surveyors used ropes as measuring instruments (see Fig. 2.2) and were referred to by the Greek philosopher Democritus as *rope-stretchers*. That is *all* the evidence there is in favor of this conjecture. The earliest Egyptian text that mentions a right triangle and finds the length of all its sides using the Pythagorean theorem dates from about 300 B.C.E., and by that time the influence of "Greek" mathematics was already established. No Egyptian text from Pharaonic times mentions even one special case of the Pythagorean theorem. Now, given that the evidence for this conjecture is really nonexistent, why is it reported as fact? Simply because it has been repeated frequently since it was originally made by the historian Moritz Cantor (1829–1920) in 1882. We know precisely the source of the rumor, but historians of mathematics have been powerless to prevent enthusiastic mathematicians from turning it into a "fact."[4]

Ratio and proportion again occur in Problems 56–60, which involve the slope of the sides of pyramids and other figures. It is of interest that there is a unit of slope analogous to the *pesu* in the problems involving strength of bread and beer. The unit of slope is the *seked*, defined as the number of palms of horizontal displacement associated with a vertical displacement of 1 royal cubit.[5] In Problem 57 a pyramid with a *seked* of 5 $\overline{4}$ and a base of 140 cubits is given. The problem is to find its height.

The *seked* given here ($\frac{3}{4}$ of 7) is exactly that of one of the actual pyramids, the pyramid of Khafre, who reigned from 2558 to 2532 B.C.E. It appears from archaeological evidence that stones were mass-produced in several standard shapes with a *seked* that could be increased in intervals of one-fourth. Pyramid builders and designers could thereby refer to a standard brick shape, just as modern architects

[4]This point was made very forcefully by Van der Waerden in *Science Awakening*. However, in his later book *Geometry and Algebra in Ancient Civilizations* Van der Waerden claimed that integer-sided right triangles, which seem to imply knowledge of the Pythagorean theorem, are ubiquitous in the oldest megalithic structures.

[5]One royal cubit was 7 palms. In our terms the *seked* is 7 times the tangent of an angle one of whose sides is vertical.

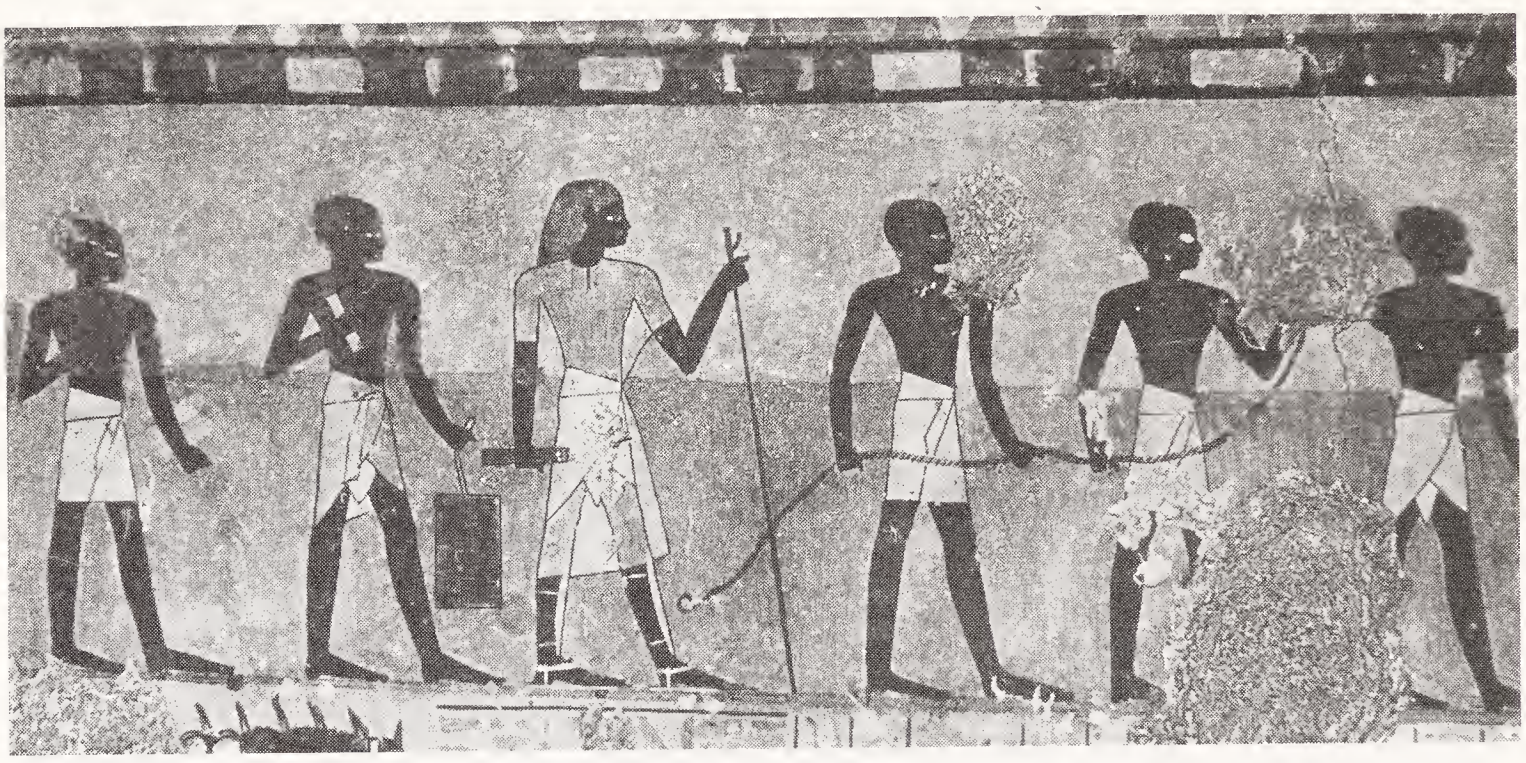


Figure 2.2: Egyptian surveyors. The Bettmann Archive.

and contractors can specify a standard diameter for a water pipe. This standard manufacturing process may explain why lateral displacement is measured in palms and vertical displacement in cubits in the definition of the *seked*. This way of measuring the dimensions will allow an annoying factor of 7 to disappear from the computations at an early stage. (The number 7 is computationally awkward in all the common bases for arithmetic.) Problem 58 gives the dimensions of the same pyramid and asks for its *seked*.

2.5.2 Volumes and Curved Surfaces

One of the most remarkable achievements of the Egyptians is the discovery of accurate ways of computing volumes. In Problem 42 we find the correct procedure used for finding the volume of a cylindrical silo, that is, the area of the circular base is multiplied by the height. Problems 44–46 calculate the volume of prisms on a rectangular base by the same procedure.

Before leaving the subject of Egyptian geometry, we shall note two problems from the Moscow Papyrus. Problem 14 from this papyrus asks for the volume of the frustum of a square pyramid, given that the side of the lower base is 4, the side of the upper base is 2, and the height is 6. The author gives the correct recipe: add the areas of the two bases to the area of the rectangle whose sides are the sides of the bases, that is, $2 \times 2 + 4 \times 4 + 2 \times 4$, then multiply by the height and divide by 3. Again, there are various conjectures as to how this knowledge was obtained.

Problem 10 of the Moscow Papyrus has been subject to various interpretations. It asks for the area of a curved surface that is either half of a cylinder or a hemisphere. In either case it is worth noting that the area is obtained by multiplying the length of a semicircle by another length in order to obtain the area. Since finding the area of a hemisphere is an extremely difficult problem, it would be

remarkable if the Egyptians even attempted it with their approximative techniques. For that reason the interpretation of the figure as half a cylinder may be plausible. The problem was translated into German by the Russian scholar V. V. Struve (1889–1965); the following is an English translation from the German:

> The way of calculating a basket, if you are given a basket with an opening of $4\ \overline{2}$. O, tell me its surface!
>
> Calculate $\overline{9}$ of 9, since the basket is half of an egg. The result is 1. Calculate what is left as 8. Calculate $\overline{9}$ of 8. The result is $\overline{\overline{3}}\ \overline{6}\ \overline{18}$. Calculate what is left of this 8 after this $\overline{\overline{3}}\ \overline{6}\ \overline{18}$ is taken away. The result is $7\ \overline{9}$. Calculate $4\ \overline{2}$ times with $7\ \overline{9}$. The result is 32. Behold, this is the surface. You have found it correctly.

If we interpret the basket as being a hemisphere, the scribe has first doubled the diameter of the opening from $4\ \overline{2}$ to 9 "because the basket is half of an egg." (If it had been the *whole* egg, the diameter would have been quadrupled.) The procedure used for finding the area here is equivalent to the formula $2d \cdot \frac{8}{9} \cdot \frac{8}{9} \cdot d$. Taking $\left(\frac{8}{9}\right)^2$ as representing $\frac{\pi}{4}$, it amounts to $\frac{\pi}{2}d^2$, or $2\pi r^2$, which is indeed the area of a hemisphere of radius r.

This value is also the lateral area of half of a cylinder of height d and base diameter d. If the basket is interpreted as half of a cylinder, the opening would be square and the number $4\ \overline{2}$ would be the side of the square. The numerical answer is consistent with this interpretation, but it does seem strange that only the lateral surface of the cylinder was given, unless the basket was open at the sides.

2.6 Pure Mathematics

Many of the problems in the Ahmose Papyrus go beyond any conceivable practical application. For instance, the table of doubles of parts gives the double of the 61st part as the 40th part plus the 244th part plus the 488th part, plus the 610th part. What object could be divided in so many ways? To some extent such problems automatically have a practical value—the mathematician does not know what specific data will occur in a practical problem and so must develop general methods and test them on examples that may be more complicated than the ones that will occur in practice. However, the papyrus also contains evidence that its author loved the subject for its own sake and enjoyed making up problems for the sheer pleasure of seeing the techniques in operation. A good example is Problem 79, whose language is somewhat obscure, but definitely requires finding the sum $7 + 7^2 + 7^3 + 7^4 + 7^5$.

2.7 Practical Mathematics

The Reisner Papyri mentioned above provide us with fascinating confirmation that this system of arithmetic actually was used in government and commerce

in Egypt. These documents contain computations preparatory to the construction of a temple and give calculations of the volume of excavation and the number of workers required. One portion of the papyrus contains a column of volumes computed from length, width, and depth, then divided by 10, which was apparently the number of cubic cubits each worker was expected to excavate per day. For example, one portion of the building was to be 8 cubits long, 4 cubits wide, and 2 cubits deep, a total of 64 "cubic cubits." This volume of 64 cubic cubits is divided by 10, resulting in $6\ \overline{4}\ \overline{10}\ \overline{20}$ worker-days of digging.

Another example of the practical use of mathematics is provided by the records of the salary distribution of the personnel at the Temple of Illahun during the Middle Kingdom and discussed in a 1902 article by L. Borchardt (1863–1938). In the record discussed by Borchardt there were 70 loaves of bread, 35 jugs of Sd beer, and $115\ \overline{2}$ jugs of Hpnw beer. This salary pool was divided into 42 equal portions, which the scribe asserted to be $1\ \overline{\overline{3}}$ loaves, $\overline{\overline{3}}\ \overline{6}$ jugs of Sd beer, and (incorrectly) $2\ \overline{\overline{3}}\ \overline{10}$ jugs of Hpnw beer. The last figure should be $2\ \overline{2}\ \overline{4}$. The small error here (amounting to an excess of $\overline{60}$ of a jug of Hpnw beer) was corrected by the scribe without comment in the final tally. The scribe records that the temple director was to receive 10 portions; the head lay priest, 3; the head reader, 6; the scribe, $1\ \overline{3}$; and the usual reader, 4. Seven priests of various sorts were to receive 2 portions each, an officer referred to as Md'w was to receive one portion, and eight various workers and watchmen were to receive $\overline{3}$ each, bringing the total to 42. R. J. Gillings, from whose book this information is taken, doubts that anyone *could* measure out the amounts recorded (the portion of Hpnw beer given to each worker is listed as $\overline{\overline{3}}\ \overline{4}\ \overline{180}$). It is worth noting, however, that the loaves are to be divided into at most 18 pieces, and Gillings himself shows that the fractions recorded could be cut evenly from 5 loaves. As for the beer, it is at least possible that the ladle used to distribute it really was only one 180th of the jug.

2.8 The Egyptian Calendar

Among the most obvious scientific applications of both counting and geometry are the establishment of a calendar and the prediction of astronomical phenomena such as eclipses. To the extent that the weather is dependent on the celestial latitude of the sun, the movement of the sun among the stars is of practical importance to terrestrial economy. A calendar is needed even by people living in small groups, and it is vital to the coordination of a large-scale economy. It is especially important in places like Egypt, where it is necessary to organize mass movements of population away from a river during spring floods and then back to the river to cultivate the soil deposited by the flood. Populations cannot be moved at a minute's notice; it takes at least several days' warning to organize such activities.

There is no doubt that the Egyptians observed the world about them with considerable accuracy, as the careful north–south orientations of some of the pyramids shows us. Now anyone who observes the sky for any extended period of time cannot help noticing the bright blue-green star Sirius, which is overhead at mid-

night during the winter season. To the Egyptians this star was the goddess Sôpdit, and they had a special reason for noticing it. Like all stars, Sirius gains about 4 minutes per day on the sun, rising a little earlier each night until finally it rises just as the sun is setting. Then for a while it cannot be seen when rising, since the sun is still up, but it can be seen setting, since the sun will have gone down before it sets. It goes on setting earlier and earlier until finally it sets just after the sun. At that point it is too close to the sun to be seen for about 2 months. Then it reappears in the sky, rising just before the sun in the early dawn. It was during these days that the Nile began its annual flood in ancient times (the floods no longer occur since the Aswan Dam was built in the 1950s). Thus the heliacal rising of Sirius (just before the sun) signaled the approach of the annual Nile flood. The Egyptians therefore had a very good basis for an accurate solar calendar, using the heliacal rising of Sirius as the epoch (day one) of the year.

The Egyptians seem originally to have used a lunar calendar with 12 lunar cycles per year. However, such a calendar is seriously out of synchronicity with the sun, by about 11 or 12 days per year, so that it was necessary to add an extra "intercalary" month every 2 or 3 years. All lunar calendars must do this, or else wander through the agricultural year. However, at an early date the Egyptians cut their months loose from the moon and simply defined a month to consist of 30 days. Their calendar was thus a "civil" calendar, neither strictly lunar nor strictly solar. Each month was divided into three 10-day "weeks" and the whole system was kept from wandering too quickly from the sun by adding five extra days at the end of the year, regarded as the birthdays of the gods Osiris, Horus, Seth, Isis, and Nephthys. This calendar is still short by about $\frac{1}{4}$ day per year, so that in 1456 years it would wander through an entire cycle of seasons. The discrepancy between the calendar and the sun accumulated slowly enough to be adjusted for, and so no serious problems arose.[6] As already mentioned, this calendar used the star Sirius to fix its first day, the star called Sôpdit by the Egyptians. When the Greeks learned of Egypt, they called this goddess Sothis. Consequently the period of 1456 years is known as the *Sothic cycle.*

Throughout most of their history the ancient Egyptians used this very simple calendar based on a year of 365 days. Some of the principles on which it is based have been incorporated in calendars used by astronomers, in particular the Julian calendar. Where computation is concerned the Julian calendar has the supreme advantage that the number of days between any two dates is simple to compute. In contrast, try to compute in your head the number of days that elapsed from April 22, 1881 to August 13, 1907.[7] From an astronomical point of view, however, it has the disadvantage that it falls 3 days behind the sun every 400 years. Thus if

---

[6]In fact, this wandering has been convenient for historians, since the heliacal rising of Sirius was regularly recorded. It was on the first day of the Egyptian year in 2773 B.C.E., 1317 B.C.E., and 139 C.E. Hence a document that says the heliacal rising occurred on the sixteenth day of the fourth month of the second season of the seventh year of the reign of Senusret III makes it possible to state that Senusret III began his reign in 1878 B.C.E. (See *Chronicle of the Pharaohs* by Peter A. Clayton, Thames and Hudson, London, 1994, pp. 12–13.) On the other hand, some authorities claim that the calendar was adjusted by the addition of intercalary days from time to time to keep it from wandering too far.

[7]The answer is 9608. If you do the computation, remember that 1900 was not a leap year.

you need to know when the winter solstice occurred in a certain year, you have a difficult computational job ahead of you, whereas with our present (Gregorian) calendar we can say with confidence that the solstice was somewhere between December 20 and December 22, no matter what the year.

2.9 Problems and Questions

2.9.1 Problems in Egyptian Mathematics

Exercise 2.1 Double the hieroglyphic number  . Check your result in modern notation.

Exercise 2.2 Multiply 27 times 42 the Egyptian way.

Exercise 2.3 (Stated in the Egyptian style.) Calculate with 13 so as to obtain 364.

Exercise 2.4 Problem 23 of the papyrus asks what parts must be added to the sum of $\overline{4}$, $\overline{8}$, $\overline{10}$, $\overline{30}$, and $\overline{45}$ in order to obtain $\overline{\overline{3}}$. See if you can obtain the author's answer of $\overline{9}\ \overline{40}$, starting with his technique of magnifying the first row by a factor of 45. Remember that $\frac{5}{8}$ must be expressed as $\overline{2}\ \overline{8}$.

Exercise 2.5 Problem 24 of the papyrus, as mentioned in the text, asks for a number that yields 19 when its seventh part is added to it, and concludes that one must perform on 7 the same operations that yield 19 when performed on 8. Now in Egyptian terms, 8 must be multiplied by $2\ \overline{4}\ \overline{8}$ in order to obtain 19. Multiply this number by 7 to obtain the scribe's answer, namely $16\ \overline{2}\ \overline{8}$.

Exercise 2.6 Multiply the result of the last problem by $\overline{7}$, add the product to the result itself, and verify that you do obtain 19, as required. (Note: The table gives $\overline{4}\ \overline{28}$ as the double of $\overline{7}$.)

Exercise 2.7 Problem 33 of the papyrus asks for a quantity that yields 37 when increased by its two parts (two-thirds), its half, and its seventh part. Try to get the author's answer: the quantity is $16\ \overline{56}\ \overline{679}\ \overline{776}$. [Hint: the table for doubling fractions gives the last three terms of this expression as the double of $\overline{97}$. The scribe first tried the number 16 and found that the result of these operations applied to 16 fell short of 37 by the double of $\overline{42}$, which, as it happens, is exactly $1\ \overline{\overline{3}}\ \overline{2}\ \overline{7}$ times the double of $\overline{97}$.]

Exercise 2.8 Verify that the solution to Problem 71 ($2\ \overline{\overline{3}}$) is the correct *pesu* of the diluted beer discussed in the problem.

Exercise 2.9 Verify that the solution $\overline{5}\ \overline{10}$ given for Problem 35 is correct, that is, multiply this number by 3 and by $\overline{3}$ and verify that the sum of the two results is 1.

Exercise 2.10 Find the height of the pyramid with base 140 cubits and *seked* equal to 5 $\overline{4}$ (Problem 57 of the Ahmose Papyrus).

Exercise 2.11 Prove that the implied formula for the volume of a frustum of a square pyramid is correct. If the sides of the upper and lower squares are a and b, and the height is h, the implied formula is the following:

$$V = \frac{h}{3}(a^2 + ab + b^2).$$

Exercise 2.12 A tropical year is the time elapsed between successive south-to-north crossings of the celestial equator by the sun. A sidereal year is the time elapsed between two successive conjunctions of the sun with a given star. Because the celestial equator is rotating (about once in 26,000 years) a tropical year is about 20 minutes shorter than a sidereal year. Would you expect the flooding of the Nile to be synchronous with the tropical year or with the sidereal year? If the flooding is correlated with the tropical year, how long would it take for the heliacal rising of Sirius to be one day out of synchronicity with the Nile flood? If the two were synchronous 4000 years ago, how far apart would they be now, and would the flood occur later or earlier than the heliacal rising of Sirius?

2.9.2 Questions about Egyptian Mathematics

Exercise 2.13 Why do you suppose the author of the Ahmose Papyrus did not choose to say that the double of the thirteenth part is the seventh part plus the ninety-first part, that is,

$$\frac{2}{13} = \frac{1}{7} + \frac{1}{91}?$$

Why is the relation

$$\frac{2}{13} = \frac{1}{8} + \frac{1}{52} + \frac{1}{104}$$

made the basis for the tabular entry instead?

Exercise 2.14 Why not simply write $\overline{13}\ \overline{13}$ to stand for what we call $\frac{2}{13}$? What is the reason for using two or three other "parts" instead of these two obvious parts?

Exercise 2.15 Could the ability to solve a problem such as Problem 35, discussed above in Section 2.4.4, have been of any practical use? Try to think of a situation in which such a problem might arise.

Exercise 2.16 We would naturally solve many of the problems in the Ahmose Papyrus using an equation. Would it be appropriate to say that the Egyptians solved equations, or that they did algebra? What does the word *algebra* mean to you? How can you decide whether you are performing algebra or arithmetic?

Exercise 2.17 Do you agree with Van der Waerden that the presence of many "Pythagorean triples" such as 6, 8, 10 is evidence that the Egyptians did know the Pythagorean theorem?

2.10 Endnotes

1. Authorities differ on spellings and dates for Egyptian rulers. The usage in this chapter follows *Chronicle of the Pharaohs* by Peter A. Clayton, Thames and Hudson, 1994. Clayton, in turn, follows the *Penguin Guide to Ancient Egypt* by William J. Mumane (1983).

2. All the information on sources and most of the ensuing discussion is taken from the book *The Rhind Mathematical Papyrus* by Gay Robins and Charles Shute, published by the British Museum in 1987 and from the monograph *Mathematics in the Time of the Pharaohs* by Richard J. Gillings (Dover reprint, published in 1982).

3. I am indebted to Milo Gardner for sending me e-mail full of interesting information on the system by which the table of doubles of parts was created.

4. Van der Waerden's interpretation of Egyptian mathematics can be found in his classic work *Science Awakening*, published originally in Dutch. The English translation was published by Wolters-Noordhoff (Groningen, 1971). Chapter 1 (pp. 15–36) is devoted to Egyptian mathematics.

5. The problem of finding a square equal to the sum of two other squares is discussed by Gillings (pp. 161–162), who cites a paper by H. Shack-Shackenburg, "Der Berliner Papyrus 6619," in *Zeitschrift für Ägyptische Sprache*, **38** (1900), pp. 135–140; **40** (1902), pp. 65ff.

6. The conjectured Egyptian squaring of the circle is taken from the book of Robins and Shute, p. 45 (*op. cit.*, endnote 2).

7. The information on the *seked* of the pyramid of Khafre is taken from the book of Robins and Shute, p. 47 (op. cit., endnote 2).

8. A detailed discussion of the hemisphere/cylinder area problem from the Moscow Papyrus can be found in Van der Waerden's *Science Awakening*. In the English translation this topic is discussed on pages 33–34.

9. The Egyptian records are taken from the book of Gillings mentioned above, pp. 124–126 and pp. 218–225. Gillings cites the article by Borchardt as "Salary Distribution for Personnel of the Temple of Illahun," *Zeitschrift für Ägyptische Sprache*, **40** (1902–1903), pp. 113–117.

10. The story that the heliacal rising of Sirius was fixed as the first day of the calendar was recorded on the outside wall of the Temple of Ramesses III at Medinet Habu. [See the book by Pierre Montet, *Everyday Life in Egypt in the Days of Ramesses The Great*, translated from the French by A. R. Maxwell-Hyslop and Margaret S. Drower (Greenwood Press, Westport, CT, 1974).]

Chapter 3

Mesopotamia

3.1 Historical Setting

In contrast to Egypt, which had a fairly stable culture throughout many millennia, the region known as Mesopotamia (Greek for "between the rivers") was the home of many successive, quite distinct, civilizations. The name of the region derives from the two rivers, the Euphrates and the Tigris, that flow from the mountainous regions around the Mediterranean, Black, and Caspian seas into the Persian Gulf. In ancient times this region was a very fertile floodplain, although it suffered from an unpredictable climate. It was repeatedly invaded and conquered, and the successive dynasties spoke and wrote in many different languages. The convention of referring to all the mathematical texts that come from this area between 2500 B.C.E. and 300 B.C.E. as "Babylonian" gives undue credit to a single one of the many dynasties that ruled over this region. Although many different peoples invaded this region over time, occupying different parts of it, for purposes of analysis this history may be oversimplified and divided into eight different civilizations, as follows:

1. *Sumerian.* The Sumerians were either the original inhabitants of the region or immigrants from farther east. They spoke a language unrelated to the Semitic and Indo-European groups. They held sway over this region for several hundred years, starting about 3000 B.C.E. It was the Sumerians who invented the method of writing known as cuneiform (wedge-shaped), performed by pressing a stylus into wet clay. Many of the small clay tablets containing such records dried out (or were deliberately baked to preserve them) and have kept their information for over 4000 years.

2. *Akkadian.* These people were conquerors who spoke a Semitic language and adapted the Sumerian cuneiform writing to their own language. One consequence was the compilation of Sumerian–Akkadian dictionaries, the equivalent of the Egyptian Rosetta Stone for the later deciphering of these documents. The Akkadians established a commercial empire under King Sargon (ca. 2371–2316 B.C.E.), which eventually collapsed and was re-

placed by a system of city-states in which the city of Ur at the mouth of the Euphrates was dominant.

3. *Amorite.* The Amorites, like the Akkadians, spoke a Semitic language. They invaded the area just before 2000 B.C.E. and established a number of small kingdoms, of which Assyria was the first to become prominent, but was soon succeeded by Babylon under Hammurabi (1792–1750 B.C.E.)

4. *Hittite.* The Hittites expanded from the west, the region now called Turkey. They spoke a language of the Indo-European family (the family to which English belongs). By 1650 B.C.E. they had established a kingdom to rival the Amorites, and in 1595 they sacked the city of Babylon. The Hittite civilization collapsed around 1200 B.C.E. due ultimately to pressure from the west exerted by the "Sea Peoples," known to us from the Bible as the Philistines.

5. *Assyrian.* The Sea Peoples, although they caused the collapse of the Hittite Empire, did not occupy the portion of Mesopotamia that had been part of that Empire. Instead, an empire based in the old city of Assyria began to grow and expand as far as its very well organized army and clever diplomacy could sustain it. The Assyrians eventually controlled a large portion of the region between the Mediterranean and the Persian Gulf, including present-day Palestine and parts of northern Egypt. Since this empire included the city of Babylon, it absorbed a great deal of the culture associated with that city. The Assyrian empire was finally conquered by the Chaldean King Nebuchadnezzar (605–562 B.C.E.).

6. *Chaldean.* This empire, although very short-lived (ca. 625–539 B.C.E.), is well-known in the West because of Nebuchadnezzar, who is mentioned in the books of Kings, Jeremiah, and Daniel in the Bible. It was Nebuchadnezzar who conquered Jerusalem in 597 B.C.E. and took the King of Judah and his followers into exile in Babylon. This civilization exerted a great influence on the writers of the Bible, especially the customs of the Chaldean court, where astrology was taken seriously.

7. *Persian.* As is well known from the Book of Daniel, the Chaldean empire was conquered in 539 B.C.E. by the Persian king Cyrus the Great. Cyrus repatriated the exiles from Jerusalem and ordered the rebuilding of the Temple. The Persians, who speak an Indo-European language, have had an unbroken civilization since that time, although one subject to many changes of dynasty and religion. We shall see them coming into the story of mathematics at various points.

8. *Seleucid.* The high period of culture in mainland Greece coincided with the rise of the Athenian Empire in the middle of the fifth century B.C.E. The Athenian Empire was perceived as a threat by the Spartans, who brought it down through the Peloponnesian War (429–404 B.C.E.). By that time, however, Greek scholarship and the Greek language were well established

as intellectual forces. When the Macedonian kings Philip and Alexander conquered the territory from mainland Greece to India and northwest Africa, they consciously attempted to spread this culture. As a result, intellectual centers grew up in widely separated places where scholars, not all Greek by birth, wrote and argued in the Greek language. Paradoxically the three best known Greek mathematicians, Euclid, Archimedes, and Apollonius, lived and worked in Egypt, Sicily, and Turkey.

When Alexander died in 323 B.C.E., his empire was divided among three of his generals. Besides the original Macedonian kingdom centered at Pella just north of Greece, there were two other regions with centers in Egypt and the Fertile Crescent. Egypt was ruled by the general Ptolemy Soter (the last of his heirs was Cleopatra, who presided over the incorporation of Egypt into the Roman Empire under Julius Caesar) while the regions around the Fertile Crescent were ruled by general Seleucus, and thereby became known as the Seleucid Kingdom.

3.2 Cuneiform Texts

Of the many thousands of cuneiform texts scattered through museums around the world, a few hundred have been found to be mathematical in content. These texts come mostly from the period of late Akkadian, Amorite, and Hittite dominance and from the Seleucid period. Deciphering them has not been an easy task, although the work was made simpler by mutilingual tablets that were created because the cuneiform writers themselves had need to know what had been written in earlier languages. It was not until 1854 that enough tablets had been deciphered to reveal the system of computation used, and not until the early twentieth century were significant numbers of mathematical texts deciphered and analyzed. The most complete analysis of these is the 1935 two-volume work by Otto Neugebauer (1899–1992), *Mathematische Keilschrifttexte*, recently republished by Springer-Verlag. Neugebauer also wrote a popular exposition of both Babylonian and Egyptian science under the title *The Exact Sciences in Antiquity*, a book highly recommended for the general reader.

To discuss the clay tablet texts we shall consider the following questions:

1. What kinds of problems are addressed?

2. What systematic procedures are used to solve these problems?

3. What procedures are taken for granted, as something the reader would automatically know?

4. For what purpose was the writer engaged in doing mathematics?

3.3 The Number System

The most striking discovery about the Babylonian number system was that the cuneiform writers used a positional system based on 60. Digits up to 9 are represented by vertical strokes, and then the number 10 is represented by a boomerang-shaped figure. Thus far the system resembles the Egyptian, and we might expect that new symbols would be invented for 100, 1000, etc. Here is where the surprise comes in. Numbers are written using the symbols for 1 and 10 only up to 59. For the number 60 the vertical stroke is repeated, and thereafter numbers are recorded more or less as we record them in the decimal system, using the physical location of a symbol in relation to other symbols as the guide to its value. For example, using our system, in the number 372 the 7 stands for 7 tens, that is, 70, and the 3 stands for 3 hundreds, that is, 300. The cuneiform system is similar, except that (1) places represent powers of 60 instead of powers of 10 and (2) there is no symbol for 0 at the end of a number and, in the earlier texts, no symbol for 0 between two other digits. Hence there is no way to distinguish 73 from 703, except by context. This much turned out to be fairly easy to infer from the tablets because many of them bear symbols in a precise logical order that can only mean the tablets were multiplication tables.

Since we wish to look at "Babylonian" mathematics on its own terms but not to handicap ourselves by having to use the Babylonian symbols, we shall adopt a compromise similar to the one we used for representing the Egyptian *parts*. We shall write numbers through 59 with our standard symbols, but larger numbers will be written in the sexigesimal system, separating the places with commas and putting a semicolon where our system would use a decimal point. Thus we shall write the number 193 as 3, 13 (meaning $3 \times 60 + 13$) and the number 7275 as 2, 1, 15 (meaning $2 \times (60)^2 + 1 \times 60 + 15$). Sexigesimal fractions are particularly easy. Thus what we would write as $\frac{1}{4}$ can be written as ;15 (meaning 15 sixtieths), etc. Various conjectures have been advanced as to the origin of this system. The most logical explanation seems to be that a people counting by tens came into contact with a people counting by twelves, since the least common multiple of 10 and 12 is 60. We have already seen that the Egyptians divided weeks into 10 days, but years into 12 months and day and night into 24 hours. Thus the basis for a sexigesimal system—the simultaneous use of tens and twelves—was present in Egypt as well. Commercially this basis exists in American society also, where feet are divided into 12 inches, and eggs and pencils are sold by the dozen, yet the currency is decimalized. It is known that the monetary system in use throughout Mesopotamia in early times involved division by 60 (60 shekels = 1 mina, 60 minae = 1 talent).

3.4 Babylonian Arithmetic

3.4.1 General Features

Tablets from the site of Senkereh (also known as Larsa), kept in the British Museum, contain tables of products, reciprocals, squares, cubes, square roots, and

cube roots of integers. It appears that the people who worked with mathematics in the civilizations we are discussing learned by heart, just as we do, the products of all the small integers. Of course for them a theoretical multiplication table would have to go as far as 59×59, and the consequent strain on memory would be large (that fact may account for the existence of so many written tables). Just as most of us learn, without being required to do so, that $\frac{1}{3} = .3333\ldots$, the Babylonians wrote their fractions usually as sexigesimal fractions and came to recognize certain reciprocals, for example $\frac{1}{9} = 0;6,40$. As with multiplication, the labor involved for 60 reciprocals is large. Moreover the reciprocal of 7 is a problem in both decimal and sexigesimal notation in that its expansion never terminates. With a system based on 30 or 60, all the numbers less than 10 except 7 have terminating reciprocals. In order to get a terminating reciprocal for 7 one would have to go to a system based on 210, which is of course out of the question.

3.4.2 Things "Everybody Knew" in Babylon

Not only are sexigesimal fractions handled easily in all the tablets; the concept of a square root occurs explicitly, and actual square roots are approximated by sexigesimal fractions, showing that the mathematicians of the time realized that they hadn't been able to make these square roots come out even. (Whether they realized that the square root would never come out even is not clear.) For example, text AO 6484 (the AO stands for Antiquités Orientales) from the Louvre in Paris contains the following problem on lines 19 and 20:

> The diagonal of a square is 10 Ells. How long is the side? [To find the answer] multiply 10 by 0;42,30. [The result is] 7;5.

Now $0;\ 42,\ 30$ is $\frac{42}{60} + \frac{30}{3600} = \frac{17}{24} = 0.7083$, approximately. This is a very good approximation to $1/\sqrt{2} \approx 0.7071$, and the answer $7;5$ is, of course, $7\frac{1}{12} = 7.083 = 10 \times 0.7083$. The writer of this tablet seems to have known that the ratio of the side of a square to its diagonal is approximately $\frac{17}{24}$. It is rather intriguing that the approximation to $\sqrt{2}$ that arises from what is now called the Newton–Raphson method turns up the number $\frac{24}{17}$ in obtaining the third approximation. The method of approximating square roots can be understood as the following procedure. Since 1 is smaller than $\sqrt{2}$ and 2 is larger, let their average be the first approximation, that is, $\frac{3}{2}$. This number happens to be too large to be $\sqrt{2}$, but it is not necessary to know that fact to improve the approximation. Whether it errs by being too large or two small, the result of dividing 2 by this number will err in the other direction. Thus, since $\frac{3}{2}$ is too large, $\frac{2}{3/2} = \frac{4}{3}$ is too small to be $\sqrt{2}$. It therefore seems likely that the average of these two numbers will be closer to $\sqrt{2}$ than either number, that is, our second approximation to $\sqrt{2}$ is $\frac{1}{2}\left(\frac{3}{2} + \frac{4}{3}\right) = \frac{17}{12}$. Again whether this number is too large or two small, the number $\frac{2}{17/12} = \frac{24}{17}$ will err in the opposite direction, so that we can average the two numbers again and continue this process as long as we like.

The writers of these tablets realized that when numbers are combined by multiplying, adding, etc., it may be of interest to know how to recover the original

data from the result. This realization is the first step toward attacking the problem of inverting binary operations. Such a problem leads to the solution of quadratic equations in our time, although one may question whether the Babylonians solved equations. Their approach to this problem was to associate with every pair of numbers, say 13 and 27, two other numbers, namely their average ($\frac{13+27}{2} = 20$) and their *semidifference*[1] ($\frac{27-13}{2} = 7$). The average and semidifference can easily be calculated from the two numbers, and likewise the original data can be calculated from the average and semidifference. The larger number (27) is the sum of the average and semidifference: $20 + 7 = 27$, and the smaller number (13) is their difference: $20 - 7 = 13$. The realization of this mutual connection makes it possible essentially to "change coordinates" from the number pair (a, b) to the pair $\big((a+b)/2, (a-b)/2\big)$.

At some point lost to history some Babylonian mathematician came to realize that the product of two numbers is the difference of the squares of the average and semidifference: $27 \times 13 = (20)^2 - 7^2 = 351$ (or 5, 51 in Babylonian notation). This principle made it possible to recover two numbers knowing their sum and product or knowing their difference and product. For example, given that the sum is 10 and the product is 21, we know that the average is 5 (half of the sum), hence that the square of the semidifference is $5^2 - 21 = 4$. Therefore the semidifference is 2, and so the two numbers are $5 + 2 = 7$ and $5 - 2 = 3$. Similarly, knowing that the difference is 9 and the product is 52, we conclude that the semidifference is 4.5 and the square of the average is $52 + (4.5)^2 = 72.25$. Hence the average is $\sqrt{72.25} = 8.5$. Therefore the two numbers are $8.5 + 4.5 = 13$ and $8.5 - 4.5 = 4$. The two techniques just illustrated occur constantly in the cuneiform texts, and were clearly taken to be procedures familiar to everyone, requiring no explanation.

3.4.3 Applications

Like all mathematicians at all times, having discovered some basic principles that are useful for solving a limited number of problems, the Babylonian mathematicians tried to extend these principles to the limit of their applicability. In so doing they were able to reduce a large number of problems to the form in which the sum and product or the difference and product of two unknown numbers are given. We shall consider just one example, a famous one that has been written about by many authors. The problem in question occurs on a tablet from the Louvre in Paris, known as AO 8862.

A loose translation of the text of this tablet (made from Neugebauer's German translation) reads as follows:

---

[1] This word is coined because English contains no one-word description of this concept, which must otherwise be described as half of the difference of the two numbers. It is clear from the way in which the number constantly occurs that the original writers of these tablets automatically looked at this number along with the average when given two numbers as data. However, there seems to be no word in the Akkadian, Sumerian, and ideogram glossary given by Neugebauer to indicate that the writers of the clay tablets had a special word for these concepts. In the translations given by Neugebauer they are obtained one step at a time, by first adding or subtracting the two numbers, then taking half of the result.

> I have multiplied the length and width so as to make the area. Then I added to the area the amount by which the length exceeds the width, obtaining 3,3. Then I added the length and width together, obtaining 27. What are the length, width and area?
>
> 27 3,3 the sums
> 15 length
> 3,0 area
> 12 width
>
> You proceed as follows:
>
> Add the sum (27) of the length and width to 3,3. You thereby obtain 3,30. Next add 2 to 27, getting 29. You then divide 29 in half, getting 14;30. The square of 14;30 is 3,30;15. You subtract 3,30 from 3,30;15, leaving the difference of 0;15. The square root of 0;15 is 0;30. Adding 0;30 to the original 14;30 gives 15, which is the length. Subtracting 0;30 from 14;30 gives 14 as width. You then subtract 2, which was added to the 27, from 14, giving 12 as the final width...

The text continues, verifying that these numbers do indeed solve the problem. Naturally, this text requires some commentary. Indeed most students find it completely baffling at first. Knowing the general approach of the Babylonian mathematicians to problems of this sort, one can understand the reason for dividing 29 in half (so as to get the average of two numbers) and the reason for subtracting 3,30 from the square of 14;30 (as we saw above, the difference between the square of the average and the product will be the square of the semidifference of the two numbers whose sum is 29 and whose product is 3,30, that is, 210). What is not clear is the following: Why add 27 to the number 3,3 in the first place, and why add 2 to 27?

Let us see if we can get inside the head of the unknown author of this problem. Van der Waerden (in *Science Awakening*) claimed that the operation of adding the difference between length and width to area, being a logically absurd operation (since one cannot add things of different types), was proof that this is a problem of pure algebra. He further noted that the words for length (*us*) and width (*sag*) are indeclinable, and claimed that they were being used like variables x and y in modern mathematics. On that basis, concluded Van der Waerden, "we may safely set [this problem] down as a set of two equations in two unknowns." Now it is quite true that if you do what Van der Waerden suggested, you will obtain the equations

$$\begin{aligned} xy + (x - y) &= 183 \\ x + y &= 27. \end{aligned}$$

If you then add the resulting equations, as seems to be indicated by adding the two numbers given in the problem, the resulting equation

$$xy + 2x = 210,$$

does indeed say that the product of two more than the width $(y + 2)$ and the length (x) is 3,30. This result suggests that we consider a new problem with width increased by 2. In that new problem we know that the product of length and width is 3,30 and the sum of length and width is 29. We have thus arrived at a new problem that can be thought of as a "standard form" problem: find two unknown numbers given their product and their sum (in this case 210 and 29, respectively). Undoubtedly we are seeing here an early example of a procedure mathematicians undertake constantly: to compile as many variants as possible of certain standard problems, so as to reduce the number of solution techniques that must be learned to a minimum. The question is whether the original writer wrote down these symbolic expressions and added them in order to obtain the reduced, standard-form problem.

It is clear that some process closely related to Van der Waerden's description must have been what the author had in mind, since the rest of the problem is merely a repetition of the standard procedure for finding two numbers given their sum and product, as discussed above. Nevertheless the author need not have been thinking of length, width, and area as abstract variables. It is quite possible that some geometric interpretation was used as a guide to the solution. Particularly when thinking of the numerical value of a length rather than the abstract concept of length, one might well have used the word *length* as a shorthand way of referring to a rectangular *area* of the given length and having width equal to 1. If so, adding the difference between length and width to the area would have meant constructing the upper gnomon (L-shaped region) in Fig. 3.1, while "the sum of length and width" would represent the lower gnomon. It is then clear that the two gnomons fit together to form a rectangle whose length is that of the original rectangle, but whose width is larger by 2 units. In favor of this interpretation we can point out that when subtracting the 2 after solving the transformed problem the author subtracts it from the new width (14) rather than from the length (15). Algebraically this 2 might just as well have been subtracted from the 15, giving the solution of length 14 and width 13, but to do so one would have to regard the length in the original problem as the width in the new problem and the width in the original problem as 2 less than the length in the new problem. The fact that the author doesn't do this suggests that the words *width* and *length* retained some of their concrete geometric interpretation, and that perhaps the author had a picture like Fig. 3.1 in mind when thinking about this problem.

3.4.4 The Nature of Babylonian Algebra

The question whether the Babylonian mathematicians developed algebra is a matter of definition. What does *elementary algebra* mean to you? If it means the study of *equations* written using abstract symbols for the unknown numbers and solved using a fixed set of algorithms, then almost certainly the Babylonians did *not* develop algebra. On the other hand, what are you really doing when you solve, say, a quadratic equation

$$ax^2 + bx + c = 0?$$

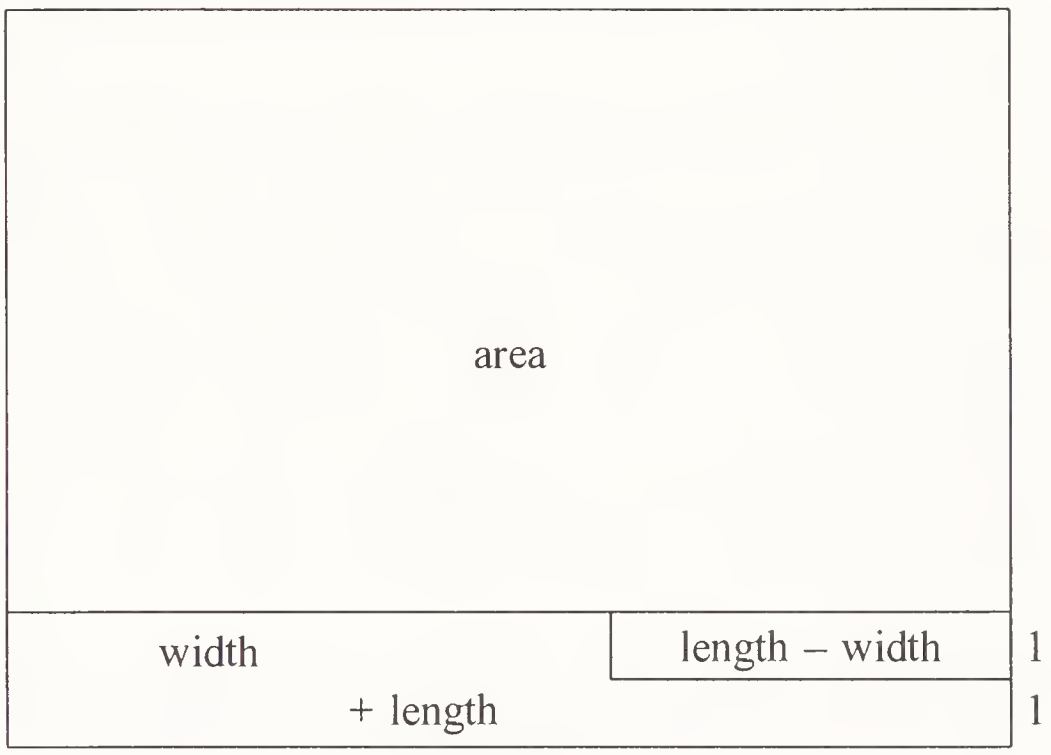


Figure 3.1: Reduction of a problem to standard form.

What data (input) must you have to solve the problem, and how is the solution (output) related to the data? It seems at first sight that the data for the problem are a, b, and c, but this is misleading. The equation is not a quadratic equation if $a = 0$; and if $a \neq 0$, you can divide by a, getting an equation with the same roots, but only two bits of data, namely

$$x^2 + Ax + B = 0,$$

where $A = b/a$ and $B = c/a$. The data for the problem are therefore A and B. The solution consists of two roots r_1 and r_2. How are these roots related to the data? Well, since the equation is

$$0 = (x - r_1)(x - r_2) = x^2 - (r_1 + r_2)x + r_1 r_2,$$

we see that $A = -(r_1 + r_2)$ and $B = r_1 r_2$. Hence, in essence *the problem of solving a quadratic equation is the problem of finding two numbers when given their sum and their product.*

In this form, as we have seen, the Babylonians certainly did develop algebra. And, as Neugebauer pointed out, the old Sumerian ideograms served a very important function as mathematical symbols. For that reason he transcribed ideograms as ideograms in his German edition of the texts, even in cases when it was clear that they were *read* as if they were Akkadian.

Let us look at this question from yet another point of view. In our educational system the break between arithmetic and algebra is clear. In arithmetic one uses symbols for numbers, but a given symbol such as the ideogram 5 always represents the same (known) number. The transition from arithmetic to algebra is marked by the use of letters instead of number symbols; moreover, each of the new letter

symbols stands for a variable or an unknown number, which will change from one setting to another. This change defines the difference between algebra and arithmetic for most people.

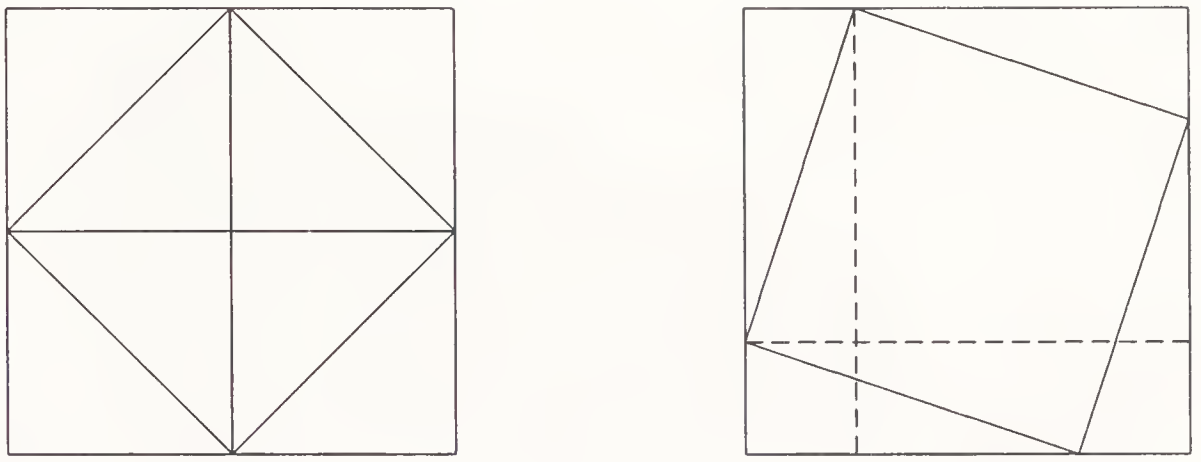
A more fundamental difference between arithmetic and algebra, however, is the following. Arithmetic consists of certain unary and binary operations such as addition, subtraction, multiplication, division, squaring, cubing, and taking square and cube roots. Whenever a sequence of arithmetic operations is performed on a set of data in order to produce a set of results, the question naturally arises how one could go the other direction and recover the original data from the results. If we define arithmetic to be the process of applying these standard operations to known numbers, we can define elementary algebra to be the study of the opposite problem of recovering the input when the operations applied and the output are known.

With this second definition it can certainly be said that the Babylonians did algebra. We have already seen how they recovered two numbers given their sum and product or their difference and product. A further example may be provided by certain tablets that give the sum of the squares and the cubes of integers. These tablets may have been used for finding the numbers to which this operation was applied in order to obtain a given number. In our terms these tablets make it possible to solve the equation $x^3 + x^2 = a$, a very difficult problem indeed.

3.5 Babylonian Geometry

As in Egyptian geometry, the primary problem in Babylonian geometry is that of area and volume. In contrast to the case of Egypt, however, we have hard proof that the Babylonians knew the Pythagorean theorem in full generality at least a thousand years before Pythagoras. They were thus already on the road to finding more abstract properties of geometric figures than mere size. How might they have discovered the Pythagorean theorem? The following fanciful story (based on Plato's dialogue *Meno*) is merely a plausible way in which a person of some mathematical ability might have made this discovery. We preface the story with a word of warning, however. Please keep in mind that the story is only a conjecture, not to be made into a "fact" like Cantor's story of the Egyptian "application" of the Pythagorean theorem. It gives the *psychologically* simplest derivation of the Pythagorean theorem. As any mathematician knows, however, theorems are often discovered in a way that is not psychologically the simplest. Once a theorem is stated and proved, it often happens that a much simpler proof is discovered. One should therefore not infer that the following explanation is "the way it happened."

Suppose that for purposes of surveying, taxation, or amusement you find it necessary to construct a square twice as large as a given square. How would you go about doing so? You might double the side of the square, but you would soon realize that doing so actually quadruples the size of the square. If you drew out the quadrupled square and contemplated it for a while, you might be led to join the midpoints of its sides in order, that is, to draw the diagonals of the four copies of the original square. Since these diagonals cut the four squares in half, they will



(a) (b)

Figure 3.2: Doubling a square (a); The Pythagorean theorem (b).

enclose a square twice as big as the original one (Fig. 3.2). Now it is not at all unlikely that someone, either for practical purposes or just for fun, discovered this way of doubling a square. If so, someone "playing" with the figure, might have considered the result of joining in order the points at a given distance from the corners of a square instead of joining the midpoints of the sides. Doing so creates a square in the center of the larger square surrounded by four copies of a right triangle whose hypotenuse equals the side of the center square (Fig. 3.2); it also creates the two squares on the legs of that right triangle and two rectangles that together are equal in area to four copies of the triangle. (In Fig. 3.2 one of these rectangles is divided into two equal parts by its diagonal, which is the hypotenuse of the right triangle.) Hence the larger square consists of four copies of the right triangle plus the center square. It also consists of four copies of the right triangle plus the squares on the two legs of the right triangle. The inevitable conclusion is that *the square on the hypotenuse of any right triangle equals the sum of the squares on the legs*. This is the Pythagorean theorem, and it is used in many places in the cuneiform texts. Whether or not this fictitious story is the way in which this theorem came to be discovered, it is a fact that the figure on the left in Fig. 3.2, and possibly the one on the right as well, occurs on the British Museum tablet BM 15 285, which dates from the period of the Amorite civilization (Fig. 3.3).

In general it can be said that Babylonian geometry, like its Egyptian counterpart, was regarded more as an application of mathematics than as mathematics proper. The primary emphasis was on areas and volumes, and the cuneiform tablets contain computations of some of the same volumes (frustum of a pyramid, for example) that are computed in the Ahmose Papyrus. As Neugebauer puts it, " 'geometry' is no special mathematical discipline, but is treated on an equal level with any other form of numerical relation between practical objects." We have seen that the Egyptian method of finding the area of a circle was to square eight-ninths of the diameter (corresponding to a value of π equal to $\frac{256}{81}$). The commonest method in the Babylonian tablets was to take 3 times the square on the radius (corresponding to $\pi = 3$), although a later tablet used the value $3\frac{1}{8}$. For a general quadrilateral the procedure for finding area was to take the product of the averages of the opposite sides, again showing the prominence of the average of two quantities in

Figure 3.3: Cuneiform text BM 15 285. Copyright British Museum.

Babylonian mathematics. This formula, however, is not exact from the point of view of Euclidean geometry; it is a good approximation only for quadrilaterals whose angles are nearly right angles, so that the opposite sides are nearly equal.

Volumes were handled similarly. For example, in many cases the volume of the frustum of a pyramid was found (incorrectly) by taking the average of the upper and lower bases and multiplying by the height, although the correct procedure is also found in some tablets.

3.6 Astronomy and the Calendar

Since one of the earliest scientific applications of mathematics, involving first numbers, then geometry, was to astronomy, the history of mathematics in early times cannot be told completely without some reference to astronomy. At the stage of history now being discussed astronomy is largely a matter of arithmetic: counting the days between full moons, solstices, etc. Keep in mind that the Copernican system that forms the basis of our picture of the universe is an imaginative construct

suggested only by sophisticated mathematical reasoning. Ordinary observation suggests a geocentric universe, and that is what we shall now discuss.

Anyone not prejudiced by "book-learning" who observed the sky carefully over a period of time would notice several phenomena that need to be explained.

1. The vast majority of stars seem to rotate in perfect circles at a constant rate of speed about a fixed axis whose direction, by definition, is north–south. Thus they can be pictured as stuck to a large sphere with its center in the earth; this sphere is called the *celestial sphere*.

2. The sun seems to fall behind the stars by a very small amount each day. In addition the sun moves north and south in a cyclic pattern, each cycle taking one year. (Here we assume that the term *year* has a meteorological meaning in terms of the air temperature; otherwise this cycle is merely the definition of a tropical year.) The path of the sun through the stars does not seem to vary from year to year. It is a great circle on the celestial sphere (called in modern astronomy the *ecliptic*).

3. The moon falls even farther behind the stars than does the sun. In fact the sun passes the moon every 29 or 30 days. The lighted face of the moon changes regularly, showing that it shines by reflecting the light of the sun. The moon follows nearly the same path through the stars (the ecliptic) as the sun, but wobbles above and below this path. When the moon crosses this path in its full phase, it is eclipsed. The sun is also eclipsed at certain times by the moon, and this happens only when the moon crosses the ecliptic in its new phase. However, there are times when the moon crosses the ecliptic in its new phase, yet no eclipse is observed.

4. There are five "wandering" stars, brighter than average, whose motions through the sky are very irregular. They all seem to follow roughly the same route through the stars as the sun and moon. However, unlike the sun and moon, which always move eastward relative to the stars, these *planets*, from the Greek word *planan* ($\pi\lambda\alpha\nu\acute{\alpha}\nu$), meaning *to wander* or *go astray*, sometimes stop and move westward for a few weeks, then resume their eastward march.

These phenomena would be of interest to anyone with a curiosity about the world, and some of them would have practical effects as well. The main practical effects from the early period will be discussed at this point. We shall reserve the more sophisticated Seleucid period astronomy to be discussed along with Greek astronomy. The main point to be noticed in these "early days" is that geometry is not really involved. The whole procedure for predicting the motions of the heavenly bodies is purely a matter of counting days between the recurrence of various phenomena—usually conjunction with the sun or first visibility after a conjunction. For the moon these conjunctions (new moons) marked the beginning of the months by which religious festivals and civil holidays were regulated. Where the economy depended on agriculture, and hence on the solar year, it was important

to keep these months in harmony with the years. The planetary phenomena have no practical importance, but were nevertheless observed and counted because they were thought to have an influence on human destiny (astrology).

The main item of practical importance in early astronomy is the establishment of the relation between a lunar month and a solar year. Normally there are 12 full moons in each solar year. However, in ancient times months and years were very much out of balance, and for civil purposes kings had to declare a thirteenth (intercalary) month occasionally in order to maintain the balance. This practice was still being carried out as late as the middle sixth century B.C.E., with the Chaldean king Nabonadi and the Persian kings Cyrus and Cambyses declaring intercalary months when required. However, during the reign of the Persian king Artaxerxes II (404–358 B.C.E.) a regular lunar calendar was declared, having a 19-year cycle consisting of 12 years of 12 months and 7 years of 13 months. This cycle is still a feature of lunar calendars today. (An error of one day will accumulate on this calendar in 6840 years.)

We know indirectly that the Mesopotamians bequeathed a large legacy of astronomical observations to the Hellenistic natural philosophers, since the astronomer Claudius Ptolemy, who lived in the second century C.E., in choosing an epoch (time zero) for computing the mean motion of the sun, says in passing that he uses the reign of Nabonassar for this purpose, since "that is the era beginning from which the ancient observations are, on the whole, preserved to our own time."

The Nabonassar mentioned by Ptolemy is an Assyrian king whose 14-year reign began in 746 B.C.E. The records used by Ptolemy have not come down to us through any other channel. However, large numbers of astronomical and mathematical tablets written in Uruk and Babylon during the Seleucid period have been unearthed, and these tablets give us a glimpse of two clever schemes—theories in the modern sense of unifying systems of explanation for natural phenomena—for predicting the motion of the sun, moon, and planets. We shall reserve these theories for a later chapter. The time period in which the Ptolemaic theory was developed is later than the period of this chapter, and the problems involve astronomical phenomena somewhat more subtle than those listed above, such as the fact that some new moons are 29 days apart and others 30 days apart, and that the sun's motion along the ecliptic is not at constant velocity (using the stars as the standard of constant velocity). At this point we shall merely note that by the reign of Artaxerxes II ordinary observation had established that 19 solar years are almost exactly equal to 235 lunar months ($235 = 12 \cdot 12 + 7 \cdot 13$). From the year 380 B.C.E. onward this cycle was the basis of the lunar calendar used in Mesopotamia.

3.7 Problems and Questions

3.7.1 Problems in Babylonian Mathematics

Exercise 3.1 Write the number 345.75 in the sexigesimal notation adopted in this chapter.

Exercise 3.2 Find two numbers whose sum is 5 and whose product is $\frac{56}{9}$ using the procedure sketched in the text.

Exercise 3.3 Find two numbers whose difference is $\frac{1}{2}$ and whose product is $\frac{99}{16}$ using the procedure in the text.

Exercise 3.4 Solve the following problem from the cuneiform tablet BM 85 196, dating from the time of the Hittite civilization or perhaps earlier (numbers in square brackets have been reconstructed, having been effaced from the tablet itself). Then explain the author's solution.

> A beam of length 0;30 GAR is leaning against a wall. Its upper end is 0;6 GAR lower than it would be if it were perfectly upright. How far is its lower end from the wall?
>
> Do the following: Square 0;30, obtaining 0;15. Subtracting 0;6 from 0;30 leaves 0;24. Square 0;24, obtaining 0;9,36. Subtract 0;9,36 from [0;15], leaving 0;5,24. What is the square root of 0;5,24? The lower end of the beam is [0;18] from the wall.
>
> When the lower end is 0;18 from the wall, how far has the top slid down? Square 0;18, obtaining 0;5;24. . . .

3.7.2 Questions about Babylonian Mathematics

Exercise 3.5 What do the two problems of recovering two numbers from their sum and product or from their difference and product have to do with quadratic equations as we understand them today? Can we conclude that the Babylonians "did algebra"?

Exercise 3.6 You can easily verify that the solution of the problem from tablet AO 8862 (15 and 12) given by the author is not the only possible one. The numbers 14 and 13 will also satisfy the conditions of the problem. Why didn't the author give this solution?

Exercise 3.7 Of what practical value are the problems we have called "algebra"? Taking just the quadratic equation as an example, the data can be construed as the area and the semiperimeter of a rectangle and the solutions as the sides of the rectangle. What need, if any, could there be for solving such a problem?

Exercise 3.8 Read Plato's dialogue *Meno*, in which the problem of doubling a square is discussed. Note that mathematics is not the point of the dialogue. What *is* the role of mathematics in this dialogue?

3.8 Endnotes

1. Much of this chapter is based on the work of Neugebauer, whose books *The Exact Sciences in Antiquity*, *Mathematische Keilschrifttexte*, and *A History*

of Ancient Mathematical Astronomy are masterpieces of scholarship and beautiful writing. In particular the first of these, which is aimed at the general educated reader, is a fascinating concise account of mathematics and astronomy in Egypt and Mesopotamia and should be read before any other book on the subject.

2. The quotation from the tablet AO 6484 is taken from *Mathematische Keilschrifttexte*, Part 1, p. 100.

3. Van der Waerden's discussion of AO 8862 is on pp. 63–65 of *Science Awakening*.

4. Neugebauer's remark on ancient geometry is taken from *The Exact Sciences in Antiquity*, p. 44.

5. The quotation from Ptolemy is taken from the recent edition of the *Almagest* edited by G. J. Toomer (Springer-Verlag, New York, 1984), p. 166.

Chapter 4

The Early Greeks

4.1 Introduction

The last two chapters have sketched some of the mathematical knowledge that existed around the shores of the Mediterranean Sea in the two millennia B.C.E. Many different peoples shared in this knowledge and helped to create it. Its transmission to us, however, came through a particular people, the Greeks, who introduced an innovation that has no parallel in any other time or place: the creation of systematic, logically deductive mathematical theories. Logic and common-sense reasoning are inherent in all of mathematics, of course, but a meticulous attention to assumptions and rules of inference, controlling the exposition of the subject, is uniquely Greek. All other peoples solved problems; the Greeks proved theorems. The word *theorem* is probably related to *theasthai* ($\theta\varepsilon\hat{\alpha}\sigma\theta\alpha\iota$), meaning *to see*. The value of such an approach, not only for avoiding inconsistency, but also for suggesting new paths to explore, will be seen throughout the rest of this book. Its universal appeal to mathematically inclined minds is attested by the fact that the mathematicians of China, India, and Japan eventually began to practice this style of mathematics after they became aware of it through contact with Western traders and missionaries. The direct intellectual heirs of the Greeks, the medieval Muslims and modern Europeans, created geometry in the style of Euclid, and attempted to give their own mathematical innovations in other areas (algebra and calculus) the same logical rigor that they found in geometry.

The complexity of society should make us suspicious of any facile explanations of a phenomenon such as Greek geometry. One important factor contributing to the appearance of this innovation among the Greeks, however, is the existence of schools of philosophy in the Greek city-states. These schools, in turn, can be explained at least partly by the fact that some of the Greeks grew wealthy through commerce with foreigners rather than having control of the resources and labor of their own land, as was the case in Egypt. Effective commerce requires realism in dealing with the world. It also leads to questioning one's own traditions and mores. These ingredients, combined with the leisure time for reflection made possible by

increased wealth, encourage a probing, philosophical outlook on all aspects of life. Whether anything more is required to explain the uniqueness of Greek geometry is a question left to the reader.

4.2 Sources

Before we embark on an exploration of Greek mathematics, however, we need to say a few words about the sources of our knowledge. The mathematics we shall be discussing in this chapter and the four following was created in the millennium from 500 B.C.E. to 500 C.E. It is therefore considerably later than the mathematics we have discussed up to now. Paradoxically, our sources for the later mathematics are less direct than those from the earlier times. Egyptian papyri and Mesopotamian clay tablets have proved to be much more durable than the materials on which Greek mathematics was written. The oldest surviving copies of almost any Greek treatise in mathematical science are typically no more than 1000 years old. These documents are, with good reason, considered to be faithful copies of the originals, which were written in some cases more than a thousand years earlier. Errors do creep in, however, and careful scholarship is required to establish the authentic text of any of the famous Greek treatises. Moreover, some documents did not survive at all, and their existence is known only because they are mentioned by later commentators. Indeed a major part of what is known about early Greek mathematics is due to quotations and summaries in the work of later commentators. Here are some of the more important of these commentators. All dates given for their lives are only approximate.

1. Marcus Vitruvius (first century B.C.E.) was a Roman architect who wrote an extremely influential treatise on architecture in 10 books. He is regarded as a rather unreliable source for information about mathematics, however.

2. Plutarch (45–120 C.E.) was the author of the *Parallel Lives of the Greeks and Romans*, in which he compares famous Greeks with eminent Romans who engaged in the same occupation, such as the orators Demosthenes and Cicero. Shakespeare relied on his account of the lives of many people, for example Julius Caesar, even describing the miraculous omens that Plutarch reported as having occurred just before Caesar's death. Plutarch is important to the history of mathematics for what he reports on natural philosophers such as Thales.

3. Theon of Smyrna (ca. 100 C.E.) was the author of an introduction to mathematics written as background for reading Plato, a copy of which still exists. It contains many quotations from earlier authors.

4. Diogenes Laertius (third century C.E.) wrote a comprehensive history of philosophy, *Lives of Eminent Philosophers*, which contains summaries of many earlier works and gives details of the lives and work of many of the pre-Socratic philosophers.

5. Iamblichus (285–330 C.E.) was the author of many treatises, including 10 books on the Pythagoreans, 5 of which have been preserved.

6. Pappus (ca. 300 C.E.) wrote many books on geometry, including a comprehensive treatise of eight mathematical books. He is immortalized in calculus books for his theorem on the volume of a solid of revolution. Besides being a first-rate geometer in his own right, he wrote commentaries on the *Almagest* of Ptolemy and the tenth book of Euclid's *Elements*.

7. Proclus (412–485 C.E.) is important for our story as the author of a commentary on the first book of Euclid, in which he quotes long passages from a history of mathematics (now lost) by Eudemus, a pupil of Aristotle.

8. Simplicius (500–549 C.E.) was a commentator on philosophy. His works contain many quotations from the pre-Socratic philosophers.

9. Eutocius (ca. 700 C.E.) was a mathematician who lived in the port city of Askelon in Palestine and wrote an extensive commentary on the works of Archimedes.

Most of these commentators wrote in Greek. Knowledge of Greek sank to a very low level in western Europe as a result of the upheavals of the fifth century. Although learning was preserved by the Church and much of the Bible was written in Greek, a Latin translation (the Vulgate) was made by Jerome in the fifth century. From that time on Greek documents were preserved mostly in the Eastern (Byzantine) Empire. After the Muslim conquest of North Africa and Spain in the eighth century some of these Greek documents were translated into Arabic and circulated in Spain and the Middle East. From the eleventh century on, as secular learning began to revive in the West, scholars from England, France, and Germany made journeys to these centers and to Constantinople, copied out manuscripts, translated them from Arabic and Greek into Latin, and tried to piece together some long-forgotten parts of ancient learning. We shall encounter these Medieval and Renaissance humanists in later chapters. The task they began will never be complete, and scholars continue to seek more old manuscripts right down to the present day. To mention just one example, in 1906 the Danish scholar J. L. Heiberg (1854–1928), who established the definitive text of many Greek scientific treatises, investigated a report of a mathematical treatise in Constantinople. This treatise proved to be a long-lost work of Archimedes, revealing that the great mathematician did not adhere exclusively to the rigorous principles of the Euclidean tradition, but possessed in addition an intuitive and highly fruitful method of discovery. A more recent example is the discovery of an Arabic version of four books of Diophantus' *Arithmetike*.

It is thus to the later commentators and the labors of antiquarian scholars since late Medieval times that we owe most of what we know about the ancient Greeks. We are forced to trust expert authority for most of what we believe. Therefore it is somewhat misleading to say that we personally *know* the things we believe. Rather we *know* that authoritative scholars have made certain assertions, and we *trust* those

Figure 4.1: Architecture uses mathematics both to solve practical construction problems and to create structures of great beauty. A classical European illustration of this phenomenon is the Parthenon, whose front is shown here. The Bettmann Archive.

scholars. Only a few experts, each in a narrow field, truly know anything, in the sense of being able to summarize the evidence on which the knowledge is based. This situation, though unavoidable, is nevertheless disquieting—after all, scholarly opinion does change, sometimes even reversing itself. Moreover it is unsatisfying to realize that what we really know is not, for example, that Thales predicted an eclipse, but that documents believed to be authentic by scholars report that Thales predicted an eclipse.

Since the life story of the documents themselves is incredibly complicated in some cases, we shall give just one or two examples. After that we shall merely report what the experts say as fact. The treatise now known as Ptolemy's *Almagest*, which was written about 150 C.E. and was the standard treatise on astronomy for nearly 1400 years, exists in nine manuscripts, the earliest of which was written in the ninth century, in other words 700 years after the original of which it is presumably a copy. The story is the same for most other works: the extant manuscripts are copies, sometimes translations into Arabic or Latin of works originally written in a form of Greek that had become a dead language centuries earlier.

The consequences of these facts are twofold: first, anyone who wishes to study "Greek mathematics" seriously has to know Arabic; second, it is necessary to compare as many independent manuscripts as can be found in order to compensate for scribal errors, and even then it is obvious from common sense that many copy errors (due to the imperfect understanding of the copyists) have crept in. For

example, G. J. Toomer, who has recently edited and commented an English edition of the *Almagest*, has said that, "there are whole classes of textual matter which must... be regarded as interpolations."

The question of authenticity of sources is very vexed; and if we were to pursue it in every case, we would wind up in a labyrinth without a torch to light our way. Without years of training in the relevant languages and an intimate knowledge of the issues involved, our judgments as to authenticity would be worthless. We therefore take as given the documents that have been worked up by historians who specialize in these areas. In so doing we recognize that our knowledge is based on certain assumptions that we cannot ourselves justify.

4.3 The Beginnings of Greek Mathematics

The period we are calling "Early Greek" is actually rather late in the history of ancient Greece. The Greeks invaded and settled the Greek peninsula in several waves some time during the second millennium B.C.E. The period of time we are concerned with in the present chapter begins with the philosopher Thales around 600 B.C.E. and ends around the time of Plato in 350 B.C.E. This period is marked by an increasing sophistication of mathematics, leaving a number of intriguing problems to be solved by future generations of mathematicians.

In this chapter our emphasis is still on arithmetic and geometry, especially the latter. Algebra in the form known to us is still far in the future. Although the mathematics involved is much more sophisticated than the mathematics of Egypt and Mesopotamia, the study of it becomes easier because it is possible to detect certain unifying themes. Among these themes are (1) the beginnings of proof and formal deduction, (2) a systematic attempt to develop a theory of proportion, and (3) an attempt to use mathematics to construct scientific theories.

As mentioned in Chapter 1, the concept of proportion plays a very large role in mathematics and its applications. If there is a single thread that runs through the history of mathematics from earliest times right down to the twentieth century, that thread is the idea of proportion. We have already seen that the notion is implicit in the Egyptian method of multiplication and in the *pesu* and *seked* problems of the Ahmose Papyrus. Much of Greek geometry is an attempt to discover proportions between different geometric figures. In other situations where the notion of proportion is not immediately applicable, we find attempts to apply it indirectly through procedures like the "Merton rule" (which you will meet when we study Medieval European mathematics). The calculus can be understood as an attempt to apply direct proportion locally and approximately when it does not apply over larger intervals. The whole subject of linear transformations of vector spaces is an elaboration of the idea of proportion, and even in the sophisticated problems of twentieth-century mathematics a frequently applied technique for studying complicated dynamic systems is to begin with a "linear" approximation, that is, one in which the idea of proportion is applicable.

Now proportion is applicable to both arithmetic and geometry. Four numbers are in proportion $a : b :: c : d$ if the quotients a/b and c/d are equal; alternatively

one can use the criterion that the products ad and bc are equal. If the four quantities a, b, c, and d represent line segments, the corresponding fact ought to be that a rectangle of sides a and d has the same area as a rectangle of sides b and c. Geometric proportion is important both in science, as we shall see in Chapter 7, and in art, where it is used in the design of beautiful buildings such as the Parthenon (see Fig. 4.1).

The attempt to transfer the relatively straightforward arithmetic theory of proportion to geometry, combined with a demand for rigorous proof, led to the discovery of incommensurables at an early date. In the words of the historian of mathematics J. J. Gray, this discovery was perhaps "the first good piece of pure mathematics." It marks the decisive point at which a universal sort of intuitive arithmetic and geometry becomes infused with logic and suitable for organization into systematic treatises.

4.4 Early Philosophers

Our story begins around the year 600 B.C.E. with a group of philosophers associated with the Greek commercial cities of the eastern Mediterranean. The mainland of Greece is rather infertile, and the Greeks were consequently engaged in planting colonies to relieve the pressure of an increasing population. Their commercial interests led to a way of life that encouraged interest in the world, a liberation from rigid social customs, and an eagerness to find new ways of doing things. The earliest Greek thinkers of whom we have a record are associated with the Greek colonies of Miletus, Samos, and Chios, near the western coast of what is now Turkey.

These Greek commercial cities came into conflict with the Persians in the early fifth century B.C.E., and the history of the resulting wars was written, along with a large number of fascinating stories and legends, by Herodotus, who was a child at the time of the wars. We shall begin our story with him.

4.4.1 Thales

Herodotus mentions Thales, the first philosopher/mathematician we have to deal with, in several places. Discussing the war between the Medes and the Lydian king Croesus, which had taken place in the previous century, he says that an eclipse of the sun frightened the combatants into making peace. Thales, according to Herodotus, had predicted that an eclipse would occur no later than the year in which it actually occurred. Herodotus goes on to say that Thales had helped Croesus to divert the river Halys so that his army could cross it.

These anecdotes show that Thales had both scientific and practical interests. His prediction of a solar eclipse (which, according to the astronomers, occurred in 585 B.C.E.) seems quite remarkable. Although solar eclipses occur regularly, they are visible only over small portions of the earth, so that their regularity is difficult to discover and verify. Lunar eclipses, however, exhibit the same period

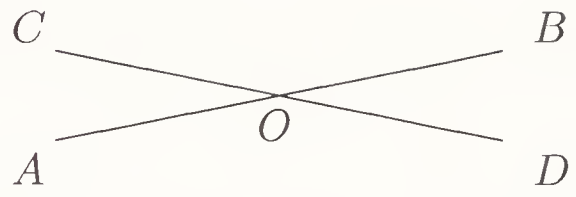


Figure 4.2: Equality of vertical angles.

as solar eclipses and are easier to observe. Eclipses recur in cycles of about 235 lunar months (19 solar years), a period that seems to have been known to many ancient peoples. Among the cuneiform tablets from Mesopotamia there are many that discuss astronomy, and Ptolemy uses Mesopotamian observations in his system of astronomy. Thus Thales could easily have acquired such knowledge in various places.

Thales must have traveled widely and learned what there was to know from the older civilizations of the region. He was also, according to Eudemus, the first real geometer, said to be responsible for the following facts:

1. The vertical angles formed by two intersecting lines are equal.
2. The base angles of an isosceles triangle are equal.
3. A circle is bisected by any line through its center.
4. Two triangles are congruent if they have two angles and the side between them equal.

Historians also associate Thales with the theorem that an angle inscribed in a semicircle is a right angle. According to Diogenes Laertius, a Roman historian named Pamphila, who lived in the time of Nero, credits Thales with being the first to inscribe a right triangle in a circle. To achieve this construction one would obviously have to know that the hypotenuse of the inscribed triangle is a diameter. Diogenes Laertius goes on to say that others attribute this construction to Pythagoras.

It is not clear in what sense Thales "knew" these facts. Had he *proved* them in the modern sense? If so, from what premises? Perhaps he invented a way of showing that they follow from more obvious facts. For example, one can show that a pair of vertical angles are equal by observing that each adds up to a straight angle with one of the other pair of vertical angles formed by the same intersecting lines, as in Fig. 4.2, where one can easily see that $\angle AOC + \angle COB = 180° = \angle DOB + \angle COB$, and hence that $\angle AOC = \angle DOB$. It may be that to Thales these propositions were simply self-evident fundamental principles that have significant practical applications.

In giving Thales credit for the theorem that a circle is bisected by a diameter Proclus also says why the theorem is true. According to Proclus,

> The cause of this bisection is the undeviating course of the straight line through the center; for since it moves through the middle and

> throughout all parts of its identical movement refrains from swerving to either side, it cuts off equal lengths of the circumference on both sides.

This argument is just the kind of vague intuitive argument decried by geometry teachers for centuries, and one is surprised to find a prominent mathematical philosopher offering it. Proclus, however, presents it only as an intuitive argument and follows it with a more rigorous geometric argument. This argument is still unfortunately not entirely clear, but the idea is apparently to reflect the portion of the circle on one side of the diameter and then argue that the reflected points must lie on the other portion of the circle since all of the radii are equal. Whether Proclus' argument is the one Thales is supposed to have given is also not clear.

Plutarch reports that Thales traveled to Egypt and was able to calculate the height of the Great Pyramid by using the proportion between his own height and shadow and measuring the length of the shadow of the pyramid. The mathematics involved in doing this would not have been beyond the Egyptians themselves, so that it is possible that Thales learned some mathematics in Egypt. Diogenes Laertius says that Thales calculated the height of the pyramid by waiting until his shadow was exactly as long as he was tall, then measuring the length of the shadow of the Great Pyramid. There are practical difficulties in executing this plan, connected with the fact that one could not get into the Pyramid to measure the distance from the center to the tip of the shadow directly. One might use the Pythagorean theorem to measure the distance from the center of the pyramid to the point where its outer wall intersects the line to the tip of the shadow. It is certainly possible that Thales knew the Pythagorean theorem. There are ways of computing this distance from proportion, however, that do not involve the Pythagorean theorem.

4.4.2 Pythagoras

About half a century later than Thales another early mathematician, named Pythagoras, was born on the island of Samos. No books of Pythagoras survive, but many later writers mention him, including Aristotle. Diogenes Laertius, in his work *Lives of Eminent Philosophers*, devotes a full chapter to the life of Pythagoras, including several of his alleged previous incarnations and two contradictory stories of his death. Diogenes Laertius mentions several books that Pythagoras is said to have written and tells us the names of his wife, daughter, and son.

If the stories about Pythagoras can be believed, he, like Thales, traveled widely, to Egypt and Mesopotamia. He gathered about him a large school of followers, who observed a mystical discipline and devoted themselves to contemplation. This group of people gave us some words that we use nearly every day, words that frame the world in a particular way for us, of which we are seldom consciously aware. Among these words is the word *theory*.

Diogenes Laertius quotes a certain Hellenistic philosopher named Alexander who claimed that the Pythagoreans generated the world from *monads* (units). By

adding a single monad to itself, they generated the natural numbers. By allowing the monad to move, they generated a line, then by further motion the line generated plane figures (polygons), the plane figures then moved to generate three-dimensional figures (polyhedra). From the regular polyhedra they generated the four elements of earth, air, fire, and water.

To the Greeks 1 was not a number, but it did generate numbers by the process of successive addition. From numbers arose lines. This idea seems very natural to us, since we have been taught analytic geometry. In Greek times, however, it led to certain difficulties, to be discussed below. Using the first geometric dimension (a line) to generate two- and three-dimensional space, and then deriving matter from physical space, the Pythagoreans thus generated the whole universe out of their monads. Moreover, this cosmology was mathematical in nature, since it was based on arithmetic and geometry. The way in which matter was believed to arise from geometry was particularly interesting, involving the five regular solid figures. According to the writer Aëtius, who lived around 100 C.E., Pythagoras said that the earth arose from the cube, fire from the pyramid (tetrahedron), air from the octahedron, water from the icosahedron, and the sphere of the universe from the dodecahedron. Proclus also credits Pythagoras with the discovery of the five regular solids, though other writers claim that these solids were first discussed by Theatetus, a contemporary of Plato.

The best known of the followers of Pythagoras is Philolaus, who lived in the fifth century B.C.E. Although he first lived with the Pythagoreans in Croton in southern Italy, he fled to Tarentum to escape the wrath of the citizens of Croton against the Pythagoreans. (The Pythagoreans were regarded as a cult that ensnared young people.) He became a wandering philosopher, propagating Pythagoreanism wherever he went. He is the author of a book *On Nature* (now lost) that is the basis for a good deal of what other ancient writers reported about the Pythagorean philosophy. This book presented a cosmology in which the sun, moon, Mercury, Venus, Mars, Jupiter, Saturn, the earth, and "counterearth" move around a central fire.

Philolaus' book is the first astronomical theory in recorded history, and it is remarkable that it is not geocentric. In Book II of his work *On the Heavens* Aristotle says that the Pythagoreans placed the most valuable substance—fire—at the most important place in the universe, which they called the "guardhouse of Zeus." He says sarcastically that there is no need to be so disturbed about the universe or to post a guard at its center. Later followers of Pythagoras claimed that Aristotle had misunderstood the doctrine. Like most of Pythagoreanism, Philolaus' book contained both scientific and mystical elements. For example there were several reasons for introducing the counterearth and the central fire. First, on mystical grounds it was believed that the number of heavenly bodies must be 10, the basis of arithmetic. Second, it had been observed that eclipses of the moon could occur while the sun was above the horizon. This would seem to be impossible if the eclipse is merely the shadow of the earth falling on the moon. With the Pythagorean system the eclipse could be understood as the shadow of the counterearth on the moon. Third, the counterearth explains the fact that eclipses of the moon are more commonly observed than eclipses of the sun. The fact that the

central fire could not be observed was explained by saying that the earth always turns the same side to it, just as the moon always turns the same side to the earth.

Pythagorean Arithmetic

The Pythagoreans held a more abstract view of numbers than any we have encountered so far. Our knowledge of Pythagorean number theory is based on several sources, including Books VII–IX of Euclid's *Elements* and a treatise on arithmetic by the neo-Pythagorean Nicomachus of Gerasa, who lived about 100 C.E. On the basis of these documents we can make the following observations about the Pythagorean arithmetic.

To begin with, the Pythagoreans made the elementary distinction between odd and even numbers. Having made this distinction, they proceeded to refine it, distinguishing between even numbers divisible by 4 (evenly even) and those that are not (even$\times$odd). They went on to classify odd numbers in a similar way, coming thereby to the concept of prime and composite numbers, and what we now call pairs of relatively prime numbers. The notion of a relational property was difficult for Greek philosophers, and Nicomachus expresses the notion of relatively prime numbers somewhat confusingly, referring to three species of odd numbers: the prime and incomposite, the secondary and composite, and "the variety which, in itself is secondary and composite, but relatively is prime and incomposite." This way of writing seems to imply that there are three kinds of integers, prime and incomposite, secondary and composite, and a third kind midway between the other two. It also seems to imply that one can look at an individual integer and classify it into exactly one of these three classes.

Like Nicomachus, Euclid devotes his three books on number theory to the mysteries of divisibility theory, spending most of the time on proportions among integers, and on prime and composite numbers. Only at the end of Book IX does he prove a theorem of a different sort, giving a method of searching for perfect numbers (numbers, such as 6 and 28, that are equal to the sum of their proper divisors).

The study of prime numbers has been one of continuing importance in mathematics right down to the present day, and the Pythagoreans must be given credit for beginning this study. Another aspect of their study of numbers has been far less fruitful, however, namely the study of figurate numbers. The Pythagoreans distinguished triangular numbers (1, 3, 6, 10, 15, 21,...) square numbers (1, 4, 9, 16, 25, 36,...), pentagonal numbers, and the like and proved abstract theorems, such as the theorem that the sum of two successive triangular numbers is a square number (for example, $15 + 21 = 36$). It is difficult to see how anyone could have predicted which of these two seedlings nourished by the Pythagoreans—the study of divisibility theory, and the study of figurate numbers—would grow into a mighty tree. The two were planted together, and no doubt both seemed important to their founders. Yet the study of prime numbers has led to enormous amounts of profound mathematics, while, except for squares, the study of figurate numbers has never been more than a curiosity.

In connection with the theory of divisibility and proportion for integers, it seems likely that the Pythagoreans would have known how to find the greatest common divisor (factor) of two numbers. There is evidence for this assertion, in that such a procedure is described in Proposition 2 of Book VII of Euclid's *Elements*, and the three books of the *Elements* concerned with number theory are believed to be Pythagorean. This procedure, now known as the *Euclidean algorithm*, deserves a detailed explanation.

For definiteness, we shall imagine that the two quantities whose greatest common *measure* is to be found are two lengths, say a and b. Suppose that a is longer than b. (If the two are equal, then clearly their common value is also their greatest common divisor.) Here is an elaboration of the procedure described in a few words by Euclid:

1. Replace the pair (a, b) by the pair $(a - b, b)$. Then the greatest common measure of $a - b$ and b is also the greatest common measure of a and b. For if a length c divides both $a - b$ and b (say, $a - b = rc$ and $b = sc$ for integers r and s), then c also measures a [since $a = a - b + b = rc + sc = (r + s)c$]. Thus any common measure of $a - b$ and b is also a common measure of a and b, and clearly the argument works in reverse. That is, (a, b) and $(a - b, b)$ have the *same* common measures. In particular if $a - b = b$, then this common value (b) is the greatest common measure. If $a - b \neq b$, we start over with the new pair. We shall denote the new pair by (a_1, b_1) and assume that $a_1 > b_1$.

2. We can now repeat the argument with the new pair, in which one of the elements is shorter than the larger element of the original pair by an amount equal to the shorter element of the original pair. It is clear that, if the argument is repeated, leading to the sequence $(a_1, b_1), \ldots, (a_n, b_n), \ldots$ either an equal pair eventually occurs or else the larger element a_n will eventually become less than half of the original larger element a. For one can see easily that the *shorter* element after one subtraction is less than half of the original larger element, and repeated subtraction of this shorter element from the other will eventually leave a remainder that is even shorter. Each pair produced will have the same common measures as its predecessor, and hence the same common measures as the original pair (a, b).

3. Since the greatest common measure of a and b divides both a_n and b_n, it follows that a_n cannot be smaller than this common divisor. We have noted, however, that either $a_n = b_n$ for some n or $a_n < \frac{1}{2}a$ for some n. Hence if the process does not terminate, a_n eventually becomes arbitrarily small. It follows that *if there is a common measure of a and b, the process must terminate by producing an equal pair in a finite number of steps.*

An example will make all this clear. Let us find the greatest common measure (divisor) of 24 and 488. Clearly a common measure does exist, namely the integer 1. Starting with the pair $(488, 24)$, we subtract the smaller from the larger, getting the new pair $(464, 24)$. Since this pair is not equal, we repeat the process. The

repeated subtraction described in the pure algorithm can be shortened by division: 488 divided by 24 gives a quotient of 20 and a remainder of 8. Thus we would apply the algorithm 20 times, getting the successive pairs $(488, 24)$, $(464, 24)$, $(440, 24)$,..., $(32, 24)$, $(8, 24)$. At this point we start with the pair $(24, 8)$ and get successively $(16, 8)$, $(8, 8)$. Having finally reached an equal pair, we conclude that the greatest common divisor of 488 and 24 is 8.

Proportion and Measurement

The greatest common divisor plays an important role in the theory of proportion. If two integers are to be used to express the ratio of one object to another, it is advisable to divide out their greatest common measure. Thus it would be foolish to say that two lines were in the ratio of 36 to 54, when one could divide both of these numbers by 18 and say that the ratio is 2 to 3. Now, in order to *find* the ratio in the first place for objects such as line segments, the simplest thing is to find the greatest common measure and see how many times it is contained in each. That is what we do when we use any calibrated measuring instrument, a ruler, for example. We take the smallest calibration (say, one millimeter), which is regarded as a divisor of the length we are measuring. This assumption is acceptable since we are seeking only approximation, although it is not strictly speaking true. Since the smallest calibration certainly divides the unit length (say, one meter), it provides us with a common measure of the object to be measured and the unit of measurement. Thus if we say that a line is seven millimeters long, we are really expressing its ratio to the standard meter as $7 : 1000$. In this way any length can be compared with the standard length, and so any two lengths can be compared with each other. It follows that any two *measured* quantities of the same kind (lengths, areas, mass, etc.) will have a common measure, simply because of what we take measurement to mean.

Pythagorean Geometry

Before discussing the clash between arithmetic and geometry in Pythagoreanism, we need to reconstruct their geometry as well as we can, since, like arithmetic, it played an important role in their cosmology. From Proclus and other later authors we have a glimpse of a fairly sophisticated Pythagorean geometry, intimately intertwined with their characteristic mysticism. For example, Proclus reports that the Pythagoreans regarded the right angle as ethically and aesthetically superior to acute and obtuse angles, since it was "upright, uninclined to evil, and inflexible." Right angles, he says, were referred to the "immaculate essences" while the obtuse and acute angles were assigned to divinities responsible for changes in things. Thus the Pythagoreans had a bias in favor of the eternal over the changeable, and they placed the right angle among the eternal things since, unlike acute and obtuse angles, it cannot change without losing its character.

Proclus mentions two specific parts of geometry as being Pythagorean in origin. One is the theorem that the sum of the angles of a triangle is two right angles. The other is a portion of Euclid that is not generally taught any more, the topic

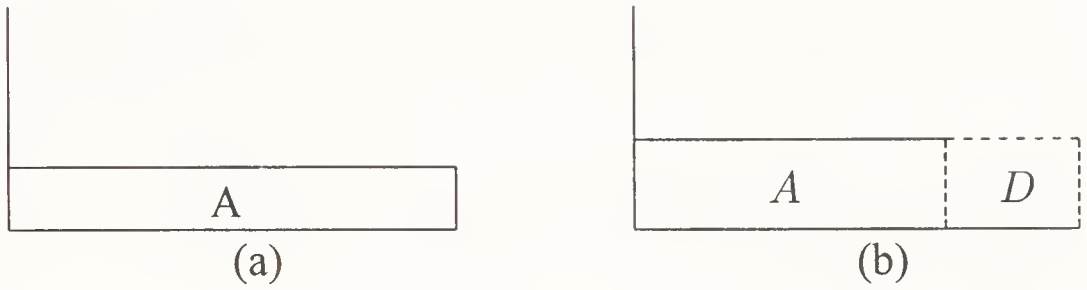


Figure 4.3: Application (a); application with defect (b).

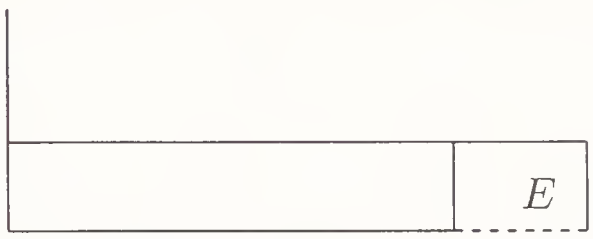


Figure 4.4: Application with excess.

of application of areas. There are three such problems: (1) given a straight line segment of length a and a second line passing through its endpoint, to construct a parallelogram having the side of length a as base, the other line as a side, and a prescribed area (*application*); (2) given the prescribed area, to construct a parallelogram equal to that area on part of the base a and having the second line as a side in such a way that the parallelogram needed to fill up a parallelogram on the entire base (called the *defect*) will have a prescribed shape (*application with defect*); and (3) given the prescribed areas, to construct a parallelogram on a side containing the base a having the second line as a side and such that portion of the parallelogram extending beyond the base a (the *excess*) will have a prescribed shape (*application with excess*). These constructions are shown in Figs. 4.3 and 4.4 for the case when the angle in which the area (A) is to be inscribed is a right angle and the ratio of the sides of the defect (D) or excess (E) is $2:3$. (The area A is not labeled in Fig. 4.4 since the excess E forms a part of it.) Proclus cites Eudemus in asserting that the solution of these problems was an ancient discovery of the Pythagoreans.

The first of these problems amounts to finding the second side of the parallelogram, given its area, one of its sides, and the angle between the sides. In the important case when the two lines are perpendicular and the excess or defect is a square, the second and third problems amount to finding two unknown quantities (lengths) given their sum and product (application with deficiency) or given their difference and product (application with excess). As we saw in the preceding chapter, in modern terms these problems amount to quadratic equations, and numerical procedures for solving them occur constantly in the cuneiform tablets. Undoubtedly this Babylonian mathematics was known to the early Greeks. The question that we cannot answer definitely is: How did these problems come to appear in geometric form in Euclid's treatise? Is there a reason why Euclid did

not write about the numerical version of the problems? Is that reason connected with the fact that the solution of these problems involves extracting square roots, which usually were not *numbers* in the sense understood by the Greeks? Or do these constructions have some purely geometric origin unconnected with the earlier numerical problems? The geometric problems are stated for parallelograms in such a way that no simple numerical interpretation of them exists.

We cannot give definite answers to these questions, but it may well be that certain difficulties we are about to discuss convinced the Greeks that not every ratio could be regarded as a number, so that the theory of proportion in geometry had to be constructed in a more complicated manner than the theory of proportion in arithmetic.

4.4.3 Zeno of Elea

Although we have some idea of the geometric results proved by the Pythagoreans, our knowledge of their interpretation of these results is murkier. How did they conceive of geometric entities such as points, lines, planes, and solids? Were these objects physically real or merely ideas? What properties did they have? Some light is shed on this question by the philosophical critics of Pythagoreanism, and we shall now discuss the ideas of the most prominent of these critics.

The Pythagoreans began with a mystical faith in the power of numbers to explain the universe. At the same time, by developing geometry, they were conjuring up the nemesis of arithmetic. It turned out that the Pythagorean view of geometry and number contained paradoxes within itself, which were starkly pointed out by the philosopher Zeno of Elea. Zeno died around 430 B.C.E., and, as usual in such cases, we do not have any of his works to rely on, only expositions of them by other writers. Aristotle, in particular, says that Zeno gave four puzzles about motion, which he called the Dichotomy (division), the Achilles, the Arrow, and the Stadium. Here is a summary of these arguments in modern language, based on Book VI of Aristotle's *Physics*.

1. *The Dichotomy*. Motion is impossible because before an object can arrive at its destination it must first arrive at the middle of its route. But before it can arrive at the middle, it must travel one-fourth of the way, etc. Thus we see that the object must do infinitely many things in a finite time in order to move.

2. *The Achilles*. (So-called because the legendary warrior Achilles chased the Trojan hero Hector around the walls of Troy, overtook him, and killed him.) If given a head start, the slower runner will never be overtaken by the faster runner. For, before the two runners can be at the same point at the same instant, the faster runner must first reach the point from which the slower runner started. But at that instant the slower runner will have reached another point ahead of the faster. Hence the race can be thought of as beginning again at that instant, with the slower runner still having a head start. Clearly the race will "begin again" in this sense infinitely many times with the slower

runner always having a head start. Thus, as in the dichotomy, infinitely many things must be accomplished in a finite time in order for the faster runner to overtake the slower.

3. *The Arrow.* An arrow in flight is at rest at each instant of time. That is, it does not move from one place to another during that instant. But then it follows that it cannot traverse any positive distance because successive additions of zero will never result in anything but zero.

4. *The Stadium.* (In athletic stadiums in Greece the athletes ran from the goal to a halfway post and then back. This paradox seems to have been inspired by imagining two lines of athletes running in opposite directions and meeting each other.) Consider two parallel line segments of equal length moving toward each other with equal speeds. The speed of each line is measured by the number of points of space it passes by in a given time. But each point of one line passes *twice* that many points of the other line in the same time as the two lines move past each other. Hence the velocity of the line must equal its double, which is absurd.

Given the modern outlook on the world, it is difficult to appreciate the problem that these paradoxes created for the Pythagoreans. There are two reasons for our difficulty in understanding the paradoxes. First, mathematicians have worked out ways of avoiding these paradoxes, and our view of the world is now such that the paradoxes cannot easily be stated. Second, we tend to regard such puzzles as recreation, not to be taken seriously. In ancient Greece those who chose to spend their time in schools (a leisure class) regarded thinking about such puzzles as important work.[1] In our modern democracies, in contrast, large numbers of people attend universities and wonder why pedantic professors waste their time with such pointless word-spinning. Our purpose in presenting these paradoxes is to see what issues they raise for the development of mathematics. In order to do that, we must clear our minds of modern concepts of motion inherited from Newtonian mechanics and the popularized theory of relativity.

The stadium paradox, for instance, seems transparent to a twentieth-century mind. The velocity of each row of bodies relative to the other is double its velocity relative to the ground, and there is no mystery here. It seemed otherwise to the Greeks, for whom the velocity of a moving point was not relative to another object, but rather was a measure of (proportional to) the quantity of *space* it passed in a given time. From this point of view, it does indeed seem a contradiction that the velocity can be two things. As noted above in connection with relatively prime numbers, Greek philosophers had difficulty with relational properties. We have seen Nicomachus' awkward description of relatively prime integers as being "in between" prime and composite integers—composite in the absolute sense but prime

[1]One must be careful in generalizing about the ancient Greeks, however. Our records are biased toward the views of the scholars who wrote the records. It is quite possible that most of Plato's students were young aristocrats who went to the Academy because their parents wanted them to be cultured. Like the sons of the British nobility who were educated at Eton, many of them probably preferred hunting and carousing to study.

with respect to each other. There is a further difficulty in the present situation in that those who translate the original Greek into modern languages cannot be sure of the meaning of the original language. It seems likely that, because the problem is a paradox, its statement was confused by those who attempted to report it, until now it is impossible to know how the paradox was originally stated. It is possible to make a real paradox out of this thought experiment by assuming that time consists of indivisible instants. If the bodies are such that each body in motion moves past a body at rest in exactly one instant of time, then two bodies in oppositely directed motion move past each other in half an instant, which is an impossibility, since instants cannot be divided in half. However, this paradox seems rather far from the language in which Aristotle reports the argument.

Similarly in the arrow paradox, Newtonian physicists would agree that in a given instant the arrow does not move. They would not agree, however, that it is "at rest" at that instant, that is, its velocity at that instant is zero. As for the dichotomy and Achilles paradoxes, any modern mathematician would point out that, although it is true that the traveling object must begin shorter and shorter journeys before it can get anywhere and the race between the fast and slow runner must begin infinitely many times before the fast runner overtakes the slow one, each new beginning requires less time than the one before, and the total sum of the times required is finite. Thus the modern world disposes of these paradoxes.

The situation was different for the Pythagoreans, however. They had built their system on lines "made up" of points, and now Zeno was showing them that space cannot be "made up" of points in the same way that a building can be made of bricks. For assuredly the number of points in a line segment cannot be finite. If it were, the line would not be infinitely divisible as the dichotomy and Achilles paradoxes showed it must be; moreover the stadium paradox would show that the number of points in a line segment equals its double. There must therefore be an infinity of points in a line. But then each of these points must take up no space; for if each point occupied some space, an infinite number of them would occupy an infinite amount of space. But if points occupy no space, how can the arrow, whose tip is at a single point at each instant of time move through a *positive* quantity of space? From these difficulties the Pythagoreans apparently found no escape. A continuum whose elements are points seemed to be needed for geometry, yet it could not be thought of as being made up of points in the way that discrete collections are made up of individuals.

4.4.4 The Problem of Incommensurables

The difficulties pointed out by Zeno lay in the background of Pythagorean geometry. That is, they affected the interpretation of geometric theorems, but not the logical validity of their derivation from first principles. There was, however, a difficulty within the geometry that the Pythagoreans were practicing, only waiting for its chance to spring forth as soon as someone examined the matter closely enough. We now turn to examine this problem as it must have arisen in the fifth century B.C.E. Of course, we do not know how the discovery was made, and what

follows is merely one possible way it may have happened. (We shall look at other possibilities in the next chapter.) As background for the discussion we assume the elements of number theory as reflected by Nicomachus in his *Arithmetic*, and we assume a certain amount of knowledge of elementary geometry on the part of the Pythagoreans. The quotations from Proclus show that this assumption is a safe one. Indeed it is safe to say that the Pythagoreans knew most of the standard theorems on congruence of figures and the relations between chords, arcs, and tangents on a circle. The conflict we are about to discuss probably originated in the theory of similar figures.

To exhibit the conflict, we return to the problem of proportion in geometry and attempt to apply the Euclidean algorithm described above to find certain proportions. Because any two *measured* quantities of the same kind are commensurable—they have as a common measure the smallest calibration on the measuring instrument—the problem we are about to discuss is not a practical problem. It arises rather from reflecting very deeply on the idealizations that make up the essence of mathematics. Thus we find that one of the first fruits of careful thought is confusion—the attempt to be more than ordinarily clear about the process of comparing lengths led to the first dilemma in the history of mathematics. We now describe that dilemma. The Euclidean algorithm works for positive integers, because positive integers are made up of "atoms" (a Greek word meaning *indivisible*). That is, there is no positive integer smaller than 1, which is a common divisor of all positive integers. It seems to common sense that the same can be said about lines. Lines are made up of points, which seem to play the same atomic role in geometry that the number 1 plays in arithmetic. Once we have a common measure of two objects, we can talk about their ratio (the ratio of the number of times each is divisible by the common measure). *Therefore*, it seems, the Euclidean algorithm should enable us to find the ratios among the parts of simple geometric figures, in particular between the sides and diagonals of squares, pentagons, and hexagons. These figures were fundamental to the Pythagoreans. As we saw above, they identified the physical elements of the world with the regular solids: the tetrahedron with fire, the octahedron with air, the icosahedron with water, and the cube with earth. The first three of these regular solids have faces that are triangles, and the cube has square faces. The dodecahedron, which has pentagonal faces, was identified with the universe itself. The pentagon thereby became a mystical figure for the Pythagoreans. The ratio between its sides and diagonals must have been a subject of intense interest. Indeed, that ratio is said to occur as the ratio of sides of rectangles in many of the classical buildings, such as the Parthenon (see Fig. 4.1), though others find the side and diagonal of a square to be the underlying "theme" for the same structure. Let us apply the Euclidean algorithm to the side and diagonal of a regular pentagon to see if we can discover their greatest common measure. Consider the pentagon $ABCDE$ in Figure 4.5 with diagonals AC and AD drawn. It is easy to see that the two diagonals trisect the angle at A, since the three angles formed by the sides and diagonals meeting at A are inscribed in equal arcs of the circle. Since the base angles of the triangle ACD are inscribed in arcs twice as long, these angles are each equal to twice angle CAD. Knowing, as the Pythagoreans did, that the sum of the angles of a

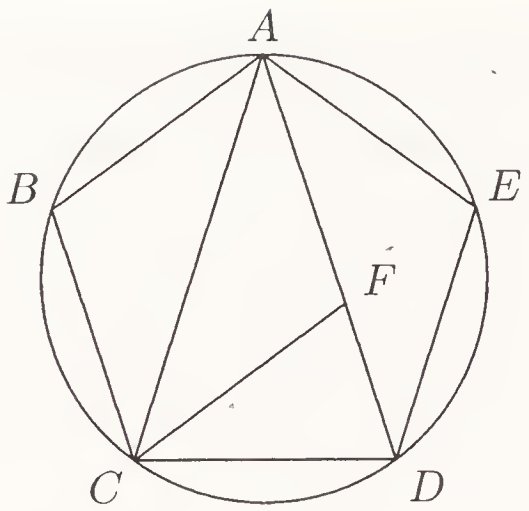


Figure 4.5: Diagonal and side of a regular pentagon.

triangle is 180°, we conclude that angles ACD and ADC are each 72° and angle CAD is 36°. Now if the bisector of the angle ACD meets AD at the point F, the triangle CDF will have an angle of 36° at C. Angle CFD will therefore be 72°, and hence the triangle CDF will have the same angles as the original triangle ACD. It follows that DF is the side of a (smaller) pentagon whose diagonals all equal CD (and CF). Now angle ACF is also 36°, and so triangle ACF is isosceles ($AF = CF$). Hence $AD - CD = AD - CF = AD - AF = DF$.

Now to apply the Euclidean algorithm, we could represent the diagonal and side of the pentagon $ABCDE$ by the pair (AD, CD). The algorithm would cause us to replace this pair by $(AD - CD, CD)$. We have just seen that this pair is the same as the pair (DF, CD). But we have also just seen that *this new pair also forms the side and diagonal of a pentagon!* Thus, no matter how many times we apply the procedure of the Euclidean algorithm, the result will always be a pair consisting of the side and diagonal of a pentagon. Therefore in this case the Euclidean algorithm will *never* produce an equal pair of lines. We know, however, that it *must* produce an equal pair if a common measure exists. We conclude that *no common measure can exist for the side and diagonal of a pentagon.* These two lengths are said to be *incommensurable.* A similar argument can be applied to the side and diagonal of a square, only in that case the algorithm must be applied twice before the new pair is the side and diagonal of a smaller square.

Whether by this argument or some other, the Greeks discovered the existence of incommensurable pairs of line segments before the time of Plato. For Pythagorean metaphysics this discovery was disturbing: number, it seems, is *not* adequate to explain all of nature. A legend arose that the Pythagoreans attempted to keep secret the discovery of this paradox. However, scholars believe that the discovery of incommensurables came near the end of the fifth century B.C.E., when the original Pythagorean group was already defunct.

The existence of incommensurables is one of the two horns of the dilemma that led the Greek philosophers to speculate on the metaphysical underpinnings of mathematics. The other horn is the belief that any two line segments *should* have a common measure, since both are made up of identical points (atoms of a sort). This is the first clash between the discrete and the continuous in mathematics, and

the war between them—between arithmetic and geometry—has raged intermittently ever since. The best efforts of mathematical "diplomats" have never resulted in anything better than a temporary and uneasy truce.

The existence of incommensurables throws doubt on certain oversimplified proofs of geometric proportion. When two lines or areas are commensurable, one can describe their ratio as, say, $5 : 7$, meaning that there is a common measure such that the first object is five times this measure and the second is seven times it. A proportion such as $a : b :: c : d$ then is the statement that both of the ratios $a : b$ and $c : d$ are represented by the same pair of numbers.

Now this theory of proportion is extremely important in geometry if we are to have such theorems as Proposition 1 of Book VI of Euclid's *Elements*, which says that the areas of two triangles or two parallelograms having the same height are proportional to their bases, or the theorem (Book XII, Proposition 2) that the areas of two circles are proportional to the squares on their diameters. Even the simplest constructions, such as the construction of a square equal in area to a given rectangle or the three application problems mentioned above, may require the concept of proportionality of lines. Because of the extreme importance of the theory of proportion for geometry, the discovery of incommensurables made it imperative to give a definition of proportion without relying on a common measure to define a ratio.

The problem was a deeply philosophical one. What is ultimately desired is a theory of *proportion*, but for that purpose it was necessary to have a clear idea of a *ratio*. As long as two objects are commensurable, their ratio can be thought of as the ratio of two positive integers (what we would call a rational number). Without commensurability, it would seem that we could define equality of two ratios, at least for lines, by saying $a : b :: c : d$ if the rectangle on a and d has the same area as the rectangle on b and c, as suggested above. However, this approach would define equality of ratios, yet leave the ratios themselves undefined. As we shall see, the ultimate solution chosen by the Greek mathematicians did, in effect, the same thing, although Euclid tried to blur this fact with a rather vague definition of a ratio. The solution of this problem, due to Eudoxus, will be discussed in the next chapter.

At this point we leave the Pythagoreans. They made some notable advances in the theory of numbers and elementary geometry, leading to fundamental and far-reaching mathematical theories. In trying to be logical and clear they succeeded in uncovering new and unsuspected difficulties in their subject, difficulties that later generations of mathematicians would have to resolve. The paradoxes of Zeno showed that there must be a difficulty somewhere within the Pythagorean philosophy. However, they did not contradict any single Pythagorean doctrine, only the total collection of doctrines. They could therefore be ignored by those who merely wished to prove theorems. The problem of incommensurables, however, contradicted assumptions explicitly made in proofs of theorems on proportion. Proofs based on the use of a common measure were, as a result of the discovery of these paradoxes, overtly fallacious; something had to be done to repair them. Once those repairs were made, the way was clear for the construction of one of the epoch-making advances in human knowledge, the first systematic, deductive

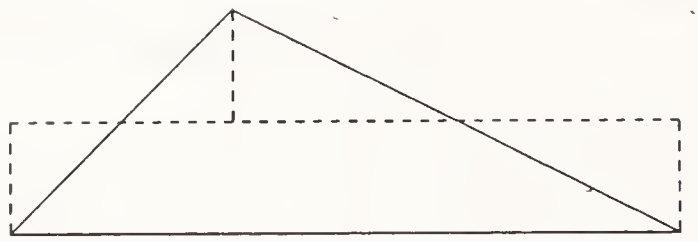


Figure 4.6: Transforming a triangle into a rectangle.

theory, as expounded in the immortal textbook of Euclid. We shall take up this part of the story in the next chapter. For the remainder of the present chapter we shall explore other aspects of the early development of geometry.

4.5 Other Greek Geometry

Because the success of Euclid's treatise on geometry made earlier treatises obsolete, we have limited knowledge of what geometry consisted of in the fifth and fourth centuries. Nevertheless it is plausible that the proportions and congruences to be found among circles and rectilinear plane figures were well known. Since the proofs of such theorems depend on a theory of proportion, we can picture the situation somewhat as follows. The elements of plane geometry involving the transformation of areas (to be discussed below) and the construction of certain lines, such as the tangent to a circle, were well established by the late fifth century B.C.E. At this point geometers were faced with two sets of problems: first, to extend the transformation of rectilinear areas to areas bounded by curves and to three dimensions (transformation of volumes); second, to reinstate the jeopardized theory of proportion by constructing a theory that would apply to both commensurable and incommensurable magnitudes. Only the first of these concerns us at present.

It is an elementary construction to transform a triangle into a rectangle, that is, to partition the triangle into a trapezoid and two smaller triangles that can then be reassembled into a rectangle, as in Fig. 4.6.

The elementary construction of the mean proportional between two lengths, illustrated in Fig. 4.7, then makes it possible to construct a square equal in area to a rectangle. Hence one can construct a square equal to any triangle. Then, using the Pythagorean theorem, which makes it possible to construct a square equal to the sum of two other squares, one can easily see how to construct a square equal to any figure that can be triangulated (partitioned into a finite number of triangles). Since any polygon can be triangulated, we see that *it is possible to construct a square equal in area to any polygon.* All this theory, known as *quadrature* (squaring), must have been known to the Pythagoreans. Now any mathematician, surveying this scene, would immediately attempt to extend these results further. In particular two problems naturally arise: (1) construct a square equal to a given circle (quadrature of the circle) and (2) construct a cube equal in volume to any given closed polyhedron. A third classical problem arises in carrying out plane

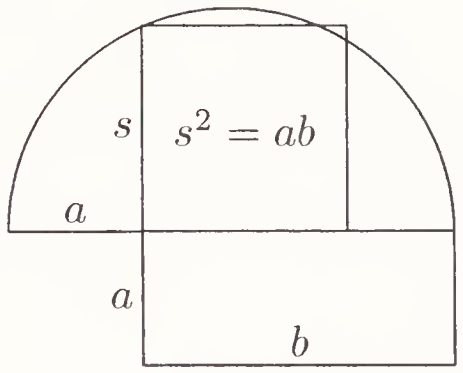


Figure 4.7: Transforming a rectangle into a square ($s^2 = ab$)

constructions. It is frequently necessary to divide a line into a finite number of equal parts, and the construction for doing so is well known. Dividing angles (arcs), however, is not so easy. The construction for half of an angle or arc is well known and simple, but the construction for one-third of an angle is more elusive. Trisecting the angle was the third classical problem of geometry.

These problems were always understood in the sense of finding chords or arcs. Thus constructing a square or cube meant constructing the length of its side. Although later writers have posed these problems so as to limit the allowable methods, the Greeks may not have imposed such limitations. One would naturally try to solve these problems using only straight lines and circles, since these were the simplest devices and had been successfully used to solve many other problems. If no success was achieved after a certain amount of effort, it was natural to look for other devices or to modify the goal slightly, and that is in fact what happened.

Several of the sources mentioned above, including Proclus and Vitruvius, report that the problem of squaring the circle was worked on by the philosopher Anaxagoras around 440 B.C.E., when he was imprisoned in Athens by the enemies of Pericles. Vitruvius also relates that Anaxagoras worked together with the philosopher Democritus (one of the two originators of the atomic theory of matter) on geometric problems in the design of scenery for the theater. Having no details of the contents of this work, however, we mention it only to show the sort of activity that was taking place in the fifth century B.C.E.

4.5.1 Hippocrates of Chios

We know from the accounts of later writers that a number of authors worked on the classical problems. One of these was Hippocrates of Chios (not to be confused with the famous physician Hippocrates of Cos), who lived in the second half of the fifth century B.C.E. and is thought to have died in Athens. His work was not preserved, but fortunately it was described in detail in Eudemus' history of mathematics, and Simplicius considered this passage important enough to quote it at length in his commentary on Aristotle's *Physics*. As often happens with an intractable mathematical problem, mathematicians tried an indirect or partial approach. In the case of squaring the circle, Hippocrates was successful in squaring

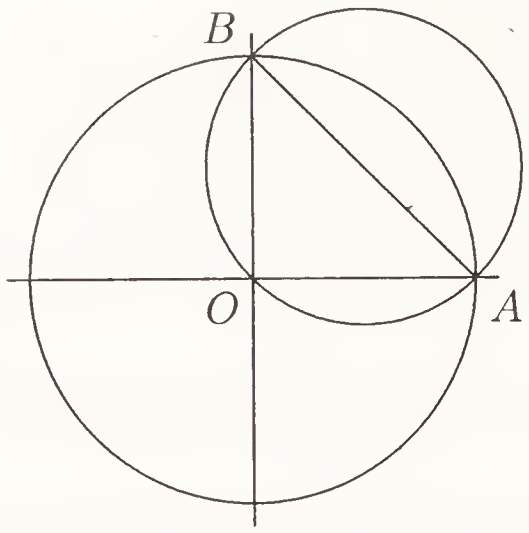


Figure 4.8: Quadrature of a lune.

certain regions between two overlapping circles. Such regions resemble crescent moons and are therefore called *lunes*. As a sample of Hippocrates' work, found in Simplicius and attributed to Eudemus, consider the following simple quadrature of a lune.

Figure 4.8 shows a small circle whose diameter AB is the chord on a $90°$ arc of a larger circle centered at O. The diameter of the smaller circle is therefore $\frac{1}{\sqrt{2}}$ times the diameter of the larger circle. Since the areas of circles are proportional to the squares on their diameters, the area of the smaller circle is half the area of the larger circle. In particular the semicircle whose diameter is AB equals the quarter-circle whose radii are OA and OB. If we subtract the segment of the larger circle that is common to both of these regions, we find that the remainders are equal. That is to say, the triangle OAB is equal in area to the lune that is inside the smaller circle and outside the larger one.

Thus an area bounded by circular arcs can be proved equal to a rectilinear area. This result, although not a solution of the problem of squaring the circle, is at least progress in that direction. Hippocrates worked out the quadrature of many simple lunes, but always it appeared that the quadrature was possible only because of a subtraction. Some mysterious complication in the circle was canceled out when part of one circle was subtracted from another, so that the difference of two circles could be squared, but not a single circle.

A similar piecemeal approach characterized the extension of the theory of measurement to three dimensions. Remembering our conjecture that the Pythagorean theorem may have been discovered in solving the problem of doubling a square, we can see that it would be natural to begin by trying to construct a cube equal to the union of two identical cubes. This is the classical problem of duplicating the cube, and Hippocrates of Chios is said to have worked on this problem also. Proclus says that Hippocrates was the first to reduce this problem to the problem of finding two numbers between two given numbers so that the four would be in continued proportion. This reduction caused the problem to be known as the problem of two mean proportionals.

Both Theon of Smyrna and Eutocius tell colorful stories about the origin of this problem. According to these authors, the citizens of Delos learned from an

oracle that a plague afflicting the city would be lifted only when they doubled the size of an existing altar. Theon goes on to say that the Delians consulted Plato about ways of accomplishing this construction, but Plato told them that the altar was a red herring—the gods really just wanted the Delians to pay more attention to the study of geometry. Eutocius quotes a letter allegedly from the mathematician Eratosthenes, who will be discussed in a later chapter, relating a similar story about doubling the size of a tomb. The letter goes on to say that Hippocrates' reduction of the problem to the construction of two mean proportionals was of no value: "the puzzle was by him turned into no less a puzzle." Modern scholars claim, however, that this letter was a forgery.

4.6 Problems and Questions

4.6.1 Problems in Greek Geometry

Exercise 4.1 The repeated subtraction in the Euclidean algorithm is usually shortened by simply dividing the larger quantity by the smaller and taking the divisor and remainder as the new pair. Illustrate this procedure by finding the greatest common divisor of 189,189 and 13,923.

Exercise 4.2 What is the ratio of the diagonal of a regular pentagon to its side? Find this number (d/s), given that $d : s :: s : (d - s)$. It is known as the *Golden Section*. What is its approximate numerical value, expressed as a finite decimal number?

Exercise 4.3 Prove the statement in the text that the problem of constructing a rectangle of prescribed area on part of a given base a in such a way that the defect is a square is precisely the problem of finding two numbers given their sum and product (the two numbers are the lengths of the sides of the rectangle). Similarly prove that the problem of application with square excess is precisely the problem of finding two numbers (lengths) given their difference and product.

Exercise 4.4 Show that the problem of application with square excess has a solution for any given area and any given base. What restrictions are needed on the area and base in order for the problem of application with square defect to have a solution?

Exercise 4.5 Use an argument similar to that in the text to show that the side and diagonal of a square are incommensurable. That is, show that the Euclidean algorithm, when applied to the diagonal and side of a square requires only two steps to produce the side and diagonal of a smaller square, and hence can never produce an equal pair. To do so, refer to Fig. 4.9.

In this figure $AB = BC$, angle ABC is a right angle, AD is the bisector of angle CAB, and DE is drawn perpendicular to AC. Prove that $BD = DE$, $DE = EC$, and $AB = AE$. Then show that the Euclidean algorithm starting with the pair (AC, AB) leads first to the pair $(AB, EC) = (BC, BD)$, and then

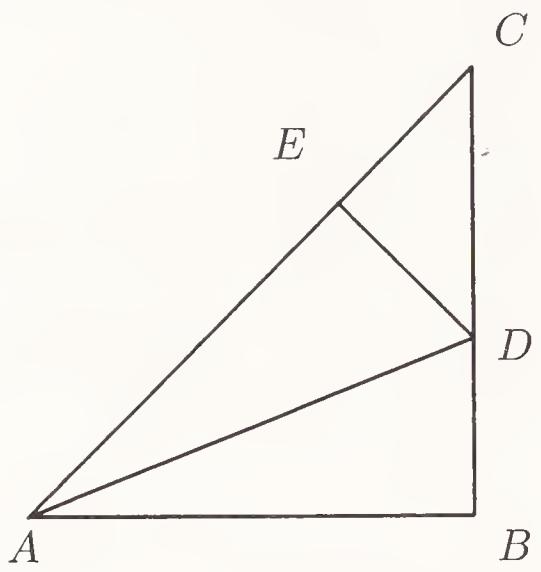


Figure 4.9: Diagonal and side of a square.

to the pair $(CD, BD) = (CD, DE)$, and these last two are the diagonal and side of a square.

Exercise 4.6 Proposition 1 of Book VI of Euclid's *Elements* asserts that two triangles having the same altitude have areas proportional to their bases. Draw two different-shaped triangles of equal altitude but with bases in the ratio of $5 : 3$. Prove that their areas are in the ratio of $5 : 3$ also. You may use the fact that two triangles with equal bases and equal altitudes have equal areas, but do *not* use the formula $A = \frac{1}{2}bh$. Instead, divide the two bases into 5 and 3 equal pieces and draw lines from the vertices to the endpoints of these pieces. How could you prove this theorem if the bases were not commensurable?

4.6.2 Questions about Early Greek Mathematics

Exercise 4.7 This exercise will take you to the library. Choose a Greek mathematician and find out what is said about him or her, on what authority, and (if possible) what original documentary sources for this information exist and where they are housed. Here are suggestions: Dinostratus, Hippias of Elis, Antiphon, Autolycus, Hypatia (the only woman mathematician of ancient times whose name is recorded by the standard sources).

Exercise 4.8 It was stated above that Thales might have used the Pythagorean theorem in order to calculate the distance from the center of the Great Pyramid to the tip of its shadow. How could this distance be computed without the Pythagorean theorem?

Exercise 4.9 Try to construct your own proofs of the theorems credited to Thales. Don't strain to remember the axioms of geometry—in Thales' time these axioms had surely not been formulated. Just try to deduce these statements logically from other principles that seem more obvious.

Exercise 4.10 In discussing the validity of various achievements credited to the Pythagoreans, the nineteenth-century historian of mathematics Moritz Cantor wrote the following.

> We shall not hesitate to ascribe to Pythagoras himself certain things belonging particularly to the history of mathematics. Among these is the *Pythagorean theorem*, which we shall attribute to him under any circumstances. It may be that no weight should be given to the testimony of Vitruvius, Plutarch, Diogenes Laertius and Proclus because of their late dates, even though they all agree. Nevertheless those whom Proclus cites as his defenders carry much more weight: "Those who wish to tell of ancient times," whether this phrase means Eudemus, as is commonly assumed, or not. Most convincing of all to us is the indirect confirmation in the list of old mathematicians. It is there stated explicitly that Pythagoras discovered the theory of irrationals. But such a theory would be completely impossible—the study of irrationals would be unthinkable—unless the theorem about the squares of the sides of a right triangle were known beforehand; and one would be in an even more difficult position if, by not crediting Pythagoras with this discovery, one were forced to assume it to be even older than Pythagoras.

How do we know that Cantor's conclusion here is wrong? What is wrong with his reasoning?

Exercise 4.11 State the paradoxes of Zeno in your own words and tell how you would have advised the Pythagoreans to modify their system in order to avoid these paradoxes.

Exercise 4.12 Do we share any of the Pythagorean mysticism about geometric shapes? Think of the way in which we refer to an honorable person as *upright*, or speak of getting a *square deal*, while a person who cheats is said to be *crooked*. Are there other geometric images in our speech that have ethical connotations?

Exercise 4.13 The Pythagoreans occupy a place in music history very similar to their place in the history of mathematics, that is, many legends have accumulated, and they have been credited by some scholars with achievements that other scholars regard as "too advanced" to be believable in their time. They are said to have recognized that the intervals we call the octave, the fifth, and the fourth are produced by clamping a plucked string so that half, two-thirds, or three-fourths of the string is allowed to vibrate. They thus associated tones with ratios, and supposedly declared that the pleasing tones are associated with the division of the string into commensurable lengths. Would you expect them to write a composition in which the octave would be "bisected," that is, one containing a tone that is the geometric mean between a given tone and its octave? (On a piano such a tone, relative to middle C and the C above it, is the F# between them.) If you have a piano, play the chord C-F#-C, and decide whether you consider it a pleasing sound.

Exercise 4.14 Was the overall effect of the discovery of incommensurables to advance or retard the development of mathematics? Keep in mind that it stimulated the discovery of a new theory of proportion, but also caused the Greek mathematicians to avoid regarding lengths as numbers. Might the Greeks have developed analytic geometry if not for this difficulty?

Exercise 4.15 In the period discussed in the present chapter we find two kinds of mathematical activity. One kind, represented by the attempt to extend the theory of the measurement of rectilinear plane figures to curvilinear and solid figures, is an attempt to discover new facts and enlarge the sphere of mathematics. The other, represented by the discovery of incommensurables, is an attempt to bring into sharper focus the theorems already proved and to test the underlying assumptions of a theory. Are these kinds of activity complementary, opposed, or simply unrelated to each other?

Exercise 4.16 In the discussion of Hippocrates' quadrature of a lune we used the fact that the areas of circles are proportional to the squares on their radii. Could Hippocrates have known this fact? Could he have proved it?

Exercise 4.17 In both the Egyptian and Babylonian documents we found many problems that we would now regard as algebra problems. What reasons can you give for the absence of such problems among the Greek writers? Were they taken for granted as problems that had already been solved? Were they overshadowed in importance by the new systematic geometry being developed? Did they appear in disguised form? Or is it possible that documents discussing such problems once existed, but have all been lost?

4.7 Endnotes

1. Toomer's edition of the *Almagest* was published in America by Springer-Verlag (New York, 1984). The quotation is from p. 5.

2. Gray's comment on the importance of the discovery of incommensurables is in the book *Ideas of Space. Euclidean, Non-Euclidean, and Relativistic*, 2nd ed. (Clarendon Press, Oxford, 1989), p. 15.

3. Herodotus' remarks on Thales can be found in his book *The History*, translated by David Grene (University of Chicago Press, 1987), pp. 67–68.

4. Proclus' commentaries are available in an annotated English translation by Glenn R. Morrow (Princeton University Press, 1970).

5. Diogenes Laertius' description of the Pythagorean cosmology can be found in the Loeb Classical Library edition of his *Lives of the Philosophers*, Vol. 2 (Putnam, New York, 1925), pp. 341–343. The cosmology is also discussed in the book by Walter Burkert, *Lore and Science in Ancient Pythagoreanism*, translated by Edwin L. Minar, Jr. (Harvard University Press, 1972). In this

book the difficulty of ascertaining what the Pythagoreans actually believed is amply demonstrated.

6. An English translation of Nicomachus' *Arithmetic* by Professor Martin Luther D'Ooge was annotated and published after the translator's death by F. E. Robbins and L. C. Karpinski (Macmillan, New York, 1926). The passages on the classification of numbers are found in Chapters XI–XIII, pp. 201–204.

7. Proclus' citation of the Pythagorean view of the ethical qualities of angles is from the edition mentioned above, pp. 106–107.

8. Zeno's paradoxes can be found in many sources. They are discussed in Aristotle's *Physics*, 239b and in the commentaries on this book by Simplicius and others. See *Simplicius: On Aristotle's Physics 6*, translated by David Konstan (Cornell University Press, 1989).

9. The discussion of Hippocrates' quadrature of lunes is based on an extended quotation by Simplicius, taken here from Vol. I of *Selections Illustrating the History of Greek Mathematics*, with a translation by Ivor Thomas (Harvard University Press, 1939), pp. 235–253.

Chapter 5

The Euclidean Synthesis

By the fourth century B.C.E. geometry had attained the status of a theory, that is, a body of knowledge given coherence by logical relations among its parts. Like all theories, it produced a number of unanswered questions and contained unresolved difficulties. We have seen some of these questions and difficulties in the last chapter in the form of the three classical problems, the paradoxes of Zeno, and the problem of incommensurables. Only the last of these was in pressing need of resolution at the time, and a solution was found by mathematicians working in Athens in the time of Plato. Once this difficulty was overcome, the way was clear for a comprehensive summary of the subject to be written, giving it a permanent place in human culture. The construction of this synthesis forms the subject of the present chapter. We begin our discussion with the statement and resolution of the problem of incommensurables during the fourth century B.C.E. We then discuss some other issues involved in organizing the presentation, followed by the contents of Euclid's *Elements* and some geometry that goes beyond Euclid.

5.1 The Problem of Incommensurables

5.1.1 Incommensurables in Plato's Dialogues

In a dialogue called the *Theatetus* Plato gives us a glimpse of what must have been a current debate over the problem of incommensurables. Like most of Plato's characters, the title character was a real person, an Athenian whose dates are given as 414–369 B.C.E. He was a friend of Plato and a student of the Pythagorean geometer Theodorus of Cyrene (ca. 460–399), who is mentioned in the dialogue.

Theatetus reports that Theodorus proved that the side of a square of area 3, 5, etc., up to 17 is not commensurable with the side of a square of area 1 (except, of course, squares of area 4, 9, and 16), saying that, for some reason Theodorus stopped at that point. On that basis the students decided to classify numbers (what we now call positive rational numbers) into "equilateral" and "oblong" numbers. The former class consists of the squares of rational numbers, such as $\frac{25}{9}$, and the

latter are all other positive rational numbers, such as 2.

One cannot help wondering why Theodorus stopped at 17 after proving that the sides of squares of areas 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, and 15 are incommensurable with the unit length. The implication is that Theodorus "got stuck" trying to prove this fact for a square of area 17. If such is the case, what caused him to get stuck? Most assuredly the square root of 17 is irrational, and the proof commonly given nowadays to show the irrationality of $\sqrt{3}$, for example, based on the unique prime factorization of integers, works just as well for 17 as any other number. If Theodorus had our proof, he wouldn't have been stuck doing 17, and he wouldn't have bothered to do so many special cases, since the proofs are all the same. Therefore we must assume that he had some other method.

An ingenious conjecture as to Theodorus' method was provided by W. Knorr (1945–1997) in his book *The Evolution of the Euclidean Elements* (Reidel, Dordrecht/Boston, 1975). Knorr suggests that the proof was based on the elementary distinction between even and odd. To see how such a proof works, suppose that 7 is an equilateral number in the sense mentioned by Theatetus. Then there must exist two integers such that the square of the first is seven times the square of the second. We can assume that both integers are odd, since if both are even, we can divide them both by 2, and it is impossible for one of them to be odd and the other even (the fact that the square of one equals seven times the square of the other would imply that an odd integer equals an even integer if this were the case). But it is well known that the square of an odd integer is always 1 larger than a multiple of 8. The supposition that the one square is seven times the other then implies that an integer 1 larger than a multiple of 8 equals an integer 7 larger than a multiple of 8, and this is clearly impossible.

This argument carries over to show that none of the odd numbers 3, 5, 7, 11, 13, and 15 can be the square of a rational number. With a slight modification it can also be made to show that none of the numbers 2, 6, 8, 10, 12, and 14 is the square of a rational number, though this is superfluous in the case of 8 and 12, since it is already known that $\sqrt{2}$ and $\sqrt{3}$ are irrational. Notice that the argument fails, as it must, for 9: a number 9 larger than a multiple of 8 is also 1 larger than a multiple of 8. However, it also breaks down for 17, and for the same reason: a number 17 larger than a multiple of 8 is also 1 larger than a multiple of 8. Thus, even though it is *true* that 17 is not the square of a rational number, the Pythagorean-type argument just given cannot be used to *prove* this fact. In this way the conjectured method of proof would explain why Theodorus got stuck at 17.

5.1.2 The Eudoxan Solution

The existence of incommensurable lines was well known by the time of Plato, and the problem that these lines posed for the geometric theory of proportion was acute. If we cannot say that $A : B :: m : n$ for positive integers m and n, how can we assert, for example, that the areas of two circles are proportional to the squares on their diameters? The solution that Euclid chose in his exposition of

this problem was invented a few years after the time of Theodorus.

The formulation of a usable definition of proportion for continuous quantities was given by Eudoxus of Cnidos (400–347 B.C.E.). Eudoxus' solution can be understood through the following considerations. Let D be the diagonal of a square and S its side, and let A and B be any two quantities of the same type, say lengths for definiteness. What could we mean by saying that $D : S = A : B$? If the quantities were commensurable, we would say that there are integers m and n such that $D : S = m : n$, that is, $nD = mS$, and also $A : B = m : n$, that is, $nA = mB$. In other words, the integers would be used as a "common currency" to express the equality of ratios of things of various kinds. Now we know that there *are* no integers m and n such that $D : S = m : n$. However, we also know two important bits of positive information. First, for any two integers m and n we must have either $D : S > m : n$ (that is, $nD > mS$) or $D : S < m : n$ ($nD < mS$), and a similar statement holds for A and B. Second, we can find pairs of integers m, n and p, q whose ratios are respectively larger and smaller than $D : S$ in the sense just mentioned while the ratios $m : n$ and $p : q$ are as close to each other as desired, so that it seems we can *approximate* the ratio $D : S$ with rational numbers as closely as desired.

Putting these two facts together suggests that we define the proportion $D : S = A : B$ to mean that for every pair of integers m and n the ratio $D : S$ is in the same relation to $m : n$ as $A : B$, that is, if $D : S > m : n$, then $A : B > m : n$; if $D : S = m : n$, then $A : B = m : n$, and if $D : S < m : n$, then $A : B < m : n$. When the ratios are eliminated by cross-multiplying, we see that this definition of the proportion $D : S = A : B$ amounts to the following: For any two integers m and n, if $nD > mS$, then $nA > mB$; if $nD = mS$, then $nA = mB$; and if $nD < mS$, then $nA < mB$.

This is the way the definition is given in Book V of Euclid (the material in brackets is added from the discussion just given to clarify the meaning):

> Magnitudes are said to be in the same ratio, the first to the second [$D : S$] and the third to the fourth [$A : B$], when, if any equimultiples whatever be taken of the first and third [nD and nA] and any equimultiples whatever of the second and fourth [mS and mB], the former equimultiples alike exceed, are alike equal to, or are alike less than the latter equimultiples taken in corresponding order [that is, $nD > mS$ and $nA > mB$, or $nD = mS$ and $nA = mB$, or $nD < mS$ and $nA < mB$].

5.1.3 How to Apply the Eudoxan Definition

Let us look at one example of the application of Eudoxus' definition of proportion, so that we will know how the seemingly cumbersome definition was used. A fundamental result in the theory of proportion is that two triangles having equal altitudes have areas proportional to their bases. This assertion is half of Proposition 1 of Book VI of Euclid's *Elements*. Let us now examine the proof, referring to Fig. 5.1.

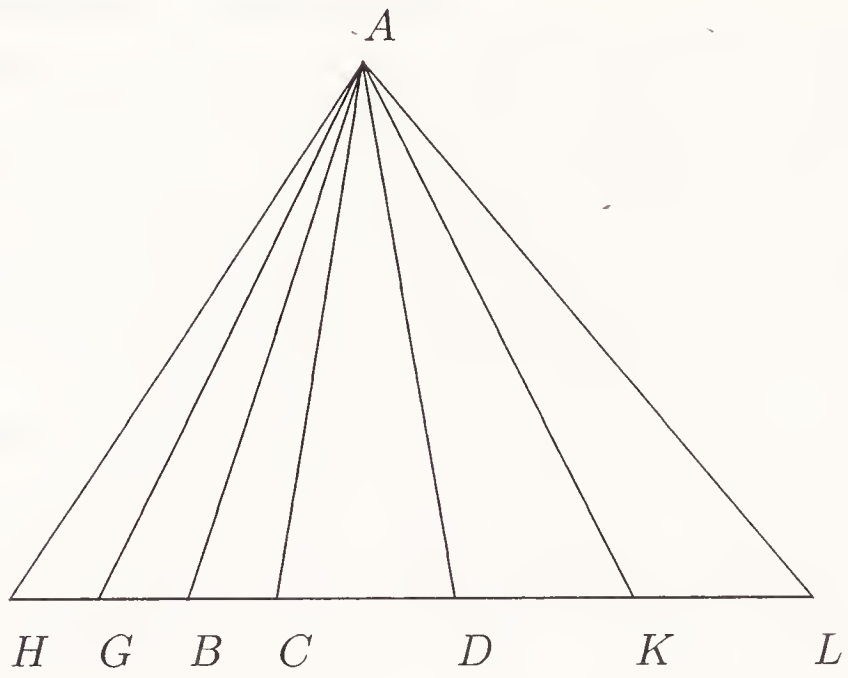


Figure 5.1: A use of the Eudoxan definition of proportion.

The proof depends on Proposition 38 of Book I, which asserts that two triangles having equal bases and altitudes are of equal area; and as a corollary, if two triangles have equal altitudes but unequal bases, the one with the larger base is the larger.

In the figure ABC and ACD are triangles having the same altitude. We want to prove that their areas have the same ratio as their bases BC and CD. To do that, we take any multiple m of the base BC (the figure shows how to do this by extending BD leftward to H so that $BC = BG = GH$, giving the triangle AHC which is 3 times triangle ABC). Similarly take some multiple n of the base CD (again this is done in the figure by extending CD to L so that $CD = DK = KL$, yielding triangle ACL equal to 3 times triangle ACD).

Then if $mCH > nCL$, triangle ACH is larger than triangle ACL, that is, $m\triangle ABC > n\triangle ACD$, and similar reasoning applies when we have $mCH = nCL$ or $mCH < nCL$, all three arguments being direct consequences of Book I, Proposition 38.

These conclusions are by definition what is meant by the proportion $ABC : ACD :: BC : CD$, which is therefore proved.

5.2 Other Issues in Geometry

5.2.1 Aristotle's View of Mathematics

Plato died in 347 B.C.E. In the quarter-century after his death the development of Greek thought, at least as it seems from the long perspective of history, was dominated by Aristotle (384–322), whose views diverged from those of Plato in certain important respects. Aristotle's writings cover a much wider field than those of Plato, embracing ethics, literature, medicine, natural science, mathematics, and

logic. Like Plato, in mathematics he seems more like a well-informed generalist than a specialist.

Theories in General

Aristotle gave a very thorough and rigorous discussion of formal inference, the validity of various kinds of deductions (syllogisms, reductio ad absurdum, etc.) in his treatise *Prior Analytics*. We are so used to the application of logic in science and mathematics that it is difficult to imagine a time when logic was a new subject. It appears that Aristotle invented a good deal of this subject. His word for deduction, for example, is *syllogismos* ($\sigma\upsilon\lambda\lambda o\gamma\iota\sigma\mu o\varsigma$), which means *reasoning together*. The more specific meaning we give to the word *syllogism* today is the result of Aristotle's emphasis on the importance of this kind of inference. Three parts of a theory are relevant to our present purpose, namely *definitions*, *postulates*, and *deduction*.

Definitions. According to Aristotle, a definition is "a statement that describes the essence of a thing." (He gives this definition of definition, or one equivalent to it, in several places.) The idea of a definition thus seems to be to codify an intuition in words. This enterprise, as we now know, has certain limitations, since a basic level of understanding is necessary *before* anything at all can be communicated. The attempt to define *everything* leads in a circle. Most twentieth-century philosophers regard definitions as constructions of pure thought, which may or may not correspond to something real, whereas Aristotle thought of his "essences" as real. He thought of definitions as discoveries rather than inventions.

In the twentieth century there has been a view of mathematics called *logicism*, associated with the British philosopher Bertrand Russell (1872–1970) according to which mathematics is merely an extension of formal logic. A related view, associated with the German mathematician David Hilbert (1862–1943) and known as *formalism*, regards mathematical propositions as purely formal assertions, having no meaning until the terms they contain are given a real-world interpretation. From these two points of view, mathematical theories begin with undefined terms that do not correspond to anything real or even intuitive. Needless to say, Greek mathematical works should not be judged by the degree to which they comply with the canons of logicism and formalism. Euclid was not striving unsuccessfully to create Hilbert's mathematics. In forming judgments about Euclid's works we must see what he was attempting to do and then decide how successful the end result was. It should be remarked that both logicism and formalism, as well as their rival, intuitionism, are *meta*mathematical theories. Practicing mathematicians, though they may subscribe to some overall view of mathematics in private, are not restricted by these principles in their daily work.

Postulates. The second important starting point for a theory is a set of *postulates*. On the modern view these are merely starting points, not statements accepted as true. For, on the modern view, the terms they contain have no meaning until they are interpreted, and mathematics proper is not concerned with interpretation. The view of Aristotle and Euclid was different. They were not consciously engaged in

building purely formal systems. To them points, lines, and planes were not mere undefined terms, but rather essential constituents of the physical world (or at least of an ideal world). The postulates taken without proof therefore had *meaning*, that is, were capable of being true or false. They were, of course, believed to be true.

Aristotle distinguishes between universal first principles, and those proper to a particular branch of science. In both cases, however, he believes the principles are in some sense true.

> Instances of first principles peculiar to a science are the assumptions that a line is of such and such a character, and similarly for the straight line, whereas it is a common principle, for instance, that if equals be subtracted from equals the remainders are equal.

Deduction. In most of Aristotle's *Prior Analytics* the illustrative examples of logical deductions involve familiar concepts such as color and familiar objects such as animals and people. Only occasionally does Aristotle give an example from geometry. One important occasion on which he does invoke geometry occurs in section 65*a* of the *Prior Analytics*, in which he discusses the question of whether parallel lines exist. It is interesting to find this question being discussed for several reasons. First, the notion of parallelism (which literally means "lying alongside") is not discussed in the dialogues of Plato. Parallel lines do exist in Euclidean geometry, of course, and most people seem to have a Euclidean intuition. The best proof of that fact is that the Euclidean form of the Pythagorean theorem was discovered independently in several different civilizations. Second, it is curious that Aristotle considered the existence of parallel lines doubtful. Why would he have doubts about something that is so clear on an intuitive level? One possible reason is that parallelism involves the infinite: parallelism asserts that two finite line segments will *never* meet, no matter how far they are extended. If geometry is interpreted physically (say, by regarding a straight line as the path of a light ray), we really have no assurance whatever that parallel lines exist—how could anyone assert with confidence what will happen if two apparently parallel lines are extended to a length of hundreds of light years? Third, Aristotle's discussion shows that he understood well the logical issues involved and the kinds of geometry in which there would be no parallel lines. He writes:

> . . . it is really not strange for the same falsehood to result by means of several assumptions, as for instance, it results that parallels intersect both if the internal angle is greater than the external and if a triangle has more than two right angles. . .

This last sentence gives an intriguing glimpse of an issue in geometry just before the time of Euclid. Aristotle knows that an exterior angle of a triangle is larger than either opposite interior angle. This fact is proved by Euclid as Proposition 16 of Book I, and we may infer that this proof was known to Aristotle. Euclid makes a number of fundamental facts depend on this proposition, among them the fact that parallel lines exist (Book I, Proposition 27), and Aristotle shows that he is aware of this connection. Proposition 16 also implies that the sum of the

angles of a triangle is at most two right angles, though this fact is not proved by Euclid, since the parallel postulate implies that the sum of the angles is exactly two right angles (Book I, Proposition 32). Aristotle's remark that there are no parallel lines (or, as he puts the matter oxymoronically, that parallels intersect) if the sum of the angles of a triangle is greater than two right angles gives an early hint as to what should be expected in a certain kind of noneuclidean geometry. Aristotle, of course, did not view the question this way, since he was sure that the angles of a triangle do not total more than two right angles. Nevertheless, in the light of noneuclidean geometry, the phrase just quoted seems amazingly prescient.

5.3 Euclid's *Elements*

By the year 300 B.C.E. the materials were available for writing a comprehensive summary of the basic parts of geometry. The order of presentation, choice of definitions, and assumptions, and the like would be decided according to the author's taste. Undoubtedly more than one such treatise was written in the late fourth century B.C.E., but all were superseded by one particular treatise that became the standard and remained so for centuries. This treatise is Euclid's *Elements* (Fig. 5.2), a textbook on the mathematics of the Pythagoreans, both arithmetic and geometry, as emended by the Eudoxan theory of proportion. This mathematics included all the standard geometry of polygons and circles, polyhedra and spheres, and the theory we would call quadratic irrationals. Since the geometry taught in high schools nowadays is a modernized approach (and considerably watered-down in terms of difficulty), it is worthwhile to look at this treatise on its own terms. We begin with a few words about its author.

5.3.1 Euclid of Alexandria

A biography of Euclid written in the twentieth century would necessarily be very brief, since almost nothing is known about the man or his life. He must have lived around 300 B.C.E., since he lived before Archimedes, whose death can be precisely dated, and worked in the research center in Alexandria, which was founded by Ptolemy I after the death of Alexander the Great in 322 B.C.E. Euclid is believed to have been invited to Alexandria by Ptolemy. There is a great deal of conjecture about him—that he may have worked in Athens before coming to Alexandria, that he was the founder of the Alexandrian mathematical school, etc., but as far as definite *knowledge* goes, he is defined for us only as the author of his books, chiefly the *Elements*, but also a few others that we shall not discuss: the *Optics*, the *Data*, the *Division of Figures*, and the *Phaenomena*.

5.3.2 General Nature of the *Elements*

Euclid's *Elements* consists of 13 books, of which only certain parts of the first 4 books are commonly taught in plane geometry nowadays. These first four books

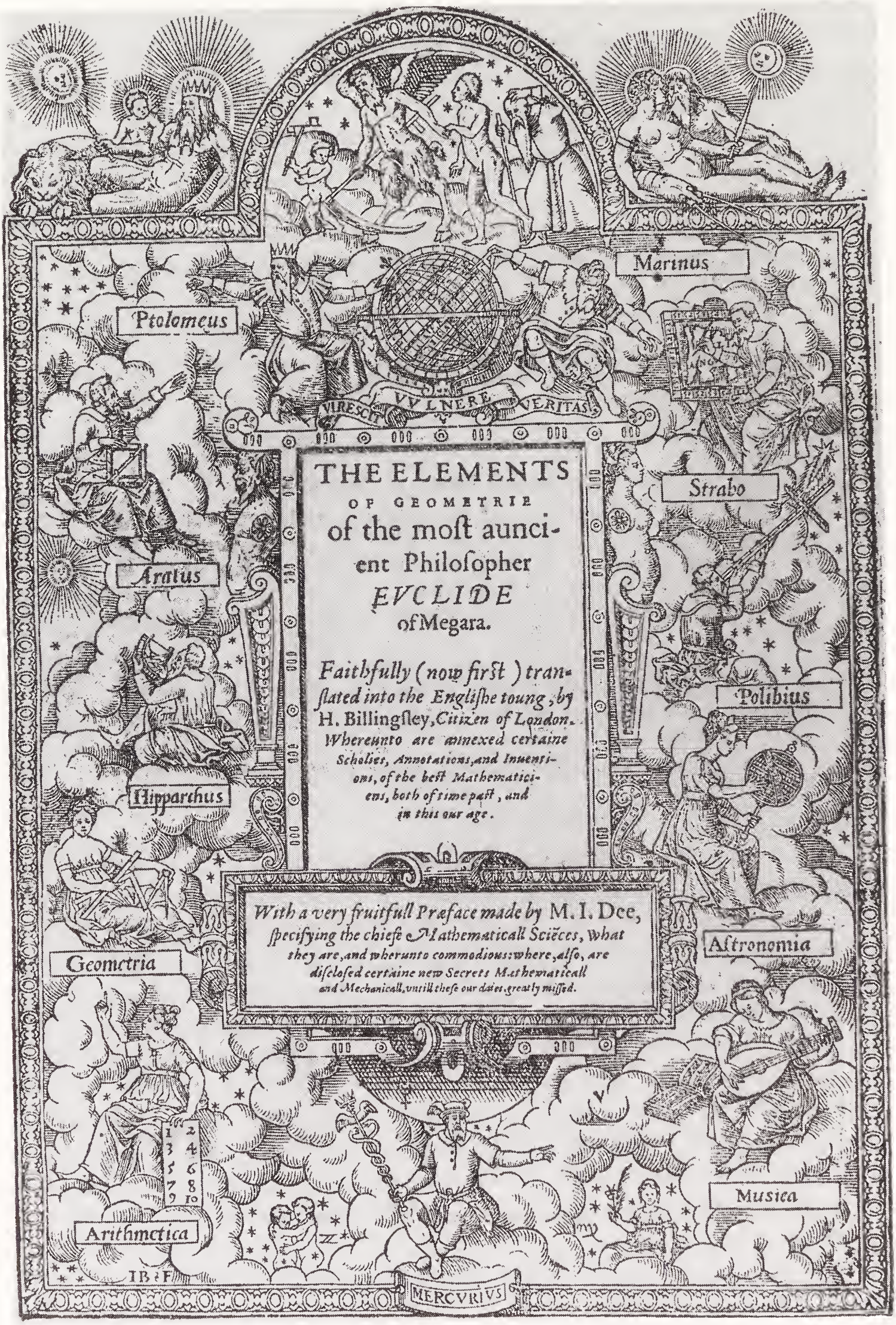


Figure 5.2: Frontispiece from Euclid's *Elements*. The Bettmann Archive.

contain the theory of angles, lines, circles, triangles, squares, etc. Book V is devoted to the Eudoxan theory of proportion, and Book VI is commonly referred to as "geometric algebra," since it contains a series of geometric constructions that are logically equivalent to the solution of certain equations. Books VII–IX contain Pythagorean number theory. Book X (the longest book of all) contains a thorough study of quadratic surds, which are irrationals of the form $\sqrt{a+\sqrt{b}}$, and Books XI–XIII are devoted to solid geometry, although they also contain some fundamental parts of plane geometry needed to establish the theory of proportion for solid figures.

The *Elements* is the earliest surviving example of a systematic treatise expounded in logical order. As such it became a kind of ideal model for scientific treatises, and philosophers and scientists strove to imitate its economy of expression. Inevitably, too, readers began to detect weaknesses in the treatise. Some of these weaknesses were the unavoidable flaws that every author must contend with. Others, as we shall see, were not really weaknesses from Euclid's point of view, merely signs that his purposes were different from those of his later critics.

D. H. Fowler, in *The Mathematics of Plato's Academy*, gives a thorough discussion of the existing manuscripts of Greek geometry. A few ostraca (shells) dated to the late third century B.C.E. have been found on Elephantine Island in the Nile River near the site of the ancient city of Syene and the modern city of Aswan, about 500 miles south of Alexandria. Some scrolls containing parts of the *Elements* were found in the ruins of Herculaneum, which was destroyed by the eruption of Mt. Vesuvius in 79 C.E., but only a few fragments have survived the attempt to unroll them. Other papyri dated to 75–125 C.E. from Oxyrhyncus, a Roman town in Egypt about 120 miles south of present-day Cairo, contain geometric propositions that are recognizably part of Euclid. The earliest complete texts, however, are a ninth-century manuscript in the Bodleian Library at Oxford and a tenth-century manuscript in the Vatican.

5.3.3 The Logical Development of Geometry

The Definitions

Looking into Euclid's *Elements* from a twentieth-century perspective, one cannot help noticing an apparently futile attempt to define the undefinable. Book I begins with 23 "definitions," of which we give only a few samples:

1. A *point* is that which has no part.

2. A *line* is breadthless length.

3. The extremities of a line are points...

23. *Parallel* straight lines are lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

The modern view is that one cannot define everything. While this view has much to recommend it as a way of understanding mathematics, it should not be

projected backwards into the minds of the mathematical pioneers. Euclid had a definite interpretation in mind for the words *point, line, plane*, etc., and the attempted definitions just given express what that interpretation was. It should be noticed that, despite modern views of what geometry "really" is, we all learn geometry *as if* it had some connection with physical space. Thus Euclid's definitions, though logically defective, have been psychologically very effective in communicating a particular interpretation of geometry for more than two millennia.

The twenty-third and last of Euclid's definitions is the definition of parallel lines. You should pause and ask yourself what definition you would have given of parallel lines. Most students, when asked, reply that parallel lines are lines that are equidistant from each other. This property of parallel lines can be proved from Euclid's postulates, but it is stronger than the definition given by Euclid, and it is not obvious that two lines can be equidistant (the assumption that they can is equivalent to the parallel postulate). About every definition one should ask whether objects satisfying the conditions of the definition exist. In Definition 22, for example, Euclid defines a square as an equilateral quadrilateral having right angles. It is not obvious that such an object exists, and indeed it turns out that the existence of rectangles is equivalent to Euclid's parallel postulate. The existence of parallel lines can be proved without the parallel postulate; but, as Aristotle pointed out in the passage quoted above, certain assumptions that he regarded as false—for example, that an exterior angle of a triangle may be smaller than an opposite interior angle—can lead to a proof that there are no parallel lines. Euclid's proof that parallel lines exist (Book I, Proposition 27) is based on two assumptions that are not made explicitly—they are part of the interpretation he gave to his terms, as we shall see below.

To summarize, Euclid chose to define parallel lines as coplanar nonintersecting lines. From Euclid's postulates, including two "hidden" postulates, it is possible to prove that parallel lines exist without using the parallel postulate. If one *defines* parallel lines differently, for example, as lines that are equidistant at all points, then the assumption that a pair of parallel lines exists is equivalent to Euclid's parallel postulate. We see, then, that a theory can be looked at from different perspectives. In one order of development a statement may be a theorem, while in another it is a definition.

The Postulates

After giving his twenty-three definitions, Euclid passes to his postulates for geometry. There are five of these: (1) [It is possible] to draw a straight line from any point to any point; (2) [it is possible] to produce [extend] a finite straight line continuously in a straight line; (3) [it is possible] to describe a circle with any center and radius; (4) all right angles are equal [congruent]; and (5) if a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on the side on which the angles are less than two right angles.

The first four of these postulates have always been accepted as reasonable starting points for geometry, but the complicated nature of the fifth postulate aroused

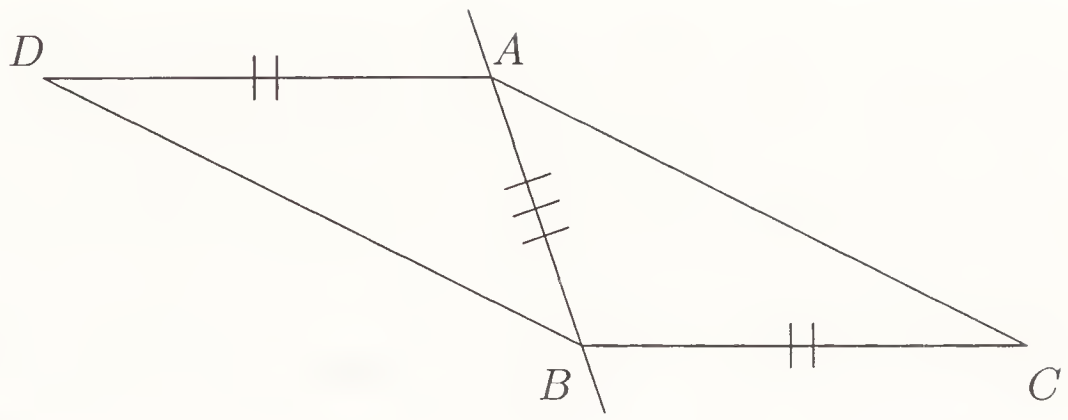


Figure 5.3: Proof that parallel lines exist (Euclid I, Proposition 27).

controversy from the very beginning. Once its rather complex language has been assimilated, it is seen to be intuitively plausible, and indeed nearly obvious to most people, for the following reason. One can prove—in fact Euclid *does* prove—that a pair of lines will be parallel under certain conditions, namely when the lines are cut by a transversal and the interior angles on one side of the transversal are together equal to two right angles. An informal proof of this result (*not* Euclid's proof) proceeds by finding a contradiction when it is assumed that the lines intersect. Since the proof is by contradiction, the illustration of it requires a figure that is impossible, so the reader will have to bear with us while we make some statements that are visually absurd. We shall assume that two lines cut by a transversal in such a way as to make the interior angles on one side equal to two right angles, nevertheless meet on that side. Thus, in Fig. 5.3, let BC be a straight line and let AC also be a straight line. Assume that $\angle ABC + \angle BAC$ equals two right angles (this is visually wrong in the figure, the first impossibility in it). Let the straight line CA be extended from A away from C to D so that $AD = BC$ (this is glaringly wrong, since CAD is clearly not a straight line, but remember, we are seeking a contradiction here—we *know* the figure is impossible). Since CAD is a straight line, it follows that $\angle DAB + \angle BAC$ equals two right angles, and therefore that $\angle DAB = \angle ABC$. But then, if we draw BD, since side AB is common and $AD = BC$, it follows that $\triangle ABC$ is congruent to $\triangle BAD$. Then $\angle ABD = \angle BAC$, and so $\angle ABC + \angle ABD$ equals two right angles also, so that DBC is also a straight line. Thus the two points D and C are joined by two distinct straight lines DAC and DBC.

Now the question arises as to just what contradiction we have obtained here. Where in his axioms did Euclid say that *only* one straight line can be drawn between two distinct points? Nowhere, apparently. This is one of the "hidden" axioms that Euclid didn't write down, but conveyed as part of the interpretation of his system. On this point the modern mathematicians who have "improved" Euclid are completely correct. Except for this assumption the great circles on a sphere satisfy all of Euclid's axioms (including the parallel postulate), and yet any two of them intersect. Since Euclid knew all about great circles on a sphere, it is clear that he intended his plane geometry to exclude this case. Let us therefore keep in mind that Euclid assumed this axiom implicitly. Do we now have a contradiction? Only if we know that C and D are really distinct points. How can we know this?

If you reflect on the figure, it may occur to you that C and D lie on opposite sides of the line AB, and hence must be distinct points. Here again, however, we run into a hidden axiom. How do we know that a line separates the plane into two disjoint half-planes? To be explicit about this point, modern mathematicians have had to add certain axioms stating that lines are *ordered* in the sense that if A, B, and C are distinct points on a line, then precisely one of these points is between the other two. Like the other hidden axiom, this assumption is intuitively clear and for that reason hard to make explicit. Without it the geometry of the projective plane satisfies all of Euclid's axioms, and again, any two lines intersect.

Thus, by slightly emending Euclid we arrive at a proof of the intuitive proposition that two lines are parallel if the interior angles they form on one side of a transversal total two right angles. Putting the same thing another way, there is no triangle (such as the hypothetical $\triangle ABC$ above) in which the sum of two of the three angles is equal to or greater than two right angles. What makes the parallel postulate reasonable in this context is that it seems to be the converse of this assertion—given a base (the transversal AB) and two angles on one side of that base totaling *less* than two right angles, those two sides will meet and form a triangle. This is the form in which Euclid states the parallel postulate, and the form most useful for the purpose of proving theorems.

The Disputed Parallel Postulate

We have seen through the quotation from Aristotle that parallelism was the subject of debate even before Euclid. The question of the proper definition of parallel lines and what it is reasonable to assume about parallelism must have been difficult to decide, and any decision would have been likely to encounter criticism. Certainly the decision actually made by Euclid did encounter criticism. We know from Proclus' *Commentary* that other mathematicians tried to find ways to eliminate this assumption, proving it as a theorem. Thus began a long enterprise that resulted ultimately in a clarification of the connection between geometry and the physical world and a clarification of the logical relations among the objects studied by geometry. This enterprise encompassed some of the best mathematics ever done, and we shall follow it from now on. At present we merely note what Proclus has to say.

Proclus claimed that the parallel postulate could be proved, in fact that it *had been* proved by Ptolemy. Unfortunately, he was unable to make good on his advertising when it came time to produce the proof. To prove the parallel postulate it would suffice to show that if parallel lines are cut by a transversal, then the interior angles on each side of the transversal total two right angles. Referring to Fig. 5.4, Proclus says the following:

> ... AF and CG are no more parallel than FB and GD, so that if the line falling on AF and CG makes the interior angles greater than two right angles, so also does the line falling on FB and GD make the interior angles greater than two right angles. But these same angles are less than two right angles (for the four angles AFG, CGF, BFG,

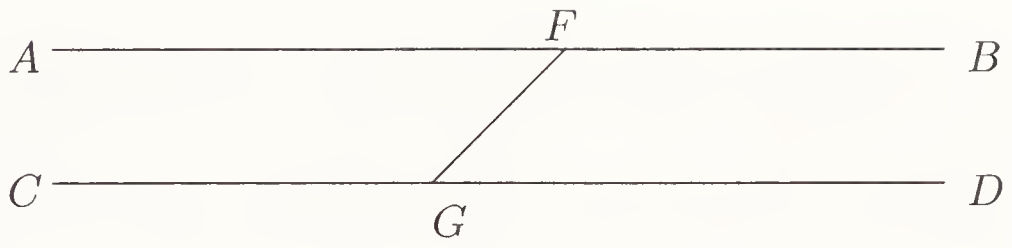


Figure 5.4: Ptolemy's "proof" of the parallel postulate.

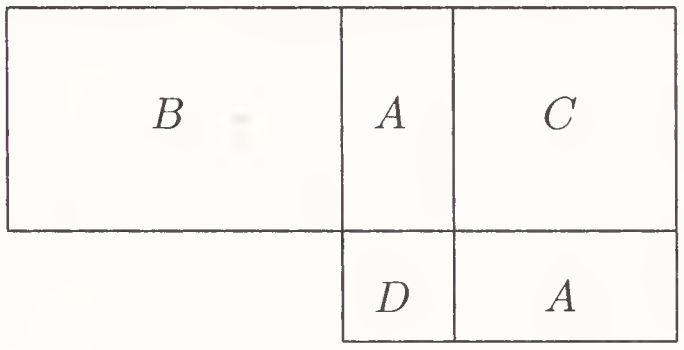


Figure 5.5: Expression of a rectangle as the difference of two squares.

and DGF are equal to four right angles), which is impossible.

The weak point in this argument is the pseudomathematical assertion that AF and CG are "no more parallel" than FB and GD. There are no degrees of parallelism. What Ptolemy and Proclus had in mind is that parallelism is *bilateral*, so that if lines do not meet, they should form a figure symmetric about any transversal. The argument, though it assumes what is to be proved, nevertheless has some positive value. It shows a certain consequence that must be accepted if the parallel postulate is denied, namely that parallelism is no longer two-sided. If the parallel postulate had been merely accepted as obvious, these consequences would never have been explored. Thus the attempt to create a deductive system, far from hindering imagination, actually stimulates it to create new ideas.

5.3.4 Contents of the *Elements*

The contents of the first book of the *Elements* are covered in the standard geometry courses given in high schools. This material involves the elementary geometric constructions of copying angles and line segments, drawing squares, etc., and the basic properties of parallelograms, culminating in the Pythagorean theorem (Proposition 47). In addition, these properties are applied to the problem of transformation of areas, leading to the construction of a parallelogram with a given base angle, and equal in area to any given polygon (Proposition 45). There the matter rests until the end of Book II, where it is shown (Proposition 14) how to construct a square equal to any given polygon.

Book II, in contrast to Book I, is neglected in most high school geometry texts. It contains geometric constructions needed to solve problems that may involve

quadratic incommensurables without resorting to the Eudoxan theory of proportion. For example, a fundamental result is Proposition 5: *If a straight line is cut into equal and unequal segments, the rectangle contained by the unequal segments of the whole together with the square on the straight line between the points of the section is equal to the square on the half.* This proposition is easily seen using Fig. 5.5, in terms of which it asserts that $(A + B) + D = 2A + C + D$, that is, $B = A + C$.

This proposition, in arithmetic form, appeared as a fundamental tool in the cuneiform tablets. For if the unequal segments of the line are regarded as two unknown quantities, then half of the segment is precisely their average, and the straight line between the points (that is, the segment between the midpoint of the whole segment and the point dividing the whole segment into unequal parts) is precisely what we called earlier the semidifference. Thus this proposition says that the square of the average equals the product plus the square of the semidifference; and that result was fundamental, as we saw in Chapter 3, for solving the important problems of finding two numbers given their sum and product or their difference and product. The geometric equivalent of these problems, however, does not appear until Book VI, Proposition 28: *To a given straight line to apply a parallelogram equal to a given rectilineal figure [polygon] and deficient by a parallelogram figure similar to a given one.* As we saw in the previous chapter, when the "defect" is a square, this problem, known as *application with defect*, is equivalent to finding two numbers having a prescribed sum and product. Proposition 29 of Book VI says: *To a given straight line to apply a parallelogram equal to a given rectilineal figure and exceeding by a parallelogram figure similar to a given one.* When the "excess" is a square, this problem (*application with excess*) is equivalent to the finding two numbers having prescribed difference and product. It always has a unique solution. We leave it to the reader to surmise why the treatment of this topic is delayed to Book VI, when it seems to mesh so neatly with the material of Book II.

These application problems are also important because of a geometric connection with the conics that we shall study in the next chapter. The Greek word for application is *parabole* ($\pi\alpha\rho\alpha\beta o\lambda\acute{\eta}$), whose roots *para*, meaning *alongside* (think of paramedics, paramilitary, and paralegals) and *bole*, meaning *throw* (think of *ballistics*), are reflected in the English cognate word *parable*, meaning a story with an application (moral). The Greek word for application with excess is *hyperbole* ($'\upsilon\pi\epsilon\rho\beta o\lambda\acute{\eta}$), which has a literary meaning of exaggerating ("overshooting"), and application with defect is called *elleipsis* ($'\acute{\epsilon}\lambda\lambda\epsilon\iota\psi\iota\varsigma$), yet another word used in literature to denote the omission of certain words that are understood without being expressed. Mathematicians know these words as the names of the three different kinds of conic sections. The reason for using these terms for conics will appear in the next chapter.

Books III and IV take up topics familiar from high-school geometry: circles, tangents and secants, and inscribed and circumscribed polygons. In particular, Book IV shows how to inscribe a regular pentagon in a circle (Proposition 11) and how to circumscribe a regular pentagon about a circle (Proposition 12), then reverses the figures and shows how to get the circles given the pentagon (Proposi-

tions 13 and 14). After the easy construction of a regular hexagon (Proposition 15), Euclid finishes off Book IV with the construction of a regular 15-sided polygon (Proposition 16).

Book V contains the all-important theory of geometric proportion based on the work of Eudoxus discussed above. Here we find (Proposition 13) the explicit construction of the mean proportional between two line segments. As already mentioned, Book VI applies this theory to the transformation of areas to solve the important problems of application with defect and excess. In Book VI Euclid also constructs the famous Golden Section (Proposition 30): *To divide a line into mean and extreme ratio.* This means to find a point on the line so that the whole line is to one part as that part is to the second part. In fact this construction is precisely that of applying an area equal to the square on the whole line segment in such a way that the excess is a square. As such it is a special case of the general problem of application with excess. The Pythagorean theorem is then generalized to cover not merely the squares on the sides of a right triangle, but any similar polygons on those sides (Proposition 31). The book finishes with the well-known statement that central and inscribed angles in a circle are proportional to the arcs they subtend.

Books VII–IX are nongeometrical and devoted entirely to Pythagorean number theory. Here, since irrationals cannot occur, the notion of proportion is redefined to eliminate the awkward Eudoxan technique. Book VII develops proportion for positive integers as part of a general discussion of how to reduce a ratio to lowest terms. The notion of relatively prime numbers is introduced, and the elementary theory of divisibility is developed as far as finding least common multiples and greatest common factors. Book VIII resumes the subject of proportion and extends it to squares and cubes of integers. Book IX continues this topic; it also contains the famous theorem that there are infinitely many primes (Proposition 20) and ends with a method of constructing perfect numbers (Proposition 36): *If the sum of the numbers* $1, 2, 4, \ldots, 2^{n-1}$ *is prime, then this sum multiplied by the last term will be perfect.* The modern statement of this fact is given in the exercises below. To see this recipe at work, start with 1, then double and add: $1 + 2 = 3$. Since 3 is prime, multiply it by the last term, that is, 2. The result is 6, a perfect number. Continuing, $1 + 2 + 4 = 7$, which is prime. Multiplying 7 by 4 yields 28, the next perfect number. Then, $1 + 2 + 4 + 8 + 16 = 31$, which is prime. Hence $31 \cdot 16 = 496$ is a perfect number. For practice the reader should continue this procedure and find the next perfect number. It is of interest that no perfect number has yet been found that is *not* generated by this procedure, although no proof exists that all perfect numbers are of this form.

Book X occupies fully one-fourth of the entire length of the *Elements*. For its sheer bulk, one would be inclined to consider it the most important of all the thirteen books, yet its 115 Propositions are among the least studied of all, principally because of their technical nature. The irrationals constructed in this book by taking square roots are needed in the theory developed in Book XIII for inscribing regular solids in a sphere. Book X begins with the operating principle of the Euclidean algorithm (Proposition 1): *Given two unequal quantities, if from the larger a quantity larger than its half is subtracted, and from that which is left a quantity larger than its half, and so forth, eventually the remaining quantity will*

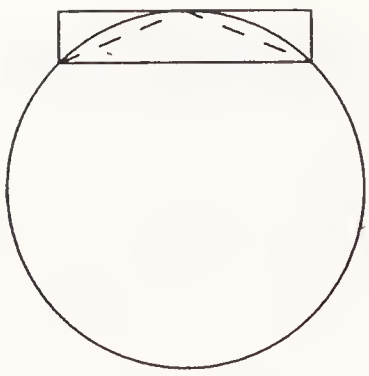


Figure 5.6: Bisecting an arc.

be less than the smaller given quantity. The way in which this process can be used to exhibit incommensurables, by showing that the Euclidean algorithm does not terminate, then follows (Book X, Proposition 2): *If, when the smaller of two given quantities is continually subtracted from the larger, that which is left never divides evenly the one before it, the quantities are incommensurable.* (Recall that we used this method of showing that the side and diagonal of a regular pentagon are incommensurable.) Proposition 1 forms the basis of what is known as the *method of exhaustion*, a way of proving certain proportionalities for curved figures (for example, that two circles are proportional to the squares on their diameters) by first proving that the proportionalities hold for similar polygons, then approximating the curved line by a polygon. In the case of circles, for instance, it is easy to prove that when an arc is bisected, the segment between the original chord and the circle is replaced by two smaller segments, which together are less than half of the original (see Fig. 5.6). Thus, if an inscribed regular polygon is thought of as approximating the circle, doubling the number of sides decreases the error of approximation by more than half. It therefore follows from this proposition that a circle can be approximated arbitrarily closely by an inscribed polygon.

Book XI contains the elementary parts of the solid geometry of planes, parallelepipeds, and pyramids. The theory of proportion for these solid figures is developed in Book XII, where one finds neatly tucked away the important theorem that circles are proportional to the squares on their diameters (Proposition 2).

The proof of Proposition 2 of Book XII is a typical instance of the use of the method of exhaustion together with the Eudoxan definition of proportion. First the Eudoxan theory is used in Proposition 1 to show that similar polygons inscribed in circles are proportional to the squares on the diameters of the circles. Then if S_1 and S_2 are the squares on the diameters of circles C_1 and C_2, respectively, it is first assumed that $S_1 : S_2 :: C_1 : D$, where D is less than the second circle. At this point similar polygons P_1 and P_2 are inscribed in the two circles, so that the area of P_2 is larger than the area D. By Proposition 1 we then have $S_1 : S_2 :: P_1 : P_2 < C_1 : D$, since $P_1 < C$ and $P_2 > D$, contradicting the original assumption. The same reasoning shows that one cannot have $S_2 : S_1 :: C_2 : D$ for any area D less than C_1, and hence the only remaining possibility is that $S_1 : S_2 :: C_1 : C_2$.

The difficulty with all arguments like this one that use the Eudoxan definition of proportion and the method of exhaustion is that one must *know* the result in

advance: the method is not a method of discovery. Such being the case, we are left wondering how such results were *discovered* in the first place. We now know that alongside the rigorous techniques that make the *Elements* such a logical masterpiece the Greeks also had an informal, intuitive method of reasoning, usually referred to as the method of indivisibles, which led them to make discoveries that could then be established rigorously by the method of exhaustion. We shall examine this method in the next chapter when we discuss Archimedes.

Book XII continues by establishing the usual proportions and volume relations for solid figures; for example, a triangular prism can be divided by planes into three pyramids, all having the same volume (Proposition 7), a cone has one-third the volume of a cylinder on the same base, similar cones and cylinders are proportional to the cubes of their linear dimensions, ending with the proof that spheres are proportional to the cubes on their diameters (Proposition 18)

Book XIII of the *Elements* is devoted to the construction of the regular solids and the relation between their dimensions and the dimensions of the sphere in which they are inscribed. The last proposition (Proposition 18) sets out the sides of these regular solids and their ratios to one another. An informal discussion following this proposition concludes that there can be only five regular solids.

5.4 Contemporaries of Euclid

The *Elements* was by no means a complete treatise of all that was known in Euclid's time. It was rather, as its name suggests, a treatment of the essential core of geometry to prepare the student to do research in the advanced topics of current research. There were other topics, in which research was still going on, that do not appear in the *Elements*. For example, Euclid never mentions conic sections in the *Elements*, even though he does mention a section of a cylinder in Book XII, Proposition 13 and is known to have written a separate treatise on this subject (it has been lost). We shall now look briefly at some of this other mathematics.

5.4.1 Menaechmus

Eutocius and Proclus both attribute the discovery of the conic sections to Menaechmus, who lived in Athens in the late fourth century B.C.E. Proclus, quoting Eratosthenes, refers to "the conic section triads of Menaechmus." Since this quotation comes just after a discussion of "the section of a right-angled cone" and "the section of an acute-angled cone," it is inferred that the conic sections were produced by cutting a cone with a plane perpendicular to one of its elements. Then if the vertex angle of the cone is acute, the resulting section (called an *oxytome*) is an ellipse. If the angle is right, the section (*orthotome*) is a parabola, and if the angle is obtuse, the section (*amblytome*) is an hyperbola (see Fig. 5.7).

Eutocius, in a commentary on a work of Archimedes, credits Menaechmus with the following solution of the problem of the two mean proportionals, that is, given two lines a and d, to find lines b and c such that $a : b :: b : c :: c : d$.

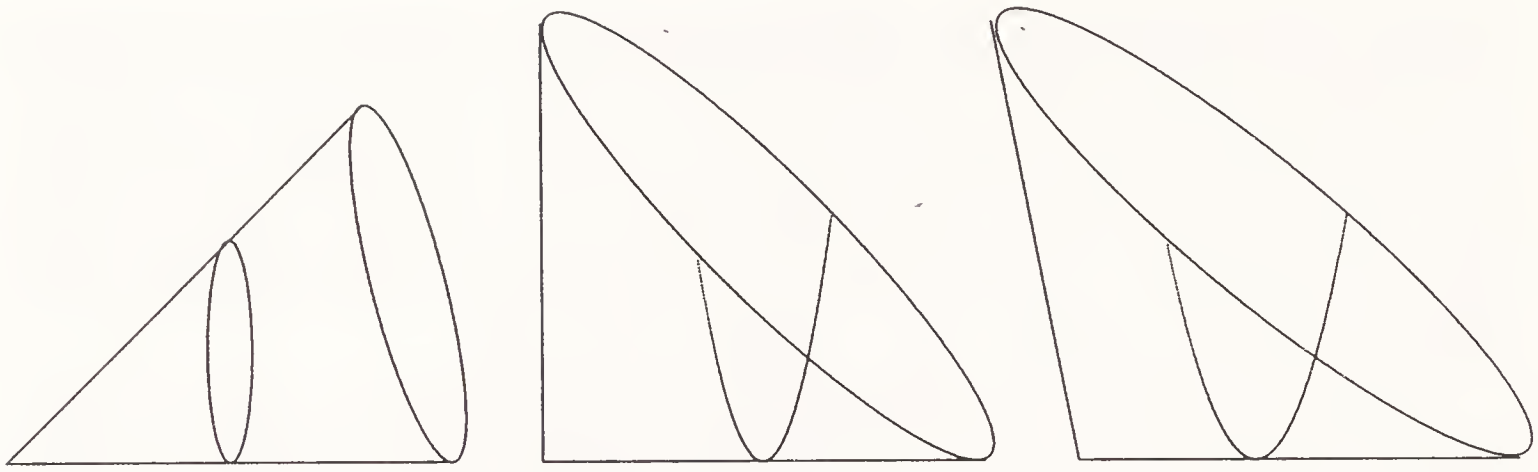


Figure 5.7: The triad of Menaechmus.

Let the two lines be a and d. Consider a pair of mutually perpendicular lines. Construct a parabola symmetric about the first line as axis and having the point of intersection of the two lines as vertex and such that the square on the line segment from each point of the parabola perpendicular to its axis equals the rectangle on a and the segment cut off on the axis. (That is, the "ordinate" is the mean proportional between the "abscissa" and the fixed line a. In our terms such a parabola would have the equation $y^2 = ax$.) Then construct a similar curve on the axis perpendicular to the axis of this parabola. (In our terms this is the parabola $x^2 = dy$ on the same set of coordinate axes.) The two parabolas intersect at a point $x = c$, $y = b$ such that $ac = b^2$ and $c^2 = bd$, that is, $a : b :: b : c :: c : d$, as required.

The properties of a parabola used here by Menaechmus are so close to the modern equation that very little distortion is involved in the simplifications introduced into the discussion above. It is therefore an interesting problem to what extent the Greek treatment of the conic sections mirrors the "analytic geometry" that is now taught. We shall look at this problem again when we discuss the main Greek treatise on the subject, due to Apollonius. Just now we are interested only in the route by which conic sections came to be of interest to the Greeks. The question we have to address is: What do mean proportionals have to do with conic sections?

The standard Greek way of producing the mean proportional between two line segments is given in Euclid, Book VI, Proposition 13. The two line segments are laid end to end in a straight line, and a circle is drawn having the combined line segment as diameter. The chord perpendicular to this diameter at the point where the two segments join is then drawn, and the mean proportional is one half of this chord (see Fig. 4.7).

This construction shows that the principle behind the construction of a mean proportional is that *the perpendicular from a point on a circle to a diameter is the mean proportional between the segments it cuts off on the diameter*. What is needed for the problem of two mean proportionals is a *family* of mean proportionals of various sizes, from which one can be chosen satisfying an additional condition. That is, we need a family of pairs of line segments x and y such that $a : x :: x : y$, from which we hope to choose a pair for which $a : x :: y : d$. From this fact

it appears that we need a family of circles of continuously varying radius. This family of circles is provided by "stacking" the circles to form a cone.

5.4.2 Archytas

Diogenes Laertius reports that Archytas, one of several eminent Greek scholars of the same name, was a contemporary of Plato and saved Plato from being put to death by Dionysius, the ruler of Syracuse. He is said to have been not only a great scholar but also a great general, never suffering a defeat. He is mentioned by many authors, including Aristotle, who wrote books (now lost) on his philosophy. He was a native of the Greek port city of Tarentum in the southeast part of Italy. Diogenes Laertius says that Archytas was a Pythagorean and that he was

> the first to bring mechanics to a system by applying mathematical principles; he also first employed mechanical motion in a geometric construction, namely when he tried by means of a section of a half-cylinder, to find two mean proportionals in order to duplicate the cube.

Plutarch says that Plato criticized Eudoxus, Archytas, and Menaechmus for considering the use of mechanical devices, which he believed were a perversion of the true purpose of geometry—to elevate the soul above the material world.

Eutocius quotes a now-lost passage from Eudemus' history of mathematics giving Archytas' solution of the problem of two mean proportionals. If this passage is an accurate report of Archytas' reasoning, one can see why Plato objected to it. For one of the points needed for the construction is located as the point of intersection of a cone with a second curve that is described only as the path of a moving point (the intersection of a cylinder with a semicircle rotating about one end of its diameter). This curve is not easily visualized or apprehended with the mind. It is not a planar curve, so that it cannot even be drawn with drafting instruments. However, if Plutarch is to be believed, Plato would have objected even if the curve could have been drawn with such instruments. He apparently accepted only reasoning that can be analyzed from fairly elementary and easily visualizable figures such as circles and straight lines.

5.5 Problems and Questions

5.5.1 Problems in Euclidean Mathematics

Exercise 5.1 Prove that the square of an odd integer is always one larger than a multiple of 8, that is, that $n^2 - 1$ is divisible by 8 if n is odd.

Exercise 5.2 Carry out the details of the reasoning using only odd-and-even principles to show that $\sqrt{11}$ is irrational. Do the same for $\sqrt{12}$, then tell how the irrationality of $\sqrt{12}$ could be proved by relying on the irrationality of $\sqrt{3}$.

Exercise 5.3 We suggested in the previous chapter that the incommensurability of the side and diagonal of a pentagon could have been discovered by applying the Euclidean algorithm and observing that it leads to a cyclic process, in which after a finite number of steps a new pair is reached having the same ratio as the original. This argument, which was given geometrically, can also be formulated numerically. For example, to prove that $\sqrt{2}$ is irrational, we can consider the pair $(\sqrt{2}, 1)$, which then yields $(1, \sqrt{2}-1)$ and then $(2-\sqrt{2}, \sqrt{2}-1)$. But this last pair has the same ratio as the original pair, as one can see by "cross-multiplying": $\sqrt{2} \times (\sqrt{2}-1) = (2-\sqrt{2}) \times 1$. Similarly, starting with $(\sqrt{3}, 1)$, three applications of the algorithm yield $(2\sqrt{3}-3, 2-\sqrt{3})$, and this is the same ratio as the original pair, as we see by cross-multiplying: $\sqrt{3} \times (2-\sqrt{3}) = 1 \times (2\sqrt{3}-3)$.

Thus in both of these cases we see that the Euclidean algorithm will never produce an equal pair. We have therefore a proof of the irrationality of the numbers in question. Construct a similar proof for $\sqrt{5}$. How long does the algorithm take to cycle in this case? Is this an argument that could have been used by the Greeks? Write a computer program to calculate the length of the cycles in the Euclidean algorithm for any integer whose square root is irrational, and compare the lengths for nonsquare integers up to 19. Does the result explain why Theodorus stopped at 17? Could he have been using this method?

Exercise 5.4 Show how to set up the problem of constructing the Golden Section as a problem involving application with excess. (It may help to phrase both the general application with excess problem and the Golden Section problem as quadratic equations.)

Exercise 5.5 You have probably been taught the parallel postulate in a different form: *Given a line l and a point P not on l there is one and only one line passing through P parallel to l.* This equivalent form of the parallel postulate is due to the Scottish mathematician John Playfair (1748–1819), who used it in his textbook on geometry. Playfair's version of the axiom is grammatically much simpler than Euclid's, and hence high-school textbooks have tended to prefer it. It is not so useful, however, in the actual proving of theorems, since a frequent use of the postulate is to show that two lines intersect. Euclid can do this by cutting the lines with a transversal and showing that the interior angles on one side are less than two right angles, whereas Playfair's version requires finding or constructing a line that intersects one of the lines and is parallel to the other. In this context and in others, Euclid's version leads to a shorter and simpler proof. Try proving, for instance, that if two parallel lines are cut by a transversal, then the alternate interior angles are equal (Euclid, Book I, Proposition 29) using first Euclid's definition, then using Playfair's. Which is easier?

Exercise 5.6 Euclid's two "hidden" postulates show up clearly in Book I, Proposition 16, in which he proves that an exterior angle to a triangle is greater than either of the opposite interior angles. Here is a sketch of the proof, based on Fig. 5.8. Tell where the two hidden postulates are used. How does the proof break down if lines are interpreted as great circles on a sphere?

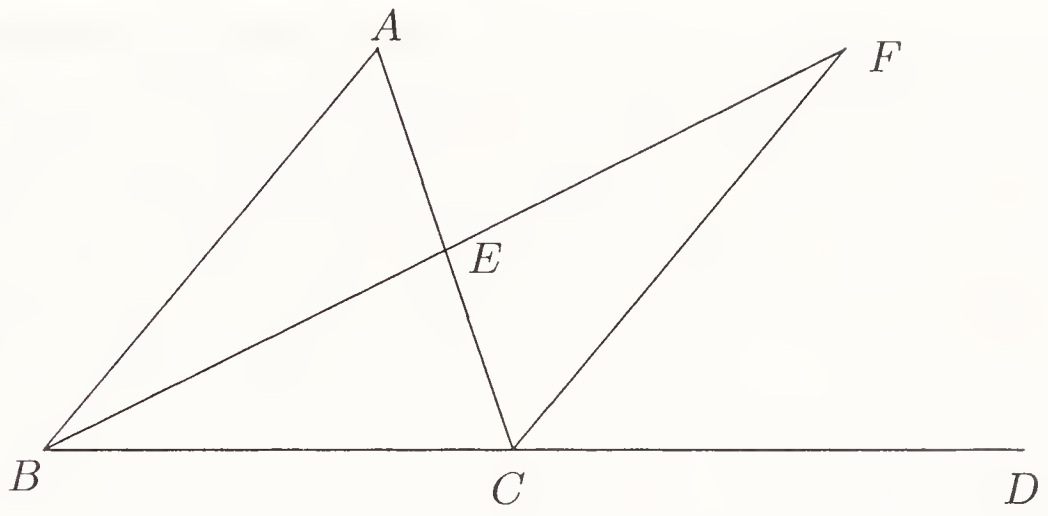


Figure 5.8: The exterior angle theorem.

Let ABC be any triangle, let side BC be extended to D, forming exterior angle ACD. We claim that $\angle ACD > \angle BAC$. For, let AC be bisected at E. Draw BE and extend it to F so that $EF = BE$, and draw CF. Then $\triangle CEF$ is congruent to $\triangle AEB$ (by side-angle-side). It follows that $\angle BAE$ equals $\angle ECF$. Thus

$$\angle BAC = \angle BAE = \angle FCE < \angle ACD.$$

Exercise 5.7 Euclid's formula for perfect numbers amounts to the statement that $2^{n-1}(2^n - 1)$ is a perfect number if $2^n - 1$ is prime. Write out the proper divisors of such a number, and prove that it must be perfect. Is your argument one that the Pythagoreans would have used?

5.5.2 Questions about Euclidean Mathematics

Exercise 5.8 Why did Euclid postpone the discussion of the problems of application with excess and defect until Book VI, when much of the "geometric algebra" needed for this topic was developed in Book II?

Exercise 5.9 The philosopher Karl Popper (1902–1994) suggested that Plato may have believed that all ratios could be expressed in terms of three kinds of line segments: (1) segments commensurable with a given unit of length, (2) the diagonal of a square having the given unit of length as its side ($\sqrt{2}$), and (3) the altitude of an equilateral triangle having the given unit length as half of its base ($\sqrt{3}$). Popper advances two reasons for this conjecture.

First, these two ratios occur constantly in Plato's mystical dialogue known as *Timaeus*. These incommensurables occur in what we commonly call the 45–45–90 right triangle (half of a square) and the 30–60–90 right triangle (half of an equilateral triangle). Their mystical use by Plato was based on the four-element cosmology of fire, air, water, and earth. These elements, as already mentioned, were identified respectively with the regular polyhedra, namely the tetrahedron, the octahedron, the icosahedron, and the cube respectively. In the *Timaeus* Plato gives a rationale for this assignment by generating these regular solids from the

two basic right triangles. He argues that every physical body must be bounded by a surface, which, if it is a regular polyhedron, can be triangulated. And, except for the dodecahedron, the faces of a regular polyhedron are either squares or triangles. Hence, if the faces are bisected, only these two triangles will occur. The idea was that, since water and air were regarded as "mean proportionals" between earth and fire the chemical processes that change one of these bodies into another should be mirrored by geometric transformations enabling one figure to be built out of the other.

Second, Plato must have been interested in finding the area of the circle. Since the inscribed octagon has area equal to $2\sqrt{2}r^2$, and the circumscribed hexagon has area $2\sqrt{3}r^2$, Plato may have believed the average of these to be the area of the circle, that is, in our terms Plato may have believed that $\pi = \sqrt{2} + \sqrt{3}$. This is in fact very close, since $\sqrt{2} + \sqrt{3} \approx 3.14626$ and $\pi \approx 3.14159$. In other words, the relative error is less than $\frac{0.005}{3} = \frac{1}{600}$.

Popper also alludes to a passage in Plato's dialogue *Greater Hippias* in which Socrates says that when two things are separately inexpressible, they may together be expressible. The word for "inexpressible" here is also Euclid's word for incommensurable: it is *arrhetos* (*'άρητος*). (The word has the same root as the word *rhetoric*.)

How plausible do you find this argument?

Exercise 5.10 In Plato's dialogue *The Laws* the participants are Cleinias (a Cretan), Megillus (a Spartan), and an unidentified Athenian visiting Crete. The three discuss several aspects of the ideal society. When they come to education (in Book Seven) the Athenian describes the Egyptian system of education in mathematics and speaks of the "deep-rooted ignorance, at once comic and shocking, that all men display in this field." The Athenian, probably speaking for Plato himself, says

> But if, as I put it, "all we Greeks" believe them to be commensurable when fundamentally they are *in*commensurable, one had better address these people as follows (blushing the while on their behalf): "Now then, most esteemed among the Greeks, isn't this one of those subjects we said it was disgraceful not to understand—not that a knowledge of the basic essentials was much to be proud of?"

The writer clearly expects his reader to be scandalized at the general level of ignorance about incommensurables. What differences between twentieth-century American culture and the culture of upper-class Greeks in the time of Plato are pointed up by this assumption on Plato's part?

Exercise 5.11 If, as seems likely, the Athenian speaking in the *Laws* represents the views of Plato himself, we may infer that *geometry*, in opening his eyes to the problem of incommensurables, has greatly enlarged his perspective rather late in life in a way that *arithmetic* alone never could have done. If such is the case, what might Plato have meant by the motto he is said to have placed above the door to his Academy: "Let no ungeometrical person come under this roof"?

Exercise 5.12 D. H. Fowler traces the history of the alleged motto of Plato's Academy in great detail. The legend seems to have begun in an oration of the Emperor Julian the Apostate in 362 C.E., who mentions an inscription over *Aristotle's* classroom and alludes to one by Plato, without saying anything about the wording of either one. The unknown author of a marginal note in a slightly later fourth-century manuscript wrote that the motto had been inscribed at the front of the school of Plato, adding that the word *ungeometrical* was a replacement for the usual formula, "Let no impure person enter," found at sacred shrines. The story of the inscription can be found in many writers from the sixth century onward, but nothing is known earlier than the fourth-century sources just mentioned. These people, of course, lived *more than 600 years after Plato.*

Seen in this context, what does the motto tell about the *emotional* significance of Plato's Academy to its pupils?

Exercise 5.13 Plato's *Theatetus* gives the twentieth-century student the opportunity to imagine what it must have been like to engage in conversation with Socrates. For most people who have had a good modern education through high-school mathematics Theatetus' classification of numbers into equilateral and oblong probably seemed pointless. After all, we *know* that all positive numbers are perfect squares, don't we? Our calculators are designed to produce an approximation to $\sqrt{2}$ at the touch of a button. Imagine yourself trying to convince Socrates of this (assuming he could speak English and knew our decimal notation for numbers). Continue the dialogue that begins below.

Modern Student. Theatetus is wrong, Socrates, *all* numbers are equilateral in the sense he defined. It is true, as Theodorus has proved, that $\sqrt{2}$ is not rational, but it is a *real* number, nevertheless?

Socrates. Really? What number is it?

Modern Student. Approximately 1.414.

Socrates. But the square of 1.414 is 1.999396, which is not 2. We already know ways of finding numbers whose squares are *approximately* 2. You claim to know a number whose square is *exactly* 2. What is this number?

Modern Student. It is the square root of 2.

Socrates. And what number is "the square root of 2"?

Modern Student. It is the unique positive real number whose square is 2.

Socrates. You are going in circles, my friend. The whole point in dispute is whether such a number exists.

Modern Student. But I know how to find the decimal expansion of this number to as many places as desired.

Socrates. Yes, but you live for only a finite length of time, and the expansion you are talking about goes on forever, so you will never know the "number" you claim to know. As I said, we Greeks *already* know how to find numbers whose squares are as close to 2 as desired. Anyway, what gives you the right to call a decimal expansion

> a number? Numbers are objects that can be added and multiplied. If you don't know the whole decimal expansion, how can you claim to calculate with it? And even if you *did* know the whole decimal expansion, how would that enable you to find the decimal expansion of the sum of two numbers? Tell me how you would find the decimal expansion of $\sqrt{2}+\sqrt{3}$, for example, if you knew the whole expansion for $\sqrt{2}$ and the whole expansion for $\sqrt{3}$.

Exercise 5.14 The Greek philosophers and mathematicians looked very suspiciously at the concept of infinity. The words we have put into Socrates' mouth in the previous exercise, objecting to infinite decimal expansions, reflect a certain attitude that one can discern behind much of their writing. Here, for instance, is what Aristotle says in his *Physics*, III. 6. 206^a9–$^b27$:

> We must keep in mind that the word "is" means either what *potentially* is or what *fully* is...
>
> But the phrase "potential existence" is ambiguous. When we speak of the potential existence of a statue we mean that there will be an actual statue. It is not so with the infinite. There will not be an actual infinite.

How do Aristotle's arguments affect the Eudoxan theory of proportion? Has Eudoxus, with his "any equimultiples whatsoever," brought the infinite into geometry? Are we dealing with an actual infinity in Aristotle's sense, when we talk about the whole set of possible multiples?

Exercise 5.15 Contrast Euclid's discussion of Eudoxus' definition of proportion, which mentions "any equimultiples whatever," with his careful statement about the extension of lines: "a line can be prolonged to any length." We nowadays teach geometry students that lines are of infinite length, but to Euclid this infinity was merely potential: any actual line was of finite length, but capable of being extended if necessary. Is Euclid being inconsistent here? Is he avoiding the actual infinity when talking about lines and allowing it when talking about proportion? Or is the latter use also only a "potential" infinity?

Exercise 5.16 Consider the claim by Proclus that the existence of incommensurables results from the existence of infinity, since if infinity did not exist, "there would be nothing inexpressible (*arrheton*) or irrational (*alogon*), features that are thought to distinguish geometry from arithmetic." The distinction between "inexpressible" magnitudes (what we call irrational square roots) and "irrational" numbers (all other irrationals, in our terms) is not pursued by Proclus. Is he justified in saying that the existence of incommensurables is due to infinity?

Exercise 5.17 Common sense seems to indicate that a logical development of a theory would have one great advantage over an informal intuitive development, namely that its conclusions would be certain, and one could therefore have much more confidence in them than in the results of vague intuitive arguments. Common

sense also seems to indicate that one would pay for this increased certainty by having to give up many appealing intuitive ideas that are too vague to be captured in a logical presentation. Are there any respects in which this common sense is the opposite of the truth? Can you think of any cases where the effect of logical development is to make conclusions seem more, rather than less, doubtful and to stimulate the creation of new intuitive possibilities rather than excluding others? You may find the history of noneuclidean geometry relevant to your answer.

Exercise 5.18 Plato apparently refers to the famous 3–4–5 right triangle in the *Republic*, 546*c*. Proclus alludes to this passage in a discussion of right triangles with commensurable sides. We can formulate the recipes Proclus attributes to Pythagoras and Plato respectively as

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

and

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

Considering that Euclid's treatise is regarded as the summation of Pythagorean mathematics, why is this topic not discussed? In which book of the *Elements* would it belong?

Exercise 5.19 Proposition 14 of Book II of Euclid shows how to construct a square equal in area to a rectangle. Since this construction is logically equivalent to constructing the mean proportional between two line segments, why does Euclid wait until Book VI, Proposition 13 to give the construction of the mean proportional?

Exercise 5.20 Why might Plato have objected to the various solutions of Archytas, Menaechmus, and others to the problem of the two mean proportionals? What aspects of his view of the world might have been in conflict with this approach to science?

5.6 Endnotes

1. All excerpts from Euclid's *Elements* are taken from the three-volume work by T. H. Heath (Dover reprint, New York, 1956).

2. The quotations from Aristotle's *Prior Analytics* are taken from the translation by Robin Smith (Hackett Publishing Co., Indianapolis, 1989), pp. 91, 93.

3. The discussion of original texts of the *Elements* is given by D. H. Fowler in *The Mathematics of Plato's Academy* (Clarendon Press, Oxford, 1987), pp. 202–220.

4. Proclus' discussion of Euclid's fifth postulate can be found in the translation of his *Commentary* by Glenn R. Morrow (Princeton University Press, 1970), pp. 150–151.

5. Ptolemy's attempted proof of the parallel postulate can be found in Proclus' *Commentary* (op. cit.), pp. 286–287.

6. Diogenes Laertius' biography of Archytas can be found in the Loeb Classical Library edition of his *Lives of the Philosophers* (Putnam, New York, 1925), Vol. 2, pp. 393–397.

7. Plutarch's report that Plato rejected the use of mechanical drawing devices can be found in the Loeb Classical Library edition of *Selections Illustrating the History of Greek Mathematics*, with a translation by Ivor Thomas (Harvard University Press, 1939), Vol. 1, pp. 387–389.

8. Popper's conjecture on Plato's mystical use of the square roots of 2 and 3 is taken from the notes at the end of volume 1 of *The Open Society and its Enemies* (Princeton University Press, 1963), pp. 251–252.

9. The quotation from Plato's *Laws* is taken from the Penguin Books edition (London, 1930), pp. 313–314.

10. The discussion of the motto of Plato's Academy is given by D. H. Fowler (op. cit.), pp. 197–202.

11. The quotation on the infinite from Aristotle's *Physics* is taken from volume 2 of *The Works of Aristotle Translated into English* by R. P. Hardie and R. K. Gaye (Clarendon Press, Oxford, 1930).

12. Proclus' discussion of right triangles with commensurable sides can be found in his *Commentary* (op. cit.), p. 428.

Chapter 6

Archimedes and Apollonius

Among the many authors of mathematical treatises in Hellenistic times there are two besides Euclid who wrote works of such lasting significance that they deserve to be looked at in some detail in an introductory history such as the present one. These two, Archimedes of Syracuse and Apollonius of Perga, form the subject of the present chapter. Both are of importance because of the profundity of their work and its influence on the subsequent development of mathematics.

6.1 Archimedes

Archimedes is one of a small number of mathematicians of antiquity of whose works we know more than a few fragments and of whose life we know more than the approximate time and place. It is ironic that the man indirectly responsible for his death, the Roman general Marcellus, is also indirectly responsible for the preservation of some of what we know about him. Archimedes lived in the Greek city of Syracuse on the island of Sicily during the third century B.C.E. and is said by Plutarch to have been a relative of King Hieron II. Since Sicily is nearly on the direct line between Carthage and Rome, it is not surprising that it became embroiled in the Second Punic War. Marcellus took the city of Syracuse after a long siege, and Archimedes was killed by a Roman soldier in the chaos of the final fall of the city. It was Plutarch's interest in Marcellus that led him to write a few lines about Archimedes.

According to Plutarch's biography of Marcellus, the general was very upset that Archimedes had been killed and had his body buried in a suitably imposing tomb. It often happens when a nation is conquered that the conquerors are insufficiently appreciative of its cultural achievements and the conquered nation is unable to preserve the relics of that culture. Such was the case with Archimedes. According to Eutocius, a biography of Archimedes was written by a certain Heracleides, who is mentioned in some of Archimedes' letters. However, no copy of this biography is known to exist today. A century after Archimedes' death his tomb had fallen into neglect. In his *Tusculan Disputations*, written in 45 B.C.E., the famous Roman

orator and statesman Cicero recalled his discovery of this tomb when was quaestor in Sicily (76 B.C.E.)

> When I was quaestor I tracked out [Archimedes'] grave, which was unknown to the Syracusans (as they totally denied its existence), and found it enclosed all round and covered with brambles and thickets; for I remembered certain doggerel lines inscribed, as I had heard, upon his tomb, which stated that a sphere along with a cylinder had been set up on the top of his grave... Slaves were sent in with sickles who cleared the ground of obstacles... So you see, one of the most famous cities of Greece... would have been ignorant of the tomb of its one most ingenious citizen, had not a man of Arpinum pointed it out.

During the Middle Ages and later Sicily was conquered many times, and the tomb of Archimedes was lost again. In popular tradition several tombs were erroneously believed to belong to Archimedes. However, the actual tomb may have been rediscovered in 1957, during an excavation.[1] Since Syracuse was taken in 212 B.C.E. and Archimedes was reported by a twelfth-century Byzantine historian named Tzetzes to have been 75 years old at the time of his death, his dates are generally given as 287–212.

There are many famous legends connected with Archimedes. These are scattered among the various sources. Plutarch, for instance, says that Archimedes made many mechanical contrivances but generally despised such work in comparison with pure mathematical thought. Plutarch also reports three different stories of the death of Archimedes and tells us that Archimedes wished to have a sphere inscribed in a cylinder carved on his tombstone. The famous story that Archimedes ran naked through the streets shouting "Eureka!" ("I've got it!") when he discovered the principle of specific gravity in the baths is reported by the Roman architect Vitruvius. The commentator Proclus gives another well-known anecdote that Archimedes built a system of pulleys that enabled him (or King Hieron) single-handedly to pull a ship through the water. Finally, Plutarch and Pappus both quote Archimedes as saying (in connection with his discovery of the principle of the lever): "Give me a place to stand and I will move the earth."

With Archimedes we encounter the first author of a considerable body of original mathematical research that has been preserved to the present day. Historians of mathematics have unanimously praised him as one of the greatest mathematical geniuses in history. Before adding our own voice to the chorus of praise, we should pause and ask what makes a mathematician great. Some criteria that one might use are the following:

---

[1]This claim was made by Prof. Salvatore Ciancio in 1965 on the basis of several criteria, including the location and date of the relics and a gold signet ring found in the crematory urn inside the tomb and bearing the ancient seal of the city of Alexandria. The famous sphere and cylinder were not part of the find. The claim was contradicted at the time by the Curator of Antiquities in Syracuse Prof. Bernabò Brea. Another counterclaim is made by D. L. Simms in "The trail for Archimedes' tomb," *Journal of the Warburg and Courtauld Institute*, **53** (1990), pp. 281–286 (reference taken from the World Wide Web). More information can be obtained at the address http://www.mcs.drexel.edu/~crorres/Archimedes/contents.html.

1. *Solving problems that others have worked on without success.* The fact that others have worked on the problems shows general agreement that the problems were important, and lack of success by others is evidence that the one who solved the problem has more than ordinary talent.

2. *Creating beautiful new theories to describe the world.* The "raw material" for such theories can come from art, physical science, social science, and mathematics itself.

3. *Discovering mathematical relations linking observations or facts that were previously thought to be unrelated.*

4. *Reorganizing and streamlining the presentation of existing theories to make their inner structure more comprehensible or more rigorous.*

5. *Suggesting lines of research leading to subsequent work of great importance.*

Thus in judging the quality of a mathematician one looks at versatility, profundity, creativity, imagination, rigor, and influence. Archimedes can be given high marks in all of these areas.

6.1.1 The Works of Archimedes

Ten of Archimedes' treatises have come down to the present, along with a "Book of Lemmas" that seems to be Archimedean, though the manuscript mentions Archimedes in the third person in some places. Some of these works are prefaced by a "cover letter" intended to explain their contents to the person to whom Archimedes sent them. These correspondents of Archimedes were: Gelon, son of Hieron II and one of the kings of Syracuse during Archimedes' life; Dositheus, a student of Archimedes' student and close friend Conon; and Eratosthenes, an astronomer who worked at the Museum in Alexandria. Like the manuscripts of Euclid, all of the Archimedean manuscripts date from the ninth century or later, usually much later. These manuscripts have been translated into English and published by various authors. A complete set of Medieval manuscripts of Archimedes' work has been published by Marshall Clagett in the University of Wisconsin series on Medieval Science.

The 10 treatises referred to above are

1. *On the Equilibrium of Planes*, Part I.

2. *Quadrature of the Parabola.*

3. *On the Equilibrium of Planes*, Part II.

4. *On the Sphere and the Cylinder*, Parts I and II.

5. *On Spirals.*

6. *On Conoids and Spheroids.*

7. *On Floating Bodies.*
8. *Measurement of a Circle.*
9. *The Sand-reckoner.*
10. *The Method.*

References by Archimedes himself and other mathematicians tell of the existence of other works by Archimedes of which no manuscripts are now known to exist. These include many works on the theory of balances and levers, optics, the regular polyhedra, the calendar, and the construction of mechanical representations of the motion of heavenly bodies.

From this list we can see the versatility of Archimedes. His treatises on the equilibrium of planes and floating bodies contain principles that are now fundamental in mechanics and hydrostatics. The works on the quadrature of the parabola, conoids and spheroids, the measurement of the circle, and the sphere and cylinder extend the theory of proportion, area, and volume found in Euclid for polyhedra and polygons to the more complicated figures bounded by curved lines and surfaces. The work on spirals introduces an entirely new class of curves, and develops the theory of length, area, and proportion for them.

Archimedes' contributions to number theory are less impressive. Except for a single Diophantine problem[2] known as the "cattle problem," his only number-theoretic work is *The Sand-reckoner*, which constructs a systematic hierarchy of numbers so as to be able to express compactly and accurately numbers of any conceivable size. Finally, the *Method* shows an entirely different approach to geometry based on what we now call infinitesimal considerations.

The profundity of Archimedes' thought, which has given the world theorems that might otherwise have remained undiscovered, will appear in the discussion of the details of some of these works. We shall reserve the "physical" works on equilibrium of planes and floating bodies until the next chapter, where we shall give a general discussion of the mathematical science of Archimedes' time. In the present chapter we confine ourselves to his pure mathematics.

We shall discuss Archimedes' geometry in increasing order of its complexity, starting with the work that seems to have been most influential in the Medieval West, the *Measurement of a Circle*, then taking up the *Quadrature of the Parabola*, *The Sphere and Cylinder*, and *On Conoids and Spheroids*. We shall then turn to some works that involve different themes, *The Sand-reckoner*, *On Spirals*, and the *Method*.

The Measurement of a Circle

The treatise on the measurement of a circle is a brief one, containing only three formally stated propositions. The first proposition, demonstrated in strict Euclidean style, is that a circle equals (in area) a right triangle with one leg equal to the

[2]A Diophantine problem is an equation in two or more unknowns for which integer solutions are to be found.

radius of the circle and the other equal to the circumference. The method of proof is exactly the method used by Euclid to prove that circles are proportional to the squares on their diameters (see the previous chapter). That is, it uses the method of exhaustion, and the trichotomy for ratios (given two ratios, either they are equal or one is larger than the other).

Proposition 1 shows that the problem of measuring the area of a circle reduces to finding the ratio of the circumference to the diameter, the number we now call π (Archimedes did not have any symbol for this number). An approximation to this number is the content of the main proposition of the treatise, Proposition 3: *Every circle exceeds three times its diameter by an amount less than one-seventh and more than* 10 *parts of* 71 *parts of the diameter.*

In our language this proposition says $3\frac{10}{71} < \pi < 3\frac{1}{7}$. The proof seems natural enough, being based on comparison of chords and tangents to very small arcs obtained by three successive bisections of a $30°$ angle. Since the procedure for finding the chord of half an arc involves square roots, Archimedes had to get careful bounds on these square roots at each stage in order to avoid error accumulation. In the end he arrived at the result

$$\pi < \frac{14,688}{4673\frac{1}{2}} = 3 + \frac{667\frac{1}{2}}{4673\frac{1}{2}} < 3\frac{1}{7}.$$

By following the same procedure, using angles with their vertex at one end of the diameter, starting with a $30°$ angle, and using the estimate $\sqrt{3} < 1351 : 780$, Archimedes found an estimate for the perimeter of the inscribed regular polygon of 96 sides, leading to the estimate $\pi > 3\frac{10}{71}$.

The approximations used for the quadrature of a circle are sometimes used as a measure of the mathematical sophistication of a culture. This criterion is naive as it stands: the numbers used for approximating and measuring depend to some extent on the base used for counting, and in any case one *might* stumble onto a very accurate value of π, say $\frac{355}{113} \approx 3.14159292$, from irrelevant aesthetic considerations or by merely guessing. A more refined version of the criterion is the *method* used for approximating π and the extent to which the approximation error can be *proved* to be small. We have seen in Chapter 2 that the Egyptians approximated the area of a circle by a method that implies $\pi = \frac{256}{81}$. This value may have been discovered by visual examination of a circle intersecting a certain square, as discussed in Chapter 2. It amounts to the equation $\pi = 3.160493827$. It is much closer than the value implied by the following passage from the Bible in I Kings (or III Kings, depending on the translation), 7:23.

> And [Solomon] made a molten sea ten cubits from the one brim to the other: it was round all about... and a line of thirty cubits did compass it round about.

The implied value $\pi = 3$ here is also found in many Babylonian tablets, although the value $3;7,30$ $(= 3.125)$ is also found in the tablets. (Some commentators claim the 10-cubit diameter refers to the outside of the rim and the 30-cubit circumference to the inside rim.) In comparison with these values Archimedes' result is much

more sophisticated, since he gives both a lower and an upper bound, which in our terms are as follows:

$$3.14084507042 < \pi < 3.14285714286.$$

What is more important than accuracy, however, is that Archimedes, building on the theory of proportion developed in Euclid, can *prove* that π lies between these two values, rather than merely *asserting* it, as is done in many other sources. (Of course, the Egyptians and Babylonians may also have arrived at their approximations through sophisticated upper and lower estimates that have been lost.) If we take the average of Archimedes' two values (3.14183582289), we obtain a value that exceeds π by only 0.008%.

Quadrature of the Parabola

In view of the lack of success the Greeks had had in squaring the circle one can appreciate that success in squaring a segment of any conic section would be considered a great achievement. Archimedes himself puts the quadrature of the parabola in this context in an accompanying letter to Dositheus. In the letter Archimedes expresses his grief over the recent death of his friend Conon, whose mathematical abilities he praises highly. He mentions various attempts at the quadrature of circles and segments of circles, saying that these efforts assumed things that could not be granted. He then announces his own success in giving a rigorous quadrature of a segment of a parabola.

This letter gives us a glimpse of Archimedes' personality. He must have been close to Conon, both personally and professionally.[3] This letter also shows why Archimedes prized this result: it was the first quadrature of a conic section and contrasted starkly with the failure to square the circle. One lemma that Archimedes used in this work asserts that, given two areas, some multiple of the first exceeds the second. This simple proposition, which denies the existence of infinitesimal areas, fits well into the modern rigorous expositions of analysis. It is reflected in the concept of Archimedean ordered fields.[4]

Archimedes' gives two proofs of his quadrature, one based on mechanical considerations of balance, the other on the method of exhaustion. In the latter approach Archimedes inscribed a maximal triangle in the segment and then did the same in each of the two smaller parabolic segments created by two sides of this triangle. He found that the two smaller triangles were each one-eighth as large as the original triangle. Hence, by repeating this operation, Archimedes was led to approximate the area by the sum of a finite geometric progression, which he could prove rigorously to be tending to $\frac{4}{3}$ of the original inscribed triangle.

---

[3] Apollonius also mentions a person by the name of Conon, probably the same person; Apollonius finds it necessary to defend Conon's mathematical prowess against a critic named Thrasydaeus.

[4] A *field* can be described informally as a structure on which addition, subtraction, multiplication, and division (except by zero) are defined and have the usual arithmetic properties. An ordered field, such as the rational numbers or the real numbers, is *Archimedean* if for any element a of the field there is a positive integer n with $n > a$.

The Sphere and the Cylinder

Archimedes' two works on the sphere and cylinder were also sent to Dositheus. In the letter accompanying the first of these he mentions the work on the quadrature of a parabola, so that the chronology of these works is completely established. Archimedes again shows his human side in this letter. After stating the most important of its contents (that the area and volume of a closed cylinder circumscribed about a sphere are each half again as large as the area and volume respectively of the sphere), he gives a little of the history of the problem, crediting Eudoxus with being the first to establish rigorously the volume of a pyramid. Archimedes considered his results on the sphere to be rigorously established, but he did have one regret:

> They ought to have been published while Conon was still alive, for I should conceive that he would best have been able to grasp them and to pronounce upon them the appropriate verdict; but, as I judge it well to communicate them to those who are conversant with mathematics, I send them to you with the proofs written out, which it will be open to mathematicians to examine.

The fact that a pyramid is one-third of a prism on the same base and altitude is Proposition 7 of Book XII of Euclid's *Elements*. Thus Archimedes could say confidently that this theorem was well established. He seems to suggest that there were other results on volumes that were *not* well established. Archimedes approached the surface area of a sphere by finding the lateral surface area of a frustum of a cone and the lateral area of a right cylinder. In our terms the area of a frustum of a cone with upper radius r, lower radius R, and side of slant height h is $\pi h(R+r)$. Archimedes phrased this fact by saying that the area is that of a circle whose radius is the mean proportional between the slant height and the sum of the two radii [that is, the radius is $\sqrt{h(R+r)}$]. Likewise our formula for the lateral surface area of a cylinder of radius r and height h is $2\pi rh$. Archimedes said it was the area of a circle whose radius is the mean proportional between the diameter and height of the cylinder.

These results can be applied to the figures generated by revolving a circle about a diameter with certain chords drawn. Archimedes showed (Proposition 22) that

$$(BB' + CC' + \cdots + KK' + LM) : AM = A'B : BA$$

in Fig. 6.1. This result is easily derived by connecting B' to C, C' to K, and K' to L and considering the ratios of the legs of the resulting similar triangles. These ratios can be added. All that then remains is to cross-multiply this proportion and use the expressions already derived for the area of a frustum of a cone. One finds easily that the area of the surface obtained by revolving the broken line $ABCKL$ about the axis AA' is $\pi AM \cdot A'B$. The method of exhaustion then shows that the product $AM \cdot A'B$ can be made arbitrarily close to the square of AA'; it therefore gives the following result (Proposition 33): *The surface of any sphere is equal to four times the greatest circle in it.*

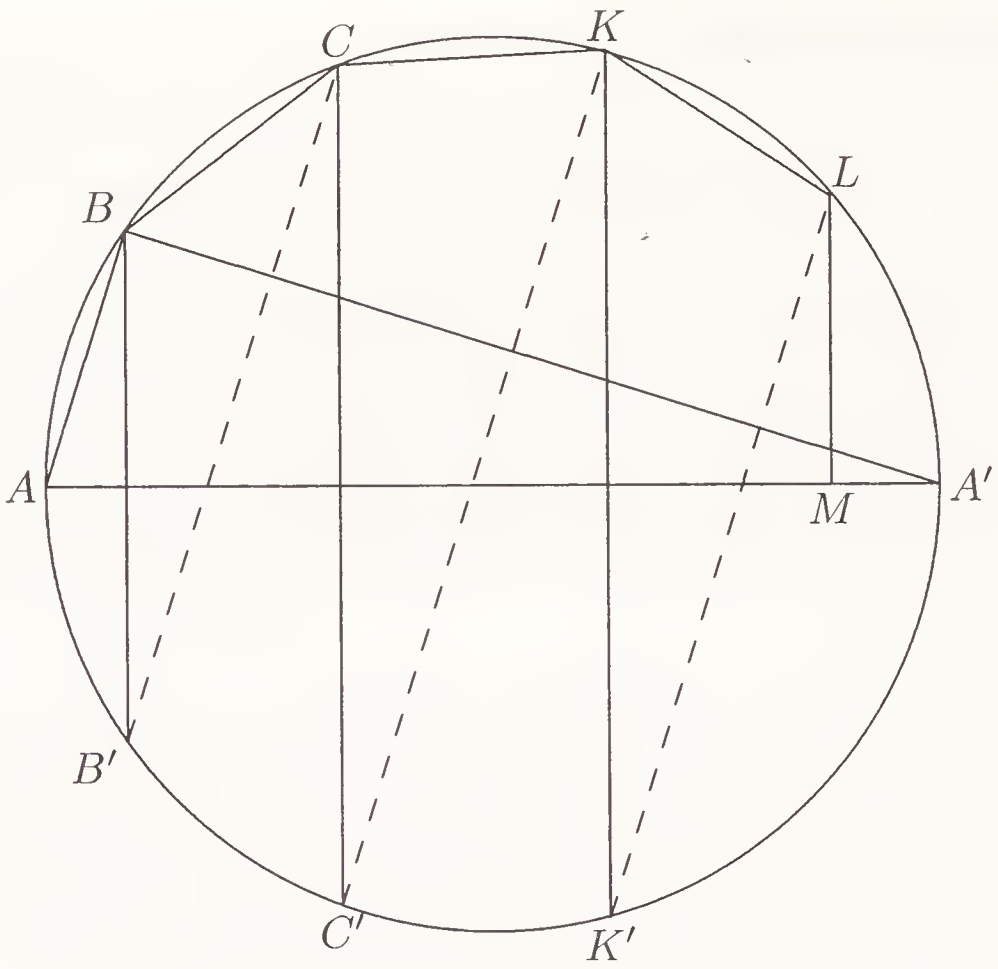


Figure 6.1: Finding the surface area of a sphere.

By the same method, using the inscribed right circular cone with the equatorial circle of the sphere as a base, Archimedes shows that the volume of the sphere is four times the volume of this cone. He then obtains the relations between the areas and volumes of the sphere and circumscribed closed cylinder. He finishes this first treatise with results on the area and volume of a segment of a sphere, that is, the portion of a sphere cut off by a plane. This argument is the only ancient proof of the area and volume of a sphere that meets Euclidean standards of rigor.

Three remarks should be made on this proof. First, in view of the failure of efforts to square the circle, it seems that the later Greek mathematicians had two standard areas, the circle and the square. Archimedes expressed the area of a sphere in terms of the area of a circle. Second, the volume of a sphere was found in other places, notably China (several centuries after Archimedes' time), but the justification for it always involved intuitive principles such as "Cavalieri's principle" that do not meet Euclidean standards. Third, Archimedes did not *discover* this theorem by Euclidean methods. He told how he came to discover it in his *Method*, which will be discussed below.

In his second treatise on the subject Archimedes attacked two sophisticated problems of solid geometry:

1. *Given a cone or cylinder, to construct a sphere of the same volume.* Archimedes points out that this problem could be solved if, given a cylinder, one could construct a cylinder of equal volume whose height equals the diameter of its base. (Such a cylinder could be circumscribed about a sphere, whose volume would then need to be increased by half. In this way the problem would reduce to the problem of two mean proportionals.) Thus the

problem of transforming the cylinder to the standard shape is also equivalent to duplicating the cube. Thus, in the typical fashion of mathematicians, he showed that one problem can be reduced to another problem that has already been studied (but not completely solved). In view of the failure to duplicate the cube and square the circle, it seems that Archimedes had to settle for two standards of volume, the cube and the sphere. The problem discussed here represents an attempt to express the volumes of certain familiar shapes in terms of one of these standards.

2. *Cut a given sphere by a plane so that the surfaces or volumes of the resulting segments have a prescribed ratio.* Archimedes remarks that in this form the problem requires an investigation to determine which data will give a problem for which a solution exists. This aspect of a problem, a general analysis to determine which data will yield a solution, was so important that Apollonius and the later commentators used a special Greek word to refer to it. That word is *diorismos* (διορισμός), from a root meaning *to divide* or *to distinguish*; we shall simply borrow this word wherever it occurs from now on. The manuscripts of Archimedes' work do not contain his solution, and Eutocius quotes other mathematicians to the effect that Archimedes did not fulfill his promise to give the solution of this last problem. Eutocius then tells of his own very interesting research, in which he hunted around through some old books and found a manuscript that seemed from its mathematical and linguistic style to be a work of Archimedes. In that work he found the solution, albeit in a very obscure form, which he was able to straighten it out with great difficulty. Eutocius was only one of several mathematicians who subsequently worked on the problem of dividing a sphere in a fixed ratio. Thus in advancing this problem Archimedes was opening up new avenues of research.

On Conoids and Spheroids

Archimedes also extended the range of geometry by studying the solids formed by revolving a conic section about its axis of symmetry. For the figures generated by revolving unbounded conics (parabolas and hyperbolas) he used the term *conoids*, while for ellipses he spoke of *spheroids*. He investigated the segments of such figures cut off by various planes. We shall not discuss this work in detail, but it will be mentioned in the next chapter in connection with Archimedes' contributions to physics.

On Spirals

The work *On Spirals* was sent to Dositheus as a follow-up report on some earlier theorems that Archimedes had sent to Conon. In the cover letter Archimedes again laments the loss of Conon, whose work on these problems was cut short by his death.

> For I know well that it was no common ability that he brought to bear on mathematics, and that his industry was extraordinary. But, though many years have elapsed since Conon's death, I do not find that any one of the problems has been stirred by a single person.

Archimedes' work on spirals is not directly connected with his other geometric work, which was a natural extension of the core of Greek geometry. The reason for Archimedes' interest in this problem is therefore a subject for conjecture. The most obvious explanation is that the spirals considered by Archimedes, which are generated by a point moving at constant speed along a ray rotating at constant angular velocity, make it possible to draw a straight line equal in length to the circumference of a given circle and to divide any angle into any number of equal parts. The spiral therefore solves two of the classical problems in the sense that, if one could draw the spiral, it would be possible to perform these constructions. Archimedes does not directly mention this connection. However, the first proposition after the preliminary lemmas on lines in arithmetic progression (Proposition 12) states explicitly that the radii drawn at equal angles to one another will be in arithmetical progression, and this is the essential property needed for trisecting an angle.

This paper is noteworthy as one of the places, along with the *Quadrature of the Parabola*, where the tangent to a curve plays an important role. The tangent to the end of the first turn of the spiral contains the hypotenuse of a right triangle having the line from the beginning of the spiral to the end of the first turn as one leg (see Fig. 6.2). By delicate use of inequalities Archimedes was able to show that the other leg of this right triangle is equal to the circumference of the circle whose radius is the first leg. The area of this triangle is therefore exactly equal to that of the circle in question. Hence it would be possible to square a circle if one could do the following two things: (1) draw a spiral from the center of the circle whose first revolution ends on the circle, and (2) draw the tangent to the spiral at the end of the first revolution. This is the first place in Greek mathematics where the tangent to a curve other than a conic section is of importance; and, as noted, it raises the problem of how such a tangent is to be drawn. (The tangent to a conic section can be drawn with straightedge and compass if the section itself has been drawn.) Even more significant in retrospect is that it connects the problem of computing length and area with the problem of constructing a tangent. In modern terms we would say it connects integration and differentiation, though these concepts are in the future, and it does not appear that Archimedes had any idea that the area problem and the tangent problem are connected.

To find the tangent to a spiral, however, is difficult. Archimedes recognized that the problem was equivalent to squaring the circle. Nowadays, of course, we use analytic methods (the derivative) to draw the tangent to any "reasonable" curve. Archimedes seems to sense the need for some tangent-drawing technique. Perhaps that is why his prefatory note laments the loss of Conon. Mathematicians like to see their ideas extended by younger minds into new areas which the limitations of their own background prevents them from creating. Archimedes must have believed that Conon could have advanced this topic.

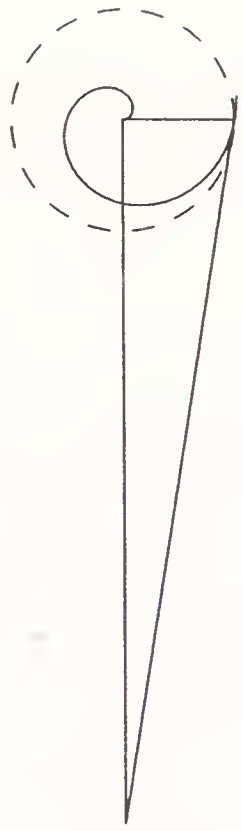


Figure 6.2: The first turn of an Archimedean spiral.

The *Sand-reckoner*

This treatise, whose contents are very different from Archimedes' other works, was apparently inspired by a specific question, as Archimedes explained in his cover letter to King Gelon: *Is the number of grains of sand required to fill up the Universe finite or infinite?* Archimedes' cover letter is of interest, not only as a preface to the mathematical work but also because of incidental remarks that reveal some of the ideas then current in astronomy. Archimedes states the geocentric cosmology as the accepted one, but mentions that Aristarchus had proposed a heliocentric model. In the geocentric cosmology, the entire Universe is the interior of the celestial sphere, to which the stars are rigidly attached. Archimedes criticizes Aristarchus for sloppy language in saying that the celestial sphere was so large that its ratio to that of the earth's orbit is that of the surface of a sphere to its center. Archimedes quite properly notes that a surface cannot have any ratio to a point. After correcting Aristarchus, he then proceeds to demonstrate that the number of grains of sand required to fill up the celestial sphere is finite. The machinery he develops to solve this problem is a cumbersome system of successive powers. While Archimedes' system of enumeration seems tiresome and hardly worth looking at nowadays, it was inspired by a real problem of stretching the human imagination. How small can a number be before it is *perceived* as zero? How large can it be before it is *perceived* as infinity? Archimedes' introduction to this work shows that some people were conflating the concepts of very large and infinite. One instance of this confusion is the view attributed to Aristarchus that the ratio of the orbit of the earth to the sphere of fixed stars is that of the center of a sphere to its surface. This view would not shock scientists who use mathematics only formally and nonrigorously, but Archimedes remarks very logically on the absurdity of the statement if taken literally. Yet we now know that the ratio is so small as to be indistinguishable from zero for practical purposes, and the

astronomer Ptolemy makes a remark to this effect in Book I, Chapter 6 of the *Almagest*: "The earth has to the senses the ratio of a point to the distances of the sphere of fixed stars."

The *Method*

Early in the twentieth century the historian of mathematics J. L. Heiberg, reading in a bibliographical journal of 1899 the account of the discovery of a tenth-century manuscript with mathematical content, deduced from a few quotations that the manuscript was a copy of a work of Archimedes. In 1906 and 1908 he journeyed to Constantinople and established the text, as far as was possible. Attempts had been made to wash off the mathematical text during the Middle Ages so that the parchment could be used to write a book of prayers. The 177 pages of this manuscript contain nearly complete texts of most of the works just discussed and a work called *Method*. The existence of such a work had been known because of the writings of commentators on Archimedes. There are quotations from it in a work of the mathematician Heron called the *Metrica* (which, however, was not discovered until 1903).

The *Method* was sent to the astronomer Eratosthenes as a follow-up to a previous letter that had contained the statements of two theorems without proofs and a challenge to discover the proofs. (This kind of game, in which a mathematician announces a theorem without giving its proof, has occurred often in the history of mathematics.) Both of the theorems involve the volume and surface of solids. In contrast to his other work on this subject, however, Archimedes here makes free use of a principle similar to one that was discovered independently in several places, particularly in China (some centuries after Archimedes' time) and in Italy during the sixteenth century. This principle, now commonly known as *Cavalieri's principle*, says that if two solids have equal cross sections at every height, then they have equal volumes. The implied reasoning is that in some sense a volume is the "sum" of its cross sections, and therefore (since "sums" of equals are equal) volumes having equal cross sections must be equal. Archimedes' *Method* is a refinement of this principle, obtained by imagining the sections balanced about a fulcrum. The reasoning is that each pair of corresponding sections will balance, and therefore the two bodies will balance.

The volume of a sphere is four times the volume of the cone with base equal to a great circle of the sphere and height equal to its radius, and the cylinder with base equal to a great circle of the sphere and height equal to the diameter is half again as large as the sphere.

Archimedes' proof is based on Fig. 6.3. If this figure is revolved about the line CAH, the circle with center at K generates a sphere, the triangle AEF generates a cone, the rectangle $LGFE$ generates a cylinder, and each horizontal line such as MN generates a disk. The point A is the midpoint of CH. Archimedes shows that the area of the disk generated by revolving QR plus the area of the disk generated by revolving OP has the same ratio to the area of the disk generated by revolving MN that AS has to AH. It follows from his work on the equilibrium of planes (discussed in the next chapter) that if the first two of these disks are hung at H,

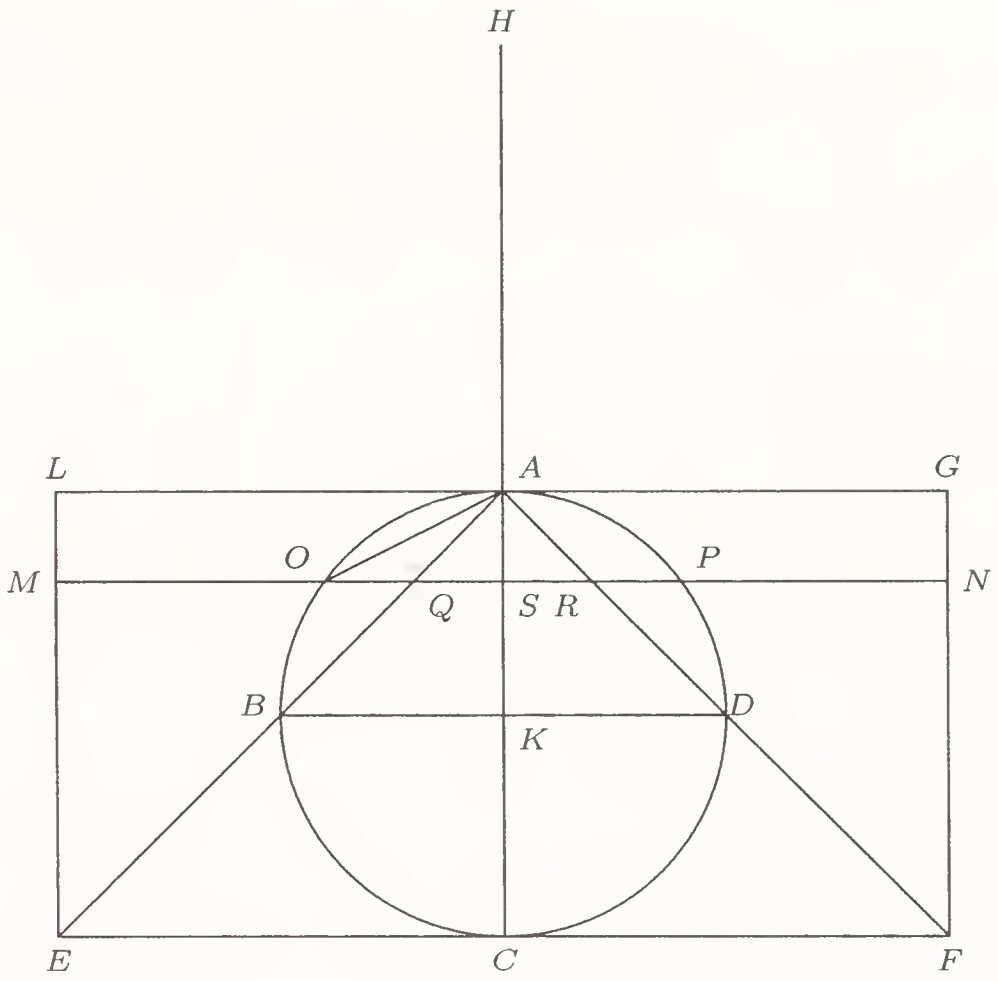


Figure 6.3: Volumes of sphere, cone, and cylinder

they will balance the third disk about A as a fulcrum. Archimedes concluded from this that the sphere and cone together placed with their centers of gravity at H would balance (about the point A) the cylinder, whose center of gravity is at K.

Therefore,

$$HA : AK = (\text{cylinder}) : (\text{sphere} + \text{cone}).$$

But $HA = 2AK$. Therefore the cylinder equals twice the sum of the sphere and the cone AEF. And, since it is well known that the cylinder is three times the cone AEF, it follows that the cone AEF is twice the sphere.

But, since $EF = 2BD$, cone AEF is eight times cone ABD, and so the sphere is four times the cone ABD.

From this fact Archimedes easily deduces the famous result allegedly depicted on his tombstone: *The cylinder circumscribed about a sphere equals the volume of the sphere plus the volume of a right circular cone inscribed in the cylinder.*

Having concluded the demonstration, Archimedes reveals that it was this method that enabled him to discover the area of a sphere. He writes

> ...judging from the fact that any circle is equal to a triangle with base equal to the circumference and height equal to the radius of the circle, I apprehended that in like manner any sphere is equal to a cone with base equal to the surface of the sphere and height equal to the radius.

The importance of the *Method* for understanding the history of Greek mathematics cannot be overestimated. Works written in the style of Euclid are like magnificent cathedrals from which the scaffolding has been removed. One can say

that obviously the stones were not simply thrown up the sides so as to land one on top of another, but the actual route by which they arrived is no longer visible. The method of exhaustion is convincing as a method of proving a theorem, but useless as a method of discovering it. With his pleasant openness Archimedes makes no attempt to conceal the routes by which he discovered his results.

6.2 Apollonius

From what we have already seen of Greek geometry we can understand how the study of the conic sections came to seem important. From commentators like Pappus we know of treatises on the subject by Aristaeus, a contemporary of Euclid who is said to have written a book on *Solid Loci*, and by Euclid himself. We have also just seen that Archimedes devoted a great deal of attention to the conic sections, usually referring to them as sections of an acute-angled, right-angled, or obtuse-angled cone. The only treatise on the subject that has survived, however, is that of Apollonius, and even for this work, unfortunately, no faithful translation into English exists. The version most accessible is that of Heath, who says in his preface that writing his translation involved "the substitution of a new and uniform notation, the condensation of some propositions, the combination of two or more into one, some slight re-arrangements of order for the purpose of bringing together kindred propositions in cases where their separation was rather a matter of accident than indicative of design, and so on." He might also have mentioned that he supplemented Apollonius' purely synthetic methods with analytic arguments, based on the algebraic notation we are familiar with. All this labor has no doubt made Apollonius more readable. On the other hand, Apollonius' work is no longer of any value to research mathematicians, and from the historian's point of view this kind of tinkering with the text only makes it harder to place the work in proper perspective.

6.2.1 Biography of Apollonius

In contrast to his older contemporary Archimedes, Apollonius remains a rather obscure figure. His dates can be determined from the commentary written on the *Conics* by Eutocius. Eutocius identifies Apollonius as a contemporary of the king Ptolemy Euergetes and defends him against a charge by Archimedes' biographer Heracleides that Apollonius plagiarized results of Archimedes. Eutocius' information places Apollonius reliably in the second half of the third century B.C.E., perhaps a generation or so younger than Archimedes.

Pappus says that Apollonius studied at Alexandria as a young man and made there the acquaintance of a certain Eudemus. It is probably this Eudemus to whom Apollonius addresses himself in the preface to Book I of his treatise. From Apollonius' own words we know that he had been in Alexandria and in Perga, which had a library that rivalled the one in Alexandria. Eutocius reports an earlier writer, Geminus by name, as saying that Apollonius was called "the great geometer"

by his contemporaries. He was indeed highly esteemed as a mathematician by later mathematicians, as the quotations from his works by Ptolemy and Pappus attest. In Book XII of the *Almagest* Ptolemy attributes to Apollonius a geometric construction for locating the point at which a planet begins to undergo retrograde motion. From these later mathematicians we know the names of several works by Apollonius and have some idea of their contents. However, only two of his works survive to this day, and for them we are indebted to the Islamic mathematicians who continued to work on the problems that Apollonius considered important. Our present knowledge of Apollonius' *Cutting off of a Ratio*, which contains geometric problems solvable by the methods of application with defect and excess is entirely based on an Arabic manuscript, no Greek manuscripts having survived. Of the eight books of Apollonius' *Conics*, only seven have survived in Arabic, and only four in Greek. None of Apollonius' other works have survived at all.

6.2.2 History of the *Conics*

The genesis of the *Conics* was reported by Pappus (five centuries after they were written). Pappus claims that Apollonius' work completed four books written by Euclid on the subject. Although he gives Apollonius fair credit for advancing the subject considerably, he makes a number of unkind remarks as to his character, accusing him of being a braggart and concealing his debt to Euclid. We, of course, are not able to sort out these internal quarrels of the ancients, and we shall not attempt to do so. As already mentioned, the first four books of Apollonius' *Conics* survived in Greek, and seven of the eight books have survived in Arabic; the astronomer Edmund Halley (1656–1743) published a Latin edition in 1710.

6.2.3 Contents of the *Conics*

Since the conic sections represent the first extension of the Euclidean theory of lines and circles to more general curves, we should expect Apollonius to discover and prove relations for them analogous to those proved for the circle in Books III and IV of the *Elements*, that is, to discuss the proportions that exist among chords and tangents of such curves. We do indeed find this, and much more besides, in the treatise. The subject is much more complicated than the theory of circles, however, since the hyperbola and parabola are unbounded curves and the hyperbola consists of two branches. Even the ellipse lacks the homogeneity of a circle, having curvature that varies from one point to another, so that the points of least and greatest curvature are of special importance.

To understand Greek mathematics the modern student must try to appreciate facts that hardly seem worthy of note nowadays, but were formidable obstacles to the Greeks. We have grown so used to "nonconstructive" mathematics and the easy application of analytic geometry that we are quite ready to accept as a plane curve any equation in two variables, even if we cannot easily compute even one point on it, for example

$$y^5 + 3x^2y^2 + 4x^7 - 2x - 8y - 1 = 0.$$

It is easy to verify that there is a point on this curve lying on the y axis somewhere between $y = 1$ and $y = 2$ and a point on the x axis somewhere between $x = 0$ and $x = 1$. This information is sufficient to convince most students that the equation defines a curve, even though we cannot begin to draw it. If we want to know more, we resort to numerical methods or computer graphics. The situation for Apollonius was very different. Ignoring for the moment the absence of algebraic notation in his time, one can substitute for the equation a *condition* and ask which points satisfy this condition. This is exactly what Apollonius is doing in the "three-line locus" and "four-line locus" problems. In some cases it may be possible to discover all the points that satisfy a condition. Those points form the *locus* (the Latin word for *place*) of the condition. *It was only through locus problems that new curves could be introduced.* Archimedes, to be sure, stretched the use of the locus method in his treatise on spirals by introducing motion into the definition, as did Apollonius (unnecessarily) in his definition of a cone, but with that exception, new curves were introduced by the Greeks only as loci, which were a *static* concept. The use of loci, rather than motion, as the device for defining curves restricted the kinds of curves that could be studied. This restriction was not overcome until the invention of analytic geometry provided an unlimited supply through equations in two variables.

The use of loci provides a basis for the study of the curves that can be defined in this way, but it also entails difficulties. Which verbal conditions will lead to a real locus? Given a condition that must be satisfied, such as those mentioned by Pappus, how can we tell whether there will be any points on the locus? This question brings up the whole circle of ideas connected with *diorismos*. It seems clear that this way of proceeding is going to impose a severe restriction on the number of curves we can consider, and some historians attribute the eventual withering of Greek mathematics in part to the shortage of useful curves.

In a preface addressed to the aforementioned Eudemus Apollonius lists the important results of his work: the description of the sections, the properties of the figures relating to their diameters, axes, and asymptotes, things necessary for *diorismos*, and the three- and four-line locus. He continues

> The third book contains many remarkable theorems of use for the construction of solid loci and for *diorismos*, of which the greatest part and the most beautiful are new. And when we had grasped these, we knew that the three-line and four-line locus had not been constructed by Euclid, but only a chance part of it and that not very happily. For it was not possible for this construction to be completed without the additional things found by us.

Since it is not feasible to look at the details of the entire work, we shall focus on just three aspects of it. First we shall consider Apollonius' construction of the conic sections and the way in which he defined them (Book I). Second, we shall look at the elementary properties of their axes and diameters, as he calls them, from Book II. Finally we shall examine some of the "remarkable" theorems from Book III, in particular the focal properties of the central conics and the problem known as the three-line locus, which Apollonius mentions in the passage just quoted.

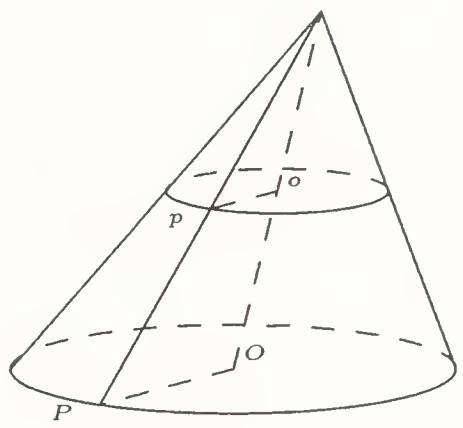


Figure 6.4: Section of a cone parallel to the base.

Definition of the Conics as Loci

The first task Apollonius faced was that of defining the conic sections as loci. As we know, these curves had previously been defined as the intersection of a cone with a plane, and such of their properties as were known had been deduced from that definition. Defining the curves as plane loci would eliminate the need for the cone and all the three-dimensional paraphernalia it entailed. Having a locus condition as the definition of the conic, one could then give an intrinsic discussion of the conic section, independently of the cone and plane that generated it, and thus one could analyze such a curve as fully as Euclid had analyzed the circle. This extension of geometry to new curves was not achieved easily. The mere description of these curves as loci is complicated.

Apollonius defined a cone as the figure formed by a family of straight line segments, all passing through a common endpoint (called the *apex* of the cone) and also intersecting a fixed (generating) circle in a plane not containing the apex. The line through the apex and the center of the generating circle forms the *axis* of the cone. A plane containing the axis of the cone intersects the portion of the cone between the apex and the generating circle in a triangle called an *axial triangle* having as base a diameter of the circle and as opposite vertex the apex of the cone. Apollonius' first important result is not surprising (Book I, Proposition 4): *All sections of a cone parallel to its base are circles*. This is easily proved (see Fig. 6.4) by the fact that the lines op and OP from the axis to a point on the intersections of the cone with the two planes remain parallel as the point P traverses the circle in the base. Hence the ratio $op : OP$ remains constant, and since OP remains constant, so must op. The next proposition (Book I, Proposition 5), however, *is* surprising. In Fig. 6.5 the cone is cut by a plane that is perpendicular to the plane of the axial triangle, though its intersection HK with the axial triangle is not parallel to the base BC. Even so, it forms a triangle AHK having the same angles as ABC, that is, $\angle AHK = \angle ACB$ and $\angle AKH = \angle ABC$. The resulting section of the cone is said to be *subcontrary*, and *it is also a circle!* For from any point P on the section one can drop a perpendicular to the plane of the axial triangle (parallel to the base), meeting the axial triangle in a point M. Through M one can draw a plane parallel to the base (hence containing the point P), and this plane will meet the cone in a circle and the axial triangle in a line DE parallel to BC. Then,

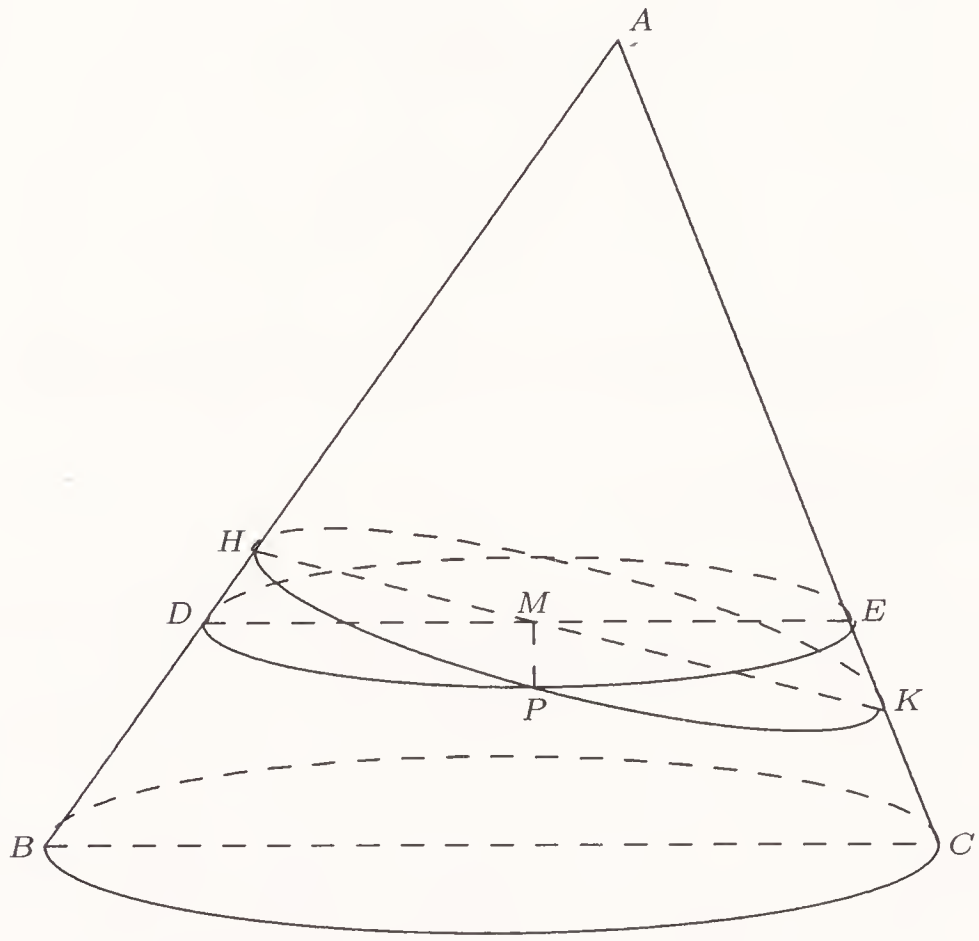


Figure 6.5: Subcontrary section of a cone.

because triangles DMH and KME are similar, $DM \cdot ME = HM \cdot MK$. But since the curve DPE is a circle, PM is the mean proportional between DM and ME, that is, $PM^2 = DM \cdot ME$. Hence PM must also be the mean proportional between HM and MK. But this property is precisely the characteristic property of a circle. Thus there are *two* families of sections of a cone consisting entirely of circles, those parallel to the base, and those that are subcontrary. All other sections are noncircular. They form the subject matter of the treatise.

With this much as background, let us look at the construction of an ellipse in Book I, referring to Fig. 6.6. (Apollonius' constructions of the hyperbola and parabola are analogous.) To do so, we consider a planar section of a cone that is neither parallel to the base nor subcontrary. The distinguishing property of an elliptic section is that the cutting plane intersects all the generators of the cone on the same side of its apex. To state this fact clearly, one needs an axial section of the cone made by a plane perpendicular to the cutting plane. The cutting plane is then required to intersect both sides of the axial triangle. Apollonius proved that there is a certain line, which he called *the upright side* (now known by its Latin name *latus rectum*) such that the square on the ordinate from any point of the section to its axis equals the rectangle applied to the portion of the axis cut off by this ordinate (the abscissa) and whose defect on the axis is similar to the rectangle formed by the axis and the latus rectum. He gave a complicated rule for constructing the latus rectum. Because of its connection with the problem of application with

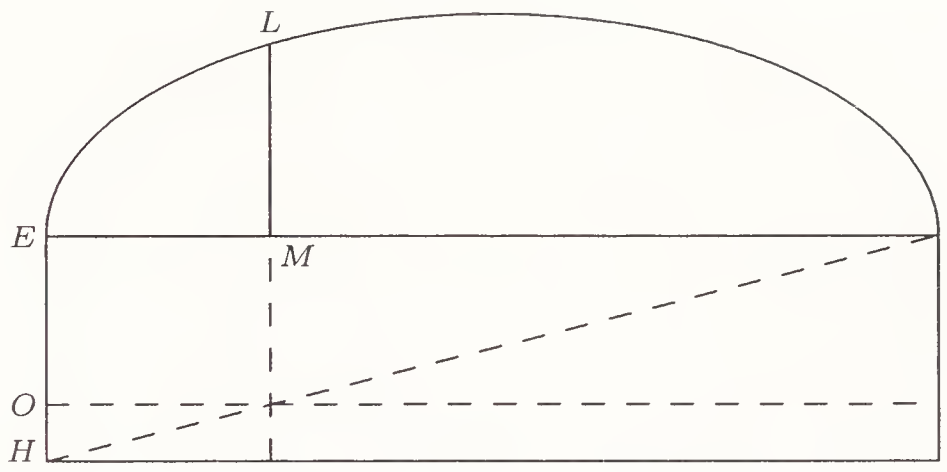


Figure 6.6: Apollonius' construction of the ellipse.

defect, he called the resulting conic section an *ellipse*. (Similar connections with the problems of application and application with excess respectively arise in Apollonius' construction of the parabola and hyperbola. These connections motivated the names he gave to these curves.)

The locus definition for an ellipse is not far removed from what we now think of as the equation of the ellipse. If we write $LM = y$ and $EM = x$ in Fig. 6.6 (so that we are essentially taking rectangular coordinates with origin at E) ,we see that Apollonius is claiming that $y^2 = x \cdot EO$. Now, however, $EO = EH - OH$, and EH is constant, while OH is directly proportional to EM, that is, to x. Thus, if we write $OH = kx$ and $EH = C$, we find that Apollonius' locus condition can be stated as the equation $y^2 = Cx - kx^2$. By completing the square on x, transposing terms, and dividing by the constant term, we can bring this equation into what we call the standard normal form for an ellipse with center at $(h, 0)$:

$$\frac{(x-h)^2}{a^2} + \frac{y^2}{b^2} = 1.$$

(In these terms the latus rectum is $2b^2/a$.) For this reason, some mathematicians say that Apollonius' methods were essentially equivalent to analytic geometry. This statement must be regarded with extreme caution, however. Apollonius did *not* have the concept of an equation nor the symbolic algebraic notation we now use, and this absence gave his work on conics a ponderous character with which most mathematicians today have little patience.

We emphasize again the "planimetric" character of Apollonius' development of the properties of conic sections. Once the locus condition that characterizes the conic is accepted, the whole three-dimensional apparatus can be dispensed with. All subsequent theorems that do not explicitly mention cones in their hypotheses will be proved using the locus condition, which, as we have just seen, is logically equivalent to what we call the equation of the curve.

Axes and Diameters

Apollonius defines a *diameter* of a conic to be a chord that bisects any chord that it intersects from a family of parallel chords. In other words, given a family of

parallel chords, the diameter is the locus of their midpoints. It is not obvious that this locus is a straight line, and that is what Apollonius proves early in the work (Book I, Proposition 7): *If a cone is cut by a plane which intersects the plane of the base in a straight line perpendicular to the base of the axial triangle, the intersection of the plane and the axial triangle is a diameter of the section.* This construction makes it easy to construct diameters by merely drawing the line through the midpoints of any two parallel chords (Book II, Proposition 44). Apollonius shows that if one diameter bisects a second, then the second also bisects the first (Book I, Proposition 15). The two diameters are then said to be *conjugates*, and the point where they meet (the midpoint of each of them) is called the *center*. Only circles, ellipses and hyperbolas have centers, and of course the center of such a *central conic* can easily be found by merely drawing two diameters (Book II, Proposition 45). If a diameter is the perpendicular bisector of a family of parallel chords, it is called an *axis* of the conic. Apollonius shows (Book II, Proposition 48) that no central conic has more than two axes. The importance of conjugate diameters shows up especially in the construction of the tangent to a central conic. The tangent at any point is parallel to the conjugate diameter to the diameter through that point, as Apollonius shows in Book I, Propositions 47 and 48. Thus drawing the tangent to a conic is a straightforward operation, given the conic, in contrast to the more complicated spirals considered by Archimedes. The importance of tangents to conic sections appears in connection with the locus problems considered later.

For the hyperbola Apollonius proves the existence of *asymptotes*, that is, a pair of lines through the center that never meet the hyperbola, but such that any line through the center passing into the region containing the hyperbola does meet the hyperbola. The word *asymptote* means literally *not falling together*.

Books I and II are occupied with finding the proportions among line segments cut off by chords and tangents on conic sections, the analogs of results on circles in Books III and IV of Euclid. These constructions involve finding the tangents to the curves satisfying various supplementary conditions such as being parallel to a given line, etc.

Foci and the Three-Line Locus

We are nowadays accustomed to constructing the conic sections using the focus-directrix property, so that it comes as a surprise that the original expert on the subject does not seem to recognize the importance of the foci. He never mentions the focus of a parabola, and for the ellipse and hyperbola he refers to these points only as "the points arising out of the application." The "application" he has in mind is explained in Book III. Propositions 48 and 52 read as follows:

(Proposition 48). *If in an ellipse a rectangle equal to the fourth part of the figure is applied from both sides to the major axis and deficient by a square figure, and from the points resulting from the application straight lines are drawn to the ellipse, the lines will make equal angles with the tangent at that point.*

(Proposition 52). *If in an ellipse a rectangle equal to the fourth part of the figure*

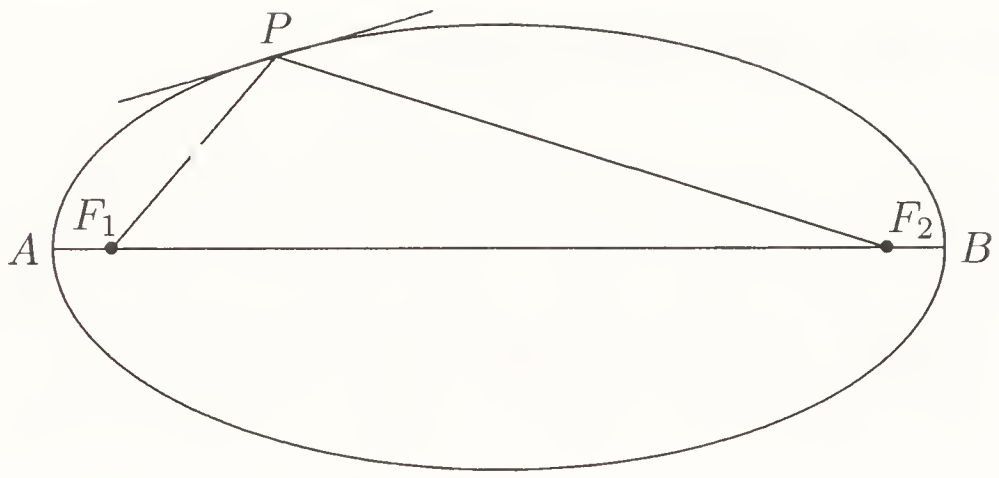


Figure 6.7: Focal properties of an ellipse.

is applied from both sides to the major axis and deficient by a square figure, and from the points resulting from the application straight lines are drawn to the ellipse, the two lines will be equal to the axis.

The "figure" referred to is the rectangle whose sides are the major axis of the ellipse and the latus rectum. In Fig. 6.7 the points F_1 and F_2 must be chosen on the major axis AB so that $AF_1 \cdot F_1B$ and $AF_2 \cdot BF_2$ both equal one-fourth of the area of the rectangle formed by the axis AB and the latus rectum. Proposition 48 expresses precisely the focal property of these two points: any ray of light emanating from one will be reflected to the other. Proposition 52 is the "string property" that characterizes the ellipse as the locus of points such that the sum of the distances to the foci is constant. These are just two of the dozens of theorems Apollonius was immodest enough to call "remarkable." Apollonius makes little use of these properties, however, and does not discuss the use of the string property to draw an ellipse.

A very influential part of the *Conics* consists of Propositions 54–56 of Book III, which contain the solution to the three- and four-line locus problems mentioned by Apollonius in his preface. Both in their own time and because of their subsequent influence the three- and four-line locus problems have been of enormous importance for the development of mathematics. To avoid excessive complexity, we merely state the three-line locus problem and tell how it was solved. The data for the problem are three lines, which for definiteness we suppose to intersect two at a time so as to form a triangle, and three given angles, one corresponding to each line. The problem requires the locus of points P such that if lines are drawn from P to the three lines, each making the corresponding angle with the given line, the square on the first will have a constant ratio to the rectangle on the other two. Apollonius shows that this problem can be solved by drawing a conic section having one side of the triangle as a chord and at the same time tangent to the other two lines at the endpoints of the chord. This conic will then be the required locus. We shall see that later mathematicians such as Pappus and Descartes set great store by these locus problems. The solution of this problem was later seen as the high-water mark of Greek geometry. It was his success in extending this problem to more than four lines that convinced Descartes of the value of his geometric methods.

6.3 Problems and Questions

6.3.1 Problems from Archimedes and Apollonius

Exercise 6.1 Show that the problem of squaring the circle is equivalent to the problem of squaring one segment of a circle when the central angle subtended by the segment is known. (Knowing a central angle means having two line segments whose ratio is the same as the ratio of the angle to a full revolution.)

Exercise 6.2 Show that the real-valued *rational functions* are an example of a non-Archimedean ordered field. A rational function is the quotient of two polynomials. One rational function $R_1(x)$ is said to be larger than another $R_2(x)$ if $R_1(x) > R_2(x)$ for all large positive values of x. An integer in this context means a constant function whose constant value is an integer. Show that in this field no integer n exceeds the function x.

Exercise 6.3 Referring to Fig. 6.1, show that all the right triangles in the figure formed by connecting B' with C, C' with K, and K' with L are similar. Write down a string of equal ratios (of their legs). Then add all the numerators and denominators to deduce the equation

$$(BB' + CC' + \cdots + KK' + LM) : AM = A'B : BA.$$

Exercise 6.4 The parametric equations of an Archimedean spiral are $x = a\theta\cos\theta$, $y = a\theta\sin\theta$. Use these equations to prove that the tangent to the spiral at the point $(2\pi a, 0)$ (corresponding to $\theta = 2\pi$) meets the y axis at the point $(0, -4\pi^2 a)$.

Exercise 6.5 Prove the proportion that Archimedes claims in Fig. 6.3. To do so use the facts that $MS = CA$ and $SQ = AS$ to establish that $MS \cdot SQ = CA \cdot AS = AO^2 = OS^2 + AS^2 = OS^2 + SQ^2$. Then use the fact that $HA = CA$ to show that $HA : AS = CA : AS = MS : SQ = MS^2 : MS \cdot SQ = MS^2 : (OS^2 + SQ^2) = MN^2 : (OP^2 + QR^2)$.

Exercise 6.6 Show that Archimedes' result on the relative volumes of the sphere, cylinder, and cone can be obtained more simply by considering the cylinder, sphere and double-napped cone formed by revolving a circle inscribed in a square about a midline of the square, the cone being generated by the diagonals of the square. In this case the area of a circular section of the cone plus the area of the same section of the sphere equals the area of the section of the cylinder since the three radii form the sides of a right triangle. (The radius of a section of the sphere cuts off a segment of the axis of rotation from the center equal to the radius of the section of the cone, since the vertex angle of the cone is a right angle. These two segments form the legs of a right triangle whose hypotenuse is a radius of the sphere, which is equal to the radius of the section of the cylinder).

Exercise 6.7 One minor work of Archimedes that we did not discuss, called the *Book of Lemmas*, contains the following trisection of the angle. In Fig. 6.8, we are

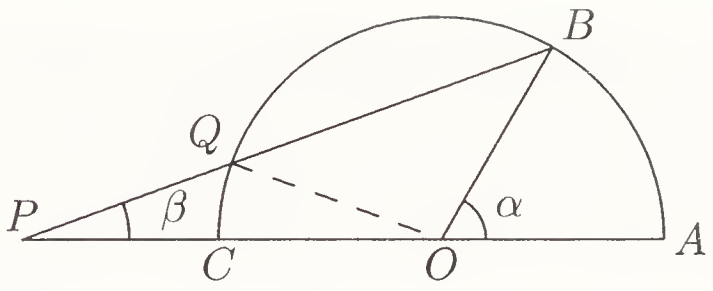


Figure 6.8: Archimedes' trisection of an angle ($\beta = \frac{1}{3}\alpha$).

given an acute angle $\alpha = \angle AOB$, whose trisection is required. We draw a circle ABC of any radius r about O, the vertex of the angle. Then, using a straightedge, we mark off on it two points P and Q separated by the distance r. Setting the straightedge down so that P is on the extension of the diameter CA, Q is on the semicircle ABC, and the point B is also on the edge of the straightedge, we draw the line PB, which contains the point Q. By drawing the radius QO, we obtain two isosceles triangles OQP and QOB. The equal angles of the first of these will be denoted β, and since the exterior angle of a triangle equals the sum of the two opposite interior angles, it follows that the equal angles of the second are equal to 2β. Therefore $\angle BPO = \beta$, $\angle PBO = 2\beta$, and again by the exterior angle theorem $\alpha = 3\beta$. That is, we have constructed an angle β equal to one-third of α. Why is this construction *not* a straightedge-and-compass trisection of the angle, which is known to be impossible?

Exercise 6.8 Justify the remark in the text that the problem of increasing the size of a sphere by half is equivalent to the problem of two mean proportionals (duplicating the cube).

Exercise 6.9 A circle can be regarded as a special case of an ellipse. What is the *latus rectum* of a circle?

Exercise 6.10 When the equation $y^2 = Cx - kx^2$ is converted to the standard form

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

what are the quantities h, a, and b in terms of C and k?

Exercise 6.11 Show from Apollonius' definition of the foci that the product of the distances from each focus to the ends of the major axis of an ellipse equals the square on half of the minor axis.

Exercise 6.12 We have seen that the three- and four-line locus problems have conic sections as their solutions. State and solve the two-line locus problem. You may use modern analytic geometry and assume that the two lines are the x axis and the line $y = ax$. The locus is the set of points whose distances to these two lines have a given ratio. What curve is this?

Exercise 6.13 Show that the apparent generality of Apollonius' statement of the three-line locus problem, in which arbitrary angles can be prescribed at which lines are drawn from the locus to the fixed lines, is illusory. (To do this, show that the ratio of a line from a point P to line l making a fixed angle θ with the line l bears a constant ratio to the line segment from P perpendicular to l. Hence if the problem is solved for all ratios in the special case when lines are drawn from the locus perpendicular to the given lines, then it is solved for all ratios in any case.)

Exercise 6.14 Show that the line segment from a point $P = (x, y)$ to a line $ax + by = c$ making angle θ with the line has length

$$\frac{|ax + by - c|}{\sqrt{a^2 + b^2}\sin\theta}.$$

Use this expression and three given lines $l_i : a_i x + b_i y = c_i$, $i = 1, 2, 3$, to formulate the three-line locus problem analytically as a quadratic equation in two variables by setting the square of the distance from (x, y) to line l_1 equal to a constant multiple of the product of the distances to l_2 and l_3. Show that the locus passes through the intersection of the line l_1 with l_2 and l_3, but not through the intersection of l_2 with l_3. Also show that its tangent line where it intersects l_i is l_i itself, $i = 2, 3$.

6.3.2 Questions about Archimedes and Apollonius

Exercise 6.15 Can you suggest any criteria for "mathematical greatness" not given in the list at the beginning of this chapter? In what ways does Archimedes fit these criteria? Give examples from the discussion of his works.

Exercise 6.16 At one point in his estimation of π Archimedes arrives at a rational approximation that he writes as $591\frac{1}{8} : 153$. Now the fraction $\frac{1}{8}$ that occurs in this expression could actually be replaced by $\frac{1}{7}$, and would yield a stronger estimate. Why do you think Archimedes prefers the weaker estimate with $\frac{1}{8}$?

Exercise 6.17 Archimedes' argument actually obtains the upper bound $\frac{14,688}{4673.5}$ for π. How much smaller is this number than $\frac{22}{7}$? Why do you think Archimedes settled for the weaker estimate $\frac{22}{7}$?

Exercise 6.18 Why didn't Archimedes give a pure existence proof for the finiteness of the number of grains of sand required to fill up the universe, based on his axiom that some multiple of any volume must exceed any other volume?

Exercise 6.19 Archimedes' *Method* is based on an appeal to physical intuition. For example, each section of the sphere and cone discussed above is balanced by a corresponding section of the cylinder. As mentioned above, a very similar technique now known as Cavalieri's principle was used both in China and in early modern Europe for finding areas and volumes. The idea is that if every pair of sections of two bodies has a certain ratio, say $2 : 1$, then the bodies themselves will

have this ratio. The difficulty from a logical point of view is that, as Zeno showed, a continuous body is not simply the "sum" of lower-dimensional pieces. We therefore cannot be sure that our intuition does not mislead us. Does Archimedes' introduction of the principle of the lever in his application of the method answer this objection?

Exercise 6.20 One reason for doubting Cavalieri's principle is that it breaks down in one dimension. Consider, for instance, that every section of a right triangle parallel to one of its legs meets the other leg and the hypotenuse in congruent figures (a single point in each case). Yet the other leg and the hypotenuse are obviously of different lengths. Is there a way of redefining "sections" for one-dimensional figures so that Cavalieri's principle can be retained? If you could do this, would your confidence in the validity of the principle be restored?

Exercise 6.21 We know that interest in conic sections *arose* because of their application to the problem of two mean proportionals (duplication of the cube). Why do you think interest in them was *sustained* to the extent that caused Euclid, Aristaeus, and Apollonius to write treatises developing their properties in such detail?

Exercise 6.22 As we have seen, Apollonius was aware of the string property of ellipses, yet he did not mention that this property could be used to draw an ellipse. Do you think that he did not *notice* this fact, or did he omit to mention it because he considered it unimportant?

6.4 Endnotes

1. The quotation from Cicero is taken from his *Tusculan Disputations*, translated by J. E. King (G. P. Putnam's Sons, New York, 1927), pp. 491, 493.

2. The information on the modern identification of Archimedes' tomb is based on *La tomba di Archimede* by Salvatore Ciancio (Ciranna, Rome, 1965).

3. The discussion of Archimedes' *Measurement of the Circle* is based on Vol. 1 of Clagett's *Archimedes in the Middle Ages* (University of Wisconsin Press, 1964), p. 49, and on T. L. Heath's *The Works of Archimedes*, Dover Reprint, New York 1953, pp. 91–99.

4. All quotations from Archimedes' prefatory letters to his works are taken from the book of Heath (op. cit.).

5. Eutocius' research on Archimedes is described in the book of Heath (op. cit.), p. 66.

6. Archimedes' near approach to the idea of a derivative in his work on spirals was pointed out in an after-dinner speech given by S. Bochner on his retirement from Princeton University in 1969.

7. Pappus' comments on Apollonius can be found in the book of T. L. Heath, *Apollonius Pergaeus. Treatise on Conic Sections* (Cambridge University Press, 1896), pp. xxxi–xxxiii.

Chapter 7

Hellenistic Mathematical Science

Both science and technology reached very high levels during the half millennium that passed between 200 B.C.E. and 300 C.E. There is not enough space here to discuss the technology of the period, and in any case this technology had very little to do with mathematics. For that reason we shall concentrate on the role played by mathematics in the formulation of the principles of certain parts of physics and astronomy. This mathematical science poses a problem for the historian. It seems to be an application of some of the simpler parts of Euclid's geometry together with the older and distinctly uneuclidean idea of attaching numbers (lengths) to line segments and arcs. Did the applied mathematicians guide the development of geometry, or did they opportunistically use the relations that geometers had developed? The texts do not tell us the answer.

We have seen that mathematical astronomy in Mesopotamia was basically numerical, a matter of counting the days between full moons, eclipses, etc. In the Greek world geometry was added to this science in a combination with computation that prefigures the subject now known as trigonometry. Geometry also penetrated into mechanics, optics, and astronomy in the time of Euclid. These three areas will form the subject matter of this chapter.

7.1 Mechanics

As mentioned in the preceding chapter, Archimedes made important contributions to the mathematization of physics, using the geometric theory of proportion to establish the basic principles of the lever and floating bodies. He was not the first to attempt a mathematical treatment of these questions, however. The subject of physics was already well advanced by the time of Aristotle, who included it among the many subjects on which he wrote treatises. We shall contrast Aristotle's explanation of the principle of the lever with that of Archimedes.

7.1.1 Aristotle's *Physics* and *Mechanics*

The *Physics*

The word *physics* is related to other ancient Greek words such as *physikos*, meaning *natural* or *inborn*. This etymology reflects the principles by which Aristotle interpreted the physical world: each identifiable object in the world has a certain *inborn nature* in terms of which its behavior is explained. This nature expresses itself as a *cause of motion*. Aristotle uses the terms *motion* and *rest* more broadly than we use them today. He defines motion as "the fulfillment of what exists potentially." Thus growth and decay are forms of motion in this sense. To preserve the distinction it is useful to refer to the common meaning of the English word *motion*, that is, change of place, as *locomotion*. In all cases motion (in the broader sense) requires (1) a mover (cause of motion), (2) a thing moved, and (3) a state or place moved *to*. Aristotle does not include a state or place moved *from*, on the grounds that it is the *goal* of the motion that determines its nature. In Book VII of the *Physics* he lays down the general principle that "everything that is in motion must be moved by something." By tracing backward from the existing motion and denying the actually infinite, he is thus led to the ultimate source of a particular motion, which must itself be eternally unmoved. The cause of motion of a particular body, however, may be the body itself; Aristotle includes this case among the possible cases of motion. For bodies moved by other bodies he distinguishes four cases: pushing, pulling, carrying, and rotating. All more complex motions, such as compressing, stretching, and combing, he says, are to be analyzed as various combinations of these things.

The principles of physics laid down thus far are qualitative and verbal. Mathematics does not enter into this *kinematic* theory (from *kinymai*, meaning *to be moved*). When Aristotle comes to formulate *dynamics* (from *dynamis*, meaning *power*), however, he does invoke mathematics, in particular the notion of proportion. He says in Book VII.5 that if a mover has moved a body B a distance Γ in time Δ, then in the same time the same mover will move a body of size $\frac{1}{2}B$ a distance 2Γ, and in time $\frac{1}{2}\Delta$ it will move the body $\frac{1}{2}B$ the full distance Γ, "for in this way the rules of proportion will be observed." Thus it appears at first sight that Aristotle is saying that a "mover" can be quantified as the product of the "body" and the distance divided by the time. If we assume that bodies are measured by their mass and that the quantified "mover" is force, this definition would make force proportional to what we now call *momentum*. This projection of modern ideas, however, will not work. For Aristotle has only an intuitive concept of what we now call instantaneous velocity. This concept occurs implicitly when he talks about objects moving faster or slower than other objects. Whenever a specific motion is in question, however, he always speaks of the distance moved in a given time, that is, the data for finding what we would call the *average* velocity. Moreover, he explicitly denies that the simple mathematical proportion that he uses to relate the motion of a body to the motion of a body half as large can be extended in the opposite direction to relate to a body *twice* as large.

The principles of physics are further developed in Aristotle's book *On the*

Heavens, where, in Book III, he discusses motion for bodies below the circle of the moon. It is in this treatise that most of the famous "erroneous" principles that everyone has heard of are formally stated. For example, in Book II, while explaining his reason for believing the earth to be circular, Aristotle reports that "a little bit of earth, let loose in mid-air, moves and will not stay still, and the more there is of it the faster it moves." Apparently this misleading observation had caused people to wonder why the *whole* earth does not similarly move, yet faster. Aristotle advances a circular earth at the center of the universe as the solution of the seeming paradox. Again, to reconcile the principle that everything in motion is moved by something with the observation that a stone thrown through the air continues to move even after it has left the hand, Aristotle says that, "the force transmits the movement to the body first by, as it were, impregnating the air," so that it is really the air that moves the object. In Book IV Aristotle distinguishes between gravity and levity (gravity is not a force, as we now think of it, but rather a property of a body, as is its opposite, levity, which we no longer believe in at all). Aristotle says that "there are things whose constant nature it is to move away from the center, while others move constantly toward the center." Whether this center is the center of the earth or the center of the universe, Aristotle says, can be left to another inquiry, since the two centers coincide. Aristotle understands that motion cannot be explained only in terms of what we call density, but that shape must play a role. He says that, while shape will not account for motion per se, it will account for the velocity of the motion. He notes that a (thin) flat piece of iron can be made to float on water, while a round piece will sink. Democritus, who had propounded an atomic theory of matter, claimed that warm bodies (atoms) could move up out of the water in sufficient numbers to support a broad piece of iron, but not a narrow one. To the objection that this effect is not observed in air Democritus (as reported by Aristotle) had replied that the atoms of air do not all move upward, so that their action was diluted, so to speak. Aristotle comes close to the modern explanation (surface tension) in saying that air is more easily divided than water, so that a body placed on a broad surface of water must "work harder" to divide the water and thus sink.

The *Mechanics*

The treatise known as the *Mechanics* was almost certainly the work of scholars who adhered to Aristotle's basic principles, but it was not written by Aristotle himself. The title of this treatise, which comes from *mechane*, meaning *machine*, is almost the exact opposite of *physics*. Indeed the author of the work says in the preface that the subject matter of mechanics is precisely those devices used to produce an effect *contrary to nature*. The most important simple machine is the lever, which makes it possible to move great weights with very little force, in seeming contradiction to nature.

The author of the *Mechanics* traces the basic principle of the lever to the circle, which embodies a unity of opposites. For when a circle is rotated about its center, it is in motion, yet the center does not move. In addition antipodal points on a rotating circle are moving in opposite directions at any given time. Two such

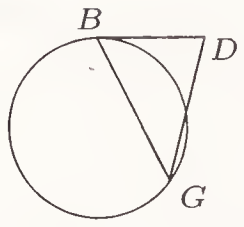


Figure 7.1: Circular motion according to Aristotle.

points are continually exchanging their motions. If one thinks of the points as being ordered, then the last point is continually becoming the first point.

The practical question about levers to which the author wishes to apply physical principles is: Why are large balances more accurate than small ones? This question is formulated in physical terms as the following: Why does a point on a rotating circle farther from the center move more rapidly than a point closer to the center? The author finds the answer to this question in the principle of combined motions:

> Now whenever a body is moved in two directions in a fixed ratio, it necessarily travels in a straight line, which is the diagonal of the figure which the lines arranged in this ratio describe.

Here we find a very clear statement of the principle known as the *parallelogram law* for addition of vectors. It is applied here to the path followed by an object moved by two forces at once, whereas in modern physics it applies to a wide variety of physical quantities. The author gives a geometric argument to justify this assertion. Then (see Fig. 7.1) the author says that circular motion cannot be compounded of two constant motions such as BD and DG, since if it were, the body would move along the chord BG in accordance with the parallelogram principle. Circular motion of the body is explained in terms of the attraction of the center of the circle. Points near the center are more strongly attracted than those farther away, and since the more interference there is, the more slowly an object moves, it follows that those points nearer the center will move more slowly than those farther away.

> This happens with any radius which describes a circle; it moves along a curve naturally in the direction of the tangent, but is attracted to the center contrary to nature. The lesser radius [of a balance beam] always moves in its unnatural direction; for because it is nearer to the center which attracts it, it is the more influenced.

The idea expressed here is contained in the concept of curvature. The curvature of a circle, as you know from calculus courses, is inversely proportional to the radius—A small circle curves more than a large one. In fact a sufficiently large circle *looks* rather straight.[1] Thus, if a point is directly above the center of a circle and moving along the circle, its natural motion will be tangential, to the right

[1] This observation was used by Aristotle in his book *On the Heavens* to reply to those who claimed the earth was flat.

or left; but if the point is attached to the center, it will be constrained to move downward (toward the center). This latter effect will be greater at points closer to the center. The author says explicitly that "one would expect the two motions (natural and unnatural) to be in proportion."

Applying this principle to a lever, the author of the *Mechanics* states that "the ratio of the weight moved to the weight moving it is the inverse ratio of the distances from the center."

The *Mechanics* is not a deductive physical theory. It does not contain any general postulates as a starting point, but is rather confined to special problems solved by *ad hoc* methods. In this respect it contrasts with the work of Archimedes we are about to discuss.

7.1.2 Archimedes' Physical Treatises

Levers

While Aristotle's explanation of the lever involves only a rudimentary amount of geometry, the treatise of Archimedes *On the Equilibrium of Planes* gives a geometric derivation of the law in accordance with the strict principles of the Eudoxan theory of proportion. Moreover, Archimedes makes no use of the large array of concepts from Aristotelian physics—he never mentions gravity or levity, nor does he make use of any principles like "whatever moves is moved by something," nor does he classify motion into pushing, pulling, carrying, and rotating. Archimedes' starting points (postulates) are mathematical and apparently based on intuitive ideas of symmetry. To give an idea of the structure of the theory Archimedes intends to create, we shall list some of his postulates.

1. Equal weights balance at equal distances (from the fulcrum); at equal distances with unequal weights the larger weight will sink.

2. If two weights balance and more weight is added to one of the two, that one will sink.

3. If two weights balance and some weight is removed from one of the two, that weight will rise.

4. If two magnitudes balance at certain distances (regardless of shape) two other magnitudes equal to them will also balance at these distances

From his postulates Archimedes deduces with strict logic that weights that balance at equal distances are equal (Proposition 1) and that if unequal weights balance, the larger weight will be closer to the fulcrum (Proposition 3). At this point Archimedes makes use of the concept of center of gravity, proving that the center of gravity of a system of two equal bodies is the midpoint of the line joining their individual centers of gravity (Proposition 4) and that if three equal magnitudes have their centers of gravity on a straight line, the outer two being equidistant from the middle one, then the center of gravity of the three-body system coincides with

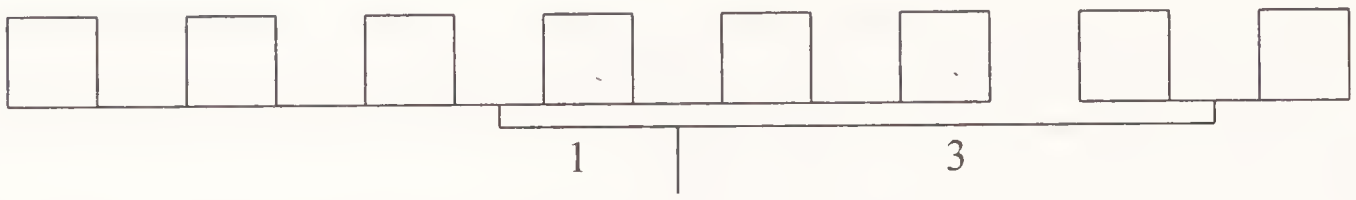


Figure 7.2: Balancing commensurable weights.

Figure 7.3: A lever with fulcrum at C.

the center of gravity of the middle body (Proposition 5). These two propositions are then extended to any even or odd number of bodies symmetrically arranged about a center.

At this point it is an easy matter to show that the inverse proportion law of the lever holds for commensurable magnitudes. The essence of the argument is that one can select a common measure of the two magnitudes such that one of the magnitudes is $2m$ times the common measure and the other is $2n$ times it. Given two distances in the same proportion as the two magnitudes, choose a common measure of the two distances such that one of them is $2m$ times the common measure and the other is $2n$ times it. Then simply imagine the two lines extending in opposite directions from a given point and marked off into $2m + 2n$ equal segments; also imagine the two magnitudes broken into $2m + 2n$ equal pieces. Placing one piece over each segment, one finds that they balance with $m + n$ segments and pieces on each side of the fulcrum. But the first body can be thought of as consisting of $2m$ of these pieces lying on a (weightless) secondary lever situated on top of the primary one, m pieces being on each side of the fulcrum in the secondary lever; and the second body likewise can be thought of as the remaining $2n$ pieces, balanced on a secondary lever with n pieces on each side. It is then clear that these secondary fulcrums are respectively n and m units from the primary fulcrum. Thus the two bodies will balance if their centers of gravity are suspended at these points (see Fig. 7.2).

There are no surprises for the modern mathematician in this treatment of the law of the lever for commensurable magnitudes. Archimedes' treatment of the subject differs from modern versions only in the logical rigor with which Archimedes deduced the fundamental principle. Adhering to the principles of the Eudoxan theory of proportion, he found it necessary to consider the incommensurable case as well. Since this law is seldom proved in modern physics textbooks, it will be of interest to see how incommensurable magnitudes are handled. Archimedes argues as follows.

Assume that the magnitudes A and B are incommensurable. Choose C on the line DE so that $A : B = DC : CE$, as in Fig. 7.3. We claim that weight A placed at E and B placed at D must balance about C as a fulcrum. Suppose not, that is, suppose that the point E sinks and the point D rises when the weights are

suspended this way. Then A is too heavy to balance B, and so some smaller weight (say, A') must balance B exactly. Between A and A' there is some weight A'' that is commensurable with B. But then the ratio $A'' : B$ is less than $DC : CE$, and so by the commensurable case the point E will rise when A'' is suspended at E and B is suspended at D. *A fortiori* the point E will rise when A' is suspended at E and B is suspended at D, contradicting the way in which A' was chosen. A similar proof, with a weight A' larger than A and A'' between A and A' and commensurable with B, shows that the point E cannot rise when the weights A and B are suspended as stated. Thus by the trichotomy we conclude that the weights will balance.

Notice that Archimedes' use of the Eudoxan theory of proportion avoids the cumbersomeness of the direct definition of proportion from Book V of Euclid. Evidently someone, perhaps Archimedes himself, discovered that arguments could be "streamlined" by making use of the following result: *Given three magnitudes A, A', and B of the same kind, with $A \neq A'$, there is a magnitude A'' of the same kind intermediate in size between A and A' and commensurable with B.* This principle can be stated in modern language by saying that the rational numbers are dense in the real numbers.

Hydrostatics

Archimedes' treatises *On Floating Bodies* are the earliest mathematical treatments of hydrostatics. He begins, as in the treatise on levers, with a simple intuitive postulate. If we imagine that the only force acting on a given point, curve, or surface in a fluid is the weight of the fluid directly above it, the part in question can be thought of as compressed by the weight of the fluid above it. If this compression is uneven between two points at the same distance from the center toward which the fluid is sinking (the center of the earth), then the fluid will move laterally from the point of higher pressure toward the point of lower pressure. In modern terms this is the definition of a fluid: a substance that cannot support a shear stress. This principle is the only postulate for the work. From this principle it is not difficult to deduce that the surface of a fluid at rest is a sphere whose center is the center of the earth (Proposition 2). Archimedes then shows that a solid of equal density with a fluid will sink to a level even with the surface but not lower. (He doesn't use the word *density*, but rather refers to solids that, "size for size are of equal weight with the fluid.") For if the solid did not sink to that level, the fluid at any level below the body would be compressed by a weight equal to the weight of fluid above any other portion of the same size and elevation, and also by the weight of the portion of the body projecting out of the fluid. Hence there would be a pressure imbalance at that level and equilibrium would not occur.

The fundamental principles of hydrostatics are all deduced from this single principle. The most important of these is Proposition 7: *A solid heavier than a fluid will, if placed in it, descend to the bottom of the fluid, and the solid will, when weighed in the fluid, be lighter than its true weight by the weight of the fluid displaced.*

This last proposition forms the basis for the famous story of Archimedes in the bath. The Roman architect Vitruvius gives this story in the preface to Book IX of his *De Architectura*. According to Vitruvius, Archimedes weighed an amount of gold equal to the weight of the king's crown and also an amount of silver equal to the weight of the crown. He then measured the amount of water displaced by these three equal weights. He found that the crown displaced more water than an equal weight of gold, but less than an equal weight of silver.

In Book II Archimedes investigated the stability of a segment of a paraboloid of revolution floating in a fluid. In particular in Proposition 2 he showed that a segment cut off perpendicular to the axis and placed in the fluid with the base above the surface of the fluid will adjust its position so that the axis of the segment is vertical, provided the material of which the segment is made has density less than that of the fluid and the length of the axis of the segment is at most three-fourths of the latus rectum p of the generating parabola. For segments whose axis A is larger than $\frac{3}{4}p$ he showed that the same proposition will hold provided the density is less than that of the fluid but not less than $\left(1 - \frac{3}{4}p/A\right)^2$. Continuing in this vein, Archimedes studied the equilibrium positions of segments of paraboloids of revolution of many shapes, describing precisely the angle between the axis and the vertical. This work is an imposing intellectual feat, leading to very elegant descriptions of physical situations that one would have found impossible to analyze without the groundwork provided by the fundamental principles of geometry.

7.1.3 Heron's Mechanical Works

The dates of Alexandrian scholar Heron are uncertain, somewhere between 150 B.C.E. and 250 C.E. The majority opinion puts him in the first century C.E. He is best remembered for having discovered how to find the area of a triangle in terms of the lengths of its sides and for having invented an early steam-powered machine. In fact he created many interesting mechanical devices besides the steam engine and wrote a treatise on surveying (*Dioptrica*). In keeping with the principles of the present chapter we shall discuss only certain theoretical work in physics at this point. In his *Mechanica*, part of which is quoted by Pappus, he considers the mechanics of a bent lever. Pappus uses this principle of Heron to discuss the problem of the power (force) required to move a weight up an inclined plane. He imagines the weight as located at the center of a sphere being rolled up the inclined plane and balanced by a fictitious weight B on the surface of the sphere at the same elevation as the center and as close as possible to the plane (see Fig. 7.4). He takes the power required as the sum of the power required to move the two weights along a horizontal surface. (This reasoning uses Aristotelian principles of physics that we no longer accept. On the modern view, except to overcome rolling friction, no force at all is required to roll a ball along a horizontal surface once it has started to roll.) Thus, although the principle of the lever was well understood in Hellenistic times, that of the inclined plane was not. Since the modern law involves only the proportions in a triangle, it seems strange that this simple principle was not discovered.

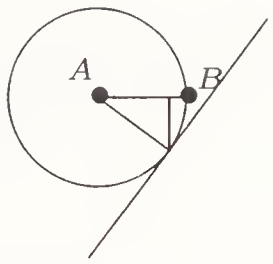


Figure 7.4: The law of the inclined plane according to Heron and Pappus.

7.2 Optics

The subject of optics, understood in Hellenistic times as the study of vision rather than light, is very naturally amenable to geometric treatment. Thus it is not surprising that Euclid wrote a textbook on the subject. A lost work of Archimedes called *Catoptrics* (reflection) is known to have existed because a remark from it is quoted by the fourth-century writer Theon of Smyrna. In the generation after Archimedes' death Diocles wrote a work *On Burning Mirrors*, which begins with two propositions related to its title and then wanders off into pure mathematics. Being so closely akin to geometry, the subject of optics has attracted the attention of many great scientists over the centuries, including the second-century astronomer Ptolemy, the Islamic mathematician Al-Haitham (Alhazen), Descartes, Huygens, and Newton, all of whom wrote treatises on optics.

Euclid: Optics

One of the surviving books of Euclid is a treatise on optics, and it appears that he, like Archimedes, wrote a book on catoptrics (reflection) that has been lost. The Euclidean style is evident in the *Optics*, which begins with a list of postulates, such as the postulate that visual rays are emitted from the eye in straight lines. (This, of course, is the opposite of the actual direction of travel of light rays, but for Euclid's geometric treatment the direction of travel is not important.) Euclid proceeds from these postulates to a large number of propositions that incorporate general principles. Typical of these propositions is Proposition 8, which asserts that the apparent sizes of equal and parallel magnitudes at unequal distances from the eye are not (inversely) proportional to those distances, that is, an object when removed to twice a given distance does not appear to be half as large. The ratio of the apparent size of the near object to the apparent size of the more distant object is *less* than the ratio of the larger distance to the smaller distance. The apparent size of an object is the angle it subtends when projected to the eye of the observer. (This is the fourth of the basic principles laid down by Euclid as a foundation of the subject.)

In precise terms (see Fig. 7.5), the angles BEA and DEG subtended at the eye E by the equal and parallel lines AB and DG are not in inverse proportion with the distances BE and DE (perpendicular to the lines AB and DG), and in fact $\angle DEG : \angle BEA < EB : ED$. Indeed by drawing the arc TZH one can see

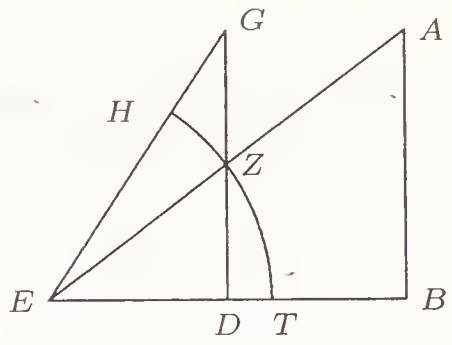


Figure 7.5: Apparent size of objects at different distances.

that $\triangle EZG$ is larger than sector EZH, while $\triangle EDZ$ is less than sector ETZ. Thus

$$\triangle EZG : \text{sector } EZH > 1 > \triangle EDZ : \text{sector } ETZ,$$

so that

$$\triangle EZG : \triangle EDZ > \text{sector } EZH : \text{sector } ETZ$$

and (adding 1 to both sides)

$$\triangle EDG : \triangle EDZ > \text{sector } ETH : \text{sector } ETZ.$$

Now the two sectors are proportional to the central angles they occupy, so that

$$\text{sector } ETH : \text{sector } ETZ = \angle DEG : \angle DEZ.$$

Since the two triangles are on the same base ED they are proportional to their altitudes, so that

$$DG : DZ > \angle DEG : \angle DEZ.$$

Since $AB = DG$ we find that $AB : DZ > \angle DEG : \angle DEZ$, and finally by similar triangles and the fact that $\angle DEZ = \angle BEA$,

$$EB : ED = AB : DZ > \angle DEG : \angle BEA.$$

In the *Optics* Euclid states 61 propositions relating to perspective. Many of these were not new at the time. Proposition 18, for example, which calls for determining the height of an inaccessible object when the sun is above the horizon, echoes the famous story told about Thales' having measured the height of the Great Pyramid by measuring the length of his own shadow and the shadow of the pyramid. It is interesting, however, that this technique does not require any angle to be measured; it relies instead on similar triangles. In that respect it resembles the method used for surveying in China, India, the Islamic world, and Medieval Europe.

An example of the type of theorem on perspective that can be found in the *Optics* is Proposition 35, which asserts that if the line from the eye to the center of a circle is perpendicular to the plane of the circle, then all the diameters of the circle will appear equal. We now know that if the problem implicit in this proposition (*What is the apparent shape of a circle viewed from outside its plane?*) is pursued,

it leads to projective geometry and a unified treatment of the conic sections. It would be too much to expect one scholar to carry mathematics so far, however. This subject was not explored in depth until modern times.

Diocles and Heron: Catoptrics (Reflection)

Diocles The original Greek version of the treatise known as *On Burning Mirrors* and said to have been written by Diocles has been lost. An English translation and facsimile of an Arabic translation, which occasionally mentions Diocles in the third person, has been published by G. J. Toomer. In the introduction to this work Diocles mentions that "when the astronomer Zenodorus came down to Arcadia and was introduced to us, he asked us how to find a mirror surface such that when it is placed facing the sun the rays reflected from it meet [in] a point and thus cause burning." This sentence gives us a hint that the mathematicians of the time were widely scattered geographically, but in constant touch by correspondence, since Diocles, living in Arcadia (in the Peloponnesus) is nevertheless well acquainted with the works of the mathematicians we have encountered and is quite current on the important questions of the time.

The title *On Burning Mirrors* was bestowed because the first two propositions of the treatise are occupied with the question just quoted and a similar question said by Diocles to have been posed to Conon regarding a mirror that would reflect all the sun's rays through the circumference of a circle. Diocles says that Dositheus, whom we have encountered as the recipient of letters from Archimedes, solved the burning mirror problem. The solution is a paraboloid of revolution, which reflects all rays parallel to the axis into the focus, the point on the axis located at a distance from the vertex equal to one-fourth of the latus rectum.

Heron Although Diocles *uses* the fact that the angle of incidence equals the angle of reflection, he does not state this fact as a basic principle or try to derive it from other hypotheses. The famous argument that a path for which the angle of incidence equals the angle of reflection is the shortest path emanating from a given point, meeting a given line, and ending in a second given point is due to Heron, whose derivation of this principle in a work entitled *Catoptrics* is the now standard one, based on Fig. 7.6. In that figure consider any hypothetical path emanating from point S, meeting line MM' in the point P', and then passing to the point E. The line segment SP' is reflected in the line MM' to produce the line segment $S'P'$ of the same length. The problem of minimizing the path length, then becomes a matter of minimizing the length of the paths from S' to E. Since the shortest such path is obviously a straight line $S'PE$, it is clear that in the shortest path $\angle SPM$ (the angle of incidence) equals $\angle MPS'$ by congruence, and $\angle MPS'$ equals $\angle EPM'$ (vertical angles). This geometric principle was thought to apply to the path of a light ray because of the Aristotelian principle that Nature does nothing in vain.

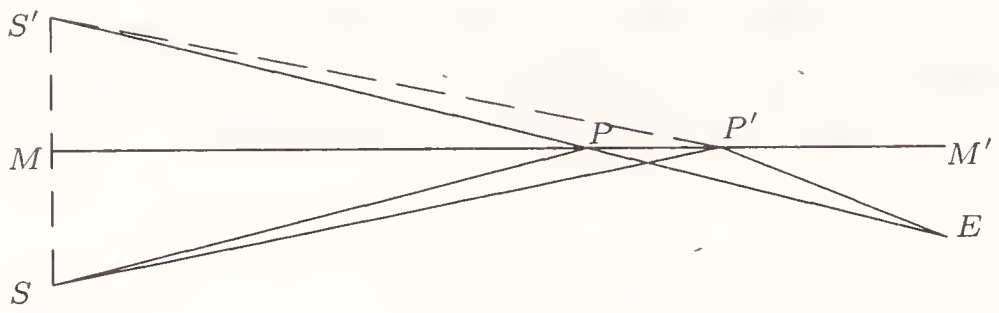


Figure 7.6: Shortest path between two points meeting a line.

Ptolemy: Dioptrics (Refraction)

The phenomenon of refraction manifests itself in the bent appearance of a stick half of which is submerged in water. This phenomenon can be explained by assuming that the brain interprets the light waves reaching the eye *as if* they had traveled along a straight line, while in fact, when the light has passed through two different media such as water and air these rays will be bent at the interface of the two media. The modern explanation of refraction is that light travels at different speeds in different media, and a light ray always moves along a path that (locally) requires less time than any nearby path. Refraction effects in crystals can be very complicated, since the angle of refraction can vary, depending on the polarization of the light. This phenomenon (double refraction), however, was not to be discovered until the seventeenth century. At the early stage we are now considering the only problem to be solved is the quantification of simple refraction in at the interface of media such as water, glass, and air. The problem is to determine the angle the refracted ray makes at the interface and set down the correspondence between the angle of incidence and the angle of refraction.

For light passing from water to air, refraction can be observed by a famous experiment, commonly shown to grade-school students. A coin is placed at the bottom of a cup and the observer moves away from the cup until the coin is concealed from view by the rim of the cup. Water is then poured into the cup, and the coin comes into view, even though it has not moved. This experiment was first described by Ptolemy in his work *Optics*. In the fifth and last of the books of this treatise Ptolemy describes a number of observations of light passing from air to water, air to glass, and water to glass. He observed the empirical fact that light passing from a denser object to one less dense is refracted away from the perpendicular, while the refraction is the opposite when the path moves in the opposite direction (since the light ray can be run in reverse). In observing this effect Ptolemy made an important contribution to the future of science. The explanation we now give for the effect—that light travels with different velocities in the two media—was not involved in Ptolemy's description of the phenomenon. The subject of dioptrics was still in its infancy and was purely observational. Ptolemy used a wheel submerged in water up to its axle in order to determine the refraction of light in passing from water to air. Based on his observations he gave the following table of the angles of refraction for angles of incidence of $10°$, $20°, \ldots, 80°$.

Angle of Incidence	Angle of Refraction
10°	8°
20°	$15\frac{1}{2}°$
30°	$22\frac{1}{2}°$
40°	29°
50°	35°
60°	$40\frac{1}{2}°$
70°	$45\frac{1}{2}°$
80°	50°

Since the differences of the angles of refraction in this table form an arithmetic sequence, one may deduce that the table is effectively the table of values of what we would call a quadratic function. To be specific, if θ_1 is the angle of incidence and θ_2 is the angle of refraction, then

$$\theta_2 = \tfrac{33}{40}\theta_1 - \tfrac{1}{400}\theta_1^2.$$

In writing down this tables Ptolemy took an important step toward the formulation of the concept of correlated variables, which are the essential elements in the modern notion of a function. It is particularly significant that the values are such that the right-hand column can be computed from the left-hand column. The computational process then amounts to the functional operation. It seems that either Ptolemy had in mind a particular *form* for the relation between the two angles, or he noticed that the angle of refraction increased out of proportion to the angle of incidence and chose the simplest way to make this happen, forming successive entries in the table by adding the terms of an arithmetic sequence rather than constant terms. The latter seems more likely on psychological grounds (it is simpler), but no direct evidence exists to confirm or refute it. If he had an idea that the relation should be of a particular type, it would be interesting to know what physical considerations led him to that belief.

7.3 Astronomy

In our discussion of astronomy in Chapters 2 and 3 we summarized a few of the celestial phenomena—full moons, new moons, and eclipses—that are obvious to everyone. These make up the content of what A. Aaboe has called "shepherd's astronomy." Merely by counting days and keeping records of these phenomena over many decades, one can construct an astronomy of some sophistication, suitable for making calendars and even predicting eclipses.

The progress of astronomy has been marked by an increasing ability to take account of small deviations from a basically simple pattern. The simplest pattern of all is the diurnal rotation of the stars. Many centuries are required before the proper motion of the stars relative to one another can be detected. Therefore we can ignore this phenomenon in an account of early astronomy. From a natural

human vantage point the stars seem to be attached to a large sphere (called the *celestial sphere* in geometric astronomy) with center at the center of the earth. This sphere rotates uniformly once every 23 hours, 56 minutes (the *sidereal day*) about an axis tilted at an angle that depends on the geographic location of the observer. At the North Pole the axis is vertical; on the equator it is horizontal. The angle this axis makes with the horizontal (the elevation of the North Star in the Northern Hemisphere) is equal to the geographic latitude of the observer.

Such is the basic pattern against which the less regular motions of other celestial bodies can be described. Actually, as was realized by Hellenistic times, this celestial sphere "wobbles" in its rotation, so that the axis of rotation traverses a circle in the sky, a phenomenon called *precession*. However this wobbling is very slow (less than 2° per century).

The simplest motion to describe against the stars is that of the sun. It seems to march eastward by a little less than one degree of arc per day, along a great circle on the celestial sphere known as the *ecliptic*. The ecliptic makes an angle of about $23\frac{1}{2}^\circ$ with the celestial equator, and the sun is north of the celestial equator during summer in the northern hemisphere of the earth and south of it during the winter. The point on the ecliptic at which the sun crosses the equator when moving from south to north is called the *vernal equinox*. On the day the sun reaches this point, around March 21, there are approximately 12 hours of sunlight and 12 hours of darkness at every point on the earth. This much astronomy can easily be observed informally.

More sophisticated astronomy must take account of certain refinements to this picture. First, the sun moves most slowly along the ecliptic in early July, just after reaching its northernmost point, and conversely it moves most rapidly in early January, just after reaching the southernmost point.[2] Second, after many centuries of observation, it was noticed that the celestial equator was wobbling (the phenomenon pf precession mentioned above), so that the vernal equinox, which by definition is the point on the ecliptic at which the sun crosses the celestial equator moving from south to north, moves along the ecliptic in the direction opposite that of the sun. It follows that this crossing occurs near different stars as time goes on. As a matter of fact, if you measure with sufficient accuracy the point on the ecliptic at which the vernal equinox occurs one year, then again the next year, you will find that the two points are different, and that the sun actually requires some 20 minutes to travel from this year's equinoctial point on the ecliptic to last year's. There is thus a slight discrepancy between the tropical year (the time between two south-to-north crossings of the celestial equator) and the sidereal year (the time required for a complete circuit of the ecliptic).

This precession of the equinoxes obviously will not cause the tropical and sidereal years to diverge very rapidly. In an average human lifetime the difference will be only about one day. Over a period of several centuries, however, a discrepancy large enough to be noticed will accumulate. Thus, when the zodiacal constellations were first used in Hellenistic times, the vernal equinox was at the beginning of

[2]The speed referred to here is measured in degrees of arc per day on the ecliptic. It does *not* refer to the north–south motion of the sun, which is most rapid around the equinoxes.

the constellation Aries. Over the 2.5 millennia that have elapsed since that time it has moved to the beginning of the constellation Pisces.[3] We are writing here as if the ecliptic were a fixed path among the fixed stars. Actually even this path "wobbles" a bit, but the wobbling is a much weaker effect than the precession of the equinoxes, and only modern astronomy is precise enough to handle it. We shall consider it to be a third-order effect, comparable to the proper motions of distant stars, which will be neglected in this book.

After the motion of the sun among the stars has been well established, the more complicated motions of the moon and planets can be considered. The common period of the sun and moon has already been commented on, and cuneiform tablets have been found that give the position in the stars at which the conjunctions and oppositions of the sun and moon (new moons and full moons) occur. These positions are given as a certain number of degrees in a zodiacal sign, each sign occupying a 30° arc of the ecliptic. From the data given in these tablets scholars have been able to infer that they were not compiled by observation. In the first place, the precision of the numbers is too good. For instance, in a famous tablet from the year 102 B.C.E., accurately dated because the tablet itself contains the statement that it was written on Day 18 of Month IX of year 209 of the Seleucid era, the conjunction for Month XII of year 207 of the Seleucid era (March of 104 B.C.E.) is said to have occurred at the position $2° 2' 6'' 20'''$ of the sign of Aries. One can hardly believe that a new moon could be observed within one degree (since the moon is not visible when so close to the sun), much less that it could have been measured to minutes, seconds, and sixtieths of a second! In the second place, the numbers fit a simple pattern, much simpler than what we know to be the actual motions of these conjunctions in the stars. The differences in position from one conjunction to the next increase or decrease by 18 seconds of arc each month, except when the increase or decrease would cause the new value of the difference to exceed $30° 1' 59''$. When that happens, the excess part of the 18 seconds is applied in the opposite direction. Thus the arcs of the ecliptic between successive conjunctions, according to the table, lengthen and shorten in a regular progression. This kind of "Babylonian" numerical astronomy existed side by side with Hellenistic attempts to understand the motions of the celestial bodies. The question naturally arises of whether the astronomers who wrote these tables could detect the discrepancy between the theoretical values of the quantities and the observed values. If not, we can call the unrealistic precision a natural artifact of their mathematical model. If, on the other hand, they had observational data that did not fit the pattern, we must wonder whether they believed strongly enough in their mathematical model to attribute the discrepancy to observational error, or merely considered the mathematical model to be the best simple approximation to the truth.

According to Aaboe, who has performed most of the analysis on which an understanding of this early astronomy is based, scientific astronomy should be

[3]Most astrologers continue to refer to the period from March 21 to April 20 as the "sign of Aries." They rationalize this seeming inconsistency by saying that the names of the constellations were adopted only as convenient reference points when astrology became codified, while the true astrological influences are deeper than mere star patterns.

said to begin when a predictive theory is formulated that enables the positions of the celestial bodies to be *computed* rather than *extrapolated* from tables. To formulate such an astronomy was the challenge to the Hellenistic mathematicians, and we shall now examine their efforts.

7.3.1 Hipparchus

When we are reading authors who record observations of astronomical phenomena, the records themselves may provide evidence for the dating of the documents, due to proper motion of stars, precession of the equinoxes, etc. For this reason the earliest theoretical astronomer we are considering, Hipparchus, can be confidently assigned to the middle of the second century B.C.E. He made very great contributions to astronomy, as we know from other sources, even though only one of his own works is still extant (a commentary on the work of Aratus, who lived two centuries earlier). Ptolemy, in Book III of the *Almagest*, quotes at length observations either made or used by Hipparchus to determine the length of the tropical year. Hipparchus suspected that this length was not a constant number of days and fractions of a day, whereas Ptolemy assures the reader that it is. It is interesting that the earlier astronomer Hipparchus was willing to consider the possibility that the precession of the equinoxes occurs at an uneven rate, while the later scholar Ptolemy will have nothing to do with this hypothesis.

Hipparchus' worry about the constancy of the tropical year was connected with his discovery of the precession of the equinoxes. Ptolemy, in Book VII of the *Almagest*, quotes Hipparchus as saying that the bright star Spica in the constellation Virgo was about 6° ahead of the autumnal equinox, whereas it had been 8° ahead a century and a half earlier in the time of a previous astronomer at Alexandria named Timocharis. It was therefore clear that the celestial equator was not fixed among the stars, but precessed by a small amount.

7.3.2 Apollonius

Our knowledge of Apollonius' contribution to astronomy is indirect, as none of his treatises on this subject have survived. There are references to him in Ptolemy's *Almagest*, however, in connection with the explanation of planetary phenomena.

Once the observation is made that the sun seems to move faster along the ecliptic in winter than in summer, two explanations naturally come to mind for a person assuming a geocentric universe. The first is that, although the sun is moving uniformly along the ecliptic, the earth is not at the center of the ecliptic, but is displaced toward the winter solstice (the southernmost point on the ecliptic). This is the theory of *eccentric motion*. A second explanation—seemingly more complicated, but actually equivalent in simple cases—is that the sun moves along a small circle whose center moves along the ecliptic. In this way the faster and slower phases of the motion can be explained as the periods when the sun is moving in the same direction as the center of its orbit or in the opposite direction. This theory is the *epicycle* theory. The advantage of epicycles shows up in the theory

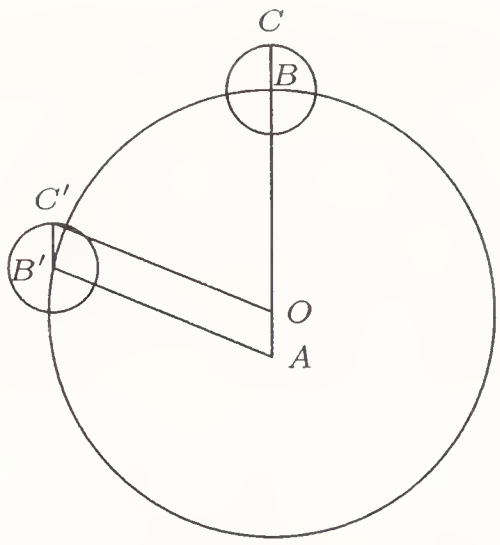


Figure 7.7: Equivalence of epicycles and eccentrics.

of planetary motion: by choosing the radius and speed of rotation on the epicycle suitably one can explain how a planet can appear to move backward among the stars (*retrograde motion*). Once the theory of epicycles has been adopted the explanation of planetary motion is a matter of "curve fitting," that is, choosing an optimal number of epicycles and their radii and periods of revolution, so as to fit observed data.

The fact that a uniform motion along a circle viewed from an eccentric point is exactly the same as a uniform motion along an epicycle combined with a uniform motion of the epicycle can be seen in Fig. 7.7, in which an epicycle of radius r moves along a circle (called the *deferent*) of radius R in such a way that the angle with which a body rotates *clockwise* on the epicycle relative to the extended radius of the deferent equals the angle with which its center has rotated *counterclockwise* along the deferent, measured from a fixed diameter of the deferent. Because $OABC$ is a parallelogram, the angle with which an observer at A at distance r from the center of the deferent sees the center of the epicycle is exactly the same as the angle at which an observer at the center of the deferent sees the point on the epicycle.

The single-epicycle, or eccentric, model is well suited for a comparatively simple motion such as that of the sun. The *path* of the sun among the stars is the ecliptic, which for our purposes is regarded as a fixed circle. Its *motion* along this path, however, is not at a uniform angular rate. It moves most slowly when passing through the constellation Gemini. Nowadays this occurs in early July, shortly after the summer solstice (the northernmost point on the ecliptic). Because of precession of the equinoxes, the summer solstice in Hellenistic times was in the constellation Cancer, so that the slowest motion of the sun occurred before the solstice, in late May. If we use the epicycle model just described, it is clear that the slowest motion occurs when the object moving on the epicycle is farthest from the observer (since at that point the rotation along the epicycle is directly opposite the rotation of the center of the epicycle along the deferent). This point is called *apogee* (farthest from earth). The opposite point is called the *perigee*. The astronomer Hipparchus placed the sun's apogee about $24°$ before the summer solstice. Using this information and the fact that (in his day), spring was $94\frac{1}{2}$ days

long while summer was $92\frac{1}{2}$ days long, Ptolemy managed to fit the sun's motion by using an epicycle and deferent whose radii were in the ratio of $1 : 24$. Such a ratio means that the sun's actual motion will never be more than about $2° 23'$ from its average motion, which is in good agreement with observation.

In Chapter 1 of Book XII of the *Almagest*, Ptolemy attributes the use of epicycles to Apollonius, quoting a particular technical lemma said to have been proved by Apollonius on the location of the points on an epicycle at which retrograde motion begins.

7.3.3 Ptolemy

The author of the standard astronomical text, originally called the *Mathematike syntaxis* (Mathematical treatise) but now better known as the *Almagest* (a hybrid word containing the Arabic definite article *al* and the Greek word *megistos*, meaning "greatest"), lived during the second century C.E. and worked at Alexandria. The treatise itself was published around 150 C.E.

There is insufficient space here to describe the whole treatise, and in any case our primary concern is with its mathematical innovations. We have already quoted parts of it above to show how the Babylonian arithmetical astronomy was refined by the Greeks. The addition of epicycles, which were a prominent feature of the Ptolemaic system, involved considerably more sophisticated geometry than mere measurement of observable positions, as in the Babylonian records. Mathematics makes a valuable contribution to the understanding of astronomy through the sophisticated combination of numbers and geometry known as trigonometry.

Trigonometry

The word *trigonometry* means *triangle measurement*, but angles are generally measured in terms of the amount of rotation they represent, that is, in terms of the ratio of the length of the arc they subtend to the circumference of the circle containing the arc. In a system that is still basically the standard one, Ptolemy divides the circumference into 360 equal parts, and measures angles in terms of those parts, that is, in degrees. The basic problem of trigonometry, from this point of view, is to determine the length of the chord subtended by a given arc and vice versa. To this end, following the Babylonian sexigesimal system, Ptolemy uses $\frac{1}{60}$ of the radius of the circle as the standard of length for chords in a given circle. The effect of this technique is that when two circles intersect, their common chord must be expressed in two different ways, in terms of the two radii. This procedure leads to constant "scaling" of lengths, and is apt to provoke an impatient reaction from the modern reader. Cumbersome though it was, however, it worked and enabled Ptolemy to give an accurate quantitative description of celestial motions.

The computation of the table of chords used by Ptolemy is an interesting exercise in numerical methods. The natural approach would be to start with an angle whose chord is known (say, $60°$), then use half-angle formulas to compute the chord of $30°$, $15°$, $7° 30'$, etc., until the desired tabular difference is achieved, after

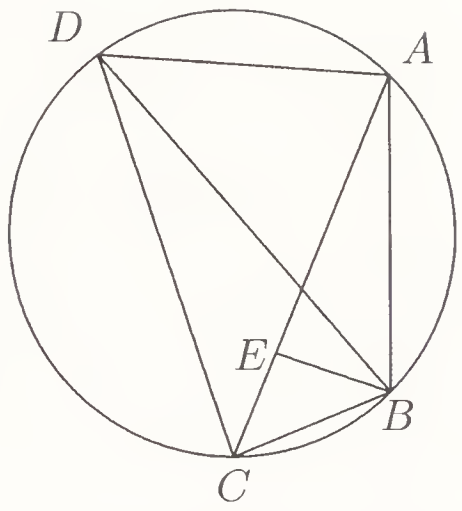


Figure 7.8: Ptolemy's theorem.

which one would build up the table in these intervals using the addition formulas for the trigonometric functions.[4] Ptolemy's approach is like this, but he does the computations very elegantly, using Ptolemy's theorem: *In a quadrilateral inscribed in a circle, the rectangle on the diagonals equals the sum of the rectangles on the two pairs of opposite sides.* To prove this theorem, draw a line BE from the vertex B to the diagonal AC such that $\angle ABE = \angle DBC$, as in Fig. 7.8. Hence $\angle EBC = \angle ABD$. Therefore, since angles BAC and BDC are both inscribed in the same arc, triangles ABE and DBC are similar; for the same reason, triangles EBC and ABD are similar. It follows that $AB \cdot CD + BC \cdot AD = AE \cdot BD + EC \cdot BD = AC \cdot BD$.

Ptolemy's theorem makes it possible to express the chord on the difference of two arcs in terms of the chords on the individual arcs. Given three points on a circle, say A, B, and C, take point D diametrically opposite one of the points, say A (see Fig. 7.9). If the chords AC and AB are given, draw the diameter AD, and the chords BC, DB, and CD. The chord AD is known, being the diameter of the circle (hence equal to 120 of Ptolemy's units). Then DB and DC can be computed using the Pythagorean theorem from the diameter and the given chords. Hence in the inscribed quadrilateral $ABCD$ both diagonals and all sides except BC are known, and so BC can be computed.

To construct his table of chords Ptolemy begins with a regular decagon inscribed in a circle. The central angles subtended by the sides of this decagon are $36°$. Because of the compass-and-straightedge construction of this figure, its side can be expressed as $\sqrt{4500} - 30$ (when the radius has length 60). Instead of repeatedly bisecting this angle, however, Ptolemy adopts an indirect strategy to find the chord of a smaller angle without having to repeat so many square roots. He uses the fact that the side of the regular pentagon inscribed in a circle (the chord of $72°$) is known from Euclid, Book XIII, Proposition 10 to be the hypotenuse of the right triangle whose legs are the radius of the circle and the side of the inscribed regular decagon. Thus this chord is approximately 70; 32, 3. Since the chord of $60°$ is obviously 60, one can then use Ptolemy's theorem to compute the

[4] In fact the algorithm by which hand calculators evaluate the trigonometric functions works roughly along these lines. Certain values are hard-wired into the calculator and others are computed by application of the addition formulas.

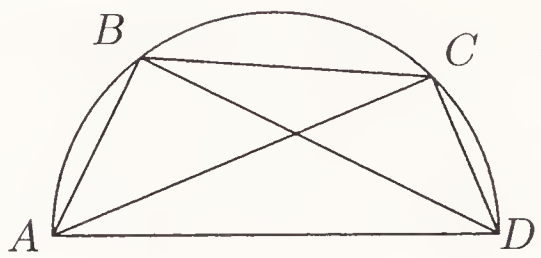


Figure 7.9: Difference of two chords.

chord of $72° - 60° = 12°$. Ptolemy then shows how to compute the chord of half an angle if the chord of the angle is known. (This is easier than deriving the trigonometric functions for half-angles: try it yourself!) In this way he is able to compute successively the chords of $6°$, then $3°$, then $1° 30'$, and finally $0° 45'$. The ingenious idea of starting from a $72°$ angle, rather than the more natural $60°$ angle, allowed Ptolemy to reach angles less than $1°$ while minimizing the roundoff error caused by approximating square roots.

Ptolemy's construction of his table misses the important angle of $1°$. This gap is not accidental: all the angles that can be found by his strategy can be constructed with compass and straightedge, but a $1°$ angle is not constructible with these instruments alone. In order to estimate the chord of $1°$, Ptolemy combines the two chords on each side of $1°$, namely $1° 30'$ and $0° 45'$ with a useful approximation theorem: *The ratio of the larger of two chords to the smaller is less than the ratio of the arcs they subtend.* Because of this proposition the chord of $1°$ is less than $\frac{4}{3}$ the chord of $0° 45'$, yet larger than $\frac{2}{3}$ the chord of $1° 30'$. In this way Ptolemy was able to establish that the chord of $1°$ is approximately $1; 2, 50$ units (where the radius is 60 units). Then, using his half-angle formula, he finds the chord of $0° 30'$, after which he is able to construct a table of chords for angles at half-degree intervals.[5]

The table of chords makes it possible to solve right triangles, in particular, to find the angles in such a triangle when given the ratio of its sides. In astronomy, however, one is always using angular coordinates on a sphere, since both the sides and angles of a spherical triangle are given as angles. It would be clumsy always to have to introduce plane triangles in order to find the parts of spherical triangles, and so Ptolemy included certain relations among the parts of spherical triangles as lemmas. These are not the laws of cosines and sines now used in spherical trigonometry, but rather two theorems that had been published half a century earlier in a work called *Sphaerike* by a certain Menelaus. With these relations it is possible to solve such problems as finding which portion of the ecliptic rises simultaneously with a given portion of the celestial equator, for example.

With this mathematical equipment and a wealth of observational data, Ptolemy was able to apply the theoretical methods invented by earlier astronomers such as Apollonius. The 12 books of the *Almagest* became the standard astronomical

[5]Two minor points may be noted in connection with this work. First, Ptolemy's use of sexigesimal notation is only partial; he does not write $1° 30'$, as we have done, but rather $1\frac{1}{2}°$. Second, in establishing the approximation for the chord of $1°$ he writes absurdly that "the chord of $1°$ was shown to be both greater and less than the same amount." But we know what he means.

treatise over a large part of the world until the seventeenth century.

7.4 Problems and Questions

7.4.1 Problems in Hellenistic Mathematical Science

Exercise 7.1 The law of the lever is usually not proved in physics textbooks today. Rather the notion of *moment* about an axis is introduced. Confining ourselves to forces all lying in a single plane, we can define this moment as the product of the distance from the point at which the force is applied to the axis and the component of the force perpendicular to the line joining the point of application to the axis, that is, $\Omega = Fr$, where Ω is the moment, F is the force, and r the distance. The moment is given a positive sign if it tends to produce counterclockwise rotation, and negative otherwise. One of the postulates of mechanics is that a body is in rotational equilibrium if the algebraic sum of the moments is zero. Apply this definition to the case when the body is a lever with two weights hung from its ends and show that it is exactly equivalent to the law of the lever as stated by the author of the *Mechanics* and by Archimedes.

Exercise 7.2 Prove that given any three magnitudes of the same kind that can be divided into arbitrarily small pieces (say, A, B, and C) with $A < B$, there is a fourth quantity D of the same kind commensurable with C and such that $A < D < B$. [*Hint:* By taking away half of C, then half of what is left, etc., in a finite number of steps one will obtain a magnitude E commensurable with C and smaller than the difference between B and A. Then some multiple of E will exceed A. Now show that the *smallest* multiple of E that exceeds A lies between A and B and is commensurable with C.]

Exercise 7.3 Give a proof of the law of the lever for incommensurable magnitudes without using the principle invoked by Archimedes (stated in the previous problem). That is, use only the definition of proportion as given by Eudoxus. [*Hint:* Suppose A and B are the two magnitudes and CD and CE are lengths such that A and B balance when suspended from D and E, respectively. You wish to show that $CE : CD = A : B$. Imagine that m and n are integers such that $mA > nB$. You need to show that $mCE > nCD$. Use the commensurable case and Archimedes' postulates to do this.]

Exercise 7.4 Using modern trigonometry, one can express Proposition 8 of Euclid's *Optics*, proved above, as an inequality between certain trigonometric functions of the two angles TEZ and TEH in Fig. 7.5. Write out this inequality. History books sometimes say (to save space) that Euclid proved this inequality. What are the advantages and disadvantages of describing Euclid's work this way? Would Euclid have recognized this description of his work?

Exercise 7.5 Describe the point where a ray of light parallel to the axis of a hemispherical mirror before reflection will strike the axis of the hemisphere after reflection.

Exercise 7.6 Diocles solved the problem of two mean proportionals using a curve that was later called the *cissoid* (from the Greek word for ivy). In terms of our analytic geometry the equation of this curve is $(a-x)^3 = y^2(a+x)$, where a is a given number. How can this curve be used to construct a line segment of length, say, $\sqrt[3]{3}a$?

Exercise 7.7 Consider the following observations: (1) an ellipse is the locus of points the sum of whose distances from two fixed points is constant (Apollonius' *Conics*, Book III, Proposition 52); (2) therefore the inside of the ellipse consists of points such that the sum of the distances is less than that constant, and the outside consists of points such that the sum of the distances is greater than that constant. Now consider the tangent at any point of the ellipse. Since the ellipse is convex, it lies entirely on one side of the tangent. Now of the rays emanating from one fixed point, meeting the tangent, and then passing to the other fixed point, all of them except the one that travels to the point of tangency must go outside the ellipse. Therefore the shortest such path is the one that goes to the point of tangency. Now use Heron's shortest-path criterion to derive a proof of the focal property of ellipses (Apollonius' *Conics*, Book III, Proposition 48): A ray of light emanating from one focus of an ellipse will be reflected to the other focus.

Exercise 7.8 According to Snell's law, the angle of incidence θ_1 and the angle of refraction θ_2 for light passing from one medium to another are related by

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2},$$

where v_1 and v_2 are the velocities of light in the respective media. For water and air the ratio of these two velocities is about $3:4$. Therefore the modern theoretical relationship is roughly $\theta_2 = \arcsin(0.75\sin\theta_1)$ instead of the value $\theta_2 = \frac{33}{40}\theta_1 - \frac{1}{400}\theta_1^2$ implied by Ptolemy's table. Notice that Ptolemy's $\frac{33}{40}$ is just 10% larger than the ratio of the two velocities. Computing the table of angles of refraction from Snell's law, we find, to the nearest half-degree, the following table.

Angle of Incidence	Angle of Refraction
10°	$7\frac{1}{2}^\circ$
20°	15°
30°	22°
40°	29°
50°	35°
60°	$40\frac{1}{2}^\circ$
70°	$45\frac{1}{2}^\circ$
80°	$47\frac{1}{2}^\circ$

Compare this table with the one given by Ptolemy. How good is the agreement? Compute the first few terms of the power series for θ_2 in terms of θ_1 and show that Snell's law says that

$$\theta_2 \approx \tfrac{3}{4}\theta_1 - \tfrac{19}{64}\theta_1^3.$$

Does this agreement reflect accurate observation on Ptolemy's part, or just good intuition?

Exercise 7.9 How do you account for the precession of the equinoxes in terms of the heliocentric theory now taught? How does the heliocentric theory account for the variation in the daily progress of the sun along the ecliptic? How does it account for the fact that the ecliptic itself wobbles slightly?

Exercise 7.10 The precession of the equinoxes can be explained by saying that the celestial equator rotates among the fixed stars. Does this mean that the earth's equator moves around on the earth also?

Exercise 7.11 Ptolemy's trigonometry refers continually to the chord of twice an arc. Let us denote the chord of twice an arc of x degrees by $C(x)$. Assuming the radius of the circle is 60 units long, show that $C(90^\circ - x) = \sqrt{(120)^2 - (C(x))^2}$.

Exercise 7.12 At present the sun reaches apogee in early July. Just to get definite numbers, we shall take some observations from 1964. On July 1 of that year the sun was 99.19° along the ecliptic, measured from the vernal equinox. Consider Ptolemy's model with the sun on an epicycle whose radius is $\frac{1}{24}$ times the radius of its deferent and assume that the sun and the centers of the epicycle and deferent are in a straight line on July 1. How far along the ecliptic will the sun be 123 days later, on November 1? The center of the epicycle progresses uniformly by $\frac{360}{365.24}$ degrees per day, and the sun moves backward on the epicycle at exactly the same rate. (The result obtained from this model is very close to the observed value, within $18'$ of arc, to be precise.)

Exercise 7.13 Another application of spherical trigonometry is in finding the location of sunrise and the length of day at different times of the year in different geographical locations. For this application one needs to consider the arcs on the celestial sphere shown in Fig. 7.10. (This figure is slightly distorted in order to give us a peek at the arc $\widehat{EQ}$, which is below the horizon.) In this figure N and E are respectively the northern and eastern points on the horizon, so that arc $\widehat{NE}$ is a 90° arc. Assume that the sun is north of the celestial equator (that is, the season is either spring or summer in the northern hemisphere). The point P is the north celestial pole, and $\widehat{PQ}$ is the 90° arc from the pole through the sun (S) to the equator, and $\widehat{EQ}$ is an arc of the celestial equator. We shall use the symbol λ for the geographic colatitude of the observer (the distance in degrees from the north celestial pole P to the point directly overhead) and $\sigma := \widehat{SQ}$ for the length of the arc from the sun perpendicular to the celestial equator, called the *declination* of the sun. We wish to know the point on the horizon at which the sun will rise, that is, the arc $\gamma := \widehat{SE}$. One can show through trigonometry that

$$\sin\gamma = \frac{\sin\sigma}{\sin\lambda}.$$

Show that this equation can be solved for γ if and only if $|\sigma| \leq \min(\lambda, 180^\circ - \lambda)$.

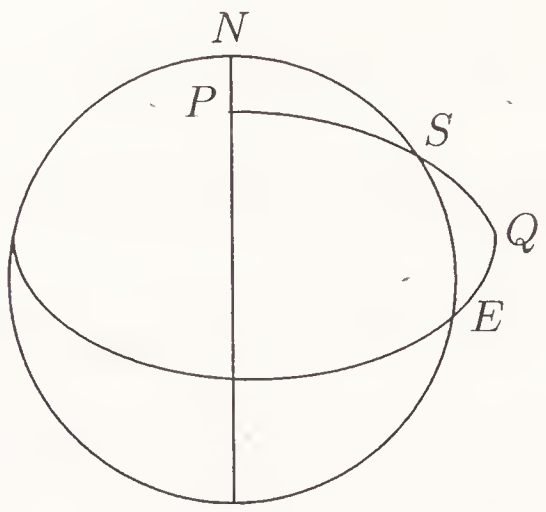


Figure 7.10: Local direction of sunrise.

Consider as an example the summer solstice (June 21), when $\sigma = 23^\circ\, 30'$, and show that in Stockholm ($\lambda = 30^\circ$) the sun rises about 53° north of east on this date.

Exercise 7.14 Ignoring the motion of the sun along the ecliptic in the course of one day (which is small), one can say that the sun describes a circle in the sky of radius $R\cos\sigma$ on a day when its declination is σ, where R is the radius of the celestial sphere (it can be taken as 1). The chord on this circle that separates the portion above the horizon (local day) from the portion below the horizon (local night) is the chord joining the two points of intersection of this circle with the horizon circle. Hence by the preceding exercise, that chord is of length $2R\cos\gamma$. Deduce that the arc of the sun's path above the horizon (in the spring and summer) is of length $180^\circ + 2\alpha$, where

$$\cos\alpha = \frac{\cos\gamma}{\cos\sigma}.$$

Compute that the center of the sun should be above the horizon in Stockholm for about 18 hours and 30 minutes on June 21. Give two independent refinements to this model, each of which helps to explain why daylight lasts longer than this computed value.

Exercise 7.15 Given three points on a circle A, B, and C bounding arcs such that $\overset{\frown}{AB} > \overset{\frown}{BC}$, prove that $\overset{\frown}{AB} : \overset{\frown}{BC} > AB : BC$. Use this result to show that the chord of 1° is larger than $\frac{2}{3}$ of the chord of $1^\circ\, 30'$ and smaller than $\frac{4}{3}$ of the chord of $0^\circ\, 45'$.

7.4.2 Questions about Hellenistic Mathematical Science

Exercise 7.16 What differences are there in the approach to physics by Aristotle and by the mathematically inclined scientists such as Archimedes, Diocles, and Ptolemy? Do these approaches differ in their effectiveness? Is a theory constructed

along the lines of one approach more effective or "better" than one constructed in the other way?

Exercise 7.17 The logic of Archimedes' solution to the problem of analyzing the king's crown is as follows. If the crown is of pure gold, it will have the same specific gravity as pure gold (and hence the crown will displace the same amount of water as an equal weight of pure gold). By observation, the crown displaces more water than an equal weight of pure gold. Therefore, it is not of pure gold. Hence it would seem that Archimedes had no need to compare the amount of water displaced by an equal weight of silver. What then is the relevance of the silver in the story reported by Vitruvius?

Exercise 7.18 In what sense is Euclid using the word "appear" when he says that all diameters of a circle whose plane is perpendicular to the line of sight will appear equal? Draw a horizontal line 5 centimeters long and a vertical line exactly the same length. Do these lines *appear* to be equal?

Exercise 7.19 Considering how heavily we have relied on the work of Neugebauer for our summary of early astronomy, it is only fair to let the reader know Neugebauer's opinion of some of the other authors whom we have quoted with respect. Here is what he says on p. 572 of his treatise *A History of Ancient Mathematical Astronomy.* Do you agree? If not, how do you refute these opinions?

> One need only read the gibberish of Proclus' introduction to his huge commentary on Book I of Euclid's "Elements" to get a vivid picture of what would have become of science in the hands of philosophers. The real "Greek miracle" is the fact that a scientific methodology was developed, and survived, in spite of the existence of a widely admired dogmatic philosophy. . . .

7.5 Endnotes

1. The discussion of Aristotle's *Physics* and *On the Heavens* is based on the book *The Works of Aristotle Translated into English*, Vol. III, edited by W. D. Ross, (Clarendon Press, Oxford 1930).

2. The discussion of the Aristotelian property of the lever is from the "Mechanical Problems" in *Aristotle. Minor Works*, translated by W. S. Hett (Harvard University Press, Cambridge, 1936), pp. 331–411. The present discussion is based on pp. 331–353.

3. Archimedes' work on the lever and floating bodies is based on *The Works of Archimedes* by T. L. Heath (Dover Reprint, New York, 1953).

4. The discussion of the inclined plane as studied by Heron and Pappus is based on *A History of Mechanics* by René Dugas, translated from the French by J. R. Maddox (Central Book Company, New York, 1955), pp. 32–35.

5. The discussion of Euclid's proof of the nonproportionality of apparent size and distance is based on *Selections Illustrating the History of Greek Mathematics*, with a translation by Ivor Thomas (Harvard University Press, Cambridge, MA, 1936), pp. 502–505.

6. Diocles' treatise can be found in the book *Diocles. On Burning Mirrors*, by G. J. Toomer (Springer-Verlag, New York, 1976).

7. The discussion of Ptolemy's work on refraction is based on the account given by George Gamow in *Biography of Physics* (Harper & Brothers, New York, 1961), pp. 20–22.

8. The discussion of early astronomy is based partly on Neugebauer's *History of Ancient Mathematical Astronomy* (3 vols.), Springer-Verlag, New York, 1975, and partly on the article by A. Aaboe, "Scientific astronomy in antiquity," which was published in *The Place of Astronomy in the Ancient World*, F. R. Hodson, ed. (Oxford University Press, London, 1974), pp. 21–43.

9. The quotation from Ptolemy's *Almagest* is taken from the edition by G. J. Toomer (Springer-Verlag, New York, 1984), p. 191.

Chapter 8

Mathematics in the Roman Empire

8.1 Introduction

The Greeks came into contact with the mathematical achievements of earlier centuries through the commercial activities of their colonies on the coast of Asia Minor. The foundations of deductive mathematics were laid in the Greek colonies of southern Italy and further developed on the mainland of Greece. When the Macedonian Empire of Philip and Alexander conquered Greece, these mathematical seedlings were transplanted to the East as far as India, and the city of Alexandria in Egypt became the most important mathematical center in the world. Alexandria was not the only center of excellence, however. The presence of such mathematicians as Archimedes in Sicily, Apollonius in Asia Minor, and Diocles in the Peloponnesus, each familiar with the work of other mathematicians in other places, shows that a wide network of mathematicians was in existence in Hellenistic times. Despite the turbulence of the Hellenistic era, these mathematicians managed to stay aware of the current state of research. One might have expected that an end to the political instability would usher in an era of expansion of mathematical research, and it is therefore surprising that events did not turn out that way.

The Roman expansion began with the incorporation of the Greek cities of southern Italy around the time of Euclid. The Romans took sides in the internecine disputes of these cities. Their intervention provoked a countermove by King Pyrrhus of Epirus, whose "victories" over the Romans cost him so dearly that he decided to abandon Italy and mount an expedition to defend the Greek cities of Sicily against the Carthaginians. At this point the Romans formed an alliance with the Carthaginians. They finally defeated Pyrrhus in 275 B.C.E. (Pyrrhus' name has become traditional; a *Pyrrhic victory* is one that costs more than it gains.) A decade later the Romans found themselves embroiled in a dispute in Sicily, in which King Hieron II of Syracuse was allied with the Carthaginians. This dispute provoked the First Punic War (the word *Punic* is the Latin word corresponding

to *Phoenician*). After early Roman successes this war dragged on for more than twenty years before the Romans defeated the Carthaginians in a sea battle off the western coast of Sicily. This war brought the Romans control of most of Sicily, the city of Syracuse being one of the exceptions. That exception was to have significant consequences for the greatest mathematician of antiquity, as we shall see.

Roman expansion led to further conflict with the Carthaginians, who were attempting to recoup their economic losses with colonies in southern Spain. The Romans seized the island of Sardinia at a time when the Carthaginians were in no position to do anything about this outright piracy. By 219 B.C.E. both sides were spoiling for a fight, and they found no difficulty in concocting a cause. The Carthaginian leader Hannibal moved his army across the Alps, winning major victories for a few years. In the end, however, the "Fabian" tactics of the Roman general Q. Fabius Maximus gained the victory for Rome in this war of attrition. The Carthaginians surrendered in 202 B.C.E.

The expansion of Rome accelerated after the Roman victory in the Second Punic War. We have mentioned this war previously only because it caused the death of Archimedes. Although Rome was victorious, the victory was a very narrow thing. The Romans had suffered catastrophic defeats by Hannibal in such disasters as the Battle of Cannae; but in the end, because of victories by Scipio Africanus and Q. Fabius Maximus, they prevailed. The Phoenician base at Carthage in North Africa remained a threat, however. To eliminate that threat the Romans undertook the pre-emptive Third Punic War (150–146 B.C.E.) in which Carthage was utterly destroyed. Having eliminated their only serious challenger, the Romans began to conquer more and more territory. The political fragmentation of the Mediterranean world under the successors of Alexander had led to a long period of unrest, with continual small wars being waged. The Roman expansion brought stability, which facilitated further expansion. The mainland of Greece was incorporated into the Roman Empire in the two generations following the Third Punic War. Egypt became a Roman province under Julius Caesar in the time of Cleopatra, the last ruling descendant of Alexander's general Ptolemy.

Under Caesar Augustus, with the Pax Romana firmly established in the entire Mediterranean world, one would expect conditions to be ideal for peaceful intellectual activities such as mathematics. The fact is, however, that the mathematics produced under the Roman Empire was much less original and profound than that which had been produced during the earlier periods of political chaos. The scholars at Alexandria and Athens, for some reason, stopped advancing knowledge with original research and spent their time commenting on and elaborating the work of the great mathematicians of the past. The reasons for this decline are a subject of speculation. Nevertheless, the four centuries from the time of Augustus to the sack of Rome were not entirely barren. They were filled with the work of the commentators listed in Chapter 4 and with a few original thinkers worthy of study. We shall examine the works of two of these mathematicians, Diophantus (dates uncertain, but probably third century C.E.) and Pappus (ca. 300 C.E.), both of whom worked at Alexandria, and we shall briefly mention some other late commentators, including the philosopher/mathematician Hypatia, the only woman mathematician

from the ancient world about whose life some details are known.

8.2 Diophantus

Little is known about Diophantus of Alexandria, other than the place where he worked. His dates are known only approximately. In his book on polygonal numbers he refers to the writer Hypsicles, who is known to have flourished around 150 B.C.E. Diophantus, in turn, is referred to by the commentator Theon of Alexandria, who lived in the fourth century. Because the third-century Bishop Anatolius of Laodicea dedicated a book to a person named Diophantus, it is conjectured that Diophantus flourished about 250 C.E.

Two works of Diophantus have survived in part, the treatise on polygonal numbers mentioned above and the work for which he is best known, the *Arithmetike*. Like many other ancient works, including those of Euclid, Apollonius, Ptolemy, and the commentators Proclus and Theon, these managed to survive because of the efforts of a ninth-century Byzantine mathematician named Leon, who organized a major effort to copy and preserve these works. There is little record of the influence the works of Diophantus may have exerted before this time.

According to the introduction to the *Arithmetike*, this work consisted originally of 13 books, but until recently only 6 were known to have survived; it was assumed that these were the first 6 books, on which Hypatia (the daughter of Theon of Alexandria) wrote a commentary. However, recently 4 more books have been found in Arabic manuscript, which proved from internal evidence to be books 4 through 7. It thus appears that we now have the first 7 books and 3 others. As is usual in such cases, the oldest extant manuscripts date to Medieval times, specifically to a monk named Maximus Planudes, who lived in Byzantium at the end of the thirteenth century, and wrote a commentary on Books I and II.

8.2.1 Characteristics of Diophantus' Algebra

The *Arithmetike* is the earliest treatise that is recognizable as algebra, and it earned Diophantus the name "Father of Algebra." Still, it is not algebra as taught in high schools or universities today. The similarities and differences between Diophantus and modern algebra are almost equally noticeable. We begin by describing the similarities:

1. *Symbolism.* Diophantus began by introducing a symbol for a constant unit $\overset{o}{M}$, from *monas* ($\mu o \nu \acute{\alpha} \varsigma$), along with a symbol for an unknown number ς, conjectured to be an abbreviation of the first two letters of the Greek word for number: *arithmos* ($\grave{\alpha} \rho \iota \theta \mu \acute{o} \varsigma$). For the square of an unknown he used Δ^{υ}, the first two letters of *dynamis* ($\Delta \acute{\upsilon} \nu \alpha \mu \iota \varsigma$), meaning *power*. For its cube he used K^{υ}, the first two letters of *kybos* ($K \acute{\upsilon} \beta o \varsigma$), meaning *cube*. He then combined these letters to get fourth ($\Delta^{\upsilon}\Delta$), fifth (ΔK^{υ}), and sixth ($K^{\upsilon}K$) powers. For the reciprocals of these powers of the unknown he invented names by adjoining the suffix *-ton* ($-\tau o \nu$) to the names of the corresponding

powers. These various powers of the unknown were called *eida* (ἐίδα), meaning *species*. Diophantus' system for writing down the equivalent of a polynomial in the unknown consisted of writing down these symbols in order to indicate addition, each term followed by the corresponding number symbol (for which the Greeks used their alphabet). Terms to be added were placed first, separated by a pitchfork (⫛) from those to be subtracted. T. L. Heath conjectured that this pitchfork symbol is a condensation of the letters lambda and iota, the first two letters of the Greek root meaning *less* or *leave*. Thus what we would call the expression $2x^4 - x^3 - 3x^2 + 4x + 2$ would be written $\Delta^{\upsilon}\Delta\bar{\beta}\varsigma\bar{\delta}\ \overset{o}{M}\ \bar{\beta}\ ⫛\ K^{\upsilon}\bar{\alpha}\Delta^{\upsilon}\bar{\gamma}$.

2. *Emphasis on equations.* The central idea in all of Diophantus' problems is the use of descriptions of an unknown number to determine the number. The procedure is analogous to replacing the name "George Washington" with the descriptive phrase "first President of the United States." Since these descriptive expressions represented numbers, it was necessary to tell how to add and multiply them. Diophantus provides rules for such manipulations, explaining that a subtracted term multiplied by a subtracted term produces an added term. (Since he did not have the concept of zero or negative numbers, Diophantus could speak only of terms *subtracted from* other terms.) Whether this step amounts to an explicit invention of equations can be debated. To study this question it is necessary to look at the original Greek in which Diophantus wrote. For example, a standard translation of part of the Preface to Book I of the *Arithmetike* reads as follows:

> Next, if there result from a problem an equation in which certain terms are equal to terms of the same species, but with different coefficients, it will be necessary to subtract like from like on both sides until one term is found equal to one term.

A more literal translation of the same passage goes as follows:

> Now in this way, if from a certain problem some species become equal to the same species, but not in the same multitude, it is necessary to subtract the same from each of the parts until one species becomes equal to one species.

Clearly the *idea* of an equation is present here. Diophantus has rules for manipulating equations; and if his abbreviation for *equals* (ἴσ.) is replaced by our "=," we can see that the translator does not distort very much in replacing Diophantus' expression

$$\varsigma\bar{\eta}\ \overset{o}{\mathrm{M}}\ \bar{\delta}\ \text{ἴσ.}\ \square^{\omega}$$

by

$$8x + 4 = \text{a square.}$$

Notice, however, that the equals sign here really means that the expression *belongs to a certain class*. This equation does not express the equality of two numbers.

3. *Subject matter.* The problems in the *Arithmetike* consist of finding one or more unknown numbers satisfying certain conditions. Frequently the condition is given as the result of applying certain operations to those numbers. The first problem of the treatise, for example, is *to separate a given number into two* [*other numbers*] *having a given difference*. In other words, to find two numbers given their sum and difference. This type of problem, as we argued in Chapter 3, is the essence of algebra. Moreover the problems are stated by formally equating (in words) two different expressions for the same unknown number, as we have just seen. Thus the equation becomes the central *tool* for solving problems.

4. *Algorithmic techniques.* Problems in Diophantus are usually solved by beginning with the statement that two formally different descriptions represent the same number. Diophantus gives the rules for transposition and cancellation to make it possible to convert such statements into simpler statements of the same type. For example, for the problem just mentioned Diophantus illustrates the general procedure with a single example in which the sum is 100 and the difference is 40. Taking the smaller of the two numbers as the unknown, he observes that twice the unknown plus 40 must equal 100. That is, the larger number must be the unknown plus 40, and since the sum of the larger and the smaller numbers must be 100, he has, except for the unimportant difference between his notation and ours, the equation $2x + 40 = 100$. Diophantus has already stated the rules for solving such an equation in his introduction; without repeating himself on this point, he merely notes that $x = 30$, so that the required numbers are 30 and 70.

The differences between the *Arithmetike* and modern algebra are also quite noticeable, however.

1. *Restricted symbolism.* Diophantus' use of symbolism is rather sparing by modern standards; he often uses words where we would use symbolic manipulation. For this reason his algebra was described by the nineteenth-century German historian of mathematics G. H. F. Nesselmann (1811–1881) as an intermediate "syncopated" phase between the earliest "rhetorical" algebra, in which everything is written out in words, and the modern "symbolic" algebra. The peculiarity of the "syncopated" phase is that the symbols are abbreviations for words, rather than ideograms like our modern symbols $+$ and $\div$.

   A further point to be noted in connection with the restricted symbolism is that Diophantus did not handle equations with unspecified coefficients. That is, he did not write the analog of $ax^2 + bx + c = 0$ to discuss a general quadratic equation (and of course he didn't classify his equations as linear and quadratic anyway). The whole Greek alphabet, plus three additional

symbols, was used for writing the specific numbers $1, \ldots, 9$, $10, \ldots, 90$, and $100, \ldots 900$. He would therefore have had to invent still more symbols and rules of manipulating them in order to conduct a discussion on the modern level of abstraction. He also never employed a symbol for a second unknown, even though many of his problems require finding several unknown numbers. This notational restriction led to some constraints on his methods of solution.

2. *Restricted role of equations.* For Diophantus the equation is a basic tool, but it is not itself the subject of investigation. He investigates different types of *problems* but does not classify *equations*. This difference from modern high-school algebra is related to the fact that he cannot handle generic equations (with parameters for coefficients).

3. *Indeterminate problems.* Many of Diophantus' problems have an infinite family of solutions. Instead of seeking unknown numbers when given the result of performing certain operations on them, Diophantus frequently seeks numbers *of a certain form* or *satisfying certain conditions*. For example, Problem 11 of Book II is *to add the same number to two given numbers so as to make each of them a square.* A problem of this sort is not an equation.

4. *Restricted concept of number.* Diophantus' concept of a number is the strict Greek notion, corresponding to what we now call a positive rational number. This requirement forces Diophantus into the activity known as *diorismos*, that is, stating restrictions on the allowable data for a problem so as to ensure that a positive rational solution exists. The geometric equivalent of the problem of finding two numbers given their sum and product (applying a rectangle equal to a given area to a line in such a way that the defect is a square) has a solution only when the given area is at most equal to the square on half of the line. This restriction was the diorismos for this problem. No diorismos is necessary for the geometric equivalent of finding two numbers given their difference and product (where the square is an excess rather than a defect), since there is always a solution to the geometric problem. However, the arithmetical version of the problem *does* require a diorismos, since the positive solution can be found only by taking a square root, an operation that may lead outside the class of rational numbers. This problem happens to be Problem 30 of Book I. Since the numerical solution involves taking the square root of the sum of four times the product and the square of the difference, the diorismos requires that this last number be a square. In illustrating the procedure by example Diophantus chooses data satisfying this condition (product 96, difference 4, so that the procedure involves finding the square root of 400).

5. *Incomplete solutions.* Requiring positive rational solutions sometimes has the effect of reducing an infinite class of solutions of an indeterminate problem to a finite class. The fact that the solutions Diophantus sought had to be positive rational numbers distinguished his work from the so-called geometric algebra found in Euclid. This restriction has remained in mathematics and

caused the name of Diophantus to be attached to the problem of finding all rational solutions to an equation in more than one unknown. For example, there are no positive rational numbers x, y, and z satisfying the equation

$$x^4 + y^4 = z^2.$$

A *Diophantine equation* is not intrinsically different from any other equation. What makes an equation "Diophantine" is the fact that only rational (or integer) solutions are considered. It should be noted, however, that Diophantus does not try to find *all* solutions to his problems. He produces one solution as a representative sample of the method and then leaves it to the reader to look for others.

6. *Connection with number theory.* Because the existence or nonexistence of rational solutions to an indeterminate equation depends on the arithmetic properties of the coefficients, such equations relate algebra to number theory. For example, the number of ordered pairs of integers (x, y) satisfying $x^2 + y^2 = n$, $y \geq 0$, and $x > 0$ equals the difference between the number of divisors of n that leave a remainder of 1 when divided by 4 and the number of divisors of n that leave a remainder of 3 when divided by 4.[1]

8.2.2 Contents of the *Arithmetike*

In view of the recent discovery of four lost books mentioned above, prevailing beliefs about the arrangement of problems in the *Arithmetike* had to be modified. Even with this new information the division of the work into separate books is difficult to explain in terms of the contents or methods of solution. There seems to be very little difference, for example, between the last problem of Book II and the first problem of Book III. The former (Problem 35 of Book II) is *to find three numbers such that the square of any one decreased by the sum of all three gives a square.* The latter is *to find three numbers such that the sum of all three decreased by the square of any one of them gives a square.* Perhaps if we had a complete pristine version of the *Arithmetike* free of all interpolations and commentaries, we might see the pattern that Diophantus himself insists is present. He says in his introduction that the different types of problems proceed from simple to complex, which is true, and are distinguished from one another (one wonders how, since there are no subheadings or explanations to tell when one type ends and another begins). Whatever the organizing principle of the work, if we adhere to a crude distinction between determinate and indeterminate problems, we can describe the arrangement of topics in the *Arithmetike* roughly as consisting of a small set of determinate problems, which are pure algebra, followed by a large number of indeterminate ("Diophantine") problems, which link algebra and number theory.

[1] To illustrate this fact, consider for example $n = 450$, which has the divisors 1, 2, 3, 5, 6, 9, 10, 15, 18, 25, 30, 45, 50, 75, 90, 150, 225, and 450. Six of these 18 divisors (1, 5, 9, 25, 45, and 225) leave a remainder of 1 when divided by 4. Three (3, 15, and 75) leave a remainder of 3 when divided by 4. The other 9 leave a remainder of 2. Thus there should be three ordered pairs of positive integers the sum of whose squares is 450, and indeed there are: $(15, 15)$, $(3, 21)$, and $(21, 3)$.

Determinate Problems

The determinate problems in the *Arithmetike* require that one or more unknown numbers be found from conditions that we would nowadays write as systems of linear or quadratic equations. The 39 problems of Book I and the first 10 problems of Book II are of these types. Some of these problems have a unique solution. For example, Problem 7 of Book I is *from a given unknown number, subtract two given numbers so that the remainders have a given ratio.* In our terms, this says

$$x - a = m(x - b),$$

where x is unknown, a and b are the given numbers, and m is the given ratio. Since it is obvious that $m \geq 1$ if all quantities are positive and $a \leq b$, Diophantus has no need to state this restriction.

Some of the problems that are determinate from our point of view may have no positive rational solutions for certain data. In such cases Diophantus requires a diorismos to restrict the data so that positive rational solutions will exist. For example, Problem 8 of Book I is *to two given numbers to add the same unknown number so that the sums have a given ratio.* This problem amounts to the equation

$$x + a = m(x + b).$$

It is clear that if $x > 0$ and $a > b$, then $1 < m = (x + a)/(x + b) < a/b$. This restriction is the diorismos for the problem, that is, the given ratio must be larger than 1 and less than the ratio of the larger number to the smaller.

Some of the earlier problems are indeterminate because not enough conditions are imposed on the data. In such cases Diophantus obtains a determinate problem by adding a new restriction in the course of the solution. For example Problem 25 of Book I requires four numbers such that if each is increased by a given fraction of the sum of the other three, the four results are equal. Setting four numbers equal to one another gives only three equations. Diophantus assumes a value for the sum of the second, third, and fourth numbers, then uses successive elimination to express the second, third, and fourth unknown numbers in terms of the first.

In all cases when a problem requires more than one unknown to be found, more than one condition must be specified. At these places we can detect a systematic method in the *Arithmetike*. In the case of two unknown numbers, for example, Diophantus always assumes that the required numbers are expressed in terms of the one symbol he has for an unknown in a form that makes one of the conditions automatically true. The remaining condition can then be written as an equation that can be solved using the rules of manipulation stated in the introduction. For example, Problem 15 of Book I is *to find two numbers such that each after receiving from the other a given number shall bear to the remainder a given ratio.* We would express this problem as

$$\frac{x + a}{y - a} = r, \quad \frac{y + b}{x - b} = s.$$

Diophantus takes the data to be $a = 30$, $r = 2$, $b = 50$, $s = 3$. He then introduces the unknown ς, as follows. The second number (what we called y) is assumed to

be $\varsigma + 30$ and the first unknown number (x) is $2\varsigma - 30$, so that first condition is automatically satisfied. Then by the second condition, $\varsigma + 80 = 3(2\varsigma - 80)$, which leads to $\varsigma = 64$, and so the two numbers are 98, that is, $2\varsigma - 30$, and 94, that is, $\varsigma + 30$.

Indeterminate Problems

The problems of Book I are either determinate or are made so by adding new restrictions in the course of the solution. Where these problems are indeterminate the indeterminacy is the result of an insufficient number of restrictions on the output. The problems are followed in Book II by some problems that are indeterminate for a different reason. In these problems the result of applying certain operations to unknown numbers is not specified as a number, but is required only to belong to given classes of numbers, usually square numbers and occasionally cubes. A famous example of this type is Problem 8 of Book II, *to separate a given square number into two squares.* Diophantus illustrates this problem using the number 16 as an example. His method in this indeterminate problem is similar to that followed in the determinate problem discussed above. That is, he expresses the two numbers in terms of his single unknown in such a way that one of the conditions is automatically satisfied. Thus letting one of the two squares be Δ^υ (which is ς^2 in our terms) the other will automatically be $16 - \varsigma^2$. To get a determinate equation for ς, he assumes that the other number to be squared is 4 less than a multiple of ς. (The number 4 is chosen because it is the square root of 16. In our terms, it leads to a quadratic equation, one of whose roots is zero, so that the other root can be found by solving a linear equation.) Illustrating the general procedure by a particular example, he assumes that the other square is $(2\varsigma - 4)^2$. Since this number must be $16 - \varsigma^2$, he thus finds that $4\varsigma^2 - 16\varsigma + 16 = 16 - \varsigma^2$, so that $\varsigma = \frac{16}{5}$. It is clear that this procedure can be applied very generally, showing an infinite number of ways of dividing a given square into two other squares.

In the seventeenth century this particular problem achieved a fame far beyond its intrinsic importance when Fermat, who was studying the *Arithmetike*, remarked that the analogous problem for cubes and higher powers had no solution, that is, one cannot find positive rational numbers x, y, and z satisfying $x^3 + y^3 = z^3$ or $x^4 + y^4 = z^4$, etc. Fermat stated that he had found a proof of this fact, but unfortunately did not have room to write it in the margin of the book. Fermat never published any general proof of this fact, although certain special cases such as $n = 3$ and $n = 4$ are consequences of a method of proof developed by Fermat and known as the method of infinite descent. The problem was a tantalizing one because of its comprehensibility. Anyone with a high-school education in mathematics can understand the statement of the problem, and probably the majority of mathematicians dreamed of solving it when they were young. Despite the efforts of hundreds of amateurs and prizes offered for the solution, no correct proof was found for more than 350 years. On June 23, 1993 the British mathematician Andrew Wiles announced at a conference held at Cambridge University that he had succeeded in proving a certain conjecture in algebraic geometry known as the Shimura–Taniyama conjecture, (see Fig. 8.1) from which Fermat's conjecture is

known to follow. This is the first claim of a proof by a reputable mathematician using a technique that is known to be feasible, and the result was tentatively endorsed by other mathematicians of high reputation. After several months of checking some doubts arose. Wiles had claimed in his announcement that certain techniques involving what are called Euler systems could be extended in a particular way, and this extension proved to be doubtful. In collaboration with another British mathematician, Richard Taylor, Wiles eventually found an alternate approach that simplified the proof considerably, and it is now believed by experts in number theory that the problem has been solved.

To give another illustration of the same method, we consider the problem following the one just discussed, that is, Problem 9 of Book II: *to separate a given number that is the sum of two squares into two other squares*. (That is, given one representation of a number as a sum of two squares, find a new representation of the same type.) Diophantus shows how to do this using the example $13 = 2^2 + 3^2$. He lets one of the two squares be $(\varsigma + 2)^2$ and the other $(2\varsigma - 3)^2$, resulting in the equation $5\varsigma^2 - 8\varsigma = 0$. Thus $\varsigma = \frac{8}{5}$, and indeed $\left(\frac{18}{5}\right)^2 + \left(\frac{1}{5}\right)^2 = 13$.

Some of Diophantus' indeterminate problems reach a high degree of complexity. For example, Problem 19 of Book III asks for four numbers such that if any of the numbers is added to or subtracted from the square of the sum of the numbers, the result is a square number. Diophantus gives the solutions as

$$\frac{17,136,600}{163,021,824}, \quad \frac{12,675,000}{163,021,824}, \quad \frac{15,615,600}{163,021,824}, \quad \frac{8,517,600}{163,021,824}.$$

8.2.3 Diophantus' Place in Greek Mathematics

Diophantus occupies a unique position in the world of Greek mathematics. Turning away from the great achievement of the Greeks in geometry, he took his subject matter from the very oldest mathematics studied by the Greeks—the properties of the positive rational numbers. Even within this subject, his *Arithmetike* is occupied with problems of the "find a number such that..." type. He does not prove general theorems about figurate numbers or any sophisticated results on primes or divisibility. What is original in Diophantus is his introduction of the technique of writing down equations with a symbol for the unknown number. Later mathematicians wrote commentaries on the *Arithmetike*, but there is nothing else like it among the surviving documents from this period.

8.3 Pappus

Like Diophantus, the second outstanding mathematician from this period is personally obscure. He probably lived at the time of Roman decline, during the reign of the emperor Diocletian (285–305), who split the Empire into eastern and western halves, and Constantine, the first Emperor to convert to Christianity. To judge from his surviving works, Pappus was the ideal of a liberal scholar, well read in the areas of mathematics, astronomy, and geography. He wrote commentaries on

At Last, Shout of 'Eureka!' In Age-Old Math Mystery

By GINA KOLATA

More than 350 years ago, a French mathematician wrote a deceptively simple theorem in the margins of a book, adding that he had discovered a marvelous proof of it but lacked space to include it in the margin. He died without ever offering his proof, and mathematicians have been trying ever since to supply it.

Now, after thousands of claims of success that proved untrue, mathematicians say the daunting challenge, perhaps the most famous of unsolved mathematical problems, has at last been surmounted.

The problem is Fermat's last theorem, and its apparent conqueror is Dr. Andrew Wiles, a 40-year-old English mathematician who works at Princeton University. Dr. Wiles announced the result yesterday at the last of three lectures given over three days at Cambridge University in England.

Bettmann Archive

Pierre de Fermat, whose theorem may have been proved.

Within a few minutes of the conclusion of his final lecture, computer mail messages were winging around the world as mathematicians alerted each other to the startling and almost wholly unexpected result.

Dr. Leonard Adelman of the University of Southern California said he received a message about an hour after Dr. Wiles's announcement. The frenzy is justified, he said. "It's the most exciting thing that's happened in — geez — maybe ever, in mathematics."

Impossible Is Possible

Mathematicians present at the lecture said they felt "an elation," said Dr. Kenneth Ribet of the University of California at Berkeley, in a telephone interview from Cambridge.

The theorem, an overarching statement about what solutions are possible for certain simple equations, was stated in 1637 by Pierre de Fermat, a 17th- century French mathematician and physicist. Many of the brightest minds in mathematics have struggled to find the proof ever since, and many have concluded that Fermat, contrary to his tantalizing claim, had probably failed to develop one despite his considerable

Continued on Page D22, Column 1

Figure 8.1: Front-page story on the proof of Fermat's conjecture, from *The New York Times*, June 24, 1993. Copyright *The New York Times*. The Bettman Archive.

the major works of Euclid, Ptolemy, Apollonius, Archimedes, Heron, and others. These commentaries are not mere explanations of obscure points in the works, but contain extensions of the results of the great scholars. In addition he wrote, around the year 320, an original work known as the *Collection* (*Synagoge*). This work consisted of eight books, but Book I and the first part of Book II have not survived. The part that has come down to us begins in the middle of a description of Apollonius' system of writing numbers in tetrads (similar to Archimedes' *Sand-reckoner*). The *Collection* is not focused, like the *Arithmetike*, on a single topic; it ranges over a variety of topics and gives information about the authors, clarifications of their proofs, and new theorems. It can therefore be read with profit and pleasure by anyone who has a basic acquaintance with the great authors and their works.

8.3.1 Contents of the *Collection*

As mentioned above, Book I has been lost, although it almost certainly contained material similar to that in the surviving part of Book II, that is, on the rather dull topic of notation for writing large integers. The remaining books merit a more detailed discussion. In general, they explore in more detail topics that had been studied by earlier writers. Often they present generalizations of known theorems and raise questions that had not been previously studied.

The Classical Problems and the Pythagorean Theorem

Book III contains a thorough discussion of the problem of two mean proportionals, presenting the solutions of this problem by earlier authors and placing it in the context of the general theory of arithmetic, geometric, and harmonic means. Pappus described a classification of various geometric constructions, which he attributed to "the ancient geometers" as "planar" ($\grave{\epsilon}\pi\acute{\iota}\pi\eta\delta\alpha$, solvable using only circles and straight lines in a plane), "solid" ($\sigma\tau\epsilon\rho\epsilon\acute{\alpha}$, solvable with the use of conic sections, which are surfaces generated in space by circles and straight lines), and "linear" ($\gamma\rho\alpha\mu\mu\iota\kappa\acute{\alpha}$, requiring curves not generated by straight lines and circles, even in three dimensions). Among these more general curves Pappus mentions several, including spirals, the conchoid, and the quadratrix. (We have not discussed these last two curves because of lack of space; they occur frequently as problems in calculus books in connection with polar and parametric equations.) Of the problem of two mean proportionals Pappus says

> Thus, given that problems are distinguished in this way, the geometers of old were not able to construct the solution of the problem of the mean proportionals to two lines by geometric reasoning, it being by nature a solid problem, since conic sections are difficult to draw in a plane. However, they achieved this construction by use of certain wonderful hand instruments.

Book IV contains a famous generalization of the Pythagorean theorem: Given a triangle ABC and any parallelograms $ACFG$ and $BAED$ constructed on two

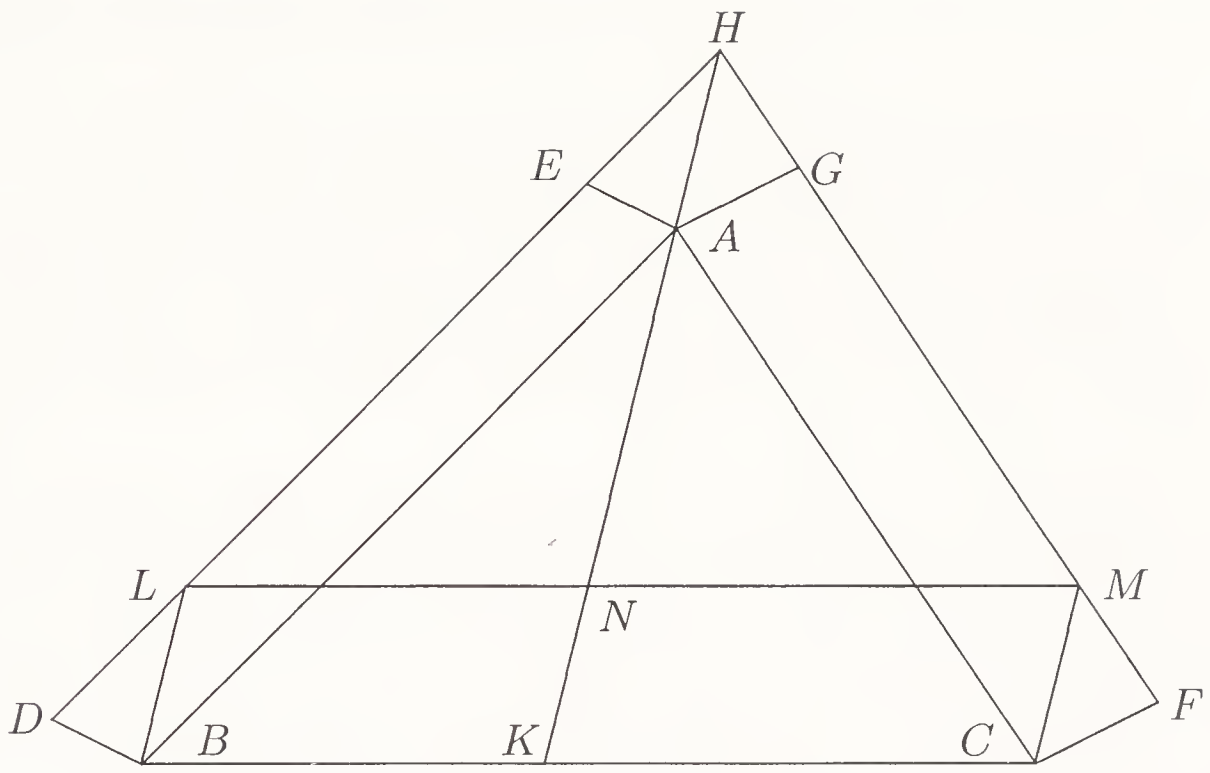


Figure 8.2: Pappus' generalization of the Pythagorean theorem.

sides, it is possible to construct (with straightedge and compass) a parallelogram $BCML$ on the third side equal in area to the sum of the other two (see Fig. 8.2). This is easily done by extending the sides FG and DE to meet at a point H, then drawing HA and extending it to meet BC in the point K. If lines are now drawn parallel to HK through B and C, meeting DE in L and FG in M, then BL and CM are both equal to AH, so that $BCML$ is a parallelogram. Now if LM is joined, meeting AK in the point N, then the parallelogram $CMNK$ equals (in area) the parallelogram $CMHA$, since both have the same base CM and the sides opposite this base both lie on the line HK, which is parallel to CM. But $CMHA$ equals $CFGA$ since both have the base AC and the sides opposite this base lie on the line HF, which is parallel to AC. Hence $CMNK = CFGA$. Likewise $BKNL = BAHL = BAED$, and so $BAED+CFGA = BKNL+KCMN = BCML$.

Book IV also contains a discussion of the other two classical problems, squaring the circle and trisecting the angle. In the course of this discussion Pappus tells much of what is known about the history of these problems and the curves used to solve them. Squaring the circle is a problem of the sort Pappus calls "linear," but trisecting the angle is a "solid" problem, and Pappus shows how to solve it by drawing a suitable hyperbola.

The Isoperimetric Problem

In Book V Pappus takes up a topic not mentioned by Euclid, but apparently discussed by the Athenian mathematician Zenodorus, whose acquaintance with Diocles was mentioned in the previous chapter. This topic is the isoperimetric problem: *Which plane figure of a given perimeter encloses the largest area? Which solid figure having a given surface area encloses the largest volume?* Pappus

introduces this problem with one of the most charming essays in the history of mathematics, one that has frequently been excerpted under the title *On the Sagacity of Bees*. Pappus speaks poetically of the divine mission of the bees to bring from heaven the wonderful nectar known as honey, and says that in keeping with this mission they must make their honeycombs without any cracks through which honey could be lost. Having also a divine sense of symmetry, the bees had to choose among the regular shapes that could fulfill this condition, that is, triangles, squares, and hexagons. They chose the hexagon because a hexagonal prism required the least material to enclose a given volume.

It was apparently Zenodorus who first stated that of all plane figures of given perimeter the circle encloses the greatest area. Pappus shows first that of two regular polygons having the same circumference, the one with the larger number of sides encloses the larger area. He then shows that a circle encloses a larger area than any regular polygon of the same perimeter. In three dimensions he shows that a sphere encloses a greater volume than any regular polyhedron having the same surface area and also greater than any cone or cylinder having the same surface area. Only then does he prove the three-dimensional analog of his comparison of polygons, that is, that if two regular solids have the same surface area, the one with the larger number of faces encloses the larger volume.

Book VI is devoted to commentary on various astronomical treatises.

Locus Problems

Book VII is of historical importance not only because of the thesis advanced in it by Pappus but even more because of the information he provides in illustrating this thesis. The subject of the book is locus problems, such as we have already encountered in Apollonius' *Conics*. In the course of the discussion Pappus mentions a number of Greek works and states the number of propositions each contains, thereby providing us with a list of many now-lost works of Greek geometers. In particular Pappus discusses the three- and four-line locus found in Book III of Apollonius' *Conics*. For these cases the locus is always one of the three conic sections. Pappus says that the locus to five or six lines, which involves a ratio of products of three lines—in other words, a ratio of two volumes—leads to curves which "cannot yet be understood in ordinary reasoning, but are merely called lines (it is not clear what they are and what properties they have)." In our terms these are cubic curves, and a really comprehensive study of them was not completed until the nineteenth century. Indeed, work continues even today on this topic in the abstract setting of modern algebra and number theory, and it was precisely conjectures about cubic curves that led to the proof of Fermat's conjecture.

In connection with this type of problem, Pappus offers an insight with the potential for enormous further advances in mathematics. Considering the locus to more than six lines, Pappus says that these conditions determine a curve. This step was an innovation, since it proposed the possibility that a curve could be uniquely determined by certain conditions without being explicitly constructible. Moreover it forced Pappus to go beyond the usual geometric interpretation of products of lines as rectangles. Noting that "there is nothing contained under more than three

dimensions" (that is, one cannot form a notion of more than three dimensions), he continues:

> It is true that some recent writers have agreed among themselves to use such expressions, but they have no clear meaning when they multiply the rectangle contained by these straight lines with the square on that or the rectangle contained by those. They might, however, have expressed such matters by means of the composition of ratios, and have given a general proof. . . .

This passage is of crucial importance in the history of mathematics, since it states both the fundamental difficulty in the development of analytic geometry within the Euclidean system and the route by which this difficulty might have been overcome. The fundamental idea of analytic geometry, already present in Euclid's books on number theory, is that numbers can be represented by line segments. The problem of incommensurables is that there seems to be an excess of line segments: Some line segments cannot be regarded as numbers. This problem might have been overcome if the algebraic operations on numbers had been suitably interpreted. Addition and subtraction have a straightforward interpretation by placing line segments end to end. The product of two line segments is interpreted as a rectangle having the segments as sides, and the quotient as their geometric "ratio." But this last notion is not clearly defined in Euclid, as we have already seen. These interpretations made it impossible to give a geometric representation of a product of more than three lines. The notion of composite proportions—a product or ratio of *ratios*—as suggested by Pappus, would have overcome these problems. Thus, instead of the "ratio" $abcd : efgh$, which had no meaning when $a, \ldots h$ were interpreted as lines, Pappus proposed considering the "compound ratio" $(a : e) \cdot (b : f) \cdot (c : g) \cdot (d : h)$, which made perfectly good sense. Although Pappus did not work out the details, this idea naturally leads to the interpretation of a number as a *ratio of line segments* rather than as a single line segment. It is significant that this step was the critical one for Descartes when he created his analytic geometry. He took as given a certain line segment regarded as unity. A number was then interpreted as a line segment having a given ratio to the unit segment, so that the product of two lines could be interpreted as a line, rather than as an area. Not by coincidence, Descartes was convinced of the value of this work precisely because it enabled him to study the locus to five and six lines.

The passage just discussed in Book VII of the *Collection* is immediately followed by Pappus' statement that the topic is of limited importance compared with some theorems he himself had proved. He then gives the following statement of such a theorem: *The ratio of* [*areas*] *completely rotated is compounded from the ratios of the* [*areas*] *rotated and the ratios of the lengths of lines from the centers of gravity similarly drawn to the axes of rotation.* The modern theorem called Pappus' theorem asserts that the volume of a solid of revolution is equal to the product of the area rotated and the distance traversed by its center of gravity (which is 2π times the length of the line from the center of gravity to the axis of rotation). In the modern form this theorem was first stated by the Swiss astronomer/mathematician

H. P. Guldin (1577–1643). Unfortunately we do not have Pappus' proof of this theorem.

The eighth and last book of the *Collection* contains a discussion of mechanical devices using the geometric theory of proportion. It is here that Pappus quotes Archimedes as saying, "*δός μοι ποῦ στῶ καὶ κινῶ τὴν γῆν.*" ("Give me a place to stand, and I will move the earth.") The basic problem is *to move a given weight with a given force.* These problems are interspersed with discussion of certain purely geometric questions such as *to draw an ellipse through five given coplanar points.*

8.4 Hypatia

The major part of mathematical writing from the period of Roman hegemony consists of commentaries. Some of these commentators—Proclus, Theon of Smyrna, Iamblichus, Simplicius, and Eutocius—have been mentioned in Chapter 4. Two others worthy of mention are Theon of Alexandria (fourth century) and his daughter Hypatia (ca. 355–415), the first woman mathematician whose name is known. Unfortunately little else is known about her. There are two primary sources for information about her life. One is a passage in a seven-book history of the Christian Church written by Socrates Scholasticus, who was a contemporary of Hypatia, but lived in Constantinople; the other is an article in the *Suda*, an encyclopedia compiled at the end of the tenth century, that is, some five centuries after Hypatia. (This work bears the traditional name *Suidas*, erroneously thought to be the name of the person who compiled it.) In addition several letters of Synesius, bishop of Ptolemais (in what is now Libya), who was a disciple of Hypatia, were written to her or mention her, always in terms of high respect. In one letter he requests her, being in the "big city," to procure him a scientific instrument (hygrometer) not available in the less urban area where he lived. In another he asks her judgment on whether to publish two books that he had written, saying

> If you decree that I ought to publish my book, I will dedicate it to orators and philosophers together. The first it will please, and to the other it will be useful, provided of course that it is not rejected by you, who are really able to pass judgment. If it does not seem to you worthy of Greek ears, if, like Aristotle, you prize truth more than friendship, a close and profound darkness will overshadow it, and mankind will never hear it mentioned...

The account of Hypatia's life written by Socrates Scholasticus, who was a later contemporary of Hypatia, occupies Chapter XV of Book VII of his *Ecclesastical History*. Socrates Scholasticus describes Hypatia as the pre-eminent philosopher of Alexandria in her own time and a pillar of Alexandrian society, who entertained the elite of the city in her home. Among that elite was the Roman procurator Orestes. There was considerable strife at the time among Christians, Jews, and pagans in Alexandria, and Cyril, the bishop of Alexandria, was apparently in conflict with Orestes. According to Socrates, a rumor was spread that Hypatia prevented Orestes

from being reconciled with Cyril. This rumor caused some of the more volatile members of the Christian community to seize Hypatia and murder her in March of the year 415.

Five centuries after the death of Hypatia a Greek encyclopedia known as the *Suda* was compiled. The *Suda* devotes a long article to Hypatia, repeating in essence what was related by Socrates Scholasticus. It adds, however, that Hypatia was the wife of the philosopher Isodoros, which is definitely not the case, since Isodoros lived at a later time. The *Suda* assigns the blame for her death to Cyril himself.

Yet another eight centuries passed, and Edward Gibbon came to write the story in his *Decline and Fall of the Roman Empire* (Chapter XLVII). In his version Cyril's responsibility for the death of Hypatia is reported as fact, and the murder itself is described with certain gory details for which there is no factual basis (the version given by Socrates Scholasticus is revolting enough, and did not need the additional horror invented by Gibbon).

A fictionalized version of Hypatia's life can be found in a nineteenth-century novel by Charles Kingsley, bearing the title *Hypatia, or New Foes with an Old Face*. What facts are known were organized in an article by Michael Deakin entitled "Hypatia and her mathematics" in the *American Mathematical Monthly*, March 1994, and a biographical study of her life by Maria Dzielska entitled *Hypatia of Alexandria* was published in 1995.

8.5 Roman Mathematics

Mathematics was not developed by the Romans, but even they found some uses for it in architecture, engineering, and geography, all of which were essential for administering the Empire.

8.5.1 Arches

Although a few arches are found here and there in Egyptian and Greek buildings, the predominant structures in both of these civilizations use post-and-lintel window and door frames, that is, a flat stone lying atop two posts. Such a construction leaves a tensile stress at the bottom surface of the lintel, which will break under its own weight if the posts are too far apart. For their bridges and tunnels the Romans used semicircular arches, which direct much of the tension outward to the posts. The arches are constructed of trapezoidal blocks called *voussoirs* arranged so that the weight on the center is passed to the blocks on each side. This structure allows much wider distances between the posts at the cost of requiring buttressing of the posts, which would otherwise be forced apart by the arch. A splendid example of this kind of Roman engineering is provided by the bridge at Nîmes, France (see Fig. 8.3). The Romans also used rows of such arches to support long tunnels and tunnel-shaped rooms in buildings. This structure is known as a *barrel vault*; its chief disadvantage is that it places considerable lateral stress on the walls and requires heavy buttressing.

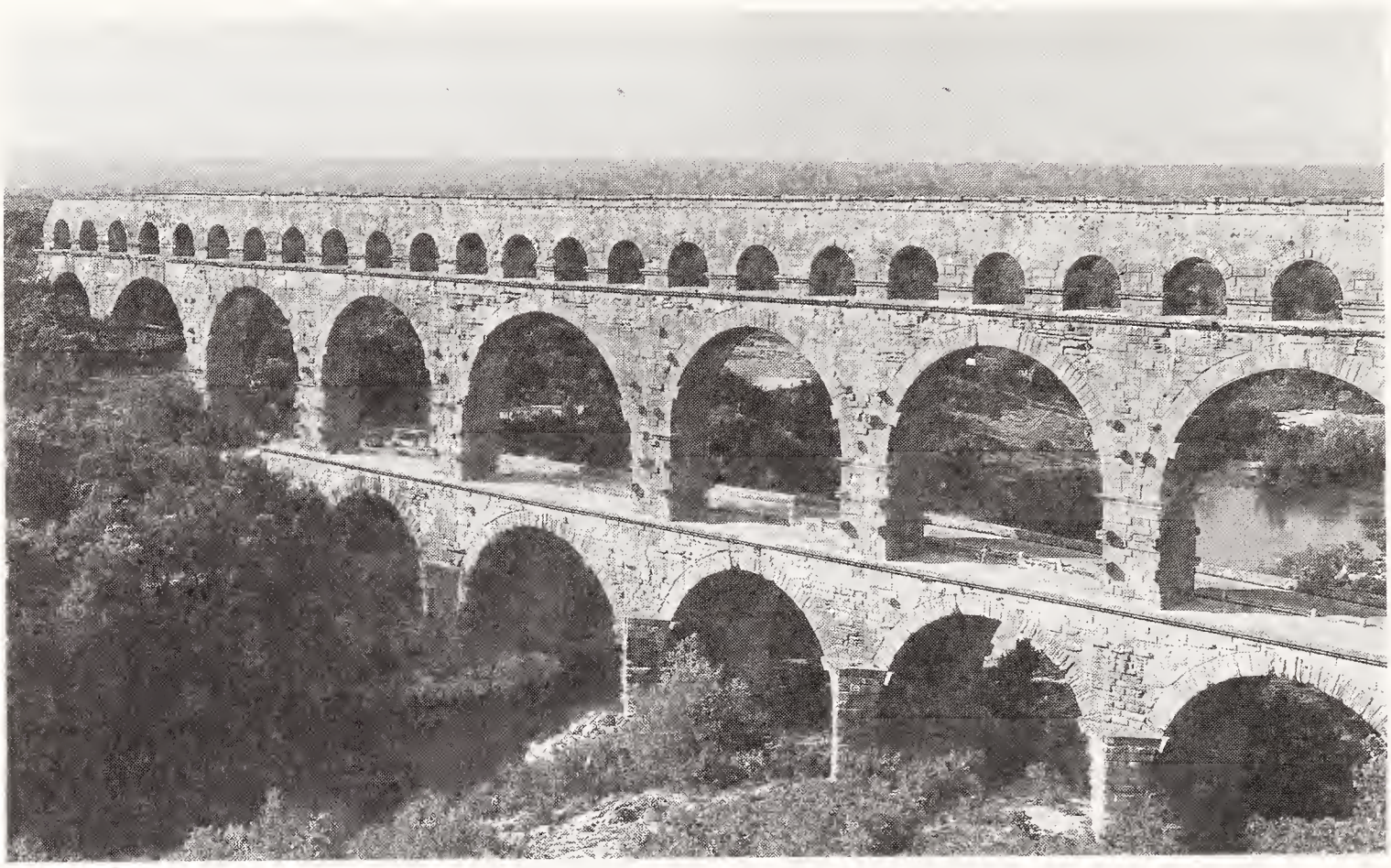


Figure 8.3: The Roman Aqueduct at Nîmes. The Bettmann Archive.

8.5.2 Mapmaking

In order to administer their extensive empire the Romans naturally needed accurate large-scale maps. Caesar Augustus commissioned his favorite admiral Agrippa to compile a map of the world. This map, which was incomplete at the time of Agrippa's death in 12 B.C.E., was completed under Augustus' direction and painted on a wall along the road in Rome now known as the Via del Corso. This wall has been destroyed, and no trace of the map remains. Such large-scale map-making requires that one take account of the curvature of the earth, and raises the problem of representing a curved surface on a plane. (This problem was to become one of the sources of differential geometry.)

Ptolemy, who lived under Roman rule in Alexandria, was the first scholar known to have looked at the problem of representing large portions of the earth's surface on a flat map. He also introduced the now-familiar lines of latitude and longitude. These lines have the advantage of being perpendicular to one another, but the disadvantage that the parallels of latitude are of different sizes. Hence a degree of longitude stands for different distances at different latitudes. Ptolemy's maps (see Fig. 8.4) are nevertheless a good example of the power of geometry for representing knowledge.

8.5.3 The *Groma*, the *Cardo*, and the *Decumanus*

From the ruins of Pompeii archaeologists have recovered a vital clue to Roman surveying, an instrument called the *groma*. The word is a mysterious one, variously thought to be a Latin corruption of the Greek *gnomon*, or derived from a conjectured word *gnorma* that was the source of both the Latin and Greek words as well as the

Figure 8.4: Ptolemy's map of the world. The Bettmann Archive.

word *normal* now used in mathematics to mean perpendicular. Writing in 1880, the German historian Moritz Cantor had only a picture of a groma, found in the tomb of a surveyor and published in 1854, on which to base his account of its use. Only one (damaged) instrument has actually been found, and it is in a room in the Naples Archaeological Museum not open to the public. A groma consisted of an iron cross with plumblines attached to the ends of each arm. It is believed that the surveyor sighted along the plumb-bobs toward an assistant who held a pole or stake. The instrument is well adapted for laying out accurate right angles, but not for measuring any other angles. As we shall see below, one can do accurate surveying without having to measure any angles except right angles.

The Romans conducted surveying in connection with the construction of roads and towns. The center of a Roman village would be at the intersection of two perpendicular roads, a north–south road called the *cardo maximus* (literally the "main hinge") and an east–west road called *decumanus maximus*, the "main tenth." Lots were laid out in blocks (*insulæ*) called "hundredths" (*centuriæ*), each block being given essentially what we call Cartesian coordinates (x, y), meaning x units *dextra decumani* (right) or *sinistra decumani* (left) and y units *ultra cardinem* (far) or *citra cardinem* (near).

8.5.4 The *Corpus Agrimensorum Romanorum*

A large collection of Roman writings on surveying was collected, translated into German, and published in Berlin in the middle of the nineteenth century. This

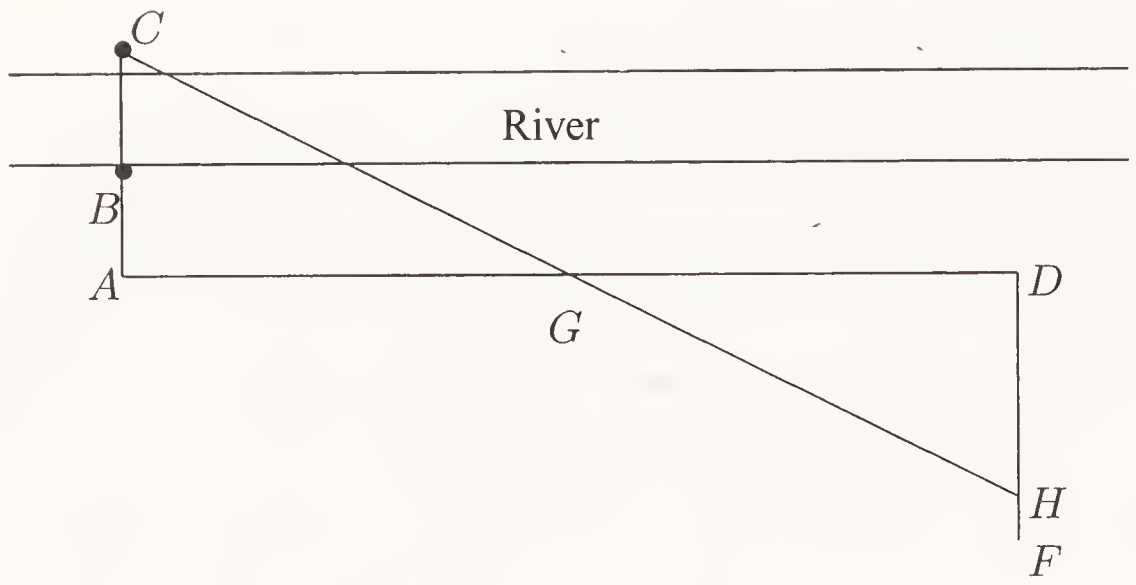


Figure 8.5: Nipsus' method of computing the width of a river.

two-volume work bears the title *Corpus Agrimensorum Romanorum*, the word *agrimensor* (field measurer) being the Latin name for a surveyor. Among the authors there contained is one M. Iunius Nipsus, who gives the following directions for measuring the width of a river (see Fig. 8.5).

You mark the point C on the opposite bank from B (a part of the procedure Nipsus neglects to mention until later), continue the straight line CB to some convenient point A, lay down the crossroads sign at A, then move perpendicularly to D, and then perpendicularly an indefinite distance to F, mark the midpoint G of AD with a pole, sight from G to C, then extend the line CG until it meets DF at H. Since the triangles AGC and DGH are congruent (by angle–side–angle), it follows that $CB = CA - AB = HD - AB$.

8.6 Problems and Questions

8.6.1 Problems from Diophantus and Pappus

Exercise 8.1 Problem 6 of Book I of the *Arithmetike* is *to separate a given number into two numbers such that a given fraction of the first exceeds a given fraction of the other by a given number.* In our terms this is a problem in two unknowns x and y, and there are four bits of data: the sum of the two numbers, which we denote by a, the two proper fractions r and s, and the amount b by which rx exceeds sy. Write down and solve the two equations that this problem involves. Under what conditions will the solutions be positive rational numbers (assuming that a, b, r, and s are positive rational numbers)? Compare your statement of this condition with Diophantus' diorismos, stated in very complicated language: *The last given number must be less than that which arises when that fraction of the first number is taken which exceeds the other fraction.*

Exercise 8.2 Compare your solution of the problem just discussed with that given by Diophantus. He illustrates the general case by dividing 100 into two parts, such

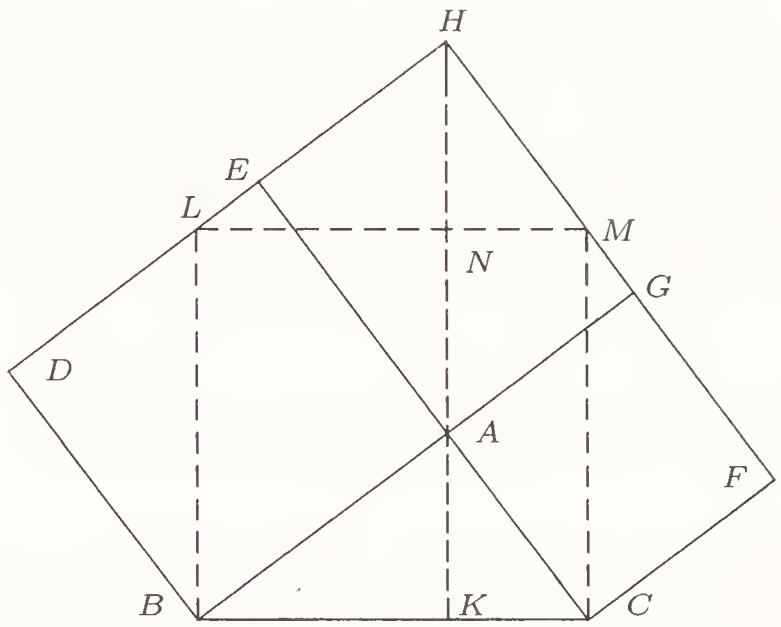


Figure 8.6: Pappus' proof of the Pythagorean theorem.

that $\frac{1}{4}$ of the first part exceeds $\frac{1}{6}$ of the second part by 20 (since 20 is less than $\frac{1}{4}$ of 100, the diorismos is satisfied). Diophantus lets the second part be $6x$, so that the first part is $4(x+20)$. Hence $10x+80=100$, so that $x=2$. The parts are therefore 88 and 12.

Exercise 8.3 Solve Problem 25 of Book I of Diophantus. *Find four numbers such that if each is increased by the same fraction of the other three the four resulting numbers are equal.* To make the problem determinate, assume that the sum of the last three numbers is 50 and that the given fraction by which all the numbers are to be increased is $\frac{1}{5}$.

Exercise 8.4 Solve Problem 19 of Book III of Diophantus (described above) by following Diophantus' approach: *In any right triangle the square on the hypotenuse increased or decreased by twice the product of the legs is a square. Therefore we must find [the legs of] four right triangles having the same hypotenuse. That is, we must find a square that is divisible into two squares in four different ways.* Diophantus' idea is, having found such a number, that is, having found four pairs (a_i, b_i), $i=1,2,3,4$, such that $a_i^2+b_i^2=c^2$, if one can arrange things so that $2(a_1b_1+a_2b_2+a_3b_3+a_4b_4)=c$, then one can take the four numbers to be $2a_ib_i$, $i=1,2,3,4$. For, the sum being c, one has $c^2 \pm 2a_ib_i=(a_i \pm b_i)^2$. The condition $2(a_1b_1+a_2b_2+a_3b_3+a_4b_4)=c$ can be obtained by scaling. Specifically, having *any* four pairs a_i, b_i satisfying $a_i^2+b_i^2=c^2$ for one and the same number c, we have, for any ς, $(a_i\varsigma)^2+(b_i\varsigma)^2=(c\varsigma)^2$. It is then merely a matter of choosing ς so that $2\big[(a_1\varsigma)(b_1\varsigma)+\cdots+(a_4\varsigma)(b_4\varsigma)\big]=c\varsigma$. This last equation is easily solved: $\varsigma=c/\big[2(a_1b_1+\cdots+a_4b_4)\big]$. The problem thus is merely to find the four Pythagorean triples with the same hypotenuse. This problem was solved in Problem 9 of Book II, which was discussed in the text. Diophantus takes $c=65$.

Exercise 8.5 Show that Pappus' generalization reduces to the usual Pythagorean theorem if the triangle ABC is a right triangle and the parallelograms on the legs

are squares, that is, show that the parallelogram on the hypotenuse is also a square (see Fig. 8.6). [*Hint:* If $\angle A$ is a right angle and $ACFG$ and $BAED$ are squares, then the triangles GHA and EAH are each congruent to the original triangle ABC, and so $\angle ACB = \angle GAH = \angle BAK$. Therefore triangle ABK is also a right triangle, that is, HK is perpendicular to BC and $CMLB$ is a rectangle. Since $CM = AH = BC$, it follows that $CMLB$ is a square.]

8.6.2 Questions about Diophantus and Pappus

Exercise 8.6 Building on our earlier definition of algebra from Chapter 3, we can characterize an algebra problem as being of the following generic type: *Certain standard arithmetic operations were performed on (unknown) numbers, and certain given numbers resulted. Find the numbers on which these operations were performed.* Diophantus' notation makes it possible to give a symbolic representation of the sequence of operations performed. Does it help in finding the solution of the problem?

Exercise 8.7 Do bees really construct hexagons in their honeycombs, or are the cells merely tangent circles whose interstices are filled with beeswax?

Exercise 8.8 Pappus, as we saw, paid little attention to the locus to more than six lines on the grounds that there are only three spatial dimensions, so that a product of four lines had no geometric meaning. Diophantus, on the other hand, considered (numerical) products of up to six factors. If these two mathematicians had combined their ideas, could they have invented parts of what we now know as analytic geometry? Sketch the outline of a project that would implement this program, showing how the notion of an unknown number could be combined with the notion of a number as a ratio of two line segments to produce the "equation of a curve."

Exercise 8.9 Pappus' assertion that the locus to five or six lines is a definite curve implicitly introduces into mathematics the notion of abstract "existence" as opposed to explicit exhibition. For instance, many constructions by Archimedes and others used lines that were assumed to exist but could not be constructed with straightedge and compass. The Greeks were reluctant to use such objects but found it unavoidable. There is a psychological difficulty in reasoning about objects that cannot be seen and processes that take place, so to speak, "offstage." Does such "offstage" action occur in high-school mathematics? Give an example from geometry and one from algebra in which one concludes that a point or a number exists, even though no method of exhibiting it is available.

8.7 Endnotes

1. The discussion of Diophantus is based on the edition of his works by T. L. Heath and the recent book by J. Sesiano, *Books IV to VII of Diophan-*

tus' Arithmetica in the Arabic Translation Attributed to Qusta ibn Luqa (Springer-Verlag, New York, 1982).

2. The quotation of Diophantus' rules for cancellation is taken from *Selections Illustrating the History of Greek Mathematics*, with a translation by Ivor Thomas (Harvard University Press, 1939), Vol. 2, p. 525.

3. The discussion of Pappus' works is based on the Latin–Greek edition of the *Collection* in three volumes published by Adolf M. Hakkert Verlag (Amsterdam, 1965).

4. The quotation of Pappus' opinion on the product of more than three lines is taken from *Selections Illustrating the History of Greek Mathematics*, Vol. 2, p. 603.

5. The quotation from a letter of Synesius is from *The Letters of Synesius of Cyrene*, translated by Augustine Fitzgerald (Oxford University Press, 1926).

6. The discussion of the chapter from Socrates Scholasticus is from the book *Ecclesiastical History. A History of the Church in Seven Books* (Samuel Bagster and Sons, London, 1844), pp. 482–483.

7. The information from the *Suda* is taken from the Greek original *Suidae Lexicon* (Teubner-Verlag, Stuttgart, 1971), Vol. 4, pp. 644–645.

8. The discussion of Gibbon is based on *The History of the Decline and Fall of the Roman Empire*, with notes by Dean Milman, M. Guizot, and Dr. William Smith (Harper & Brothers, New York, 1880), Vol. 4, pp. 646–647.

PART II

Other Mathematical Traditions

The development of mathematics in the West as a deductive system is a unique phenomenon. In other parts of the world mathematical results of considerable sophistication, involving the correct use of very intricate techniques and subtle reasoning, were obtained without the kind of formal proof demanded by Euclidean geometry. This feat seems almost more remarkable than the creation of Euclidean geometry itself. How is it possible to avoid delicate but fatal errors without the guidance of formally stated axioms and rules of inference? Perhaps the answer is that "mathematizing," like creating music, is an innate ability possessed by everyone to some degree and by a few geniuses to a high degree. Everyone knows that the music of different cultures sounds very different, yet we all recognize that there is something important that Yitzhak Perlman, Ravi Shankar, and John Coltrane all have in common. In only one culture did music lead to the symphony orchestra, and in only one culture did mathematics lead to deductive theories. Now, just as a concentration on symphonies would deprive the listener of the beauty of the koto or the sitar, and to spend all one's time studying the Impressionists would be to deny oneself the pleasure of African art, so an exclusive concentration on the standard mathematics of the modern curriculum would prevent the student from gaining the insight and pleasure that can be found in the mathematics of India, China, and Japan.

We study the mathematics of other cultures for a variety of reasons. Chief among them are the following: (1) the creators of this mathematics were exceptional geniuses whose creations deserve to be remembered; (2) their alternative ways of looking at problems cause us to rethink our own solutions; (3) some of what they did became part of the world's mathematical heritage, and its history ought to be told; and (4) some of the problems that other cultures have studied have no parallel in our own culture and are a delight to the imagination.

The different mathematical traditions are linked by the Muslim culture, which stretched from Mongolia to Spain, and built on knowledge inherited from the Greeks, Hindus, and Chinese. Muslim mathematics thus provides a natural bridge from the ancient world to the modern.

Chapter 9

The Mathematics of the Hindus

9.1 Indian Civilization

The Indian subcontinent has been inhabited for tens of thousands of years and has been the home of various civilizations for at least the last 4000 years. From archaeological excavations at Mohenjo Daro and Harappa on the Indus River in Pakistan it is known that an early civilization existed in this northwestern corner of the region for about a millennium starting in 2500 B.C.E. This civilization may have been an amalgam of several different cultures, since anthropologists recognize five different physical types among the human remains. Many of the artifacts that were produced by this culture have been found in Mesopotamia, evidence of trade between the two civilizations.

9.1.1 The Aryan Civilization

The early civilization of these five groups of people disappeared around 1500 B.C.E., and its existence was not known in the modern world until 1925. The cause of its extinction is believed to be an invasion from the northwest by a sixth group of people, who spoke a language closely akin to early Greek. Because of their language these people are referred to as Aryans. The Aryans gradually expanded and formed a civilization of small kingdoms, which lasted about a millennium. This classical civilization of northwest India exerted a strong influence on the customs and religion of India up to the present day, although it was subject to numerous stresses over the centuries due to foreign invasion.

Sanskrit Literature

The language of the Aryans became a literary language known as Sanskrit, in which great classics of literature and science have been written. Sanskrit thus played a

role in southern Asia analogous to Greek in the Mediterranean world. That is, it provided a common means of communication among scholars whose native dialects were not mutually comprehensible and a basis for a common literature in which cultural values could be preserved and transmitted. During the millennium of Aryan dominance the spoken language of the people gradually diverged from written Sanskrit. (Close modern descendants of Sanskrit are Hindi, Gujarati, Bengali, and others.) Sanskrit is the language of the *Mahabharata* and the *Ramayana*, two epic poems whose themes bear some resemblance to the Homeric epics, and of the *Upanishads*, which contain much of the moral teaching of Hinduism.

Among the most ancient works of literature in the world are the Hindu *Vedas*. The word means "knowledge" and is related to the English word *wit*. The composition of the *Vedas* began around 900 B.C.E., and additions continued to be made to them for several centuries. Some of these *Vedas* contain information about mathematics, conveyed incidentally in the course of telling important myths.

Hindu Religious Reformers

Near the end of the Aryan civilization, in the second half of the sixth century B.C.E., two figures of historical importance arise. The first of these was Gautama Buddha, the heir to a kingdom near the Himalayas, whose spiritual journey through life led to the principles of Buddhism. The second leader, Mahavira, is less well known, but more important for the history of mathematics. Like his contemporary Buddha, he began a reform movement within Hinduism. This movement, known as Jainism, still has several million adherents in India. It is based on a metaphysic that takes very seriously what is known in some Western ethical systems as the chain of being. Living creatures are ranked according to their awareness. Those having five senses are the highest, and those having only one sense are the lowest. All life is considered sacred. It is recognized that it is necessary to kill some life (at least plant life) in order to remain alive, but the most deeply initiated Jainas frequently wear cloths over their mouths to prevent the inadvertent inhalation of a small insect. They practice an extreme asceticism, recognizing fully that such practices are injurious to their health. In fact Jainism is the only religion to approve of hastening one's own death (under strictly specified conditions) in order to be free of the matter that weighs down the soul. What these principles have to do with mathematics will appear below.

9.1.2 The Maurya Dynasty

The Aryan system of small kingdoms was threatened by the invasion of Alexander of Macedon in 326 B.C.E. Although he defeated an Aryan army, Alexander was forced to return to the West by a rebellion in his own army. Five years later Chandragupta Maurya became the strongest of the local lords and founded the first empire in India. His empire was very short-lived, however, as his grandson Asoka, a convert to Buddhism, was reluctant to use military force. Asoka died in 232 B.C.E., and the last Mauryan ruler was assassinated in 185. The Mauryan

empire was then invaded from the south and simultaneously from the northwest. The invader from the northwest was a Hellenistic prince named Demetrius. He established a regime along the lines of the Seleucid rulers. His coins bore Greek engravings on one side and Indian on the other.

9.1.3 The Kushan Empire and the Gupta Empire

By the year 100 B.C.E., the kingdom of Demetrius had been conquered by invaders who established the Kushan Empire. This empire lasted for about 150 years and was marked by brilliant achievements in art, architecture, and science. It was replaced by a short-lived domination by Persia. Meanwhile in the south the Dravidian peoples, speaking Tamil, had absorbed a great deal of Hindu culture and were at the same time in contact with the Hellenistic world, as shown by the presence of Tamil words for various spices in the Greek language. When the Kushan Empire began to crumble around 220 C.E., a century of disorder followed in the north, ending with the ascension of Chandragupta I (not related to Chandragupta Maurya), who established himself as ruler of the entire Ganges Valley. This empire expanded for about a century and managed to check an invasion by the Huns in the fifth century. During this time India established colonies in southeast Asia, some of which endured as independent Indian kingdoms for over a millennium. The Gupta period was another time of brilliant literary and scientific achievement, the golden age of Sanskrit literature, in which the fifth-century author Kalidasa wrote lyric and epic poems and dramas of such quality that he has been compared with Shakespeare. At the university at Nalanda Hindu scholars studied astronomy and mathematics. Two outstanding mathematicians, whose discoveries we shall be discussing below, lived during this period.

9.1.4 Islam in India

The amazingly rapid Muslim expansion from the Arabian desert in the seventh century brought Muslim invaders to India by the early eighth century. The southern valley of the Indus river became a province of the huge Umayyad Empire, but the rest of India preserved its independence, as it did 300 years later when another Muslim people, the Turks and Afghans, invaded. The complete and destructive conquest of India by the Muslims under Timur the Lame came at the end of the fourteenth century. Timur did not remain in India, but sought new conquests; eventually he was defeated by the Ming dynasty in China. Nevertheless, India was desolated by his attack, and was decisively conquered a century later by Akbar the Lion, a descendant of both Genghis Khan and Timur the Lame and the first of the Mogul emperors. The Mogul empire lasted nearly three centuries, and was a time of prosperity and cultural resurgence. It was near the end of this period (midseventeenth century) that the Taj Mahal (Fig. 9.1) was built in Delhi by Shah Jahan as a mausoleum for himself and his favorite wife. Although Akbar himself was tolerant of non-Muslims, his successors were not. In contrast to the relative tolerance the Muslims had extended to Christians and Jews, they dealt ruthlessly

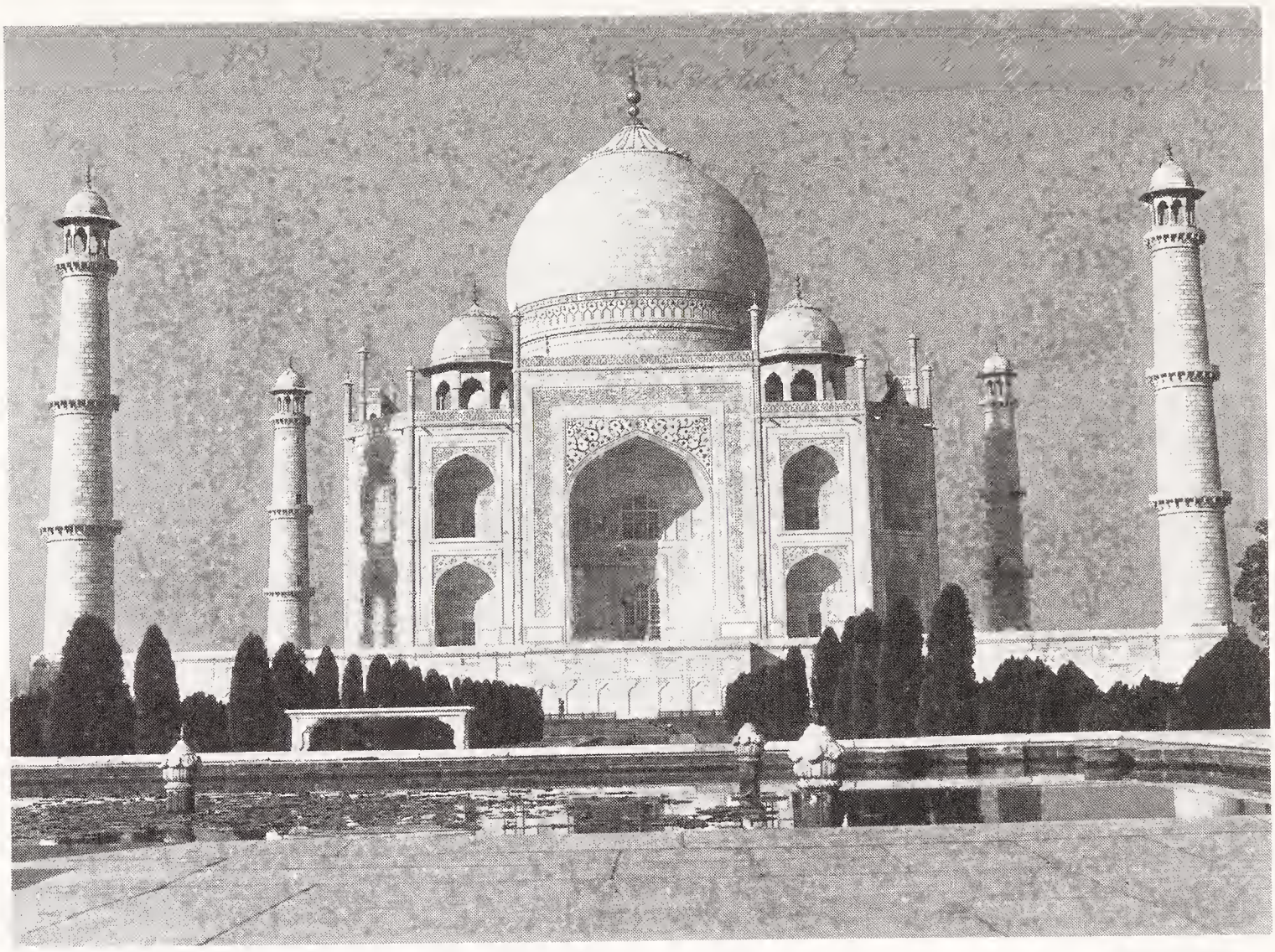


Figure 9.1: The Taj Mahal. Like the Parthenon, this famous mausoleum, built by the Muslim prince Shah Jahan in the seventeenth century, shows a sophisticated knowledge of geometric principles to solve practical problems and create beauty. The Bettmann Archive.

with the Hindus, whom they considered pagans. As a result, the Indian populace did not abandon Hinduism for Islam, as the majority had abandoned Christianity in the Middle East. India has been religiously divided since this time.

9.1.5 British Rule

During the seventeenth and eighteenth centuries British and French trading companies were in competition for the lucrative trade with the Mogul Empire. British victories during the Seven Years War left Britain in complete control of this trade. Coming at the time of Mogul decline (due to internal strife among the Muslims and ferocious resistance on the part of the persecuted Hindus), this trade opened the door for the British to make India part of their empire. British colonial rule lasted nearly 200 years, coming to an end only after World War II. British rule made it possible for European scholars to become acquainted with Hindu classics of literature and science. As a result many Sanskrit works were translated into English in the early nineteenth century and became part of the world's science and literature.

9.2 The Beginnings of Hindu Mathematics

The oldest surviving documents on Hindu mathematics are copies of works written in the middle of the first millennium B.C.E., approximately the time during which Thales and Pythagoras lived. We shall discuss two types of documents, the *Sulva Sutras*, which convey mathematical information in relation to rules for the ordering of life, and systematic treatises bearing the name *Siddhanta* in their title. The word *Siddhanta* means *that which is proved or established.* The *Sulva Sutras* are of Hindu origin, but the *Siddhantas* contain so many words of foreign origin that they undoubtedly have roots in Mesopotamia and Greece.

Since there is known to have been trade between the Middle East and India during this period, the Babylonian mathematics that had such a strong influence on Greece may also have been known in India, and there may also have been influence in the opposite direction. Mathematics was written down systematically during the Gupta empire, which was the time of Greek and Roman decline. When the Ummayads conquered part of the Indus valley in the eighth century, much of this Hindu mathematics came to the Islamic world. There it was further developed and ultimately made a contribution to the world of modern mathematics, which began in Europe during the sixteenth century.

Our story of the mathematics among the Hindus will be divided into three periods: (1) the early period of the *Sulva Sutras* and *Siddhantas* (from the sixth century B.C.E. to the fifth century C.E., but building on knowledge of much older date); (2) the period from Aryabhata (fifth century C.E.) to Bhaskara (twelfth century C.E.); and (3) the modern period.

9.3 The Earliest Period

The best-known fact about Hindu mathematics is that the decimal notation and the symbols for numerals we use today originated in India and came to Europe through the Arabs. Of course, decimal *systems* are very common throughout the world. What makes the Hindu system valuable is the place-value notation, which we have encountered already in Mesopotamia in connection with a sexigesimal system. Whether this system originated in China or India is not absolutely certain, since the two countries were in contact from a very early date, but it certainly came to the West from the Arabs, who learned it from India. In fact, one of the influential treatises by which Europeans learned about the decimal system and the symbols for digits was a treatise by the Muslim scholar Kushyar ibn Labban (ca. 971–1029), and the title he gave to the treatise is *The Art of Hindu Reckoning.*

The Hindus invented names for very large powers of 10 at an early date. One early poem, the *Valmiki Ramayana*, from about 500 B.C.E., explains the numeration system in the course of recounting the size of an army. The description uses special words for 10^7, 10^{12}, 10^{17}, and many other denominations, all the way up to 10^{55}. Essential to a place-value notation is a symbol for nothing, which is known to have been invented in India before 200 B.C.E.

9.3.1 The *Sulva Sutras*

In the period from 800 to 500 B.C.E. a set of verses of geometric and arithmetic content were written and became part of the *Vedas*. These verses are known collectively as the *Sulva Sutras*. The name means *Cord Book* or *Cord Rules*, a name that may reflect the same origin as the word *rope stretchers* used by Democritus. The root *sulv* originally meant *to measure* or *to rule*, although it also has the meaning of a cord or rope; *sutra* means *thread* or *cord*, a common measuring instrument.

Number Theory

The arithmetic content of the *Sulva Sutras* consists of rules for finding Pythagorean triples of integers, such as $(3, 4, 5)$, $(5, 12, 13)$, $(8, 15, 17)$, and $(12, 35, 37)$. It is not certain what practical use these arithmetic rules had. The best conjecture is that they were part of religious ritual. A Hindu home was required to have three fires burning at three different altars. The three altars were to be of different shapes, but all three were to have the same area. These conditions led to certain "Diophantine" problems, a particular case of which is the generation of Pythagorean triples, so as to make one square integer equal to the sum of two others.

One class of mathematical problems associated with altar building involves an altar of prescribed area in layers. In one problem from the *Bodhayana Sutra* the altar is to have five layers of bricks, each layer containing 21 bricks. Now one cannot simply divide a pile of 105 identical bricks into five layers and pile them up. Such a structure would not be stable. It is necessary to stagger the edges of the bricks. Thus, so that the outside of the altar will not be jagged, it is necessary to have at least two different sizes of bricks. The problem is to decide how many different sizes of bricks will be needed and how to arrange them. Assuming an area of one square unit (actually the unit is one square *vyayam*, about 64 square feet), the author suggests using three kinds of square bricks, of areas $\frac{1}{36}$, $\frac{1}{16}$, and $\frac{1}{9}$ square unit. The first, third, and fifth layers are to have 9 of the first kind and 12 of the second. The second and fourth layers get 16 of the first kind and five of the third.

Geometry

The geometric content of the *Sulva Sutras* encompasses many of the transformation-of-area constructions known from Euclid. In particular the Pythagorean theorem, and constructions for finding the side of a square equal to a rectangle, or the sum or difference of two other squares are given. This construction resembles the one found in Proposition 5 of Book II of Euclid rather than Euclid's construction of the mean proportional in Book VI. The Pythagorean theorem is not given a name, but is stated as the fact that "the diagonal of a rectangle produces both [areas] which its length and breadth produce separately." It is interesting that the problem of doubling a square, which we speculated in Chapter 3 might have led to the discovery of this theorem, produces a figure in the shape of one of the altars

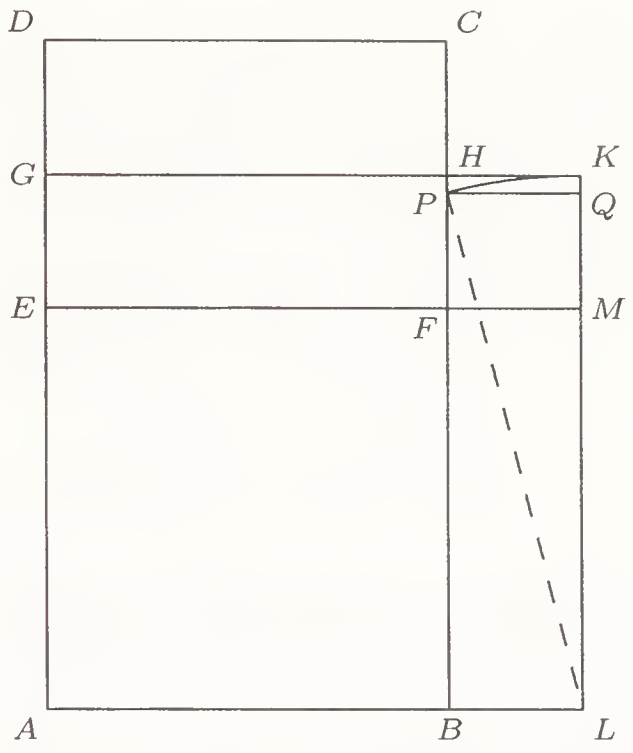


Figure 9.2: Quadrature of the rectangle in the *Sulva Sutra*.

discussed in the Vedas. Is it merely a coincidence that the problem of doubling the cube was said by the Greeks to have been inspired by an attempt to double the size of an altar?

The Hindu method of constructing of a square equal to a given rectangle (see Fig. 9.2) is as follows. Let $ABCD$ be the given rectangle, with AD longer than AB. Mark point E on AD so that $AE = AB$ and F on BC so that $BF = AB$. Draw EF, obtaining the square $ABFE$. Let G be the midpoint of ED and H the midpoint of FC. Draw GH and extend it to K so that $GK = AG$. Extend AB to L so that $AL = GK = AG$. Draw KL, obtaining the square $ALKG$. Extend EF to meet LK at M. Then the rectangle $ABCD$ equals the square $ALKG$ minus the square $HKMF$ (since the rectangle $CDGH$ equals the rectangle $BLMF$). Next choose P on BH so that $PL = KL$ (this can be done by drawing a circle with L as center and LK as radius). Draw the line from P perpendicular to LK meeting LK at Q. Then the square on LQ is the square on LP minus the square on PQ. But since $PQ = HK$ and $LP = LK$, it follows that the square on LQ is precisely equal to the rectangle $ABCD$.

To construct a square equal to a multiple of a given square, say seven times as large as a square of side a, the *Katyayana Sutra* says to construct an isosceles triangle of base $6a$ and two sides equal to $4a$. The altitude, that is, the perpendicular bisector of the base, will have length $a\sqrt{4^2 - 3^2} = \sqrt{7}a$, and hence will be the side of a square 7 times the original square.

The requirement of three altars of equal areas but different shapes would explain the interest in transformation of areas. Among other transformation of area problems the Hindus considered in particular the problem of squaring the circle. The *Bodhayana Sutra* states the converse problem of constructing a circle equal to a given square. The following approximate construction is given as the solution.

Let $ABCD$ be the square (see Fig. 9.3). From the center O of the square draw a circle with radius equal to OC. Let L be the midpoint of side BC, and let the

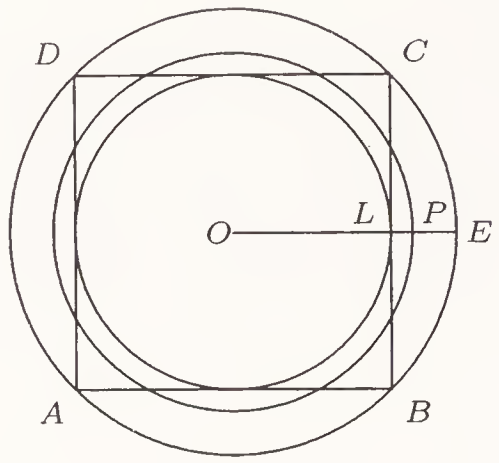


Figure 9.3: Rounding a square.

radius through L meet the circle in the point E. Choose a point P on LE one-third of the way from L to E. The point P will lie on the circle with center at O equal to the square $ABCD$. In contrast to the previously cited results on transformation of areas, which were exact, this result is only approximate. The authors, however, made no distinction between the two results. In terms that we can appreciate, this construction gives a value for π of $18(3 - 2\sqrt{2})$, which is about 3.088.

Square Roots

The Hindus also had a very good system of approximating irrational square roots. Three of the *Sulva Sutras* contain the expression

$$1 + \frac{1}{3} + \frac{1}{3 \cdot 4} - \frac{1}{3 \cdot 4 \cdot 34}$$

for the diagonal of a square of side 1 (that is, $\sqrt{2}$). If this series represents successive approximations to $\sqrt{2}$, these approximations are $1, \frac{4}{3}, \frac{17}{12}, \frac{577}{408}$. The method described in Chapter 3 gives $1, 2, \frac{3}{2}, \frac{4}{3}, \frac{17}{12}, \frac{24}{17}, \ldots$. We can only conjecture how such an approximation was obtained. One guess is the approximation

$$\sqrt{a^2 + r} = a + \frac{r}{2a} - \frac{(r/2a)^2}{2\big[a + r/2a\big]},$$

with $a = \frac{4}{3}$ and $r = \frac{2}{9}$. This approximation follows a rule given by the twelfth-century Muslim mathematician Al-Hassar.

It is not certain just how the early Hindu mathematicians conceived of irrational numbers, whether they had a name for them, or were merely content to find a number that would serve for practical purposes. By the year 1500 C.E., however, Hindu commentators were stating plainly their belief that the circumference and diameter of a circle are incommensurable.

Here we see an instance in which the Greek insistence on logical correctness was a hindrance. The Greeks did not regard $\sqrt{2}$ as a number, since they could not express it exactly as a ratio and they knew that they could not. The Hindus

may or may not have known of the impossibility of a rational expression for this number (they certainly knew that they did not *have* any rational expression for it); but, undeterred by the incompleteness of their knowledge, they proceeded to make what use they could of this number. This same "reckless" spirit served them well in the use of infinity and the invention of zero and negative numbers. They saw the usefulness of such numbers and either chose to live with, or did not notice, certain difficulties of a metaphysical character.

9.3.2 Jaina Mathematics

Like Greek mathematics, Hindu mathematics has a prominent metaphysical component. This metaphysical aspect manifests itself in various ways. One important place is in the handling of the infinite. Where the Greeks had regarded all reasoning as finite and accepted only a potential infinity, the Hindus accepted an actual infinity and classified different kinds of infinities. This part of Hindu mathematics is particularly noticeable with the Jainas, who were mentioned above. They classified numbers as enumerable, unenumerable, and infinite, and space as one-dimensional, two-dimensional, three-dimensional, and infinitely infinite. Further, they seem to have given a classification of infinite numbers remarkably similar to the modern-day theory of infinite ordinals. The idea is to progress through the finite numbers $2, 3, 4, \ldots$ until the "first unenumerable" number is reached. This number corresponds to what is now called ω, the first infinite ordinal number. Then, exactly as in modern set theory, one can consider the unenumerable numbers $\omega+1$, $\omega+2$,..., ω^2, etc. We do not have enough specifics to say any more, but there is a very strong temptation to say that the Jaina classification of enumerable, unenumerable, infinite corresponds to our modern classification of finite, countably infinite, and uncountably infinite. One must be careful, however. It appears that the Jainas considered every number to have both a successor and a predecessor, while in modern set theory not every infinite ordinal has a predecessor.

The metaphysics of the Jainas, based on a classification of sentient beings according to the number of senses possessed, led them to a mathematical topic not discussed by the Greeks. They called it *vikalpa*, and we know it as combinatorics. (The Sanskrit word *kalpa* has many meanings, among which are *possible, feasible*, and more significantly from our point of view, *ordered*. The prefix *vi-* corresponds roughly to the English prefix *dis-*, so that *vikalpa* may mean *distribution*. The occurrence of the word in the present context probably derives from the *Kalpa Sutras*, a set of Jaina verses.)

Given that there are five senses and animals are to be classified according to the senses they possess, how many different classes will there be? A typical question might be, how many groups of three can be formed from a set of five elements? We know the answer, as did the early Jaina mathematicians. In the *Bhagabati Sutra*, written about 300 B.C.E., the author asks how many philosophical systems can be formed by taking a certain number of doctrines from a given list of basic doctrines. After giving the answers for 2, 3, 4, etc., the author says that enumerable, unenumerable, and infinite numbers of things can be discussed, and, "as the number

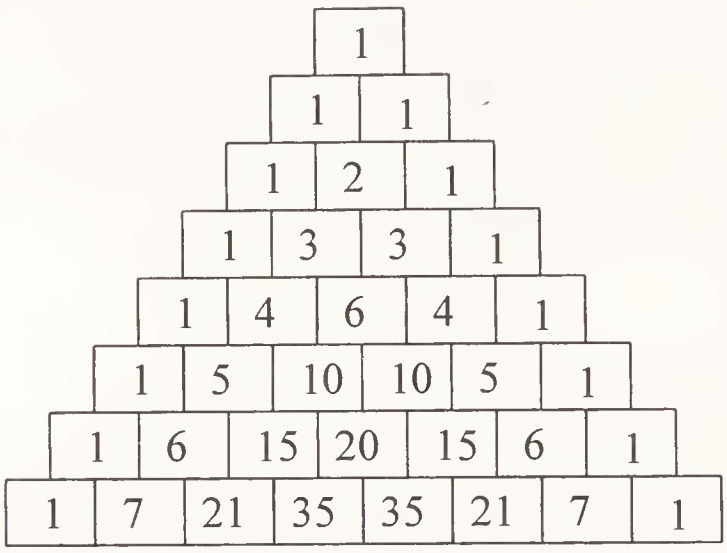


Figure 9.4: The *Meru Prastara.*

of combinations are formed, all of them must be worked out."

The general process for computing combinatorial coefficients was known to the Hindus at an early date. Combinatorial questions seemed to arise everywhere for the Hindus, not only in the examples just given but also in a much earlier work on medicine that poses the problem of the number of different flavors that can be made by choosing subsets of six basic flavors (bitter, sour, salty, astringent, sweet, hot). The author gives the answer as $6 + 15 + 20 + 15 + 6 + 1$, that is, 63. We recognize here the combinatorial coefficients that give the subsets of various sizes that can be formed from six elements. (The author did not count the possibility of no flavor at all.)

Combinatorics also arose with the Hindus in the study of literature, when a writer named Pingala in the third century B.C.E. gave a rule for finding the number of different words that could be formed from a given number of letters. This rule was written very obscurely, but a commentator named Halayudha in the tenth century C.E., explained it as follows. First draw a square. Below it and starting from the middle of the lower side, draw two squares. Then draw three squares below these, and so on. Write the number 1 in the middle of the top square and inside the first and last squares of each row. Inside every other square the number to be written is the sum of the numbers in the two squares above it and overlapping it. This, of course, is a perfect description of Pascal's triangle, 300 years before it was published in China and 700 years before Pascal. Moreover it purports to be only a clarification of a rule invented 1200 years earlier! The priority for this discovery thus belongs to the Hindus. Its Sanskrit name is *Meru Prastara* (see Fig. 9.4), which means the *spread out* or *stratified Mount Meru.*[1] We note that the inspiration for the study of this figure was quite different in China and India. In China, as we shall see in the next chapter, it came about in connection with the

[1] In Hindu mythology Mount Meru plays a role similar to that of Mount Olympus in Greek mythology. One Sanskrit dictionary gives this mathematical phrase as a separate entry.

extraction of roots and the solution of equations, whereas in India the inspiration was directly from the area of combinatorics.

9.3.3 The Bakshali Manuscript

A birchbark manuscript unearthed in 1881 in the village of Bakshali, near Peshawar, is now believed to date from the seventh century C.E. It contains some interesting algebra, including a symbol for an unknown quantity (⊂). One of the problems considered is written as follows, using modern number symbols and a transliteration of the Sanskrit into the Latin alphabet:

$$\begin{array}{cccccccccccc} \subset & 5 & & & \subset & & \subset & 7+ & & \subset \\ 1 & 1 & yu & m\bar{u} & 1 & sa & 1 & 1 & m\bar{u} & 1 \end{array}$$

which can be interpreted as, "a certain thing is increased by 5 and the square root is taken, giving [another] thing; and the thing is decreased by 7 and the square root is taken, giving [yet another] thing." In other words, we are looking for a number x such that $x + 5$ and $x - 7$ are both perfect squares. This problem is remarkably like certain problems in Diophantus. For example, Problem 11 of Book II of Diophantus is to add the same number to two given numbers so as to make each them a square. (If the two given numbers are 5 and -7, this is *exactly* the problem stated here.)

The Bakshali manuscript also contains problems in linear equations, of the sort that have had a long history in elementary mathematics texts. For example, three persons possess 7 thoroughbred horses, 9 draft horses, and 10 camels respectively. Each gives one animal to each of the others. The three are then equally wealthy. Find the (relative) prices of the three animals.

9.3.4 The Siddhantas

Just after the time of Ptolemy, in the second, third, and fourth centuries C.E., Hindu scientists were compiling treatises on astronomy known as *Siddhantas*. The word *Siddhanta* means a system and thus corresponds very nearly to the title of Ptolemy's treatise, the *Syntaxis*. One of these systems, the *Surya Siddhanta* (System of the Sun), from the late fourth century, has survived intact. Another from approximately the same time, the *Paulisha Siddhanta*, was frequently referred to by the Muslim scholar Al-Biruni (973–1048). The name of this treatise seems to have been bestowed by Al-Biruni, who says that the treatise was written by an Alexandrian astrologer named Paul. Trigonometry is an essential tool in the study of astronomy, and this subject was extremely well developed by the Hindus. It was the Hindus who discovered that the subject is simpler if you express the relations between circular arcs and chords in terms of half-chords, what are now called *sines*.

9.4 The Middle Period

Although there is a long tradition of mathematical activity in India, only a few of the mathematicians are known by name. The most prominent of these are naturally the ones who wrote treatises that survive today. We shall sample this tradition in the works of the three best known: Aryabhata, Brahmagupta, and Bhaskara.

9.4.1 Aryabhata

The earliest Hindu treatise still surviving from the second period is apparently a summary of a mathematical tradition of which no other written record survives. We are therefore in somewhat the same position in relation to this period of Hindu mathematics that we would be in relative to Greek geometry if we had only Proclus' commentary on Euclid and had lost entirely the *Elements* themselves. Besides the difficulties of translating an ancient language from obscurely written and fragmentary manuscripts, the translators were faced with an additional problem: How can we know what point the author was trying to make if we do not have access to the knowledge he took for granted? The problems of interpretation in regard to Aryabhata's work are therefore considerable. Aryabhata himself (one of at least two mathematicians bearing that name) lived in the late fifth and early sixth centuries at Kusumapura (now Pataliputra, a village near the city of Patna) and wrote a book called the *Aryabhatiya*. This work had been lost for centuries when it was recovered by the Indian scholar Bhau Daji in 1864. Scholars had known of its existence through the writings of commentators and had been looking for it. Writing in 1817, the English author Henry Thomas Colebrooke reported, "A long and diligent research of various parts of India has, however, failed of recovering any part of the... Algebra and other works of Aryabhata."

Ten years after its discovery the *Aryabhatiya* was published at Leyden and attracted the interest of European and American scholars. It consists of 123 stanzas of verse, divided into four sections, of which the first, third, and fourth, are concerned with astronomy and the measurement of time. The following examples are taken from this work.

Astronomy

In the first chapter Aryabhata begins with some astronomical facts:

> In a *yuga* the revolutions of the sun are 4,320,000; of the moon 57,753,336; of the earth eastward 1,582,237,500; of Saturn 146,564; of Jupiter 364,224; of Mars 2,296,824; of Mercury and Venus the same as those of the sun.

The word *yuga* means *yoke*. This passage is intriguing for several reasons. It refers to an eastward rotation of the earth, suggesting that the author regards the stars as fixed, with the earth rotating beneath them, while the sun, moon, and planets wander among the stars. When the stars are fixed, the resulting time periods are

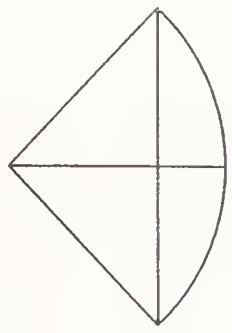


Figure 9.5: The "bowstring" (sine).

called *sidereal* periods, as discussed in Chapter 7. A *yuga* is a common period for all of the heavenly bodies: at the end of one *yuga*, they will all be in the same relative positions they were in at the beginning.

By dividing the figure given for the earth by the figure given for the sun, we find a sidereal year of 366.26 sidereal days, which is very close to the modern value. (A sidereal year is one day longer than a tropical year, since the sun makes one eastward circuit during the year, that is, the stars will have rotated east-to-west relative to the earth one more time than the sun.)

These figures give Jupiter a sidereal year of 11.86 years and Mars a year of 687 days. Again both of these figures agree with the modern values. The figure for Saturn is 29.47 years, while the best modern value is 29.46 years. Since Mercury and Venus must stay close to the sun, obviously they will make exactly the same number of circuits as the sun during any period in which they end where they began. (This is a well-known principle, known as *Rouché's theorem* in complex analysis.)

Geometry

Aryabhata gave the correct rule for area of a triangle and an incorrect rule for the volume of a pyramid. (He claimed the volume was half the height times the area of the base.) It is clear, therefore, that Aryabhata was not concerned with demonstration; we may infer that he was unfamiliar with Euclid, even though it was at least possible for him to have read Euclid. He seems to be reliable as regards plane figures, but unreliable in three dimensions. For example, he says the area of a circle is half the diameter times half the circumference, which is correct. He then says that the volume of a sphere is the area of a great circle times its own square root, which would be correct only if $\pi = \frac{16}{9}$, very far from the truth! Yet Aryabhata clearly knew a very good approximation to π. He writes

> Add 4 to 100, multiply by 8, and add 62,000. The result is approximately the circumference of a circle of which the diameter is 20,000.

This gives a value of π equal to 3.1416, which is quite accurate indeed. It is unfortunate that we do not know how this number was arrived at.

Trigonometry

Aryabhata used half-chords (sines), which occur in the earlier *Surya Siddhanta*, to study trigonometry. The *Aryabhatiya* contains a table of sines in intervals of 225 minutes of angle. The choice of the interval ($3.75°$) is undoubtedly motivated by the fact that this angle can be obtained by three angle bisections starting from $30°$. It also suggests that the discoverer didn't know Ptolemy's work, since Ptolemy had a much more sophisticated system based on the regular pentagon and was able to start with a $6°$ angle and bisect down to $\frac{3}{4}°$ intervals. We shall not examine this table, since it differs in technical respects from what we would call a table of sines, while giving equivalent information.

Aryabhata used the Sanskrit word *jya* for what we call a sine. The original meaning of this word was *bowstring*, and the reason for applying it to a chord on a circle is obvious from a glance at a drawing of the relevant geometric figure (Fig. 9.5). (The Greek word $\chi o \rho \delta \acute{\eta}$, transliterated as *chorde*, also means *string*.) A half-chord was originally called *ardhajya*, but since full chords were never used, the first syllable was eventually dropped. When the *Aryabhatiya* was translated into Arabic, this word was taken directly without translation; it simply became *jb*. In the eleventh century, when Plato of Tivoli translated a treatise on celestial mechanics by the Syrian astronomer Al-Battani from Arabic into Latin, he used the Latin word *sinus*, which denotes a cavity or fold, and corresponds to the Arabic word *jaib*, thereby giving us the word *sine*, although it has nothing to do with the original meaning of the concept.

Despite knowing a sophisticated version of trigonometry, Aryabhata discussed surveying without making any use of angles other than right angles. In fact, his method of surveying seems to be of much older date than his trigonometry. It is identical to a method that was used in China for centuries before this time and was still being used in the Muslim world and Medieval Europe many centuries later. He gives the following method of finding the height and distance of an object. Erect two poles of equal height and imagine a light at the top of the object shining down so that the poles cast shadows. Then the computations are as follows (brackets refer to Fig. 9.6).

> The distance between the ends of the two shadows $[AB]$ multiplied by the length of the shadow $[BE]$ and divided by the difference in length of the two shadows $[AF - BE]$ gives the *koti* $[BC]$. The *koti* multiplied by the length of the gnomon $[h]$ and divided by the length of the shadow $[BE]$ gives the length of the *bhuja* $[CD]$.

Algebra

Aryabhata studied some problems that we have come to regard as algebra. For example, he considered the problem of finding two numbers given their product and their difference and gave the standard recipe for solving it:

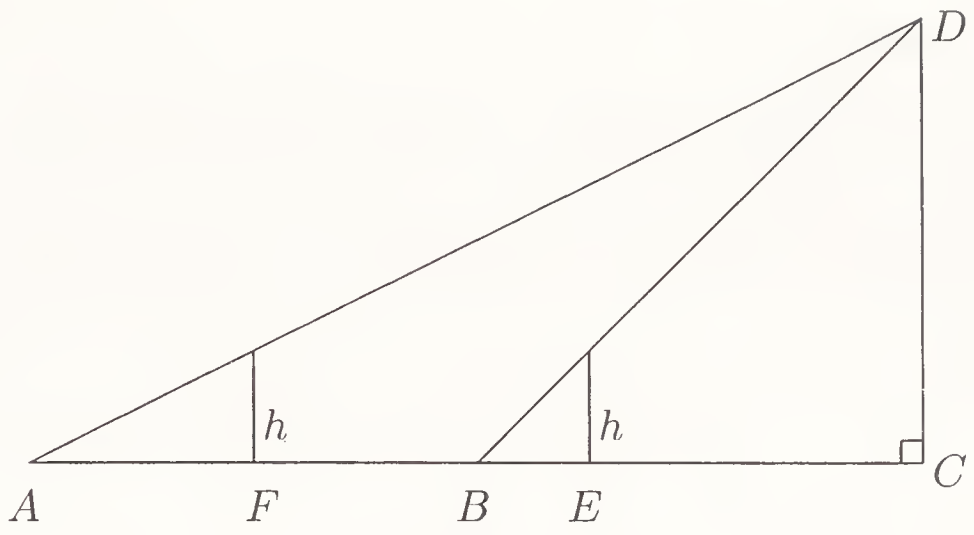


Figure 9.6: The Hindu method of surveying.

> Multiply the product by four, add the square of the difference, take the square root, then add and subtract the difference and divide the result by 2. This will yield the two numbers.

Aryabhata also considered finite arithmetic progressions and gave a rather complicated recipe for finding the sum of the terms. He gave more elegant rules for the sum of the squares and cubes of an initial segment of the positive integers.

> The sixth part of the product of three quantities consisting of the number of terms, the number of terms plus one, and twice the number of terms plus one is the sum of the squares. The square of the sum of the series is the sum of the cubes.

Here we find a completely general statement of the rules that we now write as

$$1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6},$$
$$1^3 + 2^3 + \cdots + n^3 = (1 + 2 + \cdots + n)^2.$$

Number Theory

Aryabhata considered a number theory problem connected with the theorem now called the Chinese remainder theorem (it was stated in a Chinese treatise written about a century earlier than the *Aryabhatiya*). The problem considered by Aryabhata is to find a number leaving specified remainders when divided by specified integers. However, the relevant passage in the *Aryabhatiya* is a very enigmatically stated rule, from which the problem to be solved must be inferred. The text was obscure, even to other mathematicians writing in Sanskrit, who found it necessary to clarify it with commentary, and so we shall not state it at this point.

9.4.2 Brahmagupta

About a century after Aryabhata another of the middle-period Hindu mathematicians, Brahmagupta, was born in the city of Sind, now in Pakistan. He was primarily an astronomer, but his astronomical treatise, the *Brahmasphutasiddhanta* (literally "The Corrected *Brahma Siddhanta*," contains several chapters on computation (*Ganita*). It was this work, translated into Arabic at the instigation of the Caliph Al-Mansur in the eighth century, that brought Hindu astronomy to the Muslims.

Arithmetic

Brahmagupta gives rules for handling common fractions. Although these rules are now commonplace, it should be remembered that they are by no means obvious, being unknown to the Egyptians and the Greeks. In addition, Brahmagupta's arithmetic contains some original ways of looking at many things that we take for granted. For example, to do a long division with remainder, say, $\frac{750}{22}$, he would look for the next number after 22 that divides 750 evenly (25) and write

$$\frac{750}{22} = \frac{750}{25} + \left(\frac{750}{25}\right)\frac{3}{22},$$

that is,

$$\frac{750}{22} = 30\left(1 + \frac{3}{22}\right) = 30 + \frac{90}{22} = 34\frac{1}{11}.$$

Beyond these simple operations, he also codifies the methods of taking square and cube roots, and he states clearly the rule of three. This rule used to be a part of every child's mathematical education, but has not been taught under that name in American schools for several decades. It answers familiar problems of the type, "If three bananas cost 75 cents, how much will seven bananas cost?" Here one is given three numbers and asked to find a fourth number in direct proportion. Brahmagupta names the three terms the "argument " (3), the "fruit" (75), and the "requisition" (7), and points out that the argument and the requisition must be the same kind of thing (in this case bananas). The unknown number he calls the "produce," and he gives the rule that the produce is the requisition multiplied by the fruit and divided by the argument.

Geometry

Brahmagupta also gives some geometric results; and, like Aryabhata, he is not always accurate. One of his best is the formula for the diameter of the circle through the vertices of a triangle. If a and b are two sides, and p the altitude to side c, this diameter is correctly given as ab/p. (This result, in a different mathematical language, is found in Ptolemy.) He was particularly interested in finding quadrilaterals that can be inscribed in a circle and have rational sides. He is most famous for theorems still known by his name. One of these gives a rule for finding the diagonals of a quadrilateral in terms of the sides, and the other reads as follows: *Half the sum of the sides set down four times and severally*

lessened by the sides, being multiplied together, the square root of the product is the area. In our terms this says that the area of a quadrilateral of sides a, b, c, and d is $\sqrt{(s-a)(s-b)(s-c)(s-d)}$, where s is half of the sum of the lengths of the sides. (The case when $d = 0$, which is a triangle, is known as *Heron's formula*). Brahmagupta did not mention the restriction that the quadrilateral must be inscribed in a circle.

9.4.3 Linear Congruences and *Kuttaka*

Brahmagupta gives rules for handling sums of arithmetic progressions, stating the simpler of the two rules given by Aryabhata. His most notable contribution to algebra, however, involves the systematic introduction of zero and negative numbers. He gives the correct rules for manipulating them in the eighteenth chapter of the *Brahmasphutasiddhanta*, which is devoted to a special method of solving the problem now known as the Chinese remainder problem. This topic was developed to a very high degree of sophistication.

The method is called the *kuttaka* (pulverizer). It was greatly simplified by later commentators, and we shall confine our discussion to the refined version. To explain the method let us first examine the purely mathematical problem it was designed to solve, that of finding integers x and y such that

$$ax = by + c,$$

where a, b, and c are given integers. The heart of the problem is an application of the Euclidean algorithm for finding the greatest common divisor of two integers. Let us assume that a and b are relatively prime, so that their greatest common divisor is 1. The Euclidean algorithm proceeds as follows:

$$\begin{aligned} b &= q_1 a + r_1, \\ a &= q_2 r_1 + r_2, \\ r_1 &= q_3 r_2 + r_3, \\ r_2 &= q_4 r_3 + r_4, \end{aligned}$$

and so forth, where $a > r_1 > r_2 > \cdots > 0$. To illustrate the method, let us assume that $r_4 = 1$. Then, applying the first equation, we can rewrite the desired equation $ax = by + c$ as

$$ax = (q_1 a + r_1)y + c, \quad \text{thatis, } az = r_1 y + c,$$

where $x = q_1 y + z$. Then, applying the second equation,

$$(q_2 r_1 + r_2)z = r_1 y + c, \quad \text{thatis, } r_2 z = r_1 u + c,$$

where $y = q_2 z + u$. Continuing, we find

$$r_2 z = (q_3 r_2 + r_3)u + c, \quad \text{thatis, } r_2 v = r_3 u + c,$$

where $z = q_3u + v$. Finally, since $r_4 = 1$, we get

$$(q_4r_3 + 1)v = r_3u + c, \quad \text{that is,} \quad v = r_3w + c,$$

where $u = q_4v + w$.

At this point, finding x and y amounts to solving the simultaneous equations

$$\begin{aligned} x &= q_1y + z, \\ y &= q_2z + u, \\ z &= q_3u + v, \\ u &= q_4v + w, \\ v &= r_3w + c. \end{aligned}$$

This is a system of five equations in the six unknowns x, y, z, u, v, w, and w can be arbitrary. Making the assignment $w = 0$, we nowadays would write this system as the matrix equation

$$\begin{bmatrix} x \\ y \\ z \\ u \\ v \end{bmatrix} = \begin{bmatrix} q_1y + z \\ q_2z + u \\ q_3u + v \\ q_4v \\ c \end{bmatrix}.$$

Notice that the matrix form of this system indicates how it is to be solved. The right-hand matrix tells just what to do with the rows of the left-hand matrix: $x = q_1y + z$, that is, the top row of the matrix (containing x) is obtained by multiplying the row just below it (containing y) by the first quotient (q_1) from the Euclidean algorithm, and then adding the second row below it (containing z). A similar assertion holds for the second row, using the second quotient instead of the first: $y = q_2z + u$. Obviously this procedure will work in general, at least as far as the third row from the bottom; each row is found by multiplying the row just below it by the corresponding quotient and adding the second row below it. If we adjoin a new below the bottom row and put a zero in it, this procedure even works for the second row from the bottom. The bottom row here simply contains the original data c, which is directly assigned to the variable v. That assignment gets the solving procedure started, and we proceed upwards from the bottom:

$$\begin{aligned} v &= c \\ u &= q_4c \\ z &= q_3q_4c + c \\ y &= q_2q_3q_4c + q_2c + q_4c \\ x &= q_1q_2q_3q_4c + q_1q_2c + q_1q_4c + q_3q_4c + c. \end{aligned}$$

These observations are the basis of the *kuttaka* algorithm.

The astonishing sophistication of early Hindu mathematics is shown by the fact that this matrix manipulation was known in India in the eighth century. The

kuttaka method consists of the following algorithm for solving the congruence $ax = by + c$, with $b > a > 0$ and a and b relatively prime. Write the quotients from the Euclidean algorithm (carried out until 1 appears as a remainder) in order in a column, and beneath them write the additive term (c), and below that term a zero. (The zero is inserted so that the same transformation rule applies at the beginning as in all other steps of the algorithm.) Then reduce the number of rows successively by operating on the bottom three rows at each stage. The second-from-last row is replaced by its product with the next-to-last row plus the last row; the next-to-last row is simply copied, and the last row is discarded. Thus to solve this system the *kuttaka* method amounts to the transformations

$$\begin{matrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ c \\ 0 \end{matrix} \longrightarrow \begin{matrix} q_1 \\ q_2 \\ q_3 \\ q_4 c \\ c \end{matrix} \longrightarrow \begin{matrix} q_1 \\ q_2 \\ q_3 q_4 c + c \\ q_4 c \end{matrix} \longrightarrow \begin{matrix} q_1 \\ q_2 q_3 q_4 c + q_2 c + q_4 c \\ q_3 q_4 c + c \end{matrix} \longrightarrow$$

$$\longrightarrow \begin{matrix} q_1 q_2 q_3 q_4 c + q_1 q_2 c + q_1 q_4 c + q_3 q_4 c + c \\ q_2 q_3 q_4 c + q_2 c + q_4 c. \end{matrix}$$

(A word of caution is needed at this point. If the number of quotients is odd, one must put $-c$ instead of c after the quotients, as one can see from the discussion given above. If the Euclidean algorithm had terminated with $r_3 = 1$, we would have set $v = 0$, and the equation $r_2 v = r_3 u + c$ would have meant $u = -c$. The remainders in this process appear on the same side of the equation as their predecessor, but are on the opposite side at the next step.)

The sophistication of this method does not end with matrix manipulations. The two entries in the last matrix give values of x and y satisfying the linear congruence. As we have seen, Diophantus showed how to find a particular solution of such a congruence. The Hindus, however, found *all* the solutions. They took the solutions x and y obtained by the *kuttaka* method, which were generally quite large numbers, divided x by b and y by a, replaced them by the remainders, and gave the general x and y as a pair of arithmetic sequences with differences b and a, respectively. Brahmagupta's rule for finding the solutions is more complicated than the discussion just given, since he does not assume that the numbers a and b are relatively prime. (If the greatest common divisor of a and b is not a factor of c, the problem is impossible; if it is a factor of c, it can be divided out of the problem.)

If the number of quotients in the Euclidean algorithm is odd, the last nonzero row of the matrix will be $-c$ rather than c, so that all the terms will become negative. Thus when the expressions for x and y are reduced by taking the remainders r and s when $|x|$ and $|y|$ are divided by b and a respectively, the solutions must be taken as $\xi = b - r$ and $\eta = a - s$. Brahmagupta gives this rule also. He considers such congruences with negative data as well, and is not in the least troubled by this complication. It seems clear that the name *pulverizer* was applied because the original data are repeatedly broken down by the Euclidean algorithm (they are "pulverized").

Applications of the *Kuttaka* Method

The *kuttaka* method is ideal for solving simultaneous linear congruences. For example, suppose we need to find a number n whose remainder on division with 11 is 6 and whose remainder on division by 9 is 5. We are thus trying to solve $9x + 5 = 11y + 6$, that is, $9x = 11y + 1$. Following the *kuttaka* algorithm, since $11 = 1 \cdot 9 + 2$ and $9 = 4 \cdot 2 + 1$, we perform the operations

$$\begin{matrix} 1 \\ 4 \\ 1 \\ 0 \end{matrix} \longrightarrow \begin{matrix} 1 \\ 4 \\ 1 \end{matrix} \longrightarrow \begin{matrix} 5 \\ 4 \end{matrix},$$

so that $x = 5$ and $y = 4$. The number is thus $9 \cdot 5 + 5 = 11 \cdot 4 + 6 = 50$, and the general solution is $n = 50 + 99k$, that is, n must leave a remainder of 50 when divided by 99.

Brahmagupta applied this technique to solve certain problems connected with the calendar. To see how the method might be used for this purpose, consider that a year is about $365\frac{1}{4}$ days long and a lunar month is about $29\frac{1}{2}$ days long. If we take as a unit of time one-fourth of a day, then a year is 1461 units long and a month is 118 units long. A full moon occurred on January 27, 1994. Let us now ask what is the next year in which there will be a full moon on February 14. Since February 14 is 18 days, or 72 time units after January 27, we need to solve the problem

$$118x = 1461y + 72.$$

By following the *kuttaka* algorithm, you can easily find that $y = 22$, that is, this model predicts that the moon will be full on February 14, 2016.[2] Because a sequence of simultaneous congruences can be solved two at a time, it is possible to take account of many different astronomical phenomena at once by this method.

Two remarks need to be made here. First, linear congruences were not a mere mathematical curiosity to the Hindus; they provided a method of regulating the calendar from knowledge of the periods of the heavenly bodies. Second, the essence of the pulverizer method involves use of the remainders when one integer is divided by another. If the division comes out even, this remainder is zero. It is interesting that the Sanskrit word for zero (*sunya*, meaning *empty space*) appears in Brahmagupta's treatise only in this connection. It may be that the symbol for zero was invented in connection with the pulverizer method, rather than as a place-holder in a decimal notation.

9.4.4 Bhaskara

Approximately 500 years after Brahmagupta, in the twelfth century, the mathematician Bhaskara was born on the site of the modern city of Bijapur. He is the author of a work bearing the title *Siddhanta Siromani*, in four parts; it is concerned

[2]This problem illustrates the difficulty of applying theory to the physical world. In fact the moon will be full on February 20, 2016. The reader is invited to find the flaw in the reasoning.

with algebra and geometric astronomy. Only the first of these parts, known as the *Lilavati*, and the second, known as the *Vijaganita*,[3] concern us here. Bhaskara says that his work is a compendium of knowledge, a sort of textbook of astronomy and mathematics. The name *Lilavati*, which was common among Hindu women, seems to have been a fancy of Bhaskara himself. Many of the problems are written in the form of puzzles addressed to this Lilavati. As we have already discussed most of the material contained in it in connection with Aryabhata and Brahmagupta, we confine our discussion to topics we have not yet mentioned.

Algebra

The *Lilavati* contains a collection of problems in algebra, which are sometimes stated as though they were intended purely for amusement. For example,

> One pair out of a flock of geese remained sporting in the water, and saw seven times the half of the square-root of the flock proceeding to the shore, tired of the diversion. Tell me, dear girl, the number of the flock.

Like countless other unrealistic algebra problems that have appeared in textbooks over the centuries, this story is a way of posing to the student a specific quadratic equation, namely $\frac{7}{2}\sqrt{x} + 2 = x$, whose solution is $x = 16$.

Combinatorics

Bhaskara gives a thorough treatment of permutations and combinations, which, as we know, already had a long history in India. He describes combinatorial formulas such as

$$\binom{7}{3} = \frac{7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3} = 35$$

by saying

> Let the figures from one upward, differing by one, put in the inverse order, be divided by the same in the direct order; and let the subsequent be multiplied by the preceding and the next following by the foregoing. The several results are the changes by ones, twos, threes, etc.

He illustrates this principle by asking how many possible combinations of stressed and unstressed syllables there are in a six-syllable verse. His solution is as follows:

> The figures from 1 to 6 are set down, and the statement of them, in direct and inverse order is
>
> $$\begin{array}{cccccc} 6 & 5 & 4 & 3 & 2 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{array}.$$

[3]This Sanskrit word means literally "source computation." It is compounded from the Sanskrit root *vij-* or *bij-*, which means *seed*. As we discussed in Chapter 3, the basic idea of algebra is to find one or more numbers (the "source") knowing the result of operating on them in various ways. The word is usually translated as "algebra."

> The results are: changes with one long syllable, 6; with two 15; with three, 20; with four, 15, with five, 6; with all long, 1.

Bhaskara assures the reader that the same method can be used to find the permutations of all varieties of meter. He then goes on to develop some variants of this problem, for example,

> A number has 5 digits and the sum of the digits is 13. If zero is not a digit, find the total number of possible numbers.

To solve this problem, you have to consider the possibility of two distinct digits (for example, 91111, 52222, 13333, 55111, 22333), three distinct digits (for example 82111, 73111) and count all the possible rearrangements of the digits.

Bhaskara reports that the initial syllables of the names for colors "have been selected by venerable teachers for names of values of unknown quantities, for the purpose of reckoning therewith." He proceeds to give the rules for manipulating expressions involving such quantities; for example, the rule that we would write as $(-x-1)+(2x-8)=x-9$ is written

$$\begin{gathered} ya\,\dot{1}\ ru\,\dot{1} \\ ya\,2\ ru\,\dot{8} \\ \text{Sum}\ \ ya\,1\ ru\,\dot{9}, \end{gathered}$$

where the dots indicate negative quantities. The syllable ya is the first syllable of the word for *black*, and ru is the first syllable of the word for *species*.

Bhaskara gives the usual rule for solving a quadratic equation by radicals, then goes on to give a criterion for a quadratic equation to have two (positive) roots. He also says that "if the solution cannot be found in this way, as in the case of cubic or quartic equations, it must be found by the solver's own ingenuity."

Bhaskara says explicitly (in the *Vijaganita*) that a nonzero number divided by zero gives an infinite quotient.

> This fraction $\left[\frac{3}{0}\right]$, of which the denominator is cipher, is termed an infinite quantity.
>
> In this quantity consisting of that which has cipher for its divisor, there is no alteration, though many be inserted or extracted; as no change takes place in the infinite and immutable GOD, at the period of the destruction or creation of worlds, though numerous orders of beings are absorbed or put forth.

In his astronomical work Bhaskara gives one procedure that looks like a prefiguration of infinitesimal methods, in the following statement:

> The product of the cosine of the semidiameter by the element of the radius gives the difference of the two sines.

This says, in our terms, that $\sin y - \sin x = (y-x)\cos y$, which, if x is close to y, amounts to the statement that the derivative of the sine is the cosine. There is

no known case where Bhaskara applied this approximation to other functions, and he did not develop the notion of a derivative. However, he did use infinitesimal methods similar to cylindrical approximations for finding the area of a sphere, and so anticipated the essence of the integral calculus.

9.5 The Continuing Tradition

Bhaskara lived before the Muslim conquest of India. Although dates are uncertain, other Indian mathematicians slightly later than Bhaskara found power series for the arcs corresponding to various angles. In particular an early-sixteenth-century work contains in full generality a description of a rule for forming the series now used for the arctangent function. The author (named Jyesthadeva) says that an arc θ on a circle of radius r is given by

$$\theta = r\left(\frac{\sin\theta}{\cos\theta} - \frac{\sin^3\theta}{3\cos^3\theta} + \frac{\sin^5\theta}{5\cos^5\theta} - \frac{\sin^7}{7\cos^7\theta} + \cdots\right).$$

The author says explicitly that this will work only if the arc is less than half of a quadrant of a circle, showing that he understood the concept of convergence of an infinite series. This series made it possible to compute the value of π to 10 decimal places. Commentaries written during this time express a firm conviction that the circumference and diameter of a circle are not commensurable.

India is an integral part of the modern mathematical world, possessing a large number of mathematical publications and excellent mathematicians. Indian mathematicians, working in India, Europe, and North America, have made contributions to mathematical research far out of proportion to their numbers. India has a special claim to pride in this area, both because of the general excellence of its mathematicians and because India produced one of those rare geniuses who appear only once in hundreds of years. No doubt such geniuses are born considerably more often, but only a few of those who are born with the talent are able to express it.

9.5.1 Srinivasa Ramanujan

The topic of power series is one in which Indian mathematicians had anticipated some of the discoveries in seventeenth- and eighteenth-century Europe. It was a facility with this technique that distinguished Srinivasa Ramanujan (1887–1920), who taught himself mathematics after having been refused admission to universities in India. After publishing a few papers, starting in 1911, he was able to obtain a stipend to study at the University of Madras. In 1913 he took the bold step of communicating some of his results to one of the outstanding analysts of the early twentieth century, G. H. Hardy (1877–1947) of Cambridge University. Hardy was so impressed by Ramanujan's ability that he arranged for Ramanujan to come to England. Thus began a collaboration that resulted in seven joint papers with Hardy, while Ramanujan alone was the author of some thirty others.

Unfortunately, Ramanujan was in frail health, and the English climate did not agree with him. Nor was it easy for him to maintain his devout Hindu practices

so far from his normal Indian diet. He returned to India in 1919, but succumbed to illness the following year. His notebooks were found among the papers of G. N. Watson in the 1960s (1886–1965) and finally published in the mid-1980s.

9.6 Problems and Questions

9.6.1 Hindu Mathematical Problems

Exercise 9.1 Generalize the method given in the *Sulva Sutras* for multiplying an area. What should the base and equal sides of the triangle be if the altitude is to be of length $\sqrt{n}a$, where a is the side of a given square? On what algebraic identity is this fact based?

Exercise 9.2 Solve the horse-and-camel problem described above from the Bakshali manuscript. Does a problem of this sort have any practical application?

Exercise 9.3 Solve the Bakshali manuscript problem of finding a (rational) number x such that $x + 5$ and $x - 7$ are both perfect squares. In how many ways can this be done?

Exercise 9.4 Verify that Aryabhata's rules for surveying are correct.

Exercise 9.5 Given an arithmetic sequence $a, a + d, a + 2d, \ldots, a + nd$, find an expression for the sum $S = a + (a + d) + (a + 2d) + \cdots + (a + nd)$. [You will need the expression for triangular numbers: $1 + 2 + \cdots + n = n(n + 1)/2$.] Solve this equation to obtain n in terms of S, a, and d. Then compare your result with Aryabhata's rule:

> Multiply the sum of the progression (S) by eight times the common difference (d), add the square of the difference between twice the first term (a) and the common difference, take the square root of this, subtract twice the first term, divide by the common difference, add one, and divide by two. The result will be the number of terms.

Exercise 9.6 Perform the division $\frac{980}{45}$ following the method used by Brahmagupta.

Exercise 9.7 Why is it necessary that a quadrilateral be inscribed in a circle in order to compute its diagonals knowing the lengths of its sides?

Exercise 9.8 Show that the formula given by Brahmagupta for the area of a quadrilateral is correct if and only if the quadrilateral can be inscribed in a circle.

Exercise 9.9 Show that the *kuttaka* method yields the *negatives* of the solutions of $ax = by + c$ if the total number of quotients is odd. Carry out the argument in detail assuming that $r_3 = 1$ in the example given.

Exercise 9.10 We saw above that $n = 50 + 99k$ gives all the solutions of the simultaneous equations $n = 11x + 6$ and $n = 9y + 5$. Find all solutions to the system consisting of these equations and the equations $n = 7z+2$ and $n = 5w+1$. [*Hint:* Start with the equation $7z + 2 = 99k + 50$ and proceed as in the text. The answer is $n = 2921 + 3465p$.]

Exercise 9.11 Verify the computation that implies that the moon will be full on February 14, 2016. [*Hint:* Be careful here; the number of quotients is odd.]

Exercise 9.12 Solve the following problem from Brahmagupta: What number, divided by 6, leaves a remainder of 5, and divided by 5 a remainder of 4, and by 4 a remainder of 3, and by 3 a remainder of 2? [*Hint:* As you proceed, you will find some equations in which the coefficients have a common factor. Be sure to divide out this factor, so that the *kuttaka* procedure will work as described above.]

Exercise 9.13 Brahmagupta gives the sidereal periods of many heavenly bodies, in particular the sun, which he says makes 30 circuits of the ecliptic in $10,960$ days. (In other words, the sun moves $\frac{3}{1096}$ of a sidereal year every day.) Thus he imagines the ecliptic divided into $10,960$ congruent arcs, with the sun starting at the beginning of the first arc on the first day of a $10,960$-day cycle, during which it will make 30 complete revolutions. He then asks how many days (x) have elapsed since the beginning of the 10960 day period if the sun is exactly at the end of arc number 8080. Thus he asks for an integer x such that for some integer y the equation

$$\frac{30}{10960}x = y + \frac{8080}{10960}, \quad \text{thatis,} \quad 30x = 10960y + 8080$$

holds. Use the *kuttaka* method to derive Brahmagupta's solution: $x = 1000$ days (and $y = 2$ sidereal years).

Exercise 9.14 Solve the following linear congruence problem of Bhaskara. *What quantity is it, which multiplied by* 5 *and divided by* 63 *gives a residue of* 7*; and the same multiplied by* 10 *and divided by* 63*, a remainder of* 14*?*

Exercise 9.15 Solve the Bhaskara problem of finding the number of five-digit numbers having no zero digits and sum of the digits equal to 13.

Exercise 9.16 Take $\theta/r = \arctan x$ and use the fact that $\tan t = \sin t/\cos t$ to convert Jyesthadeva's series into the Maclaurin series for $\arctan x$. How many terms of this series would be needed to compute $\arctan 0.5$ to 10 decimal places? [*Hint:* Remember that in an alternating series with terms decreasing in absolute value, the error in stopping after finitely many terms is less than the absolute value of the first neglected term but larger than the absolute value of the sum of the first two neglected terms.]

9.6.2 Questions about Hindu Mathematics

Exercise 9.17 Compare the method of squaring a rectangle in the *Sulva Sutras* with the method given in Euclid, Book II and Book VI (see Figs. 5.5 and 4.7). Are the underlying principles different or merely differently arranged?

Exercise 9.18 Show that the *Sulva Sutra* method of constructing a circle equal to a given square amounts to saying that the radius of the circle should be the weighted average of the radii of the inscribed circle (weighted as $\frac{2}{3}$) and the circumscribed circle (weighted as $\frac{1}{3}$). By looking at these two circles, make a conjecture as to the origin of this approximation.

Exercise 9.19 Compare the method of rounding a square illustrated in Fig. 9.3 with the conjectured source of the Egyptian formula for the area of a circle in Fig. 2.1. What similarities and differences do you notice?

Exercise 9.20 Compare the conjecture given in the text as to the origin of the approximation for $\sqrt{2}$ with the following, due to a later commentator of 1500 C.E. Assume that each side of the square is 12 units long. Then the diagonal has length $12\sqrt{2} = \sqrt{288} = \sqrt{17^2 - 1} \approx 17 - \frac{1}{34}$ [since $\sqrt{1-x} \approx 1 - (x/2)$]. It follows that $\sqrt{2} \approx \frac{17}{3\cdot 4} - \frac{1}{3\cdot 4\cdot 34} = 1 + \frac{1}{3} + \frac{1}{3\cdot 4} - \frac{1}{3\cdot 4\cdot 34}$. Which explanation seems more probable to you? Does either imply the other?

Exercise 9.21 What differences do you notice in the "style" of mathematics in Greece and India? Consider very particularly the importance of logic, the metaphysical views of the nature of such things as lines, circles, etc., and the interpretation of infinite objects.

Exercise 9.22 Does the division of the circle into 360 degrees by the Hindu mathematicians indicate that they received their knowledge of trigonometry from the Greeks?

Exercise 9.23 Besides the sine function, we also use the tangent and secant and their cofunctions. What is the origin of the words *tangent* and *secant* (in Latin), and why are they applied to the objects of trigonometry?

Exercise 9.24 Thinking over the mathematical traditions we have studied, do you find a point in their development at which mathematics ceases to be a disjointed collection of techniques and becomes systematic? What criteria would you use for defining such a point, and where would you place it in the mathematics of Egypt, Mesopotamia, Greece, and India?

Exercise 9.25 Does the similarity between Aryabhata's method of surveying to a method used earlier in China indicate a common source? Why does Aryabhata use this "primitive" surveying technique when he has available a table of sines?

Exercise 9.26 Aryabhata gives no proof of any of his geometric rules, some of which are correct, others not. How might he have arrived at such rules? Given

that he knew how to find the sum of the squares of the first n integers, how could he have used this sum as a guide to obtain the correct formula for the volume of a pyramid?

Exercise 9.27 Recall that Archimedes wrote the *Sand-reckoner* to prove that the universe could be filled with a *finite* number of grains of sand. The necessity of doing so shows that the Greeks had the same psychological difficulties that all people have in distinguishing clearly between "infinite" and "very large." Compare what Archimedes did (Chapter 6) with the following passage from a Jaina work, telling how to reach the largest enumerable number.

> Consider a trough whose diameter is of the size of the earth. Fill it up with white mustard seeds counting them one after another. Similarly, fill up with mustard seeds other troughs of the sizes of the various lands and seas. Still it is difficult to reach the highest enumerable number.

Exercise 9.28 If the Hindus actually used Diophantine equations to predict astronomical phenomena, they must have realized very quickly that these computations are extremely hard to use, for two reasons. First, the date of a full moon, for example, is an integer-valued function of a quantity that varies continuously. It is therefore bound to have sudden jumps. For example, suppose the moon comes into exact opposition at 12:30 A.M. on one particular date and on some later date it is in exact opposition at 11:45 P.M. Simply subtracting the dates to get the time between full moons gives an integer number of days, even though the actual time elapsed is nearly a full day longer than the computed figure. Second, with a Julian or Gregorian calendar the extra fraction of a mean solar day is added only once every four years. Therefore, if one does not know in advance how many years ahead the prediction is going to be, the phenomenon may occur nearly a full day later than computed. Thus one could be in error by almost two full days. Moreover, remainders after division are extremely unstable functions of their data (this fact is the basis of certain modern codes for encryption of information). To make the computations reasonably accurate, then, one needs a very short unit of time, say, one second. But then the question when the moon will be full on a certain date requires the solution of a separate set of equations for each second in the given day, that is, 86,400 different sets of equations! What conclusions do you draw about the application of mathematics to nature?

9.7 Endnotes

1. Much of the information in this chapter comes from two secondary sources: (a) the book *The History of Ancient Indian Mathematics* by C. N. Srinivasiengar (The World Press Private Ltd., Calcutta, 1967); (b) *Ancient Hindu Geometry*, by Bibhutibhushan Datta (Cosmo Publications, New Delhi, 1993).

2. The quotation from the *Bhagabati Sutra* is from the book of Srinivasiengar (op. cit.), p. 27.

3. The English translation of the *Aryabhatiya* is based on the Sanskrit original published in Leyden in 1874. The translator, Professor Walter Eugene Clark of Harvard University, made no claims regarding the age of the Sanskrit original. The English version is *The* Aryabhatiya *of Aryabhata* (University of Chicago Press, 1930).

4. Quotations from the works of Brahmagupta and Bhaskara are based on Colebrooke's 1817 translation: *Algebra with Arithmetic and Mensuration from the Sanscrit of Brahmegupta and Bhascara* (John Murray, London, 1817).

5. The quotation in Exercise 9.27 is from Srinivasiengar (op. cit.), p. 24.

Chapter 10

Chinese Mathematics

10.1 Introduction

The name *China* refers to a region unified under a central government but whose exact geographic extent has varied considerably over the 4000 years of its history. To frame our discussion we shall use the following convenient division into dynastic periods:[1]

1. *Prehistory.* Fossil evidence shows that the area now known as China has been inhabited by human beings for a very long time, at least 30,000 years, and prehuman hominid fossils have been found there dating back at least half a million years. Neolithic settlements at least 6000 years old have been found in the north and northwest of this area. Chinese tradition includes an early dynasty known as the Xia, but no archaeological confirmation of this dynasty has been found.

2. *The Shang Dynasty* (sixteenth to eleventh centuries B.C.E.). The use of bronze began in China about 1600 B.C.E., approximately a thousand years after it had begun in Europe. This technological innovation coincided with the beginning of the first historical dynasty, the Shang. The Shang rulers controlled the northern part of what is now China and had an extensive commercial empire.

3. *The Zhou Dynasty* (eleventh to eighth centuries B.C.E.). The Shang Dynasty was conquered by people from the northwest known as the Zhou. The Zhou empire was extensive, but broken up into many smaller domains. Around the eighth century the power of the Zhou rulers was greatly diminished, and

[1]Because of total ignorance of the Chinese language, the author is forced to rely on translations of all documents. We shall adhere to the system of writing Chinese words in the Latin alphabet with accent marks used in the translation of the book by Li Yan and Du Shiran. In this system the letter *q* is pronounced like the *ch* in *church*, *x* like the *sh* in *shoe*, *c* like the *ts* in *cats*, *z* like the *dz* in *adze*, and *zh* like the *j* in *jump*. However, we shall omit the accents used to indicate the pitch of the vowels, since these cannot be pronounced by foreigners without special training.

there was a long period of strife as their vassals struggled for supreme power. It was during this chaotic period that the great Chinese philosophers known in the West as Confucius, Mencius, and Lao-tzu lived and taught.

4. *The Period of Warring States* (403–221 B.C.E.) *and the Qin Dynasty* (221–206 B.C.E.). Warfare was nearly continuous in the fourth and third centuries B.C.E., but in the second half of the third century the northwestern border state of Qin gradually defeated all of its rivals and became the supreme power under the first Qin emperor. The name *China* is derived from the Qin. In order to maintain an efficient chain of command the emperor organized the country into 36 military provinces. He also ordered that a series of defensive walls that had been built in the past be connected to form the famous Great Wall. Traditional accounts of the great emperor say that in order to unify his people and wean them away from their provincialism he made a concerted attack on all traditions, ordering that all books be burned (with a few exceptions) and making any appeal to tradition against his authority a capital crime.

5. *The Han Dynasty* (206 B.C.E.–220 C.E.). The empire was conquered shortly after the death of the great emperor by people known as the Han, who expanded their control far to the south, into present-day Viet Nam, and established a colonial rule in the Korean peninsula. During the Han Dynasty commerce and science flourished, and trade became established between the Roman Empire and China, both of which were in their period of maximum prosperity during the second and third centuries C.E. This trade was conducted overland, with the Roman merchants traveling eastward from Syria over the Pamir Mountains and the Chinese traveling westward, the two meeting at a site in present-day Turkestan. This trade ultimately worked to the economic disadvantage of the Roman Empire, since the taste for silk and precious gems led to a trade imbalance and the devaluation of Roman currency.

6. *The Tang Dynasty* (seventh and eighth centuries). The Chinese Empire began its "decline and fall" nearly two centuries before the Roman, early in the third century with the fall of the Han Dynasty. Just as the Roman decline was simultaneous with the rise of Christianity, the Chinese decline coincided with the rise of Buddhism, which was spread by missionaries from India. This contact led to cultural exchanges as well. Hindu trigonometry may have come to China at this time, and certain Chinese geometric techniques seem to be reflected in later Hindu writing. Recovery also came some centuries earlier in China than in the West, under the Tang Dynasty. The Tang Dynasty was a period of high scholarship, in which, for example, block printing was invented. The geographical boundaries of China expanded during this period, and there was extensive commercial contact with Persia, which had only recently become an Islamic country. Decline set in after a disastrous military defeat in the year 751, as a result of which China lost Turkestan to the Prince of Tashkent.

7. *The Song Dynasty* (960–1279). The period of disorder after the fall of the Tang Dynasty ended with the accession of the first Song Emperor. Confucianism underwent a resurgence in this period, supplementing its moral teaching with metaphysical speculation. As a result a large number of scientific treatises, on chemistry, zoology, and botany were written, and—what is of most interest to us—the Chinese became the world's most advanced algebraists. In the last century of this period China was split into two rival states, the Song, and the Qin (not related to the Qin for whom China was named).

8. *The Mongol conquest and the closing of China.* The Song Dynasty was ended in the thirteenth century by the Mongol conquest under the descendants of Genghis Khan, whose grandson Kublai Khan was the first emperor of the dynasty known to the Chinese as the Yuan. As the Mongols were Muslims, this conquest brought China into contact with the intellectual achievements of the Muslim world. Knowledge flowed both ways, of course, and the sophisticated Chinese methods of root extraction seem to be reflected in the works of later Muslim scholars such as the fifteenth-century mathematician Al-Kashi. The vast Mongol Empire facilitated East-West contacts, and it was during this period that Marco Polo made his famous voyage to the Orient.

9. *The Ming Dynasty* (fourteenth to seventeenth centuries). While the Mongol conquest of Russia lasted 240 years, the Mongols governed China so incompetently that they were driven out in less than a century by the first Ming Emperor. In the Ming Dynasty Chinese trade and scholarship rapidly recovered. The effect of the conquest, however, was to encourage Chinese isolationism, which became the official policy of the later Ming emperors during the period of European expansion. The first significant European contact came in the year 1582, when the Jesuit priest Matteo Ricci arrived in China. The Jesuits were particularly interested in bringing Western science to China to aid in converting the Chinese to Christianity. They persisted in these efforts despite the opposition of the Emperor. The Ming Dynasty ended in the mid-seventeenth century with conquest by the Manchus.

10.2 Aspects of Chinese Mathematics

There is a stream of Chinese mathematical literature extending from more than 2000 years ago into the present. Throughout this long period of time there have been outstanding intellects working on mathematical problems and writing their thoughts in Chinese. The earliest period of indigenous Chinese mathematics shows the influence of the needs of administration. This mathematics consists of the geometry and arithmetic needed to solve problems in surveying, taxation, and commerce. From this practical beginning purely mathematical questions arose, just as in Greece. In pure mathematics the Chinese worked on some problems that were not considered in the West until much later, such as magic squares and

the solution of simultaneous linear congruences. In addition, the Chinese had two important mechanical computing devices that stimulated the development of arithmetic and algebra, namely counting rods and the counting board, which mimics a matrix. For example the abstract notion of a negative number, as opposed to a number subtracted from another, was formulated in China at an early date. The counting board helps to bring out the distinction between a *variable* (square on a counting board) and its *value* (number of tallies occupying the square). The period of time involved in the mathematics we shall be discussing ranges from the earliest documents (variously dated from 1200 to 100 B.C.E.) to the Yuan period (fourteenth century C.E.).

10.3 Some Important Early Documents

The origin of mathematics in China dates from the time of the mythological Emperor Yu (ca. 2100 B.C.E.), who is said to have received a diagram called the *Luo Shu*, from a tortoise in the Luo River. There are several such legends, and in one ancient chronicle Yu is described as going about "with a plumb line in his left hand and a gnomon and compass in his right" in the course of survey work as part of a flood control project. The *Luo Shu* is a 3×3 magic square, with diagrams representing the numbers from 1 to 9 in each location so that all rows, columns, and diagonals total 15:

$$\begin{array}{ccc} 2 & 9 & 4 \\ 7 & 5 & 3 \\ 6 & 1 & 8 \end{array}$$

The alternation of even and odd numbers in this magic square reveals a poetic approach to mathematics in harmony with traditional Chinese philosophy. According to the mathematician F. J. Swetz, this square reflected, "a plan of universal harmony based on a cosmology predicated on the dualistic theory of the *Yin* and the *Yang*."

From what was just quoted about the Emperor Yu one can see that the Chinese were familiar with scale drawings and the square (gnomon) and circle (compass) at a very early period. A treatise from about the fourth century B.C.E. called the *Book of Crafts* gives names to certain angles: the right angle (ju), half of a right angle (xuan), and the angles obtained by increasing the latter by 50% in three stages, that is, angles of $67^\circ\, 30'$ (zhu), $101^\circ\, 15'$ (ke), and $151^\circ\, 52'\, 30''$ (qingzhe).

10.3.1 Archaeological Data

The earliest physical evidence of mathematical activity in China comes from oracle bones dated to the Shang Dynasty. Archaeological evidence suggests that a decimal place-value system was used. (See also the Shang numerals, Fig. 10.1.)

Chinese documents from the second century B.C.E. mention the use of counting rods, and a set of such rods from the first century B.C.E. were discovered in 1970. These are the earliest known mechanical computing devices. These devices were

Figure 10.1: The Shang numerals.

used to perform the computations described in the Chinese documents from this period. The rods can be arranged to form the Shang numerals (Fig. 10.1) and thereby represent decimal digits. They were apparently used in conjunction with a counting board (a board ruled into squares, so that each column represents a particular item). A picture of such a board is almost exactly what we now call a matrix, and there is no doubt that the proficiency with which the Chinese handled linear equations was due to this system of representation. It seems reasonable that the early development of a place-value decimal system in China was facilitated by the use of these mechanical devices. This development, as we have seen, cannot be taken for granted, since none of the European civilizations made the discovery. (The cuneiform texts have a place-value system, but one based on 60 and clearly superimposed on an earlier decimal system that was not place-value, since separate symbols exist for units and tens.) Even more striking is the fact that black rods were used to represent positive numbers and red ones negative. Just when this innovation was made is uncertain, but it is mentioned in a third-century commentary known as the *Nine Chapters on the Mathematical Art*.

10.3.2 The *Arithmetical Classic*

The earliest Chinese mathematical document still in existence, the *The Arithmetical Classic of the Gnomon and the Circular Paths of Heaven* (*Zhoubi suanjing*), is concerned with astronomy and the applications of mathematics to the study of the heavens. The title apparently refers to the use of the sundial or gnomon in astronomy.

This document was begun before the third century B.C.E. and contains a famous diagram (sketched in Fig. 10.2) giving a proof of the Pythagorean theorem for the special case of a 3–4–5 triangle. (The general statement of the theorem occurs later in the document.)

The *Arithmetical Classic* dates from the Han Dynasty (206 B.C.E.–220 C.E.) and, as just mentioned, is concerned mostly with the astronomical and geographic applications of geometry, for which the Pythagorean theorem is essential. The vertical bar on a sundial was called a *gu* in Chinese, and its shadow on the sundial was called *gou*; for that reason the Pythagorean theorem was known as the *gougu* theorem. The treatise says that "The Emperor Yu can rule the country because of the gougu theorem." A commentator on this book named Zhao Shuang (third century C.E.) made advertising claims for geometry unequaled at any time before or since, even by the most enthusiastic proponents of the subject. The Emperor Yu was credited with saving his people from floods and other great calamities, and, "This is made possible because of the gougu theorem. . . ."

The use of the gnomon in surveying is a very simple exercise in proportion, but

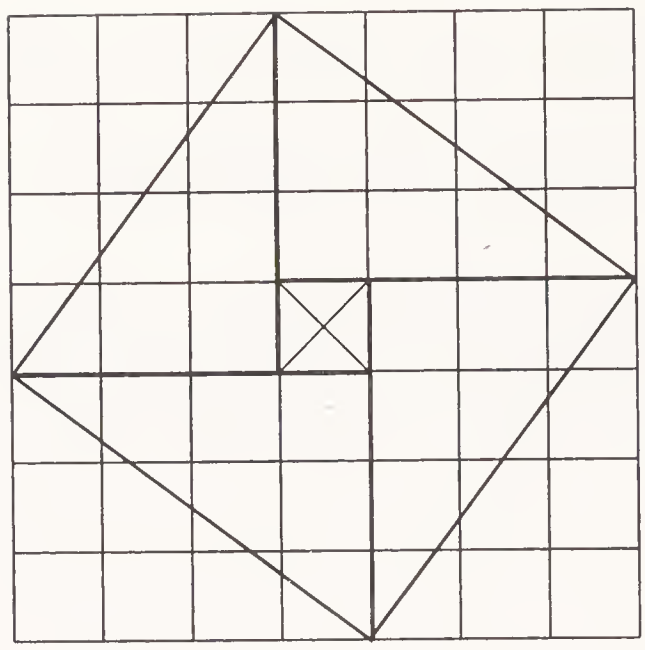


Figure 10.2: Chinese proof of the Pythagorean theorem.

observations using the gnomon cannot be made from an arbitrary location. One must be able to move to a suitable place. The instructions in the *Arithmetical Classic* for use of the gnomon are very simple: "Align the gnomon with the plumb line to determine the horizontal, lay down the gnomon to find the height, reverse the gnomon to find the depth, lay the gnomon flat to determine the distance." The principle is thus the basic principle of trigonometry, similar triangles. However, unless one has a variety of gnomons with different ratios of legs, it will be necessary to move to just the right location in order to determine, say, the height of a tall tree. Notice that this kind of surveying does not require the measurement of any angles, only lengths. It is interesting that when surveyors began to attack more complicated measurement problems, their first instinct was to make more measurements with one rigid instrument, rather than designing an instrument that would provide variable angles.

The *Arithmetical Classic* was completed in its present form some time between 100 B.C.E. and 100 C.E. As in the case of ancient Greek documents, however, the oldest extant manuscript is much more recent, dating from 1213 C.E. The astronomical content of the *Arithmetical Classic* is shown in the use of the Pythagorean theorem to calculate the height of the sun, given that at the summer solstice a stake eight *chi* high casts a shadow six *chi* long, and that the shadow length decreases by one *fen* for every thousand *li* that the stake is moved south, casting no shadow at all when moved 60,000 *li* to the south. Using the $6 : 8$ proportion, the author reasons that the sun is $80,000$ *li* high.

Although the proportions are accurate here, the geometry is slightly wrong, since the length of a real shadow does not vary directly with the distance traveled. Also, the computation is clearly based on a flat earth. Finally, the lengths are not realistic, since one *chi* is about 10 inches long (one *fen* is about 1 inch). One *li* is 1800 *chi*, that is, about 1500 feet. Thus 60,000 *li* are about 18,000 miles. Later Chinese commentators recognized this inaccuracy, and in the eighth-century C.E. an expedition to survey accurately a line of longitude found the actual lengthening (at certain latitudes) to be four *fen* per thousand *li*.

10.3.3 The *Nine Chapters* and Liu Hui

The earliest Chinese work on pure and applied mathematics, resembling in style the Ahmose Papyrus from Egypt, is the *Nine Chapters on the Mathematical Art* (*Jiuzhang suanshu*), probably assembled in its present form in the first century C.E. It is said to have been recovered during the Han Dynasty. A commentary on this work by Liu Hui dates from the third century C.E. Like the Ahmose Papyrus, the *Nine Chapters* consists of a set of pure and applied problems set out and solved. The first of the nine chapters is called "Field Measurement" (*Fang tian*). It contains the computations of areas of rather complicated shape, such as the area of a segment of a circle, a segment of a sphere, and an annulus. (The first two of these are given only approximately, and the author assumes the circumference is three times the diameter.)

The remaining eight chapters bear the following titles: "Cereals," "Distribution by Proportion," "What Width?," "Construction Consultations," "Fair Taxes," "Excess and Deficiency," "Rectangular Arrays," and "gougu" (The Pythagorean theorem). These titles and descriptions of contents suggest a compendium of engineering and administrative problems such as are found in the Ahmose Papyrus. Its role in Chinese mathematics is much more central than that of the Ahmose papyrus in Western mathematics, however. As Swetz has remarked, "its influence on Oriental mathematics may be likened to that of Euclid's *Elements* on western mathematical thought."

10.3.4 Linear Equations

The *Nine Chapters* contains 246 word problems, including the following example of what we now call linear algebra:

> There are three kinds of wheat. The grains contained in two, three and four bundles, respectively, of these three classes of wheat, are not sufficient to make a whole measure. If however we add to them one bundle of the 2nd, 3rd, and 1st classes, respectively, then the grains would become one full measure in each case. How many measures of grain does then each one bundle of the different classes contain?

The following counting-board arrangement is given for this problem.

1		2	1st class
	3	1	2nd class
4	1		3rd class
1	1	1	measures

Here the columns from right to left represent the three samples of wheat. Thus the right-hand column represents 2 bundles of the first class of wheat, to which one bundle of the second class has been added. The bottom row gives the result in each case: 1 measure of wheat. The word problem might be clearer if the final result is thought of as the result of threshing the raw wheat to produce pure

grain. We can easily, and without much distortion in the procedure followed by the author, write down this counting board as a matrix and solve the resulting system of three equations in three unknowns. The author gives the solution: a bundle of the first type of wheat contains $\frac{9}{25}$ measures, a bundle of the second $\frac{7}{25}$ measures, and a bundle of the third $\frac{4}{25}$ measures.

The significance of this problem is that in order to solve it, one must subtract bundles of one kind of wheat from bundles of another kind. Since this is a physical impossibility, one must therefore have some concept of a negative number, that is, a deficiency of a thing. The use of red counting rods to represent a deficiency and black rods for a positive amount handles this difficulty. The commentator Liu Hui explained the procedure to be followed when adding and subtracting quantities of opposite sign, or when subtracting a quantity from zero where necessary, for example, in solving the linear system of equations given above.

The solution of linear equations such as those occurring in this example requires skill at manipulating fractions, which, as we have seen, was by no means widespread in the ancient world. Of the groups we have discussed so far, only the Babylonians used a place-value system to represent fractions. The Chinese made this discovery at an early stage, and the *Nine Chapters* also contains the oldest exposition of the use of common fractions (those written in quotient form), including the idea of least common denominator. In comparison with the Ahmose Papyrus, which it resembles in the problems discussed, the *Nine Chapters* contains a much more efficient system of computation. The method of solving the problem is that of successive elimination of variables by adding and subtracting the equations from one another This method was called *fang cheng*, which meant originally "rectangular computing." It now means simply equations.

10.3.5 Square Roots and Quadratic Equations

In the *Nine Chapters* there is a method of extracting square roots that is equivalent to the computational procedure used in Europe and America until the late twentieth century, when calculators made it obsolete for computational purposes. In China, however, this computational process had a theoretical influence of great importance, and for that reason we shall discuss Problem 12 in the Chapter "What Width?". This problem requires extracting the square root of 55,225. The procedure is reasoned out as follows.

Since the number of digits in n^2 is either twice the number in n or one less than twice that number, we group the digits in pairs, writing 5 52 25. It is then clear that the square root is between 200 and 300. This information gives us the first digit (2). Thus, for a given integer N (in this case 55,225), we have found a first approximation a_1 (the first digit, in this case standing for 200) for $\sqrt{N}$. This is the initial step. The rest of the procedure consists of replacing a_1 by better and better approximations. We note that if the exact square root is $a_1 + h$, then

$$N = (a_1 + h)^2 = a_1^2 + 2ha_1 + h^2.$$

A suitable choice of h will give us a second approximation $a_2 = a_1 + h$. Noticing

that the best possible value of h would satisfy

$$h = \frac{N - a_1^2}{2a_1 + h},$$

we see that one algorithm (nowadays called the Newton-Raphson algorithm) is to take $h = (N - a_1^2)/2a_1$. For a pencil-and-paper process, this algorithm is modified slightly. Instead of the h just given the largest integer in this number (rounded downward to one significant digit) is usually taken as the "trial"; if the trial proves to be too large, it is replaced by the next smaller digit, until finally $a_2 = a_1 + h$ is chosen as large as possible with its square not exceeding N. In the present case we should take

$$h = \frac{55,225 - 40,000}{800} = \frac{15,225}{400}.$$

In practice, since we are trying to find only one digit at a time, we would use only the first digit of $\frac{152}{4} = 38$ as the trial value of h. That is we take 3 as our next digit, $(h = 30)$. Thus our second approximation becomes $a_2 = 230$. We now need the difference $N - a_2^2 = N - (a_1 + h)^2 = N - a_1^2 - 2a_1h - h^2$. Since we have already computed $N - a_1^2 = 15,225$, we need only compute $2a_1h + h^2 = h(2a_1 + h) = 30(430) = 12900$, then subtract from $N - a_1^2$, that is, $15,225 - 12,900 = 2325$. We now start over again, trying to find $a_3 = a_2 + h$, with a new h obtained by adjusting $(N - a_2^2)/2a_2 = \frac{2325}{460}$. We therefore take $h = 5$ and get $a_3 = 235$. This time when we compute $N - a_3^2 = N - a_2^2 - 2ha_2 - h^2$, we find it equal to zero, and so the computation ceases.

The computation is usually arranged as follows, suppressing final strings of zeros in each partial computation:

$$\begin{array}{r} 2\ 3\ 5 \\ \sqrt{5\,52\,25} \\ \quad 4 \\ \hline 40\ \ 1\,52\,25 \\ 43\ \ 1\,29 \\ \hline 460\ \ 23\,25 \\ 465\ \ 23\,25 \end{array}$$

This procedure rests on two observations, namely that when an approximation a to $\sqrt{N}$ is replaced by a better approximation $a + h$,

1. A good way to get a close guess to the exact h is by taking $h = (N - a^2)/2a$.
2. The error of approximation $N - a^2$ becomes $N - a^2 - h(2a + h)$. That is, the adjustment in the error when the approximation a is replaced by $a + h$ amounts to $h(2a + h)$.

In particular, in the computations above, the values of $2a$ and $2a + h$ at each stage appear on the left-hand side as the pairs $40, 43$ (standing for $400, 430$, since final zeros are suppressed) and $460, 465$. The successive values of $N - a^2$ appear as $15,225$, 2325, and 0.

This computational procedure led to an important theoretical advance in China by suggesting a way of solving the general quadratic equation

$$x^2 + px = q.$$

Indeed, the left-hand side here calls to mind the second remark just made. Casting x in the role of h and p in the role of $2a$, we see that the left-hand side represents the amount by which the error $N - a^2 = N - p^2/4$ would change if the "first approximation" $p/2$ to a certain square root $\sqrt{N}$ were adjusted to the "second approximation" $p/2 + x$. If we assume that the second approximation is exactly $\sqrt{N}$, it then follows that $N - p^2/4 = x^2 + px = q$, that is, $N = q + p^2/4$. Thus the number N can be found from the data of the problem. But then the square root of N is $x + p/2$, and so $x = \sqrt{N} - p/2$. This result can be summarized as a formula

$$x = \sqrt{q + \frac{p^2}{4}} - \frac{p}{2}.$$

This reasoning is the essence of Problem 20 of the gougu chapter of the *Nine Chapters*. The problem asks for the solution of

$$x^2 + 34x = 71,000,$$

and gives the answer as 250, found by "using the number 34 in the corollary to the square root method," that is, using 34 in the role of $2a$, and x in the role of h.

The *Nine Chapters* also contains a method for extracting cube roots, based on the fact that $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$, but this procedure did not lead to a corresponding method of solving cubic equations by radicals.

10.3.6 Geometry

The geometric formulas given in the *Nine Chapters* are more extensive than those of the Ahmose Papyrus; for example, there are approximate formulas for the volume of segment of a sphere and the area of a segment of a circle. The implied value of π, however, is $\pi = 3$. It is surprising to find this value in the *Nine Chapters*, since it is known that the value 3.15147 had been obtained in China by the first century, and the third-century commentator Liu Hui refined this value to 3.141024

Liu Hui apparently derived this value by a method similar to that of Archimedes, that is, by use of successive inscribed polygons of 6, 12, 24, 48, 96, and 192 sides, finding that this last polygon, inscribed in a circle of radius 1 *chi* would have an area of $314\frac{169}{625}$ square *fen*. (The meaning of these lengths is explained above.) This reasoning gives a value of π approximately 3.14638, very close to Archimedes' value.

In his commentary on the *Nine Chapters* Liu Hui used a method similar to the Archimedean method for finding the relative volumes of a sphere and cone inscribed in a cylinder. He considered a cube and a cylinder both circumscribed about the sphere and evaluated the ratios of the areas of planar sections of them. He knew that the ratio of the volume of the cube to that of the cylinder was $4 : \pi$.

He then reasoned that the cylinder has a horizontal cross section that is a circle, as does the sphere, and a vertical cross section that is a square, as does the cube. He argued on this basis that the volume of the cylinder is the mean proportional between the volumes of the sphere and cube. If this result were correct, the volume of the sphere would be equal to $\pi/4$ times the volume of the cylinder. In our terms this would make the volume of the sphere $(\pi^2/2)r^3$, a result which would be correct only if $\pi = \frac{8}{3}$.

10.4 The *Sea Island Manual*

In his commentary Liu Hui mentions that the last of the *Nine Chapters*, in which the theory of right triangles is developed, is inadequate. He filled the gap in an extended commentary on this chapter. Some of what is known about the mathematics of surveying in China comes from this commentary on the gougu theorem, which became separated from the rest of Liu Hui's commentary and circulated as an independent treatise known as the *Sea Island Mathematical Manual* (*Haidao suanjing*). This work consists of nine problems. Information on the mathematics of surveying is found in the first problem, from which the work is named. The problem is to compute the height and distance of a mountain on an offshore island (without getting into a boat, of course). The mathematics needed was called the method of double differences (*chong cha*). It is identical to the procedure discussed in the preceding chapter; hence there may be a common source for the use of this method in both India and China. Indeed, this method of surveying was used even more widely, being found in Islamic treatises from medieval times. The idea of making two sightings instead of measuring angles is also a feature of certain methods of surveying found in Europe during the Middle Ages.

10.5 Number Theory

The fundamental problem of divisibility, which is treated in Euclid's Books VII–IX, also occurs in Chinese treatises, in particular in a third-century treatise known as *Master Sun's Mathematical Manual* (*Sunzi Suanjing*), which contains the essence of the result still known today as the Chinese remainder theorem. The problem asks for a number that leaves a remainder of 2 when divided by 3, a remainder of 3 when divided by 5, and a remainder of 2 when divided by 7. This problem involves the notion of congruence. We are trying to find all numbers that leave the given remainders "modulo" the three given primes. The assertion that any number of such congruences can be solved simultaneously if the moduli are all pairwise relatively prime is the content of the Chinese remainder theorem. The author gives the answer to this problem as 23. We have already seen how to solve these problems using the *kuttaka* discussed by Brahmagupta. Master Sun's solution, however, is not placed in the context of a general method.

10.6 Applied Mathematics

Like other peoples, the Chinese needed an accurate calendar, and to obtain it they observed the stars and planets. Indeed, astronomical observation became one of the principal government programs, causing much of the vital information on observation to be buried in official government reports, rather than expounded in systematic treatises. Furthermore independent study of astronomy became nearly tantamount to treason, so that the effect of government support on the development of astronomy may not have been entirely positive. Nevertheless, China provides some of the best astronomical records available for much of the first millennium C.E. During this period several supernovae appeared in the sky, and were recorded only in China. Since supernovae are a sporadic phenomenon not amenable to treatment by geometric astronomy, we shall say no more about them. Likewise, although the retrograde motion of the planets was discussed in China as early as the first century B.C.E., for the sake of simplicity we shall confine our discussion of Chinese astronomy to the theory of the sun's motion and its role in establishing a calendar. Space does not permit the full discussion that would be necessary to take account of the different and conflicting cosmologies used by various Chinese astronomers.

The early Chinese treatises on the calendar, of which *Master Lu's Spring and Autumn Annals* (third century B.C.E.) is representative, describe years of 12 lunar months each and give astronomical characterizations of each of the months. As we know from our previous discussion of the calendar, it is necessary to interpolate 7 extra months in each 19-year period in order to keep the lunar calendar in harmony with the tropical or sidereal year. This 19-year cycle is noted in the *Arithmetical Classic*, as is the more refined fact that one day must be dropped from every four such cycles in order to preserve the harmony. This resemblance to the lunar calendars used in the Mediterranean world has led some scholars to conclude that the earlier Babylonian calendar was imported into China. Other scholars, however, have noted significant differences, such as the division of celestial circles into $365\frac{1}{4}$ equal parts in China rather than 360, as in the West and different groupings of stars into constellations.

The nonuniformity of the sun's motion, which was noted by the astronomer Zhang Heng around the year 100 C.E., caused difficulties in establishing the calendar and in astronomy. For both purposes it is important to know the location of the sun in relation to the stars at any given time, and considerable precision is needed. For example, since the diameter of the sun (and the moon also) is about half a degree on the celestial sphere, an error of one degree can cause the prediction of an eclipse to be entirely wrong. The Chinese astronomers therefore developed some sophisticated methods of interpolation to keep theory in harmony with observation.

The first of these methods, developed in the sixth century C.E., involves essentially quadratic interpolation. That is, observations of the sun's distance from some fixed point, usually the vernal equinox, are made at equal time intervals. The successive increments, which would all be equal if the sun moved uniformly, are found to differ, and by a larger amount than can be explained by observa-

tional errors. In other words, a linear function will not do for interpolation. The second differences, however, that is, the differences of the successive increments, are much smaller because a quadratic function can fit the data much better, and discrepancies in them can be attributed to observational error. Hence the second differences are taken as the foundation of a computational system for predicting the position of the sun. The mathematical problem then arises of reconstructing a quantity depending on time knowing its (constant) second differences. To do so, one must know not only the initial value but also the initial difference. Thus, if the initial value is 0 and the initial difference is 1, while the constant second difference is -0.1, the successive values of the differences will be 1, 0.9, 0.8, 0.7,..., and the successive values of the function will be 1, 1.9, 2.7, 3.4,.... This method was employed to create the Imperial Standard Calendar in the year 600 C.E.

As astronomy grew more sophisticated, the equipment used in making astronomical measurements became more precise. During the Yuan Dynasty a 40-foot-high tower was erected in the capital city with a perfectly level path, known as the *sky measuring scale*, leading away from its center to the north. It was flanked on both sides by a trench filled with water, so that the surface of the water could be used to verify that it was level. On this path the length of the sun's shadow could be measured very accurately and the solstices therefore determined with great precision.

Along with the better observations, the numerical techniques for representing observations grew still more sophisticated, involving third-order differences and the replacement of accumulated discrepancies between observation and linear interpolation by average daily discrepancies. This technique was used by the mathematician/astronomer Guo Shojing and others to produce the "Works and Days Calendar" in the late thirteenth century.

One can easily see how these problems might lead to the development of a systematic theory of finite differences, in which the basic problem is the reconstruction of a function from its differences of a given order plus a certain number of initial values for the function and lower-order differences. It also led to the investigation of progressions and series during the thirteenth and fourteenth centuries.

10.7 Foreign Influences

Cultural contacts between India and China began just before the Tang Dynasty. Hindu astronomers, who had been in contact with Hellenistic mathematics, brought their system of circle measurement based on degrees and the Hindu approach to trigonometry using half-chords rather than Ptolemy's chords. Unmistakable evidence of Hindu influence can be found in a 1299 treatise called *Introduction to Mathematical Studies* (*Suanxue qimeng*) by the mathematician Zhu Shijie. The author of this book introduced names for very large powers of 10, including the term "sand of the Ganges" for 10^{96}. Incidentally, this work later disappeared from China, but was preserved in Korea under the name *Sanhak Kyemong*, where it was eventually (1839) noticed and reprinted in China.

Islamic influences on China, particularly from Persia, began during the Tang Dynasty and became extensive during the time of Mongol rule. It is known from imperial records that many Arabic treatises were translated into Chinese during the Yuan Dynasty. Unfortunately, the works themselves have not survived, and so we cannot know if these works included Arabic versions of the Greek classics. Archaeologists have discovered iron plates from the Yuan Dynasty bearing 6×6 magic squares written in Arabic characters. A graphic method of computing a product (as opposed to the mechanical methods of counting rods or the abacus) also came into China from the Islamic world.

10.8 Later Developments

During the Medieval period in Europe Chinese mathematicians continued to make advances in geometry and algebra. We shall sample a few of these achievements in the present section.

10.8.1 Zu Chongzhi

The fifth-century mathematician Zu Chongzhi made outstanding contributions to mechanics, astronomy, and mathematics. Together with his son Zu Geng, he was the first to act on Liu Hui's remark that the existing computation of the volume of a sphere was incorrect and to find the volume correctly using a technique similar to Archimedes' *Method*. The two mathematicians considered a figure formed by two right circular cylinders of radius a whose axes intersect at at right angles to each other at the center of a cube of side $2a$. The two cylinders formed a figure that they called a *double umbrella* (see Fig. 10.3). A sphere of radius a with center at the center of the cube will be tangent to the double umbrella along two mutually perpendicular great circles. Now consider a horizontal section of the original cube at height h above the middle plane of the cube. In the double umbrella this section is a square of side $2\sqrt{a^2 - h^2}$ and hence area $4(a^2 - h^2)$. Therefore the area outside the double umbrella and inside the cube is $4h^2$. This is the same area as the corresponding cross section of an upside-down pyramid with a square base of side $2a$ and height a. Hence *the volume of the portion of the cube outside the double umbrella in the upper half of the cube equals the volume of a a pyramid with square base* 2a *and height* a. Since this volume is $\frac{4}{3}a^3$, it therefore follows (after doubling, to include the portion below the middle plane) that the region inside the cube but outside the double umbrella has volume $\frac{8}{3}a^3$, and hence that the double umbrella itself has volume $\frac{16}{3}a^3$.

We now compare sections of the double umbrella with those of the sphere. Each horizontal cross section of this sphere is the circle inscribed in the same section of the double umbrella. It therefore has area $\pi(a^2 - h^2)$, or, to stay closer to what seems to be the language of the original document, the ratio of its area to the area of the same section of the double umbrella is the ratio of a circle to the circumscribed square $(\pi/4)$. The volumes must therefore be in this same ratio, that is, the sphere has volume $(\pi/4) \cdot \frac{16}{3}a^3$, or $\frac{4}{3}\pi a^3$.

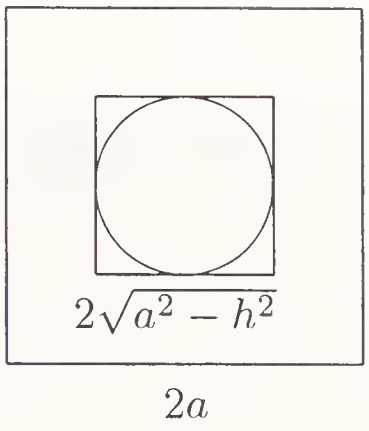


Figure 10.3: Sections of the cube, double umbrella, and sphere at height h. The area between the two squares (the sections of the cube and double umbrella) is $4h^2$.

The principle used here—that if all horizontal sections of two volumes are in a given ratio, then the volumes are in that same ratio—had been used earlier by Archimedes, as we have seen. It was revived independently in Europe a thousand years after the time of Zu Chongzi and Zu Geng and has been traditionally referred to as *Cavalieri's principle*. There is no reason to doubt that this principle was discovered independently in Europe and Asia. The idea of regarding a plane figure as a stack of lines seems to occur naturally as soon as geometry is sufficiently far advanced.

Zu Chongzhi also proved that the value of π lies between 3.1415926 and 3.1415927, which was the greatest accuracy achieved by any civilization until the time of the Islamic mathematicians.

10.8.2 Later Chinese Algebra: Higher-Degree Equations

Mathematics attained a very high level in China during the Song and Yuan dynasties, from about 1000 C.E. to 1400 C.E., which was also a high point for mathematics in the Islamic world in the West and the beginning of the European revival of learning. At this period the Chinese were the most advanced algebraists in the world. They studied equations and classified them according to degree, giving the poetic name of "celestial element" (*tian yuan*) to the unknown. Their approach to higher-degree equations reflects an understanding that the method of solving quadratic equations, based on the square-root algorithm, does not generalize easily to cubic equations. As we have seen, they had found a complete solution of quadratic equations based on the algorithm for extracting square roots at an early date. They did not find a similar algorithm for solving cubic and quartic equations. Instead they developed a method of finding a numerical approximation of a root, similar to a method that was rediscovered independently in the nineteenth century in Europe and is commonly called *Horner's method*. This method was used around the year 1200 C.E. by the mathematician Yang Hui. Because of its efficiency in finding approximate roots the Chinese were not deterred by large coefficients and high-degree equations. Yang Hui's method of solving equations is highly effective from a practical point of view. On the other hand,

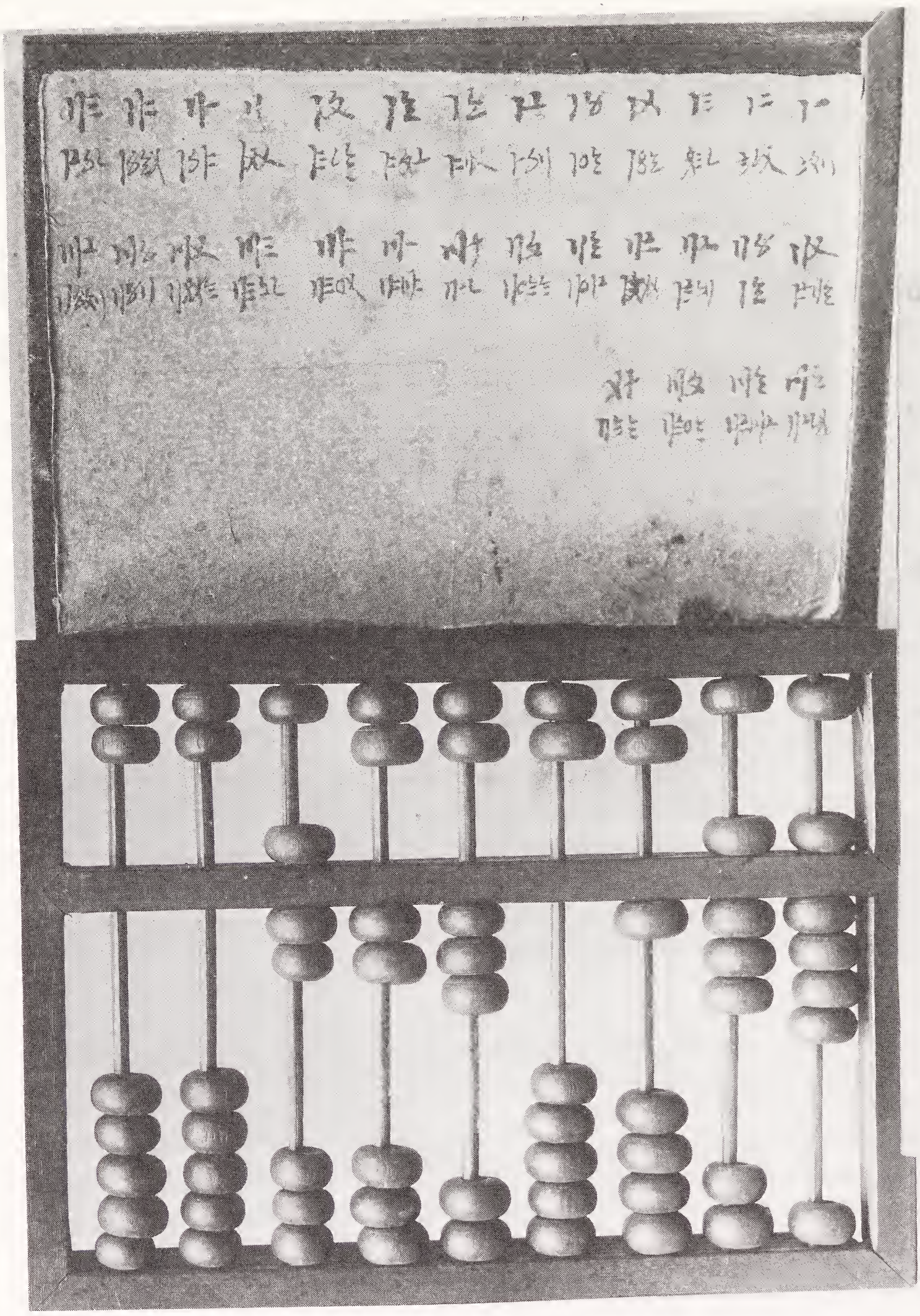


Figure 10.4: The Chinese abacus (*suan pan*). The Bettmann Archive.

the existence of effective numerical methods and computational machinery such as the abacus (Fig. 10.4) may have turned the interest of Chinese mathematicians away from the search for closed-form solutions by radicals, a search that was a powerful stimulus to mathematical advances in the Islamic world and in Europe. It led Islamic scholars such as Omar Khayyam, for example, to study the application of conic sections to the solution of such equations, presenting the solution of a cubic equation as a simultaneous solution of two quadratic equations. This kind of analysis focuses attention on the symmetries of the equation itself rather than the numerical values of its coefficients, and brings out the sequence of operations that must be performed to get from the coefficients to the roots, the so-called solution by radicals. As we shall see, such solutions by radicals can be found for cubic and quartic equations, although the radicals are often misleading as to the numerical value of the solution. The search for a closed-form solution by radicals for the fifth-degree equation, even though no such solution exists, led to the beautiful subject of Galois theory, now taught in universities throughout the world. We see here a good illustration of a principle of compensation: each decision to pursue one line of inquiry causes a different line of inquiry to be neglected, and only when several approaches to a problem have been explored can we see what is missing in each of them.

The so-called "Pascal triangle," for which Pascal is given credit because of his detailed development of its properties, appeared in a Chinese book written by Yang Hui in 1261. The figure was credited by Yang Hui to the eleventh-century mathematician Jia Xian. As we have already seen, it was known centuries earlier in India under the name *Meru Prastara*. However, it may also have been known much earlier in China. Since the two civilizations were in contact from early times, it is difficult to be sure which way any particular idea was passed.

Yang Hui is also the author of *Yang Hui's Methods of Computation*, which became one of the standard textbooks in Korea during the Yi Dynasty (fourteenth to seventeenth centuries). Like Zhu Shijie's *Introduction*, this book was lost from China for many centuries and eventually recovered because it had been reprinted in Korea.

Chinese mathematicians of this period considered geometric problems that lead to higher-degree equations, such as the following thirteenth-century problem from the *Sea Mirror of Circle Measurements* by the mathematician Li Ye (see Fig. 10.5): *Three li north of the wall of a circular town there is a tree. A traveler walking east from the southern gate of the town first sees the tree after walking 9 li. What is the diameter of the town?*

This problem is obviously concocted so as to lead to an equation of higher degree. (The diameter of the town could surely be measured directly from inside, so that it is highly unlikely that anyone would ever need to solve such a problem for a practical purpose.) Li Ye found a tenth-degree equation for the square root of the diameter of the town, but (see Fig. 10.5) it is easy to see that if the diameter of the town rather than its square root is taken as the unknown, the result is the quartic equation $x^4 + 6x^3 + 9x^2 - 972x - 2916 = 0$. This fact was pointed out by the later mathematician Li Rui (1773–1817). (Actually, the similar right triangles in Fig. 10.5 lead to the cubic equation $2r^3 + 3r^2 = 243$for the radius.)

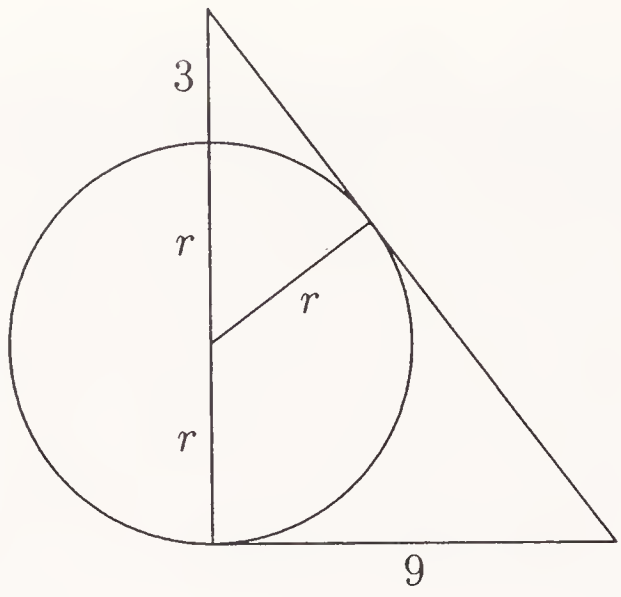


Figure 10.5: A quartic equation problem.

10.8.3 Magic Squares

Like the figurate numbers studied by the Pythagoreans, magic squares have never found any significant application. They have fascinated people for a long time, however, and their mathematical theory has been developed. The Chinese were apparently the first to develop this topic, which the Pythagoreans do not seem to have noticed. We have already mentioned the 3×3 magic square contained in the *Luo Shu*. Larger magic squares, up to 10×10, were given by Yang Hui in the thirteenth century, although the diagonal in his 10×10 magic square has the wrong sum:

1	20	21	40	41	60	61	80	81	100
99	82	79	62	59	42	39	22	19	2
3	18	23	38	43	58	63	78	83	98
97	84	77	64	57	44	37	24	17	4
5	16	25	36	45	56	65	76	85	96
95	86	75	66	55	46	35	26	15	6
14	7	34	27	54	47	74	67	94	87
88	93	68	73	48	53	28	33	8	13
12	9	32	29	52	49	72	69	02	89
91	90	71	70	51	50	31	30	11	10

Methods of constructing such squares are now known (by a modification of the "knight's move" when the number of rows and columns is odd, for example).

10.8.4 Mechanical Computation

As can be gathered from the discussions above, the Chinese used mechanical aids to computation from a very early date. The first device used was counting rods, and these were later combined with counting boards. The rods no doubt facilitated the development of the place-value decimal system, and the boards helped to develop algebra, since different locations on the board can be assigned to different powers of

the unknown, with the number of rods in the location representing the coefficient of the corresponding power. There is insufficient space here to describe the gradual improvement in the "software" accompanying this "hardware." For details the reader is referred to the book by Yan and Shiran *Chinese Mathematics, A Concise History*, Clarendon Press, Oxford, 1987, pp. 177–184. The rods were eventually replaced with beads, which were then strung together to produce the famous device we now know as the abacus. This last invention came during the Yuan Dynasty (fourteenth century). In countries where modern calculators are not available the abacus is still used in commerce and taught in schools; and the speed that a competent operator can achieve is amazing to one who has grown up in a society where only electronic cash registers are seen.

10.9 The Modern Era

When Western mathematics and science entered China with the Jesuits in the seventeenth century, the result was a conflict between traditionalists and modernists among Chinese scholars. In the end, like many other nations, China joined the common world mathematical culture, and made outstanding contributions to that culture. The routes by which European mathematics came to China were various, but when the Chinese began actively seeking more information about Western mathematics during the nineteenth century, one of their chief sources was translations from the *Encyclopedia Britannica*. A large number of Chinese scholars, both in China and abroad, have made outstanding discoveries in the many fields of modern mathematics, and their names have become associated with some of the most profound results yet attained in analysis, algebra, differential geometry, and other areas.

10.10 Problems and Questions

10.10.1 Chinese Mathematical Problems

Exercise 10.1 Compare the following loosely interpreted problems from the *Nine Chapters* and the Ahmose Papyrus. First, from the *Nine Chapters*: five officials went hunting and killed five deer. Their ranks entitle them to shares in the proportion $1 : 2 : 3 : 4 : 5$. What part of a deer does each receive?

Second, from the Ahmose Papyrus (Problem 40): 100 loaves of bread are to be divided among 5 people (in arithmetic progression), in such a way that the amount received by the last two (together) is one-seventh of the amount received by the first three (together). How much bread does each person receive?

Exercise 10.2 Carry out the solution of the bundles of wheat problem in the text. Is it possible to solve this problem without the use of negative numbers?

Exercise 10.3 Find $\sqrt{451,584}$ by the method of the text.

Exercise 10.4 Without using the formula given in the text, but repeating the reasoning that accompanies it, solve the equation $x^2 + 8x = 65$.

Exercise 10.5 The *gougu* section of the *Nine Chapters* contains the following problem. "Under a tree 20 feet high and 3 in circumference there grows a vine, which winds seven times the stem of the tree and just reaches its top. How long is the vine?" Solve this problem. [*Hint:* Despite appearances, the number π isn't involved here. Imagine that the tree is a perfect cylinder (the usual unrealistic assumption needed to get a solvable mathematical problem); then imagine that you have cut down the tree and rolled it on the ground to unwind the vine in a straight line.]

Exercise 10.6 Another right-triangle problem from the *Nine Chapters* is the following. "There is a string hanging down from the top of a pole, and the last 3 feet of string are lying flat on the ground. When the string is stretched, it reaches a point 8 feet from the pole. How long is the string?" Solve this problem. (You can also, of course, figure out how high the pole is from this information.)

Exercise 10.7 The most famous of all the problems from the *Nine Chapters* is the "broken bamboo" problem: A bamboo 10 feet high is broken and the top touches the ground at a point 3 feet from the stem. What is the height of the break? Solve this problem, which reappeared several centuries later in the writings of the Hindu mathematician Brahmagupta.

Exercise 10.8 Find the smallest positive number that leaves remainders of 3, 4, and 6 when divided by 8, 11, and 15 respectively.

Exercise 10.9 Explain how Fig. 10.2 can be used to prove a particular case of the Pythagorean theorem.

Exercise 10.10 The thirteenth-century work called the *Sea Mirror of Circle Measurements* by the mathematician Li Ye contains the following problem. [*Assume there is a circular fort of unknown diameter and circumference.*] *One person walks out of the south gate 135 steps and another person walks out of the east gate 16 steps, and then they see each other.* [*What is the diameter?*]

Draw a figure and set up the equation to solve this problem. Solve the equation by guessing a solution. Then explain the steps in the solution given by the author, as follows. [The entries in each column represent coefficients of powers of the unknown, but some of them are negative powers, that is, powers of $\frac{1}{x}$.]

Let the unknown be the radius of the fort; lay it down and add to it the southward steps, getting the *gu*.[2]

1
135

Then add to it the eastward steps, getting the *gou*.

[2]This expression represents $r + 135$, and similarly for subsequent boxes.

1
16

Multiply the *gou* and the *gu* together, getting

1
151
2160

Divide by the unknown, getting the hypotenuse[3]

1
151
2160

Multiply this by itself, getting the square of the hypotenuse, and place this on the left:

1
302
27,121
652,320
4,665,600

Multiply the *gou* by itself, getting

1
32
256

and multiply the *gu* by itself, getting

1
270
18,225

The two configurations added give

2
302
18,481

which is the same value [previously obtained for the square of the hypotenuse]. Cancel it [with the previous value][4]

−1
0
8640
652,320
4,665,600

which is a fourth-degree equation giving 120 steps as the radius of the fort.

What similarities with your own process of solution do you notice? What differences?

---

[3]Because the radius of the circle (the unknown) is the altitude of the right triangle, the product of the legs equals the product of the unknown and the hypotenuse (both are equal to twice the area of the triangle). This expression thus represents $r + 151 + 2160/r$.

[4]This final expression can be thought of as set equal to zero. It stands for $-r^2 + 8640 + 652,320r^{-1} + 4,665,600r^{-2}$.

Exercise 10.11 Solve the equation for the diameter of a town considered by Li Rui. [*Hint:* Since $x = -3$ is an obvious solution, this equation can actually be written as $x^3 + 3x^2 = 972$.]

10.10.2 Questions about Chinese Mathematics

Exercise 10.12 How do the rules for manipulating negative numbers in the *Nine Chapters* compare with Diophantus' rules for adding and multiplying expressions in an unknown quantity?

Exercise 10.13 Obviously one can find the successive digits of the square root of a number by trial and error. For example, in the problem given in the text of finding $\sqrt{55,225}$, we find that $200^2 = 40,000 < 55,225 < 90,000 = 300^2$, so that the first digit is 2. Then by trial and error, perhaps starting with 250 as a "guess," we soon discover that $230^2 = 52,900 < 55,225 < 57,600 = 240^2$, so that the second digit is 3, etc. In view of this fact, what is the advantage of learning an algorithm, such as the one described in the text?

Exercise 10.14 In several contexts now we have seen that a problem (for example, an equation) can be solved either approximately by numerical methods or in "closed form." For example, the equation $x^2 - 2 = 0$ has the closed-form solutions $x = \pm\sqrt{2}$ and the approximate solutions $x = \pm 1.41421$. What are the advantages and disadvantages of concentrating on one of these approaches to the exclusion of the other?

10.11 Endnotes

1. The study of Chinese mathematics in America and Europe has blossomed in the past few decades, and a number of good expositions can be found. A concise introduction can be found in the brief article by F. J. Swetz, "The Evolution of Mathematics in Ancient China," *Mathematics Magazine*, **52** (1), (Jan. 1979), pp. 10–19. A much more extensive, but still very readable full-length account is given in the book by Li Yan and Du Shiran, *Chinese Mathematics, A Concise History* (Clarendon Press, Oxford, 1987). A very complete treatment is given in Part 19 of *Science and Civilisation in China*, by Joseph Needham (Cambridge University Press, 1959). Most of the present chapter is based on these last two sources, together with the book by Yoshio Mikami, *The Development of Mathematics in China and Japan* (Chelsea Reprint, New York, 1913) and the book by Ulrich Libbrecht, *Chinese Mathematics in the Thirteenth Century* (MIT Press, Cambridge, MA, 1973).

2. The quotation about the *Luo Shu* is taken from the article by Swetz (op. cit.).

3. The linear algebra problem from the *Nine Chapters* is taken from the book of Mikami (op. cit.), p. 20.

4. The rules for handling negative numbers in the *Nine Chapters* are quoted from the book of Mikami (op. cit.), p. 21.

5. The derivation of the accurate value of π by Liu Hui is given in an article by Lam Lay-Yong and Ang Tian-Se, "Circle measurements in Ancient China," *Historia Mathematica*, **13** (4), (1986), pp. 325–340.

6. A complete English translation of the *Sea Island Manual* by Ang Tian Se and Frank J. Swetz was published in *Historia Mathematica*, **13** (2), (1986), pp. 99–117.

7. The discussion of Chinese astronomy is based on Joseph Needham's *Science and Civilization in China*, Cambridge University Press, 1959.

8. The Chinese use of Cavalieri's principle is explained in an article by Lam Lay-Yong and Shen Kangsheng, "The Chinese concept of Cavalieri's Principle and its applications," *Historia Mathematica*, **12** (3), (1985), pp. 219–228. A recent article by Daiwie Fu, "Why did Liu Hui fail to derive the volume of a sphere?" in *Historia Mathematica*, **18** (3), (1991), pp. 212–238, analyzes these precursors of the infinitesimal methods and shows their limitations when applied to the sphere.

9. The cubic equation problem from the *Sea Mirror of Circle Measurements* is taken from the book by Ulrich Libbrecht, *Chinese Mathematics in the Thirteenth Century* (MIT Press, Cambridge, MA, 1973), pp. 134–135.

Chapter 11

Korea and Japan

Both Korea and Japan adopted the Chinese system of writing their languages. Thus for the Orient the Chinese language played the same role that was played by Greek in the Hellenistic world, by Sanskrit in India, by Arabic in the Muslim world, and by Latin in Medieval Europe. That is, it provided a common language for scholars of many nations and a body of "classical" literature familiar to all educated people.

The influence of Chinese mathematics on both Korea and Japan was considerable. Indeed the courses of university instruction in this subject in both countries were based on reading (in the original Chinese language) the Chinese classics we have discussed in the preceding chapter. The Koreans played a role as transmitters, passing Chinese learning and inventions to Japan. (Two Korean scholars named Wang Lian-tung and Wang Pu-son journeyed to Japan in 553–554.) For many centuries both the Koreans and the Japanese worked within the system of Chinese mathematics. The earliest records of new and original work in these countries date from the 17th century. By that time mathematical activity was exploding in Europe, and Europeans had begun their long voyages of exploration and conquest. There is therefore only a comparatively brief window of time during which indigenous mathematics could grow up in these countries independently of Western influence.

11.1 Korean Mathematics

During the Koryo Dynasty of the tenth century a national university was established at Kukchagam with two professors of mathematics. The textbooks used were the *Nine Chapters* and later the *Introduction to Mathematical Studies* by Zhu Shijie. The Koreans were particularly interested in the study of equations with integer coefficients, which they called *ch'onwonsul*.[1] A 2-day examination was given in mathematics, during which the student was expected to recite whole chapters of the textbooks from memory and answer correctly at least four out of six questions posed.

[1] Recall that the Chinese name for the procedure of solving an equation was *tian yuan shu*, meaning *method of the celestial element*.

The role of Korean mathematicians was not limited to mere transmission, however. In the fifteenth century Sejong, the fourth king of the Yi Dynasty, personally checked the surveying results of his mathematicians and found them wanting. To reform the educational system he allowed the children of the nobility to study mathematics and sent Korean scholars to China to learn more mathematics. At his instigation new Korean editions of *Yang Hui's Methods of Calculation* (referred to in Korea as *Yanghwi Sanpob*) and the *Introduction to Mathematical Studies* (*Sanhak Kyemong*) were printed. Sejong is said to have observed the relation between pitch and length of a flute and to have established the Korean musical scale; comparisons with Pythagoras naturally spring to mind. The *Introduction* was reprinted in the seventeenth century, with the addition of a chapter on trigonometry.

New mathematics was created in Korea in the seventeenth century by Ch'oe Sok-jong (1646–1715), who was also active in political life (he served as prime minister six times). His mathematical interests were influenced by philosophy, and he was fascinated by magic squares. Regarding the 10×10 magic square reproduced in the preceding chapter he observed that in the first and last rows the ones digit of every entry is either 1 or 0, in the second and ninth rows it is either 2 or 9, etc. Perhaps because of the legend that the *Luo Shu* was given to the Emperor Yu by a tortoise, Ch'oe Sok-jong was interested in the tortoise shape (hexagon), and constructed some magic figures based on this shape. As mentioned in the preceding chapter, however, this topic has never blossomed into a major branch of mathematics. Nowadays Korean mathematicians work in the same areas as mathematicians of other nations.

11.2 Japanese Mathematics

11.2.1 Chinese Influence: Calculating Devices

The only surviving Japanese records date from the time after Japan had adopted the Chinese writing system. Like the Koreans, the Japanese were for a time content to read the Chinese classics. In 701 the emperor Monbu established a university system, in which the mathematical part of the curriculum consisted of 10 Chinese treatises. Some of these are no longer known, but the *Arithmetical Classic*, *Master Sun's Mathematical Manual*, the *Nine Chapters*, and the *Sea Island Manual* were among them. The evidence of Chinese influence is unmistakable in the mechanical methods of calculation used for centuries—counting rods, counting boards, and the abacus, which played an especially important role in Japan.

The Koreans adopted the Chinese counting rods and counting boards, which the Japanese subsequently adopted from them. The abacus was a technology beyond what was needed for Korean commerce. It too was exported to Japan, however, where it was eagerly adopted and improved. The abacus (*suan pan*) was invented in China, probably in the fourteenth century, when methods of computing with counting rods had become so efficient that the rods themselves were a hindrance to the performance of the computation (see Fig. 10.4). From China the invention passed to Korea, where it was known as the *sanbob*. Because it did not prove

useful in Korean business, it did not become widespread there. It did, however, pass on to Japan, where it was known as the *soroban*, which may be related to the Japanese word for an orderly table (*soroiban*). The Japanese made two important technical improvements in the abacus: (1) they replaced the round beads by beads with sharp edges, which are easier to manipulate; and (2) they eliminated the superfluous second 5-bead on each string.

11.2.2 Japanese Mathematical Innovations

It was reported by one nineteenth-century Japanese historian that the late-sixteenth-century emperor Hideyoshi sent the scholar Mori Shigeyoshi to China to learn mathematics. According to the story, the Chinese ignored the emissary because he was not of noble birth. When he returned to Japan and reported this fact, the emperor conferred noble status on him and sent him back. Unfortunately, his second visit to China coincided with Hideyoshi's unsuccessful attempt to invade Korea, which made his emissary unwelcome in China. Mori Shigeyoshi did not return to Japan until after the death of Hideyoshi, but when he did return (in the early seventeenth century), he brought the abacus with him. Whether this story is true or not, it is a fact that Mori Shigeyoshi was one of the most influential early Japanese mathematicians. He wrote several treatises, all of which have been lost, but his work led to a great flowering of mathematical activity in seventeenth-century Japan, through the work of his students.

Yoshida Koyu

Mori Shigeyoshi trained three outstanding students during his lifetime, of whom we shall discuss only the first. This student was Yoshida Koyu (1598–1672). Being handicapped in his studies at first by his ignorance of Chinese, Yoshida Koyu devoted extra effort to this language in order to read the *Systematic Treatise on Arithmetic* by Cheng Dawei, published in 1592. This book is well described by its title. It contains a systematic treatment of the kinds of problems handled in traditional Chinese mathematics, and at the end has a bibliography of some 50 other works on mathematics. Having read this book, it is said, Yoshida Koyu made rapid progress in mathematics and soon excelled even Mori Shigeyoshi himself. Eventually he was called to the court of a nobleman as a tutor in mathematics.

In 1627 Yoshida Koyu wrote his own textbook (in Japanese) based on the *Systematic Treatise*, calling his work the *Treatise on Large and Small Numbers*. Although this book was mostly derivative, it did contain a statement of what is known in modern mathematics as the Josephus problem. The Japanese version of the problem involves a family of 30 children choosing one of the children to inherit the parents' property. The children are arranged in a circle and count off by tens; the unlucky children who get the number 10 are eliminated, that is, numbers 10, 20, and 30 drop out. The remaining 27 children then count off again. The children originally numbered 11 and 22 will be eliminated in this round, and when the second round of numbering is complete, the child who was first will

have the number 8. Hence the children originally numbered 3, 15, and 27 will be eliminated on the next round, and the first child will start the following round as number 3. The problem is to see which child will be the last one remaining. Obviously solving this problem in advance could be very profitable, as the original Josephus story indicates.[2] The Japanese problem is made more interesting and more complicated by considering that half of the children belong to the couple and half are the husband's children by a former marriage. The wife naturally wishes one of her own children to inherit, and she persuades the husband to count in different ways on different rounds. The problem was reprinted by several later mathematicians.

The *Treatise on Large and Small Numbers* concluded with a list of challenge questions and thereby stimulated a great deal of further work. Here are some of the questions:

1. There is a log of precious wood 18 feet long whose bases are 5 feet and $2\frac{1}{2}$ feet in circumference. Into what lengths should it be cut to trisect the volume?

2. There have been excavated 560 measures of earth, which are to be used for the base of a building. The base is to be 3 measures square and 9 measures high. Required, the size of the upper base.

3. There is a mound of earth in the shape of a frustum of a circular cone. The circumferences of the bases are 40 measures and 120 measures and the mound is 6 measures high. If 1200 measures of earth are taken evenly off the top, what will be the height?

4. A circular piece of land 100 [linear] measures in diameter is to be divided among three persons so that they shall receive 2900, 2500, and 2500 [square] measures respectively. Required, the lengths of the chords and the altitudes of the segments.

These problems were solved in a later treatise, which in turn posed new mathematical problems to be solved; this was the beginning of a tradition of posing and solving problems that lasted for 150 years.

---

[2]Josephus tells us that, faced with capture by the Romans after the fall of Jotapata, he and his Jewish comrades decided to commit mass suicide rather than surrender. Later commentators claimed that they stood in a circle and counted by threes, agreeing that every third soldier would be killed by the person on his left. The last one standing was duty bound to fall on his sword. According to this folk legend, Josephus immediately computed where he should position himself in order to be that last person, but decided to surrender instead of carrying out the bargain. Josephus himself, however, writes in *The Jewish Wars*, Book III, Chapter 8 that the order of execution was determined by drawing lots and that he and his best friend survived either by chance or by divine intervention in these lots. The mathematical problem we are discussing is also said to have been invented by Abraham ben Meir ibn Ezra (1092–1167), better known as Rabbi Ben Ezra, one of many Jewish scholars who flourished in the Caliphate of Cordoba.

11.2.3 Isomura Kittoku

A great many treatises on mathematics were written in seventeenth-century Japan, resulting in the creation of an indigenous mathematics called *wasan*. The word comes from *Wa*, meaning Japan, and *san* meaning mathematics. (The word *san* is apparently related to the Chinese word *suan*, since the modern Japanese word for mathematics is *sugaku*.) This *wasan* bypassed many topics found in Euclid in favor of area and volume problems that were too difficult for Euclid.

In particular the attempt to approximate the volume of a sphere by cylindrical shells can be seen in Fig. 11.1. This figure is taken from a 1684 edition of a work known as *Ketsugi-sho* (literally "combination book"), first published in 1660 by Isomura Kittoku, a student of a student of Mori Shigeyoshi. The method is explained by the author as follows:

> If we cut a sphere of diameter 1 foot into $10,000$ slices, the thickness of each slice is 0.001 feet, which will be something like that of a very thin paper. Finding in this way the volume of each of them, we sum up the results, 10,000 in number, when we get 532.6 measures [that is, a volume of 0.5326 cubic feet]. Besides, it is true, there are small incommensurable parts, which are neglected.

If we make allowance for what may be inaccuracies in the translation (the word *incommensurable*, for example, seems to be inappropriate), the method is perfectly sound as an approximation, and the figure is accurate to all the decimal places given. This technique raises an important question about the level of sophistication of Japanese mathematics at this period. It is a good thing, of course, to realize that the ratio of the circumference of a circle to its diameter is the same for all circles. It is a further advance to speculate on the value of this number and its relation to other numbers, both for theoretical and computational purposes. The value $\pi = \sqrt{10}$, for example, was used in China, India, and Japan at various times. In the problem we are now discussing, Isomura Kittoku has exhibited another constant, the ratio of the volume of a sphere to the cube on its diameter. He knows that this ratio is the same for all spheres, and is approximately 0.5326. The question that naturally comes to mind is: *Did Isomura Kittoku know that this second constant is* $\pi/6$*?* Perhaps not when the *Ketsugi-sho* was first published, since at that time he believed $\pi = \sqrt{10} = 3.162$. But only 3 years later (1663) another mathematician, Muramatsu Mosei, published a work called the *Sanso* based on the Chinese *Introduction to Mathematical Studies* (*Suanxue*, apparently the source of the word *Sanso*), in which he used the approximation technique of repeatedly doubling the number of sides of a polygon inscribed in a circle of unit diameter to estimate the circumference. Starting with a square and finishing with a polygon of 31,768 sides, he found the perimeter of this polygon to be 3.141592648777698869248. If this number is taken as the circumference of the circle, it is correct to eight places ($\pi = 3.14159265\ldots$). Once this value is known, it would be a natural conjecture that the constant for the sphere (0.5236) is $\frac{\pi}{6}$. In other words, the two constants are very simply related to each other. This technique of obtaining extraordinary precision and using it to perform numerical

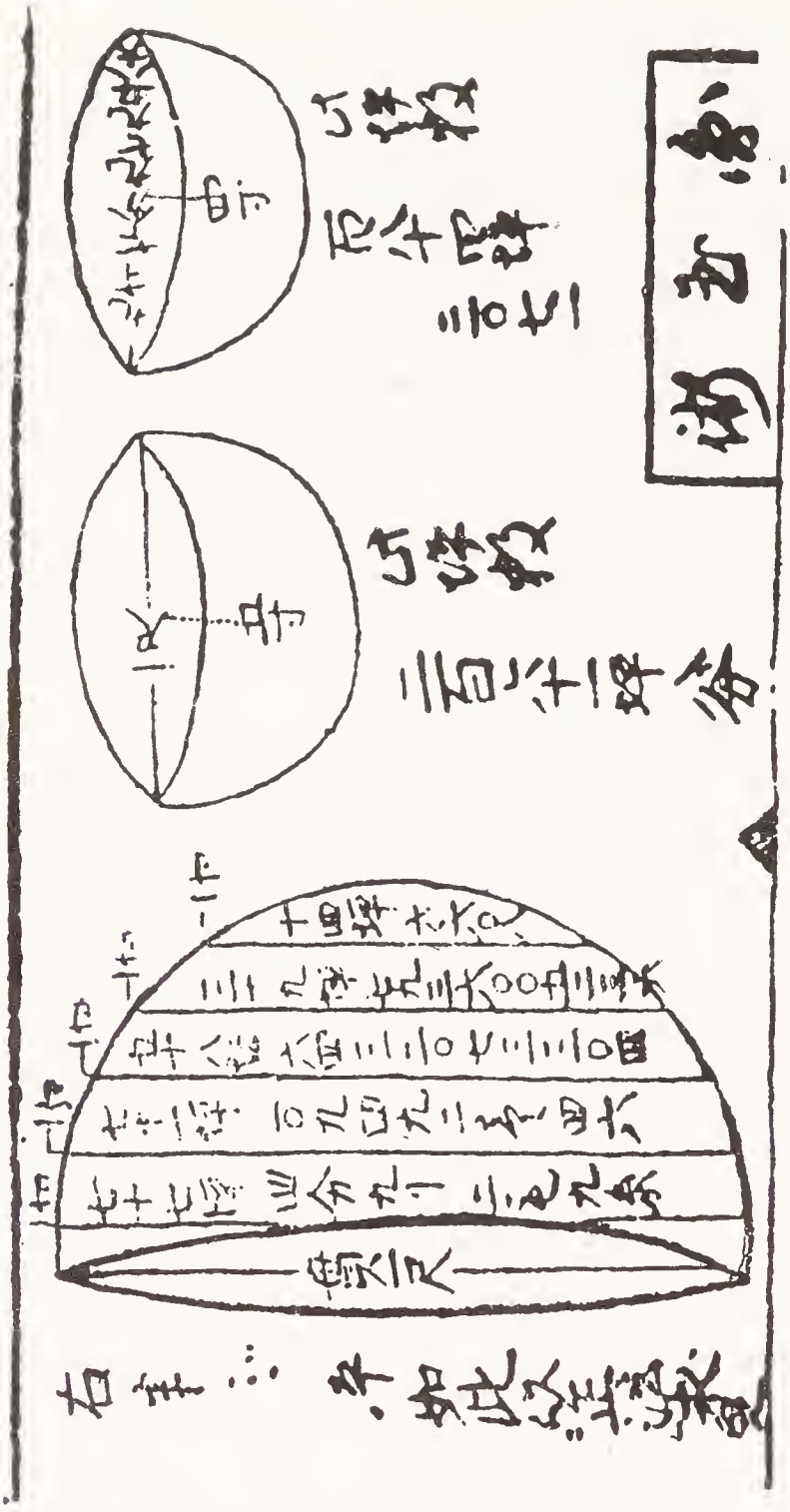


Figure 11.1: Isomura Kittoku's computation of the volume of a sphere (oriented sideways). Stock Montage, Inc.

experiments which provide the basis for general assertions appears elsewhere in Japanese mathematics, in particular in the summation of infinite series.

That Isomura Kittoku did indeed know the relation between these two constants can be inferred from his work on the surface area of the sphere. He had stated incorrectly in the original (1660) edition of the *Ketsugi-sho* that the surface area of a sphere is one-fourth the square of the circumference. He had cited several previous authorities for this statement. In our terms, he would be saying that $A = \pi^2 r^2$, which is about $9.9r^2$, although of course he would have believed it to be $10r^2$, since he believed $\pi = \sqrt{10}$. By the time of the second edition in 1684 he had realized that this value of π is wrong, probably because of the work of Muramatsu Mosei. This time he used a spherical shell method, taking two concentric spheres with diameters 10 and 10.002. By his formula, their volumes would be 523.6 and 523.9142 .... The difference between the volumes should be

approximately 0.001 times the area of the sphere of diameter 10, since the distance between the concentric spheres is 0.001. Thus a sphere of diameter 10 should have a surface area approximately $314.2\ldots \approx 100\pi$. From this numerical experiment, Isomura Kittoku concluded correctly that a sphere of diameter 10 should have an area of 100π, and in general the surface area of a sphere is π times the square of its diameter.

11.2.4 Japanese Algebra

Another impetus to the development of mathematics in Japan came with the arrival of the Chinese "method of the celestial element" (*tian yuan shu*), which spread to Korea as *ch'onwonsul* and thence to Japan as *tengen jutsu*. This form of algebra, adapted for work on a counting board, was expounded in the *Introduction to Mathematical Studies* by Zhu Shijie and the *Sea Mirror of Circle Measurements* by Li Zhi, both of which were standard textbooks in Korea during the fourteenth century. The first of these became part of the standard Japanese curriculum before the seventeenth century.

When Japanese mathematicians began to develop this subject in the seventeenth century, they made some advances on what they had learned from the Chinese. In 1666 Sato Seiko wrote a treatise called the *Kongenki* in which he recognized the possibility of more than one solution to an equation. His contemporary Sawaguchi Kazuyuki asserted that when the equation of a problem has more than one solution, there is something wrong with the data that lead to the problem. For example, Sato Seiko had posed the following problem: *There is a circle from within which a square is cut, the remaining portion having an area of 47.6255. If the diameter of the circle is 7 more than the square root of a side of the square, it is required to find the diameter of the circle and the side of the square.* If the diameter of the circle is d and the side of the square is s, we would write this problem as the equations

$$\begin{aligned} \frac{\pi d^2}{4} - s^2 &= 47.6255, \\ d &= 7 + \sqrt{s}, \end{aligned}$$

which has the natural solutions $d = 9$, $s = 4$ (when π is taken as $3.14\overline{2}$), but also the "unnatural" solutions $d = 7.8242133\ldots$, $s = 0.67932764\ldots$. Sawaguchi Kazuyuki removed the difficulty by making the area 12.278 instead of 46.6255 and the difference between the diameter of the circle and the side of the square 4 instead of 7. Then the only positive solutions are $d = 6$ and $s = 4$. Of course, this way of dealing with the ambiguity of multiple roots is really only a way of avoiding the problem.

11.2.5 Seki Kowa

One figure in seventeenth-century Japanese mathematics stands out far above all others, a genius who is frequently compared with Archimedes, Newton, and Gauss.

His name was Seki Kowa, and he was born around the year 1642, the same year in which Isaac Newton was born in England. The stories told of him bear a great resemblance to similar stories told about other mathematical geniuses. For example, one of his biographers says that at the age of five Seki Kowa pointed out errors in a computation that was being discussed by his elders. A very similar story is told about Gauss. Being the child of a samurai father and adopted by a noble family, Seki Kowa had access to books. He was mostly self-educated in mathematics, having paid little attention to those who tried to instruct him; in this respect he resembles Newton. Like Newton, he served as an advisor on high finance to the government, becoming examiner of accounts to the lord of Koshu. Unlike Newton, however, he was a popular teacher and physically vigorous. He became a shogunate samurai and master of ceremonies in the household of the Shogun. He died at the age of 66, leaving no direct heirs. His tomb in the Buddhist cemetery in Tokyo was rebuilt 80 years after his death by mathematicians of his school. His pedagogical activity earned him the title of *Sansei*, or *Arithmetical Sage*, a title that was carved on his tomb. Although he published very little during his lifetime, his work became known through his teaching activity, and he is said to have left copious notebooks.

Seki Kowa made profound contributions to several areas of mathematics. He was primarily an algebraist who converted the celestial element method into two more sophisticated and subtle methods of handling equations, known as *the method of explanation* and the *method of clarifying things of obscure origin*. He kept this latter method a secret. According to some scholars, his pupil Takebe Kenko (1664–1739) refused to divulge the secret, saying, "I fear that one whose knowledge is so limited as mine would tend to misrepresent its significance." However, other scholars claim that Takebe Kenko did write an exposition of the latter method, and that it amounts to the principles of cancellation and transposition (see below).

Seki Kowa took up the challenge that Sawaguchi Kazuyuki had avoided and considered equations with more than one root, even negative roots. He classified equations as perfect (having precisely one real root), varied (more than one root, but all roots of the same sign), mixed (both positive and negative roots), and rootless. He was aware that only an equation of even degree can be rootless. He also wrote a treatise on the calendar (a commentary on an earlier treatise), in which he used black and red symbols to distinguish positive and negative numbers.

Algebra

There can be no doubt about Seki Kowa's prowess in solving algebraic problems. Consider, for example, his solution of the fourteenth problem of Sawaguchi Kazuyuki: *There is a quadrilateral whose sides and diagonals are* u, v, w, x, y, *and* z [as shown in Fig. 11.2].

It is given that

$$\begin{aligned} z^3 - u^3 &= 271 \\ u^3 - v^3 &= 217 \\ v^3 - y^3 &= 60.8 \end{aligned}$$

$$\begin{aligned} y^3 - w^3 &= 326.2 \\ w^3 - x^3 &= 61. \end{aligned}$$

Required, to find the values of u, v, w, x, y, z.

Seki Kowa does not tell the reader any details of the solution. He gives only a bare outline:

> Take the "celestial element" for z, from which the expressions of the cubes of u, v, w, x, and y may be derived.
> Then eliminate x^3, leading to an equation of the 18th degree.
> Next eliminate w^3, leading to an equation of the 54th degree.
> Next eliminate y^3, leading to an equation of the 162nd degree.
> Next eliminate v^3, leading to an equation of the 486th degree.
> Now by eliminating u^3 two equal expressions result from which the final equation of the 1458th degree is obtained.

The fact that the six quantities are the sides and diagonals of a quadrilateral provides one equation that they must satisfy, namely:

$$\begin{aligned} (uw)^2[(u^2+w^2)-(v^2+x^2)-(y^2+z^2)] + (vx)^2[-(u^2+w^2)+ \\ +(v^2+x^2)-(y^2+z^2)] + (yz)^2[-(u^2+w^2)-(v^2+x^2)+ \\ +(y^2+z^2)] + uvy + vwz + wxy + vxz = 0. \end{aligned}$$

This equation, together with the five given conditions, provides a complete set of equations for the six quantities, and this system of equations can be solved, as Seki Kowa showed. Such equations were solved numerically by the Chinese using Yang Hui's method, the calculations being performed on a counting board. Historians of Japanese mathematics report that for equations of such prodigiously high degree a counting board the size of an entire room was ruled into small squares. As remarked by the twentieth-century Japanese historian Yoshio Mikami, "Perseverance and hard study were a part of the spirit that characterized Japanese mathematics of the old times," Nevertheless one cannot help thinking that such problems must have been a powerful stimulus to the invention of a compact notation for equations, and Seki Kowa made a contribution in this direction with his methods.

Seki Kowa always appealed to reason and logic and explained his solutions. His influence might have led to the development of algebra as a deductive system comparable to Euclid's geometry. However, deductive systems do not seem to have had much appeal for the practitioners of *wasan*. The historian of mathematics T. Murata reports that, having seen Chinese translations of Euclid, they were repelled by the great amount of fuss required to derive elementary facts.

As mentioned above, one of Seki Kowa's contributions to algebra was called by him the *method of clarifying things of obscure origin* (the Japanese phrase is also translated as *method for revealing the true and buried origin of things*). This method is simply the principles of transposition used in algebra for solving equations. Seki Kowa himself kept the method secret during his lifetime, but it

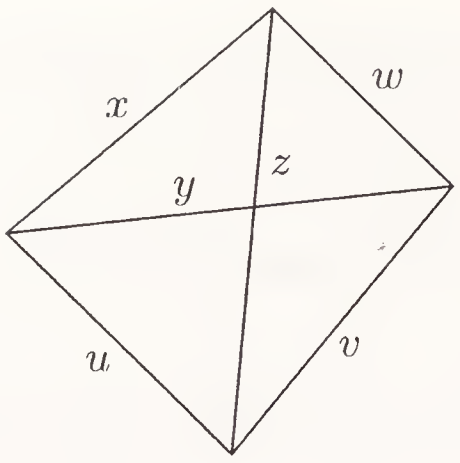


Figure 11.2: Sawaguchi Kazuyuki's quadrilateral problem.

was written about by his pupil Takebe Kenko, who called the method *tenzan*, a combination of two Chinese symbols *ten* (restore) and *zan* (strike off). These operations are familiar to us as transposition and cancellation. This method represents a shift in focus from the use of counting boards and counting rods for representing the equation to a purely graphical representation. Accompanying this shift was a great improvement in notation. Seki Kowa wrote fractions more or less as we do, except that he would write $a|b$ where we write b/a. For expressions such as a^4b^7 he would write $\begin{matrix} ab \\ 36 \end{matrix}$, meaning that a is to be multiplied by itself three times and b is to be multiplied by itself six times. (Of course, he didn't use the letters a and b.)

Determinants

Seki Kowa alone is given the credit for inventing one of the central ideas of modern mathematics, namely determinants. This concept is usually introduced in connection with linear equations, but Seki Kowa developed it in relation to equations of higher degree as well. The method is explained as follows. Suppose we are trying to solve two simultaneous quadratic equations

$$\begin{aligned} ax^2 + bx + c &= 0 \\ a'x^2 + b'x + c' &= 0. \end{aligned}$$

When we eliminate x^2, we find the linear equation

$$(a'b - ab')x + (a'c - ac') = 0.$$

Similarly, if we eliminate the constant term and divide by x, we find

$$(ac' - a'c)x + (bc' - b'c) = 0.$$

Thus from two quadratic equations we have derived two linear equations. Seki Kowa called this process *folding*. The same method makes it possible to replace n equations of degree n by n equations of degree $n - 1$.

We have written out explicit expressions for the simple 2×2 determinants here, for example,

$$\begin{vmatrix} a & c \\ a' & c' \end{vmatrix} = ac' - a'c,$$

but, as everyone knows, the full expanded expressions for determinants are very cumbersome even for the 3×3 case. It is therefore important to know ways of simplifying such determinants, using the structural properties we now call the *multilinear property* and the *alternating property*. Seki Kowa knew explicitly how to make use of the multilinear property to take out a common factor from a given row. He not only formulated the concept of a determinant but also knew many of their properties, including how to determine which terms are positive and which are negative in the expansion of a determinant.

11.2.6 Beginnings of the Calculus in Japan

The traditional Japanese mathematicians considered many problems of area and volume and developed techniques for solving them that look very much like the approximating sums for integrals. They were thus well on the way to discovering the integral calculus. The calculus was invented and highly developed in Europe by the end of the seventeenth century; however, none of this knowledge had reached Japan, and the work of the Japanese mathematicians is undoubtedly original and independent of European work.

The essence of the calculus is the use of infinitesimals or limits. Now passage to a limit can sometimes be performed by summing an infinite series. This step was taken by Seki Kowa's pupil Takebe Kenko, who wrote that Seki Kowa disliked complicated theories, but that he himself succeeded in finding the quadrature of the circle. Takebe Kenko's method was an ingenious discovery of the relation between the square of half of an arc $a^2/4$, the height h of the arc, and the diameter d of the circle. He began with height $h = 0.000001$ and $d = 10$, finding the square of the arc geometrically with accuracy to 53 decimal places. Approximating the half-arc by its chord and applying the Pythagorean theorem gives the approximation hd for $a^2/4$. Having 53 decimal places at his disposal to correct this value, he noticed that the correction that had to be added to obtain the more accurate result was approximately $\frac{1}{3}h^2$. Taking this as a first crude correction and successively refining the result, he observed that each successive corrective term was obtained by multiplying its predecessor by $(h/d)\frac{8}{15}$, $(h/d)\frac{9}{14}$, $(h/d)\frac{32}{45}$, $(h/d)\frac{25}{33}$,... . At this point he was able to guess the general law, which we would express by saying that corrective term number n is obtained by multiplying corrective term number $n-1$ by $(h/d)[2n^2/(n+1)(2n+1)]$. Putting these corrections together as an infinite series leads to the expression

$$\frac{a^2}{4} = dh\left[1 + \sum_{n=1}^{\infty} \frac{2^{2n+1}(n!)^2}{(2n+2)!} \cdot \left(\frac{h}{d}\right)^n\right],$$

In our terms $a = 2d \arcsin\left(\sqrt{h/d}\right)$. Notice what a lucky example this turned out to be: The quantity a cannot be expressed as a power series in h and d, but

a^2 can be so expressed. In using this numerical approach, Takebe Kenko was expressing a faith (which turned out to be justified) that the coefficients of the power series are rational numbers that satisfy a fairly simple recursive formula.

This series solves the problem of rectification of the circle, and hence all problems that depend on knowing the value of π. Takebe Kenko obtained this result about 15 years before the series for the arcsine was obtained by Leonhard Euler.

If the techniques just used were the basis of all reasoning on infinite series, it does not seem likely that any great generality could be obtained. There is evidence, however, for another form of passage to the limit that resembles modern methods. In an eighteenth-century manuscript of unknown authorship called the *Rolls of Heaven and Earth* (attributed to Seki Kowa by one Japanese historian of mathematics, though this claim is disputed by others), one finds a method of computing the volume of a spherical segment by slicing it into a *general* number of thin sections, then taking the limit as the number of sections tends to infinity. This technique involves knowing that the sum of the first n integers is $n(n+1)/2$ and that the sum of their squares is $n(n+1)(2n+1)/6$. Since this fact was known more than a thousand years earlier in India, it was very likely known to the Japanese algebraists. The limiting process is not general, but uses only the fact that a quotient of polynomials $p(x)/q(x)$, where q is of higher degree than p, must tend to zero as x tends to infinity. Thus we do find the essential concept of calculus (passage to a limit) in *wasan*.

11.2.7 Western Contacts

Because Japanese mathematics developed rather late, its greatest flowering being in the seventeenth century, the question of contact between the West and Japan and the sharing of ideas has been considered by historians. Determinants, for example, began to come into focus in Europe at exactly the same time that Seki Kowa was studying them. They were introduced in a letter written by Leibniz in 1693.

The policy of the Ming emperors in China was isolationist. In the seventeenth century the shoguns adopted an even stricter policy, one which could be more easily enforced in a small island kingdom such as Japan. The attempt to emigrate from Japan was considered treason, punishable by death. Christian missionaries were banned, and the practice of Christianity made illegal. European access to Japan was strictly controlled; only the Dutch were allowed to come for trade, and only at Nagasaki. As in China, these bans were not lifted until the nineteenth century, as the result of the threat of military action by Western powers. In such an atmosphere the exchange of ideas was very difficult, and it is not surprising that the Japanese and Europeans duplicated each other's work to some extent. No absolutely watertight ban on contact has ever been successful, however, and it is known that at least two Japanese students were studying in Leyden during the 1650s and 1660s. Their complete biography, however, is missing, and it is not known whether they returned to Japan.

There seems to have been a mutual complacency in the West and Japan that

hindered scholarly contact. On the Japanese side this complacency was accompanied by a fear of social disruption from an invasion of foreign ideas, and the isolation was enforced by rigid government decrees, while in Europe from the seventeenth century on the level of intellectual activity was so high that Europeans felt no need to look outside their own borders for ideas. A book calling itself a history of the subject of determinants (it is actually better described as a catalog of papers on the subject, with commentaries), was written by the South African mathematician Thomas Muir (1844–1934) in 1905. Although this book consists of four volumes totaling some 2000 pages, it does not mention Seki Kowa, the true discoverer of determinants!

Commercial contact was bound to result in some cultural penetration, however limited, and Western mathematical advances came to be known little by little in Japan. By the time Japan was opened to the West in 1854 Japanese mathematicians were already aware of many European topics of investigation. In joining the community of nations for trade and politics Japan also joined it intellectually. In the early nineteenth century Japanese mathematicians were writing about such questions as the rectification of the ellipse, a subject of interest in Europe at the same period. By the end of the nineteenth century there were several Japanese mathematical journals publishing work (in European languages) comparable to what was being done in Europe at the same period, and a few European scholars were already reading these journals to see what advances were being made by the Japanese. In the twentieth century, this trickle of Japanese work into Europe became a flood, and Japanese mathematicians have been represented among the leaders in nearly every field of mathematics.

11.3 Problems and Questions

11.3.1 Problems in Japanese Mathematics

Exercise 11.1 Solve the Japanese Josephus problem, assuming that the children wear labels $1,\ldots, 30$ to begin with. What label will be on the last child left? Is there a method or formula by which this answer could be arrived at without performing the experiment?

Exercise 11.2 One Japanese problem is the following: *There is a right triangle whose hypotenuse is* 6*, and the sum of whose area and the square root of one side is* 7.2384*. Required the other two sides.* (Problem No. 64 of the *Kongenki*). Letting x be the side whose square root is mentioned, derive the equation

$$(x^4 - 36x^2 + 4x + 209.57773824)^2 = 3353.24381184x,$$

then analyze the solution offered by Sawaguchi Kazuyuki. Let the unknown be the first side. Square it and subtract the result from the square of the hypotenuse. The difference is the square of the second side. Multiply it by the square of the first side, to obtain four times the square of the area. Denote this quantity by A. Let B equal four times the first side. Square the sum of the area and the square root of

the first side, multiply by four, and subtract A and B from the result. The square of the difference is four times AB. Denote this number by X, thus obtaining an equation of degree 8, which can be solved to yield the unknown.

The two sides are then given as 5.76 and 1.68. Verify that $x = 5.76$ is indeed a solution. [Note also that $(5.76)^2 + (1.68)^2 = 36$, as required.]

Exercise 11.3 Among the problems stated by Isomura Kittoku was the following (Problem 41): *There is a log 18 feet long, the diameter of the extremities being 1 foot and 2.6 feet respectively. This is wound spirally with a string 75 feet long, the coils being 2.5 feet apart. How many times does the string go around it?* How does this problem relate to the simpler Chinese problem of a vine winding around a tree (Exercise 10.5) from the previous chapter? Can it be solved by an analogous technique? What is the answer to this problem?

Exercise 11.4 Problem 85 of Isomura Kittoku is to find the length of one axis of an ellipsoid of revolution if the other axis is 1.8 feet long and the volume is 2.422. Does it make a difference whether the given axis is the axis of revolution? What is the solution?

Exercise 11.5 The first problem posed by Sawaguchi Kazuyuki is as follows. *In a large circle three smaller circles are inscribed, so that each is tangent to the other two and to the larger circle. Two of the inscribed circles are the same size and the third has a diameter* 5 *units larger than them. The area inside the largest circle and outside the three smaller ones is* 120 *square units. What are the diameters of the various circles?* Let the unknowns be the diameter d of the smallest two circles and the diameter D of the outside circle. Show that these quantities must satisfy the equations

$$(8d^2 + 20d + 50)D = 4d^3 + 40d^2 + \left(\frac{720}{\pi} + 125\right)d + \left(250 + \frac{4800}{\pi}\right);$$

$$D^2 = 3d^2 + 10d + 25 + \frac{480}{\pi}$$

(This problem was solved by Seki Kowa, who found an equation of degree 6 for d.)

11.3.2 Questions about Japanese Mathematics

Exercise 11.6 How is it possible that many Japanese authorities believed the area of the sphere to be one-fourth the square of the circumference, that is, $\pi^2 r^2$ rather than the true value $4\pi r^2$? Smith and Mikami assert in *A History of Japanese Mathematics* (Open Court Publishing Co., Chicago, 1914) that the error arose as follows: one can imagine a globe sliced along equally spaced lines of longitude and flattened out, so as to form very many thin "wedges" with curved sides tangent to one another and height πr, approximately. The equator (of length $2\pi r$) runs directly through the middle of all these wedges. If these wedges are cut in half along the equator, the two parts can be fitted together like sawteeth, to form

approximately a rectangle of sides $2\pi r$ and $\frac{1}{2}\pi r$. Why doesn't this argument give the correct result? Is there any way of seeing that it doesn't work without the use of rigorous mathematics?

Exercise 11.7 If a *circle* is sliced into sectors and the sectors are spread out by laying the circumference down along a straight line, the result is a large number of thin triangles of height approximately r and total base $2\pi r$. Therefore the area of the circle is πr^2. Why does this argument give the correct result when the analogous argument for the sphere (given in the previous problem) does not? Can we expect to find the lengths of arbitrary curves by inscribing broken lines in them? Can we expect to find the areas of arbitrary surfaces by inscribing triangles in them? Is this a part of mathematics where logical rigor is essential to avoid mistakes?

Exercise 11.8 Would a problem such as Exercise 11.2 have been considered a sensible problem to the Greeks? What meaning can be attached to the square root of a side of a square? Do these considerations suggest that the Japanese problems were purely arithmetical in nature, not related to the solution of real-world problems? How do you think the Japanese mathematicians would have reacted if they had read Pappus' statement that a product of more than three lines is impossible?

Exercise 11.9 Problem 84 of Isomura Kittoku is to find the length of the minor axis of an ellipse whose area is 758.940625 and whose major axis is 38. What must have been known about ellipses in order for such a problem to be formulated? How could it have been solved in the methodology of *wasan*, that is, using the approximative techniques discussed above?

Exercise 11.10 Many of the algebra problems considered by the Japanese mathematicians require that more than one unknown be found. Yet the celestial element method makes no provision for more than one unknown. Read again the solutions given by the Japanese mathematicians to see how they handled such problems, How do their methods compare with those of Diophantus?

Exercise 11.11 Why is Seki Kowa the central figure in Japanese mathematics? Are comparisons between him and his contemporary Isaac Newton justified?

Exercise 11.12 What is the justification for the statement by the historian of mathematics T. Murata that Japanese mathematics was not a science but an art?

Exercise 11.13 Why might Seki Kowa and other Japanese mathematicians have wanted to keep their methods secret, and why did their students, such as Takebe Kenko, honor this secrecy?

Exercise 11.14 For what purpose was algebra developed in China and Japan? Was it needed for science and/or government, or was it an "impractical" liberal-arts subject, on a par with the Confucian classics?

11.4 Endnotes

1. The information on Korean mathematics is based on a series of articles by Kim Yong-Woon that appeared in *Korea Journal*, **18** (7–9), (1973), pp. 16–39, and on the article "Pan-paradigm and Korean Mathematics in the Choson Dynasty," which appeared in *Korea Journal* in March 1986, pp. 25–46.

2. The section on Japanese mathematics is based on the book by Yoshio Mikami, *Mathematics in China and Japan*, the original 1913 edition of which was reprinted by Chelsea Publishing Co. (New York, 1961), and the book by David Eugene Smith and Yoshio Mikami, *A History of Japanese Mathematics* (Open Court Publishing Co., Chicago, 1914). Japanese names are given surname first, following these books. A recent article entitled "Indigenous Japanese Mathematics, Wasan," by Tamotsu Murata, gives the surname last, and also gives variant versions of some of the names and dates. Murata's article can be found in the *Companion Encyclopedia of the History and Philosophy of Mathematical Science*, Vol. I (Routledge, London and New York, 1994), pp. 104–110.

3. The note on the Josephus problem is based on the book by W. Ahrens, *Mathematische Unterhaltungen und Spiele* (Teubner, Leipzig, 1901), pp. 286–287, and on *Josephus: The Jewish War*, Gaalya Cornfield, general editor, Benjamin Mazar and Paul L. Maier, consulting editors (Zondervan Publishing House, Grand Rapids, MI, 1982), pp. 238–241.

4. Seki Kowa's solution of the equation of degree 1458 is quoted from the book by Smith and Mikami (op. cit.), pp. 100–101.

Chapter 12

Islamic Mathematics

12.1 The Expansion of Islam

During the period from 700 to 1300 C. E. the most important advances in science and mathematics in the West came in the lands under Muslim rule. Starting as a small and persecuted sect in the early seventh century, by mid-century the Muslims had expanded by conquest as far as Persia. They then turned West and conquered Egypt, all of the Mediterranean coast of Africa, and the island of Sicily. Moorish influence on Spanish architecture is evident in the Alhambra (Fig. 12.1).

12.1.1 The Umayyads

A palace revolution among the Islamic leaders led to the triumph of the first dynasty, the Umayyad (sometimes spelled Ommiad) in the year 660. Under the Umayyads Muslim expansion continued around the Mediterranean coast and eastward as far as India. This expansion was checked by the Byzantine Empire at the Battle of Constantinople in 717. In the West a Muslim general named Tarik led an army into Spain, giving his name to the mountain at the southern tip of Spain—Jabal Tarik, known in English as Gibraltar. The Muslim expansion in the West was halted by the Franks under Charles Martel at the Battle of Tours in 732. In 750 another revolution resulted in the overthrow of the Umayyad Dynasty and its replacement in the East by the Abbasid Dynasty. The Umayyads remained in power in Spain, however, a region known during this time as the Caliphate of Cordoba.

12.1.2 The Abbasids

Al-Mansur, the second of the Abbasid caliphs, built the capital of the new dynasty, the city of Baghdad, on the Tigris River. Both the Abbasids and the Umayyads cultivated science and the arts, and mathematics made advances in both the Eastern and Western parts of the Islamic world. The story of Islamic mathematics begins

in the city of Baghdad in the reign of two caliphs. The first of these was Harun Al-Raschid (786–809), a contemporary of Charlemagne. The second is the son of Harun Al-Raschid, Al-Mamun (813–833), whose court life provided the setting of the *Thousand and One Nights*.

12.1.3 The Turkish and Mongol Conquests

Near the end of the tenth century a group of Turkish nomads called Seljuks migrated from Asia into the Abbasid territory and converted to Islam. Gradually the Seljuks began to seize territory from the Abbasids, and in 1055 they took over Baghdad. It was their advance into Palestine that provoked the first Crusade in 1096. The Seljuks left the Abbasids as the nominal rulers of the empire, but in the thirteenth century both Abbasids and Seljuks were conquered by the same Mongols who had overrun Russia and China. The Mongol conquest of Iraq was particularly devastating, since it resulted in the destruction of the irrigation system that had supported the economy of the area for thousands of years. As in China, the Mongol rule was short-lived and was succeeded by another conquest, this time by the Ottoman Turks, who conquered Constantinople in 1453 and remained a threat to Europe until the nineteenth century.

12.1.4 Islamic Mathematics

The Islamic empire was unchallenged for 300 years in the East and six hundred in Spain. During this period Islamic mathematicians assimilated the science and mathematics of their predecessors and made their own unique additions and modifications to what they inherited. For many centuries they were the people who had the most extensive texts of the works of Archimedes, Apollonius, and Euclid and strove to advance beyond the point reached by these illustrious Greek mathematicians. The Greek mathematicians, however, were not the only influence on them. From earliest times the Caliph was in diplomatic contact with India, and one of Harun Al-Raschid's contributions was to obtain translations from Sanskrit into Arabic of the works of Aryabhata, Brahmagupta, and others. Some of the translators took the occasion to write their own mathematical works, and so began the Islamic contribution to mathematics. Besides the Arabic translations of many Greek works of which the originals have been lost, the modern world has inherited a considerable amount of scientific and mathematical literature in Arabic. This language has given us many words relating to science, such as *alcohol, alchemy, zenith*, and the mysterious names of the stars such as *Altair, Aldebaran, Algol*, and *Betelgeuse*.

12.2 Al-Khwarizmi

In the ninth century the caliph Al-Mamun established at Baghdad a "House of Wisdom," a research institute to which scholars were invited. There were Hindu

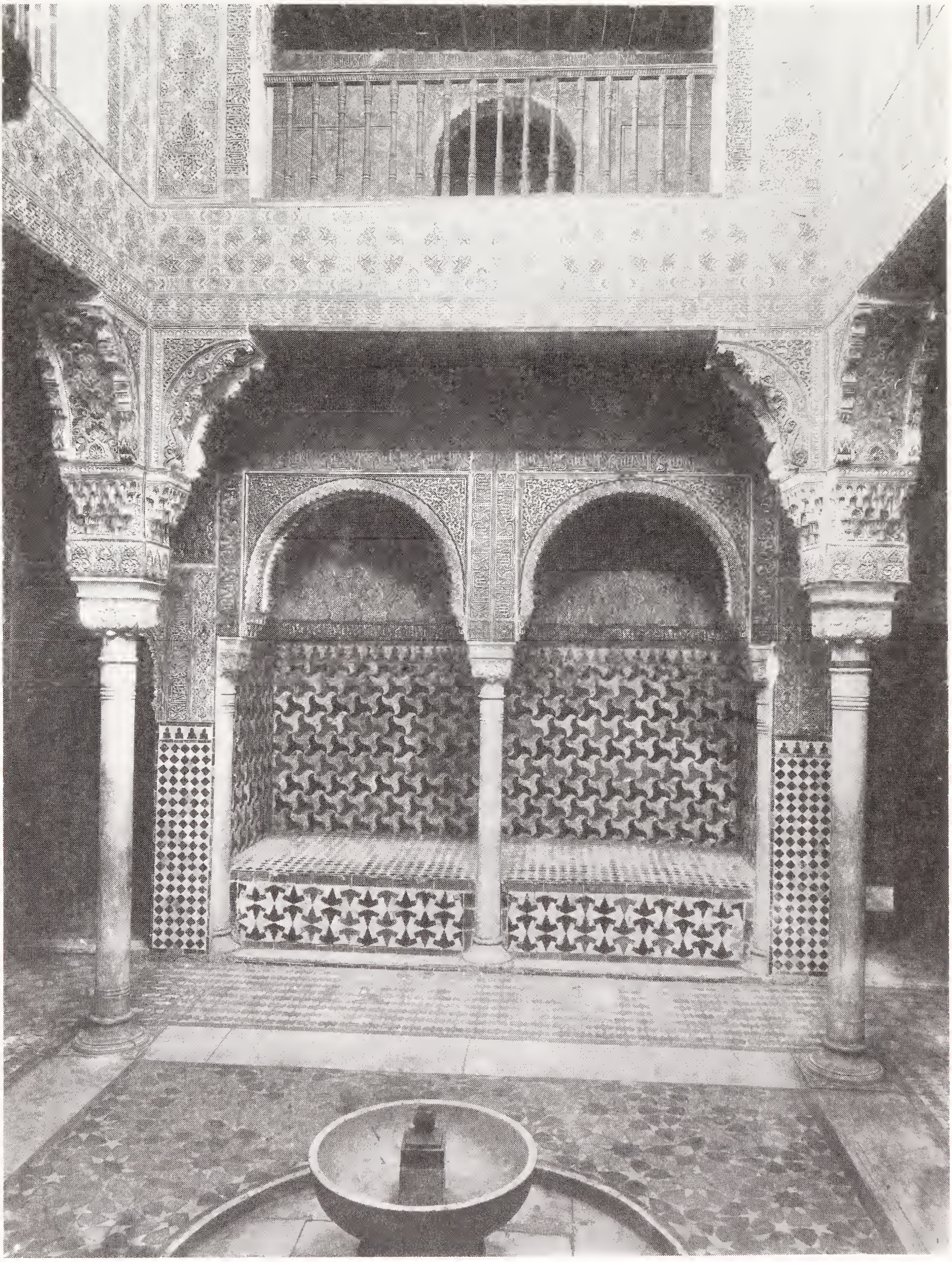


Figure 12.1: The Alhambra. Like the Parthenon, the Taj Mahal, and the Roman aqueducts, this building uses principles of geometry to determine a shape that is both strong and beautiful. In addition, its walls are decorated with abstract geometric patterns that incorporate symmetries of plane figures. The Bettmann Archive.

scholars at this institute in the early days, so that some aspects of Hindu algebra must have been known from the very beginning. In particular the notion of zero came with these scholars. The Sanskrit word *sunya* (empty) was translated into Arabic as *sifr*. This word came into Latin as *zephyrum* and ultimately into English as the words *zero* and *cipher*. Among the early scholars was a mathematician and astronomer from the territory now known as Uzbekistan. His name was Muhammad ibn Musa Al-Khwarizmi (Muhammed, son of Moses, from Khorezm, 780–850). He wrote several works, not all of which have survived. Among the works attributed to him with high probability is an *Art of Hindu Reckoning*, which was so influential in its Latin translation that the Hindu numerals came to be known in Europe as Arabic numerals. The Arabic original of this work no longer exists, and it is not known who translated it into Latin. The Latin manuscript was published in 1857 by an Italian nobleman who discovered it in the library of Cambridge University. The reason for attributing the original to Al-Khwarizmi is that the Latin text begins, "Al-Khwarizmi has said..." (Dixit algoritmi...). In this way a mathematician of 1200 years ago has given his name to one of the central concepts of modern mathematics and computer science. It was through the works of Al-Khwarizmi that the techniques of arithmetic came to be known in Europe as algorism; and the name has remained, in the modern form of *algorithm* to denote any systematic procedure for solving a problem in a finite number of steps.

12.2.1 Algebra

Although Al-Khwarizmi's treatise on arithmetic did not survive in the original Arabic, his *Algebra* did, and gave us a second word of central importance in science. This work, which is the main source of the fame of Al-Khwarizmi, bears the Arabic title: *Kitab fi al-jabr wa'l-muqabala.* This title contains the origin of the word *algebra.* The words refer to restoring or reuniting (one meaning of *muqabala* is a meeting place).[1] The title refers to the processes performed on algebraic expressions in order to solve an equation, that is, keeping the equation in balance by performing the same operation on both sides, or more specifically, gathering all like terms on the same side of the equation. The word *jabr* originally referred to adding the same positive quantity to both sides of an equation so as to remove negative terms, while *muqabala* meant canceling like terms from the two sides of the equation.

In his preface Al-Khwarizmi describes his book:

> ...a short work on Calculating by the rules of completion and reduction, confining it to what is easiest and most useful in arithmetic, such as men constantly require in cases of inheritance, legacies, partition, law-suits, and trade, and in all their dealings with one another, or where the measuring of lands, the digging of canals, geometric computation, and other objects of various sorts and kinds are concerned... .

[1] Recall the discussion of the Japanese *tenzan* in the previous chapter.

As this preface indicates, the work is intended to be "practical mathematics" in the sense of many modern books bearing that title. The subject matter is not confined to algebra, although the technique of setting up and solving an equation is a common thread throughout the book. Because of the Hindu connection mentioned above, there is no doubt that the basic ideas of algebra came to the Muslims from India. As we saw in Chapter 9, the *Vijaganita* (source computation) of Brahmagupta contained all the necessary elements: symbols (color names) for the unknown and rules for manipulating expressions involving unknowns. The development of these ideas by Brahmagupta, however, was rudimentary, covering only the case of a few quadratic equations and some linear systems of equations in more than one unknown.

In this earliest Muslim algebra, the extent of the subject was even more limited, however, an indication that the writers were themselves still striving to understand the work of the earlier Hindu scholars. For example, Al-Khwarizmi does not use negative numbers as data in his equations, although he recognizes the rules for operating with negative numbers, and at one point refers to negative roots. This absence prevented the theory of equations from becoming as unified as it might have been. It is, however, real algebra, since the central concept is an unknown appearing in one or more formal expressions representing data, from which the unknown number is to be determined. Certainly Al-Khwarizmi is much closer to our way of thinking than Diophantus was. (Recall that the Diophantine "equations" were not always thought of as equations; Diophantus asked such questions as how a number could be divided into the sum of two squares.) The notation used by Al-Khwarizmi, however, is entirely rhetorical, with no symbolism of any kind. He always uses a geometric figure to illustrate his solution of an equation. Consider, for example, his solution of the following problem:

> *A square and 10 roots equal 39 dirhems.* [A *dirhem* was a monetary unit.]

Al-Khwarizmi's solution of this problem is to draw a square of unspecified size (the side of the square is the desired unknown) to represent the square (Fig. 12.2). To add 10 roots, he then attaches to each side a rectangle of length equal to the side of the square and width 2.5 (since 4 times 2.5 equals 10). The resulting cross-shaped figure has, by the condition of the problem, area equal to 39. Now to fill in the four corners of the figure (literally "completing the square"), he adds 4 squares, each 2.5 on a side, having total area $4 \times (2.5)^2$ or 4×6.25, that is, 25. Since $39 + 25 = 64$, the completed square has side 8. Since this square was obtained by adding rectangles of side 2.5 to the original square, it follows that the original square had side 3.

Because negative numbers were not considered as data for a problem, it was necessary for Al-Khwarizmi to consider separately various classes of quadratic equations: square plus root equals number, square plus number equals root, square equals root plus number, etc. Each type called for its own technique, illustrated with rectangles and squares, as in Fig. 12.2. Al-Khwarizmi did not develop the theory of cubic equations. Roughly the first third of the book is devoted to various examples of pure mathematical problems leading to quadratic equations, causing

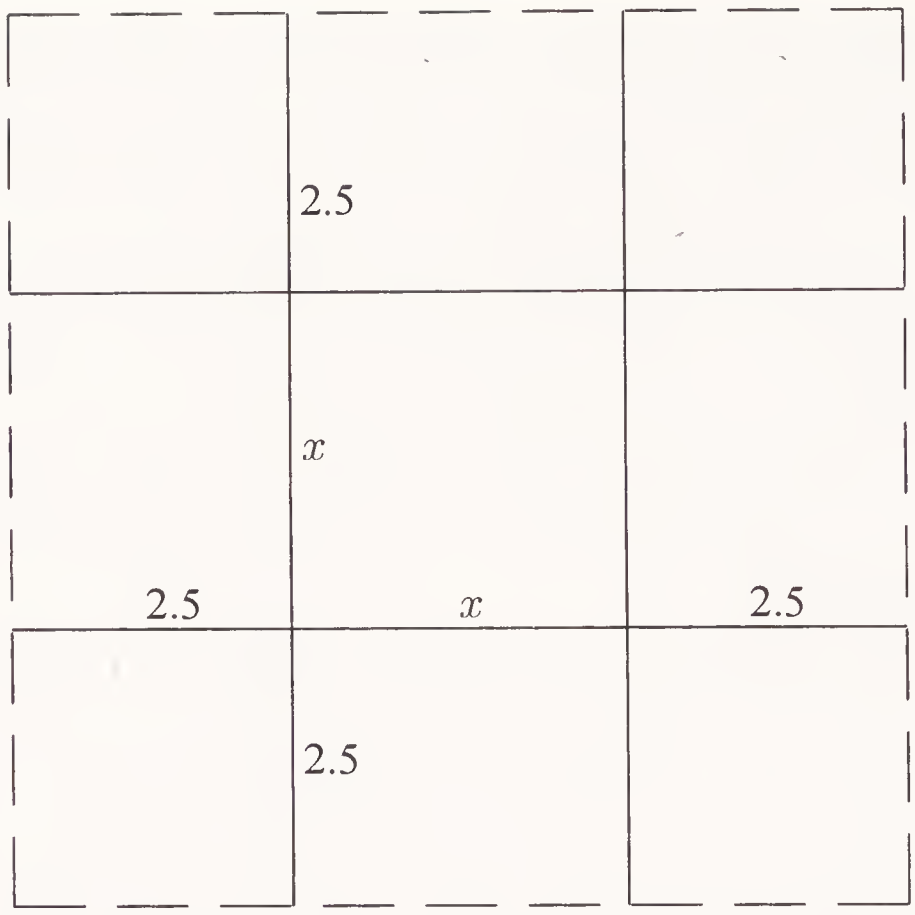


Figure 12.2: Al-Khwarizmi's solution of "square plus 10 roots equals 39."

the reader to be somewhat skeptical of his claim to be presenting the material needed in commerce and law. In fact, there are no genuine applications of quadratic equations in the book.

12.2.2 Geometry

As the problem discussed above indicates, this algebra is very geometric, and part of Al-Khwarizmi's book is devoted to formulas for areas and volumes. Al-Khwarizmi's indebtedness to the Hindus is shown in his use of the terms *bow* and *arrow* to refer to an arc of a circle and the perpendicular from the midpoint of the arc to its chord. Al-Khwarizmi gives three different ratios for the number π.

> In any circle the product of its diameter multiplied by $3\frac{1}{7}$ will be equal to the periphery. This is the rule generally followed in practical life, though it is not quite exact. The geometricians have two other methods. One of them is, that you multiply the diameter by itself; then by 10, and hereafter take the root of the product; the root will be the periphery. The other method is used by the astronomers among them: it is this, that you multiply the diameter by 62,832, and then divide the product by 20,000; the quotient is the periphery. Both methods come very nearly to the same effect.

Thus for the number π Al-Khwarizmi gives the Archimedean value $\frac{22}{7}$, the traditional $\sqrt{10}$, and Aryabhata's 3.1416.

Al-Khwarizmi gives formulas for the volume and area of many simple geometric figures, all of which are correct from the modern point of view.

12.2.3 Applications

Let us now consider how well Al-Khwarizmi lives up to his advertised purpose of providing the solution of practical problems. It may be worthwhile also to compare his algebra with the geometric algebra found in Euclid. The two are quite similar in that the problems solved are confined to linear and quadratic equations. That the motivation for them was different is shown by the fact that Al-Khwarizmi refers to the unknown as a *root* (*jadhr* in Arabic), while the Greeks had always called it a *side* (*pleurá* in Greek). The Greek name suggests that the unknown was thought of as the side of a square or cube. Latin works translated from Arabic always use the word *radix* (root) while those translated from Greek use the word *latus* (side). These considerations suggest that the motivation for Al-Khwarizmi's algebra was in arithmetic rather than geometry, but the questions remain: How did algebra arise? What motivated the problems and techniques discussed by Al-Khwarizmi? Was there some practical or commercial problem requiring their use?

When we read the *Algebra* in this light, we notice a peculiar fact—Quadratic equations never occur in the applications either in geometry or commerce. Actually there is no need to be surprised by this fact. In everyday life the average person today *never* needs to solve a quadratic equation, and things were not any different in earlier eras. Indeed, an inspection of any algebra text written in the last thousand years will not disclose a single practical application of quadratic equations in everyday life, although many books fraudulently claim such applications.

But if quadratic equations have no practical applications (outside of technology, of course), what about linear equations? Here there definitely are practical applications. A single linear equation, however, can be solved using only the standard four operations of addition, subtraction, multiplication, and division. The contribution that algebra can make to such applications is marginal at best. Nevertheless, there are occasions when the analysis really calls for algebra, and Al-Khwarizmi found many such cases in problems of inheritance, which occupy more than half of his *Algebra*. We now give a sample.

> A man dies leaving two sons behind him, and bequeathing one-fifth of his property and one dirhem to a friend. He leaves 10 dirhems in property and one of the sons owes him 10 dirhems. How much does each legatee receive?

Before studying Al-Khwarizmi's solution, consider for a moment how this estate would be settled under modern law. The man's estate would be considered to consist of 20 dirhems, the 10 dirhems cash on hand, and the 10 dirhems owed by one of the sons. The friend would be entitled to 5 dirhems (one-fifth plus one dirhem), and the indebted son would owe the estate 10 dirhems. His share of the estate would be one-half of the 15 dirhems left after the friend's share is taken out, or $7\frac{1}{2}$ dirhems. He would therefore have to pay $2\frac{1}{2}$ dirhems to the estate, providing it with cash on hand equal to $12\frac{1}{2}$ dirhems. His brother would receive $7\frac{1}{2}$ dirhems.

Now the notion of an estate as a legal entity that can owe and be owed money is a modern European one, alien to the world of Al-Khwarizmi. Apparently in

Al-Khwarizmi's time money could be owed only to a *person*. What principles are to be used for settling accounts in this case? Judging from the solution given by Al-Khwarizmi, the estate is to consist of the 10 dirhems cash on hand, plus a *certain portion* (not all) of the debt the second son owed to his deceased father. This "certain portion" is the unknown in a linear equation, and is the reason for invoking algebra in the solution. It is to be chosen so that *when the estate is divided up, the indebted son neither receives any more money nor owes any to the other heirs*. This condition leads to an equation that can be solved by algebra. Al-Khwarizmi explains the solution as follows (we put the legal principle that provides the equation in capital letters):

> Call the amount taken out of the debt *thing*. Add this to the property; the sum is 10 dirhems plus *thing*. Subtract one-fifth of this, since he has bequeathed one-fifth of his property to the friend. The remainder is 8 dirhems plus $\frac{4}{5}$ of *thing*. Then subtract the 1 dirhem extra that is bequeathed to the friend. There remain 7 dirhems and $\frac{4}{5}$ of *thing*. Divide this between the two sons. The portion of each of them is $3\frac{1}{2}$ dirhems plus $\frac{2}{5}$ of *thing*. THIS MUST BE EQUAL TO THING. Reduce it by subtracting $\frac{2}{5}$ of *thing* from *thing*. Then you have $\frac{3}{5}$ of *thing* equal to $3\frac{1}{2}$ dirhems. Form a complete *thing* by adding to this quantity $\frac{2}{3}$ of itself. Now $\frac{2}{3}$ of $3\frac{1}{2}$ dirhems is $2\frac{1}{3}$ dirhems, so that *thing* is $5\frac{5}{6}$ dirhems.

One of the more intriguing aspects of the *Algebra* is the mixture of practical legal considerations with mathematics. For example, Al-Khwarizmi considers the case of a man who marries while in his final illness and pays a marriage settlement of his entire property in the amount of 100 dirhems, 10 dirhems of which was his wife's dowry. His plans are upset, however, as his wife dies first, leaving one-third of her property to a third party, after which the husband dies. There are then three sets of claimants to the 100 dirhems: (1) the third party, (2) the wife's direct heirs (her family), and (3) the husband's direct heirs (his children or parents). How is the estate to be divided up?

The translator of Al-Khwarizmi's work has suggested that the many arbitrary principles used in these problems were introduced by lawyers to protect the interests of next-of-kin against those of other legatees.

12.3 Abu Kamil

A commentary on Al-Khwarizmi's *Algebra* was written by the mathematician Abu Kamil (ca. 850–930). His exposition of the subject contained none of the legacy problems found in Al-Khwarizmi's treatise, but after giving the basic rules of algebra, it listed 69 problems of considerable intricacy to be solved. For example, a paraphrase of problem 10 is as follows:

> The number 50 is divided by a certain number. If the divisor is increased by 3, the quotient decreases by $3\frac{3}{4}$. What is the divisor?

Abu Kamil is also noteworthy because many of his problems were copied by Leonardo of Pisa (Fibonacci) in his thirteenth-century treatise on algebra, one of the first works to introduce the mathematics of the Muslims into Europe.

12.4 Thabit ibn Qurra

About two generations later than Al-Khwarizmi another great commentator and mathematician worked in Baghdad translating Greek and Syriac treatises and taking the opportunity to carry out his own mathematical research. This mathematician, Thabit ibn Qurra (836–901), is the only source for three of the books of Apollonius' *Conics*, and he also translated many works of Archimedes, Euclid, Ptolemy, and others. In the course of making these translations he generated a good deal of mathematics of his own, especially in number theory and geometry.

12.4.1 Number Theory

We have already mentioned the standard way of generating perfect numbers in Chapter 5, namely the Euclidean formula $2^{n-1}(2^n - 1)$, whenever $2^n - 1$ is a prime. Thabit ibn Qurra found a similar way of generating pairs of *amicable* numbers, that is, pairs of numbers such that each is the sum of the proper divisors of the other. His formula is

$$2^n(3 \cdot 2^n - 1)(3 \cdot 2^{n-1} - 1) \text{ and } 2^n(9 \cdot 2^{2n-1} - 1),$$

whenever $3 \cdot 2^n - 1$, $3 \cdot 2^{n-1} - 1$, and $9 \cdot 2^{2n-1} - 1$ are all prime. The case $n = 2$ gives the pair 220 and 284. No one knows just how many new cases can be generated from this formula, but there definitely are some. For example, when $n = 4$, we obtain the amicable pair $17,296 = 16 \cdot 23 \cdot 47$ and $18,416 = 16 \cdot 1151$. Indeed if we add up the divisors of 17,296, we find

$$\begin{aligned} 1 + 2 + 4 + 8 + 16 + 23 + 46 + 92 + 184 + 368 + 47 + 94 + \\ + 188 + 376 + 752 + 1081 + 2162 + 4324 + 8648 = 18,416, \end{aligned}$$

and if we add up the divisors of 18,416, we find

$$1 + 2 + 4 + 8 + 16 + 1151 + 2302 + 4604 + 9208 = 17,296.$$

Unlike some other number-theory problems such as the Chinese remainder theorem, which arose in a genuinely practical context, the theory of amicable numbers is an offshoot of the theory of perfect numbers, which was already a completely "useless" topic from the beginning. It did not seem useless to the people who developed it, however. The tenth-century mystic Al-Majriti of Madrid recommended as a love potion writing the numbers on two sheets of paper and eating the number 284, while causing the beloved to eat the number 220. (He claimed to have verified the effectiveness of this charm by personal experience!)

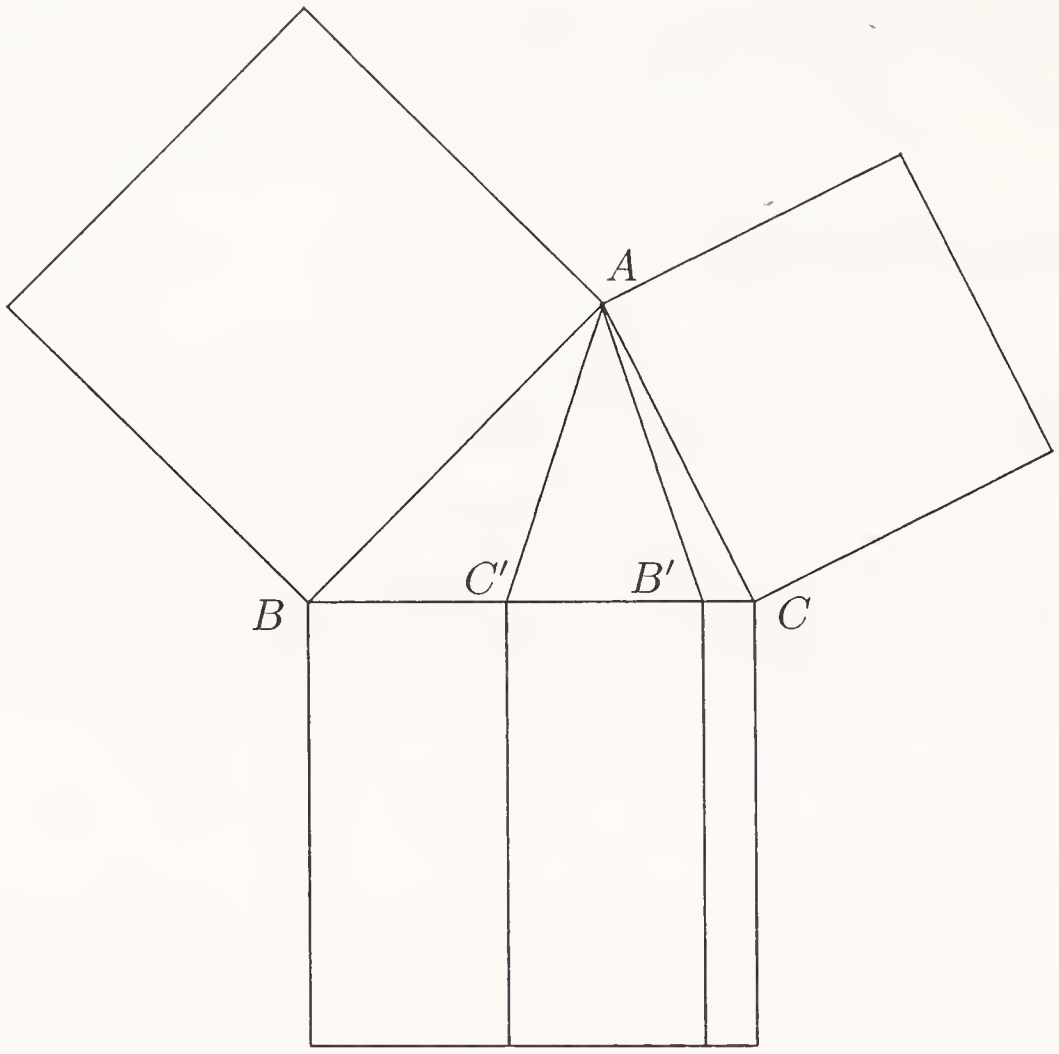


Figure 12.3: Thabit ibn Qurra's Pythagorean theorem.

12.4.2 Geometry

Thabit ibn Qurra also contributed a famous generalization of the Pythagorean theorem, different from the generalization by Pappus we discussed previously. The new theorem is easily derived from similar triangles. Consider a triangle ABC whose longest side is BC The idea is to copy angle B with A as vertex and AC as one side, extending the other side to meet BC in point C', then to copy angle C with A as vertex and BA as one side, extending the other side to meet BC in point B', so that angle $AB'B$ and angle $AC'C$ both equal angle A. It then follows that the triangles ABB' and ACC' are similar to the original triangle, and so $\overline{AB}^2 = \overline{BC} \cdot \overline{BB'}$ and $\overline{AC}^2 = \overline{BC} \cdot \overline{CC'}$, hence

$$\overline{AB}^2 + \overline{AC}^2 = \overline{BC}(\overline{BB'} + \overline{CC'}).$$

The case when angle A is acute is shown in Fig. 12.3.

12.4.3 Other Work

Thabit ibn-Qurra was a versatile scholar whose contributions to mathematics are difficult to summarize. They are known to us through remarks made by later authors, who credit him with an angle trisection that is basically the same as that of Pappus (using a hyperbola). He independently rediscovered some of Archimedes'

quadratures and found the volume of the figure obtained by revolving a parabolic segment about its axis. He also wrote philosophical essays on the nature of number and geometry, trying, like the Jainas, to make sense of the infinite. He speculated on the seeming paradox that both the even and the odd numbers are infinite, and he claimed that the number of even numbers was only half of the number of all numbers.

12.5 Omar Khayyam

Nearly everyone has heard of the *Rubaiyat* of Omar Khayyam, and most people have at one time had occasion to memorize the opening lines of its translation by the English poet Edward Fitzgerald ("A Book of Verses underneath the Bough, A Jug of Wine, a Loaf of Bread and Thou,..."). This multitalented Persian of the eleventh century (1050–1123) was also a distinguished scientist. He wrote his scientific works in Arabic, which at the time was the language of science. Omar Khayyam's *Algebra* gives a thorough classification of quadratic and cubic equations and shows how to solve them geometrically.

The influence of Greek geometry on Omar Khayyam's algebra is seen in his denial of the possibility of forming a fourth power (Diophantus' *dynamis dynamis*). To Omar Khayyam, the unknown had to be represented by a line segment, the product by a rectangle, etc., in the strict Euclidean tradition. He made the following assertion:

> I say what is called *square square* by algebraists in continuous magnitude is a theoretical fact. It does not exist in reality in any way.

12.5.1 The Cubic Equation

Omar Khayyam did not have modern algebraic symbolism. He lived within the confines of the universe constructed by the Greeks. His classification of equations, like Al-Khwarizmi's, is conditioned by the use of only positive numbers as data. For that reason his classification is even more complicated than Al-Khwarizmi's, since Omar Khayyam considers cubic equations as well as quadratics. Nevertheless, even though he illustrates his solutions with geometry, it is clear that the object of study is the equation, which it was not in Euclid.

Omar Khayyam shows how to handle many varieties of cubic equations. We shall illustrate these techniques with one particular example, his solution of a cubic equation by use of the rectangular hyperbola. The particular form considered is *cubes plus squares plus roots equal number*, or, as we would phrase it, $x^3 + ax^2 + bx = c$. In keeping with his geometric interpretation of magnitudes as line segments, Omar Khayyam had to regard the coefficient b as a square, so that we shall write b^2 rather than b. Similarly he regarded the constant term as a solid, which without any loss of generality he considered to be a rectangular prism whose base was an area equal to the coefficient of the unknown. In keeping with this reduction we shall write b^2c instead of c. Thus Omar Khayyam actually considered

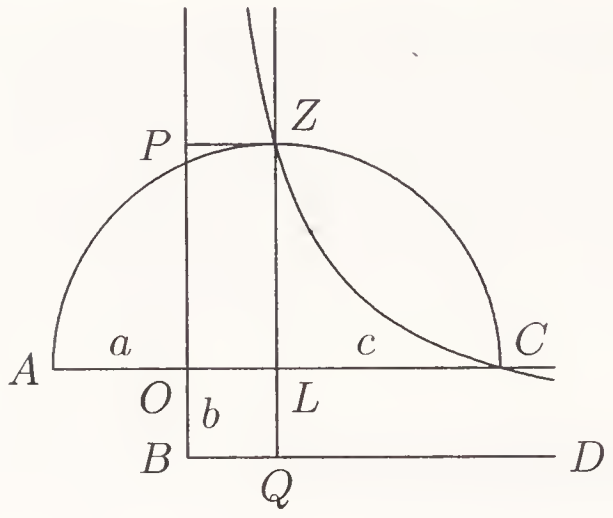


Figure 12.4: Omar Khayyam's solution of $x^3 + ax^2 + b^2x = b^2c$.

the equation $x^3 + ax^2 + b^2x = b^2c$. where a, b, and c are data for the problem. His solution is illustrated in Fig. 12.4. He drew a pair of perpendicular lines intersecting at a point O and marked off $OA = a$ and $OC = c$ in opposite directions on one of the lines and $OB = b$ on the other line. He then drew a semicircle having AC as diameter, followed by the line DB through B perpendicular to OB (parallel to AC), and the rectangular hyperbola through C having DB and the extension of OB as asymptotes. This hyperbola intersects the semicircle in the point C and in a second point Z. From Z he drew ZP perpendicular to the extension of OB and ZQ perpendicular to DB and intersecting AC at L. Then ZP represented the solution of the cubic. This fact follows quite easily from two other facts. The first of these is an elementary property of the hyperbola, namely that the product of the perpendiculars from points on the hyperbola to the asymptotes is constant, that is, $\overline{ZP} \cdot \overline{ZQ} = \overline{OC} \cdot \overline{OB} = cb$. Thus $bc = \overline{ZP} \cdot \overline{ZQ} = \overline{ZP} \cdot (\overline{ZL} + b)$, which can be written as $b \cdot (c - \overline{ZP}) = \overline{ZL} \cdot \overline{ZP}$. The second fact is the fundamental property of half-chords to a diameter, that $\overline{ZL}^2 = \overline{CL} \cdot \overline{LA} = (c - \overline{ZP}) \cdot (a + \overline{ZP})$. We can rewrite these two equations as proportions:

$$\overline{ZL} : (c - \overline{ZP}) = b : \overline{ZP}$$
$$\overline{ZL}^2 : (c - \overline{ZP})^2 = (\overline{ZP} + a) : (c - \overline{ZP}).$$

Squaring the first of these proportions and substituting into the second gives

$$b^2 : \overline{ZP}^2 = (\overline{ZP} + a) : (c - \overline{ZP}),$$

and cross-multiplying the proportion gives

$$\overline{ZP}^3 + a\overline{ZP}^2 + b^2\overline{ZP} = b^2c,$$

showing that $\overline{ZP}$ is indeed a solution.

We now ask a fundamental question: In what sense did Omar Khayyam solve the cubic equation? Certainly his method gives a *graphical* solution *provided one can draw a rectangular hyperbola through a given point having given asymptotes.* That, however, is a large *proviso.* We have no instrument that will do this. In this

sense Omar Khayyam's solution is a theoretical one, phrased in geometric terms. To find the numerical value of the root, one would have to perform a physical measurement.

Still, it is interesting that the solution of a cubic equation can be represented as the intersection of two simple geometric figures. It is worth emphasizing this point, precisely because it clarifies the difference between algebra as understood by Omar Khayyam and algebra as we know it today. For us a *numerical* answer in terms of numerical data is the only real solution. To what extent can Omar Khayyam's solution provide the kind of result we would demand?

For the answer to this question, we return to Omar Khayyam's solution, that is, the line segment ZP, and we ask what we know about it numerically. In proving his solution correct Omar Khayyam established two geometric facts about ZP. We can express these facts by the two equations

$$\overline{ZP} \cdot (\overline{ZL} + b) = bc,$$
$$\overline{ZL}^2 = (c - \overline{ZP}) \cdot (a + \overline{ZP}).$$

Thus, to find ZP numerically, it is necessary to eliminate ZL from these two equations and solve for ZP (that is, one must solve a pair of simultaneous quadratic equations in two unknowns). However, *any attempt to solve these equations merely leads back to the original cubic equation!*

Thus, from a computational point of view, Omar Khayyam's solution is circular, a mere restatement of the problem. We shall see that the cubic equation has a long history of solutions that in some cases turn out to be mere restatements of the problem, and in fact no method of solution exists (or can exist) that reduces the solution of every cubic equation with real roots to the extraction of real square and cube roots of real numbers.

12.6 The Foundations of Geometry

As the passage from Omar Khayyam quoted above shows, the Islamic mathematicians knew Euclid well enough to speculate on his defects. The most glaring of these defects, as it seemed to them, was the parallel postulate. The Islamic scholar Ibn Al-Haitham (950–1039), known traditionally in the West through his Latin name of Alhazen, attempted to prove the parallel postulate. His argument is illustrated in Fig. 12.5.

The argument runs as follows. Given two lines perpendicular to line AB at A and B, it will be proved that every perpendicular from one of them to the other is equal to AB. Thus in Fig. 12.5 AG and DB are drawn perpendicular to AB, and GD is perpendicular to DB. The claim is that $\overline{GD} = \overline{AB}$. To establish this claim, line segments GA and DB are doubled, to GE and DT respectively. It is easy to prove that $ET \perp BT$ and that ET is congruent to GD. (Simply draw the two diagonals GB and BE and use congruent triangles.) Al-Haitham claimed to have established that if the line segment ET is kept perpendicular to line DT and the point T is moved, then the other point E will trace a straight line. Thus

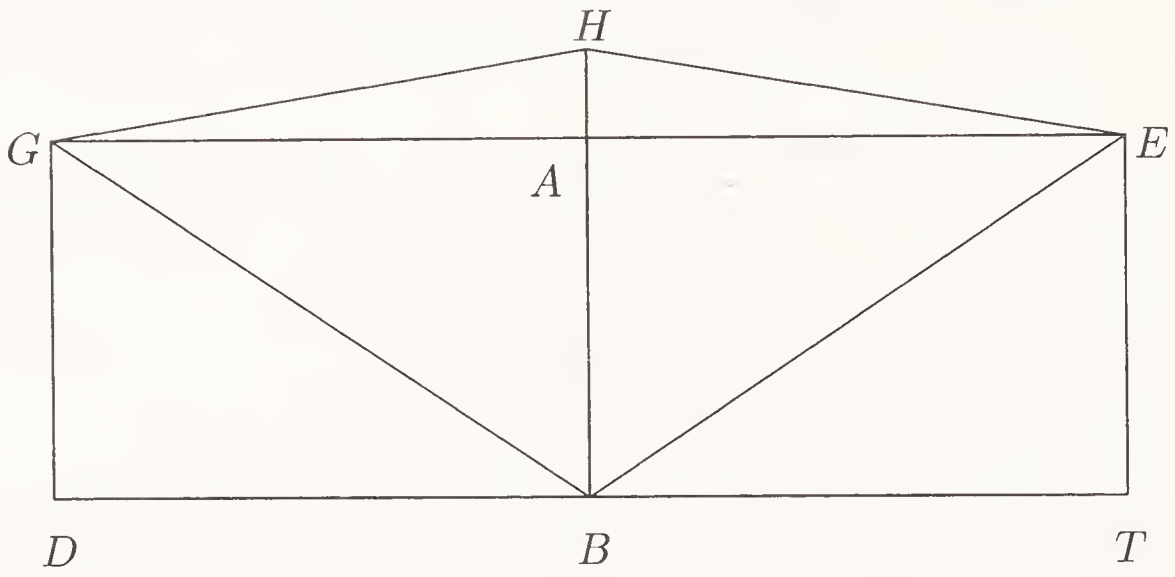


Figure 12.5: Al-Haitham's "proof" of the parallel postulate.

if ET is longer than AB, when T is moved to B, line segment ET will occupy an extension BH of the line segment BA. Then when T arrives at D, TE will coincide with DG. Hence, since it has traced a straight line, there will be two straight lines GAE and GHE joining the points G and E.

Omar Khayyam raised a number of questions for the benefit of those who do not see why this argument is wrong: He asked how a line could move while remaining perpendicular to a given line, and more generally how geometry and motion could be connected. Even admitting that Euclid allowed a line to be generated by a moving point and a surface by a moving line, he pointed out that Al-Haitham was requiring something more.

Mathematicians do not accept Al-Haitham's argument as valid today. The weakness of the argument lies in the claim that the set of points equidistant from a line is itself a line. This claim is equivalent to the parallel postulate. In hyperbolic geometry the curve described by Al-Haitham is called an *equidistant curve* and is *not* a straight line. The advance in Al-Haitham's work (based on an earlier idea of Thabit Ibn Qurra) is that it shows the relation between parallelism and the angle sum of a quadrilateral. The quadrilateral $DGET$, which has two equal sides perpendicular to a third side, is nowadays called a *Saccheri quadrilateral*, and the two quadrilaterals $AGDB$ and $AETB$, into which it is divided, having three right angles, are called *Lambert quadrilaterals*. These quadrilaterals are basic to the study of noneuclidean geometry.

12.7 Later Islamic Science

As the Christian West went on the offensive in the twelfth century, the Islamic world suffered increasing disruptions, and scientific research declined. Nevertheless, in the thirteenth century the astronomer Nasir-Eddin Al-Tusi (1201–1274) continued to work on the problem of the parallel postulate, introduced all the trigonometric functions that we know today, and discovered the interesting theorem that if a circle rolls without slipping inside a circle twice as large, each point

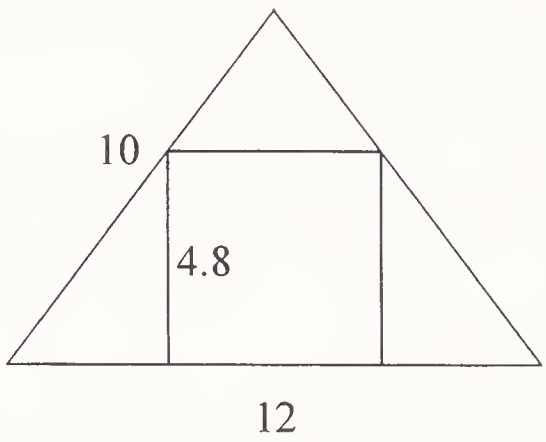


Figure 12.6: A problem from Al-Khwarizmi.

on the smaller circle moves back and forth along a diameter of the larger circle. (This is easy to prove, and an interesting exercise in geometry.) This kind of result might easily have led to the consideration of more general locus problems if social and political conditions had allowed Islamic scholarship to be sustained.

12.8 Problems and Questions

12.8.1 Problems in Islamic Mathematics

Exercise 12.1 Let x replace Al-Khwarizmi's word *thing* in the legacy problem discussed in the text. Replace each of the statements about *thing* by an equation and explain how the statements solve the equation. Compare the amount of money the three legatees receive under this solution with the amount they would receive under modern law.

Exercise 12.2 Solve the following geometric problem of Al-Khwarizmi: *Given an isosceles triangle with base 12 and legs each equal to 10, inscribe a square inside the triangle with one side along the base and the other two vertices on the legs. What is the side of the square?* First work out your own solution, then explain how Al-Khwarizmi obtained the equation

$$48 = x^2 + x\left(6 - \frac{x}{2}\right) + \frac{x}{2}(8 - x) = 10x,$$

so that $x = 4.8$. (See Fig. 12.6.)

Exercise 12.3 Solve the following legacy problem from Al-Kwharizmi's *Algebra*: *A woman dies and leaves her daughter, her mother, and her husband, and bequeaths to some person as much as the share of her mother and to another as much as one-ninth of her entire capital. Find the share of each person.* [It was understood from legal principles that the mother's share would be $\frac{2}{13}$ and the husband's $\frac{3}{13}$.]

Exercise 12.4 Solve the problem of Abu-Kamil in the text.

Exercise 12.5 Can the pair of amicable numbers 1184 and 1210 be constructed from Thabit ibn-Qurra's formula? Find one new pair that can be constructed from this formula, beyond those already mentioned.

Exercise 12.6 Explain how Thabit ibn-Qurra's generalization of the Pythagorean theorem reduces to that theorem when angle A is a right angle. What does the figure look like if angle A is obtuse? Is there an analogous theorem if BC is not the longest side of the triangle?

Exercise 12.7 One form of noneuclidean geometry, known as doubly elliptic geometry, is formed by replacing the plane with a sphere and straight lines with great circles, that is, the intersections of the sphere with planes passing through its center. Let one "line" (great circle) be the equator of the sphere. Describe the equidistant curve generated by the endpoint of a "line segment" (arc of a great circle) of fixed length and perpendicular to the equator when the other endpoint moves along the equator. Why is this curve not a "line"?

Exercise 12.8 If you know some modern algebra, explain why it is not surprising that Omar Khayyam's geometric solution of the cubic cannot be turned into an algebraic procedure. [*Hint:* Fill in the details of the following argument. Consider a cubic equation with rational coefficients, but no rational roots, such as $x^3 = 2$. The procedure for eliminating one variable between the two quadratic equations representing the hyperbola and circle is a rational one (it involves only multiplication and addition). Since the coefficients of the two equations are rational, the result of the elimination will be a polynomial equation with rational coefficients. If the root is irrational, that polynomial will be divisible by the minimal polynomial for the root over the rational numbers. However, a cubic polynomial with rational coefficients but no rational roots is itself the minimal polynomial for its roots.]

Exercise 12.9 Al-Haitham's attempted proof of the parallel postulate is fallacious because in noneuclidean geometry two straight lines cannot be equidistant at all points. Thus in a noneuclidean space the two rails of a railroad cannot both be straight lines. Assuming Newton's laws of motion (an object that does not move in a straight line must be subject to some force), show that in a noneuclidean universe one of the wheels in a pair of opposite wheels on a train must be subject to some unbalanced force at all times. [Note: The spherical earth that we live on happens to be noneuclidean. Therefore the pairs of opposite wheels on a train cannot both be moving in a great circle on the earth's surface.]

Exercise 12.10 Prove Nasir-Eddin's theorem about a circle rolling inside a second circle twice as large. Can you think of a mechanical application of this theorem?

12.8.2 Questions about Islamic Mathematics

Exercise 12.11 Why did Al-Khwarizmi include so much material on the solution of quadratic equations in his treatise, when he had no applications for them at all?

Exercise 12.12 Contrast the modern Western solution of the Islamic legacy problem with the solution of Al-Khwarizmi. Is one solution "fairer" than the other? Can mathematics make any contribution to deciding what is fair?

Exercise 12.13 Assuming that one had some practical application for cubic equations, would Omar Khayyam's solution of the cubic be adequate for practical purposes, or would one need some numerical procedure such as the Chinese were using about the same time for solving such equations?

Exercise 12.14 If numerical methods of solving equations can satisfy all practical needs, what value can there be in the efforts of people such as Omar Khayyam and the many European mathematicians who worked along the same lines to reduce the solution to the operations of arithmetic and the extraction of roots?

12.9 Endnotes

1. The distinction between the use of *side* and *root* for the solution to an equation was pointed out by Ali Abdullah Al-Daffa, in *The Muslim Contribution to Mathematics* (Humanities Press, Atlantic Highlands, NJ), 1977, p. 52.

2. The formula for amicable numbers attributed to Thabit ibn-Qurra was published by Franz Woepke in an article entitled "Notice sur une théorie ajoutée par Thábit ben Korrah à l'arithmétique spéculative des Grecs" in the *Journal Asiatique* (October-November 1852), pp. 420–429. [Cited by Cantor, *Vorlesungen über Geschichte der Mathematik* (Teubner, Leipzig, 1880), Vol. 1, p. 631.]

3. The story of the use of 220 and 284 as a love potion is told by Cantor (op. cit.), Vol. 1, p. 631.

4. Omar Khayyam's remark on the impossibility of a geometric product of four factors was taken from the source book by J. J. Gray and J. Fauvel, *The History of Mathematics: a Reader* (Macmillan Press, New York, 1987), pp. 225–226.

5. The attempted proof of the parallel postulate by Ibn Al-Haitham is quoted by Gray and Fauvel (op. cit.), pp. 235–236.

6. Omar Khayyam's criticism of Al-Haitham is quoted in full by Gray and Fauvel (op. cit.), p. 236.

PART III

Modern Mathematics

Modern mathematics began in fifteenth-century Italy with the development of algebra in a form that begins to resemble what is now taught in high schools. Before that time there was a period of several hundred years when Europeans gradually absorbed the mathematics that had been invented in the Islamic world and recovered as much as possible of the Greek heritage. The Medieval order in Europe had been based on the twin authorities of the Pope and the Emperor. In the fifteenth century this order broke down and was followed by a chaotic period of rival nation-states, wars of religion and territory, and the age of discovery. The complex cultural flowering known as the Renaissance and the scientific revolution developed from intellectual seeds planted in the Medieval universities. Accompanying the European colonial expansion, European science came to the older civilizations of Asia in the eighteenth and nineteenth centuries; these civilizations added to the sum of human knowledge, and the result was the larger scientific community that now exists.

The volume of mathematics created since the sixteenth century far exceeds all that we have studied up to now (and we have omitted important parts of even that story). Moreover, mathematics becomes increasingly sophisticated from this time on, and the essence of a mathematical achievement often cannot be comprehensibly summarized. For that reason as the story progresses we shall be forced to give fewer and fewer technical accounts of mathematical advances and concentrate instead on the impact of the work.

The background to modern mathematics lies in the Medieval period, when scholars assimilated the knowledge of the Islamic world and recovered some of the Greek works. Already in the fourteenth century European mathematicians were contributing new ideas of fundamental importance, such as the representation of variable quantities on a coordinate system. These Medieval advances were followed by the brilliant discovery in sixteenth-century Italy of algebraic techniques for solving cubic and quartic equations. The late sixteenth and early seventeenth centuries brought the invention of analytic geometry, projective geometry, and logarithms.

In the seventeenth and eighteenth centuries ideas that had been used individually for centuries were combined in new ways to produce the calculus, which was then applied to study an immense variety of physical phenomena. The calculus raised a number of questions, whose study led to the development of the theories of functions of a complex variable and functions of a real variable. In the nineteenth century progress was made on the solution of many old puzzles about Euclidean geometry and infinity, as well as the logical underpinnings of the calculus. Other mysteries involving the solvability of equations by radicals and the nature of probability were effectively solved or greatly clarified.

Analysis, which seemed to have acquired a rigorous foundation in the work of such mathematicians as Augustin-Louis Cauchy (1789–1856) and Karl Weierstrass (1815–1896), proved more resistant to ultimate clarity than had been expected, when the set theory created by Georg Cantor (1854–1918) in the late nineteenth century generated paradoxes. Our story comes to a close with a survey of the vast world of twentieth-century mathematics.

Chapter 13

Medieval Europe

13.1 The Early Middle Ages

The decline of cities in the West as the authority of the Roman Emperor failed was accompanied by a decline in scholarship. Only in the monasteries was learning preserved. As a result documents from this period tend to be biased toward issues that concern the clergy. Natural science and mathematics declined in importance among educated people and were replaced by theology, interpretation of scripture, hagiography, and church history. Nevertheless, science and mathematics were not entirely forgotten.

13.1.1 Boethius

The philosopher Anicius Manlius Severinus Boethius (480–524) wrote Latin translations of many classical Greek works of mathematics and philosophy and watered down their harder parts to adapt them to the intellectually degenerate time in which he lived. His works on mathematics proper are very elementary expositions of the simpler parts of Nicomachus and Euclid and are confined mostly to topics of use in measurement or related to philosophy; they fit the classical quadrivium of arithmetic, geometry, music, and astronomy, as depicted in Fig. 13.1.

Arithmetic and Geometry

The only topic discussed by Boethius that is not in Euclid's *Elements* is the abacus (a ruled board, not the device we now call an abacus). In this drastic abridgment the elaborate logical system of Euclid is lost entirely. The influence of this very simple mathematics on the imagination of people in the Middle Ages can be gauged from the last Canto of Dante's *Divine Comedy*, which describes the poet's vision of heaven:

> . . . As one,
> Who versed in geometric lore, would fain

Measure the circle; and, though pondering long
And deeply, that beginning, which he needs,
Finds not: e'en such was I, intent to scan
The novel wonder, and trace out the form,
How to the circle fitted, and therein
How placed: but the flight was not for my wing;
Had not a flash darted athwart my mind,
And, in the spleen, unfolded what it sought.
Here vigour fail'd the towering fantasy:
But yet the will roll'd onward, like a wheel
In even motion, by the Love impell'd,
That moves the sun in Heaven and all the stars.

Music and Astronomy

Boethius' work on astronomy is also derivative, based on Greek sources, and omits all the harder parts of Ptolemy's treatise. In addition, he wrote an influential book with the title *De institutione musica* that is of interest in the history of mathematics, since it adopts the Platonic (Pythagorean) point of view that music is a subdivision of arithmetic. Boethius divides the subject of music into three areas: *Musica Mundana*, which encompasses the "music of the spheres," that is, the regular mathematical relations observed in the stars and reflected in the sounds of nature; *Musica Humana*, which reflects the orderliness of the human body and soul; and *Musica Instrumentalis*, the music produced by physical instruments, which exemplify the principles of order noticed by the Pythagoreans, particularly in the simple mathematical relations between pitch and length of a string. For over a millennium such ideas had a firm grasp on writers such as Dante and scientists such as the seventeenth-century mathematician and astronomer Johannes Kepler. Indeed, *De institutione musica* was used as a textbook at Oxford until the eighteenth century. Kepler actually *wrote* the music of the spheres as he conceived it.

13.1.2 The Carolingian Empire

From the sixth to the ninth centuries a considerable amount of classical learning was preserved in the monasteries in Ireland, which had been spared some of the tumult that accompanied the decline of Roman power in the rest of Europe. From this source came a few scholars to the court of Charlemagne to teach Greek and the quadrivium (arithmetic, geometry, music, and astronomy) during the early ninth century. Charlemagne's attempt to promote the liberal arts, however, encountered great obstacles, as his empire was divided among his three sons after his death. In addition the ninth and tenth centuries saw the last waves of invaders from the north—the Vikings, who disrupted commerce and civilization both on the continent and in Britain and Ireland until they themselves became Christians and adopted a settled way of life. Nevertheless, Charlemagne's directive to create cathedral

Figure 13.1: The quadrivium, from Boethius' *Arithmetic*. Foto Marburg/Art Resource.

and monastery schools had a permanent effect, leading eventually the synthesis of observation and logic known as modern science.

13.1.3 Gerbert

In the chaos that accompanied the breakup of the Carolingian Empire and the Viking invasions the main source of stability was the Church. A career in public life for one not of noble birth was necessarily an ecclesiastical career, and church officials had to play both pastoral and diplomatic roles. That some of them also found time for scholarly activity is evidence of remarkable talent.

Such a talent was Gerbert of Aurillac. He was born to lower-class but free parents in south-central France some time in the 940s. He benefited from Charlemagne's decree that monasteries and cathedrals must have schools and was educated in Latin grammar at the monastery of St. Gerald in Aurillac. Throughout a vigorous career in the Church that led to his coronation as Pope Sylvester II in the year 999 he worked for a revival of learning, both literary and scientific. His work as secretary to the Archbishop of Reims was reported by a monk of that city named Richer, who described an abacus (counting board) constructed to Gerbert's specifications. It was said to have been divided into 27 lengths, and Gerbert astounded audiences with his skill in multiplying and dividing large numbers on this device.

While revising the curriculum in arithmetic Gerbert wrote a tract on the use of the abacus in which the Hindu–Arabic numerals were first introduced into northern Europe. This innovation caught on very slowly and required reintroduction several times.

Mathematical Activities

In some early letters written addressed to the monk Constantine of Fleury just before he became Abbot of Bobbio, Gerbert discusses some passages in Boethius' *Arithmetic*, and in the last letter written before he became pope, he writes to Adalbold of Liège about an inconsistency in Boethius' work. He discusses an equilateral triangle of side 30 and height 26 (since $26 \approx 15\sqrt{3}$), whose area is therefore 390. He says that if the triangle is measured by the arithmetical rule given by Boethius, that is, in terms of its side only, the rule is "one side is multiplied by the other and the number of one side is added to this multiplication, and from this sum one-half is taken." In our terms this would give area $s(s+1)/2$ to an equilateral triangle of side s. We recognize here the formula for a triangular number. Thus, guided by arithmetical considerations and figurate numbers, one would expect that this formula should give the correct area. However, in the case being considered, the rule leads to an area of 465, which is too large by 20%. Gerbert correctly deduces that Boethius' rule actually gives the area of a cross section of a stack of rectangles containing the triangle in question and that the excess results from the pieces of the rectangles sticking outside the triangle. He includes a figure to explain this point to Adalbold.

We can see from this discussion by one of the leading scholars of Europe to what an elementary level scientific and mathematical knowledge had sunk a thousand years ago. From these humble beginnings European knowledge of science underwent an amazing growth over the next few centuries.

13.1.4 Geometry

A picture of the level of geometric knowledge in the eleventh and twelfth centuries, before there was any major influx of translations of Arabic and Greek treatises, can be gained from an early twelfth-century treatise called *Practica geometriae* (The Practice of Geometry), attributed to Master Hugh of the Abbey of St. Victor in Paris.

The content of the *Practica geometriae* is aimed at the needs of surveying and astronomy and resembles the treatise of Gerbert in its content. This geometry, although elementary, is by no means unsophisticated. It discusses similar triangles and spherical triangles, using three mutually perpendicular great circles to determine positions on the sphere. After a discussion of the virtues and uses of the astrolabe, the author takes up the subjects of "altimetry" (surveying) and "cosmimetry" (astronomical measurements).

The discussion of "altimetry" is a straightforward application of similar triangles to measure inaccessible distances. The section on "cosmimetry" is of interest for two reasons. First, it gives a glimpse of what was remembered of ancient work in this area; and second, it shows what techniques were used for astronomical measurements in the twelfth century. The author begins by giving the history of measurements of the diameter of the earth, saying that the earth seems large to us, due to our confinement to its surface, even though "Compared to the incomprehensible immensity of the celestial sphere with everything in its ambit, earth, one must admit, seems but an indivisible point."

These views had been expressed by Ptolemy as justification for idealizing the earth as a point in his astronomy, and, of course, they are completely in accord with modern knowledge of the size of the cosmos. The author then goes on to discuss in detail the history of measurements of the circumference of the earth. He tells the famous story of Eratosthenes' measurement of a degree of latitude,[1] and mentions that Eratosthenes had overestimated the length of a degree by about 25%.

The author of the *Practica geometriae* continues by calculating the height of the sun by use of similar triangles. To do this, one must know the distance from the point of measurement to the point where the sun is directly overhead, then measure the length of the noontime shadow cast by a pole of known height. The author says that the Egyptians should be given credit as the first to compute solar altitude this way and that they were successful because their country was flat and close to the sun! The figure cited for the diameter of the sun's orbit (this is geocentric

[1] From the difference in altitude of the sun at Alexandria and Cyene Eratosthenes deduced that Alexandria was $7\frac{1}{2}$ degrees north of Cyene. By measuring this distance and dividing by $7\frac{1}{2}$, he calculated the length of a degree.

astronomy) is $9,720,181 + \frac{1}{2} + \frac{7}{22}$ miles. Using the value $\pi = \frac{22}{7}$, the author computes the length of the sun's orbit as $30,549,142\frac{5}{6} + \frac{1}{42}$ miles. (Needless to say, this number is only about 6% of the true value.)

13.2 The High Middle Ages

By the end of the eleventh century, the Medieval order in Europe had produced considerable prosperity, despite many local wars. The monasteries in particular were prosperous centers of both piety and learning.[2] European expansion began with the Christian reconquest of Spain from the Muslims (which took four centuries to complete) and the Crusades against the Turks in Palestine. In their sometimes violent rivalry with paganism and Judaism for adherents in the early centuries of the Christian era, Christian leaders had used appeal to reason as one of their strategies. This appeal had become unnecessary after Christianity became the official religion of the Empire at the end of the fourth century, but logical and theological disputation had continued in order to maintain doctrinal unity in the face of the heresies that were constantly arising. In the Middle Ages, faced with a rival religion that had displaced Christianity for several centuries in most of the Middle East, Christian scholars sought ways of competing with Islam in addition to war, whose outcome could not be guaranteed. Once again they cultivated reasoned debate. Centuries of debate in the monasteries and cathedral schools about fine points of theology produced scholars of formidable forensic ability. The attempt to settle fine metaphysical questions had led to a habit of scrutinizing arguments down to the tiniest hidden assumptions and reading ancient documents carefully to garner support from the authorities of antiquity. Although the attempt to establish the Christian faith on a foundation of pure reason would nowadays be considered a mistake by most people, this activity is an important part of the intellectual tradition that produced modern science when it was brought to bear on questions about the physical world. As Prof. David Lindberg expresses the situation, "...natural philosophy could not be separated from the rest of philosophy and, therefore, shared the fate of the larger whole of which it was a part."

13.2.1 The Revival of Mathematics

By the midtwelfth century European civilization had absorbed much of the learning of the Islamic world and was nearly ready to embark on its own explorations. This was the zenith of papal power in Europe, exemplified by the ascendancy of the popes Gregory VII (1073–1085) and Innocent III (1198–1216) over the emperors and kings of the time. The Emperor Frederick I, known as Frederick Barbarossa because of his red beard, who ruled the empire from 1152 to 1190, tried to maintain the principle that his power was not dependent on the Pope, but was ultimately

[2]The term *center of learning* is relative, however. The library holdings of most monasteries amounted to a few dozen books, compared with the thousands available in the libraries of the Islamic world at the same period.

unsuccessful. His grandson Frederick II (1194–1250) was a cultured man who encouraged the arts and sciences. To his court in Sicily he invited distinguished scholars of many different religions, and he corresponded with many others. He himself wrote a treatise on the principles of falconry. He was in conflict with the Pope for much of his life and even tried to establish a new religion, based on the premise that "no man should believe aught but what may be proved by the power and reason of nature," as the papal document excommunicating him stated.

13.2.2 Leonardo of Pisa

Into this empire, in the city of Pisa in the year 1170, there was born a man named Leonardo, the son of Gulielmo (William). Leonardo says in the introduction to his major book that he accompanied his father on an extended commercial mission in Algeria with a group of Pisan merchants. There, he says, his father had him instructed in the Hindu–Arabic numerals and computation, which he enjoyed so much that he continued his studies while on business trips to Egypt, Syria, Greece, Sicily, and Provence. Upon his return to Pisa he wrote a treatise to introduce this new learning to Italy.

The *Liber Abaci*

Leonardo's greatest work bears the title *Liber abaci* (The Book of the Abacus). As a document intended to promote the use of Hindu–Arabic numerals it is not a happy effort. Many of the problems reflect the routine computations that must be performed when converting currencies. These are applications of the rule of three that we find already in Brahmagupta and Bhaskara. Many of the other problems are purely fanciful and taken directly from Abu-Kamil. Moreover, from our present perspective Leonardo made things harder than they really needed to be by frequently expressing fractions as unit fractions. Here is a sample problem: *One-quarter and one-third of a tree lie below ground, a total of 21 palmi in length. What is the length of the tree?* The author imagines the tree divided into 12 equal parts, so that 7 of these parts are underground, then uses the proportion $7 : 21 = 12 : x$ to find the length of the tree.

The reader can easily figure out that the tree is 36 *palmi* high. Leonardo goes on to give a description of the rule of three for solving such proportions. The *Liber abaci* was not published in printed form until the nineteenth century.

The *Liber Quadratorum*

Leonardo wrote other books on mathematics that, when compared with the writings of Gerbert, show the extent to which scientific knowledge had increased over two centuries. In particular, his *Liber quadratorum* (Book of Squares) reflects the new vigor of intellectual life. In the prologue, addressed to the Emperor Frederick II, Leonardo says that he had been inspired to write the book because John of Palermo, whom he had met at Frederick's court, had challenged him to find a square number

such that if 5 is added to it or subtracted from it the result is again a square. This question inspired him to reflect on the difference between square and nonsquare numbers. He then notes his pleasure on learning that Frederick had actually read one of his previous books, and uses that fact as justification for writing on the challenge problem.

The *Liber quadratorum* is written in the spirit of Diophantus and shows a keen appreciation of the conditions under which a rational number is a square. For example, the ninth of its 24 propositions is, *Given a nonsquare number that is the sum of two squares, find a second pair of squares having this number as their sum.* Leonardo's solution of this problem involves a great deal of arbitrariness, since the problem does not have a unique solution (see Exercise 13.4).

Leonardo's Contribution to Mathematics

Leonardo of Pisa was a gifted mathematician who wrote treatises on both algebra and number theory. Moreover his approach to algebra begins to look modern, in that he uses letters to stand for unknown numbers. In one of his works called the *Flos* (Blossom) he considers the case of the cubic equation that we would write as $x^3 + 2x^2 + 10x = 20$. This equation has a unique positive root, which he gives in sexagesimal notation correct to six places. In addition, his approach to the problem contains a very important original element: he shows by using divisibility properties of numbers that there cannot be a rational solution or a solution obtained using only rational numbers and square roots of rational numbers. This kind of reasoning represents a new way of looking at equations, and one that was to be very fruitful for the subsequent development of algebra.

The securest basis of Leonardo's fame is a single problem from the *Liber abaci*:

> *How many pairs of rabbits can be bred from one pair in one year, given that each pair begins to breed in the second month after its birth, producing one new pair per month?*

By brute-force enumeration of cases, the author concludes that there will be 377 pairs, and "in this way you can do it for the case of infinite numbers of months."

The sequence generated here, namely 1, 1, 2, 3, 5, 8,..., in which each term after the second is the sum of its two predecessors, has been known as the *Fibonacci* sequence ever since the *Liber abaci* was first printed in the nineteenth century. The name Fibonacci seems to have been bestowed on Leonardo by the nineteenth-century historian of mathematics Guillaume Libri (1803–1869), under the mistaken impression that Leonardo's father was named Bonaccio (Fibonacci means "son of Bonaccio"). Bonaccio was apparently the family name, so that Fibonacci is equivalent to "of the Bonnaci." The Fibonacci sequence has been an inexhaustible source of identities, and many curious representations of its terms have been obtained, and there is a mathematical journal, the *Fibonacci Quarterly*, named in honor of Leonardo.

13.2.3 The Academic World

Scientific research, like all culture, depends on patronage. Scholars and artists have to be supported if they are to have time for their creative work; and this support, for those who are not of independent means, involves convincing someone with money that creative work is worth paying for. Leonardo of Pisa may have acquired the means to support his mathematical hobby through his commercial activity; he is known to have received a salary from the city of Pisa, connected with his teaching and consulting activity. Other scholars were supported at the court of Frederick II. Just why kings and emperors supported the arts and sciences by founding academies of sciences and universities is an interesting question to which there are many answers. The crudest answer is that they expected to gain something from the work of scholars and artists. We nowadays expect scientific advances to cure disease or provide us with new inventions to make our work easier and our leisure time more amusing. However, very little science and almost no mathematics actually has this effect. A more plausible answer is that real political power and influence depends on a number of very subtle factors, one of which is prestige. The huge palaces at Versailles, Charlottenburg, Windsor, and other places were intended to impress visiting dignitaries with the wealth and power of the realm. A sufficiently impressive display could have the effect of deterring aggression. Apart from such utilitarian explanations, however, there is a basic human sense of beauty that finds palaces, music, philosophy, and science satisfying. We may give some credit to the taste of monarchs, and assume that at least part of the reason for their support of academies of sciences and universities was a genuine desire to elevate the culture of the people.

Let us now return to more mundane matters and examine some of the work that was going on in universities during this time.

Reason in Philosophy and Theology

The rationalist program is exemplified by Anselm of Canterbury (1033–1109), who offered the "ontological" argument for the existence of God,[3] in contrast to earlier theologians, who appealed to an inspired reading of the Bible or to personal revelation or Church authority. This rationalizing work was continued in France by Peter Abélard (1079–1142), who wrote *Sic et non* (Yes and No), a collection of conflicting opinions by authorities on theological matters, to exhibit the necessity of reason in theology. Natural philosophy was not at first a major theme in this program. Indeed, one of the most influential metaphysical treatises of the times was Plato's *Timaeus*. The *Timaeus* contains a great deal of unscientific mysticism. It also, however, contains the crucial idea of a rationally ordered cosmos in which mathematics is the key to understanding.

Science as we know it, a blend of observation, experiment, and conceptual modeling, gained a foothold in the Medieval universities. By the year 1200 there

[3]In a somewhat distorted summary, this argument is that a greatest conceivable being must exist; for, if such a being did not exist, a greater one could be conceived, namely one having all the properties of the nonexisting greatest conceivable being, and also the property of existing.

were universities at Oxford, Paris, and Bologna. These universities grew out of cathedral schools, but developed in different ways. In Bologna students were in control of the administration, and were able in some cases to dictate the curriculum. From the time of the universities onward there is a very good documentary trail in which scholars tell about themselves and their motives; and it is clear from these documents that simple human curiosity is very often the strongest motive of all. Once a question has been asked, whether from urgent practical needs or simply as an attempt to relieve boredom, it acquires an importance of its own; and some scholar is almost certain to devote whatever time is necessary to finding the answer, whether the result justifies the effort in practical terms or not.

Although the most prominent philosophers of the late Roman Empire and the early Middle Ages were Platonists, the Platonic temperament is inclined toward mysticism and not at all methodical. The kind of close reasoning we find in Medieval philosophy is much more characteristic of Aristotle's cataloguing and classifying. Which was cause and which effect we do not have space to discuss, but the fact is that from the thirteenth century on Aristotle plays a much more prominent role than Plato in Medieval scholarship, which becomes increasingly concerned with analyzing the world through Aristotelian categories. This change encountered some early resistance, since Aristotle had been used as a source of philosophical support by the Islamic scholars to promote theological doctrines (such as the nonexistence of individual souls and the impossibility of miracles) that contradicted Christian orthodoxy. The teaching of Aristotelian doctrine was banned at the University of Paris in 1210, and this ban was reiterated in 1231 by Pope Gregory IX. Gradually ways were found to harmonize Aristotle with Christian theology, however, and in 1255 the study of Aristotle was made mandatory in Paris. Aristotle's metamorphosis within a single generation from being forbidden to being mandatory is an indication of some fundamental change of interest and outlook in the centers of learning. Let us now examine some of these new interests.

13.2.4 Jordanus Nemorarius

Little is known about the life of Jordanus Nemorarius, the thirteenth-century author of many works on mathematics and physics. Even the surname Nemorarius seems to suggest a pseudonym (*nemo* is Latin for *nobody*). What we do know is that he lived after the first Latin translation of Al-Khwarizmi's *Algebra* in 1145 and before the year 1250, when he is mentioned in a list of books called the *Biblionomia*. His work shows considerable progress in algebra in comparison with the work of Al-Khwarizmi. In one of his works entitled *De numeris datis* (On Given Numbers), for example, the well-known elementary fact that two numbers can be found if their sum and difference are known is generalized to the theorem that any set of numbers can be found if the differences of the successive numbers and the sum of all the numbers is known. A large variety of data sets that determine numbers then follows, for example, *if the sum of the squares of two numbers is known, and the square of the difference of the numbers is known, then the numbers can be found.* The four books of *De numeris datis* contain about a hundred such results.

13.2.5 Medieval Physics

The adoption of Aristotelian metaphysics meant that henceforth natural philosophy would become a logical subject in which the concepts of cause and effect were central. This way of looking at the world took a powerful hold everywhere, and today it is a basic element in the everyday thought of most people. In Medieval times these two aspects of the study of motion led to the important distinction between kinematics (the observed motion, which is an effect) and dynamics (the force causing the motion). It was possible to study kinematics independently of dynamics. Much of the history of mechanics can be interpreted as the attempt to provide a theoretical cause (dynamics) for an observed motion (kinematics).

The Science of Weights

The works of Archimedes were translated into Latin in the thirteenth century, and his work on the principles of mechanics was extended. One of the authors involved in this work was Jordanus Nemorarius, the author of several works on statics for which manuscripts still exist dating to the actual time of composition. The most sophisticated of these works bears the title *Liber Jordani de ratione ponderum* (Jordanus' Book on the Ratio of Weights). In this work, which consists of four parts, Jordanus begins with some Aristotelian principles that look suspicious, such as the postulate "That which is heavier descends more quickly." Nevertheless he improves on the results of Archimedes and Heron. In one result from Part 1 he shows that if two arms of different lengths suspended at different angles from a fulcrum have equal horizontal projections, then equal weights suspended from their ends will balance. This is a generalization of the basic Archimedean principle that equal weights suspended from the ends of two horizontal arms of equal length will balance. Moreover it recognizes implicitly the principle that the horizontal projection determines the moment of a weight suspended from an arm. Actually Jordanus did not have the concept of moment, but he did have the notion of "heavier (or lighter) in position," which fulfills the same function in the analysis of problems in statics.

This principle is used to obtain the final result of Part 1: *If two weights descend along diversely inclined planes, then, if the inclinations are directly proportional to the weights, they will be of equal force in descending.* Referring to Fig. 13.2, Jordanus states that the weights W_1 and W_2 will balance if $W_1 : W_2 = \overline{DC} : \overline{DA}$. This is precisely the modern law of the inclined plane. Recall that Heron of Alexandria had been mistaken in his analysis of this problem (Chapter 7), and Pappus also had given an erroneous solution.

The Merton Scholars

At Merton College, Oxford, in the midfourteenth century there was an active group of scholars with an interest in mathematics and physics. Among them were Thomas Bradwardine (1295–1349), William Heytesbury (1313–1372), and Richard

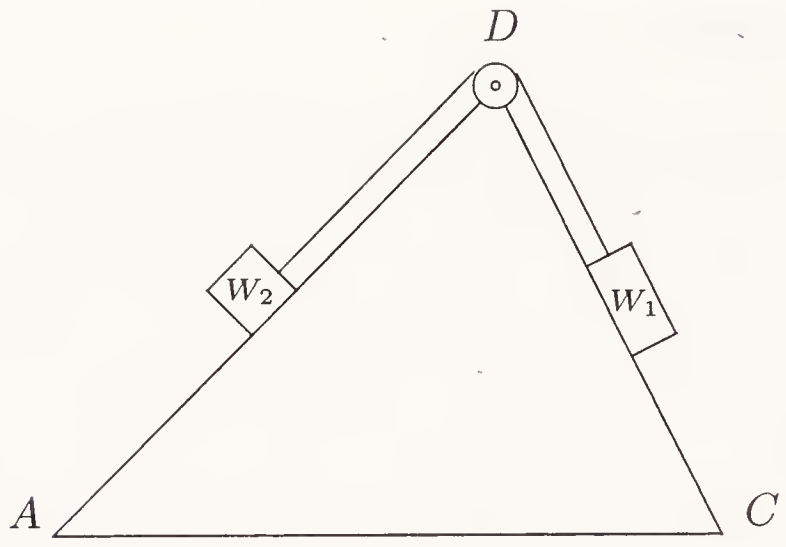


Figure 13.2: The law of the inclined plane ($W_1 : W_2 :: \overline{DC} : \overline{DA}$).

Swineshead.[4] They formulated and studied the concept of instantaneous velocity and uniformly accelerated motion. In contrast to the case of motion at constant velocity (the well-known rule $d = rt$ taught in junior high schools), the relation between distance and time is not simple when the velocity is not constant. To get a handle on more complicated types of motion, it is necessary to generalize step by step. The first natural generalization to be considered was that of uniform acceleration. Galileo, three centuries later, made the claim that falling bodies near the earth's surface undergo uniformly accelerated motion. The Merton scholars gave no way of attaching a *number* to instantaneous velocity (unless the velocity is constant). However, the Aristotelian categories came to their aid in formulating the intuitive idea. They thought of velocity as a quality of motion, and they recognized that qualities could exist in greater or lesser degree, i.e, objects can be hotter or colder, denser or rarer, and so forth, so that velocity played a role in discussing motion analogous to that of temperature in discussing heat or density in discussing weight.

In retrospect we can see that this is an extremely useful insight. The unity of the three examples given here is provided by the mathematical notion of proportion—the amount of matter per unit volume is by definition density, the thermal energy per unit volume (total translational kinetic energy of molecules) is proportional to the absolute temperature, and distance traversed is proportional to time when the velocity is constant. These examples also show the difficulty of the mechanical problem in comparison with the definitions of density and temperature, since in elementary situations the density of matter being considered is constant throughout a given sample and the temperature is also uniform, while velocity varies considerably even in simple problems. It would be many centuries before physics could consider the study of bodies in which the temperature or density varied from point to point. The first attempt to handle a "nonlinear" problem in physics occurs here in the problem of nonconstant velocity. This point also brings us back to a theme we addressed in connection with Greek science and geometry—the fundamental and pervasive presence of the idea of proportion in science, and consequently the

[4]Dates uncertain. There may have been as many as three scholars named Swineshead at Oxford during the fourteenth century.

fundamental importance of the arithmetic operations of multiplication and division, which correspond to this notion.

Without having a precise definition of instantaneous velocity, the Merton scholars defined uniformly accelerated motion as motion in which the velocity increases by equal increments in equal increments of time. For this kind of motion a rule known as the *Merton rule* was eventually distilled:

> *A body moving with a uniformly accelerated motion moves in a given time exactly the same distance it would move at constant velocity equal to its instantaneous velocity at the midpoint of the time interval under consideration.*

This rule was illustrated in many specific examples in a book by Swineshead called *Liber calculationum*. It was to play a very important role in the future of European mathematics and mechanics, becoming the object of study by the brilliant fourteenth-century scholar Nicole of Oresme. It provide one of the most important examples in the "new" science of mechanics created by Galileo, who claimed it was the kind of motion undergone by freely falling bodies.

The Concept of Force

It was mentioned in Chapter 7 that in Aristotelian physics the size of a force (mover) was measured by the distance an object of given size could be moved in a given time. If we anachronistically introduce the more precise concept of mass, we can use this property as a definition of force, that is, $F = kmd/t$, where k is a constant of proportionality, m the mass, d the distance moved, and t the time of the motion. Since d/t is just the velocity, we might say that force is proportional to mass times velocity. Actually the mass is irrelevant if the body is being rolled and not lifted; its function in this relation is to measure the resistance to the force (inertia), and so a more general way of phrasing the relation is $F = kRv$, where R is the resistance and v the velocity. In a treatise written in 1328 bearing the title *Tractatus proportionum* Thomas Bradwardine presented a variety of arguments in favor of a new way of thinking about force, resistance, and velocity. He argued that the Aristotelian relation allows any force to impart at least *some* velocity, no matter what the resistance, which is absurd if the resistance is greater than the force. According to Bradwardine, velocity is proportional to the ratio of the motive force to the resisting force.

This geometric proportionality means that in order to double a velocity it is necessary to square the ratio of motive force to resistance, while to get half the velocity, one would take the square root of the ratio. In this way the motive force must always be larger than the resistance, no matter how small the velocity required. Thus Bradwardine's principle avoids the paradox of a force overcoming a resistance larger than itself. By taking this step Bradwardine had introduced the concept of fractional powers. For one might wish to increase velocity by any real ratio, in which case it would be necessary to raise the ratio of motive force to resistance to the corresponding power. The challenge of realizing this operation

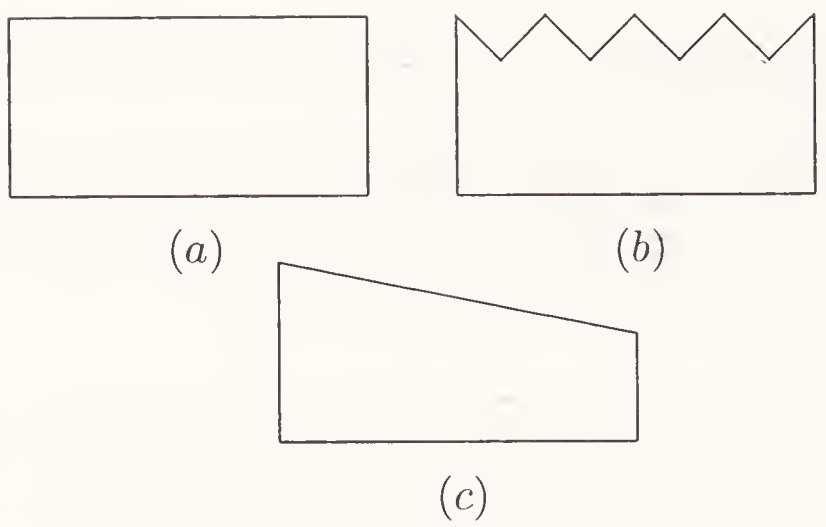


Figure 13.3: Oresme's classification of quadrangles: (a) uniform; (b) difform; (c) uniformly difform.

was undertaken by one of Bradwardine's successors, to whose works we now turn.

13.2.6 Nicole of Oresme

One of the most distinguished of the Medieval philosophers was Nicole of Oresme (1323–1382), whose clerical career brought him to the office of Bishop of Lisieux in 1377. Oresme had a wide-ranging intellect that covered economics, physics, and mathematics, as well as theology and philosophy. He considered the motion of physical bodies from various points of view, some of which physicists would nowadays not bother to consider. For example, it is an obvious fact that a falling body accelerates, but according to what law? We now believe that (neglecting air resistance, variations in the gravitational field of the earth, relativistic effects, and the like) the acceleration is constant, that is, the velocity is directly proportional to the time during which the body has been falling. However, how do we know that the velocity isn't proportional to the distance fallen? It might, in fact, have any one of infinitely many different mathematical relations to the time or the distance fallen. When mechanics was in its infancy, all of these possibilities had to be kept in mind.

Mathematical Physics

Oresme arrived at his results in a very Aristotelian way, by considering qualities. In a work entitled *Quaestiones super geometriam Euclidis* he discusses three kinds of "altitudes" for quadrangular figures. These are *uniform*, *difform*, and *uniformly difform*. The first of these simply means a figure all of whose altitudes to some base are equal, that is, a rectangle (although Oresme also uses the term *uniform* in a similar sense to describe the curvature of a circle). The second means any irregular figure. The third is explained by Oresme in terms that imply that the figure is bounded above and below by straight lines that are not parallel to each other (see Fig. 13.3).

The innovation made by Oresme was his use of these figures to represent motion. He argues that a two-dimensional figure must be used to represent any quality, one dimension to indicate its physical location (along a line), the other its intensity. He called these *longitude* and *latitude*, respectively. The total quality was thus to be represented as an area. Applying this principle to motion, he arrives at a representation of the distance traveled as an *area* whose two dimensions are longitude (representing the time of travel) and the latitude (representing the velocity). Anticipating an objection on the part of his reader, he notes that Aristotle had used lines to represent time.

Oresme formulated the Merton rule and for the first time in history explicitly used one line to represent time, a line perpendicular to it to represent velocity, and the area under the graph (as we would call it) to represent distance. He knew that it would be necessary to convince his readers that an area could be used to represent a line, and so he appealed to the ancients for examples of such usage. His conclusions were to be reiterated and further developed 250 years later in the mechanics of Galileo and the analytic geometry of Descartes.

The graphic representation of relationships between variable quantities was developed in more detail in Oresme's work *Tractatus de configurationibus qualitatum et motuum*. A treatise entitled *Tractatus de latitudinibus formarum* (Treatise on the Latitudes of Forms) was written shortly after the death of Oresme. (This work was once attributed to Oresme himself, but experts in the subject do not think this attribution is correct. The work is certainly based on what Oresme wrote, however.) Thus Oresme anticipated some of the ideas of analytic geometry, in particular the idea of mutually perpendicular coordinates and the use of coordinates to study complicated curves. This work was done long before a clear notion of a function had been formulated and is an extraordinary advance for one person to have made.

Ratios

Oresme followed up on Bradwardine's *Tractatus proportionum* with his own *Tractatus de proportionibus proportionum*. A reading of this work reveals that tremendous progress had been made in geometry since the time of Gerbert. Oresme presents his material in logical order with definitions, postulates, and propositions, together with rigorous proofs. Moreover he is careful to keep the distinction between commensurable and incommensurable, and he refers to Euclid's fifth book, which had not been mentioned by Boethius. Indeed, Oresme was even more advanced than the average twentieth-century person, in that he recognized a logical difficulty in talking about a power of, say, $\frac{1}{2}$ that equals $\frac{1}{3}$. He showed that two such ratios are incommensurable if there is no mean between the numbers of the greater ratio (for example, 1 and 3; the number 3 has no rational root of any order). In modern education students are taught how to use the rules of exponents and not encouraged to ask what is meant by, say, $\sqrt{2}^{\sqrt{3}}$. If only rational numbers are considered to exist, it will be very unusual if one can raise a given rational number to a given rational power, as would be required by Bradwardine's rule.

13.3 The Late Middle Ages

The fourteenth century was a period of turmoil in Europe as the Plantagenet kings of England attempted to assert claims to rule in France. Until this conflict was settled there was intermittent war from 1328 until 1452 (the Hundred Years' War). Even more devastating calamities were to come. In the midfourteenth century came the first of many epidemics of bubonic plague, which is estimated to have killed more than one third of the population. Needless to say, the survivors of this epidemic were deeply affected by the catastrophe. In view of these conditions it would be remarkable if cultural progress continued. The advance of European culture paused for a few decades, then resumed strongly in the fifteenth century. That part of the story forms our next chapter.

13.4 Problems and Questions

13.4.1 Problems in Medieval Mathematics

Exercise 13.1 What happens to the discrepancy between the two rules for area of an equilateral triangle discussed by Gerbert, if the side s gets very large? Does the relative discrepancy increase or decrease? (One formula is the correct formula $A = s^2\sqrt{3}/2$; the other is the incorrect triangular number formula $A = s(s+1)/2$.) Compare the discrepancy with the true value, take the limit as $s \to \infty$, and express the result as a percent.

Exercise 13.2 The *Practica geometriae* tells how to find the height of a distant object even if there is an obstacle between the observer and the object: After obtaining the ratio of height to distance as already shown, move back to a second position and repeat this operation. Then measure the distance between the two points of observation, and compute the distance from the object to the first point of observation by comparison of these measurements. (In our terms, given that $h/d_1 = a_1$ and $h/d_2 = a_2$, where a_1 and a_2 are known, if we also know $d = d_2 - d_1$, we can find d_1, and hence h. Carry out this computation for the example given in the *Practica geometriae*, where $h/d_1 = \frac{1}{3}$, $h/d_2 = \frac{1}{4}$, to find the ratio of h to d_1. (The author does not give the value of $d_2 - d_1$.) What similarities and differences do you notice in comparison with the method of surveying used in India and China?

Exercise 13.3 Readers with a strong stomach may wish to solve the following problem from Leonardo of Pisa: *A lion can eat one sheep in 4 hours; a leopard requires 5 hours; and a bear requires 6 hours. If a single sheep were given to all three, how long would it take them to devour it?* [The correct answer is $1\frac{23}{37}$ hours.]

Exercise 13.4 Leonardo's solution to the problem of finding a second pair of squares having a given sum is explained in general terms, then illustrated with a special case. He considers the case $4^2 + 5^2 = 41$. He first finds two numbers

(namely 3 and 4) for which the sum of the squares *is* a square. He then forms the product of 41 and the sum of the squares of the latter pair, obtaining $25 \cdot 41 = 1025$. Then he finds two squares whose sum equals this number, namely 31 and 8 or 32 and 1. He thus obtains the results $\left(\frac{31}{5}\right)^2 + \left(\frac{8}{5}\right)^2 = 41$ and $\left(\frac{32}{5}\right)^2 + \left(\frac{1}{5}\right)^2 = 41$. Find another pair of rational numbers whose sum is 41 following this method. Why does this method work?

Exercise 13.5 If the general term of the Fibonacci sequence is a_n, show that $a_n < a_{n+1} < 2a_n$, so that the ratio a_{n+1}/a_n always lies between 1 and 2. Assuming that this ratio has a limit, what is that limit?

Exercise 13.6 Let the Fibonacci sequence $\{a_n\}_{n=0}^{\infty}$ be given by $a_0 = 0$, $a_1 = 1$, $a_2 = 1$, $a_3 = 2$, $a_4 = 3$, etc., as in the text above, and define b_n for $n = 1, 2, \ldots$, by

$$b_n = \left(\frac{1+\sqrt{5}}{2}\right)^n + \left(\frac{1-\sqrt{5}}{2}\right)^n.$$

Prove that $b_n = a_{n-1} + a_{n+1}$, for $n = 1, 2, \ldots$.

Exercise 13.7 Consider Problem 27 of Book I of *De numeris datis*: *Two numbers are given whose sum is 10. If one is divided by 4 and the other by 2, the product of the quotients is 2. What are the two numbers?* Solve this problem in your own way, then solve it following Jordanus' recipe, which we paraphrase as follows. Let the two numbers be x and y, and let the quotients be e and f when x and y are divided by c and d respectively; let the product of the quotients be $ef = b$. Let $bc = h$, which is the same as fce or fx. Then multiply d by h to produce j, which is the same as xdf or xy. Since we now know both $x + y$ and xy, we can find x and y. [Jordanus used letters preceded and followed by a period for his variables, such as $.a.$ and $.b.$.]

Exercise 13.8 From Jordanus' rule of the inclined plane, suppose given a triangular frame with a level base of any size, and two other sides of length 35 and 64. If a weight of 80 kilograms lies on the side of length 35, how much weight lying on the side of length 64 will be required to keep the weight from sliding up or down if the two weights are joined by a rope passing over the vertex opposite the base? (Neglect friction.)

Exercise 13.9 Suppose $\angle C$ in Fig. 13.2 is a right angle. What does the law of the inclined plane become in this case, stated in terms of the angle A?

Exercise 13.10 State the general law of the inclined plane in terms of angles A and C in Fig. 13.2.

Exercise 13.11 It is an observational fact that a body heavy enough so that air resistance can be neglected in its fall will undergo nearly constant acceleration during free fall. In fact, its velocity will increase by 9.8 meters per second every second. Using the Merton rule, how far will such a body fall in 8 seconds?

13.4.2 Questions about Medieval Mathematics

Exercise 13.12 Dante's final stanza (quoted above) uses the problem of squaring the circle to express the sense of an intellect overwhelmed, which was inspired by his vision of heaven. What resolution does he find for the inability of his mind to grasp the vision rationally? Would such an attitude, if widely shared, affect mathematical and scientific activity in a society?

Exercise 13.13 What is the significance of ruling a board into 27 columns to make an abacus, as Gerbert is said to have done? Does it indicate that there was no symbol for zero?

Exercise 13.14 One frequently repeated story about Christopher Columbus is that he proved to a doubting public that the earth was round. What grounds are there for believing that "the public" doubted this fact? Which people in the Middle Ages would have been likely to believe in a flat earth? Consider also the frequently repeated story that people used to believe the stars were near the earth. Is this view of Medieval people plausible in the light of the *Practica geometriae*?

Exercise 13.15 The use of copious symbols as in Exercise 13.7 is typical of both Leonardo of Pisa and Jordanus Nemorarius. If you compared your own solution of the problem with that of Jordanus, you must have found that his solution is horribly cumbersome. What was "missing" from his algebra that makes the problem so much easier to solve nowadays? [*Hint:* The unknowns in the original problem are not really referred to as "two numbers whose sum is 10," but as "two parts of 10 that are to be found." What is the psychological and notational difference between these two ways of describing the numbers?]

13.5 Endnotes

1. The discussion of the study of science in Medieval universities is based partly on the book by David C. Lindberg, *The Beginnings of Western Science* (University of Chicago Press, 1992).

2. The quotation from Dante's *Divine Comedy* is from the Harvard Classic Edition (Collier, New York, 1909).

3. Richer's comments on Gerbert are quoted by Harriet Pratt Lattin in *The Letters of Gerbert* (Columbia University Press, New York, 1961), p. 46.

4. Gerbert's remark on Boethius' formula for the area of a triangle can be found in *The Letters of Gerbert*, cited above.

5. The section on the *Practica geometriae* is based on the translation and annotation of this work by Frederick A. Homann, S. J. (Marquette University Press, Milwaukee, 1991).

6. The problems quoted from the *Liber abaci* are taken from the source book by John Fauvel and J. J. Gray, *The History of Mathematics: A Reader* (Macmillan Press, New York, 1987), pp. 241–243.

7. The discussion of the *Liber quadratorum* is based on the recent annotated translation by L. E. Sigler, Academic Press, New York, 1987.

8. The discussion of Bradwardine's use of proportion is based on *De proportionibus proportionum* by Oresme, translated by Edward Grant (University of Wisconsin Press, Madison, 1966). The quotation from Bradwardine occurs in a footnote on p. 18.

9. The discussion of Oresme's *Questions on Euclid's Elements* is based on *Nicole Oresme and the Medieval Geometry of Qualities and Motions*, edited with translation and commentary by Marshall Clagett (University of Wisconsin Press, 1968), pp. 527–545.

Chapter 14

The Renaissance

The term *Renaissance* is cultural rather than chronological. The Renaissance began in Italy in the fifteenth century and spread northward over the next few centuries. The advance of science and scientific method, accompanied by the fragmentation of the Christian Church, led to a complete change in the world-view of educated people by the year 1700. This 300-year period saw an astonishing growth in science, paralleled by a rapid growth in mathematics. Although many details must necessarily be left out, we shall sample as much of this exciting period as space permits. There are three main themes that we shall be following: (1) the continued development of algebra, through the solution of the cubic and quartic equations; (2) new ways of computing products, made necessary by the high precision of the trigonometric tables used in astronomy; and (3) the beginning of projective geometry.

14.1 Algebra and Trigonometry

14.1.1 Regiomontanus

The work of translating the Greek and Arabic mathematical works took several centuries to complete. One of the last to work on this project was Johann Müller of Königsberg (1436–1476), better known by his Latin name of Regiomontanus, a translation of Königsberg (King's Mountain). Although he died young, Regiomontanus made valuable contributions to astronomy, mathematics, and the construction of scientific measuring instruments. He studied in Leipzig while a teenager, then spent a decade in Vienna and the decade following in Italy and Hungary. The last 5 years of his life were spent in Nürnberg. He is said to have died of an epidemic while in Rome as a consultant to the Pope on the reform of the calendar.

Regiomontanus checked the data in copies of Ptolemy's *Almagest* and made new observations with his own instruments. He laid down a challenge to astronomy, remarking that further improvement in theoretical astronomy, especially the theory of planetary motion, would require more accurate measuring instruments.

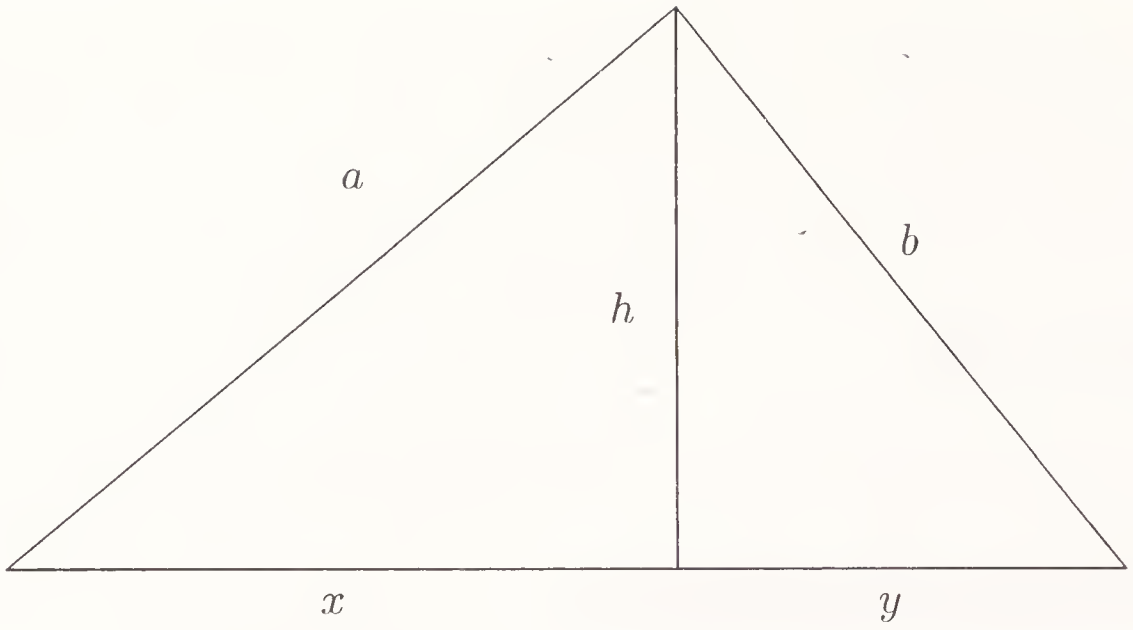


Figure 14.1: Triangle problem from Regiomontanus.

He established his own printing press in Nürnberg, so that he could publish his works. These works included several treatises on pure mathematics. He established trigonometry as an independent branch of mathematics rather than a tool in astronomy. In doing so, having considerable contact with right triangles and the Pythagorean theorem, he was frequently called upon to solve quadratic equations. For example, he considered the problem of solving a triangle given one side, the altitude to that side, and the ratio of the other two sides. Suppose the unknown sides are a and $b = ra$, where r is the ratio of the sides; suppose also that the known side is c, and it is divided into parts x and y by the known altitude h. Then we have the following relations (see Fig. 14.1):

$$\begin{aligned} a^2 - x^2 &= h^2, \\ b^2 - y^2 &= h^2, \\ ra - b &= 0, \\ x + y &= c, \end{aligned}$$

which leads to the general biquadratic equation for a,

$$\frac{(1-r^2)^2}{4c^2}a^4 - \left(\frac{1+r^2}{2}\right)a^2 + \frac{1}{4}c^2 + h^2 = 0.$$

Regiomontanus solved such equations rhetorically, as Al-Khwarizmi had done.

The main results we now know as plane and spherical trigonometry are in his book *De Triangulis Omnimodis*, although not exactly in the language we now use. For example (Book II, Theorem 1): *In every rectilinear triangle the ratio of one side to another side* [*equals*] *that of the right sine of the angle opposite one of the sides to the right sine of the angle opposite the other side.* We know this fact as the law of sines for plane triangles. His proof, however, is different from ours, because, while we think of the sine as the ratio of the opposite side to the hypotenuse in a right triangle, Regiomontanus thought of it as half the chord of twice the arc. In particular the value of the sine depended on the size of the circle,

and Regiomontanus was careful to say that sines could be compared only in circles of the same size. One important difference between Regiomontanus' trigonometry and ours is his exclusive use of sines rather than cosines and tangents. In fact it is rather intriguing that *De Triangulis*, which was begun in 1462, does not mention the tangent function, since it is known that Regiomontanus was using the tangent only 2 years later. Of course any one trigonometric function will suffice for solving triangles.

Regiomontanus states all of his theorems in words, never once writing out anything that looks like an equation. His proofs seem to indicate that he had a sense of humor. For instance, Theorem 6 of Book II states that if the three angles of a triangle are known, the ratios of its sides can be found. As proof of this fact Regiomontanus says, "This theorem presents no difficulty unless Theorem 1 above was carelessly passed over. . . ."

Although he never used the cosine directly, Regiomontanus used an equivalent function called the *versed sine* (which is the outermost of the two portions of a radius cut off by a sine and can be thought of as 1 minus the cosine). Using this function he was the first to state the law of cosines for spherical triangles. In a spherical triangle, that is, the figure formed by three great circles on a sphere, both angles and sides are measured as arcs (since the sides are arcs). The law is stated by Regiomontanus as follows (Book V, Theorem 2):

> In every spherical triangle that is constructed from the arcs of great circles, the ratio of the versed sine of any angle to the difference of two versed sines, of which one is the versed sine of the side subtending this angle while the other is the versed sine of the difference of the two arcs including this angle, is as the ratio of the square of the whole right sine [that is, the square of the radius] to the rectangular product of the sines of the arcs placed around the mentioned angle.

If the triangle has sides a, b, and c, and the angle opposite side a is α, this fact can be expressed as the trigonometric equation

$$\frac{R(1-\cos\alpha)}{R(1-\cos a)-R(1-\cos(b-c))}=\frac{R^2}{R\sin b\cdot R\sin c},$$

which easily reduces to

$$\cos a=\cos b\cos c+\sin b\sin c\cos\alpha.$$

This is the spherical law of cosines as taught nowadays.

14.1.2 Chuquet

The French Bibliothèque Nationale is in possession of the original manuscript of a comprehensive mathematical treatise written at Lyons in 1484 by one Nicolas Chuquet. Little is known about the author, except that he describes himself as a Parisian and a man possessing the degree of Bachelor of Medicine. The treatise

consists of four parts: a treatise on arithmetic and algebra called *Triparty en la Science des Nombres*, a book of problems to illustrate and accompany the principles of the *Triparty*, a book on geometrical mensuration, and a book of commercial arithmetic. The last two are applications of the principles in the first book.

Algebra in the *Triparty*

There are several new things in the *Triparty*. One is a superscript notation similar to the modern notation for the powers of the unknown in an equation. The unknown itself is called the *premier* or "first." Algebra in general is called the *rigle des premiers* or "rule of firsts." Chuquet listed the first 20 powers of 2 and pointed out that when two such numbers are multiplied their indices are added. Thus he had a clear idea of the laws of integer exponents. A second innovation in the *Triparty* is the free use of negative numbers as coefficients, solutions, and exponents. Still another innovation is the use of some symbolic abbreviations. For example, the square root is denoted R^2 (R for the Latin *Radix*, or perhaps the French *Racine*). The equation we would write as $3x^2 + 12 = 9x$ was written $.3.^2\ \bar{p}.12.\ egaulx\ a\ .9.^1.$ Chuquet called this equation impossible, since its solution would involve taking the square root of -63.

Chuquet gave an interesting way of getting rational approximations to irrational square roots, which he called the *rule of intermediate numbers*. For example, knowing that the square root of 6 is between $\frac{7}{3}$ and $\frac{5}{2}$, he adds the numerators and denominators to obtain a number in between these two, that is, $\frac{12}{5}$. As this number is too small, he pairs it with $\frac{5}{2}$ again, getting $\frac{17}{7}$, which is still too small. The next approximation is $\frac{22}{9}$, which is just slightly too small. Then $\frac{27}{11}$ is a bit too large, so that the next step is $\frac{27+22}{11+9}$, that is, $\frac{49}{20}$. Chuquet carries on with this process until he reaches the approximation $\frac{485}{198} = 2.4494949\ldots$, whereas the actual root is $2.4494897\ldots$.

Chuquet's approach to algebra and its application can be gathered from one of the illustrative problems in the second part (Problem 35). This problem tells of a merchant who buys 15 pieces of cloth, spending a total of 160 ecus. Some of the pieces cost 11 ecus each, and the others 13 ecus. How many were bought at each price?

If x is the number bought at 11 ecus apiece, this problem leads to the equation $11x + 13(15 - x) = 160$. Since the solution is $x = 17\frac{1}{2}$, this means the merchant bought $-2\frac{1}{2}$ pieces at 13 ecus. How does one set about buying a negative number of pieces of cloth? Chuquet said that these $2\frac{1}{2}$ pieces were bought on credit!

Luca Pacioli

The progress of algebra was not steadily upward. Written at almost the same time as Chuquet's *Triparty* was a work called the *Summa de Arithmetica, Geometrica, Proportioni et Proportionalita* by Luca Pacioli (or Paciuolo) (1445–1517). Since Chuquet's work was not printed until the nineteenth century, Pacioli's work is believed to be the first printed work on algebra. In comparison with the *Triparty*,

however, the *Summa* seems less original. The steps that Chuquet had taken toward an efficient way of writing a polynomial in an unknown were lacking in the *Summa*, except for the use of p for plus and m for minus. Other than that, Pacioli has only a few abbreviations, such as co for *cosa*, meaning *thing* (the unknown), ce for *censo* (the square of the unknown), and æ for *æquitur* (equals). Despite its inferiority to the *Triparty*, the *Summa* was much the more influential of the two books, because it was printed. It is referred to by the Italian algebraists of the early sixteenth century as a basic source.

14.1.3 Solution of Cubic and Quartic Equations

In Europe algebra was confined to linear and quadratic equations for many centuries, whereas the Chinese and Japanese had not hesitated to attack equations of any degree. The difference in the two approaches is a result of different ideas of what constitutes a solution. This distinction is easy to make nowadays: the European mathematicians were seeking an exact solution using only arithmetic operations and root extractions, what is called *solution by radicals*.

Our last visit with cubic equations (except for one equation considered by Leonardo of Pisa) was the discussion of the geometric solution by Omar Khayyam. At that time we remarked that, although the solution is graphically correct, being presented as the intersection of a circle and an hyperbola, any attempt at an algebraic solution of the corresponding set of two simultaneous equations describing the two curves merely leads back to the original cubic equation. This fact was fully appreciated by mathematicians at the time, and the algebraic solution of the cubic was regarded as impossible or at least very difficult. The Italian algebraists of the early sixteenth century brought a change in this way of thinking.

Scipione del Ferro (1465–1525)

The credit for the discovery of a method of solving (certain) cubic equations belongs rightly to a Professor (Lector) at the University of Bologna, Scipione del Ferro, who discovered, around the year 1500, how to solve equations of the type "cube plus things equal number," what we would phrase as $x^3 + px = q$, where p and q are positive numbers. He communicated this discovery under an oath of secrecy to his son-in-law A. Nave and to another mathematician named Antonio Maria Fior. Fior used this knowledge to build his own academic reputation by challenging others to contests, which of course he would win because of the method he learned from del Ferro.

Niccolò Tartaglia (1500–1557)

Fior overreached himself in 1535, when he challenged Niccolò Fontana of Brescia, known as Tartaglia (the Stammerer) because a wound he received as a child when the French overran Brescia in 1512 left him with a speech impediment. Fior challenged Tartaglia to solve a set of thirty problems, among which were finding a

number which yields 6 when its cube root is added to it, and finding where to cut a tree 12 *braccia* high in such a way that the part left standing will be the cube root of the part cut off at the top. What Fior had not counted on is that much of the difficulty of a mathematical problem lies in not knowing *whether* a solution exists. Once it is known that a problem is solvable, it often happens that many people are able to discover independent proofs of it. Tartaglia discovered how to solve the equation and so won the contest.

Gerolamo Cardano (1501–1576)

One of the many eccentric creative geniuses of the time in Italy was a young man whose abilities had brought him the office of Rector of the University of Padua at the age of 25. This man, Gerolamo Cardano (Fig. 14.2), was writing a book on mathematics in 1535 when he heard of Tartaglia's victory over Fior. He naturally wished to include the secret of the cubic in his book, and he wrote asking permission, which Tartaglia at first refused, hoping to work out all the details of all cases of the cubic and write a treatise himself. Algebra had taken a step backward from the time of Chuquet, in that all terms had to be positive. There were therefore a total of thirteen possible types of cubic equations, each of which required its own method of solution. In 1539 Tartaglia, according to his own account, confided the secret of one kind of cubic to Cardano after Cardano swore a solemn oath never to publish them without permission and gave Tartaglia a letter of introduction to the Marchese of Vigevano. Tartaglia revealed a rhyme by which he had memorized the procedure.

The verses Tartaglia had memorized say, in modern language, that to solve the problem $x^3 + px = q$, one should look for two numbers u and v satisfying $u - v = q$, $uv = (p/3)^3$. The problem of finding u and v is that of finding two numbers given their difference and their product, and of course, that is merely a matter of solving a *quadratic* equation, a problem that had already been completely solved. Once this quadratic has been solved, the solution of the original cubic is $x = \sqrt[3]{u} - \sqrt[3]{v}$.

To see how this method works in a particular example, consider the equation $x^3 + 132x = 1267$. Following Tartaglia's method, we need to find numbers u and v such that $u - v = 1267$ and $uv = \left(\frac{132}{3}\right)^3 = (44)^3 = 85,184$. Following the venerable procedure for finding two such numbers, we recall that $u + v = \sqrt{(u-v)^2 + 4uv} = \sqrt{(1267)^2 + 4 \cdot (18,184)} = \sqrt{1,605,289 + 340,736} = \sqrt{1,946,025} = 1395$. Now that we have both $u - v$ and $u + v$, it is easy to see that $u = \frac{1267+1395}{2} = 1331$ and $v = \frac{1395-1267}{2} = 64$. The solution is therefore $x = \sqrt[3]{1331} - \sqrt[3]{64} = 11 - 4 = 7$. This answer can then easily be checked.

Tartaglia did not claim to have given Cardano any proof that this procedure works. It was left to Cardano himself to find the demonstration. Cardano kept his promise not to publish this result until 1545. However, as Tartaglia delayed his own publication, and in the meantime Cardano had discovered the solution of other cases of the cubic himself and had also heard that del Ferro had priority anyway, he published the result in his *Ars Magna* (Great Art), giving full credit

Figure 14.2: Gerolamo Cardano, from his *Arithmetic*. Stock Montage, Inc.

to Tartaglia. Tartaglia was furious, and started a bitter controversy over Cardano's alleged breach of faith.

Cardano's *Ars Magna* contains a very thorough discussion of the thirteen kinds of cubic equations. Some of them yield very easily to a recipe like the one just given. Others cause trouble in some cases. For example, consider the case $x^3 = px + q$. Cardano says (in words, not letters) that if $(p/3)^3 \leq (q/2)^2$, then the solution can be given by following a recipe (which we write anachronistically as a formula)

$$x = \sqrt[3]{\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3}} + \sqrt[3]{\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 - \left(\frac{p}{3}\right)^3}}. \qquad (14.1)$$

The condition $\left(\frac{p}{3}\right)^3 \leq \left(\frac{q}{2}\right)^2$ amounts to a *diorismos* for this problem. When this condition is not met, Cardano said,

> the solution of this problem can be found by the Aliza rule which is discussed in the book of geometrical problems.

[The word *Aliza* is mysterious. Prof. J. F. Porto da Silveira has said (in an e-mail message to the author) that the Italian historian of mathematics Gino Loria (1862–1954) claimed the word was an Arabic word meaning *difficult*.]

Chapter XXV gives some rules for dealing with this "irreducible" case of the cubic, as it later came to be called. The rules in this chapter are, as Cardano said, not "general" since they do not work for all possible data. He considers the examples of the equations $x^3 = 16x + 21$ and $x^3 = 4x + 15$. His technique is to add or subtract a constant on both sides so that a common factor can be divided out. In the first case, for example, adding 27 to both sides gives $x^3 + 27 = 16x + 48$, so that $(x + 3)(x^2 - 3x + 9) = 16(x + 3)$. Thus the equation becomes the quadratic $x^2 = 3x + 7$, which everyone knew how to solve. In the second case, subtracting 27 leads to $x^3 - 27 = 4x - 12$, so that $x^2 + 3x + 9 = 4$, which to Cardano was absurd. Cardano had clearly wanted to solve all cubic equations, and he did not give up easily in the face of such problems. Indeed, he was the first to consider seriously the possibility of roots for quadratic equations $ax^2 + bx + c = 0$ when the discriminant $b^2 - 4ac$ is negative. In the *Ars Magna* he considers the problem of finding two numbers whose sum is 10 and whose product is 40. This is a quadratic equation problem that does not satisfy the required *diorismos* that the discriminant be nonnegative. Following the usual recipe given that $x + y = 10$, and $xy = 40$, to find $x - y$, one must write $x - y = \sqrt{(x + y)^2 - 4xy} = \sqrt{-60} = 2\sqrt{-15}$, and so Cardano is willing to speculate that the two numbers are $5 + \sqrt{-15}$ and $5 - \sqrt{-15}$. After exploring the geometric ramifications of this idea, he concludes that, "this final point is... as subtle as it is useless."

As Cardano points out elsewhere, the technique for deciding what to add in order to solve the equation $x^3 = px + q$ is equivalent to solving the equation $x^3 + q = px$. Now this may actually be easier psychologically, in that one may be able to guess a solution to the second equation but not to the first. As a general method of solving this case of the cubic, however, it is circular.

Ludovico Ferrari (1522–1565)

Cardano's student Ludovico Ferrari worked with him in the solution of the cubic, and between them they had soon found a way of solving certain fourth-degree equations. Thus armed, they did not shrink from combat with Tartaglia. Once again Tartaglia found himself challenged with a set of 30 problems, among which were some algebraic problems of a new kind, such as Number 17, *Divide* 8 *into two parts such that the product times the difference of the parts shall be as large as possible, proving everything*, and Number 21, *Find six quantities in geometric proportion such that the double of the second plus the triple of the third equals the square root of the sixth.* Others were philosophical, such as Number 22, which asks for an exposition of a passage in Plato's *Timaeus*, and Number 30: *Is unity a number?*.

Tartaglia replied that in problem Number 17 the required two parts were $4 + \sqrt{5\frac{1}{3}}$ and $4 - \sqrt{5\frac{1}{3}}$ and that the solution to Number 21 was the series with ratio $\sqrt[3]{47 + \sqrt{12}} + \sqrt[3]{47 - \sqrt{12}}$, saying that there was no point in writing out all six terms, since no skill was required to do that. He objected to the philosophical problem Number 22 on the ground that it was not a mathematical question.

Ferrari's riposte was scornful. He pointed out that Tartaglia had neglected the two most important words in Problem 17: *proving everything*.

Ferrari's solution of the quartic was included near the end of Cardano's *Ars Magna*. Counting cases as for the cubic, one finds a total of 20 possibilities. The principle in most cases is the same, however. The idea is to make a perfect square in x^2 equal to a perfect square in x by adding the same expression to both sides. For example, Cardano gives the example

$$60x = x^4 + 6x^2 + 36.$$

It is necessary to add to both sides an expression $rx^2 + s$ to make both sides squares, that is, so that both sides of

$$rx^2 + 60x + s = x^4 + (6 + r)x^2 + (36 + s)$$

are perfect squares. Now the condition for this to happen is well known: $ax^2 + bx + c$ is a perfect square if and only if $b^2 - 4ac = 0$. Hence we need to have simultaneously

$$3600 - 4sr = 0, \quad (6 + r)^2 - 4(36 + s) = 0.$$

Solving the first of these equations for s in terms of r, substituting in the second, and clearing the denominator leads to the equation

$$r^3 + 12r^2 = 108r + 3600.$$

This is a cubic equation called the *resolvent* cubic. Once it is solved, the original quartic breaks into two quadratic equations upon taking square roots and adding an ambiguous sign.

Significance of the Solution of the Cubic

The mathematicians involved in the solution of the general cubic equation had reason to be proud of themselves. To find a numerical process that solves a cubic equation exactly was genuinely new, something that had escaped both the ancient Greeks and the Islamic scholars. This work raised several important issues that should be mentioned.

1. The problem is not a practical one. We have already seen that even solving quadratic equations is of little practical use except in astronomy.

2. The Cardano recipe for solving an equation sometimes gives the solution in a rather strange form. For example, Cardano says that the solution of $x^3 + 6x = 20$ is $\sqrt[3]{\sqrt{108} + 10} - \sqrt[3]{\sqrt{108} - 10}$. This is correct, but would you know that this number is actually 2?

3. The procedure does not always work. For example, the equation $x^3 + 6 = 7x$ has to be solved by guessing a number that can be added to both sides so as to produce a common factor that can be canceled out. The number in this case is 21, but there is no *algorithm* for finding such a number.

4. For equations of the type $x^3 + 6 = 7x$ the algebraic procedures for finding x involve square roots of negative numbers. The search for an algebraic procedure using only real numbers to solve this case of the cubic continued for some three hundred years, until finally it was shown that no such procedure can exist.

5. It was a significant fact that knowledge of algebra increased two steps at a time. After the earliest days when linear and quadratic equations could be solved the next leap is the one we have just seen, where cubic and quartic equations can be solved. Likewise, it was a significant fact proved by Pappus that the three- and four-line loci were all conic sections, and, as Omar Khayyam had shown, conic sections suffice to solve cubic equations (and by implication quartic equations also, although Omar Khayyam did not know this). The parallel here could not fail to impress a well-read mathematician, and we shall see that Descartes noticed this fact.

6. The solution of cubic and quartic equations was a good piece of mathematics in that it settled an interesting open question and raised others of equal interest, while pointing out a possible method of attack on the new questions. The most natural of these is: How does one solve the fifth-degree equation? Two and a half centuries were to pass before this question was answered partially, and a full three centuries before an actual solution of the fifth-degree equation was found.

Rafael Bombelli (1526–1572)

In comparison with the preceding centuries the level of mathematical activity in Italy during the first half of the sixteenth century was astonishing. In addition to those already mentioned we must also mention an engineer in the service an Italian nobleman. This engineer, Rafael Bombelli, is the author of a treatise on algebra which appeared in 1572 (it was written about 1560). In the introduction to this treatise we find the first mention of Diophantus in the modern era. Bombelli says that, while all authorities are agreed that the Arabs invented algebra, he, having been shown the work of Diophantus, credits the actual invention to the latter. Bombelli attacked the irreducible case of the cubic, which as we have seen, leads to the cube root of a complex number. Since imaginary numbers had been rejected in connection with quadratic equations, and the modern symbolism had not yet been invented, Bombelli was forced to build from the ground up. He invented the name "plus of minus" to denote a square root of -1 and "minus of minus" for its negative. He did not think of these two concepts as different numbers, but rather as the *same* number being added in the first case and subtracted in the second. What is most important is that he realized what rules must apply to them in computation: plus of minus times plus of minus makes minus and minus of minus times minus of minus makes minus, while plus of minus times minus of minus makes plus. Such were the first attempts to make sense of these numbers. Bombelli had no systematic way of taking the cube root of a complex number. In considering the equation $x^3 = 15x + 4$, he found by applying the formula that $x = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}$. In this case, however, Bombelli was able to work backward, since he knew in advance that one root is 4; the problem was to make the formula *say* "4." Bombelli had the idea that the two cube roots must consist of real numbers together with his "plus of minus" or "minus of minus." Since the imaginary parts in the sum of the two cube roots must cancel out and the real parts must add up to 4, it seems obvious that the real parts of the cube roots must be 2. In our terms, the cube roots must be $2 \pm x\sqrt{-1}$ for some x. Then since the cube of the cube roots must be $2 \pm 11\sqrt{-1}$ (what Bombelli called 2 plus 11 times "plus of minus"), it is clear that the cube roots must be 2 plus "plus of minus" and 2 minus "plus of minus," that is, $2 \pm \sqrt{-1}$. As a way of solving the equation, this is circular, but it does allow the formula to make sense even in the irreducible case.

Notation

All original algebra treatises written up to and including the treatise of Bombelli are very tiresome for the modern student, who is familiar with symbolic notation. For that reason we have allowed ourselves the convenience of modern notation when doing so will not distort the thought process involved. In the years between 1575 and 1650 several innovations in notation were introduced that make treatises written since that time appear essentially modern. The symbols $+$ and $-$ were originally used in bookkeeping in warehouses to indicate excess and deficiencies; they first appeared in a German treatise on commercial arithmetic in 1489, but

were not widely used in the rest of Europe for another century. The sign for equality was introduced by a Welsh medical doctor, physician to the short-lived Edward VI, named Robert Recorde (1510–1558). His symbol was a very long pair of parallel lines, because, as he said, "noe 2. thynges, can be moare equalle." The use of abbreviations for the various powers of the unknown in an equation was eventually recovered from Diophantus, but there was a further step that needed to be taken before algebra became a mathematical subject on a par with geometry. Our discussion of sixteenth-century algebra will conclude with that step.

François Viète (1540–1603)

François Viète, a lawyer who worked as tutor in a wealthy family and later became an advisor to Henri de Navarre (the future king Henri IV), found time to study Diophantus and to introduce his own ideas into algebra. Viète can be credited with several crucial advances in the subject. In his book *Artis Analyticae Praxis* (The Practice of the Analytic Art) he begins by giving the rules for powers of binomials (in words). For example, he describes the fifth power of a binomial as, "the fifth power of the first [term], plus the product of the fourth power of the first and five times the second,... ."

As this quotation shows, Viète appears to be following the tedious route of writing everything out in words. However, the introduction is followed by five books of "zetetica" [from the Greek word *zetein* ($\zeta\eta\tau\acute{\epsilon}\iota\nu$), meaning *seek*]. The mention of "roots" in connection with the binomial expansions was not accidental. Viète studied the relation between roots and coefficients in general equations, though he was somewhat handicapped in this enterprise, since he did not recognize the negative and imaginary roots. His approach was to see how to find the roots given various information about them, for example (Zetetic XV of the second book), he says that, given the product of the roots and the difference between their cubes, the roots will be found.

The five books of zetetics are followed by "Treatises on the Understanding and Amendment of Equations." These treatises, despite such awkward neologisms as "zetetics," "plasmatic modification," and "syncrisis," contain several very important advances in algebra. The first is a general discussion of the structure of equations. By using vowels to represent unknowns and consonants to represent data for a problem, Viète finally achieved what was lacking in earlier treatises: a convenient way of talking about general data without having to give specific examples. His consonants could be thought of as representing numbers that would be known in any particular application of a process, but were left unspecified for purposes of describing the process itself. His first example was the equation $A^2 + AB = Z^2$, in other words, a standard quadratic equation. According to Viète these three letters are associated with three numbers in direct proportion, Z being the middle, B the difference between the extremes, and A the smallest number. In our terms, this says that $Z = Ar$ and $B = Ar^2 - A$. Thus the general problem reduces to finding the smallest of three numbers A, Ar, Ar^2 given the middle value and the difference of the largest and smallest. Viète had already shown how to do that in his books of zetetics.

This kind of analysis showed Viète the true relation between the coefficients and the roots. For example, he knew that in the equation $x^3 - 6x^2 + 11x = 6$, the sum and product of the roots must be 6 and the sum of the products taken two at a time must be 11. This observation still did not enable him to solve the general cubic equation, but he did study the problem geometrically and show that any cubic could be solved provided one could solve two of the classical problems of antiquity: constructing two mean proportionals between two given lines and trisecting any angle. As he concluded at the end of his geometric chapter, "It is very worthwhile to note this."

14.2 Prosthapheresis and Logarithms

The increased accuracy of astronomical instruments, among other applications, led to a need to multiply numbers having a large number of digits. Now it is well-known that the amount of labor involved in multiplying two numbers increases as the product of the number of digits, while the labor of adding increases according to the number of digits in the smaller number. Thus, multiplying two 15-digit numbers requires over 200 one-digit multiplications, while adding the two numbers requires only 15 such operations (not including carrying). Obviously multiplication is going to be more error-prone as well. Hence astronomical measurements and the solution of triangles with high precision could be greatly facilitated if the operation of multiplication could be simplified. Two methods of achieving this simplification were invented in the late sixteenth century, and we shall now examine them.

14.2.1 Prosthapheresis

Like a steam-driven sawmill that feeds its engines with its own wood shavings, trigonometry provided the first method of simplifying its own computations. The key turned out to be in the tables of sines and cosines that were causing the problem in the first place. The process was called *prosthapheresis*, from two Greek words meaning addition and subtraction. There are hints of this process in several sixteenth-century works, but we shall quote just one example. In his *Trigonometria*, first published in Heidelberg in 1595, the theologian and mathematician Bartholomeus Pitiscus (1561–1613), posed the following problem: *To solve the proportion in which the first term is the radius, while the second and third terms are sines, avoiding multiplication and division.* The problem here is to find the fourth proportional x, satisfying $r : a = b : x$, where r is the radius of the circle, and a and b are two sines (half-chords) in the circle. We can see immediately that $x = ab/r$, but, as Pitiscus says, the idea is to avoid the multiplication and division, since in the trigonometric tables of the time a and b might easily have seven or eight digits each.

The key to prosthapheresis is the well-known formula

$$\sin\alpha\cos\beta = \frac{\sin(\alpha+\beta)+\sin(\alpha-\beta)}{2}.$$

This formula is applied as follows: If you have to multiply two large numbers, regard one of them as the sine of an angle, the other as the cosine of a second angle. (Since Pitiscus had only tables of sines, he had to use the complement of the angle having the second number as a sine.) Add the angles and take the sine of their sum to obtain the first term; then subtract the angles and take the sine of their difference to obtain a second term. Finally divide the sum of the two terms by 2 to obtain the desired product. To take a very simple example, suppose we wish to multiply 155 by 36. A table of trigonometric functions shows that $\sin 8^\circ\, 55' = 0.15500$ and $\cos 68^\circ\, 54' = 0.36000$. Hence

$$36 \times 155 = 10^5 \frac{\sin 77^\circ\, 49' + \sin(-59^\circ\, 59')}{2} = \frac{97748 - 86588}{2} = 5580.$$

In general some significant figures will be lost in this kind of multiplication. For large numbers this procedure saves labor, since multiplying even two 7-digit numbers would tax the patience of most people nowadays. A further advantage is that prosthapheresis is less error-prone than multiplication. Its advantages were known to the Danish astronomer Tycho Brahe (1546–1601), who used it in the astronomical computations connected with the extremely precise observations he made at his observatory during the latter part of the sixteenth century.

14.2.2 Logarithms

The problem of simplifying laborious multiplications, divisions, root extractions, etc., was being attacked at the same time in another part of the world and from another point of view. The connection between geometric and arithmetic proportion had been noticed earlier by Chuquet, but the practical application of this fact had never been worked out. The Scottish laird John Napier, Baron of Murchiston (1550–1617) tried to clarify this connection and apply it. His work consisted of two parts, a theoretical part based on a continuous geometric model, and a computational part, involving a discrete (tabular) approximation of the continuous model. The computational part was published in 1614. However, Napier hesitated to publish his explanation of the theoretical foundation. Only in 1619, 2 years after his death, did his son publish an English translation of Napier's theoretical work under the title *Mirifici logarithmorum canonis descriptio* (A Description of the Marvelous Rule of Logarithms). The word *logarithm* means *ratio number*, and it was from the concept of ratios (geometric progressions) that Napier proceeded.

The Theoretical Model

In order to explain his ideas Napier resorted to the concept of moving points. He imagined one point P moving along a straight line from a point T toward a point S with decreasing velocity such that the ratio of the distances from the point P to S at two different times depends only on the difference in the times. (Actually he called the line ending at S a sine and imagined it shrinking from its initial size TS, which he called the radius.) A second point is imagined as moving along a

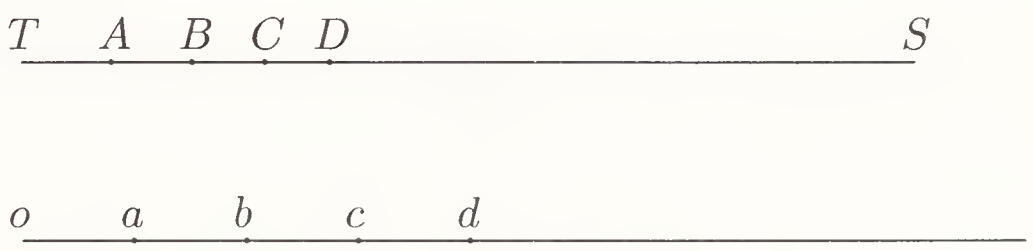


Figure 14.3: Geometric basis of logarithms

second line at a constant velocity equal to that with which the first point began. These two motions can be clarified by considering Fig. 14.3.

The first point sets out from T at the same time and with the same speed with which the second point sets out from o. The first point, however slows down, while the second point continues to move at constant speed. The figure shows the locations reached at various times by the two points: When the first point is at A, the second is at a, when the first point is at B, the second is at b, etc. The point moving with decreasing velocity requires a certain amount of time to move from T to A, then the same amount of time to move from A to B, from B to C, from C to D, etc., and $TS : AS = AS : BS = BS : CS = CS : DS$, etc. In the same amount of time required for this first point to move from T to A the second point moves from o to a, from a to b, etc.

The first point will never reach S, since it keeps slowing down, and its velocity at S would be zero. The second point will travel indefinitely far, given enough time. Because the points are in correspondence, the division relation that exists between two positions in the first case is mirrored by a subtractive relation in the corresponding positions in the second case. Thus this diagram essentially changes division into subtraction, and of course multiplication into addition. The top scale in Fig. 14.3 resembles a slide rule, and this resemblance is not accidental: a slide rule is merely an analog computer that incorporates a table of logarithms.

Napier's definition of the logarithm can be stated in the modern notation of functions by writing $\log(AS) = oa$, $\log(BS) = ob$, etc., in other words, the logarithm increases as the "sine" decreases. These considerations contain the essential idea of logarithms. The quantity Napier defined is not the logarithm as we know it today. If points T, A, and P correspond to points o, a, and p, then

$$\overline{op} = \overline{oa} \log_k \left(\frac{\overline{PS}}{\overline{TS}} \right),$$

where $k = \dfrac{\overline{AS}}{\overline{TS}}$.

Computational Considerations

The geometric model just discussed is theoretically perfect, but of course one cannot put the points on a line into a table of numbers. It is necessary to construct the table from a finite set of points; and these points, when converted into numbers, must be rounded off. Napier was very careful to analyze the maximum errors that

could arise in constructing such a table. Referring again to Fig. 14.3, he showed that oa, which is the logarithm of AS, satisfies

$$TA < oa < TA\Big(1 + \frac{TA}{AS}\Big).$$

(These inequalities are simple to prove, since the point describing oa has a velocity larger than the velocity of the point describing TA but less than TS/AS times the velocity of that point.) Thus the tabular value for the logarithm of AS can be taken as the average of the two extremes, that is, $TA\big[1 + (TA/2AS)\big]$, and the relative error will be very small when TA is small.

Napier's death at the age of 67 prevented him from making some improvements in his system, which are sketched in an appendix to his treatise. These improvements consist of scaling in such a way that the logarithm of 1 is 0 and the logarithm of 10 is 1, which is the basic idea of what we now call common logarithms. These further improvements to the theory of logarithms were made by Professor Henry Briggs (1561–1630), who was in contact with Napier for the last two years of Napier's life and wrote a commentary on the appendix to Napier's treatise. As a consequence, logarithms to base 10 came to be known as Brigssian logarithms.

14.3 Projective Geometry

In art the fifteenth century was a period of great innovation in which a large number of beautiful paintings were produced. In an effort to give the illusion of depth in two-dimensional representations some artists looked at geometry from a new point of view, studying the projection of two- and three-dimensional shapes in two dimensions to see what properties were preserved and how others were changed. A description of such a procedure (based partly on the work of his predecessors) was given by Leon Battista Alberti (1404–1472) in a treatise entitled *Della pictura*, published posthumously in 1511.

The essence of the idea is that if the eye is thought of as being at the same height as a point P (Fig. 14.4) above a horizontal plane, parallel horizontal lines in that plane receding from the imagined point where the eye is located can be drawn as rays emanating from P, giving the illusion that P is infinitely distant. The application to art is obvious: Since the canvas can be thought of as a window through which the scene is viewed, if you want to draw parallel horizontal lines as they would appear through a window, you must draw them as if they all converged on the point P (the vanishing point). Thus a family of lines having one thing in common (passing through P) projects to a family having a different common property (parallelism). It is clear that lines remain lines under such a projection. However, perpendicular lines will not remain perpendicular, nor will circles remain circles. The later discovery of projective invariants built these rudimentary ideas into a useful and beautiful mathematical structure.

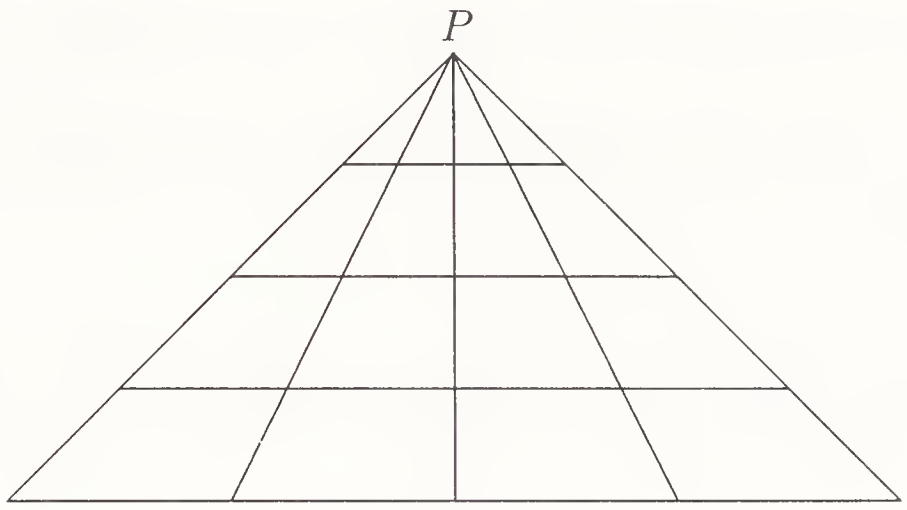


Figure 14.4: Projection from a point P.

14.4 Problems and Questions

14.4.1 Problems in Renaissance Mathematics

Exercise 14.1 Solve the triangle problem quoted from Regiomontanus with data $h = 125$, $c = 250$, $r = .8165$.

Exercise 14.2 The triangle construction problem cited above in connection with Regiomontanus shows how to find the other two sides of a triangle given its base and altitude and the ratio of the two sides. Are there any restrictions on the data of this problem, or can it be solved given any three values of base, altitude, and side ratio? [Recall the Greek notion of *diorismos*, a discussion of the data allowable in a problem. What is the *diorismos* for this problem?]

Exercise 14.3 Suppose you wish to build two ramps leaning against each other, and having their other ends 48 feet apart, with height 20 feet, in such a way that a weight of 100 kilograms on one ramp will exactly balance a weight of 200 kilograms on the other when the weights are connected by a rope passing over the point where the ramps meet. How long should the ramps be? [Remember the law of the inclined plane from the last chapter.]

Exercise 14.4 Use the spherical law of cosines to compute the number of degrees in a great circle from New York to Paris, given the following geographic information. New York lies at 41° N, 74° W and Paris lies at 49° N, 2° E. How far is it from New York to Paris, given that one degree of a great circle is about 69 miles? [Let side a of the triangle be the great circle joining Paris and New York. Let sides b and c be the lines of longitude joining these two cities to the North Pole (90° N).]

Exercise 14.5 Use Chuquet's method to find an approximation to $\sqrt{5}$ given that $\frac{20}{9} < \sqrt{5} < \frac{23}{10}$.

Exercise 14.6 Solve the equation $x^3 + 60x = 992$ using the recipe given by Tartaglia.

Exercise 14.7 How can you *prove* that $\sqrt[3]{\sqrt{108}+10}-\sqrt[3]{\sqrt{108}-10}=2$?

Exercise 14.8 Was Tartaglia correct in his solution of the problem of finding two numbers whose sum is 8 such that the product of the numbers multiplied by their difference is maximal?

Exercise 14.9 Show that solving the equation $x^3+q=px$ makes it possible to find a number $r^3=pr-q$ that can be added to both sides of $x^3=px+q$, leading to the equation $x^3+r^3=px+q+r^3$, which has $x+r$ as a factor on both sides.

Exercise 14.10 This exercise and the six following are intended to clarify certain facts that Cardano, Tartaglia, and the others saw only dimly. This extra insight is provided by modern algebraic notation.

The general cubic equation

$$Ax^3+Bx^2+Cx+D=0, \quad A\neq 0,$$

is equivalent to a monic equation

$$x^3+ax^2+bx+c=0,$$

with $a=B/A$, $b=C/A$, $c=D/A$. The substitution $x=y-(a/3)$, then reduces the problem of solving the original equation to the simpler problem of finding y such that

$$y^3+py+q=0, \tag{14.2}$$

where $p=b-(a^2/3)$ and $q=c-(ab/3)+(2a^2/27)$. Considering the identity

$$(u-v)^3+3uv(u-v)+(v^3-u^3)=0,$$

we see that $y=u-v$ will be a solution of Eq. 14.2 provided u and v can be chosen so that

$$3uv=p, \quad v^3-u^3=q.$$

In terms of the new variables $z=v^3$ and $w=u^3$, we thus need only find z and w, given that $z-w=q$ and $zw=p^3/27$. Hence solving the general cubic equation requires four operations: (1) dividing by the leading coefficient; (2) substituting $x=y-(a/3)$ and rewriting the equation in terms of y as $y^3+py+q=0$; (3) solving the (quadratic) equation $z-w=q$, $zw=p^3/27$; (4) taking the cube roots of z and w and setting $x=\sqrt[3]{w}-\sqrt[3]{z}-(a/3)$.

Follow this procedure to solve the equation

$$1000x^3-6000x^2+16,950x-19,944=0.$$

Exercise 14.11 Show that the procedure of the preceding exercise leads to the general formula

$$y=\sqrt[3]{-\frac{q}{2}+\sqrt{\frac{q^2}{4}+\frac{p^3}{27}}}+\sqrt[3]{-\frac{q}{2}-\sqrt{\frac{q^2}{4}+\frac{p^3}{27}}}$$

for a solution of $y^3 + py + q = 0$. Apply this formula to the following equations: $y^3 + 84y - 279 = 0$, $y^3 + 5y - 42 = 0$, $y^3 - 7y - 6 = 0$. Notice that the root produced is the same (3) for the first two cases; why then is the answer given in such a simple form for the first and such a complicated form for the second? Why does the formula fail for the third case? Find the solution(s) of the third equation by guessing.

Exercise 14.12 Plot the graph of $z = y^3 + py + q$ for various values of p and q. Notice that if $p \geq 0$, this function is always increasing, hence has exactly one real zero, while if $p < 0$, it has a maximum at $y = -(-p/3)^{1/2}$ and a minimum at $y = +(-p/3)^{1/2}$. If the value of $y^3 + py + q$ is negative at the first of these values (case 1) or positive at the second value (case 2), then the equation still has only one real root. In the first case this condition amounts to $\left[-(-p/3)^{1/2} \right]^3 - p(-p/3)^{1/2} + q < 0$. Since $-p$ is positive, we can write $-p = \sqrt{p^2}$ and rewrite this equation as $-\frac{1}{3}(-p^3/3)^{1/2} + (-p^3/3)^{1/2} + q < 0$ or $\frac{q}{2} < -\frac{1}{3}(-p^3/3)^{1/2}$. Note that q must be negative if this condition is to hold, since the right-hand side of this last inequality is negative. For the second case we have similarly $\frac{q}{2} > \frac{1}{3}(-p^3/3)^{1/2} > 0$. The two cases can then be combined. The equation has precisely one real root if and only if $(q^2/4) + (p^3/27) > 0$. (Note that this condition automatically holds if $p > 0$, and it holds when $p = 0$ if and only if $q \neq 0$.) Compare this result with the cases in the preceding problem. The expression $(q^2/4)+(p^3/27)$ is called the *discriminant* of the cubic $y^3+py+q = 0$. In terms of the discriminant, when does the formula for solving the cubic work, and when does it break down?

Exercise 14.13 In the preceding exercise we found a necessary and sufficient condition for the cubic equation $y^3 + py + q = 0$ with real numbers p and q to have exactly one real root. A cubic equation can also have exactly two real roots; in that case one of the roots will be a double root. Show that in this case the discriminant equals zero, but the cubic formula continues to produce a solution of the equation. Does the formula "pick out" the single root or the double root?

Exercise 14.14 Conclude from the preceding exercises that the cubic formula for equations with real coefficients breaks down if and only if the discriminant is negative, and this happens if and only if there are three distinct real roots. (If we think of the cubic formula as a genie that answers our request for a solution of the equation, the genie doesn't have any way of choosing one root rather than another, so it has a nervous breakdown.)

Exercise 14.15 We have just seen that the formula for solving a cubic equation with real coefficients and three real roots leads to the problem of finding the cube root of a complex number. Can this operation be reduced to algebraic operations that involve only real numbers? Consider, for example, the analogous problem of finding a complex number $x + iy$ whose square is $a + ib$, where x, y, a, b are all real. We need to solve the two equations

$$x^2 - y^2 = a, \quad 2xy = b.$$

If $b = 0$, this is merely the problem of finding the square root of a real number, which we already know how to do (but the answer will be an imaginary number if $a < 0$). Hence assume $b \neq 0$, so that $x \neq 0$ and $y \neq 0$ also. Solving the second equation for y in terms of x, and substituting into the first equation leads to the biquadratic equation $x^4 - ax^2 - \frac{1}{4}b^2 = 0$, and hence we find that with a suitable choice of sign

$$x = \sqrt{\frac{\sqrt{a^2 + b^2} + a}{2}}, \quad y = \pm\sqrt{\frac{\sqrt{a^2 + b^2} - a}{2}}.$$

Thus the problem of taking the square root of a complex number reduces to taking square roots of nonnegative real numbers.

Try the same procedure with cube roots, that is, try to solve simultaneously

$$x^3 - 3xy^2 = a, \quad 3x^2y - y^3 = b.$$

What equation results when you eliminate one of the variables between these two equations? Remembering that the cubic formula requires you to solve these equations simultaneously in order to find the solution of a cubic equation having three real roots, can it truly be said that the cubic formula solves the problem in this case? Is it not rather a circular process?

Exercise 14.16 If you know the polar form of complex numbers $z = r\cos\theta + ir\sin\theta$, show that the problem of taking the cube root of a complex number is equivalent to simultaneously solving two of the classical problems of antiquity, namely the problem of two mean proportionals, and the problem of trisecting the angle. (Recall that Viète had mentioned this fact.)

Exercise 14.17 Consider Viète's problem of finding three numbers in direct proportion given the middle number and the difference between the largest and smallest. Show that this problem amounts to finding x and y given $\sqrt{xy}$ and $y - x$. How do you solve such a problem?

Exercise 14.18 Multiply 78,642 by 9753 using a five-place table of sines (use only the sine column of your trigonometric table). [There are two ways to proceed, since you can regard the number 9753 as either 0.97530 or as 0.09753. Do the problem both ways. Remember, you need to find angles α and β such that $\sin\alpha = 0.78642$ and $\sin(90° - \beta) = 0.97530$. You then take the average of $\sin(\alpha + \beta)$ and $\sin(\alpha - \beta)$.] Check your work with a calculator or by hand computation.

Exercise 14.19 How can prosthapheresis be used to avoid division?

14.4.2 Questions about Renaissance Mathematics

Exercise 14.20 Why did Chuquet choose to interpret the negative amount of cloth bought as if it had been bought on credit? Would it not have been more logical to interpret it as a certain amount of cloth *sold*? Could the use of negative numbers be made consistent with Chuquet's interpretation?

Exercise 14.21 List all 13 possible cubic equations, given that the coefficients must be positive, but the terms can be on either side of the equation.

Exercise 14.22 If you were a teacher making up problems for your pupils to practice solving cubic equations, how would you construct examples for which the cube roots "come out even" as in the first example of Exercise 14.11, avoiding the messy case in the second example? [*Hint:* Look at the identity in u and v on which the solution is based.]

Exercise 14.23 Summarize in your own words the meaning of the solution of the cubic equation. In what sense is the problem solved? What operations must one be able to perform in order to use the method? What restrictions on data are there?

Exercise 14.24 Although complex numbers are now taught to high-school students in connection with the solution of quadratic equations, mathematicians were able to ignore them in that context at first. Why was this possible? Why did the solution of cubic equations force mathematicians to deal with square roots of negative numbers when quadratic equations had not done so?

Exercise 14.25 Why is notation such an important component of mathematics? Is it true that "the medium *is* the message," as the Canadian scholar Marshall McLuhan (1911–1980) was often quoted as saying? Does the style in which an idea is expressed *change* the idea? Consider this question in relation to the problem of solving an equation as stated nowadays and as stated in the sixteenth century.

Exercise 14.26 As we saw, by the late sixteenth century two methods were available for simplifying laborious multiplications and divisions by changing them into addition and subtraction. The first was prosthapheresis, based on the properties of the trigonometric functions. The second was logarithms, created on the basis of the theory of proportion, but, as we now know, essentially based on the laws of exponents. Are these two methods really different? If not, what connection is there between them?

14.5 Endnotes

1. The quotations from Regiomontanus are taken from the translation of *De Triangulis Omnimodis* by Barnabas Hughes (University of Wisconsin Press, Madison, 1967).

2. The section on Chuquet is based largely on *Mathematics from Manuscript to Print, 1300–1600*, edited by Cynthia Hay (Oxford University Press 1988), especially the article by G. Flegg, "Nicolas Chuquet–an introduction," pp. 59–72.

3. The discussion of Cardano's *Ars Magna* is based on the translation by T. Richard Witmer (MIT Press, 1968). The quotation on the Aliza problem is on p. 103.

4. The discussion of the dispute between Tartaglia and Ferrari is based on the corresponding readings in *The History of Mathematics: A Reader* by J. Fauvel and J. J. Gray (Macmillan, New York, 1987).

5. The discussion of the work of Viète is based on the English translation *The Analytic Art* (Kent State University Press, 1983). The quotation on the binomial expansion of the fifth power is on p. 41 of that book.

6. The discussions of prosthapheresis and logarithms are based on the selections in *A Source Book in Mathematics* by David Eugene Smith (Dover Reprint, New York, 1959).

Chapter 15

The Calculus

The watershed in the history of mathematics is the invention of the calculus. It synthesized nearly all the algebra and geometry that had come before and generated problems that led to most of the mathematics studied today. Although calculus is an amalgam of algebra and geometry, it soon developed results that were indispensible in other areas of mathematics. Even theories whose origins seem to be independent of all forms of geometry—combinatorics, for example—turn out to involve concepts such as generating functions, for which the calculus is essential.

Elements of the calculus had existed from the earliest times in the form of infinitesimal methods in geometry, and such techniques were refined in the early seventeenth century. In this way the raw materials for the calculus were available by the middle of the seventeenth century; the invention of the calculus was more like focusing a camera than painting on a blank canvas.

15.1 Analytic Geometry

The idea of representing numbers by lines is a very old one, occurring even in Euclid's books on number theory. The principle of using a line to represent a *variable* number, which associates geometry with algebra, can be seen in Apollonius' *Conics*. It was the basis of Omar Khayyam's solution of the cubic equation, and is explicit in the writings of Nicole of Oresme. As often happens when an idea gradually becomes recognized, the idea for the final step occurred nearly simultaneously to two people.

15.1.1 Pierre de Fermat

The works of Diophantus and Pappus were among the favorite reading of a lawyer at Toulouse named Pierre de Fermat (1601–1665). Despite his busy life of public service, Fermat found time to study the works of these two authors and reflect on them. What he made of Diophantus will be discussed in the next chapter. Just now it is the influence of Pappus that is important, particularly the things Pappus wrote

about loci. From Pappus' description of the treatise of Apollonius on loci, Fermat attempted to reconstruct the results that this treatise must have contained. By the year 1630 Fermat had discovered most of the principles of analytic geometry. The important innovation, which Pappus had not known, was the concept of an equation. Where Pappus had spoken of loci, which are verbal descriptions of conditions that a point must satisfy, Fermat thought of equations, which most of the verbal descriptions really are. At the beginning of his book *Ad locos planos et solidos isagoge* (*Introduction to Plane and Solid Loci*) he describes the situation as follows:

> Whenever two unknown magnitudes appear in a final equation we have a locus, the extremity of one of the unknown magnitudes describing a straight line or a curve...
>
> It is desirable, in order to aid the concept of an equation, to let the two unknown magnitudes form an angle, which usually we would suppose to be a right angle...

We see here some of the basic results still taught in analytic geometry today. Fermat followed the notation of Viète, using vowels to denote variables and consonants to denote constants. He describes the general equation of a line as $da = be$, where a and e stand for what we would call x and y. Fermat was careful to observe the required physical dimensions in his equations by ensuring that every term contained the same number of letters. For the square of a quantity Fermat used the Roman numeral II to indicate the exponent. Thus he described the equation of the hyperbola as $ae = z^{\mathrm{II}}$. After showing how to obtain equations for the parabola and circle, Fermat says that he has been able in this way to reconstruct all of the propositions of the second book of Apollonius' *On Plane Loci*. He concludes by stating a generalization of the two-line locus problem: *Given the position of any number of lines; if from some definite point lines be drawn forming given angles with the given lines, and the sum of the squares of all the segments is equal to a given area, the point will describe a solid locus* [*conic section*] *of determined position.*

Fermat did not publish his discoveries on analytic geometry during his lifetime, though they were circulated among scholars in manuscript form and finally published in 1679.

15.1.2 René Descartes

The person who is popularly credited with being the discoverer of analytic geometry was the philosopher René Descartes (1596–1650), one of the most influential thinkers of the modern era. He was educated in the Jesuit school at La Flèche and at the university at Poitiers, where he studied law. Having obtained his law degree, he "drifted" for some time, serving in the army, traveling and studying. He was past forty when he wrote his philosophical treatise *Discours de la méthode*, to which *La géométrie* was an appendix. However, many of the ideas contained in it had been written down as early as 1620, when he seems to have had a mystical flash

of insight, which he wrote down in a Latin work entitled *Rules for the Guidance of Thought*.

It is in the appendix to his *Discours*, however, that the fundamental ideas of analytic geometry appear in detail. In the opening words of *La géométrie* we find the point of view from which Descartes regarded his work. He saw the difference between geometry and arithmetic not only as the contrast of the continuous and the discrete, but also in the more methodical principles of arithmetic as compared with geometry. Arithmetic consisted of just five operations, and all else depended on these basic concepts, whereas geometry had no such neat structure. Descartes intended to provide one. As he says,

> Any problem in geometry can easily be reduced to such terms that a knowledge of the lengths of certain straight lines is sufficient for its construction. . . in geometry, to find required lines it is merely necessary to add or subtract other lines; or else, taking one line which I shall call unity in order to relate it as closely as possible to numbers, and which can in general be chosen arbitrarily, and having given two other lines, to find a fourth line, which shall be to one of the given lines as the other is to unity (which is the same as multiplication); or, again, to find a fourth line which is to one of the given lines as unity is to the other (which is equivalent to division); or, finally, to find one, two, or several mean proportionals between unity and some other line (which is the same as extracting the square root, cube root, etc., of the given line). And I shall not hesitate to introduce these arithmetical terms into geometry, for the sake of greater clearness.

Here Descartes takes an important step that Fermat did not take, by introducing a unit of length. As he says in the passage just quoted, this step makes it possible to represent the product of two lines as a *line* rather than a rectangle. That approach freed him from the necessity of having the same number of factors in all terms in an equation. After showing the simple geometric constructions for product, quotient, and square roots, he notes that

> unity can always be understood, even when there are too many or too few dimensions; thus, if it be required to extract the cube root of $a^2b^2 - b$, we must consider the quantity a^2b^2 divided once by unity and the quantity b multiplied twice by unity.

By this apparently simple step Descartes had used algebra to introduce arithmetic into geometry. Nowadays mathematicians would say that he had shown how to make directed line segments into a field. To illustrate these ideas Descartes showed how to solve the quadratic equation $z^2 = az + b^2$.

In contrast to the notation of all of his predecessors, Descartes' notation looks extremely modern. His convention that letters at the beginning of the alphabet stand for data and letters at the end stand for variables or unknowns was adopted as the standard and has remained down to the present with only a few improvements.

Like Fermat, Descartes used his analytic geometry to attack the problems of Pappus, in particular the three- and four-line locus problems, for which he found

the general equation to be $y^2 = ay - bxy + cy - dx^2$. This problem was the showpiece of *La géométrie*. He quoted Pappus at length and explained the kinds of curves that can be expected with different numbers of lines. He gave many examples showing how a motion described geometrically can be translated into equations and how equations can be analyzed to describe the resulting locus. How deeply he had penetrated into algebra is clear from his remarks on equations. He says explicitly that the best way to consider an equation is to set all the terms on one side and zero on the other and that the number of distinct roots equals the degree of the equation. This assertion implies that he was willing to consider negative and imaginary roots, and he does say that some of the roots may be "false," that is, less than nothing. He notes that a polynomial is divisible by $x - a$ if and only if a is a zero of the polynomial, and he gives what is known as Descartes' rule of signs: *An equation can have as many true* [*positive*] *roots as it contains changes of sign, from* $+$ *to* $-$ *or from* $-$ *to* $+$*; and as many false* [*negative*] *roots as the number of times two* $+$ *signs or two* $-$ *signs are found in succession.*

15.2 The Calculus

We have seen certain prefigurations of the calculus in the work of Archimedes, in the method used by Zu Chongzhi and Zu Geng to find the volume of a sphere, in the recursive methods of approximating certain geometric quantities used by the Japanese mathematicians, the infinite series expansions of the Hindus, and other places. We now wish to discuss how such diverse techniques coalesced into a unified and powerful method of solving geometric problems. Let us consider three aspects of the calculus: differentiation, integration, and infinite series. All three are taught nowadays along with certain elementary applications that tend to conceal the unity of calculus. Differentiation has an elementary application in the problem of finding maxima and minima; integration has the application of finding area and volumes; and infinite series can be used to compute the values of exponentials, logarithms, and trigonometric functions.

The unity of calculus arises from certain physical problems involving the study of changes in quantities over time. The main use of differentiation is to describe such phenomena as differential equations. Integration then becomes the method of solving these equations. Infinite series enter the picture since integration is often not possible in terms of elementary functions. (Most of the important equations of mathematical physics cannot be solved by direct integration; in such cases the equation itself is sometimes used to generate a solution in the form of an infinite series.)

Calculus was not invented all at one time. Instead the application of algebra to certain geometric problems, and the study of new curves in geometry gradually led to a number of techniques and results that began to present a pattern. A surprisingly simple and crucial step was the replacement of the subtangent (defined below) by the notion of relative rate of change (what we now call *slope*). Although the two concepts are closely related and each can be defined in terms of the other, problems can be solved much more quickly when analyzed in terms of slopes or

relative rates of change than by use of the subtangent. Here we can see a principle that will appear many times in the history of mathematics—two concepts that are *logically* equivalent in the context of a theory may be *psychologically* very different. In some cases priority disputes arise when one mathematician can claim with perfect truthfulness to have stated a principle before another (that is, to have stated something logically equivalent to it).

Some mathematicians, primarily Isaac Newton (1642–1727) in England and Gottfried Wilhelm Leibniz (1646–1716) on the Continent saw this pattern and its ramifications more clearly than others, and so became generally known as the creators of the subject. The subject was not perfected by Newton and Leibniz, however, and the full understanding of what could and could not be done with the techniques of calculus came only in the generation or two following them. We shall divide the history of the subject into three stages: (1) a period when certain geometric problems involving the tangent to various curves and the area bounded by curves were attacked by use of algebra and a hazily stated idea of passage to the limit; (2) the systematization of these isolated techniques into a unified set of algorithms—the invention of the calculus proper; (3) the consolidation of the new invention and its application to a wide variety of problems in physics and astronomy.

As heirs of the ancient Greek mathematicians, modern mathematicians could not be content for long with mere intuitive ideas. The results produced by the calculus were so spectacular that no one was prepared to abandon them, yet it was realized that the foundations of the calculus were not as secure as those of traditional geometry. As a result the search for a rigorous foundation of the calculus began as soon as the subject was systematically organized, and this search was not complete until the midnineteenth century.

15.2.1 Tangents

The main problem in finding a tangent to a curve at a given point is to find some second condition that this line must satisfy so as to determine it uniquely. Given that the line passes through the point in question, it suffices to know either a second point that it must pass through or the angle that it must make with a second line. The way in which algebra can assist in finding this condition, especially for algebraic curves, can be illustrated with several examples.

Descartes

Obviously one can find the tangent to a curve at a given point if and only if one can find the normal (perpendicular to the tangent) at that point; hence it is a matter of indifference which of these things one chooses to do. In *La géométrie* Descartes proposed the following method for finding the normal to a curve. Given the curve CE referred to an axis GA and the point C where the normal is required, consider all the circles passing through C with center P on GA. By solving the equation of a typical circle simultaneously with the equation of the curve, one finds in general

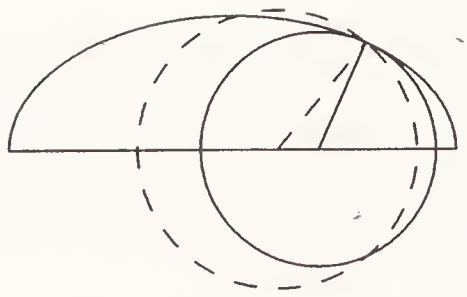


Figure 15.1: Descartes' construction of the normal to a curve.

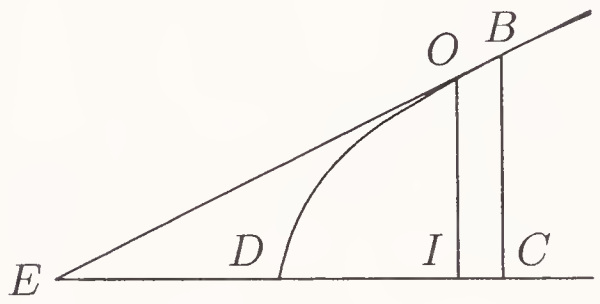


Figure 15.2: Fermat's method of finding the subtangent.

several points of intersection. If the point P is chosen so that there is only one point of intersection, then the circle will be tangent to the curve, and therefore its radius will be the required normal (see Fig. 15.1).

Descartes illustrated this method with a specific example, taking the curve CE to be an ellipse and MA the portion of the major axis from one vertex. He wrote the equation of the ellipse by translating Apollonius' definition into symbolic form: $x^2 = ry - (r/q)y^2$. Here r is the latus rectum of the ellipse and q its diameter; x is the ordinate and y the abscissa, measured from the end of the axis.

Fermat

Fermat had attacked the problem of finding maxima and minima of variables even before the publication of Descartes' *Géométrie*. As his works were not published during his lifetime, but only circulated among those who were in a rather select group of correspondents, his work in this area was not recognized for some time. His method is very close to what is still taught in calculus books today. The difference is that, where we now use the derivative to find the slope of the tangent line, that is, the tangent of the angle it makes with a reference axis, Fermat looked for the point where the tangent intercepted that axis. If the two lines did not intersect, obviously the tangent was easily determined as the unique parallel through the given point to the given axis. In all other cases Fermat needed to determine the length of the projection of the tangent on the axis from the point of intersection to the point below the point of tangency, a length known as the *subtangent*. In a letter sent to Marin Mersenne (1588–1648) and forwarded to Descartes in 1638 Fermat explained his method of finding the subtangent.

In Fig. 15.2 the curve DB is a parabola with axis CE, and the tangent at B meets the axis at E. Since the parabola is convex, a point O between B and E on the tangent lies outside the parabola, and since the abscissas measured along the axis are proportional to the squares of the ordinates measured perpendicular to the axis, it follows that $\overline{CD} : \overline{DI} > \overline{BC}^2 : \overline{OI}^2$. (Equality would hold here if $\overline{OI}$ were replaced by the portion of it cut off by the parabola.) Since $\overline{BC} : \overline{OI} = \overline{CE} : \overline{EI}$, it follows that $\overline{CD} : \overline{DI} > \overline{CE}^2 : \overline{EI}^2$. Then abbreviating by setting $\overline{CD} = g$, $\overline{CE} = x$, and $\overline{CI} = y$, we have $g : g - y > x^2 : x^2 + y^2 - 2xy$, and cross-multiplying,

$$gx^2 + gy^2 - 2gxy > gx^2 - x^2y.$$

Canceling the term gx^2, and dividing by y, we obtain $gy - 2gx > -x^2$. Since this inequality must hold for all y (no matter how small), it follows that $x^2 \geq 2gx$, that is, $x \geq 2g$ if $x > 0$. Choosing a point O beyond B on the tangent and reasoning in the same way would give $x \leq 2g$, so that $x = 2g$. Since x was the quantity to be determined, the problem is solved. Actually we have slightly distorted Fermat's words here. He referred to a previous argument and simply said that $gy^2 + x^2y$ would become equal to $2gxy$. We know, of course that this equality really holds only when $y = 0$, and hence his next step, dividing by y, is not legitimate. However, he clearly had in mind the idea of a limit of positive quantities, rather than dividing by zero. The ideas were new and difficult to express clearly.

In this paper Fermat asserted, "And this method never fails. . . ." This assertion provoked an objection from Descartes, who challenged Fermat with the curve now known as the folium of Descartes, having equation $x^3 + y^3 = 3axy$.

The Cycloid

Since analytic geometry is an application of algebra to geometry, one would expect that the first curves studied would be algebraic curves; and indeed such is the case in the writings of Fermat and Descartes. In fact Descartes was rather disdainful of nonalgebraic curves such as the spiral and the quadratrix, saying that they are generated by two motions whose relationship to each other cannot be determined exactly, and therefore should be dismissed. One such curve, which had first been noticed in the early sixteenth century by an obscure mathematician named Charles Bouvelles (ca. 1470–ca. 1553), is the cycloid, the curve generated by a point on a circle (called the generating circle) that rolls without slipping along a straight line. This curve is easily pictured by imagining a painted spot on the rim of a wheel as the wheel rolls along the ground. Since the linear velocity of the rim relative to its center is exactly equal to the linear velocity of the center, it follows that the point is at any instant moving along the bisector of the angle formed by a horizontal line and the tangent to the generating circle. In this way, given the generating circle, it is an easy matter to construct the tangent to the cycloid. This result was obtained independently around 1638 by Descartes, Fermat, and Gilles Personne de Roberval (1602–1675), and slightly later by Evangelista Torricelli (1608–1647), a pupil of

Galileo Galilei (1564–1642). This approach represents yet a third (kinematic) way of constructing tangents, independent of the methods of Descartes and Fermat.

15.2.2 Lengths, Areas, and Volumes

Seventeenth-century mathematicians had inherited two conceptually different ways of applying infinitesimal ideas to find areas and volumes. One was to regard an area as a "sum of lines." The other was to approximate the area by a sum of regular figures and try to show that the approximation got better as the individual regular figures got smaller. The rigorous version of the latter argument—the method of exhaustion, was tedious and of limited application.

Cavalieri's Principle

In the "sum of lines" approach a figure whose area or volume was required was sliced into parallel sections, and these sections were shown to be equal to, or constant multiples of, corresponding sections of a second figure whose area or volume was known. The first figure was then asserted to be equal to (or a constant multiple of) the second. The principle was formally stated by Bonaventura Cavalieri (1598–1647), a Jesuit priest and a student of Galileo. At the time it was customary for professors to prove their worthiness for a chair of mathematics by a learned dissertation. As part of his application for a position at the University of Bologna in 1629, Cavalieri submitted a work with the title *Geometria indivisibilibus continuorum nova quadam ratione promota* (Geometry Advanced in a New Way by the Indivisible Parts of Continua). In this work, published in 1635, Cavalieri asserted that figures lying between two parallel lines and such that all sections parallel to those lines have the same length must have equal area.

This principle is now called Cavalieri's principle. The idea of regarding a two-dimensional figure as a sum of lines or a three-dimensional figure as a sum of plane figures was extended by Cavalieri to consideration of the squares on the lines in a plane figure, then to the cubes on the lines in a figure, etc. What Cavalieri has in mind, in the case of a sum of squares, is the volume of a figure whose cross-section at height h parallel to a given plane equals the square of the line cut off by the plane figure at that same height.

To illustrate these ideas in a simple case, consider the two triangles into which a diagonal divides a parallelogram. It is obvious that the two are congruent, and hence have equal area. Cavalieri shows this by pairing the section in one of them a given distance above the lower base with the section of the other the same distance below the upper base. Since the sections are the same, it follows that the sum of the lines in each triangle is half the sum of the lines in the parallelogram.

Passing to the squares of the lines inside these triangles is trickier. Referring to Fig. 15.3, we can see that $\overline{RT}^2 + \overline{TV}^2 = 2\overline{RS}^2 + 2\overline{TS}^2$, since $\overline{RS} = \frac{1}{2}(\overline{RT} + \overline{TV})$ and $\overline{TS} = \overline{RT} - \overline{RS} = \frac{1}{2}(\overline{RT} - \overline{TV})$. Hence, if $\square(\cdot)$ denotes the sum of the squares of the lines inside a given figure, then

$$\square(AEC) + \square(CEG) = 2\square(ABFE) + 2\square(MEF) + 2\square(MBC).$$

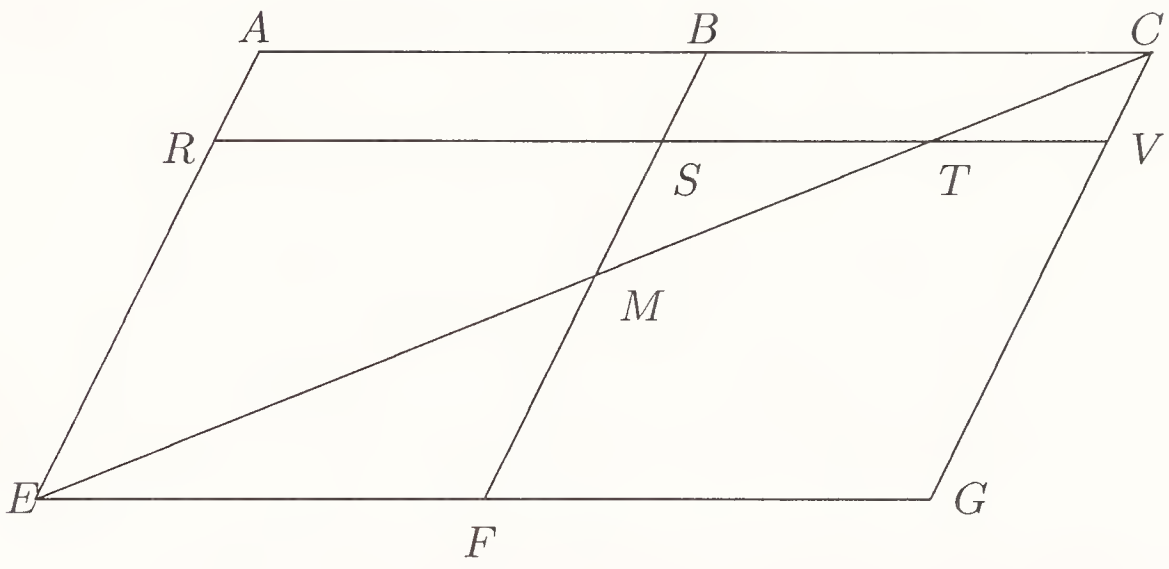


Figure 15.3: Cavalieri's principle.

Because of symmetry it is obvious that $\square(AEC) = \square(CEG)$, and $\square(MEF) = \square(MBC)$. Hence

$$\square(CEG) = \square(ABFE) + 2\square(MEF)$$

Now by pairing the line parallel to BF at distance h from E in the triangle MEF with the line at distance $2h$ from E in the triangle CEG, one finds the lines to be in the ratio of 2 to 1; hence their squares are in a ratio of 4 to 1, and since there are twice as many such lines in CEG as in MEF, it follows that $\square(CEG) = 8\square(MEF)$, and so $\frac{3}{4}\square(CEG) = \square(ABFE)$. Since each section of $ABFE$ is half of the same section of $ACGE$, it follows that $\square(ACGE) = 4\square(ABFE) = 3\square(CEG)$. That is, the sum of the squares of the lines in each of the two triangles is one-third of the sum of the squares of the lines in the whole parallelogram. The latter is a^2h, where a is the base and h the height of the parallelogram. Hence $\square(CEG) = \frac{1}{3}a^2h$. By continuing this process, Cavalieri eventually concluded that the sum of the nth powers of the lines in one of the triangles is $1/(n+1)$ times the sum of the nth powers of the lines in the parallelogram. When the parallelogram is a square, so that $h = a$, this result foreshadows the formula we know as an integral: $\int_0^a x^n\,dx = a^{n+1}/(n+1)$.

Area of the Cycloid

Cavalieri's principle was soon applied to find the area of an entirely new curve. The curve known as the cycloid was mentioned above in connection with tangents. Around 1630 Mersenne proposed using the cycloid as a test case for the new methods of indivisibles being used. This curve had been named by Galileo, who wrote to Cavalieri in 1640 that he had studied it 50 years earlier. On the basis of geometrical considerations he had conjectured that the area under one arch was three times the area of the generating circle, but experiments with physical models had convinced him that it was less than three times. He then suspected that the area was incommensurable with the area of the generating circle.

Galileo's intuition was better than he knew; in fact the area under one arch of a cycloid is exactly three times that of the generating circle, as was already

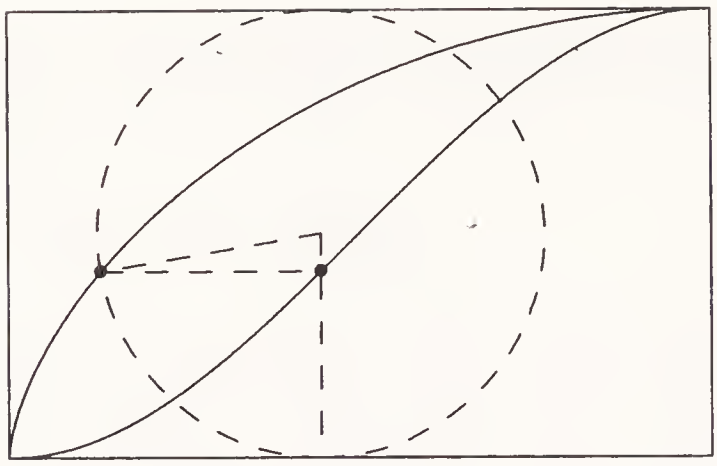


Figure 15.4: Roberval's quadrature of the cycloid.

known to Roberval by the time Galileo wrote to Cavalieri. Roberval, who found the tangent to the cycloid, also found the area beneath it by a clever use of the method of indivisibles. He considered along with half an arch of the cycloid itself a curve he called the *companion* to the cycloid. This curve is generated by a point that is always directly below or above the center of the generating circle as it rolls along and at the same height as the point on the rim that is generating the cycloid. As the circle makes half a revolution (see Fig. 15.4), the cycloid and its companion first diverge from the ground level, then meet again at the top. Symmetry considerations show that the area under the companion curve is exactly one-half of the rectangle whose vertical sides are the initial and final positions of the diameter of the generating circle through the point generating the cycloid. But by definition of the two curves their generating points are always at the same height, and the horizontal distance between them at any instant is the corresponding horizontal section of half of the generating circle. Hence by Cavalieri's principle the area between the two curves is exactly half the area of the circle. Now the rectangle has height equal to the diameter of the circle and length equal to half its circumference. Its area is therefore twice the area of the generating circle. Half of it (the area below the companion curve) is exactly equal to the area of the generating circle. Therefore the area under this half-arch of the cycloid is 1.5 times the area of the generating circle, and so the area under the full arch is three times the area of the generating circle.

Solids of Revolution

Cavalieri's method of indivisibles was intended to give exact results for areas and volumes. The intuitive idea of infinitesimals, however, is based on finite approximations. That point of view was adopted by Johannes Kepler (1571–1630). In 1615 he wrote a work entitled *Nova stereometria doliorum vinariorum* (A New Volume Measure for Wine Barrels), in which he studied the volumes of various solids of revolution. A fundamental preliminary needed in this context is the value of π. Kepler quoted Archimedes' value $\frac{22}{7}$, and in his proof of it he broke the circle into very small arcs, which he regarded as straight lines. This result is only approximate, as Kepler well knew, since no arc of a circle, no matter how short, is

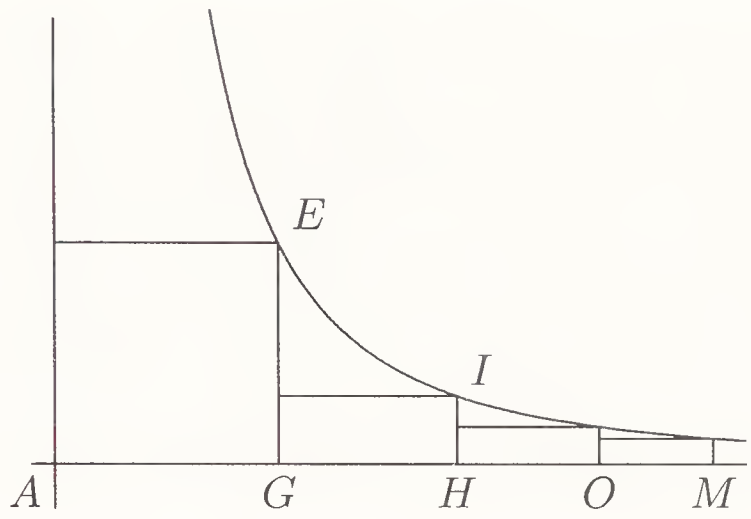


Figure 15.5: Fermat's quadrature of a generalized hyperbola.

really straight. Among the theorems he proved was Theorem 18, which is a special case of the theorem attributed to Pappus: *Any ring with circular or elliptic cross section is equal to a cylinder whose altitude equals the length of the circumference which the center of the rotated figure describes, and whose base is the same as the cross section of the ring.* Kepler's proof of this fact involved cutting the ring into an infinite number of very thin disks.

Rectangular Approximations and the Method of Exhaustion

Besides the method of indivisibles (Cavalieri's principle), mathematicians of the time also applied the method of polygonal approximation to find areas. In 1640 Fermat wrote a paper on quadratures in which he found the areas under certain figures by a method that he saw could easily be generalized. He considered a "general hyperbola," as in Fig. 15.5, a curve referred to asymptotes AR and AC and defined by the property that the ratio $AH^m : AG^m = EG^n : HI^n$ is the same for any two points E and I on the curve; we would describe this property by saying $x^m y^n = \text{const}$. The ordinary conic hyperbola is obtained when $m = n = 1$. Fermat showed how to find the area by constructing a sequence of ordinates and assuming that they are sufficiently close together that the rectangles such as $GE \times GH$ and the portion cut off under the curve $GHIE$ can be regarded as approximately equal "following the method of Archimedes." As Fermat stated, it sufficed to make this remark only once; there was no need to repeat it, "and insist constantly upon a device well known to mathematicians."

For the particular class of curves he was dealing with, Fermat found that it was useful to arrange the points G, H, O, M,... as a geometric sequence, the reason being that the ratio of successive ordinates would be equal to a ratio of powers of the abscissas. Since the abscissas are in geometric progression, the rectangles are also, and therefore it is very easy to find upper and lower bounds for them. In this way Fermat found that the area under the curve $x^2 y = \text{const}$ from EG to infinity equals the rectangle $BAGE$. He also pointed out that the problem can be solved similarly for all hyperbolas except the conical hyperbola $xy = \text{const}$ (for which the area is infinite).

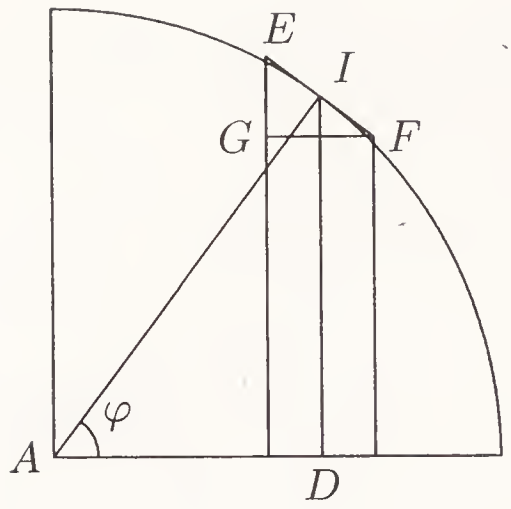


Figure 15.6: Sum of the sines in a quadrant.

Powers of Sines

Among the many geniuses of seventeenth-century France one of the most interesting is Blaise Pascal (1623–1662). Attracted by both the harmonies of the physical world and religious mysticism, he wrote brilliantly on both subjects. His *Pensées* have provided inspiration to countless readers, while his mathematical achievements have inspired only the few able to read them. We have seen above how Cavalieri found the "sums of the powers of the lines" inside a triangle. In 1659 Pascal did the same for the "sums of the powers of the lines inside a quadrant of a circle." Now a line inside a quadrant of a circle is called a sine, so that he found the sum of the powers of the sines of a quadrant of a circle. In modern terms, where Cavalieri found $\int_0^a x^n\,dx = a^{n+1}/(n+1)$, Pascal found $\int_\alpha^\beta (R\sin\varphi)\,Rd\varphi = R(R\cos\alpha - R\cos\beta)$.

Pascal's insight was that if any tangent is drawn to the circle, such as EF in Fig. 15.6, and two lines are drawn from its ends perpendicular to the base of the quadrant, then the small triangle EFG around the point of tangency is similar to the triangle AID, so that the rectangle $EF \times DI$ equals the rectangle $FG \times AI$. In the language of infinitesimals FG represents $dx = d(R\cos\varphi)$ and EF represents $ds = d(R\varphi)$, the arc length on the circle. Of course, $AI = R$ and $DI = R\sin\varphi$.

Since the tangent is very nearly equal to the arc, and for "infinitely thin" slices is exactly equal to it, the conclusion, in our terms, is that if the sines are multiplied by infinitely small sections of an arc and the products are added, the result is equal to a rectangle whose base is the projection of the arc parallel to the sines and whose height is the radius of the circle. In other words, since an infinitely small section of an arc represents what we think of as $R\,d\varphi$, where $d\varphi$ is an infinitely small central angle, $\displaystyle\int_\alpha^\beta R\sin\varphi\,Rd\varphi = R(R\cos\alpha - R\cos\beta)$.

This argument, like Kepler's, is subject to attack on the grounds that a small arc is set equal to a small piece of the tangent to it. Pascal recognized that this point might provoke an objection, and he did his best to answer it, saying that even though equality does not hold for a finite number of sines, it became true for an infinite number, "because then the sum of all the equal tangents differs... from the sum of all the equal arcs by less than any given quantity."

In other words, Pascal was appealing to the same reasoning as was used in the ancient method of exhaustion, where two numbers or ratios were shown to be equal because the assumption that they differ by any positive amount leads to a contradiction. The difficulty is that, even though the individual arcs are well approximated by the individual tangents, a great quantity of them are being added, so that the errors may possibly accumulate and not disappear in the limit at infinity. Archimedes and Euclid, when they used the method of exhaustion, were careful to give the details, and never spoke of the circle as being an infinite number of infinitely short lines (Aristotle had strictly warned against confusing these two qualitatively different things). Some clarification of Pascal's reasoning was sure to be demanded eventually.

Powers of x

A close approximation to the modern method of integrating to find the area under the curve $y = x^n$ was used by the English mathematician John Wallis (1616–1703). In 1655 he published his *Arithmetica infinitorum*, in which he found the sums of the first few powers of initial segments of the integers, formulas equivalent to

$$\frac{\sum_{k=1}^{n} k}{(n+1)n} = \frac{1}{2}; \quad \frac{\sum_{k=1}^{n} k^2}{(n+1)n^2} = \frac{1}{3} + \frac{1}{6n};$$

$$\frac{\sum_{k=1}^{n} k^3}{(n+1)n^3} = \frac{1}{4} + \frac{1}{4n},$$

and so on up to the sum of the sixth powers. By writing the sums this way he concluded that if the curve $y = x^r$ is approximated with rectangles, with the portion under the curve from $x = 0$ to $x = a$ being broken into n intervals of length a/n, the sum of the areas of the rectangles will tend to $a^{r+1}/(r+1)$. Wallis went further and made the same case for the use of fractional powers and even irrational powers r.

15.2.3 The Relation between Tangents and Areas

Except for the use of infinitely short lines in both problems, there seems to be no natural relation between the tangent problems and the area problems just discussed. Just how deep this relation lies can be seen by Archimedes' near brush with it, when he pointed out that the tangent to a spiral at the end of its first turn forms the hypotenuse of a right triangle equal in area to the circle through the point of tangency. It is clear from his silence that he does not suspect that the tangent and area problems are systematically related for all curves. The very deep relation between the two might eventually have been guessed by comparing formulas for area and the slope of the tangent; but that step was far away, since tangents were not originally determined in terms of their slopes. Certain specific problems gradually brought this relation closer to the surface, such as the problem of finding the curves on a globe cutting all meridians of longitude at the same angle (these

curves are called *loxodromes*) or finding a curve all of whose subtangents are the same length. In both of these problems one is trying to construct a curve given certain information about its tangent line: its angle of inclination in the case of a loxodrome, the length of its projection on an axis in the second. Once mathematicians turned their attention to the problem of constructing a curve given information about its subtangents they were headed in the right direction to make this discovery.

The first clear statement of a relation between tangents and areas appears in 1670 in a book entitled *Lectiones Geometricae* by Isaac Barrow (1630–1677), a professor of mathematics at Cambridge and later chaplain to Charles II. Barrow carefully gives the credit for this theorem to "that most learned man, Gregory of Aberdeen" (James Gregory, 1638–1675). Barrow states several theorems resembling the fundamental theorem of calculus. The first theorem (Section 11 of Lecture X) is the easiest to understand. Given a curve referred to an axis, Barrow constructs a second curve such that the ordinate at each point is proportional to the area under the original curve up to that point. We would express this relation as $F(x) = (1/R)\int_a^x f(t)\,dt$, where $y = f(x)$ is the first curve, $y = F(x)$ is the second, and $1/R$ is the constant of proportionality. If the point $T = (t, 0)$ is chosen on the axis so that $(x - t) \cdot f(x) = RF(x)$, then, said Barrow, T is the foot of the subtangent to the curve $y = F(x)$, that is, $x - t$ is the length of the subtangent. In modern language the length of the subtangent to the curve $y = F(x)$ is $\left|F(x)/F'(x)\right|$. This expression would replace $(x - t)$ in the equation given by Barrow. If both $F(x)$ and $F'(x)$ are positive, this relation really does say that $f(x) = RF'(x) = (d/dx)\int_a^x f(t)\,dt$.

Later, in Section 19 of Lecture XI, Barrow shows the other version of the fundamental theorem, that is, that if a curve is chosen so that the ratio of its ordinate to its subtangent (this ratio is precisely what we now call the derivative) is proportional to the ordinate of a second curve, then the area under the second curve is proportional to the ordinate of the first.

In these results Barrow had discovered a theorem logically equivalent to the central fact of the calculus. He also developed some change-of-variable theorems, such as a result (stated in terms of the subtangent) equivalent to the formula

$$\int \frac{y}{y'}\,dy = \int y\,dx.$$

Nevertheless, he had not invented calculus; he formulated his results in terms of the subtangent. What makes calculus a flexible and powerful tool is the use of differential equations, but before they could be introduced the useful but clumsy subtangent had to be replaced by the derivative. This step is the crucial one taken by Newton and Leibniz.

15.2.4 Infinite Series and Products

The methods of integration requiring the summing of infinitesimal rectangles or all the lines inside a plane figure led naturally to the consideration of infinite series.

Several special series were known by the midseventeenth century. For example the Scottish mathematician James Gregory published a work on geometry in 1668 in which he stated the equivalent of

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots.$$

(We have already seen that this result was known in India at least 150 years earlier.) As another example, the Italian priest Pietro Mengoli (1625–1686) discovered the sum of the alternating harmonic series. In our terms this sum is

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots.$$

Likewise at least two infinite product expansions were known by this time for the number π. One, given by Viète, is

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}}\cdots.$$

The other, due to Wallis, is

$$\frac{2}{\pi} = \frac{1 \cdot 3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdots}{2 \cdot 2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdots}.$$

Viète's formula results from inscribing polygons in the circle, starting with a square and continually doubling the number of sides. Wallis obtained his result by trying, like Pascal, to sum the sines inside a quadrant of a circle.

Isaac Newton

It was the binomial series that really established the use of infinite series in analysis. The expansion of a power of a binomial leads to finite series when the exponent is a nonnegative integer, and to an infinite series otherwise. This series, which we now write in the form

$$(1+x)^r = 1 + \sum_{k=1}^{\infty} \frac{r(r-1)\cdots(r-k+1)}{1\cdots k} x^k,$$

was discovered first by Isaac Newton (1642–1727) around 1665, although, of course, he expressed it in a different language. In a 1676 letter to Henry Oldenburg (1615–1677), the Secretary of the Royal Society, Newton wrote

> The extractions of roots are much shortened by the theorem
>
> $$\overline{P+PQ}\Big|\frac{m}{n} = P\Big|\frac{m}{n} + \frac{m}{n}AQ + \frac{m-n}{2n}BQ + \frac{m-2n}{3n}CQ + \frac{m-3n}{4n}DQ + \text{etc.}$$

> where $P + PQ$ stands for a quantity whose root or power or whose root of a power is to be found, P being the first term of that quantity, Q being the remaining terms divided by the first term and m/n the numerical index of the powers of $P + PQ\ldots$ A stands for the first term $P|\frac{m}{n}$, B for the second term $\frac{m}{n}AQ$, and so on... .

Newton's explanation of the meaning of the terms A, B, C,..., means that the kth term is obtained from its predecessor via multiplication by $\{[(m/n) - k]/(k + 1)\}Q$. He stated explicitly that $\frac{m}{n}$ could be any fraction, positive or negative.

Gottfried Wilhelm Leibniz

The codiscoverer of the calculus, Gottfried Wilhelm Leibniz (1646–1716), also practiced summing infinite series, starting from a known sum and deriving a series from what he called the *harmonic triangle*. He had been led to this triangle during the 1670s by reading the works of Pascal. Pascal had written a treatise on the Pascal triangle (which, as we know, had been discovered centuries earlier in India and China). He developed in detail many of the properties of the binomial coefficients that make it up and in particular gave its most prominent property, the fact that each term not in the first row or column is the sum of the number on its left and the number directly above it. It follows that each term is the difference of the term directly below it and the term to the left of that term. Leibniz' harmonic triangle started with the reciprocals of the integers in its first row and column. Thereafter each term was the difference of the term directly above it and the term directly to the right of that term. For comparison, here are the two triangles:

1	1	1	1	1	1	...	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	...
1	2	3	4	5	...		$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{20}$	$\frac{1}{30}$	...	
1	3	6	10	...			$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{30}$	$\frac{1}{60}$	...		
1	4	10	...				$\frac{1}{4}$	$\frac{1}{20}$	$\frac{1}{60}$	...			
1	5	...					$\frac{1}{5}$	$\frac{1}{30}$	...				
1	...						$\frac{1}{6}$	...					

15.2.5 The Synthesis

The results we have just examined show that parts of the calculus were already explicitly recognized by the midseventeenth century, like the pieces of a jigsaw puzzle lying loose on a table. What was needed was someone to see the pattern

and fit all the pieces together. The unifying principle was the concept of a derivative, and that concept came to Newton and Leibniz independently and in slightly differing forms.

15.2.6 Isaac Newton

Isaac Newton was born prematurely on Christmas day in 1642; his parents were minor gentry, but his father had died before his birth. The midwives who assisted at the birth are said to have predicted that the child would not live out the day. (Medical predictions are notoriously unreliable, and this one was wrong by 85 years!) He was 6 years old when the English Civil War began, and the rest of his childhood was spent in that turbulent period. He attended a neighborhood school, and though not particularly a good student, exhibited enough talent to inspire his uncle to send him to Cambridge University, which he entered about the time of the restoration of Charles II to the throne. Although he was primarily interested in chemistry, he did buy and read not only Euclid but also some of the current treatises on algebra and analytic geometry. From 1663 on he attended the lectures of Isaac Barrow.

Due to an outbreak of plague in 1665 he returned to his family home at Woolsthorpe, and during the next two years, while the University was closed, he alternated between Woolsthorpe and his rooms in Cambridge, pursuing his own mathematical and physical researches. He was a careful observer and experimenter, and this period was, as he later recalled, the most productive of his life. Besides the binomial theorem already discussed, he discovered the general use of infinite series and what he called the method of fluxions. He also made discoveries in physics that will be discussed in a later chapter. At the moment we concentrate on the fluxions and infinite series.

First Development of the Calculus

Newton first developed the calculus in what we would call parametric form. Time was the universal independent variable, and the relative rates of change of other variables were computed as the ratios of their absolute rates of change with respect to time. Newton thought of variables as moving quantities, and focused attention on their velocities. To illustrate, he regarded o as a small time interval and used p for the velocity of the variable x, so that the change in x over the time interval o was op. Similarly, using q for the velocity of y, if y and x are related by $y^n = x^m$, then $(y+oq)^n = (x+op)^m$. Both sides can be expanded by the binomial theorem. Then if the equal terms y^n and x^m are subtracted, all the remaining terms are divisible by o. When o is divided out, one side is $nqy^{n-1} + oA$ and the other is $mpx^{m-1} + oB$. Ignoring the terms containing o, since o is small, one finds that the *relative* rate of change of the two variables, q/p is given by $q/p = (mx^{m-1})/(ny^{n-1})$; and since $y = x^{\frac{m}{n}}$, it follows that $q/p = (m/n)x^{(m/n)-1}$. Here at last was the concept of a derivative.

Newton recognized that reversing the process of finding the relative rate of

change provides a solution of the area problem. He was able to find the area under the curve $y = ax^{m/n}$ by working backwards. He considered a curve with ordinates y whose area is z, with the area and the abscissa x related by $z = \big[n/(m+n)\big]ax^{(m+n)/n}$. When x is regarded as the independent variable, if it is increased by o, the area will be increased by oy, since that is the area of a rectangle of base o and height y. Thus $z + oy = \big[n/(m+n)\big]a(x+ox)^{(m+n)/n}$. Following the standard procedure of expanding by the binomial theorem, subtracting z and $\big[n/(m+n)\big]ax^{(m+n)/n}$, dividing by o, and ignoring the terms that still contain o, he found $y = ax^{m/n}$. Since the curve determines and is determined by the curve proportional to the areas, it follows that if $y = ax^{m/n}$, then the area is $z = \big[n/(m+n)\big]ay^{(m+n)/n}$. In this result we see Barrow's work beginning to take a form more like what we are used to seeing in calculus textbooks. It is easy to recognize the formula $\int ax^r\,dx = ax^{r+1}/(r+1)$ here, with $r = m/n$.

Fluxions and Fluents

Newton's "second draft" of the calculus was the concept of fluents and fluxions. A *fluent* is a moving or flowing quantity; its *fluxion* is its rate of flow, what we now call its velocity or derivative. In a work written in Latin in 1671 and published in 1742 (an English translation appeared in 1736), he replaced the notation p for velocity by $\dot{x}$, a notation still used in mechanics and in the calculus of variations. Newton's notation for the opposite operation of finding a fluent from the fluxion has been abandoned: where we write $\int x(t)\,dt$, he wrote $\acute{x}$.

In the opening chapter of the *Fluxions* Newton first explains some operations needed for finding roots of equations, then states the two basic problems: Given an expression for distance in terms of time, compute the velocity, and given an expression for velocity in terms of time, compute the distance. He mentions that the use of the word *time* is merely a convenient way of speaking of a general independent variable. His exposition of the calculus, as in his first conception of it, was parametric, that is, all variables are assumed to depend on some parameter conveniently called time. The first problem is *The relation of the flowing quantities to one another being given, to determine the relation of their fluxions.* The rule given for solving this problem is to arrange the equation that expresses the given relation (assumed algebraic) in powers of one of the variables, say x, multiply its terms by any arithmetic progression (that is, the first power is multiplied by c, the square by $2c$, the cube by $3c$, etc.), and then multiply by $\dot{x}/x$. After this operation has been performed for each of the variables, the sum of all the resulting terms is set equal to zero.

Newton illustrated this operation with the relation $x^3 - ax^2 + axy - y^2 = 0$, for which the corresponding fluxion relation is $3x^2\dot{x} - 2ax\dot{x} + a\dot{x}y + axy - 2y\dot{y} = 0$, and by numerous examples of finding tangents to well-known curves such as the spiral, and the cycloid. Newton also found their curvatures and areas. The combination of these techniques with infinite series was vital, since fluents often could not be found in finite terms. For example, Newton found that the area under the curve $\dot{z} = 1/(1+x^2)$ was given by Gregory's series $z = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \cdots$.

Later Reflections on the Calculus

The technique of using fluxions and fluents was simple and the richness of the benefits to be derived from it made a mathematician's paradise. This paradise, however, contained a serpent, namely its dubious logical foundations. Operations were performed using increments in an independent variable, division, for example, which are not allowed if the increment is zero. But then, after the division is performed, these increments are set equal to zero. Is this not self-contradictory? Similarly, in his use of infinite series, Newton asserted that a series would vanish if all of its terms vanished. This statement is no doubt true, but in its applications the terms vanish because their denominators become infinite. Can finite quantities flow out to infinity? If they do, will sums and products made from them flow to the "right" values, especially when there are infinitely many terms? These questions would eventually have to be answered, and for a time even good mathematicians occasionally got the answer wrong.

Newton later made another attempt to explain fluxions in terms that would be more logically acceptable, calling it the "method of first and last ratios." In his great treatise on mechanics, the *Philosophiae Naturalis Principia Mathematica*, published in 1687, he explained a ratio of two fluxions as follows.

> *Quantities, and the ratios of quantities, which in any finite time converge continually toward equality, and before the end ot that time approach nearer to each other than by any given difference, become ultimately equal.*
>
> If you deny it, suppose them to be ultimately unequal, and let D be their ultimate difference. Therefore they cannot approach nearer to equality than by that given difference D; which is contrary to the supposition...

Newton offered this reasoning as a salve to the logical conscience of those who found the approach through indivisibles dubious. In fact he came close to stating the modern concept of a limit, when he described the "ultimate ratios" (derivatives) as "limits towards which the ratios of quantities decreasing without limits do always converge, and to which they approach nearer than by any given difference." Here one can almost see the "arbitrarily small ε" that plays the central role in the concept of a limit.

Newton's Later Career

Newton did not enjoy publishing and kept his brilliant discoveries from the plague years mostly to himself. When Barrow left Cambridge in 1669 to become chaplain to Charles II, Newton was elected his successor as the Lucasian Professor. In 1672 he was elected a member of the Royal Society. It was only at the urging of the astronomer Halley a decade later that he was persuaded to write his masterpiece, the *Principia*, and Halley had to take responsibility for the expense of printing it and settling a quarrel with the physicist Robert Hooke over priority for the inverse-square law of gravitation. The strain of organizing this large systematic treatise

took a heavy toll on Newton's peace of mind. In 1692 he suffered a prolonged bout of nervous irritability, from which he recovered only in 1694. After this illness he never again initiated any original research into scientific questions, though he continued to work on specific problems. Entering a career in public life, he became warden of the mint in 1695 and master of the mint in 1699. In 1703 he became president of the Royal Society. He invested his money shrewdly and left a huge fortune at his death.

Newton's eccentricities were part and parcel of his genius, which is unquestionably above nearly every other thinker in history. His nervous irritability made him unpleasant company sometimes, and it is not surprising that he never married. He spent a great deal of time in arcane alchemical research and in trying to penetrate the mysteries of Biblical prophecy. If he learned anything from these researches, it has not been appreciated by posterity. He died in 1727 and is buried at Westminster Abbey in London.

15.2.7 Gottfried Wilhelm von Leibniz

The codiscoverer with Newton of the calculus was, like Newton, a man involved in public life, but a much more amiable character. The philosopher Bertrand Russell, who had studied Leibniz and understood him better than anyone else, proclaimed him not an admirable man. According to Russell, Leibniz developed a profound philosophy, which he kept secret, knowing that it would not be popular, and published instead only a fatuous optimism aimed at winning friends. Leibniz, the optimistic philosopher, was parodied in the character of Dr. Paingloss in Voltaire's *Candide*. As was the case with Newton, Leibniz had wide-ranging interests as a youth and focused on mathematics only in early adulthood. He was born in Leipzig and entered the university there in 1661, at the age of 15. Like Descartes, Fermat and Viète, he studied the law, but was considered too young to be awarded the degree of doctor of laws when he finished his course at the age of 20. He entered the service of the Elector of Mainz as a diplomat and finally came to serve the Electors of Hannover for four decades, including the future King George I of Britain, who succeeded Queen Anne in 1714.

During his lifetime France was nearing the zenith of its power on the Continent, while Germany was divided and weak. As servant of several German princes, Leibniz attempted to shield Germany from the power of the French by diverting the interests of Louis XIV toward a holy war in Egypt. It was during a mission to Paris in 1672 that Leibniz became interested in mathematics and began to read the writings of Pascal. The following year he visited London and met some members of the Royal Society, including the secretary Henry Oldenburg and the librarian James Collins (1625–1683). He kept a diary of this journey on a sheet of paper ruled into columns headed Chemistry, Mechanica, Magnetica, Botanica, etc. Under mathematics the notes are very sparse, containing only a reference to a general method of finding tangents, probably derived from the lectures of Barrow, which he had bought.

From this time on Leibniz studied mathematics in earnest and within a decade

had derived most of the calculus in essentially the form we know it today. His approach to the subject, in particular the delicate notion of the meaning to be assigned to the limiting ratio of two quantities as they vanish, is quite different from Newton's. Leibniz believed in the reality of infinitesimals, quantities so small that any finite sum of them is still less than any assignable positive number, but which are nevertheless not zero, so that one is allowed to divide by them. The three kinds of numbers (finite, infinite, and infinitesimal) could, in Leibniz' view, be multiplied by one another, and the result of multiplying an infinite number by an infinitesimal might be any one of the three kinds. This position was rejected in the nineteenth century, but was resurrected in the twentieth century and made logically sound. It lies at the heart of what is called "nonstandard analysis," a subject that has not penetrated the undergraduate curriculum. The radical step that must be taken in order to believe in infinitesimals is a rejection of the Archimedean axiom that for any two positive quantities a sufficient number of bisections of one will lead to a quantity smaller than the second. This principle was essential to the use of the method of exhaustion, which was one of the crowning glories of Euclidean geometry. It is no wonder that mathematicians were reluctant to give it up.

Leibniz was influenced by the writings of Pascal and Barrow and was interested in the triangle that appears in Pascal's paper on the summation of sines (and had appeared in connection with the tangent problem in the work of both Fermat and Barrow). The principle for constructing this triangle was the same in the works of both Fermat and Barrow: First find a finite right triangle with two vertices on the curve, that is, let the hypotenuse be a chord of the curve and the sides parallel to the coordinate axes. The finite triangle gives a slope whose numerator is a small difference in y and whose denominator is a small difference in x. When one takes account of the equation of the curve, it often happens that the small difference in x can be canceled from the numerator and denominator, leaving an expression independent of this difference, plus a second expression that still contains the difference as a factor. The first of these expressions thus determines the slope of the tangent line at the point in question as a perfectly well-defined finite entity.

It was Leibniz who invented the expression dx to indicate the difference of two infinitely close values of x, dy to indicate the difference of two infinitely close values of y, and dy/dx to indicate the ratio of these two values. This notation was beautifully intuitive and is still the preferred notation for thinking about calculus. Its logical basis at the time was questionable, since it avoided the objections listed above by claiming that the two quantities have not vanished at all, but have yet become less than any assigned positive number. However, at the time consistency would have been counterproductive in mathematics and science. At the heart of Leibniz' calculus was the characteristic triangle, in which the horizontal side is the infinitesimal change in x denoted dx, the vertical side is dy, and the hypotenuse ds is the infinitesimal change in arc length on the curve.

The integral calculus and the fundamental theorem of calculus, flowed very naturally from Leibniz' approach. Just as Leibniz had been able to sum the rows of the harmonic triangle because of the collapsing property of the sums, that is,

$$\tfrac{1}{1\cdot 2}+\tfrac{1}{2\cdot 3}+\tfrac{1}{3\cdot 4}+\cdots=\left(1-\tfrac{1}{2}\right)+\left(\tfrac{1}{2}-\tfrac{1}{3}\right)+\left(\tfrac{1}{3}-\tfrac{1}{4}\right)+\cdots=1,$$

he could argue that the ordinates to the points on a curve represent infinitesimal rectangles of height y and width dx, and hence finding the area under the curve—"summing all the lines in the figure"—amounted to summing infinitesimal differences in area dA, which collapsed to give the total area. Since it was obvious that on the infinitesimal level $dA = y\,dx$, the fundamental theorem of calculus was an immediate consequence. Leibniz eventually abbreviated the sum of all the increments in the area (that is, the total area) using an elongated S, so that $A = \int dA = \int y\,dx$.

Nearly all the basic rules of calculus for finding the derivatives of the elementary functions and the derivatives of products, quotients, etc., were contained in Leibniz' 1684 paper on his method of finding tangents. However, he had certainly obtained these results much earlier. His collected works contain a paper written in Latin with the title *Compendium Quadraturae Arithmeticae*, to which the editor assigns a date of 1678 or 1679. This paper shows Leibniz' approach through infinitesimal differences and their sums and suggests that it was primarily the problem of squaring the circle and other conic sections that inspired this work. The work consists of 49 propositions and two problems. Most of the propositions are stated without proof; they contain the basic results on differentiation and integration of elementary functions, including the Taylor series expansions of logarithms, exponentials, and trigonometric functions. Although the language seems slightly archaic, one can easily recognize a core of standard calculus here.

Despite the attractiveness of infinitesimal methods, Leibniz did not simply throw logical caution to the winds. In some places he showed how to justify his arguments by the method of exhaustion. In others he attempted to explain his rules for working with infinitesimals. Most students who have struggled to understand the later concept of a limit will be attracted to the apparent simplicity of infinitesimals. The difficulty with infinitesimals occurs in the so-called indeterminate forms, in which an infinite number is multiplied by an infinitesimal or one infinity is subtracted from another. The result may be finite, infinite, or infinitesimal, and hard work may be involved in deciding which is the case. The fundamental principle laid down by Leibniz in his *Compendium* (Proposition 20) is that

> if $V + X$ and $V + Z$ have a finite ratio (not unity) and X and Z are finite, then V is also finite, while if one of X and Z is infinite [Leibniz always assumes they are not both infinite], then V is infinite.

Leibniz put this principle to work in studying what he called paraboloids and hyperboloids, the curves whose equations are $y^m = a^{m-n}x^n$ and $x^n y^m = a^{m+n}$. (The cases $m = 1$, $n = 2$ in the first equation and $m = n = 1$ in the second give respectively the ordinary parabola and hyperbola.) Leibniz referred a general hyperboloid to a pair of asymptotes as axes and showed (Proposition 21) that "the rectangle whose sides are an infinitesimal abscissa and an infinite ordinate is infinite, finite, or infinitesimal according as the exponent of the ordinate is less than, equal to, or greater than the exponent of the abscissa." To Leibniz this infinitely long, infinitely thin rectangle was a real object, consisting of half-lines parallel to

the asymptotes and between the asymptotes and the curve.[1] We would think of it as the limiting area of a rectangle having sides along the asymptotes, one corner being at the intersection of the asymptotes and the opposite corner on the curve. If we write the equation of Leibniz' hyperboloid as $y = kx^{-\frac{n}{m}}$, this rectangle has area $xy = kx^{1-\frac{n}{m}}$, and Leibniz' assertion is verified: When $x = \infty$, this is infinite, finite or infinitesimal (zero) according as $m > n$, $m = n$, or $m < n$ (for Leibniz, x was the ordinate and y the abscissa). From this result Leibniz deduced (Proposition 22) the important result that in any hyperboloid (except the conic hyperbola), the area "under the curve" along one asymptote is finite, while the area along the other is infinite. In his derivation he took V to be the area "under the curve" from some point P out to infinity, Z the infinitely long, infinitely thin rectangle just mentioned, and X the rectangle having sides along the asymptotes, and opposite corners at P and the point of intersection of the asymptotes. He asserted correctly that $(V + X)/(V + Z) = m/n$, from which the result followed by Proposition 20.

The *Compendium* contains the basic results on integration, for example (Proposition 25), in Leibniz' notation:

$$\overline{\int x^n \, dx} = \frac{\overline{\text{diff.}(x)^{\frac{n+1}{\cdot}}, x^{\frac{n+1}{\cdot}}}}{n+1}.$$

Even with the full power of derivatives and integrals to work with, however, Leibniz needed the infinite in yet a third form, namely infinite series. His results on the harmonic triangle are stated in the *Compendium*, and many of his results are derived by integrating the geometric series or the binomial series (which Leibniz knew very well from his correspondence with Oldenburg and Collins). For example (Proposition 35): *A circle is to the inscribed square... as*

$$\frac{1}{1-\frac{1}{4}} + \frac{1}{9-\frac{1}{4}} + \frac{1}{25-\frac{1}{4}} + \cdots$$

is to 1. One may wonder why Leibniz uses an infinite series here, since it is well known that one can express all integrals whose worst irrationality is the square root of a quadratic in terms of logarithms and trigonometric functions. Certain other papers of Leibniz give the answer: these functions are known only through tables calculated laboriously from certain properties of the functions. The power series gives first of all a new way of computing the tables. More than that, however, it provides a way of calculating the values on the spot, wherever needed, thus giving a much more satisfactory numerical approximation to the ratio π for example, and making it more practical to dispense with tables entirely. In one of his papers from 1691 Leibniz emphasized that he had given an arithmetic quadrature of the circle that eliminated any need for trigonometric tables.

Our names for some of the functions considered by Leibniz have changed. The notion of the exponential function, for example, was not used by Leibniz; instead he let y be the logarithm of x and then gave the expansion of x in terms of y. With

[1] If strict rigor is applied, the existence of such a rectangle contradicts the definition of asymptotes.

that caveat we can assert nevertheless that the *Compendium* contains the equivalent of the series expansions

$$\begin{aligned}
\ln\left(\frac{1}{1-x}\right) &= \frac{x}{1} + \frac{x^2}{2} + \frac{x^3}{3} + \cdots \\
\ln(1+x) &= \frac{x}{1} - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \\
a\left(\exp\left(\frac{y}{a}\right) - 1\right) &= \frac{y}{1} + \frac{y^2}{2!a} + \frac{y^3}{3!a^2} + \cdots \\
\cos a &= 1 - \frac{a^2}{2!} + \frac{a^4}{4!} - \cdots \\
\sin a &= a - \frac{a^3}{3!} + \frac{a^5}{5!} - \cdots,
\end{aligned}$$

and other well-known Taylor series. Leibniz also states the rules for working with series whose terms alternate in sign and decrease in absolute value (Proposition 49).

Later Reflections on the Calculus

Leibniz, like Newton, was forced to answer objections to the new methods of the calculus. In the *Acta Eruditorum* of 1695 Leibniz published (in Latin) a "Response to certain objections raised by Herr Bernardo Niewentiit regarding differential or infinitesimal methods." These objections were three: (1) that certain infinitely small quantities were discarded as if they were zero (this principle was set forth as fundamental in the following year in the textbook of calculus by the Marquis de l'Hospital); (2) the method could not be applied when the exponent is a variable; and (3) the higher-order differentials were inconsistent with Leibniz' claim that only geometry could provide a foundation. In answer to the first objection Leibniz attempted to explain different orders of infinitesimals, pointing out that one could neglect all but the lowest orders in a given equation. To answer the second, he used the binomial theorem to demonstrate how to handle the differentials dx, dy, dz when $y^x = z$. To answer the third Leibniz noted that one should not think of $d(dx)$ as a quantity that fails to yield a (finite) quantity even when multiplied by an infinite number. He pointed out that if x varies geometrically when y varies arithmetically, then $dx = (x\,dy)/a$ and $ddx = (dx\,dy)/a$, which makes perfectly good sense.

Leibniz' Later Career

Leibniz' diplomatic career, had it transpired in the twentieth century, might have gained him the Nobel Peace Prize. He was not only a staunch European, working to revive the moribund Empire (to which Napoleon administered the coup de grace in 1806) but was widely read and interested in Oriental culture. As we have seen, determinants were discovered in Japan about this time. Leibniz was the first to introduce them into Europe, in a letter to the Marquis de L'Hospital in 1693. He

gives no sign that he learned of determinants through reading accounts of Oriental mathematics, and one must presume that he thought of the idea himself. No particular attention was paid to determinants at the time in any case, and they had to be rediscovered in the next century.

In his work as a diplomat he naturally came to notice the difficulties of communicating through natural languages and sought a better design for language, one based on symbols. This topic, in fact, was the subject of Leibniz' first mathematical work, *De arte combinatoria* (1666), which was prefaced by an essay claiming to prove by mathematics the existence of God. (The argument is based on the concept of a Prime Mover, and was not new. What was new was the invocation of mathematics in defense of its validity.)

Leibniz was an indefatigable organizer not only in politics but also in science. He was instrumental in the formation of the Berlin Scientific Society and urged the founding of similar societies in Vienna and Dresden. In 1714, near the end of his life, he traveled to Russia and met the Tsar (Peter I). Hoping to increase contacts between western Europe and Russia, he proposed the founding of an Academy of Sciences in Russia. In the last year of Peter's life (1724) this Academy was duly founded at St. Petersburg and staffed by 11 imported Western European scientists.

15.2.8 The Disciples of Newton and Leibniz

Newton and Leibniz had disciples who carried on their work. Among Newton's followers were Roger Cotes (1682–1716), who oversaw the publication of a later edition of Newton's *Principia* and defended Newton's inverse square law of gravitation in a preface to that work. He also fleshed out the calculus with some particular results on plane loci and considered the extension of functions defined by power series to complex values, deriving the important formula $i\phi = \log(\cos\phi + i\sin\phi)$, where $i = \sqrt{-1}$. Another of Newton's followers was Brook Taylor (1685–1731), who developed a calculus of finite differences that mirrors in many ways the "continuous" calculus of Newton and Leibniz and is of both theoretical and practical use today. Taylor is famous for the infinite power series representation of functions that now bears his name. It appeared in his 1715 treatise on finite differences. We have already seen, however, that many particular "Taylor series" were known to Newton and Leibniz; Taylor's merit is to have recognized a general way of producing such a series in terms of the derivatives of the generating function.

Leibniz also had a group of active and intelligent followers who continued to develop his ideas. The most prominent of these were the Bernoulli brothers Jakob (sometimes referred to in the literature as James or Jacques, 1654–1705) and Johann (sometimes referred to in the literature as John or Jean, 1667–1748), citizens of Switzerland, between whom relations were not always cordial. They investigated problems that arose in connection with calculus and helped to systematize, extend, and popularize the subject. In addition they pioneered new mathematical subjects such as the calculus of variations, differential equations, and the mathematical theory of probability. A French nobleman, the Marquis de l'Hospital (1661–1704), took lessons from Johann Bernoulli and paid him a salary in return for the right

to Bernoulli's mathematical discoveries. As a result, Bernoulli's discovery of a way of assigning values to what are now called indeterminate forms appeared in L'Hospital's textbook *Analyse des infiniment petits* (1696), and has ever since been known as L'Hospital's rule. Like the followers of Newton, who had to answer the objections of Bishop Berkeley (see Chapter 18 below) Leibniz' followers encountered objections from Michel Rolle (1652–1719), which were answered by Johann Bernoulli with the claim that Rolle didn't understand the subject.

There is some irony in this claim of Bernoulli's, since Rolle's theorem is equivalent to the mean-value theorem. Rolle's approach, however, was purely algebraic. He described a certain operation on polynomials exactly as Newton had described the fluxion, multiplying a given polynomial termwise by the terms of an arithmetic sequence (the constant term is multiplied by zero, the linear term by a, the square term by $2a$, and so on), then dividing by the independent variable. The result, which Newton called a fluxion, he called a cascade. In a 1691 treatise on the solution of polynomial equations he noted that the cascade always has a root between any two roots of the polynomial (Rolle's theorem).

The Priority Dispute

One of the better-known and less edifying incidents in the history of mathematics is the dispute between the disciples of Newton and those of Leibniz over the credit for the invention of the calculus. Although Newton had discovered the calculus by the early 1670s and had described it in a paper sent to James Collins, the librarian of the Royal Society, he did not publish his discoveries until 1687. Leibniz made his discoveries a few years later than Newton, but published some of them earlier, in 1684. Newton's vanity was wounded in 1695 when he learned that Leibniz was regarded on the Continent as the discoverer of the calculus, even though Leibniz himself made no claim to this honor. In 1699 a Swiss immigrant to England, Nicolas Fatio de Duillier (1664–1753), suggested that Leibniz had seen Newton's paper when he had visited London and talked with Collins in 1673. (Collins died in 1683, before his testimony in the matter was needed.) This unfortunate affair poisoned relations between Newton and Leibniz and their followers. In 1711–1712 a committee of the Royal Society (of which Newton was President) investigated the matter and reported that it believed Leibniz had seen certain documents that in fact he had not seen. Relations between British and Continental mathematicians reached such a low ebb that Newton deleted certain laudatory references to Leibniz from the third edition of his *Principia.* This dispute confirmed the British in the use of the clumsy Newtonian notation for more than a century, a notation far inferior to Leibniz's elegant and intuitive symbolism. Eventually even the British came to prefer the term *integral* to *fluent* and *derivative* to *fluxion.*

State of the Calculus around 1700

Most of what we now know as calculus—rules for differentiating and integrating elementary functions, solving simple differential equations, and expanding func-

tions in power series—was known by the early eighteenth century. Nevertheless, there was much unfinished work. We list here a few of the open questions:

Nonelementary integrals Differentiation of elementary functions is an algorithmic procedure, and the derivative of any elementary function whatsoever, no matter how complicated, can be found if the investigator has sufficient patience. Such is not the case for the inverse operation of integration. Many important elementary functions such as $(\sin x)/x$ and e^{-x^2} are not the derivatives of elementary functions. Within the sphere of algebraic functions, although all rational functions could be integrated (provided the polynomials in the numerator and denominator could be factored) and even the square roots of quadratic polynomials made no difficulty, more complicated integrals involving cube roots or the square roots of cubic polynomials could not usually be expressed in terms of elementary functions. Since such integrals turned up in the analysis of some fairly simple motions, such as that of a pendulum, the problem of these integrals became pressing.

Classification and solution of differential equations Although integration had originally been associated with problems of area and volume, because of the importance of differential equations in mechanical problems the solution of differential equations soon became the major application of integration. The general procedure was to convert an equation to a form in which the derivatives could be eliminated by integrating both sides (reduction to quadratures). As these applications became more extensive, more and more cases began to arise in which the natural physical model led to equations that could not be reduced to quadratures. The subject of differential equations began to take on a life of its own independent of the calculus.

Foundational difficulties The philosophical difficulties connected with the use of infinitesimal methods were paralleled by mathematical difficulties connected with the application of the algebra of finite polynomials to infinite series. These difficulties were hidden for some time, and for a blissful century mathematicians and physicists operated formally on power series as if they were finite polynomials. They did so even though it had been known since the time of Oresme that the partial sums of the harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ grow arbitrarily large.

15.3 Problems and Questions

15.3.1 Problems in the Early Calculus

Exercise 15.1 Verify Descartes' assertion that the line segment MO is the solution of the equation $z^2 = az + b^2$ if $LM = b$ and OPL is a circle of radius $\frac{1}{2}a$ with center at N and tangent to LM at L. (See Fig. 15.7.)

Exercise 15.2 Find the normal to the ellipse considered by Descartes with $r = 4$, $q = 1$ at the point $(1, \frac{1}{2})$.

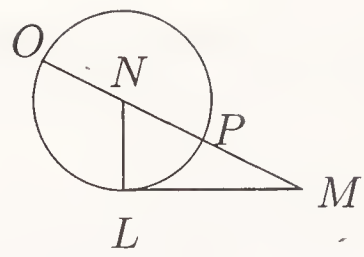


Figure 15.7: Geometric solution of a quadratic equation.

Exercise 15.3 Find the normal to the curve $y = \sqrt{x}$ at the point $(1, 1)$, using the x-axis as the coordinate line, following Descartes' method.

Exercise 15.4 Find the length of the subtangent to the ellipse $(x^2/9)+(y^2/4) = 1$ at the point $(3\sqrt{3}/2, 1)$, using Fermat's method.

Exercise 15.5 Consider an ellipse with semiaxes a and b and a circle of radius b, both circle and ellipse lying between a pair of parallel lines a distance $2b$ apart. For every line between the two lines and parallel to them, show that the portion inside the ellipse will be a/b, times the portion inside the circle. Use this fact and Cavalieri's principle to compute the area of the ellipse. [This result was given by Kepler.]

Exercise 15.6 Let $P_n = \cos(\theta/2)\cos(\theta/4)\cdots\cos(\theta/2^n)$ and $Q_n = 2^n \sin(\theta/2^n)$. Prove by induction that $P_nQ_n = P_{n-1}Q_{n-1} = \cdots = P_1Q_1 = \sin\theta$. Since $Q_n \to \theta$ as $n \to \infty$, it follows that $P_n \to (\sin\theta)/\theta$. Take $\theta = \frac{\pi}{2}$ and use the formula $\cos(\varphi/2) = \left[(1+\cos\varphi)/2\right]^{1/2}$ to derive Viète's formula.

Exercise 15.7 By taking $\varphi = \pi/4$ in Cotes' formula $i\phi = \log(\cos\phi + i\sin\phi)$, deduce that $\log(1+i) = \log(\sqrt{2}) + i(\pi/4)$. By integrating the series $1/(1+t) = \sum_{n=0}^{\infty} t^n$ from $t = 0$ to $t = i$ and comparing real and imaginary parts, derive Mengoli's sum of the alternating harmonic series and Proposition 35 of Leibniz' *Compendium*.

Exercise 15.8 Show that the point at which the tangent to the curve $y = f(x)$ intersects the y axis is $y = f(x) - xf'(x)$, and verify that the area under this curve (more precisely, the integral of $f(x) - xf'(x)$ from $x = 0$ to $x = a$) is twice the area between the curve $y = f(x)$ and the line $ay = f(a)x$ between the points $(0, 0)$ and $(a, f(a))$. (This result was used by Leibniz to illustrate the power of his infinitesimal methods.)

15.3.2 Questions about the Early Calculus

Exercise 15.9 What might have been Descartes' reason for introducing his geometric method in the course of explaining his method in philosophy? Why would a book on geometry be relevant to a treatise on philosophy?

Exercise 15.10 As we saw, the Chinese and the early Greeks had known the principle called Cavalieri's principle. What claim does Cavalieri have to the name of this principle? How would you assign credit for this principle?

Exercise 15.11 Recall that Eudoxus solved the problem of incommensurables by changing the definition of proportion, or rather, *making* a definition to cover cases where no definition existed before. Newton's "theorem" asserting that quantities that approach each other continually (we would say monotonically) and become arbitrarily close to each other in a finite time must become equal in an infinite time assumes that one has a definition of equality at infinity. What is the definition of equality at infinity? Since we cannot *actually* reach infinity, the definition will have to be stated as a potential infinity, that is, a statement about all possible finite times. Formulate the definition, and then compare Newton's solution of this difficulty with Eudoxus' solution of the problem of incommensurables.

Exercise 15.12 Compare the use of ratios in modern times with Leibniz' discussion of them in his *Compendium*. How would Euclid have responded to Leibniz' arguments if he could have read them?

Exercise 15.13 Draw a square and one of its diagonals. Then draw a very fine "staircase" by connecting short horizontal and vertical line segments in alternation, each segment crossing the diagonal. Clearly the total length of the horizontal segments is the same as the side of the square, and the same is true of the vertical segments. Now in a certain intuitive sense these segments approximate the diagonal of the square, seeming to imply that the diagonal of a square equals twice its side, which is absurd. Does this argument show that the method of indivisibles is wrong? How could Cavalieri, for example, have defended his method against such criticism?

Exercise 15.14 The reader may have noticed that Viète, Fermat, Descartes, and Leibniz were all trained in the law. The law thus seems to have provided the world with a large number of mathematicians. The number of composers who were trained for the law is also impressive, including George Frederick Handel, Carl Phillip Emmanuel Bach, Robert Schumann, and Peter Ilyich Tchaikovsky. Can you think of anything that mathematics, law, and music have in common that would account for the apparently large number of people who excel in all three?

15.4 Endnotes

1. The discussions of all the mathematicians in this chapter have been based on their published collected works and some excerpts that have been gathered in various collections of sources. The following are the source books used:

   (a) D. J. Struik. *A Source Book in Mathematics, 1200–1800* (Harvard University Press, 1969).

(b) David Eugene Smith. *A Source Book in Mathematics* (Dover Reprint, New York, 1959).

(c) Ronald Calinger, ed. *Classics of Mathematics* (Prentice-Hall, Englewood Cliffs, NJ, 1995).

(d) John Fauvel and Jeremy Gray. *The History of Mathematics. A Reader* (Macmillan, London, 1987).

2. The quotation from Fermat on plane loci is taken from the source book of David Eugene Smith (op. cit.), p. 389.

3. The quotation from Descartes is taken from the Dover translation of *La géométrie*, pp. 2–5.

4. The discussion of Cavalieri's work is based on the source book of Struik (op. cit.).

5. The quotation from Fermat on the quadrature of hyperboloids is taken from the book of Calinger (op. cit.), p. 375.

6. The quotation from Pascal on the sums of sines is taken from the book of Calinger (op. cit.), p. 182.

7. Newton's statement of the binomial theorem is taken from the source book of David Eugene Smith (op. cit.), p. 225.

8. The quotation from Newton's *Principia* is taken from the Motte-Cajori translation (University of California Press, 1966), Vol. 1, p. 38.

Chapter 16

Seventeenth-Century Mathematics

While the invention of calculus was the most important mathematical event of the modern era, it was by no means the only new development in the seventeenth century. In this chapter we shall look at several other significant changes in mathematics during this period: the development of projective geometry and probability, some advances in algebra and number theory, the first Western study of combinatorics, the invention of calculating machines, and the establishment of scientific societies and journals.

16.1 Geometry

We have seen that projective geometry began in the Renaissance, inspired by the desire of artists to represent three-dimensional scenes more realistically. During the seventeenth century this subject had two proponents who produced some results of great significance. The work was temporarily eclipsed by the spectacular development of analysis, but eventually came to be appreciated in the nineteenth century.

16.1.1 Desargues

The French architect and engineer Gérard Desargues (1591–1661) studied the projections of figures in general and the conic sections in particular. Knowing the way in which conics were originally created by the intersection of a cone with a plane, he saw that any projection of a conic section from one plane to another would remain a conic section. Similarly, if a triangle is projected from a point outside its plane onto a plane that intersects its own plane, then each side of the triangle and the projection of that side will lie in the plane determined by that side and the center of projection. If no side is parallel to its projection, then the line

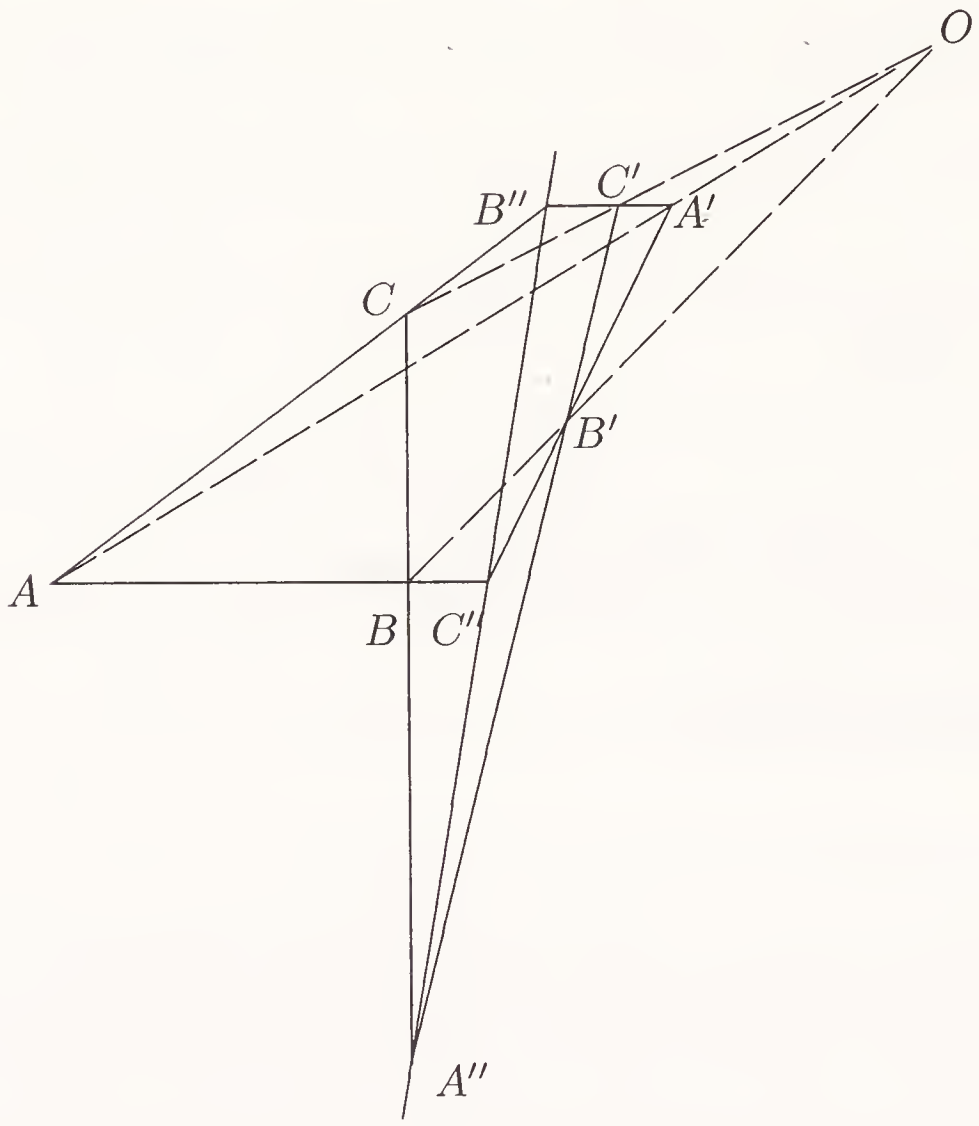


Figure 16.1: Desargues' theorem.

containing each side of the triangle will intersect the line through its projection, giving three points of intersection. Each point in which the extension of a side meets the extension of its projection lies in both the plane of the triangle and the plane of its projection. It follows that these three points of intersection will lie on a straight line, namely the line of intersection of the plane of the triangle and the plane of its projection. To make this argument clear see Fig. 16.1, in which triangle $A'B'C'$ is a section of the projection of triangle ABC from the point O. Sides AC and $A'C'$, when extended, meet in point B''; the extensions of sides AB and $A'B'$ meet in C''; and the extensions of sides BC and $B'C'$ meet in A''. The three points A'', B'', and C'' are collinear. This theorem remains true when both triangles are in the same plane, although the proof is more difficult.

The theorem just stated is a simple theorem in solid geometry; what makes it the starting point for projective geometry is the case when one side of the triangle is parallel to its projection. In that case it is easy to see that these two sides will also be parallel to the line of intersection of the two planes. For this case Desargues established a convention that a family of parallel lines in a plane has a fictitious point in common, nowadays called the *point at infinity*. The fictitious points of a plane (one for each family of parallel lines) form a fictitious line, called the *line at infinity*. This convention is extremely useful in geometry. Desargues, however, expressed his ideas so badly, with neologisms and generally clumsy notation, that hardly anyone paid any attention to it.

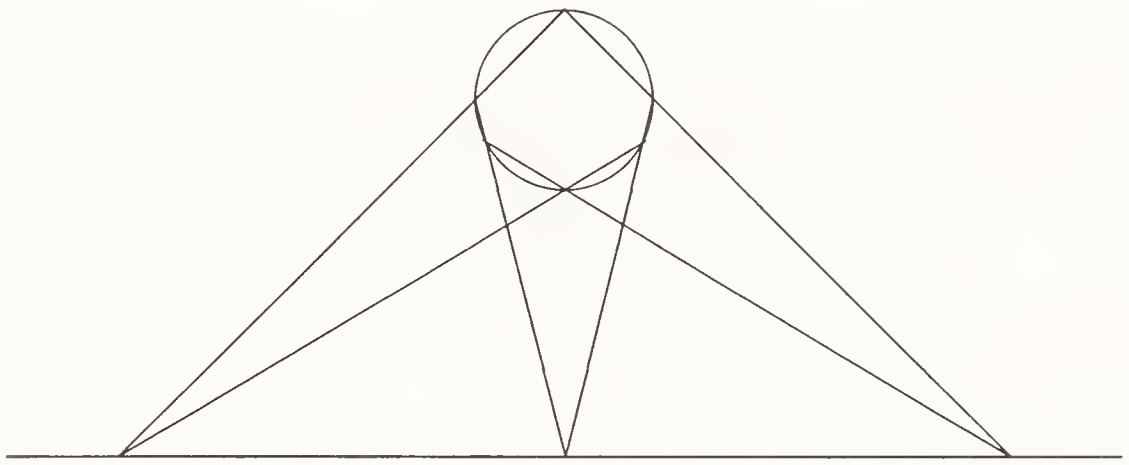


Figure 16.2: Pascal's theorem.

16.1.2 Pascal

It cannot be said, however, that *no one* paid any attention to Desargues. His treatise, which he called a *brouillon projet* (rough draft), was published in 1639. In that year Blaise Pascal was a youth of 16, but extremely precocious. He wrote an *Essay pour les coniques* in which he proved some fundamental theorems about projective invariants. He knew of Desargue's work, calling him "one of the great geniuses of this time and well versed in mathematics, particularly in conics." He further states, "I owe the little that I have found on this subject to his writings and...have tried to imitate his method, as far as possible... ."

Like Desargues, Pascal defines a family of lines all meeting at the same point or all parallel as being of the same *ordonnance*. Such a family is now called a *pencil* or a *sheaf*. To explain Pascal's work, we note that if an irregular hexagon is inscribed in a circle, the lines containing the pairs of opposite sides will normally meet in three collinear points (see Fig. 16.2). If one pair of opposite sides happens to be parallel and the other two pairs are not, the line determined by the points of intersection of the two pairs of nonparallel opposite sides will be parallel to the other two sides, that is, it will pass through the point at infinity associated with that family of parallel lines. If two pairs of opposite sides are parallel, then the third pair is also, and so all three pairs of opposite sides meet in the line at infinity.

This theorem is projectively invariant, since the lines in an *ordonnance*, when projected, become the lines in another *ordonnance*. (If a family of lines all intersect in a point, and they are projected from that point, they will project to a family of parallel lines.) Since an ellipse can be projected to a circle, it follows that if a hexagon is inscribed in an ellipse, the points of intersection of the pairs of opposite sides are collinear.

16.2 Probability

One of the most interesting and intuitively difficult parts of mathematics is the theory of probability. In its elementary parts, in which the possible outcomes of an observation or experiment are finite in number and equally likely, nearly everyone remembers some confusing introduction to the subject. Unfortunately

such introductions entangle probability with combinatorics in such a way that the sophisticated counting methods, which are only a tool in the analysis, cause the student to lose sight of the fundamental probabilistic ideas. The mathematical subject has a rich and interesting history, dating back to Cardano, whose *Liber de ludo* (Book on Gambling) was published about a century after his death. In this book Cardano introduces the idea of assigning a probability p between 0 and 1 to an event whose outcome is not certain. The principal applications of this notion were in games of chance, where one might bet, for example, that a player could roll a 6 with one die given three chances. The subject is not developed in detail in Cardano's book, much of which is occupied by descriptions of the actual games played. However, Cardano does state the multiplicative rule for a run of successes in independent trials. Thus the probability of getting a six on each of three successive roles with one die is $\left(\frac{1}{6}\right)^3$. Most important, he recognized the real-world application of what we call the law of large numbers, saying that when the probability for an event is p, then after a large number n of repetitions, the number of times it will occur does not lie far from the value np. (That is, it is not certain that the number of occurrences will be near np, but "that is where the smart money bets.")

The problem that inspired much of the subsequent development requires some imagination to appreciate. An analogy may perhaps make the difficulty clearer. In discussing the legacy problems of Al-Khwarizmi, we noted that a debt owed to a deceased person by one of the heirs entered a kind of legal limbo. Only a certain portion of the debt became part of the inheritance, and that portion was chosen in accordance with a legal principle. A similar fate overtakes money that has been put up at stake on a bet. Once the bet is made, by the gamblers' unwritten code, the stakes do not belong to anyone. After the bet is settled, the whole amount belongs to the winner, and of course, before the bet was made, each player owned the amount of the stake he/she put up. In the meantime, however, after the bet is made and before it is settled, a player cannot unilaterally withdraw from the bet and recover her or his stake. What happens then if the game is interrupted? How are the stakes to be divided? The principle that seemed fair was that, *regardless of the relative amount of the stake each player had bet, a player should recover only a portion of the stakes equal to that player's probability of winning at the moment the game was terminated.* The translation of this principle into francs and sous involves computing the probability of winning at each point of a game, what we now call conditional probability. This operation is different for each game and usually involves the combinatorial counting techniques the reader has no doubt encountered.

16.2.1 Fermat and Pascal

A French nobleman, the Chevalier de Méré, who was fond of gambling, proposed to Pascal the problem of dividing the stakes in a game where one player has bet that a six will appear in eight rolls of a single die, but the game is terminated after three unsuccessful tries. Pascal wrote to Fermat that the player should be

allowed to sell the throws one at a time. If the first throw is foregone, the player should take one-sixth of the stake, leaving five-sixths. Then if the second throw is also foregone, the player should take one-sixth of the remaining five-sixths or $\frac{5}{36}$, etc. In this way, Pascal argued that the fourth through eighth throws were worth $\frac{1}{6}\left[\left(\frac{5}{6}\right)^3+\left(\frac{5}{6}\right)^4+\left(\frac{5}{6}\right)^5+\left(\frac{5}{6}\right)^6+\left(\frac{5}{6}\right)^7\right]$.

Now, to keep the reader from going astray, let it be said that this expression *is* the value of those throws *before* any throws have been made. If, after the bets are made but before any throws of the die have been made, the bet is changed and the players agree that only three throws shall be made, then the player holding the die should take this amount as compensation for sacrificing the last five throws. However, Fermat saw clearly that if the bet were to be changed *after* three unsuccessful throws, the matter was different. For at this point the player's probability of winning on the fourth throw is one-sixth. [Before the game started the player's probability of winning on the *fourth* throw was $\frac{1}{6}\left(\frac{5}{6}\right)^3$. This number is smaller by a factor $\left(\frac{5}{6}\right)^3$ representing the probability that the player will not win the game on the first three throws.] Fermat expressed the matter as follows:

> ...the three first throws having gained nothing for the player who holds the die, the total sum thus remaining at stake, he who holds the die and who agrees not to play his fourth throw should take $\frac{1}{6}$ as his reward. And if he has played four throws without finding the desired point and if they agree that he shall not play the fifth time, he will, nevertheless, have $\frac{1}{6}$ of the total for his share. Since the whole sum stays in play it not only follows from the theory, but it is indeed common sense that each throw should be of equal value...

Pascal immediately wrote back to Fermat, proclaiming himself satisfied with Fermat's analysis and overjoyed to find that "the truth is the same at Toulouse and at Paris." In the course of this correspondence Pascal and Fermat both realized that the combinatorial coefficients that occur in the arithmetical triangle (Pascal's triangle) play an important role in the computation of probabilities when dealing with equally likely events and repeated independent trials.

16.2.2 Christiaan Huygens

A treatise on probability was written in 1657 by the Dutch mathematician and scientist Christiaan Huygens (1629–1695). Next to Newton, Huygens was the greatest scientist of his era, and we shall study a few of his contributions to physics in a later chapter. He was a man of great breadth, who wrote a treatise on music and showed considerable talent as an artist. Huygens' tract *De ratiociniis in ludo aleae* (On Reasoning in a Dice Game) was a compendium of the results of Fermat and Pascal. Huygens, however, was able to consider multinomial problems, involving three or more players, to which Pascal's triangle did not apply (he used a recursive procedure on the number of players). The idea of an *estimate of the expectation* is due to Huygens.

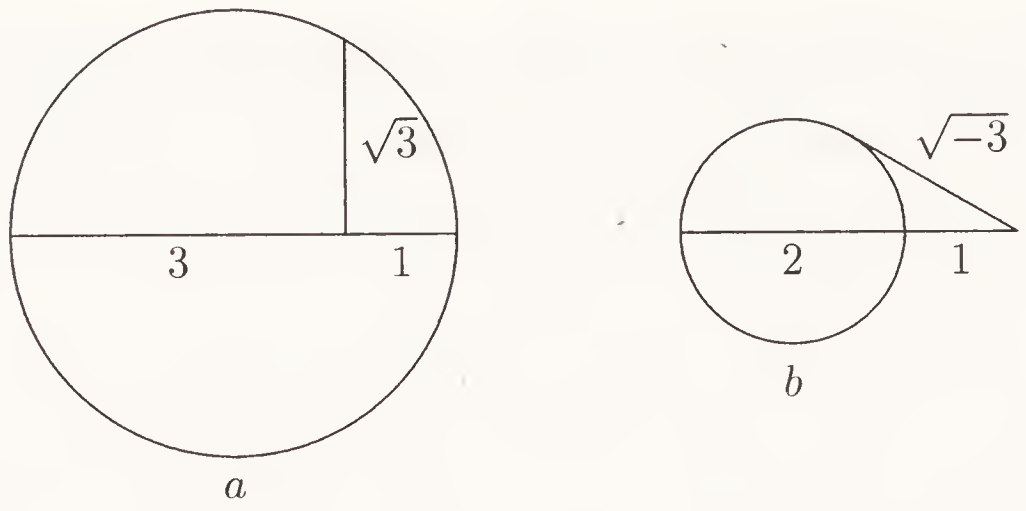


Figure 16.3: (a) Mean proportional between 3 and 1; (b) mean proportional between 3 and -1.

16.3 Algebra

16.3.1 Relations between Coefficients and Roots

The relation between coefficients and roots in an algebraic equation became clearer during the seventeenth century as a result of the work of several mathematicians. We mention only Albert Girard (1595–1632), a Frenchman who served the Prince of Orange as a military engineer. Though mostly concerned with equations having integer coefficients, Girard stated correctly that an equation $x^n + a_{n-2}x^{n-2} + \cdots = a_{n-1}x^{n-1} + a_{n-3}x^{n-3} + \cdots$ would have n roots, and that the sum of the roots would be a_{n-1}, the sum of the products taken two at a time would be a_{n-2}, etc. Girard was not in a position to prove this assertion, and his language hints that there can be exceptions if any of the coefficients equal zero. Nevertheless he stated a plausible conjecture that remained for later generations to prove. The coefficients a_k, which Girard called *factions*, are now called elementary symmetric polynomials. When $n = 3$, the equation has roots r_1, r_2, r_3, $a_2 = r_1 + r_2 + r_3$, $a_1 = r_1r_2 + r_1r_3 + r_2r_3$, and $a_0 = r_1r_2r_3$. Girard's work led to the view that the problem of solving an equation is the problem of expressing variables $x, y, z, \ldots$ in terms of the symmetric functions of these variables. For example, in the case of two variables with $x \geq y \geq 0$ we can express x and y in terms of the symmetric functions $x + y$ and xy as follows:

$$x = \frac{(x+y)}{2} + \sqrt{\frac{(x+y)^2}{4} - xy}$$

and

$$y = (x+y) - x = \frac{(x+y)}{2} - \sqrt{\frac{(x+y)^2}{2} - xy}.$$

Girard's discovery was duplicated by Newton and further developed to give upper and lower bounds on the real roots of an equation.

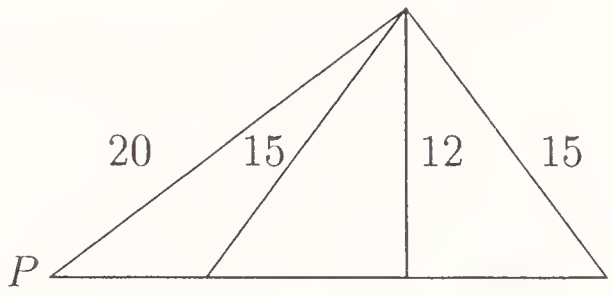


Figure 16.4: Wallis' geometric solution of a quadratic equation.

16.3.2 Imaginary Numbers

Girard's theorem is true only in the context of complex numbers, and so a full proof of it awaited a better understanding of these numbers. We have seen that Cardano and Bombelli were willing to consider such numbers, and developed some of their properties. Yet the logical foundation for such entities remained obscure. In an attempt to make these numbers more familiar, the English mathematician John Wallis pointed out that, while no positive or negative number could have a negative square, nevertheless it is also true that no physical quantity can be negative, that is, less than nothing. Yet negative numbers were accepted and interpreted as retreats when the numbers measure advances along a line. Wallis thought that what was allowed in lines might also apply to planes, pointing out that if 30 acres are reclaimed from the sea, and 40 acres are flooded, the net amount "gained" from the sea would be -10 acres. He proposed representing $\sqrt{-bc}$ as the mean proportional between $-b$ and c. Now the mean proportional is easily found for two positive line segments b and c. Simply lay them end to end, use the union as the diameter of a circle, and erect the sine to that diameter at the point where the two segments meet. When one of the numbers $(-b)$ was regarded as negative, Wallis regarded the negative quantity as an oppositely directed line segment. He then modified the construction of the mean proportion between the two segments. When two oppositely directed line segments are joined end to end, one end of the shorter segment lies between the point where the two segments meet and the other end of the longer segment, so that the point where the segments meet lies outside the circle passing through the other two endpoints. Wallis interpreted the mean proportional as the tangent to the circle from the point where the two segments meet. Thus whereas the mean proportional between two positive quantities is represented as a sine, that between a positive and negative quantity is represented as a tangent. (See Fig. 16.3.)

Wallis applied this procedure in an analogous "imaginary" construction problem. First he stated the following "real" problem. Given an isosceles triangle whose equal sides are 15 units long and in which the altitude to the base is 12, let P be a point on the extension of the base at distance 20 from the vertex of the triangle (see Fig. 16.4). How far is P from each endpoint of the base? Using the midpoint of the base as a reference point and applying the Pythagorean theorem twice, one easily expresses these numbers as 16 ± 9, that is, 7 and 25. This construction is the well-known method of solving quadratic equations geometrically, given earlier by Descartes. Wallis pointed out that this construction always works

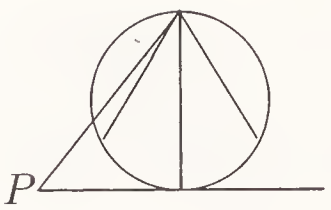


Figure 16.5: Wallis' solution of a quadratic with complex roots.

when the roots are real, whether positive or negative. He then proposed reversing the data, in effect considering an impossible isosceles triangle with equal sides 12 units long and altitude to the base equal to 15. This "imaginary" triangle was built around the altitude 15. The "base" was taken to be the line perpendicular to the altitude at one of its ends, and the equal sides were drawn as chords 12 units long in a circle having the altitude as diameter. Wallis pointed out that, although the algebraic problem has no real solution, a fact verified by the geometric figure (see Fig. 16.5), nevertheless one could certainly draw the line segments from the point P to the two vertices. These line segments could therefore be interpreted as solutions of the equation. This was the first realization that complex numbers would have to be interpreted as line segments in a plane, a discovery made again a century later by the Norwegian surveyor Caspar Wessel (1745–1818).

16.4 Number Theory

Number theory was another topic that engaged the mathematicians of the early modern era. In this area also they soon discovered that they could go beyond the ancients. We have already remarked that Fermat wrote in the margin of his copy of Diophantus that the sum of two positive rational cubes could not be a rational cube, and so on (Fermat's last theorem). Although Fermat never communicated what he believed his proof of this fact to be, he did devise a method of proof—the method of infinite descent—by which many facts in number theory can be proved, including the case of cubes and fourth powers in Fermat's last theorem. The basis of the method can be explained in the abstract by sketching a proof of the case of a fourth power. Actually the proof shows that there can be no positive integers x, y, z, such that $x^4 + y^4 = z^2$. Supposing that such numbers do exist, we assume that z is the smallest positive integer for which there exist positive integers x and y satisfying this equation. Then no two of x, y, and z have a common prime factor (otherwise the fourth power of this factor could be divided out, producing a smaller z). This means that two of the numbers are odd and one is even, and in particular that z is odd (since the square of an even number is divisible by 4, but the square of an odd number leaves a remainder of 1 when divided by 4). Assume that x is odd and y is even. Then $x^4 = (z + y^2)(z - y^2)$. Since z and y have no common factor, it follows that $z + y^2$ and $z - y^2$ also have no common factor. Since they are relatively prime and their product is a fourth power, each of the factors is a fourth power, that is, there exist (odd) integers u, v, such that $z - y^2 = u^4$, $z + y^2 = v^4$, and $uv = x$. Now $(v^2 - u^2)(v^2 + u^2) = v^4 - u^4 = 2y^2$,

and since $v^2 - u^2$ and $v^2 + u^2$ have no common prime factor except 2, there exists a factorization of y ($y = \omega\zeta$) such that either

$$\begin{aligned} v^2 + u^2 &= \zeta^2 \\ v^2 - u^2 &= 2\omega^2 \end{aligned}$$

or

$$\begin{aligned} u^2 + \omega^2 &= v^2 \\ u^2 + v^2 &= 2\zeta^2 \end{aligned}$$

The first possibility can be ruled out, since the sum of two odd square integers cannot be a perfect square. It then follows from the first equation in the second pair that u, ω, and v form a Pythagorean triple, and as Exercise 16.4 shows, this means there exist integers ξ and η such that $u = \xi^2 - \eta^2$, $\omega = 2\xi\eta$, and $v = \xi^2 + \eta^2$. Then the second equation says that $\xi^4 + \eta^4 = \zeta^2$. Since $\zeta < y < z$, this contradicts the assumed minimality of z.

16.5 Combinatorics

Leibniz, whose interest in Oriental cultures has already been mentioned, was the first to introduce into European mathematics certain topics that had been studied earlier in Asia. There is no indication that Leibniz knew these topics had been studied in Asia, but his mind coincidentally worked along the same channels as the earlier Asian mathematicians. We have already mentioned his introduction of determinants. We now come to a topic pioneered by the Hindu mathematicians: combinatorics. We have seen that Hindu mathematicians had computed the number of different kinds of lines of poetry that could be formed with a given number of stressed and unstressed syllables. This kind of problem was the origin of the modern subject of combinatorics, which has found numerous applications. A major impetus to such studies was Leibniz' publication of *De arte combinatoria* in 1666. In this work Leibniz gave tables of the number of permutations of n objects. There are many very curious aspects of this work. Although written mostly in Latin, it is rather polyglot; Leibniz occasionally and unaccountably breaks into Greek or German, and the tables are labeled with Hebrew letters. For permutations Leibniz used the word *numerus* to denote the size of the set from which objects are chosen, and *exponent* (literally *placing out*) for the number of objects chosen. The total number of permutations of a number of objects he called its *variationes*, and for the number of combinations of a set of objects taken, say, 4 at a time, he wrote *con4natio*, an abbreviation for *conquattuornatio*. (The case of 2 provides the modern word *combination*.) These combinations, now called binomial coefficients, were referred to generically as *complexiones*. Leibniz' first problem was *given the numerus and the exponent, find the complexiones*. In other words, given n and k, find the number of combinations of n things taken k at a time.

It is interesting that Leibniz, like the Hindu mathematicians before him, applied combinatorics to poetry and music. He considered, for instance, the hexameter lines possible with the Guido scale *ut, re, mi, fa, sol, la,* finding a total of 187,920.[1]

De arte combinatoria contains 12 sophisticated counting problems and a number of exotic applications of the counting techniques. It appears that Leibniz intended these techniques to be a source by which all possible propositions about the world could be generated. Then, combined with a good logic-checker, this technique would provide the key to all knowledge. His intent was philosophical as well as mathematical, as evidenced by his claimed mathematical proof of the existence of God at the beginning of the work.

16.6 Computing Machines

The graphic arithmetic that had vanquished the counting board a few centuries earlier still had certain laborious aspects connected with multiplication and division, which mathematicians kept trying to simplify. We have already seen two efforts in this direction, the use of prosthapheresis and the invention of logarithms. The fact that logarithms change multiplication into addition and that addition can be performed mechanically by sliding one ruler along another led to the development of rulers with the numbers arranged in proportion to their logarithms (slide rules). Napier himself designed a system of rods for this purpose. This linear system was soon supplemented by a system of sliding circles. Such a circular slide rule was described in a pamphlet entitled *Grammelogia* written in 1630 by Richard Delamain, a mathematics teacher living in London. Delamain urged the use of this device on the grounds that it made it easy to compute compound interest. Two years later the English clergyman William Oughtred (1574–1660) produced a similar description of a more complex device. Oughtred's "circles of proportion," as he called them, gave sines and tangents of angles in various ranges on eight different circles.

Machines for performing addition mechanically are easy to design, but multiplication and division require more sophistication. A machine for performing the operations of arithmetic was designed by Pascal. (A version of the machine built in 1652 and signed by Pascal is in the Conservatoire des Arts et Métiers in Paris.) Actually Pascal was not the first to design a calculating machine. Such a machine had been designed in the year Pascal was born by a Tübingen professor of mathematics and astronomy named Wilhelm Schickard (1592–1635), who wrote to Kepler about his discovery. Schickard's machine could also do multiplication, although it was necessary for the operator to do some counting.

The basic principle of such machines was a set of gear teeth, 10 around each circle, with every tenth gear tooth longer than the others, so that it would engage the adjacent gear once on each rotation and advance it one-tenth of a turn. If each gear is attached to a calibrated plate, the plates read off the decimal digits of a

[1]The first five of these tones are the first syllables of a Medieval Latin chant on ascending tones. The replacement of *ut* by the modern *do* came later.

number. The long tooth then automatically carries or borrows when performing addition and subtraction. The mechanical problem to be overcome in designing such a machine is to reduce the wear from the unbalanced load that results from the carrying operation. Pascal's machine was designed with counterweights that are raised higher and higher as each gear records larger numbers, then drop back when the gear goes full circle.

Pascal's machine was improved on by Leibniz, who gave a design for a machine that would multiply and divide more efficiently. Pascal's machine had focused on the idea of addition. Leibniz' machine consisted of a set of gears, each of which meshed with an identical gear attached to a dial indicating a digit of the number to be multiplied. To use it a "multiplier box" was inserted whose gears meshed with smaller gears attached to the digits of the number to be multiplied. The gears in the multiplier box had diameters equal to 2, 3, 4, 5,... times the size of the given gears. Then if, say, the "4" gear were turned once, the gears representing the number to be multiplied would each turn four times, and the long teeth on these gears would trip the counters of the registry gears four times. The multiplier box could be removed and shifted so that the number would be multiplied by 40 when the "4" gear was rotated. By 1674, with the help of a young French metalworker, Leibniz had produced a practical mechanical model of this machine. It is worth pointing out that here, at the very beginning of the history of the computer, it was already influencing mathematics: it was through designing such a machine that Leibniz was led to wonder if a machine could be made to check logic as well.

16.7 Societies and Journals

Science became a societal enterprise during the seventeenth century. Before that time mathematicians, like other scholars, had supported themselves or found a patron. During the seventeenth century scholars with a common interest formed their own societies such as the Accademia dei Lincei in Italy and a group in correspondence with Mersenne in Paris. Scholarly activity began to be concentrated in universities, and the major contributors to mathematics came more and more often from the ranks of professors. In this era of monarchies kings and queens became patrons of science partly to enhance the prestige of their realms and partly because of the economic and military value of the inventions that scholarship could produce. The Royal Society in London was formed in 1662, the Academy of Sciences of Paris in 1666. The Academy of Sciences of Berlin was founded under Friedrich Wilhelm in 1714. (Leibniz had been instrumental in the founding of the Berlin Scientific Society in 1700.) Tsar Peter I chartered the St. Petersburg Academy in 1724.

The network of correspondence was gradually replaced by a system of journals. The minutes of the meetings of the societies became the first outlets for research papers. The journals provided the same public audience as for a book, but in a format suitable for writing about smaller parts of a subject than one would cover in a treatise. The *Acta Eruditorum*, for example, was founded in Leipzig in 1682 and immediately became an outlet for the work of Leibniz and the Bernoullis.

16.8 Problems and Questions

16.8.1 Problems from the Seventeenth Century

Exercise 16.1 Prove that if two pairs of opposite sides of a hexagon inscribed in a circle are parallel, then the third pair of sides is also parallel. (Hence the points of intersection of all three pairs of opposite sides belong to a single line, namely the line at infinity.)

Exercise 16.2 Draw four rays emanating from a single point, and draw two lines, one intersecting the four rays in points A, B, C, D (in order) and the other intersecting them in A', B', C', D'. Prove that the cross ratios of the four points are equal:

$$\frac{AC \cdot BD}{BC \cdot AD} = \frac{A'C' \cdot BD''}{B'C' \cdot A'D'}.$$

Exercise 16.3 Leibniz gave a determinant condition on the numbers a_i, b_i, c_i that was necessary and sufficient for the equations

$$\begin{aligned} a_1 + b_1 x + c_1 y &= 0 \\ a_2 + b_2 x + c_2 y &= 0 \\ a_3 + b_3 x + c_3 y &= 0 \end{aligned}$$

to have a solution. What is this condition?

Exercise 16.4 Prove that if x, y, and z are relatively prime integers such that $x^2 + y^2 = z^2$, with x and z odd and y even, then there exist integers u and v such that $x = u^2 - v^2$, $y = 2uv$, and $z = u^2 + v^2$. (Imitate the arguments in the text above.)

Exercise 16.5 Use the method of infinite descent to prove that $\sqrt{3}$ is irrational. [Assuming $m^2 = 3n^2$, where m and n are positive integers having no common factor, that is, they are as small as possible, verify that $(m - 3n)^2 = 3(m - n)^2$. Note that $m < 2n$ and hence $m - n < n$.]

Exercise 16.6 Show that $\sqrt[3]{3}$ is irrational, by assuming $m^3 = 3n^3$ with m and n positive integers having no common factor. [Show that $(m-n)(m^2+mn+n^2) = 2n^3$. Hence, if p is a prime factor of n, then p divides either $m-n$ or m^2+mn+n^2. In either case p must divide m. Since m and n have no common factor, it follows that $n = 1$.]

Exercise 16.7 Suppose that x, y, and z are positive integers, no two of which have a common factor, none of which is divisible by 3, and such that $x^3 + y^3 = z^3$. By reasoning as in the proof that the equation $x^4 + y^4 = z^2$ is impossible, show that there exist integers p, q, r, such that $z-x = p^3$, $z-y = q^3$, and $x+y = r^3$. Then, letting $m = r^3 - (p^3 + q^3)$ and $n = 2pqr$, verify from the original equation that $m^3 = 3n^3$, which by the previous exercise is impossible if m and n are nonzero. Hence $n = 0$, which means that $p = 0$ or $q = 0$, or $r = 0$, that is, one of x and y, and z equals 0. Conclude that no such positive numbers x, y, and z can exist.

16.8.2 Questions about Seventeenth-Century Mathematics

Exercise 16.8 Why is it not feasible simply to allow a gambler to interrupt a game, recover his/her stake, and leave? Could a real game take place if this were allowed?

Exercise 16.9 The amount of money Pascal would erroneously have allowed the shooter to reclaim for giving up the last five of eight attempts to make a point is *smaller* than the amount Fermat correctly allowed. The amount Pascal would have allowed is the correct amount only if the shooter agrees to give up those last five shots *before* making any shots. Yet after those shots have been made, it is the shooter's *opponent* who has withstood the risk of losing on the first three shots. Why should the opponent be willing to pay *more* for those shots *after* undergoing the risk of losing than *before*?

16.9 Endnotes

1. The material of this chapter, like that in the preceding chapter, is based on the following source books:

   (a) D. J. Struik, *A Source Book in Mathematics, 1200–1800* (Harvard University Press, 1969).

   (b) David Eugene Smith, *A Source Book in Mathematics* (Dover Reprint, New York, 1959).

   (c) Ronald Calinger, *Classics of Mathematics* (Prentice-Hall, Englewood Cliffs, NJ, 1995).

   (d) John Fauvel and Jeremy Gray, *The History of Mathematics. A Reader* (Macmillan, London, 1987).

2. The section on projective geometry is based on the article by Morris Kline, "Projective Geometry," *Scientific American* (Jan. 1955), reprinted in *Mathematics. An Introduction to its Spirit and Use* (Freeman, San Francisco, 1979).

3. The quotation from Pascal's *Essay pour les coniques* is taken from Smith's source book (op. cit.), p. 329.

4. The quotations from the correspondence of Pascal and Fermat on probability are taken from Smith's source book (op. cit.), pp. 546–565.

5. The discussion of Wallis' work on imaginary numbers is based on Smith's source book (op. cit.), pp. 46–54.

6. The discussion of the slide rule is based on Smith's source book (op. cit.), pp. 156–164.

7. The discussion of computing machines is based on the book *The Computer from Pascal to Von Neumann* by Herman H. Goldstine (Princeton University Press, 1972).

Chapter 17

Beyond the Calculus

Up to the seventeenth century mathematics looks like a number of small rivulets meandering here and there. The rapid expansion of discoveries in the seventeenth century caused those rivulets to swell into a wide river by the beginning of the eighteenth century. This growth continued, so that by the beginning of the twentieth century the river had become a mighty flood. To describe the history of the two centuries from the calculus to the many subject areas that make up twentieth-century mathematics, we shall trace the growth process in the area of calculus and then see how these new ideas fostered the development of other traditional areas such as geometry and algebra. We are about to embark on a quick tour of a large amount of material. We have three goals in mind. First, we shall try to trace the origins of the mathematics that forms most of the current undergraduate curriculum. Second, we wish to make a survey of the new mathematics created in the eighteenth and nineteenth centuries, whether or not it is currently taught to undergraduates, in order to give a general idea of what was done and why. Third, we shall try to trace as many interconnections as possible, to show that mathematics grew as a unified organism. Its roots are very far away from its branches, but the two are definitely connected. The main "trunk" that joins all these branches into a single organism is the calculus.

17.1 The Calculus and Its Outgrowths

17.1.1 Expositions of the Calculus

Most of what is now called calculus was invented in the last half of the seventeenth century and organized systematically in the early decades of the eighteenth. The importance of calculus in scientific research can be seen in the work of Newton and Leibniz and was amply demonstrated by the great eighteenth-century scientists, who laid down the principles that we now think of as "Newtonian" physics. A considerable amount of this work is due to one Swiss family, the Bernoullis. We shall have space here to discuss only four members of this illustrious family, the

brothers Jakob and Johann and Johann's sons Daniel (1700–1782) and Niklaus (1695–1726). It was Jakob who, in a 1690 paper on the isochrone problem,[1] introduced the term *integral* still used today. Leibniz had used the Latin word *omnis* for this idea of summing up all the infinitesimal changes in a variable.

In 1692–93, together with his younger brother Johann, Jakob Bernoulli studied caustics, the envelopes[2] of reflected or refracted systems of rays of light; this work led to the appearance of the now-familiar formula for the curvature of a plane curve. Many of the well-known curves of calculus, such as the catenary, the tractrix, and the lemniscate were first discussed in the works of Jakob Bernoulli in the early 1690s. He also introduced the use of polar coordinates. He was proudest of all of his exposition of the properties of the logarithmic spiral (in our terms the curve $r = ae^{b\theta}$), which tends to reproduce itself under many common geometric transformations. In particular, the caustics resulting from reflection or refraction by a logarithmic spiral are also logarithmic spirals. He is said to have asked that this curve be inscribed on his tombstone, reminiscent of Archimedes' sphere and cylinder.

As we saw in Chapter 15, both Newton and Leibniz knew a number of particular power series expansions. Their disciples, Brook Taylor and Johann Bernoulli, discovered the general procedure for generating such series representations in terms of the derivatives of the functions represented, that is, the series now known as Taylor series. Taylor published his main work, the *Methodus incrementorum directa et inversa*, a study of differential equations and plane curves, in 1715. It was in this work that the famous vibrating string problem was first posed. It also contained the power series expansion now named after Taylor, which he had discovered in 1712. Taylor's claim to this result was disputed by Johann Bernoulli, who had made the same discovery somewhat earlier.

The ratio test for convergence of a series is due to Jean le Rond d'Alembert (1717–1783), who also noted that the variable in a power series could be thought of as representing a complex number. The extension of the calculus to complex numbers turned out to have monumental importance. It was d'Alembert who first used complex numbers to give a nearly rigorous proof of the fundamental theorem of algebra: Every polynomial of positive degree with complex coefficients must be equal to zero for some complex value of the variable.

Most students of calculus know the Maclaurin series as a special case of the Taylor series. Its discoverer was a Scottish contemporary of Taylor, Colin Maclaurin (1698–1746), whose treatise on fluxions (1742) contained a thorough and rigorous exposition of calculus. It was written partly as a response to the philosophical attacks on the foundations of calculus by the philosopher George Berkeley.

The secure place of calculus in the mathematical curriculum was established by the publication of a number of excellent textbooks. One of the earliest, the *Analyse des infiniment petits*, was published by the Marquis de l'Hospital in 1696. (L'Hospital had become interested in calculus from reading the work of Leibniz and had taken instruction from Johann Bernoulli.)

---

[1]The problem of finding the path down which a frictionless particle will slide in a constant time independent of its starting point.

[2]The envelope of a family of curves or surfaces is a curve or surface tangent to all of its members.

The Italian textbook *Istituzioni analitiche ad uso della gioventù italiana* (Analytic Principles for the Use of Italian Youth) became a standard treatise on analytic geometry and calculus, and was translated into English in 1801. Its author was Maria Gaetana Agnesi (1718–1799), one of the first women to achieve prominence in mathematics. In 1750 she became Professor of Mathematics and Natural Philosophy at the University of Bologna, one of the oldest and most respected universities in Europe. This work contains a discussion of the curve with equation $x^2y = a^2(a - y)$, called by Agnesi the *versiera*, and through an unfortunate translation known in English as the *witch of Agnesi*. (The Italian word *versiera* does mean *she-devil*, but Agnesi was probably referring to its *twisted* character.)

The definitive textbooks of calculus were written by the greatest mathematician of the eighteenth century, the Swiss scholar Leonhard Euler (1707–1783). In his 1748 *Introductio in analysin infinitorum*, a two-volume work, Euler gave a thorough discussion of analytic geometry in two and three dimensions, infinite series (including the use of complex variables in such series), and the foundations of a systematic theory of algebraic functions. The modern presentation of trigonometry was established in this work. The *Introductio* was followed in 1755 by *Institutiones calculi differentialis* and a three-volume *Institutiones calculi integralis* (1768–1774), which included the whole theory of calculus and the elements of differential equations, richly illustrated with challenging examples. Modern calculus books essentially repeat what Euler said about differential equations. It was from Euler's textbooks that many prominent nineteenth-century mathematicians such as the Norwegian genius Niels Henrik Abel (1802–1829) first encountered higher mathematics, and the influence of Euler's books can be traced in their work.

More than anyone else Euler determined the general shape of eighteenth-century mathematics. He applied mathematics to shipbuilding, geodesy, astronomy, ballistics, optics, and a variety of other areas, always manifesting a keen physical intuition. He wrote beautiful expository treatises that established much of the notation we now use. For example, the use of the letter e to denote the base of natural logarithms first occurred in a paper written by Euler around 1728, which, however, was not published until 1862. Euler used this letter in a published work on mechanics in 1836, defining it as "the number whose hyperbolic logarithm equals unity."

As the case of Maria Gaetana Agnesi shows, during the eighteenth century a few women managed to break through the social barriers that had previously confined them to domestic activities. One of the first to do so, Gabrielli Émilie, Marquise du Châtelet (1706–1749), began by studying languages. At the urging of Voltaire she undertook to translate Newton's *Principia* into French. To do so meant having to understand and explain the most advanced mathematics and science of her time. This work was not published until 1756, 8 years after an early death deprived the world of the further contributions she might have made.

Foundational Questions

The textbooks just discussed were written partly to respond to objections to the calculus. The philosophical difficulties with the foundations of the calculus were

cogently urged in a treatise by the philosopher George Berkeley (1685–1753) entitled *The Analyst* (1734). Although Berkeley had pondered the unresolved questions on the relations between a line and its points—whether lines could be analyzed into infinitely small parts or synthesized from them—as early as 1710, the immediate impetus to the publication of this work was religious. Berkeley knew of a man who refused religious rites on his deathbed because he had been convinced that theological propositions were meaningless. In response Berkeley, who had just become the Anglican Bishop of Cloyne, Ireland, undertook to show that the propositions on which current mathematics was based were not the least bit clearer to reason than those of theology. His treatise was subtitled *Discourse Addressed to an Infidel Mathematician* (the unnamed infidel was the astronomer Edmund Halley). Berkeley attacked the infinitesimals at their weakest point, showing that they were inconsistently handled in some places as if they were zero, in other places as if they were finite numbers. In a famous phrase, he referred to the ratios of infinitesimals as "ghosts of departed quantities." Despite the vigor of his attack, however, Berkeley did not doubt or wish others to doubt the truth of mathematical results. His purpose was just the opposite, to show that such methods, albeit seemingly contradictory to human reason, yet led to true results, as (he believed) theology did.

The defense of calculus was led by Maclaurin, whose treatise on fluxions developed the subject as Newton had said it could be developed, in accordance with the ancient method of exhaustion.

The difficulties with the notion of instantaneous rate of change, infinitesimals, and the like, caused some mathematicians to look for other ways of deriving the results of calculus. In his textbooks entitled *Théorie des fonctions analytiques* (1797) and *Leçons sur le calcul des fonctions* (1801) the Italian-French mathematician Joseph-Louis Lagrange (1736–1813) undertook to reformulate the calculus, basing it entirely on algebraic principles and stating as a fundamental premise that the functions to be considered are those that can be expanded in power series. In these textbooks the form of the remainder in a Taylor series now called the *Lagrange form* was introduced. With this approach the derivatives of a function need not be defined as ratios of infinitesimals, since they can be defined in terms of the coefficients of the series that represents the function. Functions having a power series representation are known nowadays as *analytic functions* (from the title of Lagrange's work). They have the important property that if their values are known over any finite interval of variation of the independent variable, no matter how short, then the coefficients can all be computed, and hence the values of the function can be computed for all values of the independent variable. This principle corresponds to the metaphysical principle that perfect information about the motion of bodies for a finite interval of time would make it possible to predict the entire future course of those bodies.

Functions

Infinitesimals were not the only foundational question connected with the calculus. The meaning of the algebraic symbols used also raised certain questions. Physical

laws were usually expressed as equations containing two or more variables representing measurable quantities. The relations among variables were pictured as curves or surfaces.

Leibniz used the word *function* to denote the relation between a dependent variable and one or more independent variables. In ordinary language a function is an operation that one carries out, and that seems to be the idea Leibniz had in mind. In 1718 Johann Bernoulli gave a definition of this concept close to the modern one, writing that "A function of a variable is... a quantity formed in any manner from this variable and constants." The phrase "formed in any manner" leaves a great deal of room for ambiguity. The question of just which operations are admitted has been disputed ever since, with more and more general operations being allowed as time passes. In his 1748 *Introductio* Euler emended the definition, saying that a function is an *analytic expression* formed from a variable and constants. At the time the only analytic expressions allowed were finite algebraic and trigonometric expressions and infinite series of powers of a variable. Thus for the most part a function meant an algebraic function. The use of power series eventually introduced into mathematics huge new classes of functions, which could be adapted to solve particular problems.

The calculus presented a problem: the rules for manipulating the symbols were agreed on as long as only finite expressions were involved, but the question was, what did the symbols *represent?* Normally the letters x, y, z, were thought of as representing continuous quantities such as lines or ratios in geometry. However, if they were thought of as numbers, there was some question as to what sort of numbers they could be. Euler explicitly stated that variables were allowed to take on negative and imaginary values. Thus, even though the physical quantities the variables represented were measured as *positive rational* numbers, the algebraic and geometric properties of negative, irrational, and complex numbers could be invoked in the analysis.

17.1.2 Differential Equations

One of the most powerful tools that can be constructed out of the calculus is the use of differential equations. It is only a small exaggeration to say that the principal advantage of the differential calculus is that it makes it possible to write down differential equations, and the principal use of the integral calculus is in solving differential equations.

Closed-Form and Series Solutions

Two kinds of problems arise in the application of differential equations to physics. First, the equations have to be manipulated into a form in which the solution is merely a matter of integration. This step is called *reduction to quadrature.* Second, the integration has to be carried out. The first step cannot in general be performed; in particular it is impossible to carry it out in the case of the equations Euler derived to describe the motion of a rigid body and for the equations of the

three-body problem. For that reason these problems attracted a great deal of interest during the nineteenth century. As for the second step, some very simple problems—pendulum motion, for example—lead to equations requiring the integration of the square root of a cubic polynomial. In such a case the reduction to quadrature is possible, but the great variety of possible behaviors for cubic polynomials in two variables made it clear that no simple formula could be found to express such integrals. The development of differential equations began to repeat the history of algebraic equations, as the early "exact" methods of solution encountered insuperable difficulties when the expressions became complicated and had to be supplemented by approximate methods. Indeed, this analogy turned out to be very deep.

Thus it soon became apparent that the natural approach to solving differential equations—to find a "closed-form" solution by replacing the differential equation between the variables by an equivalent relation not involving any differentials—was limited to a few special cases, and these cases were not adequate for the problems in physics to which mathematicians wished to apply the method. Another early method (used by Newton, for example, in his *Fluxions*) is the so-called method of undetermined coefficients, in which a power series expansion is assumed for the variable occurring in the differential equation, and the equation itself is used to determine the coefficients of the series. This method turned out to be very fruitful, both practically and theoretically. The potential practical value was clear, but the question whether there exists a power series representing a solution of a given differential equation remained open. This question led to a great deal of research in the nineteenth and twentieth centuries, and occupied some of the best minds of the period. The use of power series was followed by the use of trigonometric series, and this technique eventually led to much of modern functional analysis. Maclaurin, however, warned against too hasty recourse to infinite series, saying that certain integrals could be better expressed geometrically as the arc lengths of various curves.

Geometric Approaches

The fact that families of curves and surfaces can be defined by a differential equation means that the equation can be studied geometrically in terms of these curves and surfaces. The curves involved, known as *characteristic curves*, are useful in deciding whether it is or is not possible to find a surface containing a given curve and satisfying a given differential equation. This geometric approach to differential equations was begun by Gaspard Monge (1746–1818), who also defined the principal curvatures of a surface at a point and the notion of lines of curvature. He wrote a definitive textbook on this subject entitled *Géométrie descriptive*.

Analytic Solutions

In the early 1820s Augustin-Louis Cauchy (1789–1856) drew attention to an unspoken assumption that mathematicians had been making: that there exists a solution

to a given differential equation. The technique for solving equations had been to reduce them to quadrature (evaluation of indefinite integrals) when possible. If this reduction could not be achieved, the solution was assumed to be a convergent power series, and the coefficients were then determined by substituting the series into the equation. It was Cauchy who first asked why there should be a power series solution and, in the 1820s, gave the first rigorous proof that an ordinary differential equation has a solution. In 1841 Cauchy developed what is known as the method of majorants for proving that a solution of a partial differential equation exists in the form of a power series in the independent variables. His technique was to replace the equation by another equation that generated a power series with larger coefficients than those generated by the given equation. If the new equation could be reduced to quadratures and its solution shown to be analytic, it followed that the formal power series for the original equation also converged and hence represented an actual solution. The method of majorants was developed independently by Karl Weierstrass (1815–1896) in that same year in application to a system of ordinary differential equations. Weierstrass' goal was somewhat different from Cauchy's, however; he wanted to show that the differential equation itself could be used as the definition of a function, even if the power series representing the function could not be completely determined.

Weierstrass did not publish his work until 1894, when his collected works began to be published, and Cauchy published so much material that his 1841 papers were not noticed when interest in this question was revived in the early 1870s. Weierstrass' student Sof'ya Kovalevskaya (1850–1891) applied the method of majorants and a normalization theorem of Carl Gustav Jacobi (1804–1851) to produce an exceedingly elegant theorem giving cases in which an analytic solution exists.[3] This theorem is still a centerpiece of the theory of differential equations today, and is known as the Cauchy–Kovalevskaya theorem. Moreover, Kovalevskaya went beyond the positive result and showed its limitations with a counterexample. Weierstrass had believed that any equation of mathematical physics could be solved by assuming a power series representation of the solution and finding the coefficients. Kovalevskaya astounded him by showing that such is not always the case. In fact, the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2},$$

which describes the temperature in a long thermally insulated wire, has an analytic solution if the initial temperature distribution is shaped like a normal probability curve, for example, $u(x,0) = e^{-x^2}$, but not if the initial temperature is the versiera of Agnesi $u(x,0) = 1/(1+x^2)$. Since the two curves look very much alike when x is regarded as a real variable, the difference must be sought in their different properties as functions of a complex variable. Thus complex numbers are relevant to the study of this equation, even though the imaginary part of the complex variable t seems to have no physical interpretation.

[3]Kovalevskaya's work was partly duplicated by Gaston Darboux (1842–1917). The publicity involved with sorting out priority claims between the two led to the discovery that some of this work had been done earlier by Cauchy.

Trigonometric Series Solutions

The use of trigonometric series rather than power series to solve differential equations began in the mideighteenth century. The first major problem to be attacked using this technique was the famous vibrating string problem, represented by the one-dimensional wave equation derived by d'Alembert in 1747:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},$$

in which u represents the vertical displacement of the vibrating string above the point x at time t. In 1749 Daniel Bernoulli found the solution as a series of terms of the form $u_n(x,t) = a \sin nx \sin nct$. Trigonometric functions really came into their own, however, in the work of Joseph Fourier (1768–1830) on heat conduction.

The general solution of the one-dimensional wave equation was obtained by d'Alembert in the form $u(x,t) = f(x+ct) + g(x-ct)$. This solution can be interpreted physically as a representation of the wave disturbance u as the superposition of a wave f traveling left with velocity c and a wave g traveling right with velocity c.

Sturm–Liouville Problems

In studying the action of gravity Pierre-Simon Laplace (1749–1827) was led to what is now known as Laplace's equation in three variables. The two-variable version of this equation is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

The operator on the left-hand side of this equation is known as the *Laplacian*. Since Laplace's equation can be thought of as the wave equation with velocity $c = \sqrt{-1}$, complex numbers again enter into a physical problem. Recalling d'Alembert's solution of the wave equation, Laplace suggested that the solutions of his equation might be sought in the form $f(x+y\sqrt{-1}) + g(x-y\sqrt{-1})$. Once again a problem that started out as a real-variable problem led inexorably to the need to study functions of a complex variable.

With the use of trigonometric series, which were particularly adapted to the solution of certain equations involving the Laplacian, mathematicians were encouraged to look for other simple functions in terms of which solutions of more general differential equations could be expressed. Between 1836 and 1838 this problem was attacked by Charles Sturm (1803–1855) and Joseph Liouville (1809–1882), who considered general second-order differential equations of the form

$$[p(x)y'(x)]' + [\lambda r(x) + q(x)]y(x) = 0.$$

When a solution of Laplace's equation is sought in the form of a product of functions of one variable, the result is often an equation of this type for the one-variable functions. It often happens that only isolated values of λ yield solutions

satisfying given boundary conditions. Sturm and Liouville found that in general there will be an infinite set of values $\lambda = \lambda_n$, $n = 1, 2, \ldots$, satisfying the equation and a pair of conditions at the endpoints of an interval $[a, b]$, and that these values increase to infinity. The values can be arranged so that the corresponding solutions $y_n(x)$ have exactly n zeros in $[a, b]$, and any solution of the differential equation can be expressed as a series

$$y(x) = \sum_{n=1}^{\infty} c_n y_n(x).$$

The sense in which such series converge was still not clear, but it continued to be studied by other mathematicians, It required some decades for all these ideas to be sorted out clearly.

17.1.3 Calculus of Variations

Newton's formulation of his mechanics was not in the simplest and most polished form. In particular, although the important law $F = ma$ was stated by d'Alembert and Euler in the form of differential equations that explain large numbers of phenomena, the solutions of the equations came in a form that sometimes concealed some basic physical principles. In the middle of the eighteenth century the mathematician/philosopher Pierre de Maupertuis (1698–1759) stated a fundamental principle known as the *principle of least action*, as a guide to the behavior of the universe. This principle was also formulated by Euler in a way that made it useful in physics and mathematics. To explain it, we need to recall certain basic problems.

Many important questions in geometry and mechanics involve minimizing or maximizing not the value of a variable, but some quantity depending on the whole set of values of a variable. For example, given two rings of different sizes in space, what is the surface having them as boundary that has least area? Or, given a fixed area in a plane to be enclosed, what is the shortest curve that will enclose the required amount of area? In such questions the unknown is not a number but a functional relation. One such problem that appeared in Newton's *Principia* is that of choosing an optimally streamlined surface for a body moving through a fluid (Scholium to Theorem XXVIII in Book II). Newton was unable to solve the general problem, but could solve it within restricted classes of surfaces, such as paraboloids of revolution or frusta of cones. A similar problem, known as the *brachistochrone problem*, which involves finding the path down which a frictionless particle will slide in minimal time from one point to another under the influence of gravity, provoked some rivalry and ill-will between Johann and Jakob Bernoulli. Johann Bernoulli solved it by appealing to the least-time principle for the path of a light ray, from which the law of refraction could be derived; essentially he considered a ray of light moving in a medium in which the index of refraction is proportional to the square root of its elevation. The solution is an inverted cycloid.

In a 1744 paper entitled "Curvarum maximi minimive proprietate gaudentium inventio nova et facilis" (A new and easy way of finding curves satisfying a

maximal or minimal property) Euler solved the problem of minimizing a functional of the form $\int Z\,dx$, where Z is a function of x, y, p, q, r, etc., and the integral is evaluated with y regarded as a function of x and $p = (dy/dx)$, $q = (dp/dx) = (d^2y/dx^2)$, $r = (dq/dx) = (d^3y/dx^3)$, etc. Euler's solution reduced this problem to the differential equation

$$N - \frac{dP}{dx} + \frac{d^2Q}{dx^2} - \frac{d^3R}{dx^3} + \cdots = 0,$$

where

$$dZ = M\,dx + N\,dy + P\,dp + Q\,dq + R\,dr + \cdots,$$

In modern terms, when Z is independent of q, r, etc., this equation is written

$$\frac{d}{dx}\left(\frac{\partial Z}{\partial y'}\right) = \frac{\partial Z}{\partial y}$$

and is known as *Euler's equation.* This second-order differential equation gives only a necessary condition that the minimizing function $y(x)$ must satisfy, but usually its solutions are restricted enough that one need not look further for the solution. As an application Euler showed that one could calculate the trajectory of a body moving under a central force using this equation to minimize the integral of its velocity with respect to arc length, and that the result was the same as that obtained by Newtonian methods.

Fifteen years later Lagrange put the theory on a more systematic basis by introducing the concept of variation of a curve, analogous to the differential of a variable in calculus. If $y = y(x)$ is a curve, its variation is thought of as a small increment in y (depending on x), that is, the difference between $y(x)$ and a nearby curve, and denoted δy. The corresponding variation of the integral $I(y) = \int_a^b F(x, y, y')\,dx$ is the linear part of the actual increment when $I(y+\delta y)$ is expanded in a power series in δy. It is not difficult to see through integration by parts that

$$\delta\left(\int_a^b F(x, y, y')\,dy\right) = \int_a^b \left[\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) - \frac{\partial F}{\partial y}\right]\delta y\,dx.$$

Thus, arguing that this last expression must be zero for all δy if the integral has a minimum at y, Lagrange deduced Euler's equation.

Euler and Lagrange made the calculus of variations fundamental in mechanics by formulating what Euler called the "law of rest." That is, if the forces acting on a system of particles during a physical process are integrated with respect to distance (thereby producing what is now called the work done on the body and Euler called the *effort*), the path actually followed by the body will minimize this integral. Since this work equals the change in potential energy of the body, Euler's statement amounts to the claim that a body always moves to a state of minimum potential with respect to a given set of forces.

17.1.4 Analysis

In the nineteenth century the calculus continued to grow through generalizations of its methods and consolidation of its foundations. Both of these directions contributed to the development of calculus into what is now called analysis, a large set of topics grouped around two centers called *real analysis* and *complex analysis*. It has turned out that the processes of calculus—differentiation, integration, sequences, and series—mean different things when applied to real and complex numbers. Roughly speaking, functions of a complex variable tend to possess great regularity while functions of a real variable often exhibit pathological irregularity. Both have important applications, and there are many bridges between the two subjects. They function together extremely well in the subject known as functional analysis.

The Bifurcation of Analysis

Real analysis began its growth as an independent subject with the introduction of the modern definition of continuity in 1816 by the Czech mathematician Bernard Bolzano (1781–1848). Bolzano deduced what is now known as the *intermediate-value theorem* from this new definition of continuity. That is, if a real-valued function is continuous and takes on a negative value at one point and a positive value at a second point, then there must be some point in between where it is zero. To prove this theorem he established one of the fundamental facts about the real numbers: a bounded infinite set of real numbers must contain a convergent sequence of distinct numbers. (This result is now known as the Bolzano–Weierstrass theorem.) Bolzano was not well known in his own time, and the eventual establishment of these ideas was due to the Cauchy. Before this time mathematicians had used the word *continuous* to refer to a function given by a single analytic formula throughout its domain, as opposed to a "discontinuous" function, defined by different formulas in different places. The latter may well be continuous in the modern sense. The relation between continuity and the derivative remained mysterious for many years after this time, although Bolzano had shown that a function can be continuous even when there is no interval throughout which it is differentiable. (His paper on this subject unfortunately was not published until the twentieth century.) Eventually it was realized that functions representable by power series (Taylor's series) are differentiable any number of times and have natural extensions via the power series to functions of a complex variable. Cauchy discovered the interesting fact that the region of convergence of a complex power series is either a single point, or all the complex numbers, or a disk together with possibly some or all of its boundary. Because the function represented by a power series is necessarily "smooth," the complex variable turns out to be of limited use for "rougher" functions.

Cauchy greatly advanced the subject of complex analysis in 1825, when he introduced the notion of an integral along a contour in the complex plane. From this idea he discovered that a function having a continuous derivative in the complex

sense has a representation as an integral (the Cauchy integral formula):

$$f(z) = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{\zeta - z}\, d\zeta,$$

and from that integral the Taylor series of the function can be generated and proved to converge. In 1900 Edouard Goursat (1858–1936) showed that the assumption that the derivative is continuous was unnecessary.

From this time on analysis developed along two diverging lines. Integration, differentiation, and series representations were the heart of both real and complex analysis, but real analysis became concerned with trying to find more and more general functions to which integration was applicable; differentiation played a definitely subordinate role. For complex analysis the Cauchy integral was adequate, and there was no possibility of finding any more general functions than power series.

Real Analysis

While complex analysis involves power series, much of real analysis is connected with series of trigonometric functions. In 1807 Joseph Fourier (1768–1830) singled out the natural trigonometric series to represent a function (for convenience we consider only even functions):

$$f(x) \sim \sum_{k=0}^{\infty} b_k \cos kx.$$

If the series does converge nicely to the function $f(x)$, that is, if it can be multiplied by $\cos mx$ and integrated term by term, it is easy to see that b_m must be given for $m > 0$ by

$$b_m = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos mx\, dx.$$

The series with coefficients computed from this formula is known as the *Fourier series* of the function $f(x)$; and since mathematicians of the time hardly considered the possibility that a series could *not* be integrated term by term, the Fourier series was the only trigonometric series that was considered for representing a function. The question was whether it converged or not.

This question was studied by Peter Lejeune-Dirichlet (1805–1859), who proceeded from the intuitive consideration that each trigonometric function has a limited number of intervals on which it is increasing and decreasing. He showed that the series must converge to the function that generates it if the latter has only a finite number of maxima and minima and only a finite number of discontinuities. Moreover at a discontinuity the series converges to the average of the right- and left-hand limits of the generating function.

Dirichlet had started from the function being represented in order to prove that the Fourier series converged to it. The opposite question—starting from a convergent series, what can one say about its sum?—had been considered by Cauchy,

who claimed that the sum of a series of continuous functions was continuous. Abel, who admired Cauchy, remarked diplomatically in one of his papers that, "It appears to me that this theorem suffers exceptions." He proceeded to point out the example of the series

$$\sum_{k=1}^{\infty}(-1)^{k-1}\frac{\sin kx}{k} = \begin{cases} \frac{x}{2} & \text{if } 0 < x < \pi, \\ \frac{x}{2} - \pi, & \text{if } \pi < x < 2\pi. \end{cases}$$

Since the sum of the series is 0 when x is a multiple of π, the sum certainly cannot be continuous at those points. In noting this fact and giving a rigorous discussion of the convergence of power series, Abel was leading the way to the notion of uniform convergence, which is crucial for the preservation of continuity and for the justification of the termwise operations of differentiation and integration performed in analysis.

The question of the possible values a convergent series of trigonometric functions can have was raised by Bernhard Riemann (1826–1866). In order to answer this question he was forced to examine the concept of integration, creating thereby (in just three pages) the concept now known as Riemann integration. Riemann gave a necessary and sufficient condition for a bounded function to be integrable over an interval $[a, b]$: The function must have the property that for any $\varepsilon > 0$ there is a number $\delta > 0$ such that the total length of the intervals on which the function oscillates by more than ε in a partition of $[a, b]$ into intervals of length less than δ is less than ε. This condition is now expressed by saying that the set of discontinuities of the function must have measure zero.

It is a revealing comment on the wealth of talent that existed in Europe by this time that such brilliant work of Riemann's did not immediately become the starting point for fresh research. Riemann's work on trigonometric series was his *Probevorlesung*, the traditional lecture given on assuming a new post in a German university (in Riemann's case, at Göttingen in 1854). Riemann himself did not follow up on this work, and it was not published until 1867, the year after he died.

Algebraic Functions: Abelian Integrals

With calculus splitting into real and complex halves, the question of the appropriate mathematical entities for studying physical phenomena became important. For physical applications, in which the variables represent time and space, it seems natural to use functions of a real variable, but it often happens that the mathematics is clearer when placed in the context of complex analysis. Such was the case with the algebraic integrals now known as Abelian integrals. These are integrals of the form $\int R(x, y)\,dx$, where $R(x, y)$ is a rational function of two variables x and y constrained by a polynomial equation $p(x, y) = 0$. (A function y satisfying such an equation is said to be an *algebraic function* of x.) A simple example is the integral $\int \sqrt{1 - x^2}\,dx$, which is $\int y\,dx$, where $x^2 + y^2 = 1$. The most important Abelian integrals in Abel's time were the elliptic integrals, for example, $\int R(x, y)\,dx$, where $y^2 = x^3 + ax^2 + bx + c$. These integrals had been studied in

minute detail in a three-volume treatise by Adrien-Marie Legendre (1752–1833), who had organized them into three distinct classes according to their behavior and had shown a wealth of applications of them in physics. (They arise naturally in the equation of pendulum motion $\ddot{\theta} + \sin\theta = 0$.) He had also noticed their analogy with the trigonometric functions and on that basis had suggested that their inverses would have a simpler theory than the integrals themselves.

The inverse function of such an integral is called an *elliptic function*. Algebraic functions and their integrals were among the leading motives for creating the theory of functions of a complex variable. As already pointed out, the techniques of that theory, especially the use of power series, automatically generated a huge class of nonalgebraic (transcendental) functions to which the same techniques apply, thereby providing additional problem-solving potential at little extra cost.

When elliptic integrals are regarded as functions of complex variables, their inverse functions have the important property of double periodicity. Abel had developed a general theory of algebraic integrals along these lines. His great paper submitted to the Paris Academy of Sciences in 1827, however, was still lost at the time of his death. A shortened version of this paper that appeared in 1829 attracted the attention of Abel's rival for the honor of creating the theory of elliptic functions, Carl Gustav Jacobi.

Abel had shown that a sum of any number of definite integrals of an algebraic integrand could be reduced to the sum of a fixed number p of integrals. The limits of integration on the p integrals would be algebraic functions of the limits of the original integrals. For the case when $p > 1$, Jacobi realized that this theorem introduced some indeterminacy, since there would be two or more integrals (hence two or more upper limits of integration) to be determined from only one equation. To make the problem determinate, Jacobi introduced in 1832 a set of p independent integrands and posed the problem of finding the limits of integration simultaneously for all p integrands. This problem, known as the Jacobi inversion problem, was an open question for 25 years. Jacobi appealed for the publication of Abel's memoir, which was finally located and published in 1841.

Jacobi also discovered the tool that eventually solved the inversion problem, known as *theta functions*. These are series of functions of the form $\theta(x) = e^{-x^2-2ax}$. Theta functions can be represented by power series that converge extremely well, and quotients of them can be doubly periodic. They are thus ideally adapted for representing elliptic functions. This result led Jacobi to examine places where elliptic functions had occurred in physics and to show that the quantities involved could be very naturally expressed by theta functions. In particular, in 1849 he solved the problem of the rotation of a rigid body free of external forces using these functions.

Like many other analysts of the nineteenth century, Weierstrass made important contributions to both real and complex analysis. His elegant derivation of representations of elliptic functions from their periodicity properties is still taught today. In complex analysis he was a champion of the power series as the basic tool, on the grounds that differentiability was too vague a property to base a theory on. His idea was to start at any point with a convergent power series. To break out of the

circle of convergence for the series, the series itself could be used to compute the coefficients of another series representing the same function at a point near the boundary circle, and the circle of convergence for the new series would normally extend outside the original circle. In this way he obtained a chain of circles leading from any point in the domain of the function to any other point. Each power series in the chain was called an *element* of the given analytic function. This process is called *analytic continuation.*

A large amount of Weierstrass' work was devoted to clarifying the properties of algebraic functions of a complex variable. For such functions there are points (such as the point $z = 0$ in the case of a function w satisfying $w^3 - z = 0$ for example) at which no power series expansion in integer powers of z is possible. At such points Weierstrass gave a set of expansions in powers of $\sqrt[3]{z}$. Algebraic functions were his main interest, and the fact that the complex function theory he developed turned out to apply to transcendental functions as well was a bonus. Weierstrass had worked out a general solution to the Jacobi inversion problem and published part of it, but withdrew his second paper when Riemann published another solution based on an entirely different approach to complex analysis.

Riemann's most important contribution to complex analysis was the idea of a Riemann surface.[4] This idea can be illustrated with an example.

For most algebraic functions there was a difficulty with finding inverse functions, which were not uniquely determined, that is, many different values of one variable corresponded to a single value of the other. A single example will suffice. Everyone knows that there are two values of $\sqrt{z}$ and these values are negatives of each other. If z starts at 1 and traverses a circle in the complex plane with center at 0, the square root varying continuously with z, it will be found that when z again approaches 1 after making one circuit around the point $z = 0$, its square root approaches the negative of the value it had when starting. Thus $\sqrt{z}$ cannot be defined as a continuous function in a neighborhood of 0. Riemann's idea was to have two copies of the z plane associated with one copy of the w plane in the relation $w = \sqrt{z}$, or, equivalently $z = w^2$. Each copy of the z plane is cut along a ray starting at 0 (since there is only one value of $\sqrt{0}$) and the edges of the two planes are glued together so that z passes from one plane to the other each time it crosses the ray. In this way the correspondence between z and w is one-to-one, and the square root becomes a continuous function. Cauchy's contour integrals can be computed on the Riemann surface as easily as in the ordinary plane, and a great many difficulties are thereby cleared up.

Each functional relation has its own Riemann surface, and Riemann showed how to make a certain number of cuts in a Riemann surface so that it becomes simply connected, that is, so that any closed curve can be smoothly shrunk to a point. In so doing he created one of the sources of the subject known as algebraic topology; for the number of cuts of different kinds that it was necessary to make

[4]Some of Riemann's dissertation was anticipated by a paper of Victor Puiseux (1820–1883) published in 1850, the year before Riemann wrote his dissertation. What Puiseux lacked was the notion of a branch line connecting different sheets of an algebraic surface. He clung instead to the use of subscripts to denote the different values of an algebraic function at a given point (a notation due to Cauchy).

essentially determined the properties of the Riemann surface. Two surfaces with equivalent cuts could be mapped onto each other without folding or tearing. Riemann showed that any simply connected region except the whole plane, no matter what its shape, is equivalent to a disk, so that the theory of analytic functions in such a region amounts to the same theory in the disk, where power series can be used to represent any analytic function. This famous result is known as the *Riemann mapping theorem.*

17.2 Algebra

The relation between the roots and the coefficients of a polynomial became more and more transparent as time went on. Euler and d'Alembert both gave geometric arguments to show that every equation has a root in the complex numbers, so that no new kinds of numbers needed to be invented in order to solve equations. Euler (1732) noticed that the procedure for solving a third-degree equation for x was to let $x = u + v$, where u^3 and v^3 satisfy a quadratic equation; in this way the problem reduced to an equation of degree one lower, plus the operation of extracting the cube root. Similarly the solution of a fourth-degree equation for x can achieved by reducing it to a biquadratic equation via a transformation whose parameters can be found by solving a cubic equation (the resolvent cubic). From these considerations he was able to give a unified method of solving equations of degree up to 4. Based on this experience he proposed that, for example, the fifth-degree equation might be solved by setting $x = \sqrt[5]{u_1} + \sqrt[5]{u_2} + \sqrt[5]{u_3} + \sqrt[5]{u_4}$. He did not achieve the solution, however.

Lagrange attacked the same problem by creating auxiliary equations of higher degree but greater symmetry, which he called *reduced equations* (they are now called *resolvents*). In general the resolvent equation is of degree $n!$ for an equation of degree n, but symmetry may make it possible to reduce this degree. For $n = 3$ and 4 it can be reduced to an equation of smaller degree (2 and 3, respectively), and hence the given equation can be solved. For $n = 5$ Lagrange was able to reduce his "reduced equation" only to degree 6, which left the quintic equation still unsolved. However Giovanni Francesco Malfatti (1731–1807) used the resolvent to solve a number of particular quintic equations.

Lagrange's idea lighted the way to a complete solution of the problem. By focusing on the operations one would have to perform in order to solve an equation (substitution and reduction), Paolo Ruffini (1765–1822) showed in 1799 that no such method can solve every quintic equation. In his paper "Della insolubilità delle equazioni algebraiche generali di grado superiore al quarto" (On the unsolvability of general algebraic equations of degree higher than the fourth) Ruffini introduced a concept that he called the *permutation* of an equation, that is, the set of substitutions that leave a given function unchanged. In the form of a permutation group, this concept was to have strong influence, not only on the solution of the problem of solving equations, but in every area of mathematics.

The penetration of analysis into algebra increased after the extension of trigonometric functions to complex variables by Jakob Bernoulli and Abraham de Moivre

(1667–1754). The fundamental question *Does every equation have a solution in the complex numbers?* was answerable only after the calculus was fully extended to complex variables in the form now known as *complex analysis*. Only in this context can even cubic equations be said to be completely understandable. That development, however, occurred in the nineteenth century, and the last details of the solution of the quintic equation were not worked out until the 1990s.

17.2.1 From Equations to Groups and Fields

As just mentioned, Ruffini gave an argument purporting to show that the general fifth-degree equation is not solvable by algebraic means. Numerical approximations to the solutions can, however, be obtained. The Chinese had been finding them for centuries, and the Chinese method was discovered independently by William Horner (1786–1837), a schoolteacher at Bath. Techniques for numerical solution of equations are never perfect and continue to be improved down to the present day. Such techniques, however, seldom lead to new areas of thought. It is the "impractical" theoretical questions that lead to new mathematics. Chief among these questions are the following: How many roots does an equation have? Are these roots rational numbers, real numbers, or complex numbers? Or is some kind of hypercomplex number required? Granted that the roots are determined by the coefficients, how can one proceed from the data (coefficients) to the output (roots)? What is the relation between coefficients and roots?

From the very earliest times the answer to this last question was known for quadratic equations. It is nowadays summed up in the quadratic formula taught in high-school algebra. A path was found from coefficients to roots for cubic and quartic equations in the sixteenth century, but that path sometimes wandered through irrational and imaginary numbers, even when it started and ended in rational numbers, that is, when both coefficients and roots are integers, as in the case of the equation $x^3 - 7x + 6 = 0$. It was the theoretical question that led to the creation of the complex numbers with all its beautiful applications; the "practical" numerical solution would never have required complex numbers.

Roots Exist, but How to Find Them?

Without the adjunction of complex numbers even the question of the existence of roots would have required a very clumsy classification of equations, and some equations would have no roots at all. By the eighteenth century it was strongly suspected that equations always have solutions in the complex numbers. The earliest attempts to prove this fact, by d'Alembert, Euler, and Lagrange, were brought to perfection by Karl Friedrich Wilhelm Gauss (1777–1855), who gave four different proofs of what is now known as the fundamental theorem of algebra: *For any polynomial* $p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$ *with complex coefficients,* $n \geq 1$ *and* $a_n \neq 0$*, there is a complex number* r *such that* $p(r) = 0$.

With the theoretical existence question settled, the still-open problem of finding a path from coefficients to roots could be attacked with more confidence of

finding a solution. By the end of the eighteenth century, there was a strong suspicion that for the general quintic equation no path (formula) could be constructed from coefficients to roots that involved only algebraic operations (the operations of arithmetic, together with root extractions). As a young student in Christiania (now Oslo), Abel dreamed of finding a general formula for solving all equations using algebraic operations and for a brief while thought he had succeeded. When he realized his mistake, he produced a proof that no such formula is possible for the general quintic. In the process he found it necessary to introduce the notion of the numbers generated by given numbers, that is, all the numbers that can be formed as a finite arithmetic combination of these numbers and the integers. Abel said that these numbers formed the *domain of rationality* of the given numbers. Implicitly here we have the notion now called a field—a structure on which addition, subtraction, multiplication, and division are defined and obey the usual laws. Fields form the natural domain for *stating* equations, since a polynomial is formed using only arithmetic operations. The crucial question is whether a root extraction will require elements not in the field. A field, like the complex numbers, in which every polynomial equation has a root, is said to be *algebraically closed.*

By 1800 it was known that the coefficients are symmetric functions of the roots, and so the question of how to get back from the symmetric functions to the roots themselves involved an investigation of the symmetries of the coefficients. It was this question of symmetry that led to the creation of one of the most fundamental concepts in all of mathematics: a group.

Equations and Their Groups

The concept of a group was created by Évariste Galois (1811–1832) while still in his teens. He twice submitted a paper on the subject to the Paris Academy, but both times it was lost. In 1832, on the night before a duel that led to his death, he once again wrote out his thoughts and sent them to a friend. They were published in 1846 by Liouville. Galois' approach to the subject required several pieces of background. The first was Abel's notion of a domain of rationality generated by a given set of numbers (in applications the generating set will be the coefficients of the equation) and the possibility of enlarging that domain by adjoining new numbers (the square roots, cube roots, etc., of numbers in the domain). If one can reach a domain containing the roots of the equation in this way, the equation is said to be *solvable by radicals.*

Galois followed Lagrange in considering permutations of the roots of an equation. He introduced the term *group* to describe a set of permutations that is closed under composition. He noted that if one group contains another, then the larger group can be partitioned into what are now called right and left cosets with respect to that subgroup. These are the sets obtained from the smaller group by multiplying on the right or left by elements of the larger group. Galois singled out the case in which the right and left cosets of a subgroup are the same. A subgroup having this property is now called a *normal* subgroup, although Galois did not use this word; he spoke of a *proper decomposition.*

After Galois' memoir was understood, it was seen that solving an equation whose group contains a nontrivial normal subgroup can be reduced to solving two equations of lower degree. In this way, theoretically, the question of whether a given equation can be solved reduces to computing the group of the equation. Since the group is finite, any question about it can (theoretically) be answered by an exhaustive search. Unfortunately, it is seldom easy to calculate the group of an equation.

With criteria for solvability available, it was natural to ask for a way of listing all the equations with coefficients in a given field that can be solved by radicals. This problem was posed by Leopold Kronecker (1823–1891). Kronecker also conjectured that the only Abelian extensions of the rational numbers are the so-called cyclotomic fields (those obtained by adjoining roots of unity). This conjecture was proved in 1886 by Heinrich Weber (1842–1913).

Ancestors and Descendants of Group Theory

The notion of a group had antecedents in the work of Lagrange and Ruffini. Both groups and fields occur (but not under those names) in Gauss' famous treatise *Disquisitiones arithmeticae*, and Cauchy published a number of papers on groups of substitutions (again, not calling them groups) in which he, like Lagrange, proved theorems about the number of values taken on by a function of several variables when the arguments are permuted. The modern definition of a group first appeared in an 1854 paper by Arthur Cayley (1821–1895), who developed the theory of finite groups and listed all possible multiplication tables for groups of eight elements. The basic elements of modern "Galois theory" were published by Camille Jordan (1838–1922) in 1870.

Group theory soon became one of the giant areas of mathematics. The finite groups that Galois considered in application to algebraic equations presented many mysteries. A complete classification of these groups took a century and a half to complete, and the results have still not been presented in a single coherent exposition. Such an exposition would require several thousand pages at present, and it is hoped that shorter proofs of the main results may be discovered. From finite groups, mathematicians turned to infinite groups and continuous groups, both of which turned out to be useful in various areas of mathematics and physics.

17.2.2 Links with Analysis

The periodicity properties of trigonometric and elliptic functions amount to the property of invariance under the action of a certain discrete group of transformations of the complex plane, namely translations. There are, however, many other discrete groups of such transformations of the complex plane. The most important of them are the fractional linear transformations

$$z \mapsto \frac{az+b}{cz+d}, \quad ad-bc \neq 0,$$

where a, b, c, and d are integers. The study of functions invariant under a discrete group of fractional linear transformations was inaugurated in the early 1880s by Henri Poincaré (1854–1912). Poincaré came to this theory from the study of differential equations with algebraic coefficients. The classes of functions invariant under such a group are called *automorphic functions*.

17.2.3 Links with Number Theory

The other side of the coin, so to speak, from the question of which equations have solutions, is the question of which equations a given number may satisfy. In particular a number that satisfies an equation with integer coefficients is called an *algebraic* number. Numbers that are not algebraic are called *transcendental*. The question whether any transcendental numbers exist is by no means trivial. It was answered in the affirmative by Charles Hermite (1822–1902), who proved in 1878 that e is transcendental. His method of proof was soon adapted by Ferdinand Lindemann (1852–1939), who showed in 1881 that π is also transcendental.

17.2.4 Linear Algebra

Considering that both differentiation and integration are linear operations and that linear functions are the simplest functions from a computational point of view, one might have expected linear algebra to develop very early. Actually it is a surprisingly late bloomer. The essential elements of the subject—finite-dimensional vector spaces, linear operators, and the eigenvalue problem—did not come fully into focus until nearly the end of the nineteenth century. In contrast *multi*linear algebra, as exemplified by the theory of determinants, goes back to the time of Leibniz, who gave what is now known as Cramer's rule for solving a system of linear equations.[5] Likewise Alexandre Vandermonde (1735–1796) found a need for determinants in discussing methods of eliminating variables between simultaneous polynomial equations. It will be recalled that the Japanese mathematicians of the previous century had used determinants for this same purpose.

Another source of linear algebra came to prominence in Britain in the midnineteenth century. The problem from which it springs is related to certain problems of number theory studied by Gauss (see below), but takes on a life of its own in the form known as invariance theory. For example, given two linear polynomials $f(x, y) = ax + by$ and $g(x, y) = cx + dy$, the determinant $ad - bc$ is said to be an *invariant* of the two polynomials. The meaning of this term is that if $x = pu + qv$ and $y = ru + sv$, the new polynomials become $(ap + br)u + (aq + bs)v$ and $(cp + dr)u + (cq + ds)v$, whose determinant is $(ps - rq)(ad - bc)$; in other words, the determinant is multiplied by the determinant of the substitution. An *invariant* of a set of polynomials is defined as a homogeneous polynomial in the coefficients that is multiplied by a power of the determinant of the substitution when a linear substitution is made for the variables.

[5]Gabriel Cramer (1704–1752) gave this rule in 1750 in connection with a curve-fitting problem.

As this example shows, the study of determinants is an important part of invariant theory. The primary figures in invariant theory during the middle and late nineteenth century were Cayley and his friend James Joseph Sylvester (1814–1897).

The theory of matrices as objects on which algebraic operations could be performed began with an 1858 paper of Cayley entitled "A memoir on the theory of matrices," which told how to multiply matrices. (Cayley was guided by the idea of linear substitutions and defined matrix multiplication to correspond to composition of substitutions. Nowadays the concept of substitution has been replaced by that of a linear transformation, but the two are algebraically equivalent.) In this paper also Cayley stated the famous Cayley–Hamilton theorem, that every matrix satisfies its characteristic equation, and proved this result for 2×2 and 3×3 matrices.

Higher-dimensional spaces also entered linear algebra via the work of William Rowan Hamilton (1788–1856), the inventor of quaternions (1843), which are most simply described as numbers consisting of one real and three imaginary parts: $A = a + a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$, where a, a_1, a_2, a_3 are real numbers and $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$ and $\mathbf{ij} = \mathbf{k}$, $\mathbf{jk} = \mathbf{i}$, and $\mathbf{ki} = \mathbf{j}$. Hamilton regarded a quaternion as having two parts, one of which was the real part a; the other part $(a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k})$ he called a *vector* (the Latin word for a carrier). This vector analysis turned out to be ideally suited for application to several areas of physics, and was developed into a powerful tool by the American mathematician Josiah Willard Gibbs (1839–1903). A theory based on multidimensional geometry and having much in common with vector analysis was developed by the German mathematician Hermann Günther Grassmann (1809–1877), who called it *Ausdehnungslehre* (Theory of Extensions).

17.3 Geometry

One can distinguish many lines of development of geometry in the eighteenth and nineteenth centuries. We shall consider six of these lines.

17.3.1 Analytic Geometry

The use of three mutually perpendicular axes had been implied as early as 1679 by Philippe de la Hire (1640–1718), who gave the equation of a cone in terms of three variables. Johann Bernoulli gave the equation of a surface in three-dimensional coordinates in 1698, and 2 years later Antoine Parent (1666–1716) gave the equation of a sphere in essentially its modern form. Alexis-Claude Clairaut (1713–1765) studied "curves of double curvature" and in 1731, when he was only 18, published a treatise on curves of double curvature in three-dimensional space. This seems to have been the first time such curves were ever considered. He recognized these curves as the intersection of surfaces and therefore realized that a curve in three-dimensional space requires two equations for its description. In his 1748 treatise on infinitesimal analysis Euler expounded three-dimensional analytic geometry in a form, that, except for vector notation, is what one now finds in calculus books.

The name *analytic geometry* first came into common use in French textbooks of the early nineteenth century. In particular Gabriel Lamé (1795–1870) wrote a textbook on methods of solving geometric problems in 1818, in which he gave some of the standard notation now used, such as the equation of a plane in terms of its intercepts ($x/a + y/b + z/c = 1$). Much of this language is still retained in calculus texts today, though supplemented by vector notation.

17.3.2 Projective and Descriptive Geometry

The projective properties of figures and the projective approach to geometry in general was developed by a large number of French, German, and Italian mathematicians. Jean Victor Poncelet (1788–1867) defined two figures to be *projectives* if they could be mapped onto each other by a series of projections. The properties of a figure that are preserved under projection were its projective properties. The property of being a conic section, for example, is a projective property.

The German mathematicians August Ferdinand Möbius (1790–1868) and Julius Plücker (1801–1868) and the Swiss mathematician Jakob Steiner (1796–1863) made projective geometry one of the most important areas of German mathematical research for the middle half of the nineteenth century. Steiner adhered to the "synthetic" approach, in which the use of coordinates to prove theorems was avoided. Steiner's idea was to generate more complex figures from simpler ones in a natural order. The most elementary shapes were a series of collinear points, a pencil of lines (the lines in a plane passing through a single point), and a pencil of planes (the planes passing through a single line). Following them were coplanar points and lines, and congruences (two-parameter families) of lines and planes. These were followed by the points and lines of three-dimensional space.

Incidentally, Steiner also devoted attention to the problem of constructions with ruler and compass. In particular, he showed that if a single circle is drawn, then every construction that is possible with ruler and compass can be performed with ruler alone.

In contrast to Steiner's synthetic approach, Möbius and Plücker promoted analytic methods. Möbius introduced barycentric coordinates, described as follows. Given any three noncollinear points P_1, P_2, and P_3 in a plane, each point P inside the triangle formed by these points will be the center of mass of a unique system of three masses m_1, m_2, and m_3 for which $m_1 + m_2 + m_3 = 1$. The three masses are the barycentric coordinates of P. (If we allow zero and negative masses, then points on and outside the triangle can also be given barycentric coordinates). As one might expect from the way in which barycentric coordinates are introduced, they also lead to very simple formulations of theorems in statics.

A generalization of this technique is to use what are called *homogeneous coordinates*, introduced by Plücker. The homogeneous coordinates of a point P are obtained by taking a fixed point P_0 outside the plane and representing P as a sum $e_1 \overrightarrow{P_0P_1} + e_2 \overrightarrow{P_0P_2} + e_3 \overrightarrow{P_0P_3}$. (Barycentric coordinates can be thought of as the limiting case of homogeneous coordinates when the point P_0 goes off to infinity.) The use of such coordinates greatly simplifies the formulation of many theorems

on the projective properties of figures, and makes it easy to deal with points and lines at infinity.

Plücker's greatest achievement was the realization that space can be thought of as made up of lines or circles, rather than points. The importance of this approach is that it leads naturally to spaces (manifolds) of more than three dimensions. For example, the space of spheres in three-dimensional space is four-dimensional, since three dimensions are required to specify the center of a sphere and a fourth is needed to specify the radius.

The British mathematicians developed projective geometry using analytic methods. Cayley, in working out the theory of algebraic forms and coordinate mappings, introduced what he called a *projective metric*. Such metrics, as Felix Klein (1849–1925) later showed, made it possible to regard both Euclidean and noneuclidean geometries as special cases of projective geometry. It was this analytic approach that led Klein, as a young professor at the University of Erlangen in 1872,[6] to propose a program of studying the geometric objects that remain invariant under different groups of transformations and thus sort out the relations between the different kinds of geometry. This project is known as the *Erlanger Programm*.

17.3.3 Algebraic Geometry

Algebra provided a much more natural classification of curves than the old Greek distinction among plane, solid, and linear locus problems. A locus could be said to be of order n if the equation of minimal degree representing the curve of the locus was of algebraic degree n. One could then state reasonable theorems based on this algebraic nomenclature. A good example is provided by a theorem contained in a paper by James Stirling (1692–1770) "Lineae tertii ordinis Newtonianae," (Newtonian curves of third order, a commentary on an earlier work of Newton on cubic curves). Stirling's theorem asserts that a curve of order n is uniquely determined by $\frac{1}{2}n(n+3)$ points. This would seem to be a natural conclusion, since there are $[(n+1)(n+2)]/2$ coefficients in the general equation of degree n in two variables. Trying to fit $[(n+1)(n+2)]/2$ points with such a curve would generally force all the coefficients to be zero, and there would be no curve. To be sure of leaving one coefficient nonzero, one would attempt to make the curve pass through at most $[(n+1)(n+2)]/2 - 1 = [n(n+3)]/2$ distinct points. Thus, since the general quadratic equation in two variables represents a conic, there ought to be a unique conic passing through any five points. It is obvious from algebra that there is at least one conic passing through any five points, and almost equally obvious (intuitively) that two distinct conics cannot intersect in more than four points. (This last conclusion turns out to reflect our own carelessness in considering degenerate cases, as Exercise 17.5 below shows.)

Stirling's claim conflicted with a result of Maclaurin, however, published in his 1720 work *Geometrica organica*. In this work Maclaurin proved that in general a

[6]Until 1875 new professors at Erlangen were required to defend a thesis at a public lecture for which a printed *program* was provided as an invitation. See "Erlangen programs" by Konrad Jacobs and Heinrich Utz, in *The Mathematical Intelligencer*, **6** (1), (1984), p. 79.

curve of order m intersects a curve of order n in mn points (counting multiplicities and imaginary points suitably). This result is now commonly called *Bézout's theorem* after Étienne Bézout (1730–1783), who discussed it in works published in the late 1750s. This result leads to no surprises in the case $m = n = 2$, where one usually does find four points of intersection. The difficulty arises in the case of cubics, where Stirling's result predicts that nine points should determine a cubic uniquely, yet Maclaurin's result predicts that *two distinct* cubics will intersect in nine points. The seeming paradox was noticed by Maclaurin; it was later rediscovered by Gabriel Cramer and was not adequately explained until the nineteenth century.

The most important curves studied in analytic geometry are those whose equations take the form $p(x, y) = 0$, where $p(x, y)$ is a polynomial in two variables. Such curves are called *algebraic curves*, for obvious reasons. It is well known, for example, that a curve of degree at most two is a conic section (possibly degenerate). Under the influence of Plücker these curves came to be studied in projective space. Plücker used line coordinates, defining the equation of a one-parameter family of lines to be the equation of the envelope of that family (the curve tangent to all of them). The degree of this equation is called the *class* of the curve, while the degree of its usual point equation is called its *order*. Plücker gave a set of four equations relating the order and class of a curve to the number of its nodes, cusps, stationary points, and double tangents.

An important link between geometry and calculus was through the integration of algebraic functions, and this became a major theme of nineteenth-century mathematics after the work of Abel. The need to regard these functions as functions of a complex variable was apparent in the work of Abel and Jacobi. The geometric aspect of the subject showed up best in Riemann's 1857 paper "Theorie der Abelschen Funktionen," which introduced yet another number to classify such curves, the *genus*. An algebraic curve of genus 0 represents a function whose worst irrationality is the square root of a quadratic polynomial. Those of genus 1 are elliptic functions. This classification laid the foundation for systematic study of such curves, which can exhibit a great deal of variety. The nineteenth century saw the development of this subject nearly complete, and research in this area did not revive until after World War II, when a new and more abstract point of view led to still more profound research.

17.3.4 Differential Geometry

Along with the language of calculus there came into geometry a whole set of analytic methods that made it possible to state and solve problems that went far beyond the capabilities of the old Euclidean geometry. The notion of curvature, for example, was given by Newton as Problem 5 in his *Fluxions*. The standard measure of curvature was naturally taken to be the circle, and the problem was to determine the circle at each point that curves at the same rate as the given curve (called the *circle of curvature*; its radius is the *radius of curvature*). This problem is now well known in calculus and involves the second derivative. The

centers of all the circles of curvature of a given curve form a curve called its *evolute*. Intuitively the curve can be obtained from its evolute by imagining a string wrapped tightly around the evolute unwinding while being kept taut. This fact is expressed by saying that the original curve is the *involute* of its evolute. Such problems as finding the involute and evolute of a curve, shortest paths on a surface between two points (geodesics), etc., form the subject matter of early differential geometry.

The subject of differential geometry grew up gradually during the early nineteenth century in the research of the French mathematicians and Gauss. Olinde Rodrigues (1794–1851), for example, studied *lines of curvature* and *radii of curvature* on a surface and discovered what is nowadays known as the Gaussian curvature.

The form differential geometry was to take for the next century was largely determined by Gauss' 1828 work *Disquisitiones generales circa superficies curvas* (General Treatise on Curved Surfaces). Gauss emphasized the definition of a surface in parametric form $[x = x(p, q),\ y = y(p, q),\ z = z(p, q)]$ in preference to the implicit definition by an equation $f(x, y, z) = 0$ and showed the importance of the first and second fundamental forms

$$ds^2 = E\,dp^2 + 2F\,dp\,dq + G\,dq^2 \ \text{and} EG - F^2,$$

where

$$\begin{aligned} E &= (\partial x/\partial p)^2 + (\partial y/\partial p)^2 + (\partial z/\partial p)^2, \\ F &= (\partial x/\partial p)(\partial x/\partial q) + (\partial y/\partial p)(\partial y/\partial q) + (\partial z/\partial p)(\partial z/\partial q), \\ G &= (\partial x/\partial q)^2 + (\partial y/\partial q)^2 + (\partial z/\partial q)^2, \end{aligned}$$

for computing arc length and area on the surface.

Ideally a mapping of a surface area (a city tourist guide), for example, should be a scale drawing, so that angles would be preserved between the original surface and the length of an object and its image would be in direct proportion. No such mapping between a curved surface and a flat surface is possible, of course, and so the question arises of whether one can preserve similarity to the maximum possible extent, say by mapping a portion of a sphere onto a plane in such a way that angles are preserved at each point and the magnification (the limiting ratio of the length of a line segment ending at the point to the length of its image) is the same in every direction. Such a mapping is said to be *conformal*, and the question whether conformal mappings exist is an important one in differential geometry. Gauss showed in his treatise that this question reduced to the question whether the fundamental quantities E, F, and G for the two surfaces were proportional. He also considered the ratio of the area of a small portion of a surface to the area of its projection on a sphere of unit radius having the same tangent plane as the surface, the center of the sphere being the center of projection. He showed that the limit of this ratio is the product of the largest and smallest curvatures (the principal curvatures) at the point. A consequence is that the area of a piece of the surface can be obtained by integrating the total curvature of that piece. He went

on to find an expression for the curvature in terms of E, F, and G and their partial derivatives that was homogeneous of degree 1 (his famous *theorema egregium*).

Gauss also considered the question of the shortest curved paths along a surface from one point to another (given the name *geodesics* by Joseph Liouville in 1830). This topic links geometry with calculus of variations and hence with differential equations. A link with noneuclidean geometry came when Gauss showed that the area of a triangle whose sides are geodesics is proportional to the difference between the sum of the angles of the triangle and two right angles.

Implicit in these elegant results was the possibility of studying a surface without any Euclidean space around it, as if the surface were itself the entire universe. The fundamental quadratic forms that Gauss introduced turn out to be independent of the parameterization of the surface, and hence define what is called the *intrinsic geometry* of the surface. The possibility thereby arose that the intrinsic geometry of physical space might be noneuclidean. This possibility can be tested experimentally by measuring the angle sums of large triangles, as Gauss certainly realized. In his 1828 work on curved surfaces, mentioned above, Gauss took advantage of geodetic survey measurements to consider the angles of a very large triangle; the results showed no measurable deviation from Euclidean geometry.

17.3.5 Noneuclidean Geometry

While analytic geometry was being developed, the old problems associated with Euclidean geometry were not forgotten, especially the greatest of them all, the problem of "what to do about the parallel postulate." This problem was systematically investigated by Girolamo Saccheri (1667–1733), a Jesuit priest. In a treatise published in the last year of his life he created a quadrilateral having equal vertical sides and right angles at the base, now known as a *Saccheri quadrilateral*. Its importance lies in the fact that the line through the midpoints of two sides of a triangle is parallel to the third side. If perpendiculars are dropped to this line from the endpoints of the third side, they form a Saccheri quadrilateral. Saccheri showed easily, as anyone could, that the other two angles of the Saccheri quadrilaterals, called the *summit angles*, are congruent. He proposed to show that they are right angles, thereby proving the parallel postulate. Without much difficulty he was able to show that they could not be obtuse angles. This result follows from what Euclid showed in Book I, assuming along with Euclid that a line divides the plane into two parts and that two distinct lines can intersect in only one point. There remained the possibility that the summit angles may be acute.

Saccheri began deducing consequences of the "hypothesis of the acute angle." One of the most interesting of these—now a pillar of hyperbolic geometry—is that two coplanar lines either have one common perpendicular, or meet at some point, or continually approach each other in one direction and continually recede from each other in the opposite direction (Proposition 23). At this point Saccheri was led into reasoning "at infinity." Considering the third of the possibilities he concluded, "we have two lines which produced must run together into the same line and have at one and the same infinitely distant point a common perpendicular."

Since infinitely distant points were not part of the machinery of his argument, he fell back on intuitive ideas and argued rather vaguely that this conclusion was impossible.

Saccheri, had he only known it, was the discoverer of the noneuclidean geometry that would be rediscovered a century later by János Bólyai (1802–1860) and Nikolai Ivanovich Lobachevskii (1792–1856). Like Columbus, however, he did not recognize what he had discovered because he was pursuing a different goal. Also like Columbus, he reported to the world that he had achieved his goal. His treatise bore the title *Euclidis ab omni naevo vindicatus* (Euclid freed of every blemish).

The pioneers of noneuclidean geometry, Lobachevskii and Bólyai, used a synthetic approach in the 1820s and 1830s to create what is now called hyperbolic geometry (a name suggested by Klein). Both men proved a standard set of theorems about hyperbolic geometry, and both derived the trigonometric formulas appropriate to this geometry. The standard kinds of geometry known as elliptic, parabolic (Euclidean), and hyperbolic, can be distinguished by imagining a circle tangent to a line, and watching the circle widen and flatten out as its center moves away from the line.

There are three possibilities:

1. When the center reaches some finite point (called the *pole* of the line of tangency) the circle coincides with the line. This is the case in the geometry of a sphere, in which a small circle tangent to the equator of a sphere, becomes the equator itself if its center recedes to the pole.

2. Every point in the half-plane on the side of the tangent line containing the circle is eventually engulfed by the circle, but the circle never coincides with the tangent line, that is, the circle never reaches the line, but approaches arbitrarily closely to it. This is the case in Euclidean geometry.

3. The circle approaches a limiting curve (called a *horocycle*), and there is a region of points lying between the tangent line and the horocycle. This is hyperbolic geometry. If the horocycle is revolved about the radius through the point of tangency, the resulting surface in three-dimensional hyperbolic space is called a *horosphere*. Lobachevskii was able to prove that the geometry of the horosphere is ordinary Euclidean plane geometry, and from that fact he derived the trigonometric formulas for hyperbolic geometry.

All horocycles are congruent; their existence makes it possible to define an absolute unit of length in hyperbolic geometry and to derive a formula for the angle α at which a line transversal to a second line at a point P will be parallel to a third line perpendicular to the second at a point Q. If the distance between P and Q is d, then

$$\alpha = 2 \arctan \left(e^{-d/k} \right).$$

where the length k cannot be determined from the axioms of the geometry.

It turns out that the trigonometric relations in this geometry bear a strong resemblance to those of spherical trigonometry. For example, in a right triangle

with legs a and b and hypotenuse c the relation

$$\cosh \frac{a}{k} \cosh \frac{b}{k} = \cosh \frac{c}{k}$$

holds. This formula is analogous to the formula of spherical geometry

$$\cos \frac{a}{r} \cos \frac{b}{r} = \cos \frac{c}{r},$$

where r is the radius of the sphere. Since $\cosh x = \cos ix$ ($i = \sqrt{-1}$), an earlier remark of Johann Heinrich Lambert (1728–1777) that an alternative to Euclidean geometry could be pictured as the geometry on a sphere of imaginary radius turned out to be astoundingly accurate.

Neither Lobachevskii nor Bólyai received due recognition for this work in their lifetimes, but in the next generation the subject blossomed into a beautiful and intricate theory as Riemann, Klein, and others developed their own ideas. In 1868 Eugenio Beltrami (1835–1900) attempted to interpret hyperbolic geometry using differential geometry. He introduced a pair of mutually perpendicular lines as coordinate axes and set the coordinates of a point equal to its distances from these two lines, just as in ordinary analytic geometry. He found that the Gaussian curvature of the hyperbolic plane at every point was $-(1/k^2)$. The analogy with a sphere of radius r, which has Gaussian curvature $1/r^2$ at every point, became even more apparent. Beltrami went further and sketched an interpretation for the hyperbolic plane *within* Euclidean geometry by regarding lines as chords in a disk. The existence of such an interpretation showed that any supposed contradiction in hyperbolic geometry would imply a contradiction within Euclidean geometry itself.

Very soon other mathematicians, including Klein and Poincaré, found other interpretations for hyperbolic geometry. At the same time, Cayley, Klein, and others were developing the noneuclidean geometry that results from assuming that any two lines intersect. (Klein suggested the name "elliptic geometry" for this kind of geometry.) Spherical geometry gives a good intuitive model of elliptic geometry, except that its "lines" (the great circles on a sphere) intersect in *two* points.

17.3.6 Topology

The subject now known as algebraic topology has origins in the seventeenth and eighteenth centuries. As early as 1619 Descartes had discovered that for any closed polyhedron, such as a tetrahedron, octahedron, dodecahedron, or prism, the number of vertices plus the number of faces is always two more than the number of edges. For example, a cube has 8 vertices 6 faces, and 12 edges. The number 2, which is the excess of the number of vertices and faces over the number of edges, is now known as the *Euler characteristic* of a closed polyhedron (or the plane or a sphere, since this branch of mathematics ignores shape). This area of geometry came to be known as *geometria situs* or *analysis situs* (geometry or analysis of position) to contrast with the "geometry of magnitude" that constitutes ordinary geometry. This subject is nowadays called *topology*, from Greek words meaning study of

position. This name was first used by Johann Benedikt Listing (1808–1882) in his 1847 book *Vorstudien zur Topologie.*

Topology ignores such notions as exact distance and is concerned only with the way in which an object is fitted together, which closed curves or surfaces on the object are boundaries of higher-dimensional regions, and the like. In its early days attention was focused on numerical relations between the faces of a polyhedron and their boundary edges. The earliest example of such a relation is the Euler formula just mentioned connecting the number of vertices (N_0), edges (N_1), and faces (N_2) of a closed convex polyhedron:

$$N_0 - N_1 + N_2 = 2.$$

Obviously this relation will remain true if the polyhedron is stretched or shrunk, provided it is not torn or folded over on itself. In more precise terms, in the spirit of Klein's classification of geometry, it is a relation that is preserved under continuous one-to-one-transformations (homeomorphisms).

Euler studied such problems only occasionally. Cauchy also studied the Euler relation once or twice and extended it to nonclosed polyhedra and unions of polyhedra. A more significant generalization came in 1813 from a professor at the University of Geneva named Simon l'Huilier (1750–1840), who showed that for a closed polyhedron with p cylinders stuck through it $N_0 - N_1 + N_2 = 2 - 2p$. This formula gives the general form of Euler's relation. (After Riemann's work, the number p could be identified with the genus of the surface.) The general study of the numerical relations that could result when faces are glued together along edges to form "complexes" was undertaken by Listing in 1862.

One of the standard objects that now inhabit the world of topology, the Möbius band, was introduced by Möbius in 1861 and developed more fully in the following years. Möbius introduced the concept of "elementary relatedness" to describe a correspondence between the points of two surfaces that preserves "infinitely near" pairs of points, what is now called a *homeomorphism*. Using this kind of correspondence he found that he could classify polyhedra according to the number of boundary curves they possessed.

These topological questions turn out to be intimately related to the question of which differential forms on a surface are exact differentials, a matter of great importance in the theory of differential equations and in complex analysis. This connection first appeared in the work of Riemann on Riemann surfaces, where a major theme is the classification of the closed paths on a Riemann surface, distinguishing those that are boundaries from those that are not. Later, when group theory came to permeate the subject, this topic would be known as homology theory. A related question about a Riemann surface is the question of which curves can be deformed continuously to a point on the surface and which can be deformed into each other. When formulated in the context of group theory, this question leads to homotopy theory. Riemann began the subject by defining a surface to be simply connected if every curve on it can be shrunk to a point, doubly connected if it can be made simply connected by cutting it open in one place, etc. (For example, a disk is simply connected; a disk with its center removed is doubly connected since a cut from the center to the boundary makes it simply connected.)

The difficulty of this subject is due to the great generality of curves that have to be considered. This difficulty shows up particularly well in a famous theorem that anyone can understand, but almost no one can prove. It was asserted in 1887 by Camille Jordan that a closed curve in the plane divides the plane into two regions, the inside and the outside of the curve. His proof, however, was objected to later on, and this theorem has had a long history of insufficient proofs. It does have some proofs that are regarded as correct. All of them require the subject known as algebraic topology.

In his work on Riemann surfaces Riemann was primarily concerned with complex analysis, but he was also interested in geometry for its own sake. His work on higher-dimensional objects was left in fragments at his death and published in his collected works. His friend Enrico Betti (1823–1892) was the first to speak explicitly of a space of any number of dimensions (in 1871). The introduction of such spaces, where visual intuition was limited, led to an increasing reliance on algebraic techniques.

17.3.7 Links with Differential Equations

We have already discussed differential equations from an analytic point of view. A different way of looking at differential equations was adopted by the Norwegian geometer Marius Sophus Lie (1842–1899). Lie hoped to do for differential equations what Galois had done for algebraic equations, that is, to associate a group with each equation to determine whether it can be solved by various prescribed methods.[7] To do this, he formulated the theory of what are now called *Lie groups*. These are continuous spaces such as the torus or the three-dimensional unit sphere in four-dimensional space on which a natural group operation can be defined (if the torus is thought of as the set of pairs of complex numbers (z, w) of absolute value 1, the group operation is componentwise multiplication; the 3-sphere in four-dimensional space is the group of quaternions of unit length). Lie himself worked only with the parts of the group near the identity element; the "global" construction of the group is a twentieth-century creation. Associated with every Lie group is a purely algebraic object now known as a Lie algebra, generated in a natural way by infinitesimal operations on the group. Lie established the relations between the Lie group and the purely algebraic structure. The importance of Lie groups and Lie algebras in modern physics is enormous.

The introduction of topological ideas into differential equations came in several stages. The Cauchy–Kovalevskaya theorem asserting the existence of analytic solutions is a local theorem applicable, for example, to an equation of the form

$$\frac{dx}{X(x,y)} = \frac{dy}{Y(x,y)}$$

[7] This subject, now known as *differential Galois theory*, was studied from a different point of view by Liouville, who was able to prove that some equations do not have solutions expressible as a finite formula involving only elementary functions. For example, Bessel's equation, $x^2y'' + xy' + (x^2 - p^2)y = 0$, has elementary solutions only when p is half of an odd integer.

only at points (x, y) where one of X and Y is nonzero. It turns out that the greatest interest lies precisely at the points at which both X and Y vanish, which are called *singular points*. Cauchy's methods had a pair of strong proponents in Charles Auguste Briot (1817–1882) and Jean-Claude Bouquet (1819–1885), who worked together to develop the theory of differential equations in terms of these singular points. They classified singular points into centers, foci, nodes, and saddle points. This classification was very fruitful in the hands of Poincaré, who extended it to general first-order equations of the form $F(x, y, y') = 0$. By considering the surface $F(x, y, z) = 0$, Poincaré was able to obtain a simple equation relating the various kinds of singularities to the genus of the surface.

One source of Poincaré's work was the need to study differential equations qualitatively in situations where closed-form solutions are not possible and numerical solutions offer no insight. Another was the study of algebraic (Abelian) integrals using complex variables. From these two bases he realized the need for fundamental topological research and undertook such research during the 1890s. He was the first to introduce homology theory as it is now known, defining the boundary of a manifold as a chain of submanifolds of lower dimension. If the boundary is trivial, the manifold is called a *cycle*. He defined the kth Betti number of a manifold to be the maximum number of independent k-cycles. (A set of k-cycles is independent if no nontrivial combination of the cycles can form the boundary of any higher-dimensional manifold.) Poincaré discovered that the Euler characteristic of a surface could be expressed in terms of the Betti numbers. In this way it became possible to generalize the Euler characteristic to topological spaces of higher dimension.

Poincaré also generalized Riemann's ideas of simple connectivity to higher-dimensional objects, thereby creating homotopy theory. After considerable experimentation with arcane examples, Poincaré conjectured that a closed three-dimensional manifold whose homotopy theory is trivial must be topologically equivalent to the three-dimensional sphere in four-dimensional space. This famous conjecture remains unsolved as of the present, although its generalizations to dimensions higher than 3 have all been proved.

17.4 Probability

One of the classical works in probability is the posthumous (1713) treatise of Jakob Bernoulli called the *Ars conjectandi* (*The Art of Prediction*). Bernoulli was interested in the application of this mathematical technique to human life, and he gave a very stark picture of the gap between theory and application, saying that only in simple games such as dice could one apply the equal-likelihood approach of Fermat and Pascal, whereas in the cases of interest, such as human health and longevity, no one had the power to construct a suitable model. He recommended statistical studies as the remedy to our ignorance, saying that if 200 people out of 300 of a given age and constitution were known to have died within 10 years, it was a 2-to-1 bet that any other person of that age and constitution would die within a decade.

17.4.1 The Law of Large Numbers

Bernoulli imagined an urn containing numbers of black and white pebbles, whose ratio is to be determined by sampling with replacement. Here it is possible that you will always get a white pebble, no matter how many times you sample. However, if black pebbles constitute a significant proportion of the contents of the urn, this outcome is very unlikely. After discussing the degree of certainty that would suffice for practical purposes (he called it *virtual certainty*), he noted that this degree of certainty could be attained empirically by taking a sufficiently large sample. The probability that the empirically determined ratio would be close to the true ratio increases as the sample size increases, but the result would be accurate only within certain limits of error. More precisely, given certain limits of tolerance, by a sufficient number of trials,

> we can attain any desired degree of probability that the ratio found by our many repeated observations will lie between these limits. . . .

This last statement is the law of large numbers for what are now called *Bernoulli trials*, that is, repeated independent trials with the same probability of a given outcome at each trial. (Recall that Cardano had given a vague formulation of the same idea.) If the probability of the outcome is p and the number of trials is n, this law can be phrased precisely by saying that for any $\varepsilon > 0$ there exists a number n_0 such that if m is the number of times the outcome occurs in n trials and $n > n_0$, then the probability that the inequality $|(m/n) - p| > \varepsilon$ will hold is less than ε. In other words, one can specify an error tolerance as small as desired and a probability of exceeding that error as small as desired. If n is large enough, the probability that the proportion of trials in which the outcome occurs will differ from the probability of the outcome by more than the tolerated error will be less than the specified probability.

17.4.2 The Central Limit Theorem

The problem of the law of large numbers raised the secondary problem of estimating the sum of a segment of terms in the binomial series. This problem was attacked by Abraham de Moivre. In 1733 he wrote a paper on approximation of a sum of terms of the binomial expansion $(a+b)^n$, in which he touched on several important parts of modern probability theory. The main problem was to compute the probability that the number of occurrences of a given outcome of probability p in n trials will be between A and B. Jakob Bernoulli had shown that this probability would be $\sum_{A<k<B} \binom{n}{k} p^k (1-p)^{n-k}$, this expression being simply part of the binomial expansion $1 = (p + (1-p))^n$. The difficulty occurs in computing the binomial coefficients $\binom{n}{k} = [n!/k!(n-k)!]$. The factorials rapidly become huge, and the amount of computation involved becomes unfeasible. Bernoulli had resorted to crude estimates sufficient to establish the law of large numbers. De Moivre worked out an approximation to these factorials by focusing on the middle term in the expansion of $2^n = (1+1)^n$. He found that the ratio of this term to 2^n was approximately $2/(B\sqrt{n})$, where the natural logarithm of B (which De Moivre

called the *hyperbolic logarithm*) is given by $-1+\frac{1}{12}-\frac{1}{360}+\frac{1}{1260}-\frac{1}{1680}+\cdots$. De Moivre's friend James Stirling showed that $B=\sqrt{2\pi}$. From this result De Moivre was able to show that for large values of n, the term l places from the middle term differs from this value by a factor of $e^{-(2l^2/n)}$ approximately. (De Moivre stated this fact in terms of the logarithm; he did not mention the number we call e.) Then as a corollary he observed that, if an infinite number of trials could be carried out, with equal probability of an event occurring or not occurring at each trial, the probability that the number of occurrences of that event would be between $\frac{1}{2}n-\frac{1}{2}\sqrt{n}$ and $\frac{1}{2}n+\frac{1}{2}\sqrt{n}$ would be

$$\sqrt{\frac{2}{\pi}}\Big(1-\frac{1}{3}\frac{1}{2}+\frac{1}{2\cdot 5}\frac{1}{4}-\frac{1}{6\cdot 7}\frac{1}{8}+\frac{1}{24\cdot 9}\frac{1}{16}-\frac{1}{120\cdot 11}\frac{1}{32}+\cdots + (-1)^k\frac{1}{k!\cdot(2k+1)}\frac{1}{2^k}+\cdots\Big).$$

This corollary expresses for this particular case what is now known as the central limit theorem: If a large number of independent and identically distributed random variables, each of which has expected value 0 and standard deviation 1, is averaged, the average is approximately a normal distribution with standard deviation $\frac{1}{\sqrt{n}}$. In particular, for large values of n the approximate probability given by the series just written is now expressed by the integral

$$\frac{1}{\sqrt{2\pi}}\int_{-1}^{+1} e^{-\frac{1}{2}t^2}\,dt.$$

These considerations were the first indication of the important role to be played by the "bell-shaped curve" [the graph of $y=(1/\sqrt{2\pi})e^{\frac{-(1}{2)x^2}}$] known as the standard normal probability density. The fact that the average of suitably normalized independent samples of any distribution whatsoever is approximately normal is known as the central limit theorem in probability.

Soon after its introduction by Huygens and Jakob Bernoulli the concept of mathematical expectation came in for some critical appraisal. While working in the Russian Academy of Sciences, Daniel Bernoulli and his brother Niklaus discussed the problem now known as the *Petersburg paradox*. We can describe this paradox informally as follows. Suppose you flip a coin until heads appears. If it appears on the first flip, you win \$2, if it first appears on the second flip, you win \$4, and so on; if heads first appears on the nth flip, you win 2^n dollars. How much money would you be willing to pay to play this game? Now by "rational" computations the expected winning is infinite, being $2\cdot\frac{1}{2}+4\cdot\frac{1}{4}+8\cdot\frac{1}{8}+\cdots$, so that you should be willing to pay, say, \$10,000 to play each time. On the other hand, who would bet \$10,000 knowing that there was an even chance of winning back only \$2, and that the odds are 7 to 1 against winning more than \$10? Clearly something more than mere expectation was involved here. That something is of vital importance to the insurance industry, which makes its profit by having a large enough stake to play "games" that resemble the Petersburg paradox. The question involved is: Granted, one should expect the "expected" value of a quantity depending on chance, how

confidently should one expect it? The question of *dispersion* or *variance* of a random quantity lies beneath the surface here and needed to be brought out. It turns out that when the expected value is infinite, or even when the variance is infinite, no rational projections can be made.

17.4.3 Statistics

The subject of probability formed the theoretical background for the empirical science known as statistics. Some theoretical analysis of the application of probability to hypothesis testing and modification is due to Thomas Bayes (1702–1761), a British clergyman. The first work on statistics proper was a treatise of 1835 entitled *Physique social* by the Belgian scholar Lambert Quetelet (1796–1874). The name *statistics* comes from the records used in administering government, which provide the raw data we now call statistics.[8] Quetelet introduced certain analogies with physical concepts into social analysis, the most famous of these concepts being the "average man" (*l'homme moyen*), which he considered the exact analog of the notion of center of gravity of a physical body. The needs of statistics helped to guide the development of probability theory, which was applied to analyze large data samples by regarding each data point as having the same probability as any other data point. This technique has provided some powerful ways of testing hypotheses, and is indispensible in modern law, medicine, and many other areas.

This area, incidentally, is one of the few in which American mathematicians made significant contributions during the nineteenth century. For example, the Irish-American mathematician Robert Adrain (1775–1843) discovered the normal distribution of errors in 1808. As is too often the case with scholars working in isolation (the Russians are a good example), the discoveries are often duplicated later at large centers of research, and the second discoverer gets all the credit. Gauss published the same result in 1809, and the normal distribution is now called alternatively the *Gaussian distribution.*

17.4.4 Large Numbers and Limit Theorems

In the late eighteenth century Laplace showed rigorously what de Moivre had already stated, that the probability that the number of successes in a sequence of independent Bernoulli trials is between a and b tends to an expression given by an integral of e^{-t^2}, that is, what is now called a normal, or Gaussian, distribution. This result was the first special case of what is known as the central limit theorem. Laplace, as an astronomer, was interested in this problem as it applied to observational errors.

The law of large numbers was studied by Simeon Denis Poisson (1781–1840), who discovered an approximation to the probability of getting at most k successes in n trials, and thereby introduced what is now known as the *Poisson distribution.*

The Russian mathematician Pafnutii L'vovich Chebyshëv (1821–1894) introduced the concept of a random variable and its mathematical expectation. He is

[8] Sometimes by folk etymology a single piece of data is called "*a* statistic."

best known for his 1846 proof of the weak law of large numbers for repeated independent trials; he showed that the probability that the actual proportion of successes will differ from the expected proportion by less than any specified $\varepsilon > 0$ tends to 1 as the number of trials increases. In 1867 he proved what is now called *Chebyshëv's inequality*, that the probability that a random variable will assume a value more than k standard deviations from its mean is at most $1/k^2$; this inequality implies the law of large numbers. In 1887 Chebyshëv also gave an explicit statement of the central limit theorem for independent random variables.

The extension of the law of large numbers to dependent trials was achieved by Chebyshëv's student Andrei Andreevich Markov (1856–1922). The subject of dependent trials—known as *Markov chains*—remains an object of current research. In its simplest form it applies to a system in one of a number of states $\{S_1, \ldots, S_n\}$ which at specified times may change from one state to another. If the probability of a transition from S_i to S_j is p_{ij}, the matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix}$$

is called the *transition matrix*. If successive transitions are all independent of one another, one can easily verify that the matrix power P^k gives the probability of a transition in k steps.

17.5 Number Theory

The seventeenth-century work of Fermat in number theory was ably advanced in the eighteenth century. Fermat had conjectured that the number $2^{(2^n)} + 1$ is always a prime. This statement is true for $n = 1, 2, 3, 4$, as the reader can easily check. For $n = 5$ this number is $4,294,967,297$, and to prove that it is prime in the crudest manner—by checking all possible prime factors—one must attempt to divide it by every prime less than 65,537. In 1732 Euler found that this fifth Fermat number is divisible by 641.

Everyone knows the famous Fermat conjecture that the sum of the nth powers of two positive integers is not the nth power of a rational number unless $n = 2$. Fermat himself mentioned that his method of infinite descent could prove this for $n = 3$ and $n = 4$, but he did not write out the proof. Euler provided the proof in 1738. He also proved that every positive integer is the sum of at most four square integers and conjectured that no sum of fewer than n nth powers could be an nth power, a conjecture that was finally refuted for $n = 5$ in 1966.

A second assertion of Fermat proved by Euler is now known as "Fermat's little theorem." It asserts that any prime p divides $a^{p-1} - 1$ unless a is itself divisible by p. Euler went on to show more generally that m divides $a(a^{\phi(m)} - 1)$, where $\phi(m)$ is the number of positive integers less than m and relatively prime to m (now called *Euler's ϕ-function*).

A problem of number theory whose fame is second only to the Fermat conjecture is a conjecture of Christian Goldbach (1690–1764), who wrote to Euler in 1742 that every odd integer seemed to be a sum of at most three odd prime integers. In the form of the slightly emended proposition that every even integer larger than 4 is the sum of two primes (both necessarily odd, of course), this famous assertion is known as the *Goldbach conjecture*. In 1937 the Russian mathematician Ivan Matveevich Vinogradov (1891–1983) proved that every sufficiently large odd integer is the sum of at most three primes. In proving this result Vinogradov made use of elementary but extremely delicate estimates of the magnitudes of trigonometric polynomials, showing once again the penetration of analysis into number theory.

The elegant particular results of Fermat, Euler, and Lagrange in number theory were generalized in the course of the nineteenth century. This subject was one of Gauss' favorite objects of contemplation, and his *Disquisitiones arithmeticae* became a classical work on the properties of integers. One of his earliest discoveries as a teenager was the law of quadratic reciprocity. To state it we need the concept of congruence modulo an integer. Two integers m and n are said to be *congruent modulo* r if they leave the same remainder when divided by r (equivalently, if their difference is divisible by r). The law of quadratic reciprocity says that if p and q are two primes both congruent to 3 modulo 4, then precisely one of them is congruent to a square integer modulo the other. If one of them is congruent to 1 modulo 4, then either each is congruent to a square modulo the other or neither is congruent to a square modulo the other.

In attempting to extend the law of quadratic reciprocity to higher powers, Gauss was led to consider what are now called the *Gaussian integers*, that is, the complex numbers of the form $m + n\sqrt{-1}$. Gauss showed that the concepts of prime and composite number make sense in this context just as in the ordinary integers and that every such number has a unique representation (up to multiplication by the units ± 1 and $\pm\sqrt{-1}$) as a product of irreducible factors. Notice that no prime integer of the form $4n + 1$ can be "prime" in this context, since it is a sum of two squares: $4n + 1 = p^2 + q^2 = (p + q\sqrt{-1})(p - q\sqrt{-1})$. The generalization of the notion of prime number to the Gaussian integers is an early example of the endless generalization and abstraction that characterizes modern mathematics.

Gauss' work was carefully read by Dirichlet, who contributed several gems to this difficult area. One of these is the theorem that each arithmetic sequence in which the first term and the common difference are relatively prime contains an infinite number of primes. To prove this result, he introduced the "Dirichlet character" $\chi(n) = (-1)^k$ if $n = 2k + 1$, $\chi(n) = 0$ if n is even, along with the "Dirichlet series"

$$\sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = 1 - \frac{1}{3^s} + \frac{1}{5^s} - \frac{1}{7^s} + \cdots.$$

17.5.1 The Prime Number Theorem

Dirichlet's theorem raises the problem of a quantitative estimate of the relative number of primes among the integers. The prime numbers seem to be quite ir-

regularly distributed among the integers, but it is known that there is always a prime between n and $2n$. A good estimate of the number of primes less than or equal to a given integer N is given by $N/(\log N)$. This estimate was suggested by Gauss. Another estimate suggested by Legendre, $N/(A \log N + B)$ with $A = 1$, $B = -1.08366$, turns out to be correct only in its first term. This fact was realized by Dirichlet, but only after he had written approvingly of the estimate in print. (He corrected himself in a marginal note on a copy of his paper given to Gauss.) Dirichlet suggested $\sum_{k=2}^{N}[1/(\log k)]$ as a better approximation.

This problem was also studied by Chebyshëv. The number of primes in the finite sequence $\{1, \ldots, n\}$ is nowadays denoted $\pi(n)$. Chebyshëv proved that if $\alpha > 0$ is any positive number (no matter how small) and m is any positive number (no matter how large), the inequality

$$\pi(n) > \int_2^n \frac{dx}{\ln x} - \frac{\alpha n}{\ln^m n}$$

holds for infinitely many positive integers n, as does the inequality

$$\pi(n) < \int_2^n \frac{dx}{\ln x} + \frac{\alpha n}{\ln^m n}.$$

This result strongly suggests that $\pi(n) \sim [n/(\ln n)]$, but it would be desirable to know if there is a constant A such that

$$\pi(n) = \frac{An}{\ln n} + \varepsilon_n,$$

where ε_n is of smaller order than $\pi(n)$. It would also be good to know the rate at which $\varepsilon_n/\pi(n)$ tends to zero. (Chebyshëv's estimates imply that if A exists, it must be equal to 1.) He was able to show that in fact

$$0.92129 < \frac{\pi(n)}{n/\ln n} < 1.10555.$$

This result and its later refinements is known as the *prime number theorem*).

The full proof of the prime number theorem turned out to involve the use of complex analysis. Riemann had introduced the function known as the Riemann zeta function:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

defined by this formula for $\operatorname{Re} s > 1$ and extended by analytic continuation to all other complex numbers. It is not difficult to see that the extended function has zeros at the even negative integers. Riemann showed that a good estimate of $\pi(x)$ can be obtained if all the other zeros of $\zeta(s)$ lie on the line $s = \frac{1}{2} + it$, t real. This conjecture, still unproved, is known as the *Riemann hypothesis*. The gaps in Riemann's methods were finally circumvented by two long-lived twentieth-century mathematicians, the Belgian Charles de la Vallée-Poussin (1866–1962) and

the Frenchman Jacques Hadamard (1865–1963). The former, in particular, showed that

$$\pi(n) = \int_2^n \frac{dx}{\ln x} + \varepsilon_n,$$

where for some $\alpha > 0$ the error term ε_n is bounded by a multiple of $ne^{-\alpha\sqrt{\ln n}}$.

17.5.2 Links with Algebra

The abstract concepts of modern algebra began to appear thick and fast in the late nineteenth century. In work published from the 1870s to the 1890s Richard Dedekind (1831–1916) introduced the notion of a *field* (*Zahlkörper*) as a collection of complex numbers on which the four arithmetic operations are defined and satisfy the commutative, associative, and distributive laws familiar from arithmetic. (Nowadays such objects are called *subfields* of the complex numbers.) He also introduced the notions of a *module* as an object that is closed under addition and subtraction (nowadays we would call this a *module over the integers*) and an *ideal* as an object that is closed under addition and subtraction and under multiplication by any number, whether in the ideal or not.

17.5.3 Links with Analysis

Assaults on Fermat's last theorem continued throughout the nineteenth century. In 1847 Gabriel Lamé published a paper in which he claimed to have proved the result. Unfortunately, he assumed that complex numbers of the form $a_0 + a_1\theta + \cdots + a_{n-1}\theta^{n-1}$, where $\theta^n = 1$ and $a_0, \ldots, a_{n-1}$ are integers, can be factored uniquely, just like ordinary integers. Ernst Eduard Kummer (1810–1893) had noticed some 10 years earlier that such is not the case and had constructed what he called "ideal divisors" to save the theory. This was just one of the many ways in which the objects studied by mathematicians became increasingly abstract, and the old objects of numbers and space became merely special cases of the general objects about which theorems are proved. Kummer was the first to make general progress toward a proof of Fermat's last theorem. The conjecture that $x^p + y^p = z^p$ has no solutions in positive integers x, y, and z when p is an odd prime had been proved only for the cases $p = 3$, 5, and 7 until Kummer showed that it was true for a class of primes called *regular primes*, which included all the primes less than 100 except 37, 59, and 67. This step effectively closed off the thought that Fermat might be proved wrong.

17.6 Combinatorics

The seeds planted by Leibniz in his *De arte combinatoria* sprouted and grew during the nineteenth century as problems from algebra, probability, and topology required sophisticated techniques of counting. One of the pioneers was the British clergyman Thomas Kirkman (1806–1895). The first combinatorial problem he

worked on was posed in the *Lady's and Gentleman's Diary* in 1844: *Determine the maximum number of distinct sets of p symbols that can be formed from a set of n symbols subject to the restriction that no combination of q symbols can be repeated in different sets.* Kirkman himself posed a related problem in the same journal 5 years later: *Fifteen young ladies in a school walk out three abreast for 7 days in succession; it is required to arrange them daily so that no two shall walk twice abreast.* This problem is an early example of a problem in combinatorial design. The problem of covering each location in a square array of n rows and n columns with a symbol chosen from a set of n symbols in such a way that each symbol appears once in each row and once in each column (such an array is called a *Latin square*) is another example.

Kirkman's combinatorial work dovetailed with topology in two areas: first in the classification of polyhedra having prescribed numbers of faces meeting at each vertex, second in the theory of knots. The mathematical study of knots was impossible before algebra and combinatorics had advanced to a certain level adequate to classify graphs. (Nowadays the subject relies on even more sophisticated notions involving the connectivity of the space complementary to the knot itself.) Peter Tait (1831–1901) published a paper on knots in 1876, in which he classified all knots with seven or fewer crossings. At Tait's suggestion Kirkman (now a septegenarian) took up the study of knots and links and classified those having up to ten crossings.

17.7 Foundations of Mathematics

The relation between numbers and the line, that is, the problem of incommensurables, was finally settled by Richard Dedekind. In an 1872 work entitled *Stetigkeit und irrationale Zahlen* (Continuity and Irrational Numbers) Dedekind pointed out the formal similarity in order properties between the line and numbers. The crucial consideration is that each point on a line separates the line into two parts: the points to the right of it and the points to the left of it. Dedekind realized that these two sets of points could be used in place of the point itself in any argument. There is no value in doing so in geometry, but in terms of numbers there is a great deal of value in doing so. For there is as yet no definition of an irrational number. If the number is simply *defined* as a partition of the rational numbers such that every element in one class is smaller than every element of the other class, the result is an object that has all the properties of a number: it can be added to and multiplied by other numbers of the same type. As Dedekind said, it was now possible for the first time to prove rigorously that $\sqrt{2}\sqrt{3} = \sqrt{6}$. Before that time mathematicians had been applying the ordinary rules of arithmetic to square roots, logarithms, exponentials, and many other numbers without having a clear definition of what a real number is or a rigorous proof that these rules are correct.

Although the calculus was securely established, the structure of the real line was by no means exhaustively studied. A deeper insight into the structure of the line came from a different source. Riemann's work on trigonometric series had raised the question whether a trigonometric series that converged to zero at every

point must have all coefficients zero. Riemann had given a positive answer to this question assuming that the coefficients tend to zero. This work was extended by Georg Cantor (1845–1918). Cantor showed that the assumption that the coefficients tend to zero was unnecessary, and that the conclusion remained valid even if the series fails to converge to zero at a finite number of points. Now, given an *infinite* number of points in a finite interval, the Bolzano–Weierstrass theorem implies that the points must cluster around one or more points, which are now called *points of accumulation.* Cantor studied the matter further and discovered that the theorem remained valid assuming the series converges except at an infinite number of points, provided there were only a finite number of points of accumulation. This was the crucial step that led Cantor to the concept of a point set. A point can only be a point of accumulation relative to a collection or set of points.

Starting with a set P, Cantor used the letter P' to denote the set of its points of accumulation, known as the derived set. One can then consider P'', P''',... $P^{(n)}$, etc. Now the interesting fact is that $P^{(n+1)} \subseteq P^{(n)}$ for all n, so that the derived sets are nested. That makes it possible to consider the derived set of infinite order $P^{(\omega)}$, defining it as the set of points common to all of the derived sets of finite order. But then, one can *go beyond infinity* by considering the derived set of $P^{(\omega)}$, denoted $P^{(\omega+1)}$.

In this way Cantor had discovered the ordinal numbers. The concepts of ordinal and cardinal numbers and other mysteries of set theory occupied him for the rest of his life. He never went back to the question of uniqueness of trigonometric series. His work seemed to some mathematicians to be more philosophy than mathematics (and with good reason, since many pages of his early papers were devoted to a discussion of what philosophers had said), and many mathematicians of a conservative bent opposed it. Prominent among the latter were Poincaré and Kronecker. The question of the proper foundation of mathematics was now joined in earnest; and although set theory remains the basic language for most mathematicians today, there are schools of mathematicians, notably the intuitionists, who oppose the use of some of its principles. The subject of set theory, which was an attempt to analyze the real numbers completely, eventually split into two areas. One of these is measure theory, which is concerned with generalizing the concept of length, area, and volume from sets that are geometrically simple to more complicated ones; it is closely linked with the theory of integration and nowadays with probability theory also. The other area is descriptive set theory, which attempts to classify sets according to their complexity. It starts with the simplest sets (intervals), then passes to countably infinite unions and intersections of intervals (class 1), then countably infinite unions and intersections of sets of class 1 (class 2), etc. Descriptive set theory has generated some of the hardest problems in mathematics, such as the continuum hypothesis. (The continuum hypothesis asserts that every uncountable subset of the real numbers can be placed in one-to-one correspondence with the whole set of real numbers).

The topological concepts of compactness, connectedness, convergence, and category and their relations to measure-theoretic notions and integration were worked out during the two decades from 1890 to 1910. The knowledge needed in order to become a mathematician did not increase, however; it merely became more ab-

stract. A large number of particular facts about elementary and special functions passed out of the curriculum to make room for the new concepts. These forgotten particular facts are continually being rediscovered in the late twentieth century and occasionally published in good faith as new mathematics.

17.8 Logic and Calculating Machines

As is well known today, the construction of effective computing machines and programming languages is impossible without symbolic logic. Pioneers in this work were two British mathematicians, Augustus de Morgan (1806–1871) and George Boole (1815–1864). De Morgan was primarily a logician, who invented a symbolism consisting of capital letter-small letter pairs to denote a concept and its opposite; for example, if X stands for "human," then x stands for "nonhuman." He is best remembered for the logical laws that bear his name: "not-(A or B)" is equivalent to "(not-A) and (not-B)," and "not-(A and B)" is equivalent to "(not-A) or (not-B)." He developed a calculus in which elementary propositions could be characterized by equations.

De Morgan conducted a correspondence with William Rowan Hamilton, the inventor of quaternions, on the quantification of propositions. This correspondence inspired Boole, who had previously been occupied with more standard mathematical questions in differential equations, to write a series of books on mathematical logic. The third book, entitled *The Laws of Thought* (1849), gave a systematic exposition of symbolic logic and formed an important part of the background for the school of mathematical philosophy known as *logicism* in the early twentieth century.

The early calculating machines of Pascal and Leibniz were improved in design by Charles Babbage (1791–1871), who, according to his own account, was dreaming over a table of logarithms, when it occurred to him that all of these tables could have been computed mechanically. His desire to simplify computation had the same source as the logarithms themselves, the needs of astronomy. Babbage developed and built a *difference engine*, which could calculate values of functions at small intervals. When tested using the quadratic function $x^2 + x + 41$, it proved to be greatly superior to hand computation in speed for numbers with a large number of digits; moreover it was indefatigable and not prone to errors. The success of the difference engine led Babbage to attempt to improve it. This project was much more difficult and caused Babbage to have a breakdown in 1827.

Babbage had been inspired by the success of the Jacquard loom, which wove predetermined patterns, reading its instructions from punched cards. He designed a machine called the *analytical engine* that would accept and store data from one set of cards and instructions from another set of cards. Babbage's talent brought him great honors, despite his irascible character. He made the acquaintance of Augusta Ada Lovelace (1815–1852), the daughter of the poet Lord Byron. This talented woman had studied mathematics with Augustus de Morgan, through whose wife she came to know Babbage. Not at all intimidated by the complexity of the analytical engine, she translated and expanded an Italian account of Babbage's

work into a clear exposition of it, and is credited with having written the first computer program. Unfortunately, she died at the age of 36, not having had time to devote herself systematically to scientific investigation.

17.9 Problems and Questions

17.9.1 Problems in Postcalculus Mathematics

Exercise 17.1 Find the length of the logarithmic spiral $r = e^\theta$ from $\theta = 0$ to $\theta = \varphi$, where φ is any given angle. [Arc length in polar coordinates is given by $l(\varphi) = \int_0^\varphi \sqrt{r^2 + (r')^2}\, d\theta$, where r' means the derivative with respect to θ.] This problem was solved by Torricelli in 1640.

Exercise 17.2 Imagine a thread tightly wrapped around the logarithmic spiral $r = e^\theta$ with its end at the point $(1, 0)$. What will be the equation of the end of the thread as it is unwound, always being kept taut? [That is, find the locus of points P such that the length of the tangent from P to the point Q of tangency equals the length of the curve from $(1, 0)$ to Q. This curve is called an *involute* of the given curve.]

Exercise 17.3 Draw the *complete* graph of the equation $y^2 = x^2(x - 1)$. (Be sure not to leave out any points.) How is the point $(0, 0)$ related to the rest of the curve? This point is called a *conjugate point* of the curve.

Exercise 17.4 Consider a curve $y = f(x)$ in the plane. Its tangent line at a point (x_0, y_0) has the equation $f'(x_0)x - y = f'(x_0)x_0 - y_0$, and therefore the normal to this line has the equation $x + f'(x_0)y = x_0 + f'(x_0)y_0$. Somewhere on this line we should find the center of a circle through (x_0, y_0) that fits the curve as well as any circle can, that is, having the same curvature as the original curve. To find this center, solve this equation simultaneously with the corresponding equation at a nearby point, that is, $x + f'(x_1)y = x_1 + f'(x_1)y_1$, obtaining the relation $(x_1 - x_0) - (y - y_1)\big(f'(x_1) - f'(x_0)\big) + (y_1 - y_0)f'(x_0) = 0$. Then divide by $x_1 - x_0$ and let x_1 tend to x_0 to find the limiting value of y. Give an expression for the curvature $\big((x - x_0)^2 + (y - y_0)^2\big)^{-(1/2)}$.

Exercise 17.5 Consider the two equations

$$\begin{aligned} xy &= 0, \\ x(y - 1) &= 0. \end{aligned}$$

Show that these two equations are independent, yet will always have infinitely many common solutions. What kind of conic sections do these equations represent?

Exercise 17.6 Consider the general cubic equation

$$Ax^3 + Bx^2y + Cxy^2 + Dy^3 + Ex^2 + Fxy + Gy^2 + Hx + Iy + J = 0,$$

which has 10 coefficients. Show that if this equation is to hold for 10 different points (x, y), the only way to achieve this result (in general) would be to take all the coefficients $A, \ldots J$ equal to zero. That is, in general one cannot find a curve of order at most 3 passing through 10 different points. Show that for any nine points, however, there will certainly be solutions $A, \ldots, J$ with some of the coefficients nonzero. Use linear algebra to show that, in general, these constants would be in proportion for any two such solutions, and hence would represent the same curve. That is, there is always at least one curve of order (at most) 3 passing through any *nine* points in the plane, and generally that curve is unique.

In the same way show that, given eight points, one can always find at least two different curves of order (at most) 3 passing through those eight points. Moreover, if one has two polynomials $P(x, y)$ and $Q(x, y)$ of degree at most three whose coefficients are not proportional to each other such that the curves $P(x, y) = 0$ and $Q(x, y) = 0$ pass through these eight points, then in general any curve of order (at most) 3 passing through these points has an equation of the form $\lambda P(x, y) + \mu Q(x, y) = 0$.

Exercise 17.7 Use Euler's equation to find the shortest curve $x = x(t)$ between the points $(0, 1)$ and $(1, 2)$, that is, minimize the arc-length integral

$$\int_0^1 \sqrt{1 + \left(\dot{x}(t)\right)^2}\, dt = \int_0^1 f(t, x, \dot{x})\, dt$$

with $x(0) = 1$ and $x(1) = 2$. [*Hint:* You *know* the answer to this problem in advance.]

Exercise 17.8 Show that the differential equation

$$\frac{dx}{\sqrt{1 - x^4}} + \frac{dy}{\sqrt{1 - y^4}} = 0$$

has the solution $y = [(1 - x^2)/(1 + x^2)]^{1/2}$. Find another obvious solution of this equation.

Exercise 17.9 Use the Maclaurin series for $e^{-(1/2)t^2}$ to verify that the series given by de Moivre represents the integral

$$\frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{1}{2}t^2}\, dt,$$

which is the area under a standard normal ("bell-shaped") curve within one standard deviation of the mean, as given in many tables.

Exercise 17.10 The Petersburg paradox is one of the paradoxes of the infinite, though not of the same sort as the set-theoretic paradoxes. The infinite expected winnings in this game depend on being able to play infinitely many games. What a rational person must take into account is the likelihood of variance from the mean, which could be disastrous in a low-payoff game such as the Petersburg

game if the stakes are high. Show that if you had a large enough stake to play, you could expect to come out ahead paying \$10,000 per game if you could play $2^{10,000}$ games. How long would it take you to play this many games?

Exercise 17.11 Verify that

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5.$$

Exercise 17.12 Prove Fermat's little theorem by induction on a. [*Hint:* The theorem can be restated as the assertion that p divides $a^p - a$ for every positive integer a. Use the binomial theorem to show that $(a+1)^p - (a+1) = mp + a^p - a$ for some integer m.]

Exercise 17.13 Verify the law of quadratic reciprocity for the primes 17 and 23 and for 67 and 71.

Exercise 17.14 Show that Fourier series can be obtained as the solutions to a Sturm-Liouville problem on $[0, 2\pi]$ with $p(x) = r(x) \equiv 1$, $q(x) = 0$, with the boundary conditions $y(0) = y(2\pi)$, $y'(0) = y'(2\pi)$. What are the possible values of λ?

Exercise 17.15 Using the Maclaurin series $e^x = 1 + x + x^2/2! + x^3/3! + \cdots$ and $\cos x = 1 - x^2/2! + x^4/4! - x^6/6! + \cdots$ and the formula $\cosh x = (e^x + e^{-x})/2$, verify that $\cos(\pm ix) = \cosh x$, where $i = \sqrt{-1}$. Hence show that the hyperbolic Pythagorean theorem is the spherical Pythagorean theorem with a sphere of imaginary radius. Also use these series to show that both of the formulas become the ordinary Pythagorean theorem if $r = \infty$. Hence ordinary Euclidean geometry results from spherical or hyperbolic geometry when the space becomes flat, that is, its curvature $1/r^2$ becomes zero.

Exercise 17.16 Prove that the number of primes less than or equal to N is at least $\log_2(N/3)$, by proceeding as follows. Let $p_1, \ldots, p_n$ be the prime numbers among $1, \ldots, N$, and let $\theta(N)$ be the number of square-free integers among $1, \ldots, N$, that is, the integers not divisible by any square number. We then have the following relation, since it is known that $\sum_{k=1}^{\infty}(1/k^2) = \pi^2/6$.

$$\begin{aligned}
\theta(N) &> N - \sum_{k=1}^{n} \left[\frac{N}{p_k^2}\right] \\
&> N\left(1 - \sum_{k=1}^{n} \frac{1}{p_k^2}\right) \\
&> N\left(1 - \sum_{k=2}^{\infty} \frac{1}{k^2}\right) \\
&= N\left(2 - \frac{\pi^2}{6}\right) > \frac{N}{3}.
\end{aligned}$$

But a square-free integer k between 1 and N is of the form $k = p_1^{e_1} \cdots p_n^{e_n}$, where each p_j is either 0 or 1. Hence $\theta(N) \leq 2^n$, and so $n > \log_2(N/3)$. This interesting bit of mathematical trivia is due to the Russian-American mathematician Joseph Perott (1854–1924).

Exercise 17.17 Deduce from Gauss' *theorema egregium* that if one surface can be conformally mapped onto another, the curvatures of the two surfaces at any two pairs of corresponding points are proportional. (This fact is also referred to as the *theorema egregium*.)

Exercise 17.18 Consider a plane with two distinct points O and P singled out and identified as the complex numbers 0 and 1, respectively. Then consider all the points that can be located (as the intersection of two straight lines, two circles, or a straight line and a circle) using a straightedge and compass, starting with these two points. (A straight line can be drawn only if two points on it have already been located; a circle can be drawn only if its center and one of its points have already been located. A point is located if it is the intersection of two curves already drawn.) Show that the points corresponding to all complex numbers of the form $r + si$, where r and s are rational, are among these numbers. Show that if z is one of these points, so is $\sqrt{z}$, and that if a and b are among these points, so are $a+b$, $a-b$, ab, and (if $b \neq 0$) a/b. Conclude that these "Euclidean" numbers form a field. Show that every quadratic equation with coefficients in this field has a root in this field. Is the same statement true for cubic equations?

Exercise 17.19 Show that if every polynomial $p(z)$ with *real* coefficients has a zero in the complex numbers, then every polynomial with *complex* coefficients also has a complex zero. (This reduction is vital for some of Gauss' proofs of the fundamental theorem of algebra.) [*Hint:* If $p(z)$ has complex coefficients, consider $q(z) = p(z)\overline{p(\bar{z})}$, where the bar denotes complex conjugation.]

Exercise 17.20 Show that the quadratic formula

$$ax^2 + bx + c = 0 \text{ if } x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

is valid in any field where $1 + 1 \neq 0$. Naturally 2 here means $1 + 1$ in the field, and 4 means $1 + 1 + 1 + 1$.

Exercise 17.21 Consider the field consisting of four elements $\{0, 1, \alpha, \beta\}$ whose addition and multiplication tables are

$+$	0	1	1	β
0	0	1	α	β
1	1	0	β	α
α	α	β	0	1
β	β	α	1	0

and

$\cdot$	0	1	α	β
0	0	0	0	0
1	0	1	α	β
α	0	α	β	1
β	0	β	1	α

Does the quadratic formula enable you to solve equations in this field? What is the solution of the equation $x^2 + x + 1 = 0$? Can this solution be expressed in terms of the coefficients of the equation using only the field operations and square roots? Does the equation $x^2 + \alpha x + 1 = 0$ have a solution in this field?

Exercise 17.22 Show that the factorization of numbers of the form $m + n\sqrt{-3}$ is *not* unique by finding two different factorizations of 4. Is factorization unique for numbers of the form $m + n\sqrt{-2}$?

17.9.2 Questions about Postcalculus Mathematics

Exercise 17.23 How do you explain the following seeming paradox, based on Exercise 17.6? Nine points ought to determine *one* cubic; yet we can produce a set of nine points by simultaneously solving two different cubic equations. Given such a set of nine points, for any eight of these points any two essentially different cubic polynomials passing through those points ought to determine the whole family of cubics that pass through them. In particular the P and Q referred to above ought to determine this family. But clearly any curve of this family will also pass through the ninth point of intersection of the curves $P = 0$, $Q = 0$.

The logical conclusion is that certain sets of nine points in the plane have a peculiar property: eight of them suffice to determine the ninth. Putting it another way, given eight points, there is (generally) a ninth point such that any cubic passing through the eight points will also pass through the ninth.

Exercise 17.24 How might Descartes have discovered the Euler characteristic? Imagine drawing a connected polygon in the plane. You start with a single vertex and a single face (the whole plane) and no edges. Show that each time you add a new vertex or a new edge starting from a point already on the graph, the number of vertices and faces added equals the number of edges added, no matter how this is done.

17.10 Endnotes

1. The material in this chapter is based partly on a reading of the original documents and partly on the following sources.

   (a) Siegfried Gottwald, Hans-Joachim Ilgauds, and Karl-Heinz Schlote (eds.), *Lexikon Bedeutender Mathematiker* (Bibliographisches Institut, Leipzig, 1990).

(b) Moritz Cantor, *Geschichte der Mathematik*, Vol. 3 (Teubner, Leipzig, 1898).

(c) David Eugene Smith *History of Mathematics*, Vol. 1 (Ginn and Company, New York, 1923).

(d) Julian Lowell Coolidge, *A History of Geometrical Methods* (Clarendon Press, Oxford, 1940).

(e) Herman H. Goldstine, *A History of the Calculus of Variations* (Springer-Verlag, New York, 1980).

(f) Morris Kline, *Mathematical Thought from Ancient to Modern Times* (Oxford University Press, New York, 1972).

(g) *Mathematics of the Nineteenth Century*, A.N. Kolmogorov and A.P. Yushkevich, eds., Vol. 1: *Logic, Algebra, Number Theory, Probability Theory* (Nauka, Moscow, 1978). English translation, Birkhäuser, Basel, 1992.

(h) *Mathematics of the Nineteenth Century*, A.N. Kolmogorov and A.P. Yushkevich, eds., Vol. 2: *Geometry, Analytic Function Theory* (Nauka, Moscow, 1981). English translation, Birkhäuser, Basel, 1996.

(i) *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences* (2 Vols.), I. Grattan-Guinness, ed., Routledge, London, 1994.

2. The discussion of Bernoulli's *Ars Conjectandi* is based on the source book of Ronald Calinger, *Classics of Mathematics* (Prentice-Hall, Engelwood Cliffs, NJ, 1995), pp. 421–423.

3. The counterexample to Euler's conjecture on the sum of n nth powers (Exercise 17.11) can be found in the paper of L.J. Lander and T.R. Parkin, "Counterexample to Euler's conjecture on sums of like powers," *Bulletin of the American Mathematical Society*, **72** (1966), p. 1079.

4. The discussion of combinatorics is based on the article "T.P. Kirkman, Mathematician" by N.L. Biggs, in the *Bulletin of the London Mathematical Society*, **13** (1981), pp. 97–120.

5. The material on Babbage's computing machines is taken from *The Computer from Pascal to Von Neumann* by Herman Goldstine (Princeton University Press, 1972).

Chapter 18

Modern Mathematical Science

The story we have been telling in the last few chapters is distorted by being cut off from its roots in physical science. The interaction between science and mathematics is profound, mysterious, and beautiful. This subject is vast, and one historian of mathematics, Ivor Grattan-Guinness, has devoted more than 1500 pages to a study of the development of mathematics and science in France alone in the years from 1800 to 1840. Grattan-Guinness notes that historians have vastly overemphasized pure mathematics, to the detriment of an accurate understanding of the history of mathematics. The central role in the development of analysis and differential equations during the first half of the nineteenth century in France must be assigned to problems of physics and engineering.

Since the major applications of mathematics have been in physics, we shall look at some of the most influential of these connections, in mechanics, electricity and magnetism, and relativity. Our purpose is not to give a general account of mathematical physics, but only to illustrate the role that mathematics plays in science. The same points could have been made by considering other areas, such as optics, acoustics, or quantum mechanics.

18.1 Mechanics and Astronomy

We shall trace just one thread in the tapestry of celestial mechanics—the explanation of planetary motion—through the work of five scientists: Galileo Kepler, Descartes, Huygens, and Newton.

18.1.1 Galileo

In his *Dialogues Concerning the Two New Sciences*, written while he was under house arrest in 1638, Galileo attacked certain Aristotelian concepts of motion, not for being in conflict with observation, as is sometimes supposed, but for internal inconsistency. Salviati, the main character in the dialogue, speaks on behalf of "the author," who claims to have verified Salviati's claims by experiment. Salviati

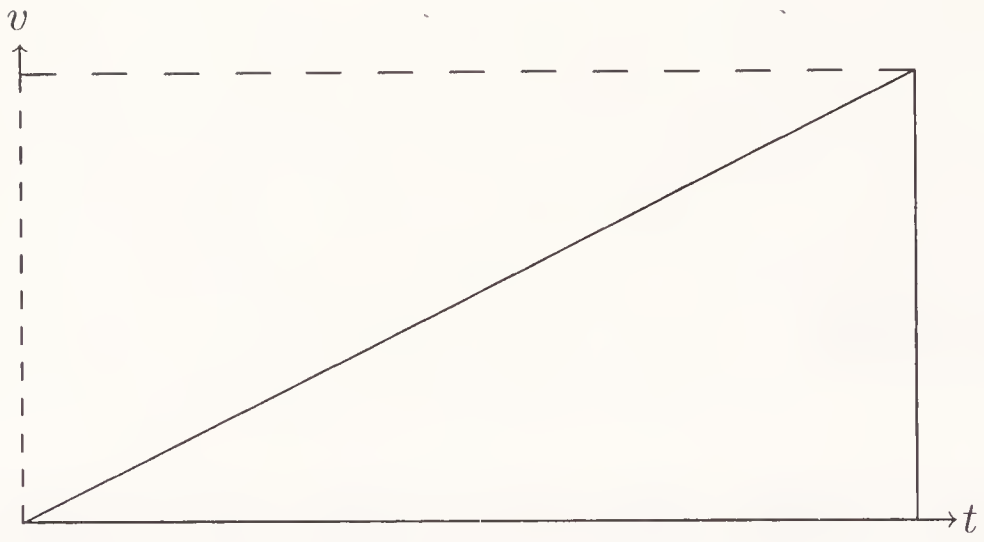


Figure 18.1: The Merton rule (velocity v as a function of time t). The distance traveled is represented by the lower triangle.

defines *naturally accelerated motion* just as the Merton scholars had defined *uniformly accelerated motion*, that is, a motion in which the increase in velocity over a time interval is proportional to the length of the interval. Salviati asserts that this kind of motion is the motion of actual falling bodies. The other participants in the dialogue, Sagredo and Simplicio, do not object to this proposition in principle, but ask to have certain difficulties disposed of, for example, the objection that such a motion must have zero velocity at its beginning and therefore could never begin. Having disposed of such objections and corrected certain conclusions erroneously drawn from the hypothesis of uniformly accelerated motion, Galileo states the Merton rule for the motion of a uniformly accelerated body and illustrates it with exactly the same figure (Fig. 18.1) that was given by Nicole of Oresme two and a half centuries earlier. Galileo noted that the distances traveled in successive equal time intervals will be in proportion to $1, 3, 5, \ldots$, and from this fact he deduced that the distance will be proportional to the square of the time interval. In fact it will be given by $s = \frac{1}{2}at^2$, where s is the distance covered and a the increase in velocity per unit time (acceleration).

Another of Galileo's contributions to the subject of physics was the idea of resolving motion into components parallel to coordinate axes according to the parallelogram law. This idea originated in the time of Aristotle, as we saw in Chapter 7, but its extensive use in physics dates from the time of Galileo. It is the earliest prefiguration of the concept of a vector. Galileo used this idea to derive the correct law of the inclined plane, which Jordanus Nemorarius had discovered three centuries earlier. He used the same principle to resolve the motion of a projectile into a uniform horizontal component and a uniformly accelerated vertical component. Since the horizontal motion is at constant velocity, while the vertical motion is uniformly accelerated (downward), it followed that vertical position would be proportional to the square of the horizontal displacement, so that the path would be a parabola.

So far the mathematics has been algebraic and clean. The crucial step toward applying calculus in physics involves imagining that irregular processes (nonuniformly accelerated motion, for example) take place on a microscopic scale over an

infinitesimal length of time. On this microlevel, velocities can be regarded as constant and curves as straight lines. Whatever laws can be derived on the microlevel can then be transferred to the macrolevel. Galileo had provided this important idea in an earlier dialogue *On the Two Great World Systems* (1632). In that dialogue Simplicio claimed that the earth could not rotate, since such a rotation would cause the inhabitants to be thrown off at a tangent, like the sparks that fly from a blacksmith's wheel. Galileo argued that there were two components of force acting on the body, the tangential force of its inertia and the centripetal force of gravity. He noted (see Fig. 18.2) that if inertia would move the body tangentially from A to B in a given time, the centripetal force will keep it on the circle by merely moving it from B to C, which is much smaller (in the limit, infinitely smaller) than the distance from A to B. Therefore any centripetal force, no matter how small, would suffice to overcome any force due to rotation, no matter how large.

This pattern of reasoning involves three assumptions:

1. observable phenomena can be regarded as the result of processes taking place on an infinitesimal level;

2. on the infinitesimal level curves can be regarded as straight lines;

3. approximations that become arbitrarily precise on a sufficiently small finite scale become true equality on the infinitesimal level.

This reasoning was elevated to the status of a scientific principle by Riemann in the midnineteenth century, when he argued that the geometry of space must be constructed from a metric given on the infinitesimal level. That is, the square of an infinitesimal length of curve ds^2 must be given as a combination of the infinitesimal increments in the coordinates, with coefficients $g_{\mu\nu}$ that may vary from point to point:

$$ds^2 = \sum_{\mu,\nu} g_{\mu\nu}\, dx^\mu\, dx^\nu;$$

and the length of a finite curve is then obtained by integrating ds.

The secret of the success of infinitesimal methods is mathematical simplicity: on the infinitesimal level one can assume that an effect is directly proportional to its cause. Thus, for example, on the infinitesimal level a planet subject to the gravitational attraction of the sun can be assumed to travel along the diagonal of a parallelogram one side of which is tangent to its orbit and the other side of which is directed toward the sun. In this way, through the mediation of the infinitesimal, the notion of direct proportion provides the bedrock of classical mechanics.

18.1.2 Kepler

Galileo's new mechanics of falling bodies developed at almost the same time as a new kinematics of celestial motion. The Copernican system had been published two decades before Galileo was born, but it had not yet triumphed and it was regarded with suspicion by both Catholic and Protestant leaders. The Danish astronomer Tycho Brahe (1546–1601) had not fully accepted it, adhering instead

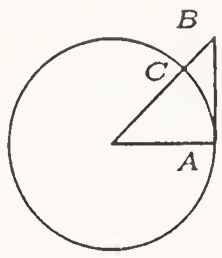


Figure 18.2: Forces on a body in circular motion.

to a compromise system in which the sun and the outer planets revolved around the earth while Mercury and Venus revolved around the sun. Brahe made a large number of observations of planetary locations that were much more precise than any made previously. After his death it was left to his associate Kepler to pore over these voluminous records and wrest from them the secrets of the solar system. In his *Astronomia nova* (1609) Kepler used the language of Cavalieri to express one of his discoveries about planetary motion. He described the area swept out by the line from the sun to the planet as "the sum of the lines from the sun to the planet." To simplify his explanation of his discovery he first allowed the orbit of a planet to be a circle with the sun off-center (eccentric). He then described its motion:

> . . . the total sum of the distances is to the time of a full period as any part of the sum of the distances is to its time.

In modern language this principle says that the area swept out in a given time interval by the line joining the sun to a planet is proportional to the length of the time interval. This rule is now called *Kepler's second law*. Kepler discovered it first because it is simpler than his other laws and consistent, as he noted, with all kinds of orbits.

Kepler's first law was much more difficult to derive. In trying to fit the data for the observations of Mars he noted that the orbit deviated inward from a certain circle, but outward from an ellipse he had placed inside the circle. He therefore looked for an intermediate curve to fit the orbit exactly, and this trick provided him with his discovery. Noting that the only mean between a circle and an ellipse is another ellipse, he concluded,

> Therefore the path of the planet is an ellipse. . .

It was Kepler who coined the name *foci* for the two points inside an ellipse now known by that name. The word *focus* is the Latin word for a hearth, and it was chosen because the sun is located at that point.

Kepler was inclined toward mysticism and waxed quite lyrical about the significance of the sun. He undertook to write the "harmony of the spheres" as music to be played. In his youth he had been intrigued by the idea that the orbits of the planets are in proportion to the radii of the spheres inscribed in and circumscribed about the five regular solids. This principle would imply that there could be no

undiscovered planets. Yet as a way of generating conjectures this seemingly irrelevant and hopelessly wrong hypothesis led him to a vital discovery, crucial for the future of physics. It kept him wondering about the distances of the planets from the sun, and caused him to wonder if their sidereal periods had any regular relation to their distances from the sun. Eventually he found such a relationship, and immediately communicated it in his *Harmonice mundi* in 1618. He reported that the idea first occurred to him on March 8 of that year, but unfortunately was rejected because of an erroneous computation. The idea came back to him on May 15, however, and this time his computations were correct. Kepler was jubilant at finding this reward for 17 years of labor among Brahe's observations, and this feeling made him immediately mistrustful of himself. However, he finally proclaimed his third law with confidence.

> ...the ratio of the periodic times of any two planets is precisely the sesquialteral ratio of the average distances, i.e., of their orbits...

In modern language the sidereal period of a planet is proportional to $a^{3/2}$, where a is the semimajor axis of the ellipse (the average of the greatest and least elongations of the planet from the focus). This third law was to play a major role in the establishment of Newtonian mechanics.

18.1.3 Descartes

The principles of mechanics were arrived at piecemeal, and the path toward them was not entirely straight. Occasionally someone would stumble on the correct and simple analysis of a phenomenon, yet reject it because it was inconsistent with certain assumptions of the time. Nevertheless, the proper description of certain mechanical phenomena gradually came to be understood. Such concepts as force, momentum, velocity, and acceleration were gradually given clear definitions and their relationships to one another were sorted out. One step on this road can be seen in a treatise published by Descartes in 1644, entitled *The Principles of Philosophy*. This treatise contains the modern definition of momentum, the law of conservation of momentum, and the law of inertia, including the fact that an unforced motion would be in a straight line, which had escaped Galileo.

> ...it seems evident to me that it was none other than God who, in his omnipotence, created matter with the motion and rest of its parts and who now conserves by his regular operations in the universe exactly the same amount of movement and rest as set out when he created it. For, though movement be only a form in inert matter, that matter nevertheless has a definite quantity of it, which never increases or decreases, even though there may be more or less of it in various parts. This is why, when one piece of matter moves twice as fast as another and the other is twice as large as the first, we must consider that there is just as much motion in the smaller as in the larger; and whenever the motion of one piece decreases, that of some other piece increases in proportion... if a body has once begun to move, we must

> conclude that it continues to move, and that it will never stop of its own accord... each piece of matter tends to continue its motion in straight lines, never in curves... .

As this quotation shows, Descartes regarded motion as a "quality" possessed by bodies, a very awkward way of thinking about it. He did not understand the composition of momenta, and as a result his analysis of the motion of bodies was mostly wrong.

Descartes' cosmology was short-lived. He was still hampered by trying to explain things qualitatively rather than quantitatively. His *Principles of Philosophy* are a mixture of brilliant insight such as the laws of momentum and inertia, side by side with utter nonsense, such as the claim that a fixed star can turn into a comet and the notion that low tides are caused by pressure from the vortex of the moon on the earth.

18.1.4 Huygens

The study of motion at constant velocity was fairly complete in ancient times, and linear motion under constant acceleration was well explained by Galileo. There remained, however, one other geometrically simple motion that needed to be explained, namely motion in a circle at constant angular velocity. Descartes had noted that unforced motion would be in a straight line. Therefore motion in a circle of radius r at speed v must be an accelerated motion; symmetry shows that the acceleration must be constant in magnitude. The problem was to find its value.

This problem and many others in mechanics were solved by Huygens. He made a thorough investigation of the motion of falling bodies and proved that the oscillations of a pendulum are not truly isochronous, as Galileo had believed. He found that a particle sliding down a hemispherical bowl would take slightly longer to reach the bottom if started from a greater height. He showed mathematically, however, that if the bowl was formed by rotating a cycloid, the ball would take the same time to reach the bottom, no matter how high it started. He also discovered that the involute of a cycloid (the curve obtained as the locus of the end of a piece of string initially taut against the cycloid and then unwound while keeping it taut) would be another cycloid. These two principles enabled him to design a pendulum clock that would theoretically keep the same time no matter how wide an arc the pendulum traversed. The top portion of the pendulum was a flexible band, and on each side of it were two strips of metal bent into the shape of a cycloid. These strips forced the end of the pendulum to swing along the involute of the cycloid, i.e, in another cycloid. (See Fig. 18.3.) Because of frictional loss in the flexing of the band, the cycloidal pendulum clock does not keep better time than an ordinary pendulum clock. However, the problem raised interest in the purely mathematical problem of the relation of a curve to its involute.

In 1673 Huygens published his discoveries in a work entitled *De horologio oscillatorio*, which also contained several theorems now central to classical mechanics. One of these was the principle that the center of gravity of a group of bodies oscillating periodically rises to its original height during each period, but

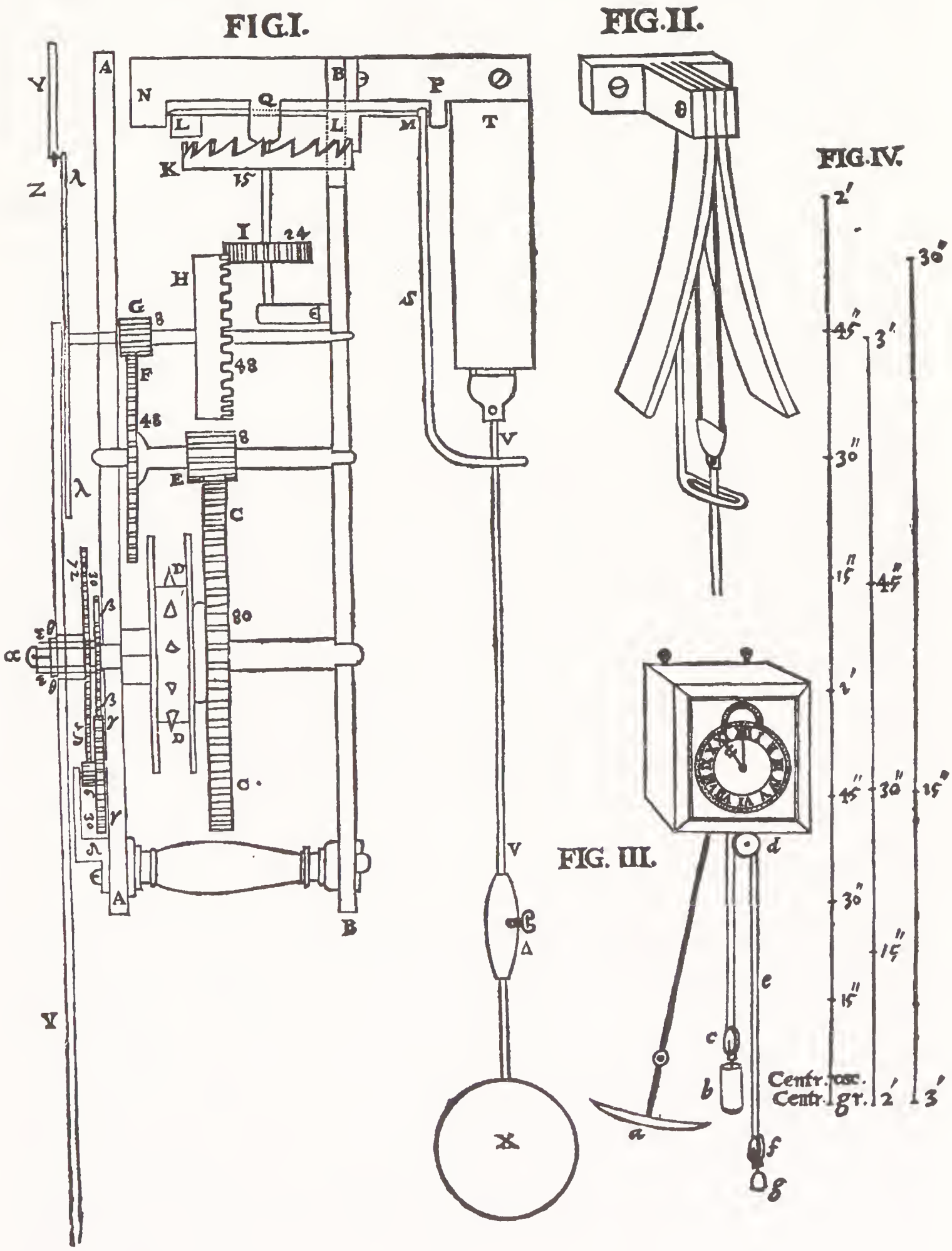


Figure 18.3: The cycloidal pendulum clock, from Huygens' *Horologium*. Stock Montage, Inc.

no higher. This theorem is one application of the law now known as conservation of energy. A second fundamental result in this work was the law of acceleration for uniform circular motion. A body in uniform circular motion can be thought of as continually falling toward the center of the circle with a constant acceleration a, and thus the result of Galileo that the distance fallen is $\frac{1}{2}at^2$ will apply on the infinitesimal level. The problem is to find a in terms of the radius r of the circle and the linear velocity v of the body.

Huygens gave this result in the fifth part of his treatise, in which he stated that when two equal masses move in unequal circles at the same speed, their centrifugal forces (accelerations) are inversely proportional to the diameters of the circles (Theorem II), while if two equal masses move on equal circles at different constant speeds, the centrifugal forces are directly proportional to the squares of the speeds (Theorem III).

If we assume that a particle is moving along a circle of radius r at constant speed v, then by the principle of inertia, without an acceleration, it would move from point A to point B, a distance vt tangent to the circle, in time t (see Fig. 18.2). (The actual point on the circle whose arc from A equals AB is not C, but for very small time intervals t the error in using the point C instead of the point reached in time t is negligible.) The distance BC that the particle "falls" is therefore $r\left[\sqrt{1+(vt/r)^2}-1\right]$, and since $\sqrt{1+x}-1$ equals $\frac{1}{2}x$ for infinitesimal values of x (we would now say that the limit of their ratio is 1), the distance BC that the particle "falls" toward the center in an infinitesimal time interval t is $\frac{1}{2}(v^2/r)t^2$. Comparing this expression with Galileo's law of motion for falling bodies $s = \frac{1}{2}gt^2$, where g is the acceleration due to gravity, we find that the acceleration is v^2/r. This is the result announced by Huygens.

18.1.5 Newton

The idea that the gravitational attraction of one body for another is inversely proportional to the square of the distance between them seems intuitively plausible if one thinks of gravity as a force that radiates from each particle of matter. The total amount of force on a sphere with center at the particle is the same for all spheres, while the area of the sphere increases as the square of the radius. Therefore, to keep the total force constant, the "concentration" or intensity of force at each point must decrease in proportion to the square of the radius. This intuitive idea, however, is not the path Newton followed to the discovery of this law. Instead he applied the results of Galileo, Huygens, and Kepler to the motion of the planets. For a body moving uniformly in a circle of radius r with speed v two things can be deduced: (1) the acceleration is v^2/r, as Huygens showed; (2) the period T of revolution is given by the equation $T = 2\pi r/v$. Then, if Kepler's third law is true, there is a constant k such that $T = kr^{\frac{3}{2}}$. Thus, writing $c = 2\pi/k$, we have $v = cr^{-(1/2)}$, and hence the acceleration is $a = v^2/r = c^2/r^2$. This is the inverse-square law of gravitational acceleration: *The force per unit mass due to gravity decreases as the square of the distance.*

To make some rough calculations using these considerations, let us compute

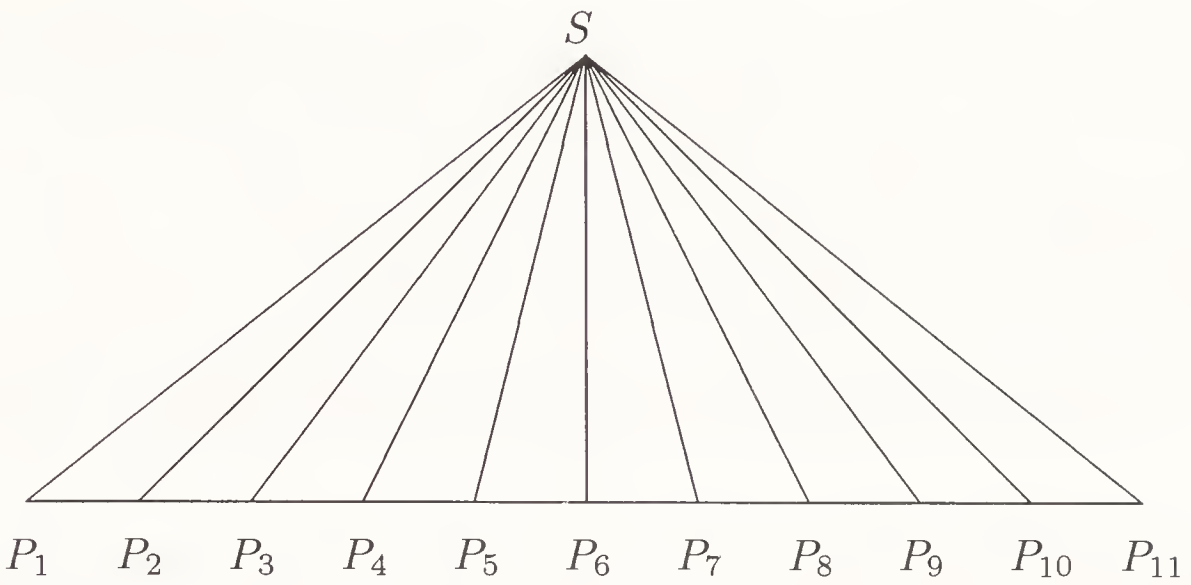


Figure 18.4: Area swept out about a center without attraction.

the acceleration of the moon. Since the earth's radius is about 4000 miles, and the moon's center is about 240,000 miles from the center of the earth (on the average), we can say that the moon is about 60 times as far from the center of the earth as are objects on the earth's surface. Since the acceleration of gravity at the earth's surface is 32 feet per second per second, the acceleration of the moon should be about $\frac{32}{3600}$ feet per second per second, that is, in one second the moon should "fall" about $\frac{16}{3600}$ feet (since the distance fallen s is given by $s = \frac{1}{2}at^2$). This is about $\frac{4}{75}$ inch, that is, $0.05333\ldots$ inch.

Let us compare this computed theoretical value with what is known from observation. The sidereal period of the moon is 27.3 days, that is, $T \approx 2,359,000$ seconds. Since the distance traversed in this time is $2\pi r$, the velocity v is $2\pi r/T$, and the distance fallen in one second is $\frac{1}{2}a = v^2/2r = 2\pi^2 r/T^2$. Since we have used inches as the unit of length in the preceding computation, we have $r = 240,000 \times 5280 \times 12$. When these numbers are inserted, we find that the observed distance fallen in one second is 0.054 inch, a remarkably close agreement of theory with observation. However, we have made many careless approximations—the moon's orbit is not exactly circular, and we have rounded off the sidereal period of the moon and the radius of the earth.

When Newton made this computation, he underestimated the radius of the earth, believing that one degree of arc on the surface was about 60 miles, when in fact it is 69.5 miles. This error threw his computation off by a noticeable amount, and the agreement between the inverse-square law and observation was not very good. However, in 1684 he happened to hear a report of new and more accurate measurements of an arc on the surface of the earth. Returning to the computation, he found the better agreement we have just discussed.

The computation of the shape of an orbit under an inverse-square law of attraction is not easy, but Newton was equal to the task. He showed that the only possible orbits (assuming a fixed central body and only one other body in orbit around it) were conic sections. Since the only closed conic section is an ellipse, it followed that the path of a planet must be an ellipse (Kepler's first law). It was this computation, which had stumped several members of the Royal Society, that

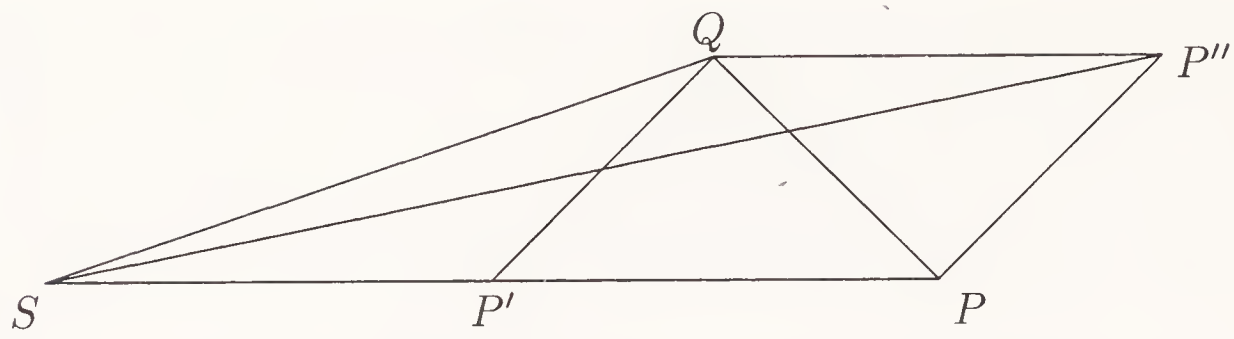


Figure 18.5: Area swept out under a centripetal force.

the astronomer Edmund Halley put to Newton in 1684. Since Robert Hooke had asked him about it in 1679, Newton was ready with the answer. Newton's quick response prompted Halley to urge him to write the *Principia*.

Kepler's third law is likewise a consequence of the inverse-square law of attraction and Newton's second law of motion, but again the mathematics needed to derive it is somewhat complicated. In the *Principia* Newton gave a simple derivation of Kepler's second law in a very general form: *for any object subject to a central attraction, whether inverse-square or some other, the radius from the center to the object sweeps out area at a constant rate.* Since the argument is simple and again illustrates so well the use of infinitesimal analysis in mechanical problems, we give a summary of it. This principle is easily derived if there is no attraction at all toward the center, since in that case, by Newton's first law, the body moves in a straight line at a constant speed, and so in equal time intervals the line from the center sweeps out triangles of equal base (the distance moved by the body) and equal height (the perpendicular distance from the center to the line of motion, as in the 11 positions shown for a "planet" not attracted by a "sun" in Fig. 18.4). For a centrally accelerated motion, shown in Fig. 18.5, Newton considered a particle originally at a point P, which the attraction of S acting alone would cause to fall to the point P' in a given (infinitely short) time, while its motion without any attraction toward S would carry it to P'' in the same time. By the parallelogram law for combining these motions, the body will actually move to the point Q in this time. Hence the area swept out will be the triangle SPQ, which clearly has the same base SP and the same height (equal to the height of the parallelogram $P'PP''Q$) as the triangle SPP'' that would have been swept out without the acceleration. In other words, central acceleration does not change the rate at which area is swept out, which therefore must be constant.

In our study of Greek mathematics we saw that such "atomistic" methods in geometry were laboriously justified by the complicated methods of Eudoxus. The reintroduction of such methods into modern science was bound to cause uneasiness. Newton offered the following excuse for not adhering to Euclidean rigor:

> These Lemmas are premised to avoid the tediousness of deducing involved demonstrations *ad absurdum*, according to the method of the ancient geometers [i.e., by the method of exhaustion]. For demonstrations are shorter by the method of indivisibles; but because the hypothesis of indivisibles seems somewhat harsh, and therefore that method

> is reckoned less geometrical, I chose rather to reduce the demonstrations of the following Propositions to the first and last sums and ratios of evanescent quantities, that is, to the limits of those sums and ratios, and so to premise, as short as I could, the demonstrations of those limits. For hereby the same thing is performed as by the method of indivisibles... . Therefore if hereafter I should happen to consider quantities as made up of particles, or should use little curved lines for right [straight] ones, I would not be understood to mean indivisibles, but evanescent divisible quantities... .

18.2 Electromagnetism and Relativity

18.2.1 Electricity, Magnetism, and Light

During the 1830s Gauss worked with the physicist Wilhelm Weber (1804–1891) on the new subject of electrodynamics. That a current in a loop would affect a compass needle had been discovered in 1820 by the Danish physicist Hans Christian Oersted; (1777–1851) the quantitative expression of the force on a magnet was expressed by Jean Baptiste Biot (1774–1862) and Felix Savart (1791–1841) in the following year. Then over the next 4 years André Marie Ampère (1775–1836) studied quantitatively the effect of one current on another. With amazing rapidity the work of Ampère was followed by an 1827 paper of Georg Simon Ohm (1787–1854), who made extremely delicate measurements of the torque on a needle due to the current in a loop. Ohm found that the relation between the torque X and the length of the conductor x of a given cross-sectional area used to carry the current were related by an equation of the form $X = a/(b + x)$, where a and b are parameters depending on the material of the conductor and the method of generating the electricity. In 1827 the British mathematician George Green (1793–1841) introduced the notion of a potential function for studying these phenomena. Without knowing of Green's work, which was republished by Lord Kelvin (William Thomson, 1824–1907) after Green's death, Gauss independently created the notion of a potential, a function defined in three-dimensional space whose partial derivatives give the components of the force at each point. The potential functions considered by Green are a special case of Gauss' potentials. Gauss was interested in studying the earth's magnetic field, and he developed the potential of this field in a series of negative powers of r (the distance from the earth's center):

$$V = \frac{P_1}{r} + \frac{P_2}{r^2} + \cdots,$$

where P_n satisfies Laplace's equation. Gauss truncated this series after four terms, from which he obtained equations yielding the strength of the earth's magnetic field. In 1845, in a letter to Weber, Gauss suggested a way of computing the interaction of two moving electric charges e and e' at distance r from each other. A year later

Weber gave the mutual force as

$$F = e \cdot e' \left(\frac{1}{r^2} - \frac{1}{c^2 \cdot r^2} \left(\frac{dr}{dt} \right)^2 + \frac{1}{c^2 \cdot r} \frac{d^2 r}{dt^2} \right).$$

This law is known as *Weber's law*. It depends on a velocity c that must be computed experimentally. In 1855 Weber and a collaborator computed this velocity as 4.3945×10^{10} centimeters per second. The following year the physicist Gustav Kirchhoff (1824–1887) noted the interesting fact that Weber's constant velocity was almost exactly $\sqrt{2}$ times the speed of light. However, this coincidence was not explored at the time.

18.2.2 Maxwell

In trying to develop a model of magnetism as a disturbance in an elastic medium referred to as a "magnetic field" the Scottish physicist James Clerk Maxwell (1831–1879) imagined the medium divided into cells surrounded by small spherical particles of electricity. By making reasonable assumptions as to the elasticity of such a medium, Maxwell computed the velocity with which a wave would propagate in it and found it to be 193,088 miles per second. He compared this number with several estimates of the velocity of light given in his day (his own estimate of 192,500 miles per second, based on aberration, and the values of 195,777 and 193,118 given by other authors) to conclude that light must be just such a disturbance in this medium. Interestingly, he seemed to be aware, as good scientists must be, that his own bias and desire for good results could be misleading. He wrote to William Thomson in December 1861 that he had made out his equations before he thought of any connection between the velocity of propagation of electromagnetic waves and the velocity of light. The evidential value of this coincidence would have been greatly decreased if he had—perhaps unconsciously—set up the equations so that the two velocities coincided. He concluded that the magnetic and lumeniferous media were the same and that,

> Weber's number is really, as it appears to be, one-half the velocity of light in millimeters per second. [Maxwell must have meant $\sqrt{2}$ where he wrote one-half.]

This event in 1861 remains one of the outstanding contributions of mathematics to the understanding of the physical world. To ordinary common sense there is not the slightest connection between electricity, magnetism, and light. Yet the mathematical analysis of measurements of electric and magnetic forces revealed that electromagnetic waves *must* travel at exactly the speed of light. The inference that light is an electromagnetic phenomenon was irresistible.

The mechanical model just described proved inadequate for several reasons, and a decade later, in 1873, Maxwell summarized the many individual results on electrical and magnetic interactions in the set of partial differential equations that now bear his name. The American mathematician Josiah Willard Gibbs (1839–1903) developed a compact notation that makes it very easy nowadays to write

Maxwell's equations in terms of the curl and divergence of two vector fields $\mathbf{E}$ (the electric field intensity) and $\mathbf{B}$ (the magnetic induction) that are related in mathematically simple ways to the current density vector $\mathbf{J}$ and charge density ρ:

$$\begin{aligned} \nabla \times \mathbf{E} &= -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \\ \nabla \times \mathbf{B} &= \frac{4\pi}{c}\mathbf{J} + \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \cdot \mathbf{E} &= 4\pi\rho. \end{aligned}$$

The intuitive brilliance of these equations is that, to one practiced in the use of vectors, they convey an immediate and vivid picture and formulate concisely some well-known experimental results. The first equation, for example, says that a magnetic field that changes over time produces an electric field. The second says that currents and changing electric fields produce magnetic fields. The third says that there are no magnetic charges, and the fourth says (when combined with the divergence theorem) that the electric flux through a closed surface is proportional to the charge contained inside that surface.

18.2.3 Relativity

Popular accounts of relativity tend to focus on the constancy of the speed of light, the famous Michelson–Morley experiment, and mechanical effects, such as the contraction of time and space for a body in motion. These observable phenomena, however, were not the starting point for relativity. The theory of relativity arose in connection with electromagnetism, specifically through the Lorentz[1] equation for the force on a particle of charge q moving in an electric field $\mathbf{E}$ and a magnetic induction $\mathbf{B}$ with velocity $\mathbf{v}$:

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}).$$

It is an obvious consequence of Newton's second law that two observers moving with constant velocity relative to each other must agree about the forces acting on any particle. Since they will *not* agree as to the velocity $\mathbf{v}$ of the particle just mentioned, they must disagree about either $\mathbf{E}$ or $\mathbf{B}$, or both. In fact it is not difficult to show that they agree about $\mathbf{B}$ and disagree about $\mathbf{E}$. Thus physicists were faced with a problem in reconciling classical mechanics, which had been very successful, with the new mathematical theory of electromagnetism. Albert Einstein (1879–1955) called attention to these problems in his 1905 paper, "Zur Elektrodynamik bewegter Körper" (On the electrodynamics of moving bodies), in which the special theory of relativity was first introduced.

> It is known that Maxwell's electrodynamics (in its current interpretation) when applied to moving bodies, leads to asymmetries which do

[1]Named after the Dutch physicist Hendrik Antoon Lorentz (1853–1928).

> not appear to be intrinsic to the phenomena. Consider, for example, the electrodynamic interaction of a magnet and a conductor. The observable phenomenon here depends only on the relative motion of the conductor and the magnet, whereas the standard interpretation draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion... Examples of this kind... lead to the conjecture... that... the same laws of electrodynamics and optics will be valid for all systems of coordinates in which the equations of mechanics hold good. We wish to make this conjecture (whose content will be called the "Principle of Relativity" from now on) into a postulate...

The success of the special theory of relativity in removing the asymmetries of electrodynamics was a powerful argument in its favor; the equivalence of matter and energy expressed by the famous equation $E = mc^2$ was an added bonus. The theory also had some experimental justification from the Michelson-Morley experiment, which had not detected any change in the velocity of light traveling in different directions relative to the moving earth.

The fact that such seemingly concrete and indisputable things as distance and time would have to be different for two observers in order for mechanics and electromagnetism to be mathematically consistent came as a surprise. The fact that two observers could agree on the magnitude of a force but disagree about its physical nature raised the question as to how two observers could know when they were measuring the same physical quantity and using the same physical laws. The special theory of relativity considers how to formulate physical laws so that two observers moving with constant relative velocity can reconcile their observations.

The more difficult problem as to how to reconcile the observations of two observers in arbitrary relative motion took another decade to consider. Here again, Einstein was guided to a large degree by the mathematics available, in this case the tensor calculus. Mathematically two different observers are represented by two different coordinate systems $\{x^i\}$ and $\{y^i\}$ related by equations $y^1 = \varphi_1(x^1, \ldots, x^n), \ldots, y^n = \varphi_n(x^1, \ldots, x^n)$. Here the superscripted variables x^i and y^j can represent any measurable physical quantities. By the well-known rules of differential calculus

$$dy^i = \sum_{j=1}^{n} \frac{\partial \varphi_i(x^1, \ldots, x^n)}{\partial x^j} dx^j = \sum_{j=1}^{n} \frac{\partial y^i}{\partial x^j}\, dx^j.$$

The differentials dx^i and dy^j are said to be *covariant* because they transform in this way. That is, in matrix language the chain rule becomes

$$\begin{pmatrix} dy^1 \\ \vdots \\ dy^n \end{pmatrix} = \begin{pmatrix} \frac{\partial y^1}{\partial x^1} & \cdots & \frac{\partial y^1}{\partial x^n} \\ \vdots & & \vdots \\ \frac{\partial y^n}{\partial x^1} & \cdots & \frac{\partial y^n}{\partial x^n} \end{pmatrix} \begin{pmatrix} dx^1 \\ \vdots \\ dx^n \end{pmatrix} = J \begin{pmatrix} dx^1 \\ \vdots \\ dx^n \end{pmatrix},$$

On the other hand, if

$$z = \psi(y^1, \ldots, y^n) = \psi(\varphi_1(x^1, \ldots, x^n), \ldots, \varphi_n(x^1, \ldots, x^n))$$

is any function of these variables, the partial derivatives of z are *contravariant*, since, for example,

$$\begin{pmatrix} \frac{\partial z}{\partial y^1} \\ \vdots \\ \frac{\partial z}{\partial y^n} \end{pmatrix} = \begin{pmatrix} \frac{\partial x^1}{\partial y^1} & \cdots & \frac{\partial x^n}{\partial y^1} \\ \vdots & & \vdots \\ \frac{\partial x^1}{\partial y^n} & \cdots & \frac{\partial x^n}{\partial y^n} \end{pmatrix} \begin{pmatrix} \frac{\partial z}{\partial x^1} \\ \vdots \\ \frac{\partial z}{\partial x^n} \end{pmatrix}.$$

Quantities that transform in these ways are known as *tensors*. Einstein suggested that different observers could agree that they were measuring the same quantity if their measurements transformed in this way. He took the position that physical laws should be stated as tensor equations, so that any law can be translated into any coordinate system.

It should be noticed that the transformations just considered involved only one differential per term. However the notion of infinitesimal distance for the geometry of a curved space that Riemann had proposed involved a product of two differentials: $ds^2 = \sum_{\mu,\nu} g_{\mu\nu}\, dx^\mu\, dx^\nu$. Investigating the transformations of such a tensor, Einstein noted that if the coefficients $g_{\mu\nu}$ were constant in some coordinate system, then a certain general tensor $R^\rho_{\mu\nu\tau}$ called the Riemann–Christoffel[2] tensor derived from this one would have all coefficients zero. The Riemann–Christoffel tensor is of rank 4, having three covariant indices and 1 contravariant index. By an algebraically simple transformation called contraction, the contravariant index ρ and the covariant index τ can be made to "annihilate" each other, leaving a covariant tensor of rank 2, which in a suitable coordinate system assumes a particularly simple form $G_{\mu\nu}$. Einstein required the gravitational fields to be such that $G_{\mu\nu} = 0$ in the absence of matter. Einstein stated explicitly that in making this choice he was guided by the fact that $G_{\mu\nu}$ was the only tensor of second rank formed from the $g_{\mu\nu}$ and their derivatives not involving any derivatives of order higher than 2 and expressible as a linear function of those derivatives. In other words, it was the mathematically simplest tensor that said anything significant about the geometry of space. Einstein's tacit assumption is that *one should seek physical laws in the simplest possible mathematical form*. This principle allows mathematics to guide the discovery of physical law. From this point of view observation and experiment do not always *suggest* physical laws by revealing patterns; sometimes they are used instead to *test* laws arrived at on esthetic and mathematical grounds. On the other hand, one of Einstein's goals in creating the general theory of relativity was to show that relativistic considerations could account for discrepancies in planetary orbits, specifically a precession of the perihelion of Mercury by 43 seconds of arc per century that could not be derived from perturbations due to the other planets. Two years before he published his general theory Einstein had rejected an earlier version when it predicted a precession of only 18 seconds per century for the perihelion of Mercury.

The equation $G_{\mu\nu} = 0$ gave a system of differential equations for the functions $g_{\mu\nu}$, and mechanics could be formulated by saying that the path of a particle in the resulting geometry would be a geodesic, a principle similar to the principle of least action in classical mechanics.

[2]Named after Riemann and Elwin Bruno Christoffel (1829–1900).

The justification of this approach was given by Einstein in his 1916 paper on general relativity. After showing that the simplest first-order approximation to his law was Newton's law of gravity, he compared the difference between his law of motion and that of Newton for the case of the planet Mercury and found that the elliptical orbit of Newtonian mechanics was replaced by a more complicated orbit. In second approximation this more complicated orbit was an ellipse whose axis rotated (precessed) by an amount he calculated to be 43 seconds of arc per century. By a harmony too improbable to be accidental, this was exactly the amount of precession that astronomers had been unable to account for as the result of perturbations due to the other planets. As Einstein said, "These facts must, in my opinion, be taken as convincing proof of the correctness of the theory."

We have discussed electrodynamics and relativity in order to show how physicists trying to explain the natural world are often guided by mathematical elegance and simplicity in conjecturing the laws of nature. One should, however, balance these successes against the wreckage of past mathematical theories (the elastic solid theory of light propagation, for example) that seemed to have great success for a time, but were eventually overwhelmed by stubborn, unresolvable difficulties. Even these outmoded theories, however, were often essential stepping stones on the way to more comprehensive and satisfying theories. The mystery of this "preordained harmony" between mathematics and the physical world continues to inspire awe in scientists.

18.3 Questions about Mathematical Physics

Exercise 18.1 How do you answer the objection that motion cannot begin if the velocity of a falling body is zero when it begins to fall?

Exercise 18.2 Show that the distance traversed under uniformly accelerated motion, starting from rest, is proportional to the square of the time elapsed, as asserted by Galileo.

Exercise 18.3 If Galileo is correct that any centripetal force will overcome any tangential force, why *do* the sparks from a grinding wheel fly off at a tangent? What would happen if the rotation of the earth gradually speeded up until a person standing on the equator weighed nothing, and then the speed of rotation increased still further? Would the person fly off at a tangent, as Simplicio had argued?

Exercise 18.4 Newtonian mechanics (neglecting friction and air resistance) predicts theoretically what Galileo claimed to have observed, that two spheres of different sizes will require exactly the same time to roll down an inclined plane. Yet this same model predicts that a hoop and a sphere will require different times to roll down the plane. What is the explanation for this difference? Did Galileo actually observe what he claimed to observe?

Exercise 18.5 Bertrand Russell, in his *History of Western Philosophy*, writes, "Kepler is one of the most notable examples of what can be achieved by patience without much in the way of genius." Is this a fair verdict on Kepler's work?

Exercise 18.6 Maxwell assured William Thomson that he had not "cooked the books" in his theory so that the theoretical speed of propagation of electromagnetic waves would turn out to be equal to the measured speed of light. Einstein, on the other hand, from the very beginning, wanted a relativistic theory of gravity that would explain the precession of the perihelion of Mercury and adjusted his physical theory until he got that result. Does this fact decrease the value of this explanation as evidence in favor of general relativity? What about our back-of-the-envelope computation of the distance the moon falls each second? Was the outcome influenced in any way (when Newton did it) by a desire to show that the acceleration due to gravity decreases as the square of the distance?

18.4 Endnotes

1. The two quotations from Kepler's *Astronomia nova* are translated from his collected works, published by C. H. Beck'sche Verlagsbbuchhandlung (Munich), Vol. 3, 1937, pp. 263, 366.

2. The quotation from Kepler's *Harmonice mundi* is translated from his collected works (sp. cit.), Vol. 5, 1940, p. 302.

3. The quotation from Descartes' *Principia Philosophiae* is translated from the French translation found in his *Œuvres*, Vol. 3 (Levrault, Paris, 1824), p. 151.

4. The quotation from Newton's *Principia* is taken from F. Cajori's revision of Motte's 1729 translation, published by the University of California Press (Berkeley and Los Angeles, 1966), p. 38.

5. Maxwell's letter to Thomson can be found in *The Scientific Letters and Papers of James Clerk Maxwell*, Vol. 1 (Cambridge University Press, 1993), p. 695.

6. The quotation from Einstein's paper on special relativity is translated from his original paper, "Zur Elektrodynamik bewegter Körper," *Annalen der Physik*, **17** (1905), p. 891; see *Collected Papers of Albert Einstein* (Princeton University Press, 1989), Vol. 2, p. 276.

7. Einstein's heuristic path to his law of gravity can be found in *The Principle of Relativity* (Dover, New York, 1952), p. 144.

8. The information on Einstein's rejected draft of a general theory of relativity is taken from *The Collected Papers of Albert Einstein*, Vol. 4, Martin J. Klein, A. J. Kox, Jürgen Renn, and Robert Schulmann, eds. (Princeton University Press, 1985).

Chapter 19

Contemporary Mathematics

The narrative up to this point has carried the story of mathematics to the end of the nineteenth century. In this final chapter we shall look at some parts of twentieth-century mathematics, emphasizing the way in which mathematics has been practiced. Both internal and external forces have helped to shape this practice, and the study of these forces involves sociology and philosophy in ways that would soon take us out of our depth if we were to attempt to account for many details. The questions deserve to be examined, however. Decisions are being taken every day, both by mathematicians and by the leaders of business and government, which determine what mathematics will be in the future. As a citizen and a potential user or practitioner of mathematics, the student ought to have some idea of how mathematics is practiced and what it can contribute to the solution of economic and social problems.

We shall begin by discussing the internal changes in the nature of mathematical research, specifically generalization, abstraction, and rigorization, after which we shall address the social and political aspects of the practice of mathematics.

19.1 Generalization and Abstraction

The most prominent feature of twentieth-century mathematics, compared with that of the past, is the high level of abstraction and generality of its results. Throughout the century there has been a concerted effort to examine the hypotheses and methods of proof of major theorems, to strip away the inessential parts and reduce hypotheses to a minimum. We shall illustrate this trend with examples from several major areas of mathematics.

19.1.1 Analysis

While the nineteenth-century mathematicians were concerned with achieving a clear definition of continuity that did not rely on vague intuition, they needed this definition only for real- and complex-valued functions of real and complex

variables. In the early twentieth century, however, Maurice Fréchet (1878–1973) pointed out that the algebraic properties of real and complex numbers really played no role in the definition of continuity. What was essential to the definition was only a notion of convergence, of *nearness*, so that it would make sense to say that the distance between $f(x)$ and $f(y)$ could be made as small as desired by choosing x and y sufficiently close together. In this way continuity could be defined on any space in which a notion of distance was defined.

Felix Hausdorff (1868–1942) noted in his 1914 book *Grundzüge der Mengenlehre* (Fundamentals of Set Theory) that nearness could be defined without mentioning the notion of distance. The concept of distance was of use in defining the interior and boundary of a set. Once that definition was made, it was possible to consider the class of open sets (sets that do not intersect their boundaries). Hausdorff pointed out that open sets could be defined in terms of a concept of *neighborhood*, which is qualitative rather than quantitative. The class of open sets is characterized by two properties: (1) any union or finite intersection of open sets is an open set, and (2) the empty set and the entire space are open. Any collection of sets with these properties defines a *topology*. Once a topology is defined on the domain and range of a function, continuity of a function (at all points) is defined by saying that the set of points that map into any open set of the range is an open subset of the domain. This definition allows much more general spaces to be studied than the spaces of real and complex variables considered previously. Throughout the twentieth century analysts, topologists, and geometers have found this abstract notion of a topological space essential in their work.

This kind of abstraction can easily get out of hand, and one of the efforts to rein it in consists of theorems showing that an abstractly defined object must actually resemble a more traditional one. A good example of this kind of theorem is the Weierstrass approximation theorem, which asserts that a function about which nothing is known except that it is continuous on an interval $[a, b]$ can be approximated with any required precision by a polynomial (in fact by a polynomial with rational coefficients, hence an eminently computable object). Thus the exceedingly abstract object known as a continuous function can be represented for all practical purposes by a polynomial, a very concrete object. However, the abstractionists have not been silenced by this result, and it too has been generalized to a modern version known as the Stone–Weierstrass theorem after its discoverer Marshall Stone (1903–1989). The Stone–Weierstrass theorem asserts that any algebra of continuous real-valued functions (such as the polynomials) defined on a compact Hausdorff space (an example of which is the interval $[a, b]$), on which the algebra separates points and for each point x contains a function that is not zero at x, is dense in the space of all continuous real-valued functions, so that any continuous function can be approximated by functions of the algebra.

A similar generalization and abstraction has shaped the development of another major property of real-valued functions: integrability. In the work of Émile Borel (1871–1956), Henri Lebesgue (1875–1941), Pierre Fatou (1878–1929), W. H. Young (1863–1942), and others in the period from 1890 to 1910, it was shown that the essential components of a theory of integration are (1) a collection of sets, each of which has a "measure" (length, area, volume, or some generalization of

these—such a set is called *measurable*) and (2) a real-valued function $f(x)$ such that the set of points x where $a < f(x) < b$ is a measurable set for each a and b (such a function is said to be *measurable*). The class of measurable sets, called a σ-field, resembles the concept of a topological space, except that measurable sets are closed under countable unions and intersections rather than arbitrary unions and finite intersections. Analysts now routinely work with abstract spaces on which both a topology and a σ-field are defined. To limit the apparent abstraction in this area Nikolai Nikolaevich Luzin (1883–1950) proved in 1915 that any measurable function is the derivative of a continuous function, and Marshall Stone proved a classification theorem showing that an abstract measure space can be modeled by combining the measurable sets in a Euclidean space with a set of discrete points.

Examples of this kind of abstraction can be given almost without limit. The classical[1] theory of Fourier series involved Fourier series of periodic functions and Fourier transforms of integrable functions. What these two topics had in common was the idea of transforming a function defined on one group (the "circle," obtained by identifying the two endpoints of a closed interval, or the real line) into a different function (its sequence of Fourier coefficients, or its Fourier transform) defined on another group (the integers or the line) called the *dual* of the original group. These groups possess a topology and a measure that are invariant under the group operations (the translate of an open or measurable set is open or measurable and has the same measure as the original set). Groups having such topologies and measures were already known—the groups called the "classical groups" after the title of a famous treatise by Hermann Weyl (1885–1955). The latter are groups of invertible matrices, the group operation being matrix multiplication. A square matrix of size $n \times n$ can be regarded as an element of n^2-dimensional Euclidean space, providing a natural metric topology on any such group. The construction of an invariant measure on any such group was carried out by Alfred Haar (1885–1933).

Even greater abstraction and generality was achieved in the work of Frigyes Riesz (1880–1956), Maurice Fréchet, and others from 1900 to 1910, by regarding functions themselves as elements of a space having the algebraic structure of a vector space on which a metric is defined. Part of the impetus to this construction came from the work of David Hilbert (1862–1943) on integral equations in this same period. Hilbert worked with the space of square-integrable functions, which is now called a Hilbert space in his honor. Riesz, Fréchet, and others worked with spaces of continuous functions. The end result was summed up in the work of Stefan Banach (1892–1945), who studied an abstract class of vector spaces having a metric; these are now called Banach spaces.

When the elements of a Banach space can be multiplied, the result is a richer structure called a Banach algebra. The theory of Banach algebras led to yet another generalization of Fourier series, as I. M. Gel'fand (b. 1913) showed in 1940 that the elements of such an algebra can be transformed into continuous functions on

[1]The word *classical* has special meanings in mathematics. It is most often applied to a famous result that has been known for a long time, originally stated in the limited context of Euclidean space or real and complex variables, to contrast it with modern abstract generalizations of the theorem. More loosely, however, it is used to distinguish any concrete original result from its later more abstract forms.

a topological space called the *maximal ideal space* of the algebra. The Fourier transform can be considered a special case of this abstract Gel'fand transform.

19.1.2 Algebra

One can trace a similar development in algebra, where abstraction and generalization began in the late nineteenth century with the creation of the concepts of ideals and fields. The study of equations, which had begun with symbols assumed to be representing complex numbers, could now be pursued with symbols representing elements of an arbitrary field. The theory of finite groups, created by Galois to decide whether equations could be solved by radicals, worked in this wider setting. The elements of a Galois group are permutations of the roots of an equation. What is essential in the study of the group, however, is its multiplication table; one need not know exactly which roots are mapped to which in order to analyze the group. Thus the modern concept of an abstract group arises, a structure whose elements are arbitrary on which a binary operation satisfying a few simple axioms is defined.

By the early twentieth century a set of standard objects—groups, rings, fields, vector spaces, and others—made up the universe of algebra. What these objects had in common was that they were composed of elements that could be combined two at a time to produce new elements. Inevitably the abstract question arose: What kinds of properties can be proved about a collection of elements on which a collection of unspecified operations is defined, each of which combines a certain number of elements to produce a new element? The result of trying to answer questions like this is the contemporary subject known as universal algebra.

In algebra also the attempt to give general structure theorems showing that an abstractly defined object can be built out of certain concrete specific parts is exemplified by one of the great triumphs of the twentieth century—the classification of all finite groups. From Galois theory came the concept of a solvable group—one having a finite nested sequence of normal subgroups (defined in Chapter 17) such that the "quotient groups" (the original groups with two elements regarded as identical if one can be obtained from the other by multiplying by an element of the subgroup) have a prime number of elements. The solvable group is built from these quotient groups in a natural way, and the structure of a group having a prime number of elements is completely understood. The major question of the century was whether every group having an odd number of elements is solvable. The affirmative answer to this question came in 1963 from John Thompson (b. 1932) and Walter Feit (b. 1930). After this giant leap, a number of smaller steps were required—to classify a small set of "sporadic" finite groups (computers were used for this work). The work is now finished and remains an impressive monument to the algebraists of this era.

Category Theory

The applications of algebra in geometry bring out certain strong analogies between the two subjects. Topological spaces correspond to homology and homo-

topy groups, though the correspondence is not perfect. Continuous mappings of one space into another (that is, mappings that preserve the topological structure) correspond to homomorphisms of the associated groups (mappings that preserve the algebraic structure). If the topological space is a surface, it may have a tangent plane, which has the structure of a vector space. Then a mapping from one surface to another that is differentiable corresponds to its differential, which is a linear transformation of the corresponding tangent spaces. Such analogies led to a new subject of study: category theory, which arose in 1942 in the work of Saunders MacLane (1909–1995) and Samuel Eilenberg (b. 1913). In category theory the basic elements are sets of "objects" and mappings among the objects called "morphisms." The objects may be vector spaces and the morphisms linear transformations, or the objects may be topological spaces and the morphisms continuous functions. All these particular objects are encompassed in the more general subject of category theory.

19.1.3 Geometry

Abstraction in geometry came from the generalization of surfaces in Euclidean space to imaginative constructs (called *complexes*) that may be physically unrealizable. These imaginative constructs are formed from primitive elements (called *simplexes* or *cells*) such as triangles and tetrahedra by imposing rules for identifying points of different elements (gluing them together in the imagination). In this way such objects as the projective plane and the Klein bottle as well as a host of higher-dimensional objects could be studied with complete clarity. (It required the point-set topology of Hausdorff and others, however, to give a logically acceptable interpretation of this kind of gluing.) Since the simplexes or cells of the complex could be specified by merely writing down a finite sequence of symbols, algebra could be used to study the geometric properties of the complex objects. The inspiration for doing so came partly from complex analysis (Riemann surfaces), in which it was necessary to integrate over paths while avoiding certain singular points. The resulting abstraction formed the foundation of the theory of homology and homotopy in algebraic topology. Since the objects were glued together from pieces of Euclidean space, it became worthwhile to investigate the properties of an object having a topology (collection of open sets) in which each point is surrounded by an open set that has the same topological structure as a ball in Euclidean space. Such an object is called a *manifold*, and manifolds are now a major topic of study in geometry and a tool in analysis.

Here also structure theorems play a role in showing that the abstract objects are not completely unknowable. The classical objects of study in geometry were surfaces in Euclidean space. A classification theorem, due to Hassler Whitney (1907–1989), asserts that a manifold of dimension m can be regarded as a subset of Euclidean space of dimension $2m+1$. Thus any one-dimensional manifold can be imbedded in three-dimensional space (as is well known intuitively, since any conceivable curve can be actually realized in a physical model using thread). On the other hand, there are conceivable surfaces—the Klein bottle and the projective

plane, for example—for which no faithful physical model can be constructed.

19.1.4 Probability

One subject that became clearer as it became more abstract is probability theory. The question of what constitutes an event or a random variable was finally answered in the early twentieth century, thanks to the advances in integration theory described above. The notion of an abstract measure space turned out to provide the key. A space whose total measure equals 1 can be called a probability space. In applications the points of such a space can be the possible outcomes of an observation or experiment. Events are sets of such points for which probability is defined, that is, measurable sets, and a random variable is a simply a measurable function $f(x)$. This interpretation of probability is essentially the one proposed by Andrei Nikolaevich Kolmogorov (1903–1987) in 1933. Probability is not simply subsumed in the theory of integration, however. It is concerned with special aspects of random variables, such as independence and stochastic processes, which are not of concern in the general theory of integration.

19.2 Foundations of Mathematics

One area of mathematics affects all the others so strongly that it deserves a thorough discussion. That area is logic and set theory.

This area arose from thinking not about mathematical problems, but about the nature of mathematics itself, in particular, the grounds on which mathematical results are accepted or rejected by the community of mathematicians. Mathematics cannot produce an algorithm for solving every problem. Mathematics textbooks would be much poorer without theorems like the fundamental theorem of algebra, which asserts that for any polynomial p there exists a complex number z (whatever *exists* means when applied to incorporeal objects like numbers) satisfying the equation $p(z) = 0$. Mathematicians tend to believe that certain statements are either true or false, even when they do not know which is the case. Such questions aroused debate around the beginning of the twentieth century, involving not only mathematicians, but also philosophers such as Alfred North Whitehead (1861–1947) and Bertrand Russell.

19.2.1 The Progress of Set Theory, 1870–1900

As we saw in Chapter 17, Georg Cantor had discovered ordinal numbers through the study of the derived sets of a set. In the 1880s he also discovered the definition of cardinal numbers through the concept of one-to-one correspondence. Galileo had noticed that large circles could be placed in such a correspondence with small circles. He thought this was merely a puzzle of the infinite. Cantor took it as the definition of infinite cardinality. Independently of Cantor other mathematicians

were also considering ways of deriving mathematics logically from simplest principles. Gottlob Frege (1848–1925), a professor in Jena, who occasionally lectured on logic, attempted to establish logic on the basis of "concepts" and "relations" to which were attached the labels *true* or *false*. He was the first to establish a complete predicate calculus, and wrote in 1884 a treatise called *Grundgesetze der Arithmetik* (Foundations of Arithmetic). Meanwhile in Italy, Giuseppe Peano (1858–1939) attempted (1889) to axiomatize the natural numbers. Peano took the successor relation as fundamental and based his construction of the natural numbers on this one relation and nine axioms, together with a symbolic logic that he had developed. The work of Cantor, Frege, and Peano attracted the notice of a young student at Cambridge, Bertrand Russell, who had written his thesis on the philosophy of Leibniz. Russell saw in this work confirmation of a thesis that he advocated throughout the rest of a long life: that mathematics is merely a prolongation of formal logic. This view, that mathematics can be deduced from logic without any new axioms or rules of inference, is now called *logicism*. It encountered difficulties from its beginning, however, as we shall now see. Even the seemingly primitive notion of membership in a set turned out to require certain caveats.

19.2.2 Paradoxes

Cantor defined equality between cardinal numbers as the existence of a one-to-one correspondence between sets representing the cardinal numbers. A set B has larger cardinality than set A if there is no function $f : A \to B$ that is "onto," that is, such that every element of B is $f(x)$ for some $x \in A$. Cantor showed that the set of all subsets of A, which he denoted 2^A, is always of larger cardinality than A, so that there can be no largest set. For if $f : A \to 2^A$, the set $C = \{t \in A : t \notin f(t)\}$ is a subset of A, hence an element of 2^A, and it cannot be $f(x)$ for any $x \in A$. For if $C = f(x)$, we ask whether $x \in C$ or not. If $x \in C$, then $x \in f(x)$ and so by definition of C, $x \notin C$. On the other hand, if $x \notin C$, then $x \notin f(x)$, and again by definition of C, $x \in C$. Since the whole paradox results from the assumption that $C = f(x)$ for some x, it follows that no such x exists, that is, the mapping f is not "onto." This argument was at first disputed by Russell, who wrote in an essay entitled "Recent work in the philosophy of mathematics" (1901) that "the master has been guilty of a very subtle fallacy." Russell thought there must clearly be a largest set, namely the set of all sets. What fallacy he thought Cantor had committed is not clear, since in a later reprint of the article he added a footnote explaining that Cantor was right.

In fact, if we consider the set of all sets, as Russell had suggested, we must, by its very definition, believe it to be *equal* to the set of all its subsets. Therefore the mapping $f(E) = E$ should have the property that Cantor says no mapping can have. Now if we apply Cantor's argument to this mapping, we are led to consider $S = \{E : E \notin E\}$. By definition of the mapping f we should have $f(S) = S$, and so, just as in the case of Cantor's argument, we ask if $S \in S$. Either way, we are led to a contradiction. This result is known as *Russell's paradox*.

19.2.3 The Debate over the Axiom of Choice

The trend toward abstraction and generalization that we discussed above has meant that much of the action in a proof takes place "offstage." That is, certain objects needed in the proof are shown to exist, but no procedure for constructing them is given. Proofs relying on the abstract existence of such objects, when it is not possible to choose a particular object and examine it, became more and more common in the twentieth century. Indeed much of measure theory, topology, and functional analysis would be impossible without such proofs. The principle behind these proofs later came to be known as *Zermelo's axiom*, after Ernst Zermelo (1871–1953), who first formulated it in 1904 in order to prove that every set could be well ordered.[2] It was also known as the principle of free choice (in German, *Auswahlprinzip*) or, more commonly in English, the axiom of choice. In its broadest form this axiom states that *there exists* a function f defined on the class of all nonempty sets such that $f(A) \in A$ for every nonempty set. (Intuitively, if A is nonempty, there exist elements of A, and $f(A)$ chooses one such element from every nonempty set.)

Zermelo made this axiom explicit and showed its connection with ordinal numbers. The problem then was either to justify the axiom of choice, or to find a more intuitively acceptable substitute for it, or to find ways of doing without such "non-effective" concepts.

A debate about this axiom took place in 1905 in the pages of the *Comptes Rendus* of the French Academy of Sciences, with arguments for and against it being contributed by a number of mathematicians and philosophers. The achievements and the program of the logicists were presented in a systematic work by Russell and Whitehead in 1910 entitled *Principia Mathematica*.

19.2.4 Formalism

A different view of the foundations of mathematics, known as *formalism*, was advanced by Hilbert, who was interested in the problem of axiomatization (the axiomatization of probability theory was the sixth of his famous 23 problems) and particularly interested in preserving as much as possible of the freedom to reason that Cantor had provided while avoiding the uncomfortable paradoxes of logicism. In the formalist view mathematics is the study of formal systems. This view involves a strict separation between the symbols and formulas of mathematics and the meaning attached to them, that is, a distinction between syntax and semantics. A given formal system consists of certain rules for recognizing legitimate formulas, certain formulas called axioms, and certain rules of inference (such as syllogism, generalization over unspecified variables, and the rules for manipulating equations). These rules make up the syntax of the language. One can therefore always tell by following clearly prescribed rules whether a formula is meaningful (well formed), and whether a sequence of formulas constitutes a valid deduction. To avoid infinity

[2] A set is *well ordered* if any two elements can be compared and every nonempty subset has a smallest element. The positive integers are well ordered. The positive real numbers are not.

in this system while preserving sufficient generality Hilbert resorted to a "finitistic" device called a *schema*. Certain basic formulas are declared to be legitimate by fiat. Then a few rules are adopted, such as the rule that if A and B are legitimate formulas, so is $[A \Rightarrow B]$. This way of prescribing legitimate (well-formed) formulas makes it possible to determine in a finite number of steps whether a formula is well formed or not.

The formalist approach makes a distinction between statements *of* arithmetic and statements *about* arithmetic. For example, the assertion that there are no positive integers x, y, z such that $x^3 + y^3 = z^3$ is a statement *of* arithmetic. The assertion that this statement can be derived from the axioms of arithmetic is a statement *about* arithmetic. The language in which statements are made about arithmetic, called the *metalanguage*, contains all the meaning to be assigned to the statements. In particular it becomes possible to distinguish clearly between what is true (that is, what can be known to be true from the metalanguage) and what is provable (what can be deduced within the object language). Two questions thus arise in the metalanguage: (1) *Is every deducible proposition true?* (the problem of consistency); (2) *Is every true proposition deducible?* (the problem of completeness).

19.2.5 Intuitionism

The most cautious approach to the foundations of mathematics, known as *intuitionism*, was championed by the Dutch mathematician Luitzen Egbertus Jan Brouwer (1881–1966). In a series of articles published form 1918 to 1928 Brouwer laid down the principles of this school of mathematicians. These principles include the rejection not only of the axiom of choice, but also of proof by contradiction. Roughly speaking, intuitionists reject any proof whose implementation leaves choices to be made by the reader. Thus it is not enough in an intuitionist proof to say that objects of a certain kind exist. One must choose such an object and use it for the remainder of the proof. This extreme caution has rather drastic consequences. For example, the function $f(x)$ defined in ordinary language as

$$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

is not considered to be defined by the intuitionists, since there are ways of defining numbers x that do not make it possible to determine whether the number is negative or positive. [For example, is the number $(-1)^n$, where n is the trillionth decimal digit of π, positive or negative?] This restrictedness has certain advantages, however. The objects that are acceptable to the intuitionists tend to have pleasant properties. For example, every rational-valued function of a rational variable is continuous.

The intuitionist rejection of proof by contradiction needs to be looked at in more detail. Proof by contradiction was always used somewhat reluctantly, since such proofs seldom give insight into the structures being studied. For example, Euclid's proof that there are infinitely many primes proceeds by assuming that the

set of prime numbers is a finite set $P = \{p_1, p_2, \ldots, p_n\}$ and showing that in this case the number $1 + p_1 \cdots p_n$ must either itself be a prime number or be divisible by a prime different from $p_1, \ldots, p_n$, which contradicts the original assumption that $p_1, \ldots, p_n$ formed the entire set of prime numbers.

The appearance of starting with a false assumption and deriving a contradiction can be avoided here by stating the theorem as follows: If there exists a set of n primes $p_1, \ldots, p_n$, then there exists a set of $n + 1$ primes. The proof is exactly as before. Nevertheless, the proof is still not intuitionistically valid, since there is no way of saying whether $1 + p_1 \cdots p_n$ is prime or not.

A purely intuitionist mathematics is obviously going to be somewhat sparser in results than mathematics constructed on more liberal principles.

19.2.6 Clarification of the Difficulties

The most influential figure in mathematical logic during the twentieth century was Kurt Gödel (1906–1978). The problems connected with consistency and completeness of arithmetic, the axiom of choice, and many others all received a fully satisfying treatment at his hands that settled many old questions and opened up new areas of investigation. In 1931, he astounded the mathematical world by producing a proof along strictly finitistic Hilbertian lines that any consistent formal language in which arithmetic can be encoded is necessarily incomplete, that is, contains statements that are true according to its metalanguage but not deducible within the language itself. The intuitive idea behind the proof is a simple one, based on the self-destroying statement that follows:

This statement cannot be proved.

If one asks whether the statement just given is *true*, the answer must be positive if the system in which it is made is consistent. For if this statement is false, then it can be proved; and in a consistent deductive system, a false statement certainly cannot be proved. Hence we agree that the statement is true, but by its very nature it cannot be proved.

The example just given is really nonsensical, since we have not carefully delineated the universe of axioms and rules of inference in which the statement is made. The word "proved" that it contains is not really defined. Gödel, however, took an accepted formalization of the axioms and rules of inference for arithmetic from *Principia Mathematica* and showed that the metalanguage of arithmetic could be encoded within arithmetic. In particular each formula can be numbered uniquely, and the statement that formula n is (or is not) deducible from those rules can itself be coded as a well-formed formula. Then, when n is chosen so that the statement, "Formula number n cannot be proved" happens to *be* formula n, we have exactly the situation just described. Gödel showed that such an n can be constructed in ways that Hilbert would have accepted. Thus, if Gödel's version of arithmetic is consistent, it contains statements that are formally undecidable, that is, true (based on the metalanguage) but not deducible. This is Gödel's first incompleteness theorem. His second incompleteness theorem is even more interesting: The assertion that arithmetic is consistent is one of the formally undecidable statements. Hence

if the formalized version of arithmetic that Gödel considered is consistent, it is incapable of proving itself so. It is doubtful, however, that one could truly formalize every kind of argument that a rational person might produce. For that reason, great care should be exercised in drawing inferences from Gödel's work to the actual practice of mathematics. In fact, over the last decade Edward Nelson of Princeton University and others have shown how to reconstruct a good portion of mathematics within a set of rules for which a finitary proof of consistency exists, and computer-implemented proof checkers have been written that can read a file written for the most popular mathematical typesetter (TEX,[3] the language in which this book has been written) and verify proofs.

19.2.7 The Aftereffects

The axiom of choice is ubiquitous in modern analysis; almost none of functional analysis or point-set topology would remain if it were omitted entirely (although much weaker assumptions might suffice). It is fortunate, therefore, that its consistency and independence of the other axioms of set theory has been proved. However, the consequences of this axiom are suspiciously strong. In 1924 Alfred Tarski (1901–1983) and Stefan Banach deduced from it that any two sets A and B in ordinary three-dimensional Euclidean space, each of which contains some ball, can be decomposed into pairwise congruent subsets. This means, for example, that a cube the size of a grain of salt (set A) and a ball the size of the sun (set B) can be written as disjoint unions of sets $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ respectively such that A_i is congruent to B_i for each i. This result (the Banach–Tarski paradox) is very difficult to accept. It can be rationalized only by realizing that the notion of existence in mathematics has no metaphysical content. To say that the subsets A_i, B_i "exist" means only that a certain formal statement beginning $\exists \ldots$ is deducible from the axioms of set theory.

The paradoxes of naive set theory (such as Russell's paradox) were found to be avoidable if the word *class* is used loosely, as Cantor had previously used the word *set*, but the word *set* is restricted to mean only a class that is a member of some other class. (Classes that are not sets are called *proper classes*.) Then in order to belong to a class A, a class B must not only fulfill the requirements of the definition of the class A, but must be known in advance to belong to some (possibly different) class.

This approach avoids Russell's paradox. The class $C = \{x : x \notin x\}$ is a class; its elements are those classes that *belong to some class and* are not elements of themselves. If we now ask the question that led to Russell's paradox—Is C a member of itself?—we do not reach a contradiction. It is true that if we assume $C \in C$, then we conclude that $C \notin C$, so that this assumption is not tenable. However, the opposite assumption, that $C \notin C$, *is* acceptable. It no longer leads to the conclusion that $C \in C$. For an object x to belong to C, it no longer suffices that $x \notin x$; it must also be true that $x \in A$ for some class A, an assumption not made for the case when x is C. A complete set of axioms for

[3] See D. E. Knuth, *A System for Technical Text* (Am. Math. Soc., Providence, RI 1979).

set theory avoiding all known paradoxes was worked out by Paul Bernays (1888–1977) and Adolf Fraenkel (1891–1965). It forms part of the basic education of mathematicians today. It is generally accepted because mathematics can be deduced from it. However, it is very far from what Cantor had hoped to create: a clear, concise, and therefore obviously consistent foundation for mathematics. The axioms of set theory are extremely complicated and nonintuitive, and far less obvious than many things deduced from them. Moreover their consistency is not only not obvious, it is even unprovable. In fact one textbook of set theory asserts of these axioms that, "Naturally no inconsistency has been found, and *we have faith* that the axioms are, in fact, consistent"! (Emphasis added.)

So then, does the practice of mathematics require faith? Do mathematicians strain at a gnat (arithmetic) and swallow a camel (set theory)? No, common sense has not been entirely abandoned. Because of weird statements like the Banach–Tarski paradox, set theoreticians have developed substitutes for the axiom of choice that allow the deduction of most of the standard mathematical results for which the axiom of choice is used but avoid the paradoxical statements. The question is: to which philosophy of mathematics do mathematicians subscribe? The fact is that most mathematicians need not take any position on these metamathematical questions in their professional work. A few actually do, working within the framework of intuitionism or a closely related school of constructivism. The majority of research mathematicians, however, have been taught set theory with all its caveats and accept it informally as a basis for communication. Even across schools of thought, although mathematicians who accept different fundamental principles will naturally not agree that each other's results are "valid," they can certainly agree that these results do follow from the premises on which they are based. This pragmatic approach of most mathematicians is strikingly shown by the fact that the intuitionist Brouwer proved theorems in topology whose proofs are not intuitionistically valid.

19.3 Professionalization

Until the nineteenth century mathematics for the most part grew as a wild plant. Although the academies of science of some of the European countries nourished mathematical talent once it was exhibited, there were no mathematical societies dedicated to producing mathematicians and promoting their work. All this changed with the French Revolution and the founding of technical and normal schools to make education systematic. The effects of this change were momentous. The curriculum shifted its emphasis from classical learning to technology, and research and teaching became linked.

19.3.1 Educational Institutions

At the time of the French Revolution the old universities began to be supplemented by a system of specialized institutions of higher learning. The most famous of

these was the École Polytechnique, founded in 1795. A great deal of the content of modern textbooks of physics and mathematics was first worked out and set down in the lectures given at this institution. Admission to the École Polytechnique was a great honor, and only a few hundred of the brightest young scholars in France were accepted each year. This institution and several others founded during the time of the French Revolution, such as the École Normale Supérieure produced a large number of brilliant mathematicians during the nineteenth century. Some of their research was devoted to questions of practical importance, such as cartography and canal building, but basic research into theoretical questions also flourished.

In Germany the unification of teaching and research proceeded from the other direction, as professors at reform-minded universities such as Göttingen (founded 1737) began to undertake research along with their teaching. This model of development was present at the founding of the University of Berlin in 1809.

This educational trend was duplicated elsewhere in the world. During his Italian campaign Napoleon founded the Scuola Normale Superiore in Pisa, which reopened in 1843 after a long hiatus. In Russia a university opened along with the Petersburg Academy of Sciences in 1726, and the University of Moscow was founded a generation later (1755) with the aim of producing qualified professionals. It was not until the nineteenth century, however, that the faculty in Moscow began to engage in research. The University of Stockholm opened in 1878 with aims similar to those of the institutions just named. In Japan an office of translations was opened in the Shogunate Observatory in 1811. It was renamed the Institute for the Investigation of Foreign Books in 1857, and became the home of a department of Occidental mathematics in 1863, taking on two Dutch faculty members in 1865. By 1869 only Western mathematics was being taught, and the teaching was being done by French and British teachers.

19.3.2 Mathematical Societies

Another illustration of the professionalization of mathematics was the founding of professional societies to supplement the activities of the mathematical sections in academies of sciences. The oldest of these is the Moscow Mathematical Society (founded in 1864). The London Mathematical Society was founded in 1866, the Japanese Mathematical Society in 1877. The American Mathematical Society (originally the New York Mathematical Society) was founded in 1888 and the Canadian Mathematical Society in 1945.

19.3.3 Journals

These educational institutions and professional societies also published their own research journals, such as the *Journal de l'École Polytechnique* and the *Journal de l'École Normale Supérieure*, in which some members of the Paris Academy of Sciences chose to publish to avoid the delays associated with the official journal of the Academy. These journals contained some of the most profound research of the nineteenth century. Other nations soon emulated the French. The German *Journal*

für die reine und angewandte Mathematik was founded by August Leopold Crelle (1780–1855) in 1826 (informally it is still called *Crelle's Journal*); the Italian *Annali di scienze matematiche e fisiche* appeared in 1850; the Moscow Mathematical Society began publishing the *Matematicheskii Sbornik* (*Mathematical Collection*) in 1866; the Swedish *Acta Mathematica* was founded in 1881. By the end of the nineteenth century there were mathematical research journals in every European country, in North America, and in Japan. The first American research journal, *The American Journal of Mathematics* was founded at Johns Hopkins University in 1881 with the British mathematician J. J. Sylvester as its principal editor, assisted by the American William Edward Story (1859–1936). The first issue of *The Canadian Journal of Mathematics* was dated 1949.

19.4 Democratization

In the seventeenth century the practice of mathematics in the West was confined to a few centers of high culture in Britain, France, Italy, Switzerland, and Germany, with only an isolated scholar making important contributions outside this area. Japan and China had excellent mathematicians and were beginning to take an interest in the mathematics being produced in the West. By the early nineteenth century this base had enlarged to include many countries on the periphery of Europe, such as Russia, Norway, and Hungary, as well as Canada and the United States. The achievements of the European mathematicians have been discussed in previous chapters, but this is an appropriate point to survey some of the development of mathematics in North America.

19.4.1 North America

Until the late nineteenth century most of the mathematics done in North America was purely practical, and to find examples of its practitioners we shall have to leave mathematics proper and delve into related areas. Commerce required a certain amount of mathematics and astronomy to meet the needs of navigation, and all the early American universities taught dialing (theory of the sundial), astronomy, and navigation. These subjects were standard, long-known mathematics, a great contrast to the rapid pace of innovation in Europe at this period. Nevertheless, to write the textbooks of navigation and calculate the tides a year in advance required some ability. It is remarkable that this knowledge was acquired by two Americans who were not given even the limited formal education that could be obtained at an American university. These two—Benjamin Banneker and Nathaniel Bowditch—are in some respects twins, and in other respects opposites. Both came from families of modest means, both had to struggle to make a living, and both eventually published works involving a knowledge of astronomy. Yet there was a difference between them, which made Banneker's struggle incomparably more difficult than Bowditch's; for Bowditch's ancestors came from Europe and Banneker's from Africa.

Benjamin Banneker (1731–1806)

In the fall of 1791 the Baltimore publishing house of William Goddard and James Angell published a book bearing the title *Banneker's Almanac and Ephemeris for the Year of our Lord 1792...* (see Fig. 19.1). The author was at the time about 60 years old, the only child of parents of African descent[4] who had left him a small parcel of land as an inheritance. For most of his life Banneker lived near Baltimore, struggling as a poor farmer with a rudimentary formal education. Nevertheless, he acquired a reputation for cleverness due to his skill in arithmetic. In middle age he made the acquaintance of the Ellicotts, a prominent local family, who lent him a few books on astronomy. From these meager materials Banneker was able to construct an almanac for the year 1791. Encouraged by this success, he prepared a similar almanac for 1792. In that year the Ellicotts put him in contact with James McHenry (who had been Surgeon General of the American Army during the Revolutionary War). McHenry wrote to the editors:

> ...he began and finished [this almanac] without the least information or assistance from any person, or other books than those I have mentioned; so that whatever merit is attached to his present performance is exclusively and peculiarly his own.

Banneker's *Almanac* was published and sold all over America in the decade from 1792 until 1802. The contents of the *Almanac* are comparable with those of other almanacs that have been published in America: on alternate pages one finds calendars for each week or month, giving the phases of the moon, the locations of the planets and bright stars visible during the period in question, and the times of sunrise, high and low tides, and conjunctions and oppositions of planets. Interspersed among these pages one finds poetry, inspirational essays, lists of roads in America, schedules of court sessions, and snatches of medical and financial advice. Thus compiling an almanac required not only mathematical talent, but also a good literary sense and knowledge of what interested the public. In Banneker's *Almanac* for 1793, one finds, for example, a proposal for the establishment of a cabinet post, Secretary of Peace.[5] There is also a long quotation from Thomas Jefferson on the evils of slavery. This first edition of the *Almanac* appeared during the George Washington's term as President, when Jefferson was serving as Secretary of State. At this time, a controversy over the location of the capital of the United States had been settled by the decision to house it in a district separate from all the states. The first surveyor for the district was Andrew Ellicott, and Banneker was one of his assistants.

Was there some intentional irony in Banneker's quotation of Jefferson? In his *Notes on Virginia* Jefferson had expressed doubts as to the intellectual equality of the races, although he kept in mind the possibility that the differences he observed were due to social circumstances. As the first *Almanac* was being published,

---

[4]Banneker's grandmother was an Englishwoman who married one of her slaves. Their daughter Mary, Banneker's mother, also married a slave, who had the foresight to purchase a farm jointly in his own name and in the name of his son Benjamin.

[5]At the time what is now called the Department of Defense was known as the Department of War.

Banneker sent a copy to Jefferson, along with a letter containing an ardent plea that Jefferson would take cognizance of the state of African Americans. He very astutely quoted Jefferson's own words from the Declaration of Independence: *We hold these truths to be self-evident...* .

Jefferson had a plan for the gradual abolition of slavery, which eventually failed because of the importance of the cotton trade to the economy. On the day he received the *Almanac* he wrote to Banneker, thanking him for providing "such proofs as you exhibit" that the black race was endowed with intelligence equal to those of other races and assuring him that he wished more than anyone else for the amelioration of their condition. He wrote to the French mathematician/philosopher the Marquis de Condorcet (1743–1794) on the very same day, sending him the copy of the *Almanac* and pointing out the moral to be gleaned from it.

Recognition came late to Banneker. The money he earned from his *Almanac* gave him some leisure in his old age, and his name was praised by Pitt in Parliament and by Condorcet before the French Academy of Sciences. Yet he was never elected to any scientific societies in America, despite having achieved considerable fame and having corresponded with well-known American scientists.

Nathaniel Bowditch (1773–1838)

Benjamin Banneker was about 40 years old and still living in obscurity near Baltimore when Nathaniel Bowditch was born in Salem, Massachusetts. His ancestors had been shipbuilders, but had accumulated no substantial amount of money by this trade. His father abandoned it and became a cooper, a trade that barely provided for his family of seven children. Nathaniel received only a rudimentary public education before being apprenticed to a ship chandler at the age of 10. Twelve years later, when Banneker's *Almanac* had been published for only a year or two, he signed on board a ship and, like Banneker, used his few intervals of leisure to study mathematics and astronomy. Bowditch was a natural teacher who enthusiastically shared his knowledge of navigation with his shipmates. With his aptitude for mathematics, he managed to get through Newton's *Principia*, learning a considerable amount of Latin on the way. Later he taught himself French, which was displacing Latin as the language of science as a result of the pre-eminence of French mathematicians and scientists.

Bowditch first gained a scholarly reputation by pointing out errors in the standard navigational tables. His abilities immediately attracted interest, and his *Practical Navigator*, first published in 1800, gained him wide recognition while he was still in his 20s. Bowditch became a member of the American Academy of Arts and Letters, and in 1818 was elected a member of the Royal Society.

With recognition came leisure time to devote to purely scholarly pursuits, a luxury denied to Banneker in his most vigorous years. For the last quarter-century of his life Bowditch labored on his monumental translation and commentary of the *Mécanique céleste* by Laplace. This work amounts really to a complete rewriting of Laplace's treatise, which shows the effects of a pronounced stinginess with ink and paper. Bowditch filled in all the missing details of arguments that Laplace had merely waved his hand at, not having the patience to write down arguments

that had sometimes taken him weeks to discover. These pursuits brought Bowditch international fame, and he died covered with honors. The *American Journal of Science* published his obituary with a portrait of him in a classical Roman tunic which it is unlikely he ever actually wore.

North America Joins the European Intellectual World

The end of the American Civil War in 1865 was followed closely by the founding of the Canadian Federation in 1867. The Federation was the result of the British North America Act, which reserved some constitutional controls for Britain. (Full independence came in 1982.) From that time on both countries experienced a remarkable cultural flowering, which included advances in mathematics. Americans and Canadians began to go to Europe to learn advanced mathematics. This early generation of European-trained mathematicians generally found no incentive to continue research upon returning home. However, they at least made the curriculum more sophisticated and prepared the way for the next generation.

The United States. In Europe there were more Ph.D. mathematicians being produced than the universities could absorb. Most of these entered other professions, but a few emigrated across the Atlantic. A scholarly coup was scored by Johns Hopkins University, which opened in 1876 with a first-rate mathematician on board, namely James Joseph Sylvester. Despite being 63 years old, Sylvester was still a creative algebraist, whose presence in America attracted international attention. One of his first acts was to found the first American mathematical research journal, the *American Journal of Mathematics*. The founding of this journal had been suggested by William Edward Story, one of the many Americans who went abroad to get the Ph.D. degree, but atypically continued to do mathematical research after returning to America. Before Johns Hopkins was founded, there had been a few graduate programs in mathematics in places such as Harvard and the University of Michigan, but now such programs began to multiply. Bryn Mawr College opened in the mid-1880s with a graduate program in mathematics. The founding of Clark University in Worcester, Massachusetts and the University of Chicago in the late 1880s and early 1890s promised that America would soon begin to make respectable contributions to mathematical research, and this promise was largely fulfilled by the 1920s with a large number of talented Americans achieving recognition in Europe. America's present position as the world leader in mathematics, however, was largely the result of the turbulence of the 1930s and 1940s, which drove many of the best European intellectuals to seek refuge far from the dangers that threatened them in their homelands.

Canada. Like American schools of the same period, English-language Canadian institutions of higher learning tended to rely on British textbooks such as those of Charles Hutton (1737–1823, a professor at the Miltary School in Woolwich). In French Canada there was a somewhat longer tradition of educational institutions, and a French calculus text written by Abbé Jean Langevin was published in 1847. For Canadians, as for Americans, the importance of research as an activity of the mathematics professor arose only after the founding of Johns Hopkins University

Benjamin Bannaker's

PENNSYLVANIA, DELAWARE, MARYLAND, AND VIRGINIA

ALMANAC,

FOR THE

YEAR of our LORD 1795;

Being the Third after Leap-Year.

PHILADELPHIA:

Printed for WILLIAM GIBBONS, Cherry Street

Figure 19.1: Title page of Benjamin Banneker's *Almanac*. The Bettmann Archive.

in 1876. In fact the early volumes of the *American Journal of Mathematics* contain several articles by two Canadians, J. G. Glashan (1844–1932), superintendent of schools in Ottawa, and G. Paxton Young (1818–1889), a professor of philosophy at the University of Toronto. Because of the close proximity of the two countries in geography and similar speech patterns, the many Canadian mathematicians who come to work in American universities and corporations are routinely mistaken for Americans. Canada has faced the same problems as the United States in establishing a basis for scientific research; but in addition, the wealth of the United States has acted to draw off a number of talented Canadians and to discourage the duplication of their activity in Canada.[6] An example of this phenomenon is Simon Newcomb (1835–1909), a native of Nova Scotia who taught school in a number of places in the United States before procuring a job at the Nautical Almanac Office in Cambridge, Massachusetts, where he attended Harvard. He eventually became Director of the Naval Observatory in Washington, and after 1884 Professor of Mathematics at Johns Hopkins.

Two Canadian mathematicians deserve special mention. The geometer H.S.M. Coxeter, a native of Britain, emigrated to Canada in 1936 and has played a leading role in Canadian research in symmetry groups and symmetric geometric objects of all kinds. John Charles Fields (1863–1932), a native of Hamilton, Ontario, received the Ph.D. from Johns Hopkins in 1887 and studied in Europe during the 1890s. In 1902 he became a professor at the University of Toronto. He wrote one book (on algebraic functions). Like many other mathematicians on the intellectual periphery of Europe, much of his activity was devoted to encouraging research in his native country. In the last few years of his life he established the Fields medals, the highest international recognition for mathematicians, which are awarded at the quadrennial International Congress of Mathematicians. Beginning in 1936, when two awards were given, then resuming in 1950, the Fields Medals have by tradition been awarded to researchers early in their careers. To date 36 mathematicians have been so honored, among them natives of China, Japan, New Zealand, the former Soviet Union, many European countries, and the United States.

19.4.2 Asia and Africa, and American Minorities

In these last few chapters we have concentrated on Europe and America. Yet in the twentieth century no one is surprised to find names like Hua, Yoshida, and Harish-Chandra, among the world mathematical leaders. The nations of Asia have assumed a prominent role in mathematical research. Within America also there is an increasing diversity of scholars. According to the *Notices of the American Mathematical Society* of December 1994, 30 of the 469 Americans who received the Ph.D. degree in mathematics during the 1993–94 academic year were Asian-Americans; this figure amounts to more than 6%, from a group that constitutes only

[6] An exception to this general rule is the history of mathematics, which enjoys relatively more institutional support in Canada than in the United States. The study of the history of mathematics in Canada received an impetus from Kenneth May (1915–1977), who left the United States during the 1950s.

3% of the population. For non-U. S. citizens receiving the Ph.D. degree in American universities the proportion of Asians was even more impressive, amounting to 329 of 590, or more than 50%.

The peoples of Africa, who have no long history of scientific activity and are struggling to industrialize and create nations at the same time, have achieved less. Yet there has been progress in Africa also. The immediate problem facing African mathematical educators after their nations achieved independence was to train a generation of mathematical teachers and create suitable textbooks. In former French colonies, for example, most of the teachers had been European French, and mathematical textbooks used problems that took for granted French institutions and geography. In replacing French texts by those in local languages, even wider cultural gulfs had to be bridged. For example, counting in the indigenous languages of Senegal was based on 5 rather than 10.

The first mathematical journal in Africa, called *Afrika Mathematica*, was founded in 1978. In the 1980s the *Nigerian Mathematical Journal* was founded, and in September 1988 the first international symposium of the African Mathematical Union (founded in 1975) was held in Arush, Tanzania.

African-Americans also are making progress in overcoming the effects of discrimination. The first African-American to obtain a doctorate in mathematics was Elbert Cox (1895–1969), who became a professor at Howard University after obtaining the doctorate. The first African-American women to receive the doctorate in mathematics (both in 1949) were Marjorie Lee Brown (1914–1979) and Evelyn Boyd Granville (b. 1924). Brown was a differential topologist who received her degree at the University of Michigan and taught at North Carolina Central University. Granville received the Ph.D. from Yale University and worked in the space program during the 1960s. She later taught at California State University in Los Angeles.

Although these early examples are inspiring, the number of African-Americans choosing to enter mathematics and science is still comparatively small. In fact the author of an article entitled "Black Women Ph.D.'s in Mathematics" in the 1980s was able to interview *all* of the people described in the title who were still alive. A career in research, after all, requires a long apprenticeship, during which financial support must be provided either by family, by extra work, or by grants and loans. For people who do not come from wealthy families, other careers, promising earlier financial rewards, are likely to seem more attractive. Undoubtedly if the average income of African-Americans were higher, more of them would choose scientific careers.

19.4.3 Women Mathematicians

The democratization of mathematics has taken a very long time to reach women. Maria Gaetana Agnesi, who was mentioned in a previous chapter, attained an appointment as professor at the University of Bologna; but for reasons that cannot be known with certainty, her mathematical research declined and she began to engage more and more in charitable work. She left the University after only

two years. The situation in France was similar. Despite his efforts at popular education, Napoleon was a believer in male dominance (his expressed opinions on the rights of women were retrograde in the extreme). As a result Sophie Germain (1776–1831) was forced to study and practice mathematics as an outsider. Her talent eventually won her a prize from the Paris Academy of Sciences and high praise from the Gauss, the greatest mathematician of the nineteenth century. The "public" routes to the world of scholarship—the educational institutions—were not available to women in northern Europe, so that only a few leisured women such as the Marquise du Châtelet, Augusta Ada Lovelace, mentioned in a previous chapter, and Mary Fairfax Somerville (1780–1872, the author of textbooks of astronomy) were able to pursue mathematical interests. In Britain, for example, women were not allowed to receive degrees at Cambridge until 1948. Yet one woman, Charlotte Angas Scott (1858–1931), was allowed to take the Tripos examination for the degree at Cambridge in 1880. She ranked eighth in the entire University and was wildly cheered by her classmates, even though the rules forbade reading her name at the graduation ceremony. Five years later she received the doctorate from the University of London. Unable to find a position in Britain, she came to America and took up a post at Bryn Mawr College.

The greatest of the nineteenth-century women, Sof'ya Vasil'evna Kovalevskaya, advanced very far in this male world through her own powerful energy and the support of reform-minded mathematicians. Despite never having been allowed to enroll in a university mathematics course, Kovalevskaya obtained the Ph.D. degree from Göttingen University in 1874, at the age of 24. (She had done her work privately with Weierstrass in Berlin, but Weierstrass knew that it would be pointless to ask the conservative University of Berlin to grant her a degree.) Kovalevskaya published two papers of fundamental importance which are still remembered a century later, attained a regular faculty position at the University of Stockholm, and won a prize competition at the Paris Academy of Sciences. These achievements would be remarkable under any circumstances. When set against the brevity of her life (she died just after her 41st birthday) and the discouragement she met at every stage of her career from her family and from every institution of society, they are awe-inspiring.

Despite Kovalevskaya's pioneering achievements, 50 years later the mathematical genius Emmy Noether (1882–1935) was unable to obtain at Göttingen the place her talents deserved, even with the enthusiastic support of Felix Klein and David Hilbert, two of the greatest mathematicians in the world. Noether spent the first 40 years of her life obtaining the position that she would easily have reached in her 20s if she had been a man of comparable talent. Like Kovalevskaya, her life at the top was tragically brief. She had barely reached the age of 50 when the Nazis took power, and because she was Jewish, she had to emigrate. Unlike the toprank male mathematicians who came to America, she was not offered a position at the universities that were the major centers of mathematical research (many of these universities were open only to men until the 1960s). She found, however, a good position at Bryn Mawr College, which she occupied for the one year of life remaining to her.

One of the most versatile mathematicians of the century was Olga Taussky-Todd

(1906–1995), who worked first in algebraic number theory (class-field theory) and later in boundary-value problems for hyperbolic differential equations, numerical analysis, and the stability theory of matrices. A close contemporary of Emmy Noether, she also worked at Göttingen in the early 1930s, as well as in Vienna and came to Bryn Mawr at the same time as Noether (1934). Then, after a few years spent in Britain, she returned to America to work at the National Bureau of Standards. She later became professor at the California Institute of Technology, retiring in 1977.

In the twentieth century there have been dozens of outstanding women mathematical researchers. A number of them came from the Soviet Union—women such as Ol'ga Aleksandrovna Ladyzhenskaya (b. 1922) and Ol'ga Arsenevna Oleinik (b. 1925), both of whom have made first-rate contributions to the theory of differential equations; Luzin numbered several talented women among his students. Two of them were Nina Karlovna Bari (1901–1961), who was the first to discover that the uniqueness properties of certain sets relative to trigonometric series can be formulated in terms of their number-theoretic properties, and Lyudmila Vsevolodovna Keldysh (1904–1976), who gave profound analyses of the hierarchy of Borel sets.

In America many of the best graduate schools were all-male until the 1960s. Despite the lack of encouragement, and even in the face of outright discouragement, some American women did manage to obtain the doctorate and find teaching positions in universities. One of the most outstanding of these was Julia Bowman Robinson (1919–1985), the first woman mathematician elected to the National Academy of Sciences. After great hardships in childhood and early adulthood, she began to study mathematics in the late 1940s and completed a doctoral dissertation with Alfred Tarski at the University of California in 1948. She contributed crucial steps to the solution of Hilbert's tenth problem: *Find, if possible, an algorithm for determining whether a Diophantine equation is solvable.*

These few examples by no means give a fair picture of women's contributions to twentieth-century mathematics. The vast extent and complexity of twentieth-century mathematics make it impossible to summarize, and the contributions of women have mostly fallen within this century. Fortunately there are now specialized studies devoted to the topic of women in mathematics.

19.5 Mathematics and Society

Until the Renaissance mathematicians and scientists in general had been self-motivated and self-supporting individuals. During the Renaissance and Enlightenment, for the sake of prestige and a monopoly of scientific discoveries, monarchs supported scholars through academies of science and universities. In Europe and Canada such support survived the transition to democracy, but in America most research was carried on at private or state-funded universities. The Federal Government did not begin supporting universities until the midtwentieth century (indeed there were no constitutional grounds for the government to do so); and in contrast to Europe, members of the American Academy of Sciences do not receive a salary for their work (they retain their occupations in whatever industry, business,

or university they are employed). In the present era, however, a large portion of mathematical research is supported by the government, either through direct grants to researchers from agencies such as the Department of Education or the National Science Foundation, or indirectly, by paying the tuition of graduate students, which is then used to pay salaries to professors, part of whose obligations is research. The wisdom and the effects of this system are matters that a responsible citizen of a democracy must attempt to judge in order to help set policies for the extent of such support.

To fill out the history of twentieth-century mathematics, we shall examine some aspects of the relation of mathematics and government in the Soviet Union, Nazi Germany, and America, to see what dangers and opportunities there are in cooperation between scholars and government.

19.5.1 The Soviet Union

After the October Revolution of 1917 the Communist Party enacted a series of measures to shore up its support among the people who had previously been at the bottom of the socioeconomic ladder. In particular, it opened up the universities to all young people except certain proscribed groups (such as the former nobility and the tsarist police).

The Attack on the Moscow Mathematical Society

The Academy of Sciences retained its independence somewhat longer than the universities, partly because the education of a new generation of intellectuals had to begin with the universities. These new intellectuals would move into the Academy only later. When the attack came, however, mathematicians suffered as much as researchers in other areas of science. The first indirect attacks on the Academy came through the Moscow Mathematical Society, and the first to suffer was the President of the Society D. F. Egorov (1869–1931). He was forced to resign one of his committee responsibilities at the University of Moscow and in 1929 was dismissed as director of the Institute for Mechanics and Mathematics at the University. As political power shifted to those loyal to the Party, Egorov came under attack from Communist students. In December 1929 he was formally censured at a meeting of graduate students, who pledged themselves to take up antireligious work. (Egorov was a prominent member of the Russian Orthodox Church.) The following year he came under attack at a University council meeting from the militant Czech Marxist Arnost Kolman (1892–1979). Kolman accused Egorov of *vreditel'stvo*, a word that literally means *damaging* and denotes activity somewhere between obstruction and outright sabotage. It is usually translated as *wrecking*. Egorov, not at all intimidated, replied that "true wrecking is nothing other than the imposition of a standard worldview on scientists," a very accurate jab at Marxist orthodoxy. Kolman was incensed that the moderator of the meeting cut off the argument at this point. In the end Egorov was arrested and sentenced to a labor camp. While being transported to serve his sentence, he deliberately starved himself to death.

The official Soviet view of this affair was summed up in the biography of Egorov that appeared in the first *Large Soviet Encyclopedia*. The editor of the *Encyclopedia* was another mathematician, Otto Yulevich Shmidt (1891–1956), an algebraist of some distinction. As an important functionary in the Soviet establishment, he found it prudent to get as far away from Stalin as possible, and so he headed several Arctic expeditions in the late 1930s. In the encyclopedia Egorov is described as follows:

> . . . author of works on analysis, number theory and other areas, not containing, however, any significant scholarly discoveries. Prominent representative of the reactionary (idealist) Moscow mathematical school. Actively struggled against the measures of the Soviet regime for reorganization of higher education and scientific institutes. After the exposure of a "Egorov conspiracy" he was removed from his post as director of the Institute of Mathematics and Mechanics and in 1930 excluded from membership in the Moscow Mathematical Society.

Unpleasant as these words are, one cannot accuse the author of lying. The judgment of Egorov's works is harsh, but not unreasonable—he was *not* a world leader in mathematics. The existence of a Egorov conspiracy is a strained interpretation of his opposition to the regime, but opposed he certainly was. Only by silence does this article lie, saying nothing about the arrest and death of Egorov.

This article is worth comparing with the post-Stalinist Brezhnev-era article on Egorov, which appeared in the 1972 edition of the *Encyclopedia*, especially since a great many intellectuals in the West were inclined to think the best of the post-Stalinist leaders. The new, rehabilitated Egorov, appeared as follows:

> Soviet mathematician, corresponding member of the Academy of Sciences of the USSR (1924), honorary member of the Academy of Sciences of the USSR (1929). Graduated from the University of Moscow (1891), professor of the University after 1903, *President of the Moscow Mathematical Society (1922–1931)* [emphasis added].

The article then goes on to depict in glowing terms the worldwide importance of Egorov's contributions to differential geometry, integral equations, calculus of variations, and functions of a real variable. Thus a victim of the regime is pictured as one of the shining stars of Soviet science and, by implication, its loyal servant. In the last sentence a deliberate lie is told to cover up the brutal treatment he was accorded.

The Luzin Affair

Egorov's student Luzin apparently did not take Kolman seriously, and this judgment cost him dearly. For Luzin was compromised. As a student he had made the friendship of Pavel Aleksandrovich Florenskii (1882–1937), a brilliant mathematician, physicist, and philosopher. Florenskii wrote his dissertation on "The Idea of Discontinuity as an Element of a Worldview." Despite the philosophical-sounding

title, the first part of the dissertation concerns singularities of algebraic curves. He never published it, however. Before he could begin his scientific career, religion won a complete victory in him, and he became a priest, though he continued to do scientific work. Kolman launched a vicious attack on him in 1933, and in that same year Florenskii was arrested and sentenced to 10 years at hard labor. Florenskii was a brave man who had been arrested by the Tsar's police in 1906 for protesting the execution of the leaders of the 1905 revolt. Stalin's interrogators, however, used methods of persuasion that civilized people would find difficult to believe (some of them are described in Solzhenitsyn's *Gulag Archipelago*). As their victims were chosen for capricious and arbitrary reasons, there was seldom any case against them that would withstand rational scrutiny. The only way to proceed with an appearance of legality (which was important to maintain the image of the regime for external propaganda) was to procure a confession. Under torture Florenskii confessed to being the leader of a Fascist organization called the Party of the Rebirth of Russia, which allegedly aimed at securing a German occupation of Russia. The plot that Florenskii described in his confession was to conclude a union between the Orthodox Church and the Roman Catholic Church through a certain German Jesuit representing the Pope. This preposterous fiction required that certain professors be implicated, and Florenskii was induced to name Luzin as one of them.

Had Luzin been a typical obscure Soviet citizen, this betrayal would have sealed his fate. Luzin, however, was an acknowledged world leader in mathematics and had many intellectual friends in France and Germany. The authorities did not wish to arrest him without any evidence of wrongdoing, and the case against him was too far-fetched to be publicized. They therefore took a different approach. For some time Luzin was left alone. Then in 1936 he was "set up." He was invited to a Moscow high school to observe the mathematics instruction and asked to write his observations in *Izvestiya* on June 27. All his life Luzin was a timid and polite man, incapable of any harsh criticism of students. He wrote a complimentary piece, called "A Pleasant Disillusionment," explaining that he had expected the usual incompetence in mathematics that is rampant in high schools and had been pleasantly surprised to find the level of achievement much higher than he had believed. In so doing he fell into a carefully prepared trap.

On July 2 the principal of the school wrote a "Response to Academician Luzin" in *Pravda*, in which he stated that Luzin had apparently forgotten that he was in a Soviet school and that he was expected to provide constructive criticism. Luzin's polite words were made to seem like a sinister attempt to sabotage the school in its efforts to improve. The whole course of events had obviously been planned in advance. The following day *Pravda* ran an article with the title "Enemies masquerading as Soviets," which accused Luzin of a long history of abuses, especially being insufficiently critical of the works of other mathematicians, publishing his best works abroad (his treatise on analytic sets had been published in French in Paris and was financed by the Rockefeller Foundation), idealizing the West, and plagiarizing the results of his students. Although this article was unsigned and Kolman later denied having anything to do with it, he is the most likely author.

Pravda continued to print articles denouncing Luzin for 2 weeks. In insti-

tutes all over Moscow emergency meetings were held to label him an enemy of the state. His first accuser at the University of Moscow was Sof'ya Aleksandrovna Yanovskaya (1896–1966), a logician and a dedicated Marxist. Her denunciations were echoed by Luzin's students Aleksandrov (1896–1982) and Kolmogorov. Within a week the Academy of Sciences had no choice but to investigate the matter. On July 7 a special commission was set up by the Presidium of the Academy. Luzin's most vicious attackers were his student Aleksandrov, the algebraist Shmidt, and the analyst Sergei L'vovich Sobolev (1908–1989). Despite the danger to themselves, two mathematicians—Sergei Natanovich Bernshtein (1880–1968) and Aleksei Nikolaevich Krylov (1879–1955)—defended Luzin, as did Academician Pëtr Kapitsa (1893–1984), a man whose courageous resistance to the Soviet regime was to be demonstrated on many occasions. Luzin also received support from abroad, especially from prominent French mathematicians, and this support may have influenced the outcome of the affair.

It seemed that Luzin was doomed, yet by some mysterious *deus ex machina* never explained,[7] the campaign against him abruptly stopped on July 13. The Presidium decided to reprimand him, and the case was closed. Luzin broke with all of his former students except two, Lyudmila Vsevolodovna Keldysh and Nina Karlovna Bari. He never forgave Aleksandrov and managed to block his election to full membership in the Academy of Sciences. (Aleksandrov was elected only after Luzin died in 1950.)

Why Did Soviet Mathematics Flourish?

It is a curious fact that the practical, applied focus of the Soviet regime and its tight control of the universities did not lead to the extinction of pure mathematics in the Soviet Union. In fact quite the opposite was the case. Despite clear evidence of discrimination against Jewish mathematicians during the Brezhnev era, the Soviet Union produced a large portion of the top mathematicians in the world for more than 50 years. In the halls of the Main Building at the University of Moscow there hangs a picture of the "Luzin tree," a sort of genealogy of the students and "grandstudents" of Luzin. The students of Luzin and their students played a prominent role in many important areas of mathematics. Of course there were many other outstanding mathematicians in Moscow not directly connected with Luzin. When these people are added to the Leningrad mathematicians and the mathematicians in other Soviet cities, the total amount of mathematical talent is prodigious.

Part of the explanation for this flourishing of "useless" pure mathematics may lie in the nature of Marxist ideology, with its tendency to see the dialectic at work in every aspect of the universe, including pure thought.[8] Undoubtedly also the regime expected to make some gains in engineering and productivity through

---

[7] In a recent article, Charles Ford and Sergei Demidov, hypothesize that the charge of spying for the Germans was inconsistent with Stalin's ambition to form an alliance with Hitler.

[8] In an earlier attack on Luzin at a 1931 conference in London, for example, Kolman claimed to have found a Marxian contradiction in Luzin's space of irrational numbers, which combined the discrete and the continuous.

mathematical advances, but in any case a brilliant constellation of scholars is good for both internal and external propaganda.

19.5.2 Mathematics in Nazi Germany

Of the two major totalitarian societies of the twentieth century, both based on dogmatically held theories, the Communist regime was the more rational, and hence sustained itself the longer in the face of internal and external opposition and the increasingly obvious incompatibility of its basic principles with reality. Its most irrational aspect was the arbitrary terror used to impose wasteful and ineffective methods of production on the populace. What it held as its goals—prosperity and freedom for all—would not have been rejected by anyone; the only question is whether those goals are achievable and if so, by what methods. The Nazi regime, in contrast, was as near to insanity as any civilized society is likely to get. It was based on demonstrably absurd theories of race, for which the foundation of belief was envy and hatred. For that reason it had almost no support outside its homeland, and failed in a few years by provoking a war. The two regimes certainly resembled each other in their organization of brutality and oppression on a mass scale, and only in this macabre aspect was the Nazi regime more "rational." It sought prosperity (not freedom) only for "Aryans" and attempted the destruction of non-Aryans; while the Soviet regime, claiming to seek prosperity for the proletariat and peasantry, pursued this aim by sending millions of proletarians and peasants to perish in labor camps.

The German Universities

The German universities rose to prominence during the nineteenth century, catching up in most respects with those in France. This blossoming of German culture coincided with the unification of the German confederacy under Bismarck. The German universities differed from those in Britain in being oriented toward research while those in Britain aimed at educating leaders for public service. (These are, of course, only general characteristics; in fact, both activities were present in both countries.) American universities in the early years followed the British model, but increasingly after the Civil War they patterned themselves on the German universities. German scholarship was widely admired, and by the end of the nineteenth century there was hardly any area of learning in which German scholars had not produced a definitive treatise. This period of German cultural upsurge was accompanied by several trends that led in different directions, and it was not clear at first what their combined effect would be.

1. *Enlargement of the professoriate.* Just as Louis XIV had effectively made it impossible for Protestants to live in France in the seventeenth century, Catholics suffered civil disabilities in Britain until 1830 and somewhat longer in Germany. The position of Jews had been precarious in all European countries for centuries. They had been expelled from England in the thirteenth

century and from Spain at the end of the fifteenth century. In the progressive nineteenth century there seemed reason to believe that these ancient prejudices were coming to an end. Particularly in the area of scholarship an ideal of universal humanity was widely felt. Jacobi was the first prominent Jewish professor in Germany and Weierstrass the first Catholic. In terms of its recognition of talent, one would believe that Germany in the nineteenth century was at least as tolerant of Jews as any other country in Europe. There was some awkwardness on a social level, but rarely does one read any outright antisemitism in the letters of the mathematicians of the time. In a letter to Gösta Mittag-Leffler (1846–1927) written June 1, 1884 Sof'ya Kovalevskaya mentions that a certain young mathematician named Meyer Hamburger (1838–1909) is a Jew. She notes that he has very little contact with other mathematicians, "mostly because he dresses so badly." This comment is revealing, showing that the social code was more concerned with manners than with ethnicity. Nevertheless, in his letters to Kovalevskaya Weierstrass made generalizations about Jewish mathematicians (his enemy Kronecker in particular) that border on prejudice.

2. *Nationalism.* In order to unify the many German principalities into a single state Bismarck had to stimulate pride in identity as a German rather than a Bavarian or a Hessian or a Prussian. This kind of national pride, reinforced by romantic philosophies, acted as a barrier to the encouragement of more universal human values. German patriotism found a seductive artistic expression in some of the operas of Richard Wagner, who was fiercely antisemitic. During World War I some patriotic German professors, among them Felix Klein, attempted to gain support for the German position among the world's intellectuals. Most of the latter, however, belonged to one of the belligerent countries, and so the effort was ineffective.

3. *Technical focus.* The German emphasis on research and the outstanding achievements of German scholars led to an overemphasis on merely technical competence and an exaggerated confidence in the applicability of the methods of physical science. Attempts to analyze human society by regarding people as members of groups are useful only to the extent that they make it possible to understand individuals better. When the groups, which are merely ideal creations, come to be treated as the basic elements of society, the results are sometimes unpleasant. In fact the social sciences were relatively neglected in the German universities. Moreover Germany was not advancing toward democracy, as Britain was. Politics was not of interest to most German students since they were not planning to participate in a political process. The attempt at a democratic revolution in Germany in 1848 had failed because of lack of organization on the part of the democrats.

Victims of the Nazis

Hitler had very little interest in science, though he was fascinated by technology. He wanted education to be aimed at producing disciplined, self-sacrificing servants

of the State. For this end the most important subject was history, naturally a history dominated by the Nazi ideology and made up of tendentiously selected facts. Hitler's program had been foreshadowed in *Mein Kampf*, and some intellectuals were already planning to flee or resist when he came to power in 1933. Albert Einstein, the most prominent Jewish scientist in Germany, came under attack in the Nazi press immediately. Being in America when Hitler came to power, Einstein resigned from the Prussian Academy of Sciences and refused to return to Germany. He went instead to Belgium and waited to see what would happen in Germany. He soon found that his property in Germany had been confiscated and that there was a price on his head.

The Nazis were not long in starting to rid Germany of "foreign" influences. The Law for the Restoration of the Career Civil Service was passed in April 1933, just 2 months after Hitler came to power. Since instructors at institutions of higher learning were considered civil servants, this law affected the universities. The basic purpose of the law was to rid the civil service of Communists and "non-Aryans." Exceptions were allowed, insisted on by President Hindenburg, for non-Aryans whose appointments began before the war, or who had served in the war, or who had lost a father or son during the war. The definition of the term "non-Aryan" was at first rather vague. Since the law was primarily aimed at Jews, a person was said to be Jewish who had a parent or grandparent who practiced the Jewish religion.

Since very few German professors were committed to political action, the main effect of this law was to prevent Jewish professors from teaching. In carrying out this policy the Nazis had strong support from students, who were organized by the Nazis into the German Students' Association. A racial purification campaign began in April 1933 and reached its climax in May with public burnings of books by non-Aryans. There were some protests at first. A few liberal newspapers expressed hope that Jews would be allowed to continue their work in Germany. Their voices were feeble, however. By the end of April the first dismissals of Jewish professors had begun with the termination of Richard Courant (1888–1972) and Emmy Noether at Göttingen. Courant was a student of Hilbert and succeeded to Klein's position when the latter retired in 1921. Courant and several other mathematicians, including Otto Neugebauer and Hermann Weyl, met with the physicists Max Born (1882–1970) and James Franck (1882–1964) at Franck's home and considered mass resignations to protest the new laws. Neugebauer and Weyl were not Jewish, though Weyl's wife was. Franck took public action and deliberately sacrificed his career in Germany. Courant was officially safe from the ethnic provisions of the law, having been wounded in the war, but he was under suspicion as a former Social Democrat. By the first week of May he was placed on leave with pay. Many outstanding scholars rallied to his support, including the physicists Max Planck (1858–1947), Werner Heisenberg (1901–1976), Erwin Schrödinger (1887–1961), and Arnold Sommerfeld (1868–1951). The mathematicians Kurt Friedrichs (1901–1982) and Hellmuth Kneser (1898–1973) appealed on his behalf directly to the central government, but to no avail. Courant later remarked that he knew by this time he would have to leave Germany, as his youngest son could not understand why he was not allowed to join the Hitler Youth. In August he accepted a

position in Cambridge. In a curious twist of fate, in October he was notified that the Civil Service Law did not apply to him, and his leave was canceled. He left Germany in November.

The mere recital of the names of outstanding German mathematicians who were victims of the Nazis provokes a sense of bewilderment at the insanity of the Nazi regime. How could any country deliberately discard so much talent, especially a country that had always appreciated ability and done so much to develop it? From Göttingen the losses included the Jewish mathematicians Edmund Landau (1877–1938), who remained in Germany without a position until his death in 1938; Paul Bernays, Landau's student and Hilbert's assistant, whom we have mentioned above in connection with set theory; Courant and Noether, already mentioned; Hans Lewy (1904–1988), Courant's student, who went first to Rome and then to Brown University, ending his career at the University of California; and Herbert Busemann (b. 1905), another student of Courant, who eventually moved to California. Weyl decided to leave for his wife's sake and out of principle; he eventually came to the Institute for Advanced Study in Princeton. Neugebauer was under suspicion for his liberal politics. Although his interests were originally in analysis [he is the co-discoverer with Harald Bohr (1887–1951) of the Bohr–Neugebauer theorem about almost-periodic solutions of differential equations], his interests shifted toward the history of mathematics. He eventually came to Brown University and became the leader of America's best-known school of history of mathematics. We have mentioned his work on the cuneiform texts in Chapter 3.

Göttingen was purged of Jewish mathematicians, leaving a greatly impoverished group of scholars to continue its brilliant tradition, headed by the elderly Hilbert. In another of history's ironic twists, Hilbert had suffered from pernicious anemia during the 1920s, a disease that had previously been fatal. He was one of the first victims to be saved by vitamin injections. During his illness he had received a blood transfusion from Courant, leading to the bitter joke that after 1933 there was only one good mathematician left in Göttingen and even he had Jewish blood.

What happened in Göttingen was repeated all over Germany. In Bonn Otto Toeplitz (1881–1940), who had studied with Hilbert, was dismissed in 1933 and emigrated to Palestine; Felix Hausdorff, a multitalented genius, remained and was allowed to work, but in 1942, given the order to be deported to a concentration camp along with his family, he committed suicide. From Munich Salomon Bochner (1899–1982) fled to England and eventually to Princeton, and Friedrich Hartogs (1874–1943) was forcibly retired. From Hamburg Emil Artin (1898–1962) left to become professor at Notre Dame and Princeton (in 1958 he returned to Hamburg), and the young Max Zorn (1906–1993) emigrated to the United States. From Halle Reinhold Baer (1902–1979) went on sabbatical and never returned, finishing his career in the United States. From Berlin Stefan Bergmann (1898–1977) emigrated first to the Soviet Union, then to the United States, as did Richard von Mises (1883–1953); Leopold Löwenheim (1878–1957) and Issai Schur (1875–1941) were forcibly retired. From Breslau Max Dehn (1878–1952) and Hans Rademacher (1892–1969) came to the United States. From Frankfurt Ernst Hellinger (1883–1950) and Otto Szász (1884–1952) emigrated to the United States. In Freiburg Alfred Loewy (1873–1935) was forcibly retired in 1933 and

Ernst Zermelo, some of whose work was discussed above, resigned his honorary professorship in protest against the Nazis in 1935. In Giessen Ludwig Schlesinger (1864–1933) was forcibly retired and died shortly thereafter. In Tübingen Erich Kamke (1890–1961) was forcibly retired in 1937. Hans Reichenbach (1891–1953) fled from Stuttgart to Turkey and ultimately to the United States. Gabor Szegö (1895–1985) emigrated from Königsberg to St. Louis in 1935. Adolf Fraenkel emigrated from Marburg to Jerusalem in 1933. Carl Ludwig Siegel (1896–1981) took an extended leave of absence from Göttingen, spending the years from 1940 to 1951 in Princeton. And so it went; every major German university lost talented professors.

As the Nazi regime expanded and the danger of further expansion increased, so did the number of refugees. When Austria was annexed to Germany in 1938, Vienna lost Kurt Gödel, Eduard Helly (1884–1943), and Karl Menger (1912–1985). Bryn Mawr College acquired Olga Taussky from Vienna and Emmy Noether from Göttingen in 1934. When Germany occupied Czechoslovakia in 1939 Karl Loewner (1893–1968) left the German University in Prague and moved to the United States. Eduard Cech (1893–1960), a professor at Brno, was incarcerated from 1941 to 1945. Marc Kac (1914–1984) left the University of L'vov in the Ukraine to move to Cornell in 1938. When Poland was partitioned between Germany and the Soviet Union and the Baltic States were annexed to the Soviet Union, there were further losses. The Polish mathematician Antoni Zygmund (1900–1992) left Vilnius in 1939, the year before Lithuania was annexed to the Soviet Union. His student Joseph Marcinkiewicz (1910–1940) died in a Soviet prison the following year.

Of the German mathematicians who remained some were Nazis; others were not. Among the prominent Nazi supporters was Oswald Teichmüller (1913–1943), who wrote the best mathematics that appeared in the Nazi-sponsored journal *Deutsche Mathematik* and died on the Eastern front. Hilbert, who never supported the Nazis, remained in Germany and died in 1943. To counter the propaganda damage done by the expulsion of so many first-rate mathematicians, the Nazi public relations organs touted the superiority of German mathematicians such as the recently deceased Felix Klein.[9] Although the Western democracies benefited intellectually from the influx of immigrants, there were not enough university positions to absorb all of them. Until December 1941 America hoped to remain neutral and was very reluctant to accept the refugees. As a result, many wasted their talents earning a living in occupations in which they were seriously underemployed. A few scholars from Germany fled eastward to the Soviet Union rather than westward. Among them were Stefan Cohn-Vossen (1902–1936), who fled from Köln to Leningrad in 1934, and Emmy Noether's brother Fritz (1884–1941), another talented mathematician, who became a professor at Tomsk, but was arrested, ironically accused of being a German spy, in 1937 and executed in 1941, during the German invasion of Russia.

What is to be learned from this horrendous story? Readers may draw their

[9]So far as one can tell from his writings, which were full of sympathy and tolerance, Klein would have been horrified to find his name used in this way. He was, incidentally, suspected of being Jewish, and a thorough investigation of his genealogy was conducted before he was made into a Nazi icon.

own conclusions. Our purpose is to sketch the colossal waste of human beings, the destruction and disruption of lives. That the Nazis did incalculable damage to their own country in their efforts to be rid of "foreign" influences in no way ameliorates the horror of the Nazi program.

In comparison with the human cost of Nazism, the intellectual cost is trivial. Nevertheless, a comparison of German mathematical journals from the nineteenth century such as the *Mathematische Annalen* (edited by Felix Klein, among others) with those from the 1930s makes it clear how much had been lost. The regime supported a journal called *Deutsche Mathematik*, edited by Ludwig Bieberbach (1886–1982). This journal was published for six years, starting in 1936, until the disruption of the war made it unfeasible. It contained some articles of respectable profundity, such as those of Oswald Teichmüller mentioned above, but was not even remotely comparable in quantity or quality with the *Mathematische Annalen* or the *Journal für die reine und angewandte Mathematik*.

19.5.3 Mathematics and American Scientific Policy

The Nazi destruction of scholarship crippled the German war effort by just enough to prevent the development of nuclear weapons. The decay of uranium through fission had been demonstrated in 1939 by Otto Hahn (1879–1968), a loyal servant of the Nazi regime. (He received the Nobel Prize for this work in 1944.) Fortunately Germany had neither the scholars nor the resources to develop the atomic bomb; otherwise the war might have been considerably prolonged. As it was, the British and Americans were able to pool their resources in a joint project that came to fruition just as Germany surrendered.

The American–Soviet rivalry after the end of the war and America's assumption of the role of leader of the Western democracies had several consequences for American science and mathematics. Suspicion of Communism in the early and mid-1950s led to the dismissal of a few professors. D. J. Struik (b. 1894), the author of a standard work on the history of mathematics and a prominent differential geometer, was placed on leave from the Massachusetts Institute of Technology for his Marxist views. (By way of apology in the 1980s the Commonwealth of Massachusetts named him an outstanding citizen.)

American fears of Communism were greatly exacerbated in October 1957 when the USSR launched the first artificial satellite. Less than 4 years later the USSR launched the first space ship with a human pilot. America was not able to match either of these feats until many months later, and the general belief was that the USSR was far ahead in space. The shock to American complacency and pride led to a vast increase in government support of science and an expansion of National Science Foundation programs of support for graduate students. The early 1960s were a period when money flowed easily to researchers in any area of science from a variety of sources. A great many researchers in pure mathematics were supported by agencies of the armed services such as the Air Force Office of Scientific Research and the Office of Naval Research.

This period of prosperity for American universities came to an end for two

reasons. The first of these reasons was the Vietnam War, which was the most unpopular war of the twentieth century. Although discontent with the war was widespread, the demonstrations against it tended to be centered in the universities, where there were large numbers of young men vulnerable to the draft. These demonstrations alienated the government and some of the public from the universities. Against this background and the horrors of such atrocities as the My Lai massacre researchers had to search their consciences to decide what measures, if any, were justified by the aim of the war—to halt the spread of Communism in Asia—and whether those measures were being reasonably applied. It became a serious question for some mathematicians whether they ought to cooperate with the armed services at such a time. Needless to say, the armed services were not likely to look favorably on requests for funding from researchers who were known for denouncing the military.

The second, less political reason for the decline in support for pure science was a philosophical debate over the soundness of such support. In the mid-1960s articles appeared in various newspapers and magazines pointing out that mathematics is not a science, and that, despite certain vague perceptions of the public, research into Riemannian manifolds had nothing to do with putting people on the moon.

The perceived need for scientific research declined sharply in the years after 1969, when America succeeded in sending several expeditions to the moon. The popular concern shifted from the space race to the problem of controlling pollution, in which science itself was suspect. At the same time the armed services, under pressure from Senators Mansfield of Montana and Proxmire of Wisconsin, were required to justify any support for research on the basis of practical military needs. The result was a precipitous decline in the level of funding for mathematical research. All these events helped to begin a debate on the role of government policy in science and the role of mathematics in that policy, a debate that continues today. The issue is at bottom one of public interest and individual interest. Does the country as a whole have an interest in the promotion of scientific research? What benefits does it bring that could not be obtained without government subsidy? Do citizens benefit from this research in a way that justifies taxing them to support it?

19.6 The World of Mathematics Today

Mathematical research is now a thriving enterprise in nearly every country in the world. Already by the end of the nineteenth century it was decided to hold international meetings. The first of these was held at Chicago in connection with the World's Columbian Exposition in 1893, and the featured speaker was Felix Klein. This meeting is often referred to as the Zeroth International Congress. At the Second International Congress, held in Paris in 1900, the acknowledged leader in several areas of mathematics, David Hilbert, gave the keynote address, listing 23 important current problems that he hoped would be solved during the twentieth century. Hilbert's calling attention to these problems made them the object of intensive research, and many were solved or shown to be unsolvable

in the course of the century. International congresses have been held regularly during the twentieth century except during the two world wars, and they are now a quadrennial event.

With thousands of talented researchers working all over the world, duplication and priority disputes were bound to proliferate. To solve these problems more and more journals and mathematical societies were founded. The mathematical societies of various countries became the clearinghouses for mathematical information, supplementing the mathematical sections of the Academies of Science. On this basis mathematics continued to grow every year except during the two world wars until it reached its present dimensions.

The scale of this enterprise can be judged from the 1993–94 membership rolls of the three largest American societies: the American Mathematical Society (AMS: 27,333 members), the Mathematical Association of America (MAA: 34,844 members), and the Society for Industrial and Applied Mathematics (SIAM: 7915 members), with a total membership of 57,075 for the three organizations. The members of these organizations are engaged in mathematics through research, teaching, and application. The number of articles written on new research is so large that no library could possibly afford to subscribe to all of the hundreds of journals in which it is printed. The *Mathematical Reviews* in America, the *Zentralblatt für Mathematik* in Germany, and the *Referativnyi Zhurnal Matematiki* in Russia each publish reviews of some 50,000 books and articles per year, written by more than 60,000 authors (many articles have more than one author). These articles and books are classified according to a scheme worked out by the American Mathematical Society and the publishers of the *Zentralblatt für Mathematik* into 61 areas, each having from three to a dozen specialties, each with 5–20 subspecialties. A reviewer is often hard-pressed to say which of these minute areas constitutes the primary subject matter of an article. A typical research mathematician may attempt to keep up with current work in a few subsubspecialties.

Obviously the centrifugal forces acting on modern mathematics are enormous. Like a carousel spinning out of control, mathematical research forces its practitioners farther and farther from one another. This problem was recognized already in the early twentieth century, and attempts were made to remedy it by publishing fairly detailed surveys of the current state of various mathematical sciences. This project, the *Enzyklopädie der Mathematischen Wissenschaften*, produced many thousands of pages of good exposition in German and was translated into French. However, the project was essentially hopeless, being aimed at a hypothetical broadly educated person, at a time when only a fairly profound specialist could begin to appreciate what was happening in a given area. One of the current attempts to preserve unity in mathematics, the *Lecture Notes in Mathematics* series published by Springer-Verlag, is aimed only at making specialized areas accessible to new researchers and nonspecialists. It now constitutes some 1500 volumes. As this example shows, one of the largest problems for modern mathematics is to get the known material organized in such a way that it is accessible to a person with a certain amount of core knowledge.

Difficult though it is, this problem is being solved. The *Mathematical Reviews* keeps cross-indices of its reviews by subject and by author (the indices alone oc-

cupy about 3500 pages each year), and the reviews themselves can be accessed electronically. Since the mid-1980s titles and authors have been included in several data bases. These aids to research make the specialist's job much easier. Researchers in many fields tend to form electronic-mail networks to keep in touch with current work all over the world. As a result, the hard copy of an article that appears in a printed journal becomes increasingly redundant. Being written as concisely as possible, a typical article is opaque to the nonspecialist or the student attempting to begin research in a field, while the specialist, in all likelihood, has already heard about the results and verified the proof based on a sketch of the method used.

Thus the branches of mathematics grow longer and thinner each year, and some seem in danger of breaking off entirely. Yet there remains a common core of mathematics. All graduate schools in America require students to have a thorough knowledge of real analysis, complex analysis, and algebra before proceeding to do research in more specialized areas, and in each small area of research there are good expository works that trace a path from this common core to the current research. With these aids and a good advisor, students continually take up rather arcane research and add to the flood of new articles each year that fill journals published in many different countries.

19.7 Problems and Questions

19.7.1 Problems in Contemporary Mathematics

Exercise 19.1 The most important property of a distance is the triangle inequality: $d(x,y) \leq d(x,z) + d(z,y)$, which says that the distance from x to y does not exceed the distance from x to z plus the distance from z to y. Consider the set of continuous functions on $[0,1]$, with the "distance" from a function $f(x)$ to the function $g(x)$ being defined in the following three ways:

(a) $d(f,g) = \int_0^1 |f(x) - g(x)|\, dx$;

(b) $d(f,g) = \left(\int_0^1 |f(x) - g(x)|^2\, dx \right)^{1/2}$;

(c) $d(f,g) = \left(\int_0^1 |f(x) - g(x)|^{\frac{1}{2}}\, dx \right)^2$.

Which of these functions $d(f,g)$ satisfy the triangle inequality?

Exercise 19.2 Bertrand Russell pointed out that some applications of the axiom of choice are easier to avoid than others. For instance, given an infinite collection of pairs of shoes, describe a way of choosing one shoe from each pair. Could you do the same for an infinite set of pairs of socks?

Exercise 19.3 Prove that $C = \{x : x \notin x\}$ is a proper class, not a set, that is, it is not an element of any class.

Exercise 19.4 Suppose the only allowable way of forming new formulas from old ones is to connect them by an implication sign, that is, given that A and B are well formed, $[A \Rightarrow B]$ is well formed, and conversely if A and B are not both well formed, then neither is $[A \Rightarrow B]$. Suppose also that the only basic well-formed formulas are p, q, and r. Show that

$$\Big[[p \Rightarrow r] \Rightarrow \big[[p \Rightarrow q] \Rightarrow r\big]\Big]$$

is well formed but

$$\big[[p \Rightarrow r] \Rightarrow [r \Rightarrow]\big]$$

is not. Describe a general algorithm for determining whether a finite sequence of symbols is well formed.

Exercise 19.5 Consider the following theorem. There exists an irrational number that becomes rational when raised to an irrational power. Proof: Consider the number $\theta = \sqrt{3}^{\sqrt{2}}$. If this number is rational, then we have an example of such a number. If it is irrational, then the equation $\theta^{\sqrt{2}} = \sqrt{3}^2 = 3$ provides an example of such a number. Is this proof intuitionistically valid?

Exercise 19.6 Show that any two distinct *Fermat numbers* $2^{2^m} + 1$ and $2^{2^n} + 1$, $m < n$, are relatively prime. (Use mathematical induction on n.) Apply this result to deduce that there are infinitely many primes. Would this proof of the infinitude of the primes be considered valid by an intuitionist?

Exercise 19.7 Suppose you prove a theorem by assuming that it is false and deriving a contradiction. What you have then proved is that either the axioms you started with are inconsistent, or the assumption that the theorem is false is itself false. Why should you conclude the latter rather than the former? Is this why some mathematicians have claimed that the practice of mathematics requires faith?

19.7.2 Questions about Contemporary Mathematics

Exercise 19.8 What are the advantages, if any, of building a theory by starting with abstract definitions, then later proving a structure theorem showing that the abstract objects so defined are really composed of familiar simple objects?

Exercise 19.9 L. E. J. Brouwer, the leader of the intuitionist school of mathematicians, is also known for major theorems in topology, including the *Brouwer fixed-point theorem*, which asserts that for any continuous mapping f of a closed disk into itself there is a point x such that $x = f(x)$. To prove this theorem, suppose there is a continuous mapping f for which $f(x) \neq x$ at every point x. Construct a continuous mapping g by drawing a line from $f(x)$ to x and extending it to the point $g(x)$ at which it meets the boundary circle (see Fig. 19.2). Then $g(x)$ maps the disk continuously onto its boundary circle and leaves each point of the boundary circle fixed. Such a continuous mapping is intuitively impossible (imagine stretching the entire head of a drum onto the rim without moving any point

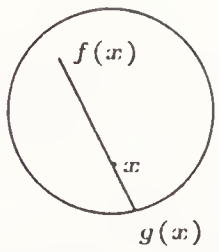


Figure 19.2: The Brouwer fixed-point theorem.

already on the rim and without tearing the head) and can be shown rigorously to be impossible (the disk and the circle have different homotopy groups). How can you explain the fact that the champion of intuitionism produced theorems that are not intuitionistically valid?

Exercise 19.10 What are the possible advantages and disadvantages of eliminating or greatly reducing the volume of journals, instead placing all articles on electronic files that can be downloaded from various information systems?

Exercise 19.11 On the basis of the geometric series $1/(1+x) = 1-x+x^2-x^3+\cdots$ Euler was willing to say that $1 - 5 + 25 - 125 + \cdots = \frac{1}{1+5} = \frac{1}{6}$. Later analysts rejected this use of infinite series and confined themselves to series that converge in the ordinary sense. However, Kurt Hensel (1861–1941) showed in 1905 that it is possible to define a notion of distance (the p-adic metric) by saying that an integer is close to zero if it is divisible by a large power of the prime number p (in the present case $p = 5$). Specifically, the distance from m to 0 is given by $d(m, 0) = 5^{-k}$, where 5^k divides m but 5^{k+1} does not divide m. The distance between 0 and the rational number $r = m/n$ is then by definition $d(m, 0)/d(n, 0)$. Show that $d(1, 0) = 1$. If the distance between two rational numbers r and s is defined to be $d(r - s, 0)$, then in fact the series just mentioned does converge to $\frac{1}{6}$ in the sense that $d(S_n, \frac{1}{6}) \to 0$, where S_n is the nth partial sum.

What does this historical experience tell you about the truth or falsity of mathematical statements? Is there an "understood context" for every mathematical statement that can never be fully exhibited, so that certain assertions will be *verbally* true in some contexts and verbally false in others, depending on the meaning attached to the terms?

Exercise 19.12 Are there true but unknowable propositions in everyday life? Suppose your class meets on Monday, Wednesday, and Friday. Suppose also that your professor announces one Friday afternoon that you will be given a surprise exam at one of the regular class meetings the following week. One of the brighter students then reasons as follows. The exam will not be given on Friday, since if it were, having been told that it would be one of the 3 days, and not having had it on Monday or Wednesday, we would know on Thursday that it was to be given on Friday, and so it wouldn't be a surprise. Therefore it will be given on Monday or Wednesday. But then, since we *know* that it can't be given on Friday, it also can't be given on Wednesday. For if it were, we would know on Tuesday that it was to be given on Wednesday, and again it wouldn't be a surprise. Therefore it must be given on Monday, we know that now, and therefore it isn't a surprise. Hence it is impossible to give a surprise examination next week.

Obviously something is wrong with the student's reasoning, since the professor can certainly give a surprise exam. Most students, when trying to explain what is wrong with the reasoning, are willing to accept the first step. That is, they grant that it is impossible to give a *surprise* exam on the *last* day of an assigned window of days. Yet they balk at drawing the conclusion that this argument implies that the originally next-to-last day must thereby become the last day. Notice that, if the professor had said nothing to the students, it would be possible to give a surprise exam on the last day of the window, since the students would have no way of knowing that there was any such window. The conclusion that the exam cannot be given on Friday therefore does not follow from assuming a surprise exam within a limited window alone, but rather from these assumptions supplemented by the following proposition: *The students know that the exam is to be a surprise and they know the window in which it is to be given.*

This fact is apparent if you examine the student's reasoning, which is full of statements about what the students *would know*. Can they truly *know* a statement (even a true statement) if it leads them to a contradiction?

Explain the paradox in your own words, deciding the question whether the exam would be a surprise if given on Friday. Can the paradox be avoided by saying that the conditions under which the exam is promised are true, but the students cannot *know* that they are true?

Exercise 19.13 Mathematical research is like any other commercial commodity in the sense that people have to be paid to do it. We have mentioned the debate over taxing the entire public to support such research and asked the student to consider whether there is a national interest that justifies this taxation. A similar taxation takes place in the form of tuition payments to American universities. Some of the money is spent to provide the salaries of professors who are required to do research. Is there an educational interest in such research that justifies its increased cost to the student?

19.8 Endnotes

1. A detailed account of the development of set theory and the issues surrounding the axiom of choice can be found in the book by Gregory H. Moore, *Zermelo's Axiom of Choice: Its Origins, Development, and Influence* (Springer-Verlag, New York, 1982).

2. Russell's comments on Cantor's proof that there is no largest cardinal number were made in an essay entitled "Recent work in the philosophy of mathematics," *International Monthly*, 1901, and reprinted as "Mathematics and the Metaphysicians" in the book *Mysticism and Logic* (Longmanns, Green, & Co., London, 1921), p. 89.

3. The comment that mathematicians have faith in the consistency of set theory can be found in the book *Introduction to Set Theory*, by J. Donald Monk (McGraw-Hill, New York, 1969), p. 22.

4. The brief sketch of Benjamin Banneker is based on his *Almanac* and the biography *The Life of Benjamin Banneker* by Silvio A. Bedini. (Charles Scribner's Sons, New York, 1972).

5. The account of Soviet mathematics is based on several sources, including *Science in Russia and the Soviet Union* by Loren Graham (Cambridge University Press, 1993); "Dmitrii Egorov: Mathematics and Religion in Moscow," by Charles Ford, *The Mathematical Intelligencer*, **13** (2), (1991), pp. 24–34; "Mathematics in Moscow in the 1930s," by S. S. Demidov (manuscript); and *Directives of the All-Union Communist (Bolshevik) Party on Public Education. A Collection of Documents from 1917 to 1947* an appendix to the journal *Soviet Teacher*, No. 2, assembled by Candidate of Pedagogical Sciences N. I. Boldyrev (Academy of Pedagogical Sciences of the Russian SFSR, Moscow/Leningrad, 1947) (in Russian).

6. The account of Nazi mathematics is based on *Scientists under Hitler* by Alan D. Beyerchen (Yale University Press, 1977) and *Midwives to Nazism* by Alice Gallin (Mercer University Press, 1986). The list of refugees from the Nazis was culled from *Lexikon Bedeutender Mathematiker* (Biographisches Institut, Leipzig, 1990). Other information came from the article "Jewish Mathematicians at Göttingen in the Era of Felix Klein" by David Rowe, in *Isis*, 77 (288), (1986), pp. 422–250 and the article "Fritz Noether—Opfer zweier Diktaturen," by Karl-Heinz Schlote, in *Schriftenreihe für Geschichte der Naturwissenschaften, Technik, und Medizin*, **28** (1), (1991), pp. 33–43.

7. More information on African-American mathematicians and women mathematicians, in particular Charlotte Angas Scott and Julia Bowman Robinson, can be found in *A Century of Mathematics in America, Part III*, edited by Peter Duren, (American Mathematical Society, Providence, RI, 1989).

8. The information on African mathematics is partly derived from a conversation with Prof. Madieyna Diouf of Université Chesikh Anta Diop in Dakar.

Answers to Selected Exercises

Exercise 1.2. Addition, subtraction, and multiplication are constantly used when making out tax returns; in that context division also occurs, although disguised as multiplication by a percent. Obviously also one uses addition when deciding how much carpet to buy to cover the floors in several rooms of a house (add the areas of the individual rooms, plus an allowance for wastage), addition and subtraction in balancing a checkbook, multiplication when computing the area of a rectangular floor or wall to be covered or painted, and division when "splitting" a restaurant check equally among a group of diners.

These operations involve proportion when, for example, deciding how much paint or varnish to buy for a given job. For example, if one gallon covers 300 square feet, the following proportion is used:

paint required : 1 gallon = area to be covered : 300 square feet.

A second example, less pleasant, concerns fines for speeding. The fine is usually court costs plus a certain amount per mile of excess speed. A third example comes from "tax brackets." The additional tax due within each income bracket is (approximately) proportional to the additional income in that bracket.

Exercise 1.5. All the numbers that we use in everyday life, including especially those that we enter into computers, are expressible using a finite number of binary digits. In other words, they are rational numbers expressible using denominators that are powers of 2. Yet human thought makes powerful use of geometry, and geometry requires incommensurables. Those who wish to comprehend as much as possible of the universe will wish to reconcile these two powerful modes of thought, the discrete and the continuous. The problem of creating a constructive foundation for analysis has occupied some very good mathematicians in the twentieth century.

Those whose desire to feel is stronger relative to their desire to understand—the philosopher Henri Bergson, for example—tend to reject discrete concepts entirely, implicitly denying the possibility that continuity can be analyzed logically. The right-brain/left-brain dichotomy that has engaged the popular imagination lately seems to mirror this dispute, but not enough is known at present to draw any definite conclusions.

Exercise 1.11. The author can think of no answer to this question other than to introduce a standard of constant velocity and use the proportionality between time

and distance in such a motion in order to infer elapsed time from the distance covered. The current standard is based on vibrations of atoms; in the past the motion of stars and the swinging of a pendulum were accepted as examples of steady motion from which elapsed time could be inferred. It is a difficult epistemological question whether this proportionality is more than a human convention, that is, whether it expresses a relation between real objects. We are on the safer ground of common sense when we compare the different standards to see if they are consistent. (They are not; according to the atomic standard, the stars are slowing down.) Congruence of time intervals must be simply *defined* to mean that equal numbers of standard time units elapse during the intervals.

Exercise 1.17. Despite the apparent difficulty of this problem, the solution is surprisingly easy and is achieved by imagining that the four lines containing the outside vertices are walls and moving the way a tennis ball would bounce off these walls.

Exercise 2.2. Using modern symbols, we write

$*$	1	42
$*$	2	84
	4	168
$*$	8	336
$*$	16	672
Result	27	1134

Exercise 2.4. When each of the fractions in the sum is multiplied by 45, the results are respectively $11\ \overline{4}$, $5\ \overline{2}\ \overline{8}$, $4\ \overline{2}$, $1\ \overline{2}$, and 1. The fractional parts here are $\overline{4}$, $\overline{8}$, $\overline{2}$, $\overline{2}$, and $\overline{2}$, which total $1\ \overline{2}\ \overline{4}\ \overline{8}$. Hence the magnified sum is $23\ \overline{2}\ \overline{4}\ \overline{8}$, while $\overline{\overline{3}}$ magnifies to 30. Thus we are lacking $6\ \overline{8}$, and so we must "calculate with 45 to obtain $6\ \overline{8}$." The scribe was apparently guided by the knowledge that $45 = 9 \times 5$ and so used a procedure similar to the following:

1	45	
$\overline{9}$	5	$*$
$\overline{45}$	1	$*$
$\overline{360}$	$\overline{8}$	$*$

When the last the entries in the left-hand column are combined, it is easy to remember that the term $\overline{45}$ is 8 times $\overline{360}$, so that $\overline{45} + \overline{360}$ is 9 times $\overline{360}$, which is $\overline{40}$.

Exercise 2.7. *Solution.* We first ask how the number 97 pops out here. The number 16 is merely a reasonable starting point, which might easily have been different. Having chosen that point and performed the indicated operations on it, the scribe would have found $36\ \overline{\overline{3}}\ \overline{4}\ \overline{28}$ (since the last two terms represent what in our language is the fraction $\frac{2}{7}$). Thus the scribe would be trying to complement $\overline{\overline{3}}\ \overline{4}\ \overline{28}$ to get 1. Following the standard procedure for doing such things, the scribe

might have used a common denominator of 84, and the expression $\overline{\overline{3}}\ \overline{4}\ \overline{28}$, when multiplied by 84, yields 80. It would then be easy to recognize that 4 parts out of 84 were lacking, that is, that the twenty-first part was needed. The problem would then be to "calculate with $1\ \overline{\overline{3}}\ \overline{2}\ \overline{7}$ so as to obtain $\overline{21}$." In other words one would like to perform a calculation having the two rows

$$\begin{array}{cc} 1 & 1\ \overline{\overline{3}}\ \overline{2}\ \overline{7} \\ & \text{result} \quad \overline{21} \end{array}$$

If we assume that the scribes had an intuitive grasp of the fact that the rows of these computations are proportional, we must believe that with the large number of computations they performed they could not help realizing that one can "interchange means and/or extremes," so that the same "result" would occur if the computation became

$$\begin{array}{cc} \overline{21} & 1\ \overline{\overline{3}}\ \overline{2}\ \overline{7} \\ & \text{result} \quad 1 \end{array}$$

This last row can be achieved by proceeding as follows:

$$\begin{array}{ccc} \overline{21} & & 1\ \overline{\overline{3}}\ \overline{2}\ \overline{7} \\ 1 & 21\ 14\ 10 & \overline{2}\ 3 \\ 1 & & 48\ \overline{2} \end{array}$$

At this point, since we are seeking a 1 in the bottom row of the right-hand column, it is natural to double the row, then divide it by 97. Dividing 2 by an odd number is precisely what the table allows one to do. (Indeed this computation suggests that division as we know it may have been thought of as multiplying by the corresponding part.)

Thus one needs the double of the 97th part. The rest is then simply a matter of looking in the table of doubles.

Exercise 2.11. The frustum can be thought of as the remainder after a smaller pyramid is chopped off of the top of a larger one. The heights of the two pyramids are, say g and $g+h$, and the proportion between g and h is derived from

$$\frac{g}{g+h} = \frac{a}{b},$$

so that $g = ah/(b-a)$. Since the volume of a pyramid is $\frac{1}{3}$ times the area of the base times the height, we find the volume of the frustum to be

$$\begin{aligned} \frac{1}{3}(b^2(g+h) - a^2 g) &= \frac{1}{3}(g(b^2-a^2) + b^2 h) \\ &= \frac{1}{3}(ah(b+a) + b^2 h) \\ &= \frac{h}{3}(a^2 + ab + b^2). \end{aligned}$$

Exercise 2.15. One imaginative possibility is that the user is a bureaucrat charged with licensing a jug for the sale of beer. Suppose the jug is emptied into the

standard state-approved jug three times and then when one-third of the jug is added, the standard jug is filled to the brim. What volume should be assigned to the jug?

Exercise 3.2. By the principles used on the tablets, the average of the two numbers is $\frac{5}{2}$, and their semidifference is

$$\sqrt{\left(\frac{5}{2}\right)^2 - \frac{56}{9}} = \sqrt{\frac{1}{36}} = \frac{1}{6}.$$

The two numbers are therefore $\frac{5}{2} \pm \frac{1}{6}$, which is to say $\frac{8}{3}$ and $\frac{7}{3}$.

Exercise 3.6. One possible answer is accident: the author was looking for only one solution, and this is the one found. A more substantive answer is that the method of solving the problem, which involved adding 2 units to the "width," would in the second case lead to a width that was larger than the length, contradicting the meaning of the word *length*.

Exercise 3.7. Let us be frank! There are no applications of quadratic equations in everyday life. Certain linear problems of great practical value—input/output analysis, for example—may lead to the need to solve higher-degree equations in order to find eigenvalues and eigenvectors, but that is a technical use. Similarly, one might wish to solve the problem of analyzing the motion of a heavy body thrown upward from a certain height h with velocity v. According to the simplified Newtonian model, neglecting air resistance, among other things, its height in meters at time t seconds after it is thrown will be $h + vt - 4.9t^2$, where h is in meters and v in meters per second. Then any question as to the time at which the object will have a given height H becomes a quadratic equation. One can hardly consider such questions "practical," yet they might have occurred to someone of a speculative bent.

Exercise 4.1. The division algorithm yields the following quotients and remainders. The last nonzero remainder is the greatest common divisor (819).

$$\begin{aligned} 189,189 &= 13 \cdot 13,923 + 8190 \\ 13,923 &= 1 \cdot 8190 + 5733 \\ 8,190 &= 1 \cdot 5733 + 2457 \\ 5,733 &= 2 \cdot 2457 + 819 \\ 2,457 &= 3 \cdot 819. \end{aligned}$$

This computation can be conveniently performed on paper by working from right to left:

	3	2	1	1	13
819	$\overline{\vert 2457}$	$\overline{\vert 5733}$	$\overline{\vert 8190}$	$\overline{\vert 13,923}$	$\overline{\vert 189,189}$
	$\underline{2457}$	$\underline{4914}$	$\underline{5733}$	$\underline{8,190}$	$\underline{139,23}$
	0	819	2457	5,733	49,959
					$\underline{41,769}$
					8,190

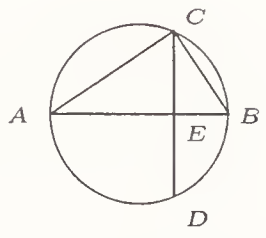


Figure A.1. Angle inscribed in a semicircle.

Exercise 4.5. Side AD is common to triangles ADE and ADB. Since AD is the bisector of $\angle CAB$, $BD \perp AB$, and $DE \perp AC$, it follows that triangles ADE and ADB are congruent by angle–angle–side. Since $\angle CDE$ must equal $\angle CAB$ (because both must be complementary to $\angle C$ in their respective right triangles), it follows that $\angle CDE$ is congruent to $\angle C$, hence that triangle DCE is isosceles. Therefore $EC = ED = BD$. We have $AB = AE$ by the congruence of triangles ADB and ADE.

Starting the Euclidean algorithm with the pair (AC, AB), we get $(AC - AB, AB) = (AC - AE, AB) = (EC, AB) = (BD, BC)$. Since $BC > BD$, our next pair is $(BC - BD, BD) = (CD, BD) = (CD, DE)$, which, as asserted, form the diagonal and side of a square.

Exercise 4.9. The proof that opposite angles are equal is obvious; each forms a straight angle with each angle of the other opposite pair.

The proof that the base angles of an isosceles triangle are equal is most simply achieved by imagining the triangle picked up and turned over, so that each base angle lies on the space previously occupied by the other.

The proof that a circle is bisected by a line through its center is proved by imagining the disk folded along the line through the center, so that the two halves of the circle coincide. They must do so, since any chord perpendicular to the diameter is bisected by the diameter. (That fact in turn can easily be seen by drawing the radii to the endpoints of the chord and observing the two congruent right triangles that result.)

The proof that two triangles are equal if two sides and the included angle of one are congruent to two sides and the included angle of the other is proved by imagining the two included angles made to coincide, and then picking up one triangle and turning it over if necessary to assure that the endpoints of the sides of the angle must also coincide. Since the third side of each triangle is determined by those two endpoints, the third sides will then coincide as well.

The theorem that an angle inscribed in a semicircle is a right angle may have come from the simpler facts that a circle is bisected by a diameter and that angles inscribed in equal arcs are equal. For example, given an angle ACB inscribed in a semicircle with diameter AB, complete the circle and draw the chord CD perpendicular to AB, meeting AB in point E. Then angles BCE and EAC are inscribed in equal arcs $\overset{\frown}{BD}$ and $\overset{\frown}{CB}$. They are therefore equal, and since $\angle CAE + \angle ACE$ must total one right angle (these angles being the two acute angles of the right triangle ACE), angle BCE and angle ACE must also total one right angle. (See Fig. A.1.)

Exercise 4.12. A person who speaks truthfully and frankly is said to be a *straight-talker*. Uncomplicated people who are regarded as dull are sometimes said to be *one-dimensional*. A broadly educated person is said to be *well-rounded*. Words such as *rectitude*, coming from the Latin, have similar roots. Topological notions enter ethics in such words as *integrity* (wholeness) and *duplicity*, the image being that a person of integrity is consistent and can be relied upon, while a duplicitous (two-faced) person may appear to be of one opinion in one context and an opposite opinion in some other.

Exercise 4.15. Both are important mathematical activities, each serving as a stimulus to the other. When intuition discovers an important new insight, the attempt begins to formulate it in the most precise and economical terms possible and to find a rigorous proof. Often the discovery of lapses in rigor leads to new mathematical research of great importance. Failure to prove a proposition can mean that examples in which the contrary proposition holds can be found. A prominent example was the discovery by Niels Henrik Abel in the 1820s that a series of continuous functions could converge to a discontinuous function, contrary to a claim of Augustin-Louis Cauchy. An even more spectacular example is the discovery of non-Euclidean geometry, which resulted from failures to prove the parallel postulate.

Exercise 5.1. The difference of the squares of two numbers is wellknown to be the product of the sum and the difference of the two numbers. If the smaller square is 1, while the larger one is odd, it follows that the difference of the two squares is the product of two successive even numbers. Since one of these two successive even numbers is a multiple of 4, it follows that their product is a multiple of 8.

Exercise 5.4. Given a line segment of length a to be divided into mean and extreme ratio, the problem is to find two segments whose difference is a and whose product is a^2. Let one of the segments be z and the other $z - a$. The required condition is that $z(z - a) = a^2$, which yields the proportion $z - a : a = a : z$, which then extends to give $z - a : a = a : z = 2a - z : z - a$. This last extension follows since $z - a : z = 1 - (a : z) = 1 - ((z - a) : a) = 2a - z : a$, so that $z - a : a = (z - a : z)(z : a) = (2a - z : a)(a : z - a) = (2a - z : z - a)$. Since $2a - z$ is the length of the segment left when the segment of length $a - z$ is subtracted from the segment of length a, we see that the larger segment has to the given segment the same ratio that the given segment has to the shorter, and that the shorter has to its complement in the given segment, that is, the given segment has been divided into mean and extreme ratio.

As mentioned, the shorter of the two segments will provide the point at which the line is to be divided.

Exercise 5.6. The assertion that $\angle FCE < \angle ACD$ relies on the figure, in particular on the qualitative fact that points A and F lie on the same side of line BD. If two lines could intersect in more than one point, then point F might very well be on the other side of line BD. This is exactly what happens, for example, if A is the pole of a great circle containing the arc $\overset{\frown}{BC}$ on a sphere and the arc $\overset{\frown}{BC}$

is longer than one-fourth of a great circle. Doubling the great-circle arc $\overset{\frown}{BE}$ will bring the point F into the opposite hemisphere from A. Even if we exclude the possibility that lines can intersect in more than one point, the figure could still be qualitatively wrong. In particular, we could not assert that B and F are on opposite sides of the line containing A and C, since removing a line may fail to divide the plane into two disconnected half-planes. Regarding antipodal points of the sphere as identical, for example, produces a geometry in which two lines (great circles) intersect in only one point, but then the point F is identified with its antipodal point, and one of these is in the same hemisphere as B relative to the great circle containing A and C. The second point of intersection of the great circle containing B and F with the great circle containing B and C (the point antipodal to B) is now identified with B, and the point F lies on the arc from B to E shown in the figure.

Exercise 5.13. It is the author's opinion that, taken on his own ground, Socrates would win this debate. The modern "construction" of the real numbers would have too many nonconstructive elements for him to accept, no matter how clearly they were explained. The formalist idea that one could "interpret" line segments as numbers and thereby turn a line into a field, would not be acceptable to Socrates on metaphysical grounds. For Socrates, if Plato reports him truly, numbers were entities having a real existence in a perfect world, as were line segments. To call a line segment a number would have seemed to Socrates to be a factual error. But unless one allows either this approach or the nonconstructive definition of a real number as a Dedekind cut or an equivalence class of Cauchy sequences, it is difficult to find an acceptable definition that allows real numbers to be used for analytic geometry.

The most promising common-sense approach is to define a real number as a decimal or binary expansion, make the usual identifications for numbers having two such expansions (or else simply exclude expansions that end in an infinite string of 0s), and attempt to describe an algorithm for adding and multiplying such expansions. The algorithm, however, is very cumbersome.

Exercise 5.17. It is precisely because Euclid was systematic and stated explicitly what others may have considered obvious that attention was focused on the parallel postulate and the possibility of questioning its validity arose. Although it is natural to imagine that people would be in doubt about a proposition until it is proved, experience shows that doubts often require a long time to surface. Proof is indeed used to allay doubt, but only after it has arisen, and in some cases doubt was not even present in the early stages of the search for a proof. For example, mathematicians took for granted that the parallel postulate was true for many centuries, but ultimately the failure to prove it led to the conclusion that it could be denied without contradiction.

Set theory gives a good example of the opposite side of this coin. There the attempt to reach ultimate clarity in the formulation of mathematical concepts led to the creation of a "foundation" for mathematics that is subject to far more doubt than the propositions allegedly derived from it.

Exercise 6.1. Suppose we could square a segment of a circle whose central angle we knew. Since the segment together with the isosceles triangle enclosed by the radii to the vertices of the segment form a sector, we could then square that sector. Then, having the ratio of the central angle of that sector to a full revolution as a ratio of two lines, we take the square root of that ratio (specifically, square the rectangle that each of the two lines forms with any given line, then take the ratio of the sides of the resulting two squares). Then the side of the required square (equal to the whole circle) is the fourth proportional whose first three terms are the square roots just constructed and the side of the square equal to the sector.

Exercise 6.3. It is apparent that each of these right triangles will have an angle inscribed in an arc subtended by a chord equal to the side of the polygon. Hence the triangles are all similar. This means that the lines BB', CC', KK' and DM are proportional to the successive line segments on the line AA' as far as M. We can therefore add numerators and denominators, getting the expression

$$\frac{BB' + CC' + \cdots + KK' + LM}{AM},$$

which equals each of these ratios. Finally, if $A'B$ is joined, we obtain yet another right triangle $A'BA$ with an angle inscribed in the arc AB, and hence yet another similar triangle, as asserted.

Exercise 6.5. If add the line CO to the figure in the text, we see that the triangle CAO is a right triangle. Hence AO^2 is the mean proportional between its hypotenuse and the segment of the hypotenuse on the same side of the altitude OS as AO, that is, we have proved that $CA \cdot AS = AO^2$. The rest is mere substitution of equals for equals, using at the last step the facts that $MN = 2MS$, $OP = 2OS$, and $QR = 2SQ$.

Exercise 6.10. This is routine algebra. Transposing everything to the left-hand side and completing the square leads to

$$k\left(x^2 - \frac{C}{k}x + \frac{C^2}{4k^2}\right) + y^2 = \frac{C^2}{4k},$$

which then becomes

$$\frac{[x - (C/2k)^2]}{(C/2k)^2} + \frac{y^2}{(C/2\sqrt{k})^2} = 1.$$

Exercise 6.11. In general, any rectangle deficient by a square has area equal to the product of the two line segments into which it divides the line segment to which it is applied. Apollonius is asserting that (what we call) the foci are points at which this product is one-fourth the product of the major axis and the latus rectum. By the geometric way in which the ellipse is defined, it is clear that the square of the ordinate at the midpoint of the major axis (that is, the square on half of the minor axis) will be exactly one-fourth of the rectangle on the major axis and the latus rectum. (Simply put, if l is the latus rectum, and a and b are half of the major and

minor axes respectively, then $2b^2 = al$. Note that for a circle, where a and b are both equal to the radius, this formula gives l as the diameter of the circle.) Hence the assertion follows.

Exercise 6.12. The distance from a general point (x, y) to the line $ax + by = c$ is well-known to be $|ax + by - c|/\sqrt{a^2 + b^2}$. Hence the general equation is

$$|y| = \frac{r}{\sqrt{a^2 + 1}}|ax - y| = q|ax - y|,$$

where r is the ratio of the two distances and $q = r/\sqrt{a^2 + 1}$. By squaring both sides, transposing the left-hand side to the right, and then factoring the difference of the two squares, we obtain the equation

$$[aqx + (1 - q)y][aqx - (1 + q)y] = 0,$$

which represents a pair of lines through the intersection of the two given lines. A pair of intersecting lines is considered a degenerate hyperbola.

Exercise 6.17. The difference between the two numbers is less than 0.0000305675, that is, the gain in accuracy achieved by using $\frac{14688}{4673.5}$ as the value of π rather than $\frac{22}{7}$ is less than 0.001%. Such a small gain is certainly not worth the extra labor.

In mathematics simplicity of a result is very important. For most applications both inside and outside of mathematics, $\frac{22}{7}$ gives satisfactory precision. There is nothing practical to be gained, even when "practicality" is extended to mean usefulness in a mathematical argument, by introducing the more cumbersome expression.

Exercise 6.20 To extend Cavalieri's principle to the computation of lengths, one could define the "zero-dimensional volume" of the point of intersection of two lines as the cosecant of their angle of intersection, so that two lines intersecting at a right angle would have an intersection of zero-dimensional volume 1 and two lines that coincide would have an intersection whose zero-dimensional volume is infinite, as one would expect. Note that the cosecant is the same for any of the four angles formed by two intersecting lines, so that this concept is unambiguously defined. For two intersecting curves one could define the volume to be the volume of the intersection of their tangents at the point of intersection.

This definition would then give consistent results for lines and curves in a plane. Incidentally, it provides a theorem about plane curves: *Let $y = f(x)$ and $y = g(x)$ be plane curves having continuously turning nonvertical tangents at each point $x = c$ for all $c \in [a, b]$ If for all $c \in [a, b]$ the cosecant of the angle of intersection of the curve $y = f(x)$ with the vertical line $x = c$ bears the ratio r to the cosecant of the angle of intersection of the curve $y = g(x)$ with the same line, then the length of the former is r times the length of the latter.* The proof is the observation that the cosecant of the angle in question is the secant of the angle between the tangent line and the horizontal, that is, it is $\sqrt{1 + (f'(c))^2}$ and $\sqrt{1 + (g'(c))^2}$ for the two curves, and the integrals of these two functions give the arc lengths of the two curves.

The need to consider this case points to a perhaps unnoticed assumption in the original statement of the principle and a possibility of generalizing it. The unnoticed assumption was that the sections of the given figures are taken inside a space of the same dimension as the figures themselves. The possibility of considering, for example, one-dimensional sections of two-dimensional figures in three-dimensional space requires some convention such as the one just introduced for zero-dimensional sections of one-dimensional figures in two-dimensional space.

Exercise 7.3. By the commensurable case, the weight mA would balance the weight A placed at distance of CD from the fulcrum if placed at a distance of $(1/m)CD$ from the fulcrum. Hence the smaller weight nB would rise if placed this distance from the fulcrum, A being at the distance CD on the opposite side of the fulcrum. Since weight B balances A at the distance CD when B is placed at distance CE, it follows that nB will balance A at distance CD when placed at distance $(1/n)CE$. Therefore if nB is placed at distances $(1/m)CD$ and $(1/n)CE$, the weight placed at the former distance will rise. That distance must therefore be the smaller distance, that is, we find

$$\frac{1}{m}CD < \frac{1}{n}CE,$$

and therefore, multiplying this inequality by mn, that $nCD < mCE$, as required. (We have used here the obvious fact that if X at distance d balances Y at distance e and Z at distance f, then Y at distance e also balances Z at distance f. Archimedes did not state this fact, but it is easy to prove.)

Exercise 7.4. The inequality in question is obtained from the inequality in the text by setting $EB = c/\tan\alpha$ and $EG = c/\tan\beta$, where $c = AB = DG$. It asserts that

$$\frac{\alpha}{\beta} > \frac{\tan\alpha}{\tan\beta}$$

when $\alpha < \beta$. In our terms this inequality asserts that the function $f(x) = \frac{\tan x}{x}$ is strictly increasing, and it follows easily from calculus, since the equation $f'(x) > 0$ is easily converted to $x > \sin x \cos x$, which is obviously true.

Euclid would have had to spend some time learning our trigonometry (which he would probably have objected to on the grounds that it assigned numbers to line segments and thereby ignored the problem of incommensurables). He might even have been forced to restrict our trigonometric functions to the literal sense of ratios of line segments before he would agree to this statement. The use of trigonometry is an anachronism that distorts the history of the mathematics. However, it is sometimes useful as the starting point for understanding the situation in which an ancient mathematician was working, although Euclid himself would not have seen this situation in the same way.

Exercise 7.6. Intersect the cissoid with the parabola $y^2 = 9ax$. The x coordinate of the point of intersection satisfies the equation

$$x^3 + 6ax^2 + 12a^2x = a^3.$$

By adding $8a^3$ to both sides, we see that this equation becomes $x + 2a = \sqrt[3]{9}a$. Since x is determined as the length of the perpendicular to the y axis from the point of intersection of the cissoid and the parabola and a was a given length, it follows that the length $\sqrt[3]{9}a$ is determined. The length $\sqrt[3]{3}a$ is the mean proportional between this length and a.

Exercise 7.12. The lines from the earth to the center of the epicycle and to the sun on November 1 form a triangle when taken together with the radius of the epicycle. The radii of the epicycle and deferent can be taken as 1 and 24, respectively, and the angle between them as $180° - \frac{360°}{365.24} \times 123 = 58.76°$. The law of cosines then gives the third side of this triangle as 23.50. The law of sines then gives the angle opposite the radius of the epicycle as $2.09°$. Subtracting this amount from the $121.24°$ of progress made by the center of the epicycle, we find that the sun has reached a point $99.19° + 121.24° - 2.09° = 218.34°$ along the ecliptic from the vernal equinox. (The observed value was $218.64°$ in 1964, so that the theoretical error was 0.3 degrees, or 18 minutes of arc.)

Exercise 7.17. Very often nonmathematicians do not fully appreciate the economy of logic that results from a close argument. It is possible that Vitruvius simply did not notice that his argument was redundant on this point. A more substantive possibility is that if the crown is known or suspected to be an alloy of gold and silver, the relative amounts of each can be determined by the exact amounts of water displaced by equal weights of the two metals. Thus Archimedes could determine precisely the sum by which the goldsmith bilked the king.

Exercise 8.2. Even with the anachronistic introduction of the modern symbol x, Diophantus' solution looks strange. He really starts with the smallest piece mentioned, one-sixth of the second quantity, then expresses the two quantities in terms of this piece. His procedure for finding that smallest piece, however, is essentially the one we would use, namely the equation $10x + 80 = 100$. Note that not having a second letter for an unknown actually simplified the solution for Diophantus, since he was forced to choose as his unknown a common currency for all the quantities mentioned. When we do the problem the modern way, we focus on x and y and miss the fact that $y/6$ is a more natural unknown for the problem.

Exercise 8.4 Very little is left to do, given the explanation accompanying the problem. We merely observe that

$$(65)^2 = (63)^2 + (16)^2 = (60)^2 + (25)^2 = (56)^2 + (33)^2 = (52)^2 + (39)^2.$$

We now choose

$$\varsigma = \frac{65}{2 \cdot 63 \cdot 16 + 2 \cdot 60 \cdot 25 + 2 \cdot 56 \cdot 33 + 2 \cdot 52 \cdot 39} = \frac{65}{12,768}.$$

The first of the four numbers is then

$$\frac{2 \cdot 63 \cdot 16 \cdot (65)^2}{(12,768)^2} = \frac{8,517,600}{163,021,824}.$$

The other three numbers are the ones given in the text and are found in the same way.

Exercise 8.8. The crucial step would have been to interpret numbers as *ratios* of lines rather than simply as lines. As we saw, Pappus came close to taking that step, but he did not pursue the matter or link it with the solution of geometric problems by Diophantus' methods (if he was even aware of those methods).

Exercise 9.2. After the transactions are performed, the three people own the following sets of animals: (1) five thoroughbred horses, one draft horse, and one camel; (2) one thoroughbred horse, seven draft horses, and one camel; and (3) one thoroughbred horse, one draft horse, and eight camels. If all three of these menageries represent equal wealth, the prices of the animals can be compared by imagining that each of them gives away one animal of each kind. The three then possess respectively four thoroughbred horses, six draft horses, and seven camels, and these all represent equal wealth.

Since the animals represent wealth, such reasoning can be used to establish the relative value of any three different currencies, given collections of mixed currencies of equal value. Whether such data actually occur in monetary transactions, however, is doubtful. It is more likely that the relative values of coins are known in advance and the problem is to mix currencies so as to obtain equal values.

Exercise 9.3. If x must be an integer, the only solution is $x = 11$, since the only two square integers whose difference is twelve are 16 and 4. If x need only be rational, there are infinitely many possibilities. One can take $x = \left[3r + (1/r)\right]^2 - 5 = 9r^2 + 1 + (1/r^2)$, where r is any rational number. Obviously $x + 5$ is a perfect square, and $x - 7 = 9r^2 - 6 + (1/r^2) =$
$big[3r - (1/r)\big]^2$.

Exercise 9.8. For a quadrilateral of sides a, b, c, and d Brahmagupta's formula can be written as the equation

$$16A^2 = 8abcd + 2a^2b^2 + 2a^2c^2 + 2a^2d^2 + 2b^2c^2 + 2b^2d^2 + 2c^2d^2 - a^4 - b^4 - c^4 - d^4.$$

Now a necessary and sufficient condition for a quadrilateral to be inscribed in a circle is that one pair of opposite angles be supplementary. (It then follows that both pairs of opposite angles have this property, since the four angles taken together must sum to four right angles.) Considering a quadrilateral with sides of length a and b on one side of a diagonal of length e and sides of length c and d on the other side, the condition that the angles on opposite sides be supplementary says that their cosines must be negatives of each other. Using the law of cosines, we find

$$a^2 + b^2 - 2ab\cos\theta = e^2 = c^2 + d^2 - 2cd\cos\varphi.$$

Now if θ and φ total two right angles, we must have

$$\cos\theta = -\cos\varphi,$$

and therefore

$$a^2 + b^2 - c^2 - d^2 = 2(ab + cd)\cos\theta,$$

so that

$$\cos\theta = \frac{a^2 + b^2 - c^2 - d^2}{2(ab + cd)}.$$

Now the area of the quadrilateral is

$$A = \frac{1}{2}(ab\sin\theta + cd\sin\varphi).$$

Hence the condition for the vertices to lie on a circle is that

$$A = \frac{1}{2}(ab + cd)\sin\theta =$$

$$\frac{1}{2}(ab + cd)\sqrt{1 - \cos^2\theta} = \sqrt{\left(\frac{ab + cd}{2}\right)^2 - \left[\frac{(a^2 + b^2) - (c^2 + d^2)}{4}\right]^2}.$$

or

$$16A^2 = 4(ab + cd)^2 - [(a^2 + b^2) - (c^2 + d^2)]^2.$$

Expanding the two squares in this last expression and gathering like terms results in precisely the formula of Brahmagupta.

Exercise 9.11. When the Euclidean algorithm is applied in the case of the equation $118x = 1461y + 72$, the quotients are 12, 2, 1, 1, 1, 1, 1, so that we begin the *kuttaka* with the matrix

$$\begin{matrix} 12 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ -72 \\ 0 \end{matrix}.$$

which yields the general solution $x = -18,144 + 1,461r$, $y = -1,512 + 118r$. Taking $r = 13$ gives the smallest positive solutions: $x = 849$, $y = 22$.

Exercise 9.16 The sines and cosines used by Jyesthadeva are different for different values of r, in other words, what Jyesthadeva calls $\sin\theta$ is what we would think of as $r\sin\theta$, while his θ itself is the length of an arc that subtends an angle whose radian measure is what we call θ. The suggested value of x gives $\tan\theta = x$ (where θ represents the angle whose arc on a circle of radius r is the arc θ in Jyesthadeva's equation). Hence

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} + \cdots.$$

In order to get $\arctan 0.5$ to ten decimal places, we need $\left[(0.5)^{2n+1}/(2n + 10\right] < 10^{-10}$, that is, $(2n + 1)\cdot 2^{2n+1} > 10^{10}$. When $n = 14$, this inequality

holds. Hence 14 terms of the series suffice. That 13 terms do not suffice follows from the fact that

$$\frac{1}{27 \cdot 2^{27}} - \frac{1}{29 \cdot 2^{29}} > 10^{-10}.$$

In fact

$$\frac{1}{27 \cdot 2^{27}} - \frac{1}{29 \cdot 2^{29}} = \frac{4 \cdot 29 - 27}{27 \cdot 29 \cdot 2^{29}} = \frac{89}{29 \cdot 27 \cdot 2^{29}} > \frac{1}{9 \cdot 2^{29}},$$

and $9 \cdot 2^{29} = 4,831,838,208 < 10^{10}$.

Exercise 9.20. Every step suggested by the later commentator seems to flow naturally from the problem itself. The earlier conjecture based on Al-Hassar's rule would be reasonable if only the suggested values of a and r could be made to seem natural. Both explanations are based on the fundamental approximation $\sqrt{1+x} \approx 1 + \frac{1}{2}x$.

Exercise 9.28. The safest conclusion is that one should *never* trust a purely mathematical conclusion until it is checked by observation. When a large number of observations tend to support the general principles of large and complicated theories, however, one gains some additional comfort and confidence. In such a case a mathematical prediction gains its plausibility from the way in which it interlocks with a large number of known phenomena.

It is seldom realized, for example, that the launching of the first artificial earth satellites in the 1950s amounted to a new test of Newtonian mechanics, on a larger scale than had been available previously. By the time this test was conducted, however, no one was even interested in its value as a test of Newtonian mechanics. So many other scientific theories had been woven together with Newtonian mechanics by that time that any challenge to it (except the small corrections introduced by relativity and quantum mechanics) would have involved serious recasting of most of modern science.

Exercise 10.1. Although the problem from the *Nine Chapters* is stated in terms of proportion and the problem from the Ahmose Papyrus in terms of an arithmetic progression, both use the same underlying mathematical construction, as we can see by stating both in modern algebraic language. First the problem from the *Nine Chapters*:

$$x + 2x + 3x + 4x + 5x = 5;$$

Next the problem from the Ahmose Papyrus, letting the "last" be first:

$$a + (a+d) + (a+2d) + (a+3d) + (a+4d) = 100$$
$$7[a + (a+d)] = (a+2d) + (a+3d) + (a+4d).$$

The difference between the two now becomes clear: The problem from the *Nine Chapters* gives implicitly (in the language of proportion) the statement that the first term and the common difference of the progression are equal ($a = d = x$). The Egyptian problem gives two independent conditions by which both the first term of the progression and the common difference can be determined.

Exercise 10.4. The left-hand side is the correction in the square-root algorithm, that is, it must be $x(2p+x)$, where p is the previous approximation to $\sqrt{N}$. Thus $p = 4$. Assuming that the correction finishes the job, it follows that $(p+x)^2 = N$, that is, that $N = p^2 + x(2p+x) = 16 + 65 = 81$. Hence $p + x = \sqrt{N} = 9$, and so $x = 9 - p = 5$.

Exercise 10.9. The 7×7 board contains 49 equal squares. Removing the center square leaves an area of 48 squares, and symmetry shows that half of it lies inside the square on the hypotenuse and half outside. The total area inside this square is thus 25.

Exercise 10.10. The interpretation of the problem is that a line tangent to the circular wall of the fort (the line joining the two people) forms the hypotenuse of a right triangle with right angle at the center of the fort. The legs of this triangle extend 16 steps eastward and 135 steps southward beyond the wall of the fort. Let the radius of the fort be r. The two portions into which the radius to the point of tangency divides the hypotenuse can be written in terms of r by using similar triangles:

$$a = \frac{r(r+135)}{r+16}, \quad b = \frac{r(r+16)}{r+135}.$$

The area of the triangle can then be computed either as half the hypotenuse times the altitude $[r(a+b)]$ or as half the product of the legs $[(r+135)(r+16)]$. Setting these two expressions equal to each other, doubling, using the relation just given for a and b, and clearing out the denominators, we obtain the quartic equation

$$r^4 - 8640r^2 - 652,320r - 4,665,600 = 0.$$

It is here that the Chinese have an advantage over us. This equation has only one positive root. At this point in our study we would have to guess it. However, in trying to locate it approximately we would search for two successive integer values of r between which the expression on the left changes sign, and that procedure would lead us to the root, so to speak "by accident." We would notice, for example, that the left-hand side is negative when $r = 100$ and positive when $r = 150$ (choosing two values for which it can be computed rather easily). The Chinese were adept at numerical solutions and found the solution $r = 120$ in that way. To find the solution by radicals requires a technique that was invented in sixteenth-century Italy (see Chapter 14).

Exercise 10.14. The chief advantage of the numerical procedure is that it applies to equations of all degrees. The only additional labor for a higher-degree equation is the additional time and complexity resulting from more coefficients and more multiplications to be performed. Rather surprisingly, that advantage is also its chief disadvantage. Looking for closed-form solutions leads to the interesting discovery that there are qualitative, not merely quantitative, differences in equations of different degrees. In particular, the solutions can be obtained from the coefficients by a finite sequence of extractions of square and cube roots for equations of degree up to 4, but for fifth-degree equations and those of higher degree no such sequence of operations will solve every equation.

Exercise 11.1. By arranging the numbers in a circle or as three rows of 10 and crossing them off sequentially, one finds them disappearing in the following sequence: 10, 20, 30, 11, 22, 3, 15, 27, 9, 24, 7, 23, 8, 26, 14, 2, 21, 16, 6, 4, 1, 5, 13, 19, 12, 29, 18, 17, 28. The last one remaining bears number 25.

There certainly is a method by which the answer can be found on a computer. One can, for example, define the following function of two integer variables inductively:

$$f(n,0) = n, \quad n = 1, \ldots, 30.$$

If $f(n,k) \geq 10$, then $f(n,k+1) = f(n,k) - 10$; otherwise $f(n,k+1) = f(n,k) + 20 - k$. The function $f(n,k)$ gives the distance from the kth integer stricken from the list to the integer n at the time the kth integer is removed. If the computer is asked to print out the first value of k for which $f(n,k) = 0$, that value will be the order in which n is crossed off the list.

Thus, for example, $f(22,0) = 22$, $f(22,1) = 12$, $f(22,2) = 2$, $f(22,3) = 2 + 20 - 2 = 20$, $f(22,4) = 10$, $f(22,5) = 0$, reflecting the fact that 22 is the fifth number crossed off the list. This procedure needs some refinement for larger values of k, when the total number of integers remaining is less than 5.

Exercise 11.5. Simply by subtracting the areas of the three smaller circles from that of the larger, we find that $480 = 4 \cdot 120 = \pi D^2 - 2\pi d^2 - \pi(d+5)^2$. This equation can easily be converted into the second of the stated equations.

If we now draw the diameter of the largest circle that passes through the point of tangency of the two smallest circles, it also passes through the center of the third inside circle. Since the line joining the center of the outside circle to the center of one of the two smallest circles has length $(D-d)/2$, it follows from the Pythagorean theorem that the distance from the center of the outside circle to the point of tangency of the two smallest circles is

$$\sqrt{\left(\frac{D-d}{2}\right)^2 - \left(\frac{d}{2}\right)^2}.$$

But the distance from the center of the third inside circle to the point of tangency of the two smallest circles is

$$\sqrt{(d+2.5)^2 - \left(\frac{d}{2}\right)^2}.$$

The distance between the centers of the outside circle and the third inside circle is precisely the difference in their radii. We therefore have

$$\begin{aligned}(D-d) - 5 &= 2\sqrt{(d+2.5)^2 - \left(\frac{d}{2}\right)^2} - 2\sqrt{\left(\frac{D-d}{2}\right)^2 - \left(\frac{d}{2}\right)^2} \\ &= \sqrt{d^2 + 20d + 25} - \sqrt{(D-d)^2 - d^2}.\end{aligned}$$

If we square both sides of this equation and cancel wherever possible, we find

$$d^2 + 5d + 5D = \sqrt{(3d^2 + 20d + 25)(D^2 - 2dD)}.$$

Squaring again, gathering like terms, and using the equation already derived for D^2 brings about the desired result.

Exercise 11.7. In the case of a circle, as in the case of Cavalieri's principle, the surface whose area is being found lies in a two-dimensional space. In general one can have some confidence in intuitive arguments of this type when the "ambient" space is of the same dimension as the figures being considered. There is, however, a qualitative difference between curves and surfaces. Curves can be rectified by inscribed polygons, regardless of the dimension of the ambient space. Surfaces in 3-space, however, cannot always be approximated by inscribed triangles.

Exercise 11.12. Although some of the geometric examples we have looked at in the exercises to this chapter and the one preceding may be of practical importance, the majority of those leading to higher-degree equations are not. The geometry seems to be included as decoration for the algebraic problem to be solved. In nearly every case one is trying to compute quantities that could be more conveniently measured directly, such as the radius of a fort or the height at which a bamboo shoot was broken. Even worse, many of the problems, such as the one involving adding the square root of a side of a triangle to the area of the triangle, are nonsensical from a practical point of view. It is likely that among a Japanese social elite knowledge of mathematical methods was a mark of refinement, just as knowledge of the plays of Shakespeare is in modern America.

Exercise 12.1. Let x be the amount of money from the debt that is to be included in the estate. The estate therefore consists of $x + 10$ dirhems. The friend is to receive $\frac{1}{5}x + 2$ of the estate; when that amount is subtracted, $\frac{4}{5}x + 8$ dirhems remain. After the extra dirhem is given to the friend, the estate consists of $\frac{4}{5}x + 7$ dirhems. Half of this, or $\frac{2}{5}x + 3\frac{1}{2}$ dirhems is the share of each son. Since the indebted son is not to receive or owe any money, this amount must equal the portion of the estate coming out of his debt, so that $\frac{2}{5}x + 3\frac{1}{2} = x$. Transposing the $\frac{2}{5}x$, we obtain $3\frac{1}{2} = \frac{3}{5}x$. When both sides are multiplied by $\frac{5}{3}$ (Al-Khwarizmi says, "each side is increased by $\frac{2}{3}$ of itself"), we find $x = \frac{5}{3}\frac{7}{2} = \frac{35}{6}$.

In current American law (probably), the estate would consist of 20 dirhems. The friend would receive 5 dirhems, and each son would be entitled to 7.5 dirhems. The son who had borrowed 10 dirhems would therefore owe the estate 2.5 dirhems. Hence the 10 dirhems cash on hand would probably be divided in the proportion of $3 : 2$ between the other son and the friend, (6 for the son, 4 for the friend). As the debt was repaid, the son would get 1.5 dirhems, and the friend 1 dirhem. Note that in Al-Khwarizmi's solution, the son gets a total of $5\frac{5}{6}$ dirhems, so he receives *more* money "up front" in the modern solution and still more when the debt is repaid. The friend would receive $4\frac{1}{6}$ dirhems under Al-Khwarizmi's solution, so he also receives more money eventually under the modern solution, although he gets less immediately. The big loser in the modern solution is the indebted son, who has to come up with 2.5 dirhems, whereas he was debt-free under Al-Khwarizmi's solution.

Exercise 12.5. The given pair cannot be constructed from the formula, since $n = 3$ already gives a number larger than 1210. It is easier to look at the known pairs

of amicable numbers and see which of them fit the formula than to test all three numbers in the formula for primality. After $n = 2$ and $n = 4$, the next n that actually yields an amicable pair is $n = 7$, which gives the pair $9,363,584$ and $9,437,056$ (rediscovered by Descartes in the seventeenth century).

Exercise 12.8. The suggested argument shows that the single equation that results when the two equations are combined must be divisible by the minimal polynomial for the roots over the rational numbers. Suppose a cubic polynomial with rational coefficients is not the minimal polynomial of one of its roots. It is then divisible by the minimal polynomial of that root. Either the minimal polynomial or the quotient obtained by dividing the cubic by the minimal polynomial is a rational polynomial of degree 1. Hence one root of the cubic must be rational.

Exercise 12.9. The argument is fairly well sketched in the statement itself. Since at least one of the pair of opposite wheels is not moving along a straight line, whichever one it is must be subject so some unbalanced force.

Exercise 12.14. The primary immediate value is the creation of beautiful mathematical structures. These structures, it turns out, can later be used to analyze problems of practical importance in physics. (Both the solution of an equation and the structure of crystals depend on an analysis of symmetries.)

The restricted scope of exact methods in comparison with numerical methods is thus more than compensated for by an enhanced understanding of the underlying mathematical reality.

Exercise 13.2. The two given values of the ratios give $d_1 = 3h$ and $d_2 = 4h$, it follows that $d_2 - d_1 = h$. The similarity with the method used in China and India is very strong. One does not measure any angles. The elimination of this labor is paid for by the need to make two measurements of distance. However, the explicit use of the ratio of height to distance is a slight deviation from the double-difference method.

Exercise 13.4. Taking the Pythagorean triple $5^2 + 12^2 = 13^2$, we form the product $41 \cdot 169 = 6929$. This number is necessarily the sum of two squares, namely $23^2 + 80^2$. Hence we also have $41 = \left(\frac{23}{13}\right)^2 + \left(\frac{80}{13}\right)^2$.

The principle on which this method is based is the identity

$$(a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2.$$

Hence if $c^2 + d^2 = g^2$, we have

$$a^2 + b^2 = \left(\frac{ac + bd}{g}\right)^2 + \left(\frac{ad - bc}{g}\right)^2.$$

Exercise 13.11. Starting from velocity 0, the velocity after 4 seconds will be $4 \cdot 9.8 = 39.2$ meters per second. This will be the average velocity for the entire 8 seconds, and so the body will fall $8 \cdot 39.2 = 313.6$ meters.

Exercise 13.14. The *general* level of education in the Middle Ages was very low, even among the nobility, few of whom could read or write. However, among those

who did have this ability or who had studied at the cathedral schools, the general picture of the world was not so simple as these stories suggest. True, medieval maps of the world seem terribly unrealistic nowadays; yet, especially after the twelfth century, when Aristotle came to be used as a standard authority, the picture of the universe held by the educated was as realistic as in Hellenistic times. In particular, the author of the *Practica geometriae* certainly knows that the sphere of the stars is incomparably larger than the earth.

Exercise 14.1. The reader will probably *not* care to preserve a pedantic accuracy in this problem. The quartic equation to be solved is

$$44.442955568025 \times 10^{-8} a^4 - 0.833336125a^2 + 31,250 = 0.$$

This equation suggests we work instead with $u = a/100$, which satisfies

$$44.442955568025u^4 - 8333.36125u^2 + 31,250 = 0.$$

With sufficient accuracy, this equation can be rewritten, dividing out the leading coefficient, as

$$u^4 - 187.5u^2 + 703.149 = 0.$$

The quadratic formula then gives

$$u^2 = 93.75 - \sqrt{8789.0625 - 703.149} = 93.75 - 89.92 = 3.83.$$

Thus $u = \sqrt{3.83} = 1.95$, and so $a = 195$, approximately.

Exercise 14.4. Note: $\cos 41^\circ = 0.75470958 = \sin 49^\circ$, $\sin 41^\circ = 0.656059029 = \cos 49^\circ$. The angle at the North Pole whose sides are the longitudinal lines of New York and Paris is 76°, and its cosine is 0.241921895. Hence the cosine of the arc from New York to Paris is

$$\cos a = (0.75470958)(0.656059029)(1.241921895) = 0.614917798,$$

and so $c = 52.05405237^\circ$. The distance is therefore about 3592 miles.

Exercise 14.6. We want two numbers whose difference is 992 and whose product is $\left(\frac{60}{3}\right)^3 = 8000$. Following the ancient method from Mesopotamia (see Chapter 3), we know that the average of the two numbers is the square root of the sum of their product and the square of half their difference, that is, the sum is $\sqrt{8000 + (496)^2} = 504$. the two numbers are therefore $504 + 496 = 1000$ and $504 - 496 = 8$. Hence the solution is $\sqrt[3]{1000} - \sqrt[3]{8} = 10 - 2 = 8$. It is easily verified that this solution is correct.

Exercise 14.13. When the discriminant is zero, the formula produces the root

$$y = -2\sqrt[3]{\frac{q}{2}}$$

and since $y^3 + py + q = \left(y + 2\sqrt[3]{q/2}\right)\left(y - \sqrt[3]{q/2}\right)^2$ (because $p = -3\sqrt[3]{q^2/4}$), it follows that the formula picks out the single root $-2\sqrt[3]{q/2}$ rather than the double root $\sqrt[3]{q/2}$.

Exercise 14.17. The three numbers are a, ar, and ar^2, where $r > 1$, $a > 0$. In terms of the x and y mentioned above, $a = x$, $ar = \sqrt{xy}$, and $ar^2 = y$. Given that $\sqrt{xy}$ is known, we have only to square it, and we know the difference and product of y and x. Hence finding them involves only our now-fast friend the quadratic equation.

Observe that a modern student would probably proceed in a slightly different way, somewhat as follows. We wish to find a and r given that we know ar and $ar^2 - a = a(r^2 - 1)$. Hence let $c = ar$, $d = a(r^2 - 1)$. We then have the quadratic equation

$$c(r^2 - 1) = dr,$$

that is, $cr^2 - dr - c = 0$, so that $r = (d + \sqrt{d^2 + 4c^2})/2c$ and $a = c/r = (-d + \sqrt{d^2 + 4c^2})/2$. Just to verify that this is right, suppose $c = 9$ and $d = 12$. Then $a = 3\sqrt{13} - 6$, $ar = 9$, and $ar^2 = 3\sqrt{13} + 6$.

Exercise 14.24. As we have seen, in solving quadratic equations, radicals leading to irrational or complex numbers arise only when the solutions are themselves irrational or complex numbers. If there is resistance to the idea of the square root of a negative number (and there was), one can argue that the formula has failed because no solution exists.

In the case of the cubic equation, however, the formula requires complex numbers even when real solutions indubitably exist. The apparent breakdown of the formula therefore requires some explanation. The result of trying to explain it was, ultimately, the creation of the subject of complex analysis.

Exercise 14.26. The connection, which was mentioned earlier as part of the work of Roger Cotes (Chapter 15), is the formula

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

which implies

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2}, \quad \cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2},$$

so that $\sin\alpha\cos\beta = \frac{1}{2}[\sin(\alpha + \beta) + \sin(\alpha - \beta)]$. This last identity is equivalent to the fact that $e^{\alpha+\beta} = e^{\alpha}e^{\beta}$.

Thus, to take our previous example, if we wished to find the product 9753×78642 using logarithms, we would find α and β such that $e^{\alpha} = 9753$ and $e^{\beta} = 78642$. These values would be approximately 9.185330209 and 11.27266119. We would then take their sum (20.4579914) and find its exponential, which would be the product of the original numbers:

$$e^{20.4579914} = 766995425.9.$$

Since the trigonometric functions were discovered before the exponential functions, this convenient property was first noticed in the language of trigonometry. Once again, it is a question of the same message being carried by two different media.

Exercise 15.2. The ellipse has the equation $x^2 = 4y - 4y^2$. We shall take the reference axis to be the y-axis. The equation of a circle centered at the point $(0, C)$

and passing through the point $\left(1, \frac{1}{2}\right)$ is $x^2 + (y-C)^2 = 1 + \left(\frac{1}{2} - C\right)^2$. This leads to the equation

$$4y - 4y^2 + (y-C)^2 = 1 + \left(\tfrac{1}{2} - C\right)^2.$$

That is,

$$3y^2 + 2(C-2)y - \left(C - \tfrac{5}{4}\right) = 0.$$

This equation has a single root when its discriminant is 0, that is, when

$$(C-2)^2 + 3\left(C - \tfrac{5}{4}\right) = 0.$$

This quadratic equation says

$$C^2 - C + \tfrac{1}{4} = 0,$$

and this relation holds if and only if $C = \frac{1}{2}$. The normal is therefore the line $y = \frac{1}{2}$.

Exercise 15.4. Suppose the tangent intersects the x axis at $(T, 0)$. The tangent then has slope $1/[(3\sqrt{3}/2) - T]$. Between two nearby points (x, y) and $(x+h, y+k)$ on the curve the secant has slope k/h. Hence we need $\left(\frac{3\sqrt{3}}{2} - T\right)k = h$. When we subtract the coordinates of the two neighboring points, we find

$$\frac{2xh + h^2}{9} + \frac{2yk + k^2}{4} = 0,$$

which yields $8xh + 4h^2 = -18yk - 9k^2$. Since $y = 1$ and $x = \frac{3\sqrt{3}}{2}$, we find, neglecting $4h^2$ and $9k^2$, that $12\sqrt{3}h = -18k$, that is, $h = -\frac{\sqrt{3}}{2}k$. Comparing this result with the previously obtained relation between h and k, we see that $\frac{3\sqrt{3}}{2} - T = -\frac{\sqrt{3}}{2}$, that is, $T = 2\sqrt{3}$.

Exercise 15.8. At the point (x_0, y_0), where $y_0 = f(x_0)$, the equation of the tangent line is

$$y - f(x_0) = f'(x_0)(x - x_0).$$

Setting $x = 0$ here, we find $y = f(x_0) - x_0 f'(x_0)$. Suppressing the subscripts now gives a new function of x defined in terms of f. The area under this new curve from $x = 0$ to $x = a$ is

$$\int_0^a f(x) - xf'(x)\,dx = \int_0^a f(x)\,dx - \int_0^a xf'(x)\,dx.$$

In the second integral we integrate by parts, taking $u = x$, $dv = f'(x)\,dx$, so that $du = dx$ and $v = f(x)$. The second integral then becomes

$$xf(x)\Big|_{x=0}^{x=a} - \int_0^a f(x)\,dx.$$

Hence the area under the new curve is

$$2\int_0^a f(x)\,dx - af(a) = 2\left[\int_0^a f(x)\,dx - \frac{1}{2}af(a)\right].$$

Since the expression $\frac{1}{2}af(a)$ represents the area under the line $ay = f(a)x$ between $x = 0$ and $x = a$, we are done.

Exercise 15.11. If we formulate "equality after an infinite time" so as to avoid the actually infinite, we find ourselves saying that for any prescribed difference there is a finite time after which the quantities will differ by less than that amount. In that respect, Newton's "proposition" becomes a mere tautology. It says that quantities that become arbitrarily close to each other in a finite time must come closer than any prescribed difference in some finite time. Thus Newton's solution of the difficulty of "indivisibles" is, like Eudoxus' solution of the difficulty of incommensurables, an attempt to make a definition that fits intuition. Newton's attempt to turn his definition into a theorem resembles Euclid's attempt to define the term *point*.

Exercise 15.13. We have mentioned previously (Exercise 6.20) that Cavalieri's principle fails for curves in the plane unless one defines the zero-dimensional volume of the intersection of two curves as the cosecant of their angle of intersection. If this definition is made, horizontal lines intersect the diagonal in a point of zero-dimensional volume $\csc 45° = \sqrt{2}$, and they intersect the staircase either in a point of zero-dimensional volume 1 or (for a finite number of lines only) in a line segment whose zero-dimensional volume is infinite. We can either argue that the principle is not valid because of these exceptional lines, or else argue that the exceptions are negligible because there are so few of them. In the latter case the ratio of the zero-dimensional volumes is $\sqrt{2} : 1$, and so the diagonal has one-dimensional measure (length) equal to $\sqrt{2}$ times the length of the side, as required.

In general, however, length, area, and volume have to be carefully defined in order to avoid paradoxes. We have come too close to disaster using this principle to be entirely confident of its validity.

Exercise 16.1. The work of proving this result is mostly contained in a lemma that asserts that chords on a circle are parallel if and only if the two arcs between them are equal. That fact, in turn is proved by drawing the line joining alternate endpoints of the two chords. The alternate interior angles formed by this transversal are inscribed in the two arcs, and are equal if and only if the chords are parallel (see Fig. A.2). Hence if we imagine a hexagon $ABCDEF$ inscribed in a circle so that sides AB and DE are parallel, we conclude that arcs $\overset{\frown}{BD}$ and $\overset{\frown}{EA}$ are equal. If in addition sides BC and EF are parallel, we conclude that arcs $\overset{\frown}{CE}$ and $\overset{\frown}{FB}$ are equal. By subtracting, we conclude that $\overset{\frown}{ED} - \overset{\frown}{BC} = \overset{\frown}{AB} - \overset{\frown}{EF}$. When we transpose the two negative terms to the opposite side, we find that arcs $\overset{\frown}{DF}$ and $\overset{\frown}{AC}$ are equal, and this says precisely that CD is parallel to FA.

Exercise 16.2. The cheapest way to get this result is to use analytic geometry. The four lines can be thought of as having the equations $y = a_i x$, $i = 1, 2, 3, 4$ (see Fig. A.3). A line that intersects all of them has equation $y = mx + b$, where $m \notin \{a_1, a_2, a_3, a_4\}$. Solving this last equation simultaneously with each of the

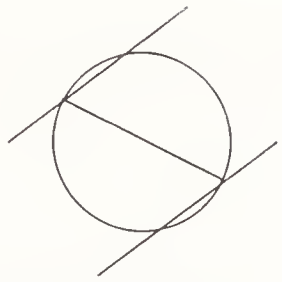


Figure A.2: Parallel chords on a circle.

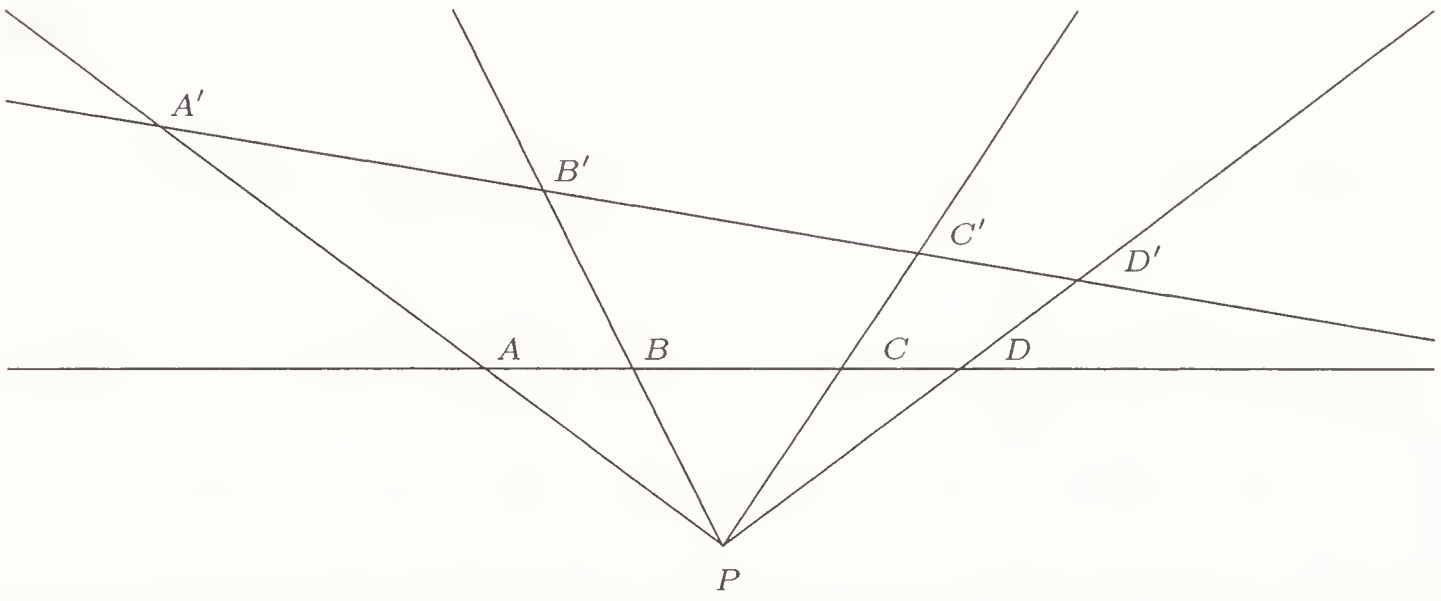


Figure A.3: Cross ratio of four lines cut by two transversals.

first four gives the four points of intersection as

$$\left(\frac{b}{a_i - m}, \frac{a_i b}{a_i - m}\right).$$

The standard distance formula in the plane then shows that the cross ratio is

$$\frac{AC \cdot BD}{BC \cdot AD} = \frac{|a_1 - a_3|\,|a_2 - a_4|}{|a_2 - a_c|\,|a_1 - a_4|},$$

which is independent of m and b. Thus the cross ratio of four concurrent lines depends only on those lines. Putting this fact another way, the cross ratio of four points on a line remains the same if projected to another line from a point.

Exercise 16.6. The assumption $m^3 = 3n^3$ implies $m^3 - n^3 = 2n^3$, that is, $(m-n)(m^2 + mn + n^2) = 2n^3$. Now if p is a prime factor of n, it must divide either $m - n$ or $m^2 + mn + n^2$. The first case implies that it divides $m = n + (m - n)$; the second case implies that it divides m^2, and hence also divides m. It follows that $n = 1$, and so the *integer* m satisfies $m^3 = 3$, which is absurd.

Exercise 16.7. The assumed equation implies that $y^3 = z^3 - x^3 = (z - x)(z^2 + zx + x^2)$. Let a be any prime divisor of $z - x$. Then $a \neq 3$ and a does not divide zx. (If a divided zx, it would also divide either z or x, and hence, since it divides $z - x$, would divide both z and x .) Then a does not divide $z^2 + zx + x^2$, since if it did, it would also divide $3xz = z^2 + zx + x^2 - (x - z)^2$. Since a^3 divides

y^3, it follows that a^3 must divide $x - z$. In this way we see that $z - x = p^3$. The proofs that $z - y = q^3$ and $x + y = r^3$ are similar.

We now have

$$\begin{aligned} x &= \frac{r^3 - (p^3 - q^3)}{2}, \\ y &= \frac{r^3 + (p^3 - q^3)}{2}, \\ z &= \frac{r^3 + (p^3 + q^3)}{2}, \end{aligned}$$

and hence, when the equation $x^3 + y^3 = z^3$ is multiplied by 8, we obtain

$$2r^9 + 6r^3(p^3 - q^3) = r^9 + 3r^6(p^3 + q^3) + 3r^3(p^3 + q^3)^2 + (p^3 + q^3)^3,$$

which easily converts into

$$r^9 - 3r^6(p^3 + q^3) + 3r^3(p^3 + q^3)^2 - (p^3 + q^3)^3 = 24p^3q^3r^3,$$

that is, $m^3 = 3n^3$, where $m = r^3 - (p^3 + q^3)$ and $n = 2pqr$. Since the only solution of this last equation is $m = n = 0$, the assertion follows.

Exercise 16.9. The value of any given shot is the same for both players at any time during the game. Although it might seem that the opponent would be entitled to *more* money as a reward for having undergone the risk of losing everything on the first three shots, that player is now betting the same stake on a game in which the chances of winning are much better than when the stakes were actually laid down. The "reward" for withstanding the risk is an improved chance of winning. If the stakes had been divided before the game was played, the share given to the shooter's opponent would have been even smaller. Hence the desire of a rational opponent to see the game continue will be stronger after the shooter has made three unsuccessful shots than before those shots were made. Put another way, "buying" the last five shots after three have been made assures the buyer of winning the remaining portion of the stake, whereas the entire stake remains at risk if those shots are bought before the game starts.

Exercise 17.2. Because of the solution of the previous problem, it is easy to see that the point P corresponding to any point Q on the logarithmic spiral can be constructed very simply, by drawing the ray from the origin to Q. At the point R where the ray intersects the unit circle $r = 1$, draw a tangent to the unit circle of length RQ $(= r - 1)$. The point P will be the end of that tangent. The reason is that the arc length from $(1, 0)$ to Q is $r - 1$, as shown in the previous problem, while the tangent to the spiral at each point makes an angle of $45°$ with the radius vector to that point, that is $(x\,dx + y\,dy)^2 = (x^2 + y^2)((dx)^2 + (dy)^2)\cos^2 45°$. This last relation is easily proved by taking $x = e^\theta \cos\theta$ and $y = e^\theta \sin\theta$, so that $dx = e^\theta(\cos\theta - \sin\theta)$ and $dy = e^\theta(\sin\theta + \cos\theta)$. It is then easy to compute that both sides equal $e^{4\theta}(d\theta)^2$.

This geometric construction of P makes it possible to find the equation of the involute. The relations among the parts of the triangle OQP show that if $P = (\rho\cos\varphi, \rho\sin\varphi)$ and $Q = (r\cos\theta, r\sin\theta)$, then

$$1 + (r-1)^2 = \rho^2$$

and

$$\tan(\theta - \varphi) = r - 1 = \sqrt{\rho^2 - 1}$$

so that

$$\begin{aligned}\varphi &= \theta - \arctan\left(\sqrt{\rho^2-1}\right)\\ &= \ln r - \arctan\left(\sqrt{\rho^2-1}\right)\\ &= \ln\left(1+\sqrt{\rho^2-1}\right) - \arctan\left(\sqrt{\rho^2-1}\right).\end{aligned}$$

Exercise 17.10. The expected payoff after playing $2^{10,000}$ games is larger than

$$2^{9999}\cdot 2 + 2^{9998}\cdot 4 + \cdots + 2^2\cdot 2^{9998} + 2\cdot 2^{9999} + 2^{10,000} = 10,000\cdot 2^{10,000},$$

which is the total amount of money paid for playing the games. The reasoning here is that half of the games (2^{9999}) can be expected to end after one flip of the coin, paying off \$2, one-fourth (2^{9998}) can be expected to end after two flips of the coin, paying off \$4, etc. The sum above thus accounts for $2^{9999} + 2^{9998} + \cdots + 2 + 1 = 2^{10000} - 1$ games. On the remaining game, you will win something, so that it is reasonable to expect a profit after playing 2^{10000} games. Since $2^{10} > 10^3$, however, this number represents over 10^{3000} games, and this is larger than the number of games than would have been played by now, even if every atom in the known universe had played a thousand games per second since the universe began. Moreover, you are likely to win big or lose big, depending on the number of games taking more or fewer flips to decide.

To cut the paradox down to more manageable size, the risk of playing 1024 games at \$10 per game (which would require a stake of \$10,240) would not be large. The fact that makes the risk in general unacceptable is that breaking even depends on the occurrence of rather rare events, such as winning \$1,024 dollars at least once in 1,024 attempts. The probability of this win on each attempt is indeed $\frac{1}{1,024}$, but the probability that it will occur at least once in 1,024 attempts is less than 64%. (The probability that this big win will not occur is about e^{-1}, which is 0.367879441, for any large number of games.) In the nearly 37% of tries in which this big win does not occur, it must be compensated for by *more than* two excess wins of \$512 (since the actual profit in two wins of \$512 is only \$1004, which is \$10 less than the profit in one win of \$1024), or more than four excess wins of \$256 (the actual profit in four such wins is only \$984), or some combination of these and smaller wins. In fact, playing at \$10 per game, you could expect to lose money on 898 of the 1,024 games and win only 126 of them.

Exercise 17.14. The equation that results from these values of p, q, and r is

$$y'' + \lambda y = 0,$$

whose general solution is $y(x) = A\cosh\mu x + B\sinh\mu x$, when $\lambda < 0$, $y(x) = Ax + B$ if $\lambda = 0$, and $y(x) = A\cos\mu x + B\sin\mu x$ if $\lambda > 0$. Here $\mu = \sqrt{|\lambda|}$. Because of the boundary conditions that $y(0) = y(2\pi)$ and $y'(0) = y'(2\pi)$, the first two of these are impossible, and λ must be the square of an integer.

Exercise 17.22. The first problem is easily done: $4 = (2+0\sqrt{-3})(2+0\sqrt{-3}) = (1+\sqrt{-3})(1-\sqrt{-3})$. One can easily show by consideration of cases that the factors here are all irreducible, that is, the only divisors of 2 and $1 \pm \sqrt{-3}$ are units (± 1 or $\pm\sqrt{-1}$).

Factorization is unique for numbers of the form $m+n\sqrt{-2}$. This is because a division algorithm exists for these numbers analogous to the Euclidean algorithm. That is, given $m + n\sqrt{-2}$ and $p + q\sqrt{-2}$, there exist $r + s\sqrt{-2}$ and $t + u\sqrt{-2}$ such that

$$m + n\sqrt{-2} = (p + q\sqrt{-2})(r + s\sqrt{-2}) + (t + u\sqrt{-2}),$$

and $t^2 + 2u^2 < p^2 + 2q^2$. [The quantity $N(m+n\sqrt{-2}) = m^2 + 2n^2$ is analogous to absolute value—indeed $N(m + n\sqrt{-2})$ *is* the squared absolute value of the complex number $m + n\sqrt{-2}$.] This division algorithm with a remainder smaller than the divisor allows the Euclidean algorithm to proceed and find a greatest common divisor for each pair of numbers. To prove that the division algorithm exists, let r be the integer nearest to the rational number $(mp + 2nq)/(p^2 + 2q^2)$ and s the integer nearest to $(np - mq)/(p^2 + 2q^2)$. (These rational fractions are the real and imaginary parts x and y of the exact quotient $x + y\sqrt{-2}$ of the two numbers, regarded as complex numbers.) Then let

$$\begin{aligned} t &= m - (rp - 2qs) \\ u &= n - (rq + ps). \end{aligned}$$

Thus

$$\begin{aligned} r &= \frac{mp + 2nq}{p^2 + 2q^2} + \varepsilon, \\ s &= \frac{np - mq}{p^2 + 2q^2} + \eta, \end{aligned}$$

where $|\varepsilon| \le \frac{1}{2}$ and $|\eta| \le \frac{1}{2}$. It then follows that

$$\begin{aligned} t &= m - p\left(\frac{mp + 2nq}{p^2 + 2q^2}\right) - p\varepsilon + 2q\left(\frac{np - mq}{p^2 + 2q^2}\right) + 2q\eta, \\ &= -p\varepsilon + 2q\eta, \\ u &= n - q\left(\frac{mp + 2nq}{p^2 + 2q^2}\right) - q\varepsilon - p\left(\frac{np - mq}{p^2 + 2q^2}\right) - p\eta, \\ &= -q\varepsilon - p\eta. \end{aligned}$$

From this we can easily compute that

$$t^2 + 2u^2 = (\varepsilon^2 + 2\eta^2)(p^2 + 2q^2) \le \left(\tfrac{3}{4}\right)(p^2 + 2q^2).$$

The existence of a division algorithm of this sort is the fundamental element in the proof that any positive integer is a unique product of primes. The standard proof of that fact, which is based on the existence of a greatest common divisor of any two elements, can be found in any elementary book on number theory.

Exercise 17.23. Given any points (x_i, y_i), $i = 1, \ldots, 8$, and unspecified real numbers x and y, the system of 10 inhomogeneous equations in 8 unknowns

$$\begin{aligned}
t_1 + t_2 + \cdots + t_8 &= 1 \\
t_1x_1 + t_2x_2 + \cdots + t_8x_8 &= x \\
t_1y_1 + t_2y_2 + \cdots + t_8y_8 &= y \\
t_1x_1y_1 + t_2x_2y_2 + \cdots + t_8x_8y_8 &= xy \\
t_1x_1^2 + t_2x_2^2 + \cdots + t_8x_8^2 &= x^2 \\
t_1y_1^2 + t_2y_2^2 + \cdots + t_8y_8^2 &= y^2 \\
t_1x_1^2y_1 + t_2x_2^2y_2 + \cdots + t_8x_8^2y_8 &= x^2y \\
t_1x_1y_1^2 + t_2x_2y_2^2 + \cdots + t_8x_8y_8^2 &= xy^2 \\
t_1x_1^3 + t_2x_2^3 + \cdots + t_8x_8^3 &= x^3 \\
t_1y_1^3 + t_2y_2^3 + \cdots + t_8y_8^3 &= y^3
\end{aligned}$$

can be solved under certain conditions. Gaussian elimination will provide (in general) two solvability conditions in the form of linear combinations of the right-hand sides that must vanish. These two conditions are cubic equations in x and y. These conditions will certainly be satisfied by (x_i, y_i) (since with these values of x and y the equations hold with $t_i = 1$, $t_j = 0$ for $j \neq i$). At this point we have derived a constructive procedure for determining (generally) two independent cubic equations satisfied by the given 8 points.

The procedure tells us something more, however. It shows us why there is no contradiction in the fact that more than 8 points may satisfy the two given cubics. For, given any additional point (x, y) satisfying these two cubics, we see that all the powers of x and y that occur when these values are substituted into the cubic are linear combinations of the corresponding powers of the original 8 points, the same linear combination for each power.

Exercise 18.3. The tangential path of sparks is due to the *sudden* cessation of the adhesive force that held the hot particles to the wheel or to the object being sharpened. Gravity does not cease abruptly when an object rises above the earth's surface.

If the earth's rotation speeded up and the adhesive forces that hold it together continued to operate, a person standing on the equator weighing nothing would effectively be in orbit at an altitude of 0 kilometers above the earth. If the rotation speeded up still further and the person stayed above the same point on the surface, the orbit would enlarge its radius, that is, the person would rise higher above that point.

Exercise 18.4. The difference between a sphere and a hoop lies in their moments of inertia about their axes. For a sphere of radius R and mass M, this moment is

$I_s = \frac{2}{5}MR^2$. For a hoop, regarded as a circle of the same mass and radius, it is $I_c = MR^2$. Each of these bodies, starting from rest and rolling down an incline that descends by a distance h, will acquire kinetic energy $E = Mgh$. The center of each will be moving forward with velocity v and each will be spinning with angular velocity ω related by

$$\omega = \frac{v}{R}.$$

(To see this, note that the time required for the body to spin once is $2\pi/\omega$, and this is precisely the time required for the center to advance by $2\pi R$, that is, $2\pi/\omega = 2\pi R/v$.) The velocity v is different for the two bodies, however, because the kinetic energy acquired is the sum of the translational and rotational energy, that is,

$$Mgh = \frac{1}{2}Mv^2 + \frac{1}{2}I\omega^2 = \frac{1}{2}\Big(M + \frac{I}{R^2}\Big)v^2.$$

This last relation shows that v is independent of R (since I/R^2 is independent of R) and also of M. It does, however, depend on the shape of the object. In fact, we find that

$$v = \sqrt{\frac{10}{7}gh}$$

for the sphere and

$$v = \sqrt{gh}$$

for the hoop. Thus a sphere will require less time to roll down the incline than a hoop. If Galileo actually performed this experiment, it was a lucky circumstance that he chose to use only spheres and not solid cylinders or cylindrical hoops.

Exercise 19.3. If C were an element of some class, it would be a set and therefore, according to the rules for set formation, it would be an element of itself if and only if it were *not* an element of itself. Since we cannot allow this situation to occur, we must forbid C to be an element of any set.

Exercise 19.5. The proof is not intuitionistically valid, since it asserts that "p or q" is true without proving either that p is true or that q is true. (In the intuitionistic propositional calculus, if $p \vee q$ is a theorem, then either p is a theorem or q is a theorem.)

Exercise 19.6. Consider the numbers $G_m = F_m - 1 = 2^{2^m}$, and observe that $G_{m+1} = G_m^2$. Hence $G_{m+k} = G_m^{2^k}$. This equality asserts that

$$F_{m+k} = 1 + (F_m - 1)^{2^k},$$

from which it follows (by the binomial theorem) that

$$F_{m+k} = F_m^{2^k} - 2^k F_m^{2^k - 1} + \cdots - 2^k F_m + 2 = QF_m + 2,$$

That is, each Fermat number, when divided by a smaller Fermat number, leaves a remainder of 2. Thus the only possible common divisors of two Fermat numbers are 1 and 2. Since all Fermat numbers are odd, 2 cannot be a common divisor.

Hence if we take all the prime divisors of Fermat numbers, we must obtain an infinite set of primes. This is short of exhibiting an algebraic formula that always generates a prime, but it does give an *algorithm* that always generates a prime. The algorithm proceeds as follows. Form the number F_m. Then divide F_m by each positive integer, starting with 3, until the first integer is reached at which the remainder is zero. Let that integer, which is necessarily prime and necessarily less than or equal to F_m, be p_m. Increment m and continue.

This algorithm ought to satisfy an intuitionist, who should confess that the primes are at least potentially infinite.

Exercise 19.7. The words *faith* and *should* are slippery ones. The agnostic position is always available to both scientists and mathematicians: it is possible to *explore the consequences* of a proposition without *affirming* the proposition. This position is not available in other areas, and it contradicts the meaning of the word *faith.* Mathematicians who use set theory, for example, can state that they are using it only to derive theorems and make no claim as to its consistency. In that respect a mathematician need not assert that we *should* draw any conclusions at all from our proofs, other than the hypothetical conclusion that "if all our assumptions are true (and hence consistent with one another), then our conclusions are also true." Thus it can be argued that the word *faith* is misapplied in both mathematics and science, at least as far as pure logic is concerned.

Where logic is satisfied, however, human psychology is not. If mathematicians did not have considerable *confidence* in the consistency of set theory, they would not use it, any more than chemists and physicists would devote large amounts of time and effort seeking a reaction (cold fusion, for example) that they did not believe possible. Thus the word *faith* comes back on the psychological level. It is a rather anemic faith, however, compared with the robust affirmations required by religions. Confidence exists in various degrees, expressed as probabilities: One can bet odds of arbitrarily high levels on the correctness of the multiplication table. As to the proof of Fermat's last theorem or the four-color conjecture, however, most mathematicians would probably not give odds of more than 1000 to 1.

Index

This pragmatic, issues-oriented history traces the discovery, solution, and application of mathematical problems

From the arithmetic of the ancient Egyptians to the intricacies of postcalculus math, *The History of Mathematics: A Brief Course* focuses on how mathematics has developed over the centuries. Roger Cooke has selected the most intriguing and significant problems in the history of mathematics and asked of each one: Why was it important? How was it solved? How was its solution applied? Did its solution lead to further advances in the field?

The carefully selected topics in this book include

- The nature and origins of mathematics
- Early Western mathematics as practiced by the Egyptians, the Mesopotamians, the Greeks, and the Romans
- Non-Western traditions, including Hindu, Chinese, Korean, Japanese, and Islamic mathematics
- The development of modern mathematics from the Middle Ages to the calculus and other seventeenth-century discoveries to today's number theory
- The relationship of modern mathematics to science
- Contemporary issues in mathematics, including the role of women and minorities

This readable, up-to-date study is ideal for undergraduate courses in mathematics and mathematics education. Everyone interested in the field will want to keep a copy of *The History of Mathematics* close at hand.

ROGER COOKE is a professor in the Department of Mathematics and Statistics at the University of Vermont. For many years he has taught a general introduction to the history of mathematics.

Cover Design: Bachner + Co.

WILEY-INTERSCIENCE
John Wiley & Sons, Inc.
Professional, Reference and Trade Group
605 Third Avenue, New York, N.Y. 10158-0012
New York • Chichester • Weinheim
Brisbane • Singapore • Toronto

ISBN 0-471-18082-3
90000
9 780471 180821