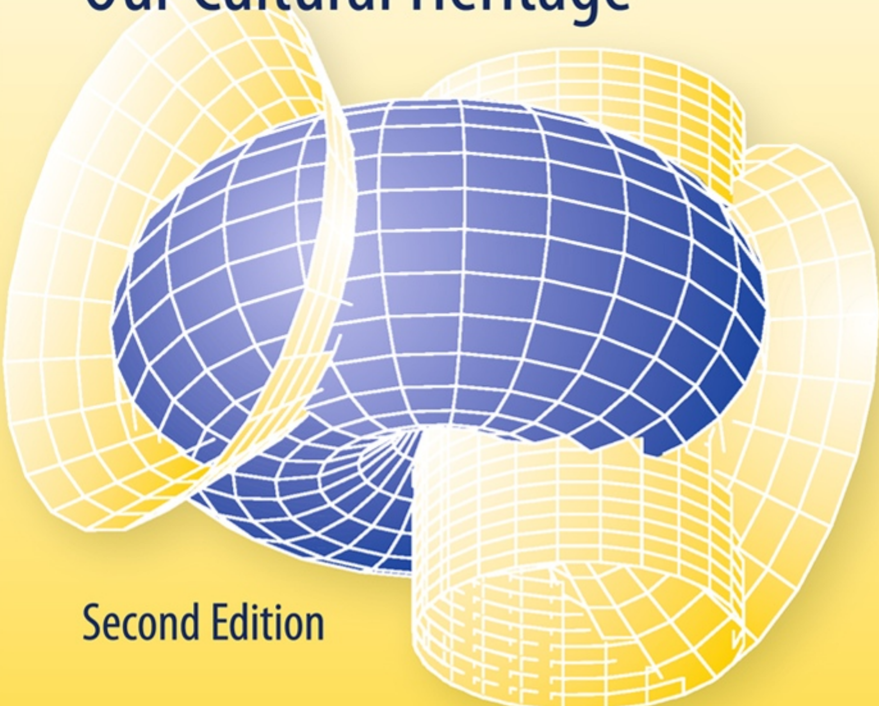
Audun Holme

Geometry

Our Cultural Heritage



Second Edition

Springer

Geometry

Audun Holme

Geometry

Our Cultural Heritage

Second Edition

Springer

Audun Holme
Department of Mathematics
University of Bergen
Johs. Brunsgt. 12
5008 Bergen
Norway
holme@mi.uib.no

ISBN 978-3-642-14440-0 e-ISBN 978-3-642-14441-7
DOI 10.1007/978-3-642-14441-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010936506

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This is a revised edition of the first printing which appeared in 2002. The book is based on lectures at the University of Bergen, Norway. Over the years these lectures have covered many different aspects and facets of the wonderful field of geometry. Consequently it has never been possible to give a full and final account of geometry as such, at an undergraduate level: A carefully considered selection has always been necessary. The present book constitutes the main central themes of these selections.

One of the groups I am aiming at, is future teachers of mathematics. All too often the texts dealing with geometry which go into the syllabus for teacher-students present the material in ways which appear pedantic and formalistic, suppressing the very powerful and dynamic character of this old field, which at the same time so young. Geometry is a field of mathematical insight, research, history and source of artistic inspiration. And not least important, an integral part of our common cultural heritage.

Another motivation is to provide an invitation to *mathematics in general*. It is an unfortunate fact that today, at a time when mathematics and knowledge of mathematics is more important than ever, the phenomena of *math avoidance* and *math anxiety* are very much present under different names all over the world. It is an important task to attempt seriously to heal these ills. Perhaps they are inflicted on students at an early age, through deficient or even harmful teaching practices. Thus the book also aims at an informed public, interested in making a new beginning in mathematics. And in doing so, learning more about this part of our cultural heritage.

The book is divided in two parts. Part I is called *A Cultural Heritage.* This section contains material which is normally not found in a mathematical text. For example, we relate some of the stories told in [28] by the Greek historian *Herodotus*. We also include some excursions into the history of geometry. These excursions do not represent an attempt at writing the history of geometry. To write an introduction to the history of geometry would be a quite different and very challenging undertaking.

To write the History of Geometry is therefore definitely not my aim with Part I of the present book. Instead, I wish to seek out the roots of the themes to be treated in Part II, *Introduction to Geometry.* These roots include not only the geometric ideas and their development, but also the historical context. Also relevant are the legends and tales – really *fairy-tales* – told about, for example, *Pythagoras*. Even if some of the more or less fantastic events in *Iamblichus'* writings are unsubstantiated, these stories very much became our *perception* of the geometry of Pythagoras, and thus became part of the *heritage of geometry*, if not of its *history*.

In Chaps. 1 and 2, we go back to the beginnings of science. As geometry represents one of the two oldest fields of mathematics, we find it in evidence from the early beginnings. The other field being *Number Theory*, they go back as far as written records exist. Moreover, in the first written accounts from ancient civilizations they present themselves as already well developed and sophisticated disciplines.

Thus we find that problems which ancient mathematicians thought about several thousand years ago, in many cases are the same problems which are difficult to handle for the students of today. As we move on in Chaps. 3 and 4, we find that great minds like Archimedes, Pythagoras, Euclid and many others should be allowed to speak to the people of today, young and old. They are unsurpassable tutors.

The mathematical insight which Archimedes regarded as his most profound theorem, was a theorem on geometry which was inscribed on Archimedes' tombstone. All of us, from college student to established mathematician, must feel humbled by it. What does it say? Simply that if a sphere is inscribed in a cylinder, then the proportion of the volume of the cylinder to that of the sphere, is equal to the proportion of the corresponding surface areas, counting of course top and bottom of the cylinder. The common proportion is $3 : 2$. This is a truly remarkable achievement for someone who did not know about integration, not know about limits, not know about... Its beauty and simplicity beckons us. How did Archimedes arrive at this result? Archimedes deserves to be remembered for this, rather for the silly affair that he ran out into the street as God had created him, shouting – *Eureka, Eureka!* But the story may well be true, his absentmindedness under pressure cost him dearly in the end.

A new addition to the present revised edition is a more extensive treatment of the Archimedean polyhedra, starting in Sect. 4.4. The Archimedean polyhedra are also treated in Sect. 6.9, where the process of finding these polyhedra is also tied to finding all semi regular tessellations of the plane. It becomes evident that the tessellations and the polyhedra are intimately connected. This relies on the pioneering work of *Johannes Kepler* (1571–1630), long after Archimedes. Our study of the Archimedean polyhedra is concluded in Sect. 20.6, where the mathematics of symmetries in space is outlined.

Pythagoras and his followers certainly did not discover the so-called *Pythagorean Theorem*. The Babylonians, and before them the Sumerians, not only knew this fact very well, they also knew how to construct all *Pythagorean triples*, that is to say, all natural numbers a, b and d such that $a^2+b^2 = d^2$. This is documented by a famous babylonian clay tablet, now in the library of Columbia University in New York, it is known today as the tablet *Plimpton 322*. The tablet was originally found or otherwise obtained by an American collector and adventurer who is a probable model for the character of *Indiana Jones*.

Some have speculated whether the Babylonians used these insights to construct the equivalent of trigonometric tables. Such tables would be simple, accurate and powerful thanks to the *sexagesimal* system they used for representing numbers. But this theory has few if any followers today, for one thing the very concept of *an angle* lay far into the future at the time of the creation of Plimpton 322.

In any case it would be a challenging project for interested college students to understand the mathematics of the Plimpton 322 tablet at Columbia University, and to correct and explain the four mistakes in it. Or perhaps even to construct the successor or the predecessor of this tablet if we allow ourselves to imagine that it were indeed one of several in a series of tablets, constituting "trigonometric tables".

So what did Pythagoras discover himself? We know nothing with certainty of Pythagoras' life before he appeared on the Greek scene in midlife. Some say that he travelled to Egypt, where he was taken prisoner by the legendary, in part infamous, Persian King *Cambyses II*, who also ruled Babylon, which had been captured by his father *Cyrus II*. Pythagoras was subsequently brought to Babylon as a prisoner, but soon befriended the priests, the *Magi*, and was initiated into the priesthood in the temple of Marduk. We tell this story as related by Herodotus and Iamblichus in [28] and [37]. However, the accounts given in these classical books are not always historically correct, the reader should consult the footnotes in [28] to get a flavor of the present state of Herodotus, *The Father of History*, who by some of his critics is called *The Father of Lies.* But Herodotus is a fascinating story-teller, and the place occupied by Pythagoras today has considerably more to do with the legends told about him than with what actually happened. So with this warning, do enjoy the story.

Euclid's Elements represents a truly towering masterpiece in the development of mathematics. Its influence runs strong and clear throughout, leading to non-Euclidian geometry, Hilbert's axioms and a deeper understanding of the foundations of mathematics. The era which Euclid was such an eminent representative of, ended with the murder of another geometer: Hypatia of Alexandria.

Chapter 5 is a new addition to this revised edition. It deals with the Arabic contribution to geometry and mathematics in the epoch following the decline of European mathematics and civilization. The central themes include the House of Wisdom in Baghdad, the life and work of Al Khwarizmi, who invented and gave rise to the name of the the field of Algebra in our sense of the word. Ibn Qurra, who was one of the first real reformers of the Ptolemaic astronomy, and also did important work in mechanics. Al Battani also did important astronomical work, and used a new method in geometry which we today call *trigonometry*. Al Wafa al Buzjani did work on mensuration, and as a method for craftsmen devised constructions with a so called *rusty compass*. Al Quhi solved a problem on spheres posed by Archimedes, by intricate procedures using methods developed by Euclid, Apollonius and Archimedes. We give a modern treatment of this material, we see methods as powerful as our own advanced college calculus! Ibn Hud who was King of Saragossa and the real discoverer of the so called Ceva's Theorem. Omar al Khayyam was both mathematician an an eminent poet and philosopher. His work on algebraic equations, as continued by Sharaf al Din, significantly paved the road for the Italian geometers who respectively found or rediscovered and found methods to solve a general equations of degree three and four by radicals. He is some times referred to as the *Voltaire of*

Persia. We finally give some glimpses of the work and biography of the fascinating mathematician ant thinker *Nasir al Din al Tusi.*

The cultural and scientific contribution of Arabic civilization is a vast and important field, and in this book it has only been possible to scratch the surface.

In Chap. 6, we describe how the foundation of present day geometry was created. Elementary Geometry is tied to straight lines and *circles*. The theorems are closely tied to constructions with straightedge and compass, reflecting the postulates of Euclid. In *higher geometry* one moves on to the more general class of conic sections, as well as curves of higher degrees.

Descartes introduced – or reintroduced, depending on your point of view – algebra into the geometry. At any rate, he is credited with the invention of the *Cartesian coordinate system*, which is named after him.

In the last section of this chapter we return to the theme of Archimedean polyhedra as well as tessellations, as already mentioned.

In Chap. 7, the last chapter of Part I, we discuss the relations between geometry and the real world. The qualitative study of *catastrophes* is of a geometric nature. We explain the simplest one among *Thom's Elementary Catastrophes*, the so-called *Cusp catastrophe.* It yields an amazing insight into occurrences of *abrupt events* in the real world.

Also tied to the real world are the fractal structures in nature. Fractals are geometric objects whose *dimensions are not integers,* but which instead have a *real number* as dimension. Strange as this sounds, it is a natural outgrowth of *Felix Hausdorff's* theory of dimensions. Hausdorff was one of the pioneers of the modern transformation of geometry, referred to in his time as the *High Priest of point-set topology.* In the end, this all did not help him. He knew, being a Jew, what to expect when he was ordered to report the next morning for deportation. This was in 1942 in his home town of Bonn, Germany. Instead of doing so, Hausdorff and his wife committed suicide.

The *Geometry of fractals* shows totally new and unexpected geometric phenomena. Amazingly, what was thought of as *pathology*, as useless curiosities, may turn out to give the most precise description of the world we live in.

In Part II, *Introduction to Geometry*, we take as our starting point the axiomatic treatment of geometry flowing from Euclid.

Considering that Euclid's original system of axioms and postulates is well over 2,000 years old, we must say that it has passed the test of modern demands to rigor remarkably well. To say the least it is the precursor to modern axiomatic theories. But the original system was set on a shaky foundation by our current mathematical standards. A clarifying explanation of the foundations was provided by Hilbert.

The search for a proof of Euclid's Fifth Postulate had gone on ever since the Elements were written, but met with no success. One version of this postulate asserts that there is one and only one line parallel to a given line through a point outside it.

A plausible approach to the problem of proving the Fifth Postulate was to assume the converse, and then derive a contradiction. This approach is usually referred to as an *indirect proof.* But instead of producing a contradiction, this relentless toil ended up producing collections of theorems belonging to *alternative geometries,* to non-Euclidian geometry. This was a highly troubling development for an age in which non-Euclidian Geometry would appear as controversial as “Darwinism” appears in some circles today.

We explain non-Euclidian Geometry in Chap. 10. But first we need to do some work on foundations. We start with *Logic and Set Theory.* In fact, the Intuitive Set Theory, even as put on a firmer foundation by *Cantor*, turned out to contain contradictions. The best known is the so-called *Russell’s Paradox*, which we explain in Sect. 8.3.

Thus arouse the need for *Axiomatic set theory*, to which we give an introduction. The aim is to give a flavor of the field without going into the technical details at all.

We then explain the interplay between *axiomatic theories* and their *models* in Sects. 8.3 and 8.4. The troubling result of *Gödel* is explained, in simplified terms, showing that a *mathematical Tower of Babel* as perhaps dreamt of by Hilbert, is not possible: Any axiomatic system without contradictions among its possible consequences, will have to live with some *undecidable* statements. This means that it may happen that statements which are perfectly legal constructions within the system, are inherently undecidable: Their truth or falsehood cannot be ascertained from the system itself.

In Chap. 9, we apply these insights to axiomatic projective geometry. This is an extensive field in itself, and a complete treatment does of course, fall outside the scope of this book. But we give a basic set of axioms, to which others may be added, thus in the end culminating with a set which determines uniquely the *real, projective plane*. This is not on our agenda here. But we do give, in some detail, two important models for the basic system of axioms. The *Seven Point Plane* and the *real projective plane* $\mathbb{P}^2(\mathbb{R})$. In Sect. 9.2, we see that the simple axioms still leads to intriguing open problems. Use of powerful computers and dexterous programming have led to new insights in axiomatic projective geometry, and there are good possibilities for further research. The question the following: *Given a projective plane* $\mathcal{P}$, as defined in the first section of Chap. 9. How many points can there be on each line? It is not difficult to see that this number m is the same for every line in the geometry $\mathcal{P}$. And we also see easily by the standard theory that m can be any power of any prime number p. But no other possible value is known. This question is related to the existence of a sufficient supply of *mutually orthogonal Latin squares*, and goes back to *Leonhard Euler*.

In Chap. 10, we are ready to explain models for non-Euclidian Geometry. In the *hyperbolic plane* there are infinitely many lines parallel to a given line through a point outside it. In the *elliptic plane* there are no parallel lines: Two lines always intersect. A model for this version is provided by $\mathbb{P}^2(\mathbb{R})$.

Plane non-Euclidian geometries have, of course, their spatial versions. This is best understood by turning to some of the basic facts from *Riemannian Geometry*, which we do in Sect. 10.5.

Chapter 11 contains some much needed mathematical tools, simple but essential. We need them for constructions to be carried out in the next chapters. The reader is advised to take the moments needed to ingest this material, which may well appear somewhat dry and barren at the first encounter.

In Chap. 12, we are then able to give coordinates in the projective plane, introduce projective n-space and discuss affine and projective coordinate systems. Again, the material may appear dry, but the reader will be rewarded in Chap. 13. There we use these techniques to give the remarkably simple proof of the theorem of Desargues. We introduce *duality for* $\mathbb{P}^2(\mathbb{R})$ and start the theory of conic sections in $\mathbb{R}^2$ and $\mathbb{P}^2(\mathbb{R})$ discussing tangency, degeneracy and the familiar classification of the conic sections. Pole and Polar belong to this picture, as well as a very simple proof of a famous theorem of Pascal. Using it, we then prove the theorem of Pappus by a classical technique known as *degeneration*, or some times as the *principle of continuity*. Here we give the first, naive, definition of an algebraic curve.

In Chap. 14, we move on to study curves of degrees greater than 2. This forms the fundament for Algebraic Geometry, and gives a glimpse into an important and very rich, active and expanding mathematical field. Here we encounter the *cubic parabola*, merely a fancy name for a familiar curve, but also the enigmatic *semi-cubic parabola*, so important in modern *Catastrophe Theory*. However, as we shall see in the following Chap. 15, from a projective point of view these two kinds of affine curves are the same. This is shown at the end of Sect. 15.5. We also learn about the *Folium of Descartes*, the *Trisectrix of Maclaurin*, of *Elliptic Curves* – which are by no means ellipses – and much more. Chapter 15 concludes with Pascal's *Mysterium Hexagrammicum*, which may be obtained as a beautiful application of Pascal's Theorem: Dualizing it the Mystery of the Hexagram is revealed.

In Chap. 16, as the title says, we sharpen the Sword of Algebra. The aim is to show how one finally disposes of the three so called *Classical Problems*. They have haunted mathematicians and amateurs for two millennia. And unfortunately, still does haunt the latter. The algebra derives in large part from the heritage of Euclid, relying as it does on *Euclid's algorithm*. This mathematics also constitutes the foundation for the important field of *Galois theory* and the theory of equations and their solvability by radicals. That theme is, however, not treated in the present book.

In Chap. 17, we use this algebra for proving that the three classical problems are insoluble: Trisecting an angle with legal use of straightedge and compass, doubling the cube using straightedge and compass, and finally we see how the transcendency of the number π precludes the squaring of the circle using straightedge and compass. Gauss' towering achievement on constructibility of regular polygons conclude the chapter. The solution of this problem by Gauss transformed the answer to a geometric question into a number theoretic problem on the existence of certain primes, namely primes of the form $F_r = 2^{2^r} + 1$, the so-called *Fermat primes*. For $r = 0, 1, 2, 3, 4$ the numbers P_r are $3, 5, 17, 257$ and 65,537. They are all primes, but then no case of an r yielding a prime is known. Gauss proved that if q is a product of such primes p_r, all of them distinct, then the regular $n = 2^m q$-gon may be constructed with straightedge and compass, and that this are precisely all the constructible cases. Thus for example the regular 3-gon, the regular 5-gon and the regular 15-gon are all constructible with straightedge and compass, as is the regular 30-gon and the regular 60-gon. The first impossible case is the regular 7-con. Now Archimedes constructed the regular 7-gon, but he used means beyond legal use of straightedge and compass. In Sect. 4.4 we have given Archimedes' construction of the regular 7-gon, the regular *heptagon*, by a so-called *verging construction*. It is not possible by the *legal use* of compass and straightedge, but may be carried out by conic sections or by a curve of degree 3. In fact, such constructions were part of the motivation for passing from *elementary geometry* to *higher geometry*.

In Chap. 18 we take a closer look at the theory of fractals. We explain the computation of fractal dimensions.

Chapter 19 contains a mathematical treatment of introductory Catastrophe Theory. We explain the Cusp Catastrophe as an application of geometry on a cubic surface. For this we also explain some rudiments of Control Theory.

The final chapter is Chap. 20. Here we return to polyhedra and tessellations, and study them in light of their groups of symmetry. This also applies to the more general situation of patterns and their groups of symmetry. We start out with the important groups of symmetries in the Euclidian plane and the Euclidian 3-space. This chapter presupposes more knowledge of linear algebra and group theory than the earlier parts of the book. Good sources and references for some of the material in this section are the books [9] and [57].

Some of the historical material giving historical context has been extended, and and a large number of illustrations have been added. In revising the historical part, I have tried to follow the guideline that when an interpretation is controversial, this should be noted, and in some cases I provide the alternatives interpretations. In particular this applies to the explanation of the numbers on the Babylonian tablet Plimpton 322, where the first edition only treats the original theory of Otto Neugebauer and his collaborators. Today this is not a justified exposition, pathbreaking as this work was at the time. It is certainly true that a controversial theory should not be presented as a fact.

The main difference from the first edition, however, is the inclusion of a large number of exercises with some suggestions for solutions. Some of the exercises are simple, others more challenging.

Several historical topics, which were not treated at all in the first edition, now have been included in the form of exercises with hints or complete solutions. In Part II the exercises are more or less of the standard type which might appear on a college test. In some cases I include complete solutions, in other cases just more or less extensive hints of just the answer, while some exercises are left open without answers.

Some of the material in this book has been published in the author's [31] and [33]. The material is included here with the permission of *Fagbokforlaget*, the publisher of [31] and [33]. A large number of the illustrations are created with the marvellous system Cinderella, [47], some of them were made by Ulrich H. Kortenkamp, one of the authors of the system. Others were made by the author, who would like to take this opportunity to thank Professor Kortenkamp for his efforts in making these illustrations, as well as for his valuable advice and assistance during this work. Some illustrations are made with the aid of the Computer Algebra system MAPLE, and finally some were made by Springer's illustrator, based on sketches by the author. In addition there are a number of images from various sources which are listed after the bibliography.

Another nice treatment of the material in Chapters 15 and 16 may be found in [16]. For children and young people [35] may be suitable as introduction to our history. An interesting historical source is [41]. My interest was much inspired by [45].

It is also a great pleasure to thank Springer Verlag, in particular the Mathematics Editor Dr. Martin Peters as well as Mrs. Ruth Allewelt, for their enthusiasm and support. Mrs. Jayalakshmi Gurupatham of SPI Publisher Services provided great support and assistance in getting the manuscript of the second edition in shape for the printer.

Some illustrations were made by Professor Ulrich H. Kortenkamp, one of the authors of the Cinderella system, some were made by the author as drawings or computer aided, and some were made by Springer's illustrator, based on sketches by the author. The maps were drawn by the illustrator of Fagbokforlaget, Bergen, Norway, publisher of two volumes on the History of mathematic by the author, namely [33] and [34]. These maps are based on sketches by the author. The remaining illustrations are acknowledged as follows:

Figures 2.2, 2.8, 3.1, 3.6, 4.45, 4.52, 5.2, 6.1, 6.2, 10.11. From Wikimedia, Wikimedia Commons is *"a media file repository making available public domain and freely-licensed educational media content (images, sound and video clips) to all."* The description page in [62] has the following entry: *This image (or other media file) is in the public domain because its copyright has expired.* The name of the creator, if known, as well as other relevant information concerning the image is linked to the image in [61].

Figure 1.1. The illustration was provided by the authors of the article cited.

Figure 8.2: Archives of the Mathematisches Forschungsinstitut Oberwolfach.

Figure 4.54: Courtesy of the Mittag-Leffler Institute, Djursholm, Sweden.

Figure 4.22. Drawing by Nils Henrik Gran, Moss, Norway.

Figures 4.42, 4.43: Downloaded from From Wikimedia. Courtesy of The General Libraries, The University of Texas at Austin. According to the collection's title page, the image is in the public domain and no permission is needed to use it.

Bergen *Audun Holme*
In the spring of 2010

Contents

Part I
A Cultural Heritage

Chapter 1
Early Beginnings

1.1 Prehistory

Mankind must have possessed knowledge about geometric phenomena as far back as our historical records take us, and undoubtedly even much further back into the twilight of prehistorical times.

Human conceptions of number and form are documented as far back as the Old Stone Age, the *Palaeolithic*.

But there is no reason to assume this as being the beginning. On the contrary. Very interesting and astonishing results obtained by Jessica F. Cantlon and Elizabeth M. Brannon at Duke University show that monkeys have the ability to think about mathematics (Fig. 1.1). The author learned about this research from an article in the New York Times in 1998. We quote the summary of the article [2]:

> *"Nonhuman animals cannot perform the precise, symbolic mathematics that humans use to take a trigonometry exam or make change at a cash register. However, some basic principles of human mathematics are employed by nonhuman animals in order to estimate quantities of objects and events in their environment. Specifically, nonhuman animals represent numerical values and perform arithmetic operations such as ordering, addition, subtraction, and estimating proportions. We discuss evidence from a whole host of animal species, especially nonhuman primates, that suggests that the mathematical minds of humans may have more in common with the minds of other animals than was once suspected."*

In some so-called *primitive cultures* there are evidence that not only have such knowledge been present, but it has actually been remarkably advanced and sophisticated.

A famous example of this is the wall paintings found in a huge limestone cave in southern France. The cave contains about 300 remarkable paintings, about 30,000 years old. At this time there were made no written records, but these paintings may have served such purposes as well.

So the cave paintings document an artistic level of achievement second to none in later ages. They also indicate an understanding of space and form which certainly would have warranted formation of mathematical and geometrical concepts, had urge or need to do so existed at this time. But that may not have been the case more than 30,000 years ago. The people of the caves led difficult lives, collecting

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_1,

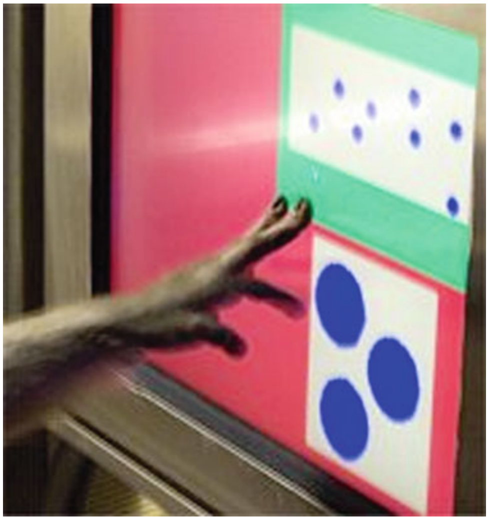


Fig. 1.1 A monkey orders numbers by pointing at a computer screen. In the experiment, the monkey has to choose the smaller numerical value and ignore the size of the dots

their livelihood hunting and gathering. The paintings may very likely have served as a magic vehicle for gaining control over nature, for casting a spell on the game thus ensuring a successful hunt. But a purpose of recording events, enumerating items, describe motion and spacial relationships may also have been present. One could say that such documents contain elements of *protogeometry*: The first stage of representing quantity, space and time, if not yet a truly abstract representation.

1.2 Geometry in the New Stone Age

Most historians of mathematics believe that mathematics and geometry did not develop until the need for it arose. This happened when the nomadic way of life ended as the sole basis for human society. This change marks the beginning of the New Stone Age or the *Neolithic*.

Replacing hunting and gathering by agriculture carried with it a completely altered way of life and a new society. Although taking place over thousands of years, this was a true revolutionary change. Living in one place, with valuable possessions and irreplaceable provisions, the need arouse for protection. Thus early urban centers were formed. Agriculture depended in many instances on artificial irrigation, through canals and irrigation channels. And protection from flooding necessitated the building of large structures as dykes. Also bulwarks in the form of walls around urban centers became necessary. Thus arose the need for engineering skills and insights.

Fig. 1.2 The Maltese Stonehenge. Photo by the author, directed by a friendly Maltese

When early humans lifted their eyes to the nightly sky, they were stirred by the same primordial emotions as their close or remote cousins among other creatures on the Earth. The light from a full moon, the millions of stars scattered across the firmament, the northern or the southern lights. Thunder and lightening in the heavens. But unlike the remote cousins, our forebears wanted to understand. And so they tied the events in the heavens with the life on the earth. The changing of seasons, heralded by certain stars appearing, becoming visible at sunset. In the north life reawakened as the sun gained strength, further south the rain at the onset of fall reawakened nature. All this contributed to the blend of early science, mathematics and astronomy with early religion, astrology and myths.

At the end of the third millennium B.C., northern Europe was still in the Neolithic period, the young Stone Age. From this time dates the Stonehenge at Salisbury Plain, Wiltshire in England. This mysterious structure may have served many purposes, one of them being as an astronomical observatory. Someone who stood at the center of Stonehenge on the morning of the summer solstice 4,000 years ago, would see the sun rising directly over a certain stone conspicuously located in the structure. Stonehenge has become an icon, but there is a very similar structure at Malta. The Maltese Stonehenge is not as well known as the British version, but makes the phenomenon even more fascinating. Other structures like this one in several locations at Malta seem to indicate a wide spread technology and civilization associated with such structures (Fig. 1.2).

There is also evidence which documents an impressive insight into astronomy or perhaps some kind of astrology. This has led some writers to speculate that there must have occurred visits by extraterrestrials from outer space: No other explanation would seem possible for the presence of such insights in so primitive societies, they argue. This line of reasoning bespeaks the prejudice and cultural arrogance in *our own society*, more than anything else. It is indeed humbling to contemplate how deep insights may have been gained, and then forgotten, time and again throughout our human history. But our brains today are the same as theirs then. The improvement probably lies in the way knowledge was eventually securely recorded and passed on to new generations, even as carefully guarded secrets of a priesthood.

1.3 Early Mathematics and Ethnomathematics

From this period we find numerous patterns of a geometric nature, best preserved on pottery but also from fragments of textiles. Today the study of such geometric activity forms part of what is known as *Ethnomathematics.* Such patterns were also used in artistic works by the Samish population in northern Scandinavia and in Russia. Much of this was destroyed by missionaries, considering it as pagan sorcery.

Several historians of mathematics have speculated on the connection between patterns of art and decoration and the development of numbers and numerology. There may well be close connections, but it is not easy to shed the preconceived ideas our modern mathematical education has provided us with, as we study this material. For example, a pattern of smaller isosceles triangles inside a larger one is aesthetically appealing. We may also use the same pattern to define the concept of *triangular numbers*. But does this mean that triangular numbers were known as a mathematical or numerical concept? Or is it fair to assume these patterns being evidence of a developed geometry at this time? Is it really mathematics?

These questions are perhaps more interesting than any proposed answers. They take us into the fascinating realm of the *Philosophy of Science*, perhaps posing insoluble enigmas. For our more practical purpose here, it is better to assume the broadest definition of *what constitutes mathematics*, thereby perhaps offending some purists but not unduly exclude an important facet of our field of study.

In any case there can be no doubt that historians of mathematics have made grave mistakes and misjudgments as to what constitutes *mathematical knowledge.* Thus for instance, even contemporary work on the history of mathematics can be seen as espousing the view that the ancient Mesopotamians "did not know the proof of the Pythagorean Theorem", simply because no clay tablet has been found containing a Greek-style proof of this fact! Is it not true that the Greek conceived of the *novel* idea of a *formal proof* because they considered mathematics as a branch of the *dialectics*, of the art of *debating*? And what if the ancient Mesopotamian Sumerian and Akkadian mathematicians were, not members of some debating-academy, but members of a *Priesthood of the Temple*, who closely guarded their wisdom from falling into the hands of the Unworthy? The mere idea of a *proof*, whose function it should be to convince the incredulous, would strike them at utterly absurd. That they still had secure knowledge in mathematics is firmly documented by the historical record, as we shall see later in this book.

Chapter 2
The Great River Civilizations

2.1 Civilizations Long Dead: And Yet Alive

In the first part of the fourth millennium B.C. a group of cultural centers developed in southwestern Asia. Probably emerging through coalescence from a web of small and, it would seem, insignificant Neolithic villages, impressive cities formed in the river valleys of the Indus, the Euphrates–Tigris and the Nile. Spreading out to form nets with other, in part more peripheral, urban centers, the classical civilizations bearing their names were born in these river valleys (Fig. 2.1).

Today their once flourishing life is attested to only by the mounds or *tells* covering their ruins. And by no means have we managed to uncover them all. Great finds are awaiting future archaeologists. The fascinating story of how these tells in some cases were persuaded to reveal their secrets, is not the subject of the present narrative. Nor is it our concern here to elaborate on the heroic endeavor and the almost superhuman tenacity, fortitude and resilience shown by those pioneers who managed to decipher the writings left behind by the vanished civilizations, thousands of years dead.

Suffice to recall that the pyramids and ruins in the Nile valley had been the objects of legends for millennia. Tales of treasures and curses.

Bedouins were roaming the deserts and fever-infested marshlands which now covered the once flourishing river basins of the Euphrates and the Tigris. They still worshiped with awe the mysterious *tells*, calling them by mystic names, the origins of which were long forgotten.

But in present day Iraq, not too far from the present city of Baghdad, were the crumbling ruins of the once so marvellous city of *Babylon*, the Gate of The Gods, the *Bab-Ili*. Once a center of science and mathematics, literature and history, astronomy and medicine, astrology and worship. As well as the bearer and center of a complex business structure which extended throughout the known world. With exquisite restaurants and a most sinful and hedonistic nightlife. In the rubble of what once were, there remained a ruin of what evidently had been a gigantic structure. Locally, it still was called reverently by the name of the ancient *Tower of Babel*.

Napoleon's invasion of Egypt led to the discovery of the Rosetta Stone and subsequently the decipherment of its inscriptions by *Jean-Francçois Champollion*.

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_2,

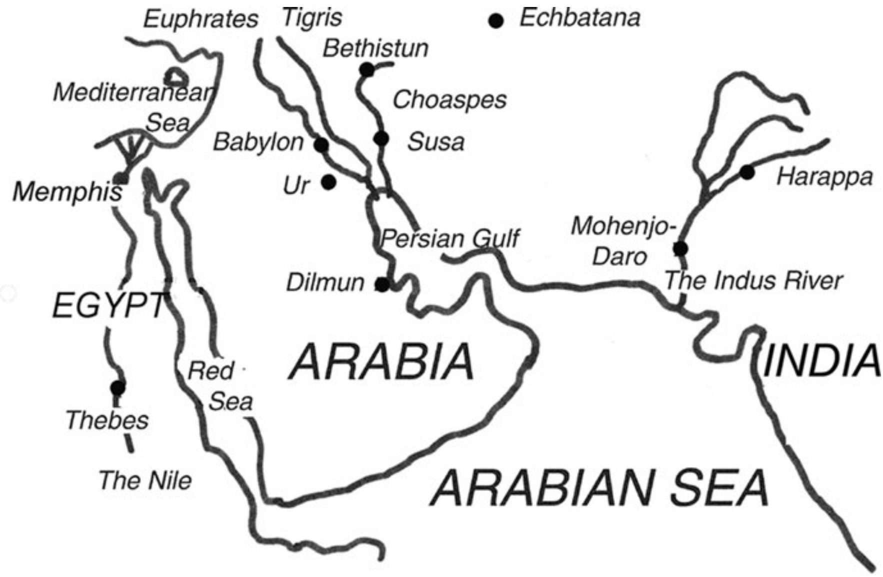


Fig. 2.1 This map shows the areas of the three Great River Civilizations

The main rivers shown are the Nile, the Euphrates and the Tigris. Euphrates flowed through the city of Babylon at this time, and the two rivers joined before they met the Persian Gulf. Also shown is the river of Choaspes, as well as the Indus River. Trade routes over land and sea connected them, and the presence of trade as well as other kinds of contact and interaction are very much in evidence. The ancient cities of Babylon and Ur of Mesopotamia are shown, Memphis and Thebes of Egypt, and Mohenjo-Daro and Harappa of the Indus Valley Civilization. Susa was the capital city of Persia before Darius founded the magnificent capital of Persepolis (in 518 B.C.). Dilmun was, according to some legends, the location of *the Garden of Eden*, central to three great religions. In ancient times this whole area was much more fertile than today, the climate was better and wildlife was abundant. Lion-hunting was a favorite activity for the Kings of Mesopotamia, as a result of which lions became extinct relatively early there.

Thus were opened up insights into events extending a thousand years beyond those recorded by the Bible and those accounted by the classical Greek historians and travellers. The decipherment of the hieroglyphic script of Egypt provided insights into all aspects of life in this ancient land.

The original script of ancient Egypt was, of course, *the hieroglyphs*. Eventually the need for a more practical and faster way of writing led to the development of the so called *hieratic* (sacred) script, a kind of simplified hieroglyphs. From the eight century B.C. a third type of script, which is even easier, is used, the so called *demotic* script, *the script of the people*. The Rosetta Stone is a fragment dating from 196 B.C., found in 1799 by one of Napoleon's officers. On it a decree had been written in Egyptian with hieroglyphs and with hieratic characters, and in Greek in demotic script. The breakthrough in deciphering the hieroglyphs and hieratic script came when Champollion could locate the names *Ptolemy* and *Cleopatra* in both

Egyptian scripts. It was then possible to compare with the Greek text, and piece together the alphabets.

The insights provided by the finds in the pyramids are subject to an important restriction. Almost exclusively all we have is material intended as support for the deceased in after-life. This is not to say that we are lacking texts pertaining to daily life or practical matters. But the criterion for preservation have been that of inclusion into the burial chamber.

Mesopotamia was now part of the once powerful Turk Ottoman Empire which was entering its final decades. The mounds and the areas around them had long been a source of building material of exquisite quality: Bricks of clay, baked at such high temperature that it surpassed by far what could be accomplished with contemporary technology. A number of these bricks were decorated with strange patterns of wedge-formed marks. These ubiquitous *decorated bricks* were excellent as building material, and served useful purposes as landfill as well as for building dykes along the banks of the Euphrates. In fact, as one of the first expeditions arrived at the old cite of Babylon to commence excavations there, workers were busily occupied with constructing huge embankments along the river banks. Using as building material not earth, but the books from the Imperial Library, housed at the Imperial Museum which had been located at the northernmost corner of ancient Babylon.

Indeed, the decorated bricks were books. The key to the decipherment of the *cuneiform* script in – or rather, on – these books, was provided by inscriptions found at Behistun near Persepolis, in the highlands of present day Persia. As in the case of the Rosetta Stone the inscriptions had been made in more than one language, one of which was Old Persian, which was known. An other version of the inscription was in the cuneiform script found on the *"decorated bricks."* Ordered by the Persian King Darius to commemorate his victories, the inscription had been carved on a great limestone cliff near the present village of Behistun, about 300 ft above the ground. Just getting an exact copy of the ancient letters was a strenuous and quite dangerous task. The story about the decipherment of the ancient script of the Sumerians and the Babylonians is suspenseful and fascinating. Building on work by the German philologist George Friedrich Grotefeld (1775–1853), the British diplomat and scholar, later to be called *the Father of Assyriology*, Henry Creswick Rawlinson (1810–1895), finally unravelled the script in 1846.

The assignment undertaken and carried out successfully by the archaeologists and linguists is humbling: A totally unknown script, writing texts in a completely unknown language, several thousand years dead. Nevertheless, the script was deciphered and the language was slowly reconstructed and pieced together. At first the results were viewed with skepticism by the scholarly world, suspecting a conspiracy of swindles. In the end a curious but convincing test was undertaken: A guaranteed new tablet was copied and given to a number of experts. Secluded they were then each required to make their individual versions of translation of the text! They passed. Even though there were discrepancies, there were enough common elements in all the translations to clearly demonstrate that they had indeed red and understood a common text which they had been given, à priori unknown to all of them.

The most mysterious of the Great River Civilizations is undoubtedly the *Indus Civilization*, contemporary to the early stages of the Mesopotamian and Egyptian ones. Excavations have uncovered what could be intriguing relations to the *Sumerian* civilization of ancient Mesopotamia.

In the time span between 2700 B.C. and 1500 B.C. the cities in the Indus valley developed into remarkable urban centers, carriers of an advanced civilization second to none from this epoch. Then decline set in, around 1,500 it is over. We do not know the cause of the demise of this great human achievement. Over the next 1,000 years a different way of life is in evidence, of a totally rural nature.

The three best known cities are *Harappa*, *Mohenjo-Daro* and *Chanhu-Daro*. The layout of the cities remarkably resemble that of ancient Mesopotamian ones: In Mohenjo-Daro the center is dominated by the *Citadel*, an elevated area surrounded by a wall about 50 ft high. Here we find *the Great Bath*, a watertight pool which may have had a sacred function. Below the Citadel lies the city, with broad avenues and more narrow side streets, arranged in a regular grid. Houses are build with baked brick, are usually two stories high around a central courtyard. All this closely resemble the layout and architecture found in Mesopotamian cities like ancient Ur and Babylon. Running water is supplied, and we find a covered system for drainage and sewage.

An intriguing feature is the lack of imposing structures immediately identifiable as *palaces* of kings or rulers. This has led some scholars to speculate that perhaps no ruling class existed at all, or possibly the ruling class harbored values which made them shun outwardly trappings of their elevated position. Also in evidence are a larger number of female sculptures, leading to hypothesizing of a matriarchal society.

More than 2,000 seals and seal impressions have been found. Again we find a close parallel to seals uncovered in Mesopotamia. As in Mesopotamia, they were carved from stone, and probably were used as the signature of the owner on various documents, letters and packets. The script, the *Harappan script,* found on them has not been deciphered as of this writing.

As of this writing no evidence of the mathematics or the geometry of this civilization has been uncovered.

This notwithstanding that evidence of their architecture and technology is everywhere. Also, they had a standardized system for weights and measure. One may, therefore, speculate that spectacular breakthroughs in revealing their science and mathematics may well lie in the future. This possibility is also borne out by the ubiquity of sophisticated geometric patterns and ornaments in decorations found throughout the Indus Valley area. These finds are of a clear *protogeometric nature*, which constitutes strong evidence of the sophistication required to support geometric ideas. Another intriguing piece of information is the following: In the first Indian mathematical text, presumably of Hindu origin, there are specific geometric rules for constructing *altars*. The tool for doing so is a set of *ropes* or *string*. The title of the work is the *Sulva-Sutra*, which means *"string-rules"*. The methods employed document knowledge of the so-called *Pythagorean Theorem* as well as *similar right triangles*. Now, the altars are supposed to be made of *burned bricks*, a technology the Hindus of that time did not possess, according to knowledgeable sources. But in the cities of the Indus civilization this technique is to be found everywhere. This has led some historians of mathematics to speculate that the Sulva-Sutra may have originated here. It certainly should be admitted that this is a speculative hypothesis, but it should be worth some serious digging at the cites in question.

2.2 Birth of Geometry as We Know It

Some historians have tended to dismiss the early science as *"merely magic and sorcery"*. But others have forcefully espoused the diametrically opposite view: *The ancients employed precisely the same method as modern scientists!* Indeed, the model of explanation they had for events in nature, for disease, for astronomical phenomena, and so on, was tried out. Corrections were attempted for the shortcomings. Eventually, through trial and error, with failures and mistakes, humanity arrived at our state of today. It may have taken a long time. Or did it really? The invention of the wheel, the first written records, may date from around the fourth millennium B.C. That makes 6,000 years up to our time. But compare that to the cave paintings of 30,000 years ago!

Amazingly, however, the earliest mathematics we encounter is qualitatively of the same nature as the mathematics of today. For no other science can one assert the same.

An important precondition for humans to be able to live in a well organized society, based on agriculture, is the existence of a reliable calendar. Indeed, without secure knowledge on the changes of the seasons, it is not possible to sow the grain or other seeds at the right time. Sowing too early may destroy the crops by nightly frost early on, and sowing too late may not leave enough time for it to ripen.

These needs were of the outmost importance, literally a question of life and death. And knowledge of a calendar is not possible without insights in astronomy, which again requires knowledge of geometry. Geometry and mathematics did also play an important role in measuring land, constructing irrigation channels or dykes along major rivers and in other engineering tasks. Some historians of mathematics speculate that the capricious and often unpredictably violent behavior of the Euphrates and the Tigris accounts for the fact that mathematics seems to have been better developed in ancient Mesopotamia than it was in ancient Egypt, where the more benign Nile behaved with exemplary regularity. But this comparison is not uncontroversial: Other historians argue that we know more about Mesopotamian mathematics than we do of the Egyptian, simply because the former was written on baked clay tablets, a practically imperishable medium, while the Egyptians wrote on papyrus which has a much shorter life under normal circumstances.

2.3 Geometry in the Land of the Pharaoh

The Egyptian civilization erected itself a proliferation of monuments in the form of huge geometric objects: The Great Pyramids. Such is the immenseness of these artifacts that some writers have speculated that they were left behind by extraterrestrials visitors to Earth. How could people without a sophisticated technology make plans for these structures, let alone carry out the actual constructions?

And the pyramids themselves have been surrounded by mysticism and speculations by puzzled observers. But the greatest pyramid of them all was not, according to some historians of mathematics, one found in the Egyptian desert. Instead, it is found on an ancient piece of papyrus, named the *Moscow Papyrus* (Fig. 2.2).

The so-called Moscow Papyrus dates from approximately 1850 B.C. The papyrus contains 25 problems or *examples*, already old when the papyrus was written. It was bought in Egypt in 1893 by the Russian collector *Golenischev*, and now resides in the Moscow Museum. The text was translated and published in 1930 by *W.W. Struve*, in [56]. This papyrus may show that the mathematical knowledge of the Egyptians went considerably further than the so-called *Rhind Papyrus* (see below) demonstrates.

In one of the problems treated there, a formula for the volume of a frustum of a square pyramid is given. If a and b are the sides of the base and the top, respectively, and h is the height, then the formula for its volume is

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

This is exactly right, and its beauty and simplicity has led some historians of mathematics to reverently refer to it as *the greatest of all the Egyptian Pyramids.*

It is often asserted that this formula was unknown to the Babylonians, thus documenting a rare instance where Egyptian mathematics surpassed the Babylonian. But whereas there does exist tablets from Babylonia where the (obviously false) formula

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

is used, there also exists at least one tablet where a formula equivalent to the Egyptian one may have been employed, for a frustum of a cone.

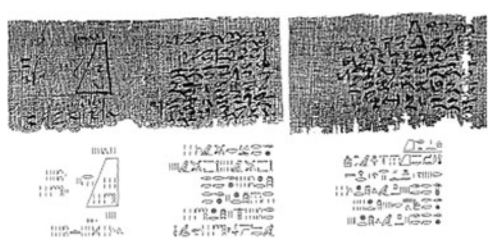


Fig. 2.2 The Moscow Papyrus with the geometry described in the text

This is according to a controversial interpretation by Neugebauer, see [58, pp. 75–76]. Much of the interpretation hinges on whether there is an error in the calculation on the tablet. By the way, some of the tablets we find from ancient Babylonia are the "papers" prepared by the students of the *Temple Schools* or the *Business Schools* which could be found in the larger cities, certainly in Babylon itself. So some sources must be treated with caution. On the other hand, there are some 22,000 tablets from the Royal Library of the last of the great Assyrian Kings *Ashurbanipal* at Nineveh.

Another explanation for the mistake might be that the correct formula can be written as

$$V = h\frac{(A + \sqrt{AB} + B)}{3},$$

where A and B denote the surface area of the base and the top (Fig. 2.3). Now $\frac{(A+\sqrt{AB}+B)}{3}$ is known as the *Heronian Mean* (named after Heron of Alexandria) of A and B. So one might speculate if two different kinds of "mean" have been confused here.

But be this as it may, the geometric insights documented by the *Greatest of the Egyptian Pyramids* is surely prodigious. It is instructive to attempt deducing this formula by our present day High School Math. We proceed as illustrated in Fig. 2.4.

We let the side of the base be a and the side of the top be b. The height of the big, uncut pyramid is $OC = T$, and the height of the small one, which has been removed, is $OC' = t$. Thus the height of the frustum is $h = T - t$, in other words the distance between the base and the top. Further $AB = a$ and $A'B' = b$ so that the similar triangles given by O, A', B', and O, A, B yield

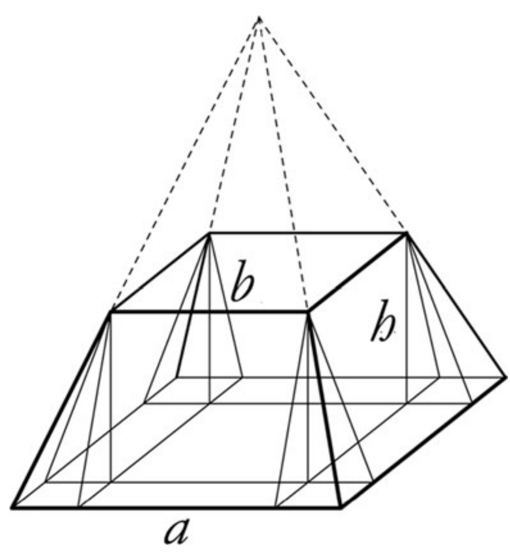


Fig. 2.3 Proof by cutting and reassembling

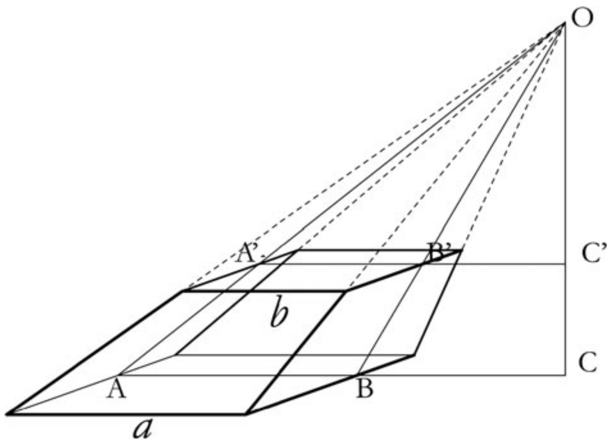


Fig. 2.4 Deducing the formula by our present day High School Math

$$\frac{T}{a} = \frac{t}{b}$$

We are now ready to compute the volume V of the frustum.

$$V = \frac{1}{3}Ta^2 - \frac{1}{3}tb^2 = \frac{1}{3}\left(\frac{T}{a}a^3 - \frac{t}{b}b^3\right) = \frac{1}{3}\frac{T}{a}(a^3 - b^3) =$$
$$\frac{1}{3}\frac{T}{a}(a-b)(a^2 + ab + b^2) = \frac{1}{3}(T - \frac{T}{a}b)(a^2 + ab + b^2) =$$
$$\frac{1}{3}(T - t)(a^2 + ab + b^2) = \frac{1}{3}h(a^2 + ab + b^2)$$

The Moscow-papyrus also contains another problem of great interest. Struve, in [56], claims that in it, Egyptian mathematicians document that they know how to compute the surface area of the sphere (actually, the hemisphere). He interprets the text as computing this area with a value for π implicitly given by the formula

$$\frac{\pi}{4} = \left(1 - \frac{1}{9}\right)^2,$$

which gives $\pi = 3\frac{13}{81} \approx 3\frac{1}{6}$. Other researchers disagree sharply with Struve's interpretation. Van der Waerden writes as follows in [58]:

> The genius of the Egyptians would have been wonderful and indeed incomprehensible, if they had succeeded in obtaining the correct formula for the area of the hemisphere.

The situation is not improved by the presence of an unfortunate hole at a decisive spot in the papyrus. Thus it must be regarded as an open question whether the

Egyptians knew the formula for the surface area of the sphere. But the claim is supported by the fact that Papyrus Rhind also does give this value for the number π.

The so-called *Papyrus Rhind* is in fact the most important papyrus for our understanding of Egyptian mathematics. It has been given this name because it was bought in Luxor by the Scottish Egyptologist *A. Henry Rhind* in 1858. Rhind, who was in poor health, had to spend some winters in Egypt. He died on his way home from his last visit there in 1863, and the papyrus was purchased from his executor by British Museum, together with another Egyptian mathematical document known as *the Leather Scroll.*

A more appropriate name for this important papyrus would be *the Ahmes Papyrus*, after the *Egyptian Scribe*[1] who copied it from a considerably older papyrus. This name is now being used more frequently. Ahmes relates on it that the original stems from the Middle Kingdom, which dates to about 2000 to 1800 B.C.

The copy by Ahmes is from around 1650 B.C. Together with the Moscow Papyrus and the Leather Scroll, the Ahmes Papyrus forms our main source for Egyptian mathematics. The Egyptians used the value given above for π, and with this value a computation which appears on the papyrus, uses the correct formula for the area of a circular disc. Altogether the papyrus has the appearance of a practical handbook of math, explaining basic methods by doing a total of 85 examples.

A very beautifully booklet has been published recently with photographs in color of the entire papyrus, transcription of the hieroglyphs and figures on it and explanation of the mathematics in a modern language. Highly recommended reading [48].

2.4 Babylonian Geometry

When we use the term *Babylonians* we actually mean the civilization residing in the whole of Mesopotamia, not just the citizens of that marvellous city Babylon. This culture was already highly developed at the time from which we find the earliest records, the ancient culture of the *Sumerians*. The main city was not Babylon, until comparatively recent times. The ancient city of *Ur* in southern Mesopotamia was the spiritual and political center for a long time. The Sumerians arrived in this region with their culture already well developed, we do not know from where. The political hegemonies shifted over time, most notably with the arrival of the Akkadians, of which the Babylonians eventually were part. But new rulers carefully preserved the old culture of the Sumerians, and the Kings carefully collected ancient books, baked clay tablets, in Libraries, and made translations into the Akkadian from the Sumerian. In fact we have preserved elaborate dictionaries for the two languages, as well as parallel translations.

The Babylonians had a sophisticated way of representing numbers and computing. They represented numbers to the base 60, in the same way as we represent

[1] *Ahmes* is the earliest individual name associated with mathematics which we know.

Fig. 2.5 Some sexagesimal digits. Above (10), (20), (30), (40) and (50). Below (1), (2), (21), (19) and (59)

numbers to the base 10. Thus for instance they would represent the number 61 as (1)(1), while the number 6,359 would be represented as (1)(45)(59).

The name *sexagesimal* comes from the Latin term *sexagesimus*, which means "sixtieth". The word *sex* is Latin for *six*. In Greek "six" is *hex*, hence the terms *hexadecimal*, meaning the number system with base 16, used extensively in Computer Science. Further, the term *hexagon* means *6-gon*. In Fig. 2.5 we have written the *sexagesimal* digits in parenthesis. Those possible digits are of course (0), (1), . . . , (59). Using a stylus usually cut from reed, the Babylonians impressed wedges on clay tablets, which were subsequently baked if the writing was to be preserved. Wedges of different shapes were used, thus making it possible to codify a large set of characters. The digits from 1 to 59 were build up of two types of wedges, in the simplest script in use (others were also present at different epochs). In Fig. 2.5 we see some digits, ending with (59). Note the mixture of base, as the individual digits in the base 60-system were represented with symbols for 1's and 10's.

The Babylonians did not directly use the digit (0) in the beginning, but did so indirectly by leaving an open space: Nothing there! But as *scribes*, writers and copiers, copied old tablets to new clay to be baked, mistakes were easily made. So to clarify matters, they started to write a symbol which meant *None* or *Not*. But trailing zeroes were not used. Thus context would have to determine whether (1) meant 3,600, 60, 1, $\frac{1}{60}$, . . . Even though we would find this clumsy, it represented a *numerology*, a representation and understanding of numbers, far superior to that of the Egyptians, Greek or the Romans.

We know a great deal about the mathematics of the Babylonians. This research was to a large degree initiated by *Otto Neugebauer* and his collaborators and associates. Like many others Neugebauer had to flee Germany during the Nazi era, and came to the United States. He uncovered and interpreted many tablets from Babylonia, and made the striking discovery of the meaning of the most famous of all tablets which have been found until now, and which we shall return to below.

While realizing that the Babylonians had admirable mathematical insights, historians of mathematics had no clear understanding of the motivation behind it. In fact, it was a widespread view that all mathematics prior to the Greek period only consisted of simple practical computations for everyday applications in trade, agriculture and simple engineering tasks. Mathematics as the science we know it, they maintained, did not exist until the advent of the Greek. This view would be espoused since it was the Greek who introduced the concept of a mathematical proof.

But it is a fundamental misunderstanding that there can be no mathematics as a science without our modern notion of *proof*. Indeed, the creative process which every research mathematician engages in when mathematics is discovered is almost the complete opposite of a formal proof. Only *à posteriori* do we mathematicians cloak our work in the formal style of *Satz–Beweis*, so beloved by some professors but equally hated by the majority of their students. Of course proofs are necessary so as to ensure correctness of results. And actually *finding* a proof of a conjecture everyone believes to be true is also very much central to mathematics, as in the case of Andrew Wiles' proof of the famous Fermat Conjecture in the last decade of the twentieth century, or *Grigori Perelman's* recent proof of the *Poincaré Conjecture.* But it really is not necessary to have produced a formal proof of a mathematical theorem in order to document complete knowledge of why the theorem is indeed true.

As it happens, a careful analysis of a baked clay tablet from ancient Babylon elucidates this point very well.

The tablet which is perhaps the most famous one, has been given the name *Plimpton 322.* It signifies that it is the tablet numbered 322 in the *A.G. Plimpton* collection at Columbia University in New York. The tablet is written in old Babylonian characters, dating from the period 1900–1600 B.C. We follow some of the description of the tablet in E. Robson [51]: The tablet is about 13 by 9 by 2 cm. Its second and third column list the smallest and largest member of Pythagorean triples, one may think of the shortest side and the hypothenuse of a right angled triangle. The final column contains the line count from 1 to 15. Unfortunately the tablet is damaged, in that a piece along the entire left edge is missing. Moreover, there is a deep indentation at the middle of the right hand side. Finally, it is also somewhat damaged at the upper left corner. So the first column is partly broken away. It may have contained either the square of the hypothenuse divided by the square of the longest side, or the square of the shortest side divided by the square of the longest side.

Whatever interpretation of these incomplete data, however, the tablet documents that the Babylonians had firm knowledge of so called Pythagorean triples.

Some claim that it has been found traces of modern glue along the rupture-edge at the left, thus indicating that it was complete at excavation, but broke thereafter in the possession of individuals with access to such amenities as glue, who attempted repairing it.

If so, it would be interesting if the missing piece could somehow be traced. It could reside in one of the many bins of unclassified and unintelligible fragments of Babylonian tablets. As it happens, this was the gravest danger facing the ancient tablets: Destruction at the time of their excavation, which was often – at least in the beginning – done quite crudely.

The tablet was acquired by an interesting character named *Edgar James Banks*, (1866–1945).[2] He was an American college professor, antiquities collector and dealer, and adventurer. He was active in the Ottoman Empire, at the end of its

[2] We follow [61] among other sources.

existence, and is probably an original for the figure of *Indiana Jones.* He started out as American consul in Baghdad in 1898, and bought many cuneiform tablets on the markets of the decaying Ottoman Empire. These he resold in carefully planned small installments, so as not to flood the international market and thus deflate prices. The tablets went to museums, libraries, universities, and theological seminaries. One of the tablets which Banks sold, was, according to the information he gave, from Senkereh in southern Irak, near the ruins of the ancient city of Larsa. He sold this tablet to Professor Eugene Smith of Columbia University in New York. Smith willed his books to the university, and the tablet is today number 322 in the A.G. Plimpton library's collection of rare books.

The contents of Plimpton 322 demonstrates that the Babylonians had firm knowledge Pythagorean triples. They probably also knew the so called Pythagorean Theorem. In what sense did they know this? In the absence of firm knowledge we may ask questions and speculate. Before proceeding with Plimpton 322, I shall present the simplest and most beautifully proof I know of this theorem.

Is it the *Babylonian proof*, the proof they knew? But they would not call it a proof, but regard it as an example of using the *rule by which certain areas may be added.* And, of course, we give the proof here in modern language and symbolism. But first we give a more conventional proof, the principle behind it might also have been known to the Babylonians, in Fig. 2.6. See Howard Eves, [14].

In Fig. 2.6 the three sides in the right triangle are labelled as above: The hypothenuse as d, the two others as a and b, where $a \geq b$. We then set

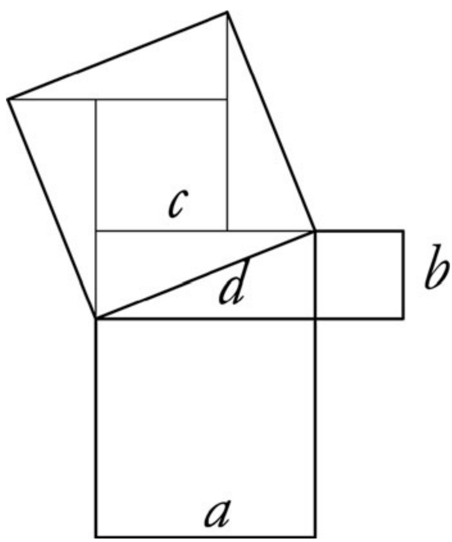


Fig. 2.6 A (very hypothetical) Babylonian proof of "Pythagoras' Theorem". The essential part of this figure, namely the subdivision of the largest square, appears in the oldest Chinese mathematical text we know, the *Chóu-pü*, from the second millennium B.C. Thus evidence suggests that this insight formed part of a common wisdom in the ancient world

$$c = a - b$$

From the figure we now see that the area of the square on the hypothenuse, d^2, is equal to c^2 plus the areas of the four right triangles congruent with the given one. As the area of a triangle is equal to *half the base times the height*, a fact well known to the Babylonians, we get

$$d^2 = c^2 + 4\left(\frac{1}{2}ab\right) = c^2 + 2ab$$

But as the Babylonians also knew,

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

which, using the formula in the case of the minus-sign, finally yields

$$d^2 = a^2 + b^2,$$

as desired.

Figure 2.6 and the corresponding proof is one possibility. A variation of the same theme, less familiar to us in our usual thinking concerning "Pythagoras' Theorem", but even more in line with the way the *Babylonians* thought, is a proof derived from Fig. 2.7.

Indeed, the Fig. 2.7 yields

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

from which follows $d^2 = a^2 + b^2$.

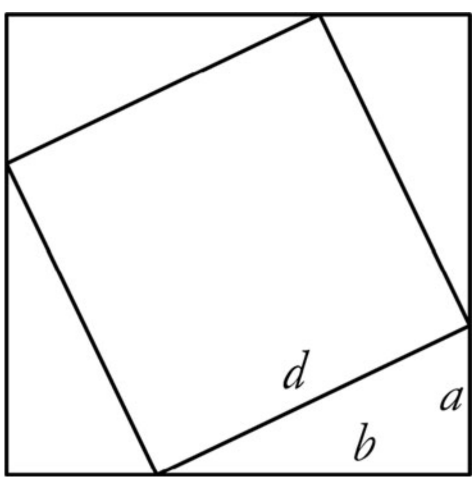


Fig. 2.7 The Putative Babylonian Proof of "Pythagoras' Theorem"

Such methods for dealing with sums of squares is well documented from Babylonian tablets. From [55, pp. 27–28], we reproduce the following example, to be found on a tablet in Strasbourg's *Bibliothèque National et Universitàire*. Phrased in modern language:

> An area A, consisting of the sum of two squares, is 1,000. The side of one square is $\frac{2}{3}$ of the side of the other square, diminished by 10. What are the sides of the square?

The Babylonians would solve this as follows, again presented in modern language: The sides of the respective squares are denoted by x and y. We then have $x^2 + y^2 = 1{,}000$, as well as the relation $y = \frac{2}{3}x - 10$. Squaring the latter yields

$$y^2 = \frac{4}{9}x^2 - 2 \cdot \frac{2}{3}x \cdot 10 + 10 \cdot 10 = \frac{4}{9}x^2 - \frac{40}{3}x + 100.$$

Substitution into the first equation yields

$$\frac{13}{9}x^2 - \frac{40}{3}x - 900 = 0.$$

Having thus transformed the geometric problem into an algebraic one, the Babylonian scholars and scribes – rather, in the present case presumably students doing their homework – could find the solution utilizing their knowledge about equations and systems of equations. The answer to the present problem is 30, the one positive solution of the equation.

The presentation of the solution starts like this: *"Square 10, this gives (1)(40) (i.e., 100). Subtract (1)(40) from (16)(40) (i.e., 1,000), this gives (15)(0) (i.e., 900)..."*

We return to Plimpton 322 (Fig. 2.8). The tablet contains a table of numbers, arranged in four columns of 15 numbers each. The rightmost column just consists of the numbers $1, 2, \ldots, 15$. The column to the left is partly destroyed by the missing part.

Today Neugebauer and Sachs' explanation is no longer generally accepted. An alternative explanation is by so called *reciprocal pairs*. The explanation is due to a number of authors. Accounts of this work, with references, may be found in Eleanor Robson [51] and [50], as well as in *Jöran Friberg's* book [15]. Friberg ties Plimpton 322 to an Old Babylonian generating rule, which has been ascribed to Pythagoras and Plato, and also appear in Euclid's Elements, Book X. Below we shall start with the explanation which was given by Neugebauer and his collaborators, then give a briefly summary of Friberg's and Robson's explanation of the method of regular reciprocal pairs.

However, the competing explanations are mathematically related, and they have similar far reaching consequences. They demonstrate first of all that the Babylonians knew ways to generate such triples (a, b, d). It is also fairly certain that they knew the so-called Pythagorean Theorem. But exactly how did they work out their list of the Pythagorean triples?

As already stated, the rightmost column only serves to *number* the entries in the other columns. But the two next columns look at first rather haphazard and arbitrary.

Fig. 2.8 The tablet Plimpton 322

At first this led some to assume that the tablet merely constituted a fragment of some business-files, which are actually very much present in quantity among the ancient tablets from Babylon. But the column to the left bears the heading "*diagonal*", while the next has the heading "*breadth*". As with most of the numbers in the first row, to the left, also the heading here is illegible. But the consensus of opinion among the experts is that the numbers constitute in some way a list of Pythagorean triples. How they are presented, however, cannot be ascertained with certainty.

Clearly, given any Pythagorean triple (a, b, d), we get another by multiplying each number by the same natural number r, obtaining (ra, rb, rd). Thus we need only to generate the so-called *primitive* Pythagorean triples, that is to say the triples where the numbers do not have a common factor >1. Now there is an elegant way of generating all possible primitive Pythagorean triples. Usually the method and its proof is attributed to *Diophantus*, as it is explained in his *Arithmetica*. But recent detective work might indicate that *Hypatia of Alexandria* deserves some of the credit for this work, see Chap. 4, Sect. 4.20, as well as [11].

In Sect. 2.7 we shall give a completely modern account of the method for finding the primitive Pythagorean triples, according to Diophantus as reconstructed by *Fermat and Newton* much later, still. The reader who is only interested in Plimpton 322, may skip that section.

Only a small fragment of this theory is really needed in the explanation of the Pythagorean triples on Plimpton 322. In particular the full statement of Theorem 1 in Sect. 2.7 is not needed, it suffices to carry out the obvious verification

$$(v^2 - u^2)^2 + (2uv)^2 = (v^2 + u^2)^2,$$

thus

$$a = v^2 - u^2,\ b = 2uv \text{ and } d = v^2 + u^2$$

form a Pythagorean triple. But when the significance of Plimpton 322 was first discovered by Neugebauer and Sachs, there was a tendency to interpret it as evidence of a much more far reaching mathematical knowledge on the *parametrization of primitive Pythagorean triples*.

2.5 The (u,v) Explanation of Plimpton 322

It is generally accepted that the tablet contains four errors. Three of them are easy to explain as a simple mistake with the stylus, whereas the fourth is more mysterious. Several explanations have been offered, but as long as we only have this one table of this type, and in view of the missing part, it is difficult to decide what the correct explanation is.

At any rate, except for these presumed errors the second and third column from the right consists of the numbers b and d described above, for the choices of u and v shown in the table presented as Fig. 2.9.

We have written the corrected numbers with the presumed erroneous ones in parenthesis.[3] We start with entry number 11: The values 2 and 1 should give $(b, a, d) = (4, 3, 5)$ which is not shown. Instead this triple is multiplied with 15, to give more palatable digits in the Babylonian number system. Next we note that the last entry, in line number 15, is not a primitive triple.

b	a "Breadth"	d "Diagonal"	No.	v	u
120	119	169	1	12	5
3456	3367	4825 (11521)	2	64	27
4800	4601	6649	3	75	32
13500	12709	18541	4	125	54
72	65	97	5	9	4
360	319	481	6	20	9
2700	2291	3541	7	54	25
960	799	1249	8	32	15
600	481 (541)	769	9	25	12
6480	4961	8161	10	81	40
60	45	75	11	2	1
2400	1679	2929	12	48	25
240	161 (25921)	289	13	15	8
2700	1771	3229	14	50	27
90	56	106 (53)	15	9	5

Fig. 2.9 The reconstructed Plimpton 322

[3] Also note that these are all *primitive* triples corresponding to the given values of v, u, with the exception of two entries. This observation is only interesting for the readers who will study Sect. 2.7.

If this explanation of the numbers on Plimpton 322 is correct, the numbers u and v would be carefully chosen. First, they would all be regular sexagesimal numbers: Their inverses are finite sexagesimal fractions. That such choices are possible at all for the entire table is due to the choice of base 60, which has the prime factors $2, 3, 5$, whereas base 10 only has $2, 5$. Thus for instance, in Babylonia they would have $\frac{1}{3} = (0) \cdot (20)$ and $\frac{1}{15} = (0) \cdot (4)$. Then the tricky long division in the sexagesimal system could be avoided in many cases, and replaced by multiplication, which they easily performed using multiplication tables on baked clay tablets.

With our base of 10, *we* have a special relationship to the numbers $3, 7$ and 13, as being, respectively, *lucky, sacred and unlucky*. The Babylonians do not seem to have offered 3 much thought, but 7 was sacred and 13 was very unlucky, *The Number of the Raven*.

As stated above the leftmost column may have contained either the square of the hypothenuse divided by the square of the longest side of the triangle, or the square of the shortest side divided by the square of the longest side. Therefore it has been speculated that this tablet might have been used in computations as equivalent to a table over $\text{cotan}(\varphi)$ or $\cos(\varphi)$ for angles φ between $44°46'$ and $31°53'$. According to Robson [51], page 112 the concept of angle is anachronistic, in that the Babylonians did not have this concept.[4] The decrement in the values of φ are not constant, and $\sec(\varphi)$ decreases by very roughly $\frac{1}{60}$ from one line to the next. Is Plimpton 322 part of a set of "trigonometric tables" for use in astronomy and engineering? Some might still like to believe that, but there is no evidence for such a usage. On the contrary, Babylonian astronomy and astrology flourished much later than the Old Babylonian Epoch, which the tablet comes from.

2.6 Regular Reciprocal Pairs, Babylonian Number-Work and Plimpton 322

The Babylonians did most of their number-work relying on tables. For example, multiplication could be carried out using the tables of squares by the formula

$$xy = \frac{1}{4}((x+y)^2 - (x-y)^2).$$

Moreover, one should note that

$$\frac{1}{4} = (0) \cdot (15),$$

and multiplication with this number is especially simple in base 60, much like multiplying by 0.2 or 0.5 in base 10.

[4] But this would not preclude that there might exit tables which have served a similar purpose to trigonometric tables.

In addition to tables of squares, the students of the ancient scribal schools had to learn sexagesimal multiplication tables by heart, and also had to learn tables of *regular sexagesimal reciprocal pairs.*[5] These tables were important for a handy conversion of a problem of division into a problem of multiplication.

As an illustration of a division using this, we look at

$$123 : 12 = 10.25,$$

with our decimal system, in modern sexagesimal notation $(10) \cdot (15)$, while the Babylonians would write the answer as $(10)(15)$.

The Babylonians would very probably *not* handle such an easy division by their Method of Reciprocal Pairs, but nevertheless, here is how it works: First observe that $12 \times 5 = 60$, thus in modern sexagesimal notation $\frac{1}{12} = (0) \cdot (5)$, and in Babylonian notation the reciprocal of (12) is (5). Since, as we would write

$$123 : 12 = 123 \times \frac{1}{12} = (2)(3) \times (0) \cdot (5),$$

the Babylonians would proceed to multiply $(2)(3)$ with (5), obtaining the answer $(10)(15)$, immediately and without having to consult tables of squares. Finally this answer has to be interpreted right, going back to the context. The correct answer is $10 + \frac{15}{60} = 10\frac{1}{4}$ rather than, for instance, $10 \times 60 + 15 = 615$.

Now we return to Pythagorean triples. We have worked above with a particular reciprocal pair, namely $(12, \frac{1}{12})$ in our notation. Now it turns out that every such pair of reciprocals x and $x' = \frac{1}{x}$ yields two rational numbers $b' = \frac{x-x'}{2}$ and $d' = \frac{x+x'}{2}$ such that with $a' = 1$ we get $a'^2 + b'^2 = d'^2$, in other words (a', b', d') is a rational Pythagorean triple. In fact, since $xx' = 1$ we get

$$\begin{aligned} a'^2 + b'^2 &= \left(\frac{x - x'}{2}\right)^2 + 1 = \frac{x^2 - 2xx' + x'^2 + 4}{4} \\ &= \frac{x^2 + 2xx' + x'^2}{4} = \left(\frac{x + x'}{2}\right)^2 = d'^2. \end{aligned}$$

With $x = 12$ we obtain $b' = \frac{143}{24}$ and $d' = \frac{145}{24}$. Scaling this rational triple we get a Pythagorean triple of integers $(24, 143, 145)$, which by the way does not appear on Plimpton 322.

Now, going back to the (u, v)-explanation, we have that

$$a = v^2 - u^2,\ b = 2uv \text{ and } d = v^2 + u^2,$$

thus

[5] See Robson [51, p. 113].

$$a' = \frac{a}{b} = \frac{v^2 - u^2}{2uv} = \frac{1}{2}(x - x'),\ b' = \frac{b}{b} = 1 \text{ and } d' = \frac{v^2 + u^2}{2uv} = \frac{1}{2}(x + x'),$$

where $x = \frac{v}{u}$ and $x' = \frac{u}{v} = \frac{1}{x}$. Hence from a mathematical point of view the two explanations are equivalent. However, the point is that regular reciprocal pairs are ubiquitous in Babylonian mathematics, whereas primitiveness and parametrization appears nowhere else. This argument alone would lead one to discard the (u, v)-version of the explanation in favor of the regular reciprocal pairs.

Friberg, in [15, p. 92], refers to the rule

$$d, b, a = \frac{x + x'}{2}, 1, \frac{x - x'}{2}$$

as the *Old Babylonian generating rule*, and he argues on page 88 for the following tentative translation of the headings of Plimpton 322, although as he states "The meaning [...] is far from obvious":

> The square of the holder for the diagonal (from) which 1 is subtracted, then [the square of the holder for] the front comes up. The square side of [the square of the holder for] the front. The square side of [the square of the holder for] the diagonal. Its line number.

2.7 Parametrization of Pythagorean Triples

We now explain the complete theory of parameterizing primitive Pythagorean triples. Let (a, b, d) be a Pythagorean triple. We then have

$$\left(\frac{a}{d}\right)^2 + \left(\frac{b}{d}\right)^2 = 1,$$

i.e., the point $(x, y) = (\frac{a}{d}, \frac{b}{d})$ lies on the unit circle which has the equation

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

So the problem is equivalent to finding all points with rational coefficients on this circle. We now pull one of today's standard tricks, taught in every class of first year calculus: We wish to find *a rational parametrization of the circle*, that is to say, to find rational expressions in some variable t, $x = \varphi(t)$, $y = \psi(t)$, such that when t varies, then $(\varphi(t), \psi(t))$ runs through all points on the circle. The trick is to let t be the slope of the line through the point $(-1, 0)$, see Fig. 2.10.

The equation of this line is

$$y = t(x + 1),$$

which we substitute into the equation for the circle, thus obtaining

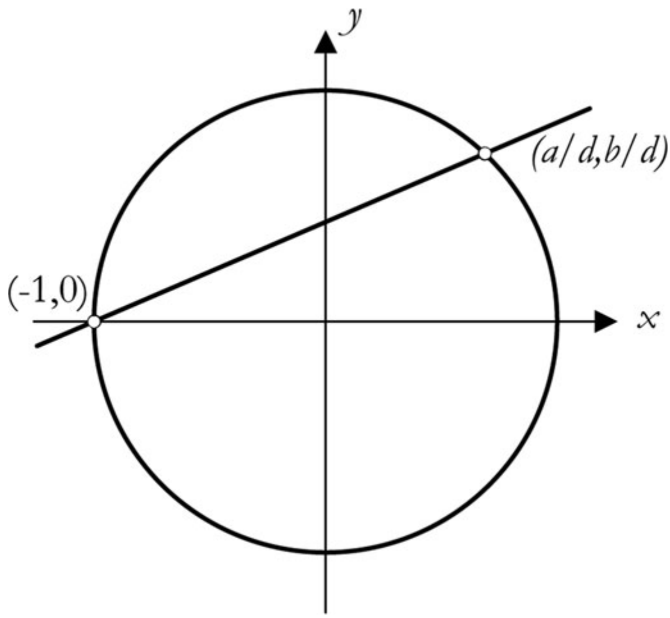


Fig. 2.10 Finding all rational points on the *circle*

$$x^2 + t^2(x+1)^2 = 1,$$

and hence

$$(1+t^2)x^2 + 2t^2x + t^2 - 1 = 0,$$

which, as $1 + t^2$ is never zero, may be written as

$$x^2 + \frac{2t^2}{1+t^2}x + \frac{t^2-1}{1+t^2} = 0.$$

Now the formula for the roots of the general second degree equation,

$$x^2 + px + q = 0,$$

is

$$x = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q},$$

which when applied to the equation in question here yields

$$x = -\frac{t^2}{1+t^2} \pm \sqrt{\left(\frac{t^2}{1+t^2}\right)^2 - \frac{t^2-1}{1+t^2}} = -\frac{t^2}{1+t^2} \pm \frac{1}{1+t^2},$$

after a short computation. We thus obtain

$$x = -1 \text{ or } x = \frac{1-t^2}{1+t^2}$$

Substituting the last solution into the equation for the line, we get

$$y = t\left(\frac{1-t^2}{1+t^2} + 1\right) = \frac{2t}{1+t^2}$$

Since the points (x, y) have rational coordinates, we may write $t = \frac{u}{v}$ for natural numbers u and v. Here we must have $v > u$ since the slope t of our line lies in the interval $<0, 1>$. Substituting this into the expressions for x and y, we obtain the following formulas:

$$x = \frac{a}{d} = \frac{v^2 - u^2}{v^2 + u^2}, \quad \text{and} \quad y = \frac{b}{d} = \frac{2vu}{v^2 + u^2}.$$

We have essentially completed all ingredients needed to prove the following:

Theorem 1. *All primitive Pythagorean triples* (a, b, d) *are given by*

$$a = v^2 - u^2, b = 2uv \quad \textit{and} \quad d = v^2 + u^2,$$

where u and v are positive integers, $v > u$, without a common factor > 1. Moreover, u and v are not both odd numbers.

Proof. First of all, numbers of the form $a = v^2 - u^2, b = 2uv, d = v^2 + u^2$ where u and v are natural numbers do form a Pythagorean triple, as is seen by computing $a^2 + b^2$. If we assume that u and v have no common factor > 1, then the triple is also primitive, *except for the possibility that* $a = 2\overline{a}, b = 2\overline{b}$ and $d = 2\overline{d}$. Indeed, a, b, d can have no other common factor than 2, and it is easily seen that this happens if and only if u and v are both odd numbers. Then the overlined numbers do form a primitive Pythagorean triple. In this case we introduce new versions of u and v by putting

$$\overline{v} = \frac{v+u}{2}, \quad \overline{u} = \frac{v-u}{2},$$

from which we find

$$2\overline{uv} = \frac{v^2 - u^2}{2} = \overline{a}, \overline{v}^2 - \overline{u}^2 = uv = \overline{b} \text{ and } \overline{v}^2 + \overline{u}^2 = \frac{v^2 + u^2}{2} = \overline{d}$$

It is not difficult to verify that $\overline{v}, \overline{u}$ have no common factor > 1, and are not both odd numbers.

Now, given a primitive Pythagorean triple (a, b, d). From the considerations preceding the formulation of the theorem, we can always find natural numbers v and u, such that

$$\frac{a}{d} = \frac{v^2 - u^2}{v^2 + u^2}, \text{ and } \frac{b}{d} = \frac{2vu}{v^2 + u^2}.$$

Unless u and v are both odd numbers, we therefore have that $a, b,$ and d must be as claimed in the theorem. If u and v are both odd, then we proceed as above, obtaining new u and v's, $\overline{v} = \frac{v+u}{2}, \overline{u} = \frac{v-u}{2}$, also without common factors, but now not both odd numbers, such that

$$2\overline{vu} = \frac{v^2 - u^2}{2}, \overline{v}^2 - \overline{u}^2 = uv, \text{ and } \overline{v}^2 + \overline{u}^2 = \frac{v^2 + u^2}{2}.$$

Thus the primitive Pythagorean triple a, b, d is described as in the theorem, but with the roles of a and b interchanged. □

Exercises

Exercise 2.1 An ancient method for computing the area of a circle is to take the average of the areas of the inscribed and the circumscribed squares. What value for π does this method correspond to?

The following exercises are modern generalizations of problems which come from Babylonian clay tablets. You are free to use all your modern algebra and calculus. See [14, pp. 58–59], for the original formulation and more information on these problems.

Exercise 2.2

(a) An Old Babylonian tablet, that is to say a tablet from the period 1900–1600 B.C., the same time as Plimpton 322, poses a problem about a *ladder* standing upright against a wall.
This problem deals with a ladder of known length b stands upright against a wall. The ladder is then allowed to slide down a known distance a. The question is how far out from the wall the lower end of the ladder will be.

(b) A similar problem comes from a much later period, namely the Seleucian epoch, about 300 B.C.–300 A.D. This problem states the following: A reed stands up against a wall, and then slides down a known distance a, which results in the lower end moving out a known distance b from the wall. The question is how long the reed is.

Exercise 2.3 Find the radius of the circumscribed circle of an isosceles triangle with sides b, b and $a > b$. On the tablet $a = 60, b = 50$. For these values, write the answer in the sexagesimal system.

Exercise 2.4 Find the sides x and y of a rectangle, when it is given that $xy = A$ and that $x^3 d = B$, where d is the diagonal. Find the answer when $A = 12$ and $B = 320(= (5)(20))$. Then compute the answer for the values on the tablet,

$A = (20)(0)$, $B = (14)(48)(53)(20)$. For these two sets of values, write the answer in the sexagesimal system.

Exercise 2.5 Find the area A of an isosceles trapezoid with bases a and b and sides s. On the tablet $a = (50), b = (14)$ and $s = (30)$. For these values, write the answer in the sexagesimal system.

Exercise 2.6 One leg of a right triangle is a. A line parallel to the other leg at a distance h from it cuts off a right trapezoid of area A. Find the lengths of the bases of the trapezoid. On the tablet $a = (50), h = (20)$ and $A = (5)(20)$. For these values, write the answer in the sexagesimal system.

Exercise 2.7 An area consisting of the sum of two squares is A. The side of one square is 10 less than $\frac{2}{3}$ of the side of the other square. What are the sides of the square? On the tablet $A = (16)(40)(= 1{,}000)$. For this value, write the answer in the sexagesimal system.

Exercise 2.8 A rectangle has area A and perimeter B. Find the lengths of the sides x and y. Take $A = (1)(40)$, $B = (1)(44)$. For these values, write the answer in the sexagesimal system.

Exercise 2.9 An Old Babylonian tablet, found at Susa, gives the ratio of perimeter and circumference of the circumscribed circle as $(0) \cdot (57)(36)$ for a regular hexagon. Use this to find an approximate value for π, written sexagesimally.

The following two problems are inspired by the Moscow Papyrus.

Exercise 2.10 The area of a rectangle is A, and the width is the fraction $\frac{p}{q}$ of its length. Find the dimensions of the rectangle. Compute the answer when $A = 12$, $p = 3$ and $q = 4$.

Exercise 2.11 The area of a right triangle is A, and one leg is m times the other. Find the dimensions of the rectangle. Compute the answer when $A = 20$ and $m = 2.5$.

The two following exercises are based on information from [58].

Exercise 2.12 The ancient Egyptians computed the area of a triangle and a trapezoid correctly. But the quadrangles were some times treated as follows: Half the sum of two opposite sides was multiplied by half the sum of the other two sides. Is this method correct? If no, when does the method yield a correct answer?

Exercise 2.13 To find the area of a circle, the Egyptians squared the diameter and multiplied by $\frac{8}{9}$. What value for π does this give?

Exercise 2.14 As stated in the text, it has been speculated that the tablet Plimpton 322 might have been used in computations as equivalent to a table over $\cot an(\varphi)$ or $\cos(\varphi)$ for angles φ between $44°46'$ and $31°53'$, or at least perhaps served a similar purpose to such a table. The decrement in the values of $\sec^2(\varphi)$ is very close to $\frac{1}{60}$ from one line to the next. Assuming that Plimpton 322 were part of such a collection, try to compute the 15 numbers the preceding tablet would have contained.

Chapter 3
Greek and Hellenic Geometry

3.1 Early Greek Geometry: Thales of Miletus

The word *geometry* is derived from two Greek words, namely $\gamma\eta$, gē, which means *earth* and $\mu\varepsilon\tau\rho o\nu$, metron, which means *measure*. Our sources on early Greek geometry (Fig. 3.1) – and mathematics in general, for that matter – are sparse. Indeed, as far as mathematical contents is concerned we have to rely on the work of the first serious *historian of mathematics*, namely *Eudemus of Rhodes*, 350–290 B.C.

He was, probably, a student of Aristotle, at any rate a close associate and collaborator of him. But Eudemus of Rhodes should not be confused with *Eudemus of Cyprus*, another philosopher associated with Aristotle. In any case, our Eudemus is known to have written three works on the history of mathematics, namely *The History of Arithmetic, of Geometry* and *of Astronomy.* All three are lost now, but were available to Hellenistic mathematicians and used to the extent that at least some of their contents is known to us today. In particular Eudemus reports, in his History of Astronomy, that *Thales of Miletus* (Fig. 3.2) predicted a solar eclipse, which is presumed to be the one which occurred on May 28, 585 B.C. But most historians of mathematics tend to be skeptical to this claim. The reason for this is that Thales is generally agreed to have been the first Greek astronomer, and that such abilities would have been unlikely at this early stage of Greek astronomy. However, it appears that the most plausible explanation is offered by *van der Waerden* in [58], where he writes:

> *The conclusion is inescapable that he must have drawn upon the experience of Oriental astronomers.*

By the way, the Greek historian *Herodotus* also makes this assertion concerning the prediction by Thales. The solar eclipse occurred during a battle fought between the Lydians, under their King *Alyattes*, and the Medians under their King *Astyages*. The war had been going on for five years, and when the eclipse occurred during an ongoing battle, the belligerent parties found it prudent to end the fighting and make peace. The Gods, evidently, did not approve of what they were doing. Thales had ties to the Lydian kingdom, and when Alyattes' son Croesus later went to war against the Persian King Cyrus II, who had meanwhile conquered the Median kingdom,

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_3,

Fig. 3.1 The School of Athens by Raphael, 1509–1510 fresco, 500 by 770 cm, in Vatican City, Apostolic Palace

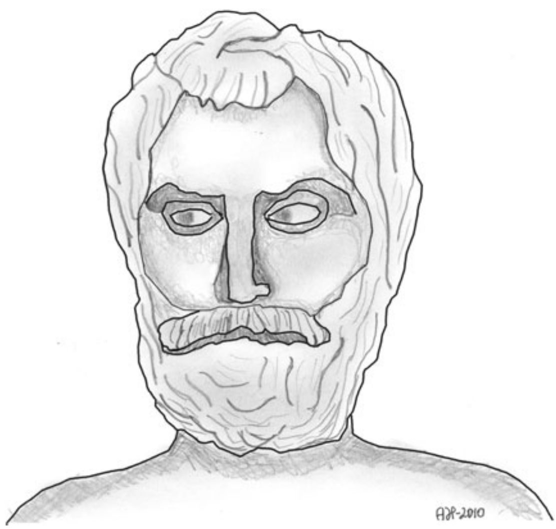


Fig. 3.2 Thales, Father of Greek Geometry. Drawing by the author, following a Greek stamp

Thales went along as an advisor to the Lydian King. Thales is credited with a clever scheme for splitting the river Halys, so that the Lydian troops could pass over.

Eudemus' historical works are lost. But their contents are, to some extent, known through later summaries. The last Greek philosopher and mathematician was *Proclus Diadochus*. He was head of the Neoplatonic Academy in Athens late in the fifth century A.D., one of the last holdouts of classic civilization. At that time Eudemus' books were still extant. As all research towards the end of the classic civilization, Proclus' research is not very original. But as part of his work at the Academy he wrote a summary of Eudemus' History of Geometry, as an introduction to his own *Comments on Euclid's Elements*, Book I. This is essentially the only surviving source on early Greek geometry, frequently referred to as *The Eudemian Summary*. There can be no doubt that Proclus amply deserves a honorable place in the history of geometry and mathematics for preserving this knowledge for posterity. Another important contribution by Proclus was the formulation of Euclid's Fifth Postulate as we state it today, usually referred to as *Playfair's Axiom*. See Sect. 4.1.

Thales is the first Greek mathematician whose name we know. He lived and worked in Miletus, a Greek city in Asia Minor, now in Turkey. He was born about 625 B.C. and died around 545 B.C., in Miletus. We may regard Thales as the *Father of Greek Geometry*. His mother was *Cleobulina*, the first woman philosopher in Greece. Thales referred to her as *The Wise One*.

There are reports that Thales was *of Phoenician descent*, but others refute this by asserting that *"... the majority opinion considered him a true Milesian, and of a distinguished family."* Do we sense a trace of bigotry here? Perhaps the infusion of some Phoenician blood through Thales did the Greeks and their science some real good...

Thales is supposed to have estimated the height of a pyramid in Egypt by measuring its shadow at the time when the shadow cast by himself was equal in length to his own height. Eudemus ascribes to Thales a method for finding the distance between two ships at sea. We do not know exactly what this method was, but van der Waerden in [58] supposes that it might be something like the method described by the Roman surveyor *Marcus Junius Nipsius*, which goes as follows:

> In order to find the distance from A to the inaccessible point B, one erects in the plane a perpendicular AC to AB, of arbitrary length, and determines its mid point D. On C one constructs a line CE perpendicular to CA, in a direction opposite of AB, and one extends it to a point E, collinear with D and B. Then CE has the same length as AB.

The rule is illustrated in Fig. 3.3.

Thales also is credited with discovering that the base angles of isosceles triangles are equal, and that vertical angles are equal. He is also said to have discovered that a diameter of a circle divides it in two equal parts. In what sense Thales "discovered" these geometrical facts is not clear, it does seem reasonable to assume that this knowledge would have predated Thales by perhaps more than a 1,000 years, in Egypt, Mesopotamia, and elsewhere in the East. He may, however, have studied this material, providing some sort of proofs for the above statements.

According to Aristotle, Thales was ridiculed by some Milesians for directing a lot of energy to activities which had no useful applications, and from which he made

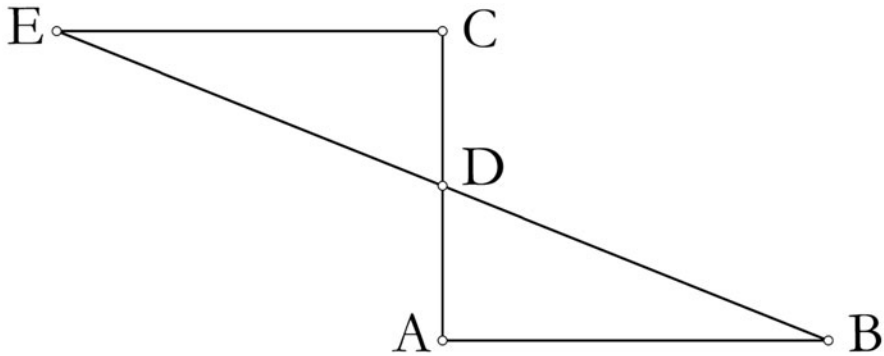


Fig. 3.3 To find the distance to an inaccessible point

no profit. Thales then decided to show them that if he had thought it worthwhile, he could do better than most of them in this regard as well. Thus, noticing signs that a bumper crop of olives was in the comings, he bought up all the presses. When the bumper crop then subsequently *did* materialize, the growers had to buy or rent presses from him, at a substantial price.

3.2 The Story of Pythagoras and the Pythagoreans

Pythagoras of Samos is a rather enigmatic figure. It is frequently asserted in texts on the history of mathematics that we know practically nothing of his life and work prior to the time when he founded the *school of the Pythagoreans* in Croton, at which time Pythagoras may have been in his mid 1950s. We do know however, that he appears at this precise point, and that he undoubtedly possessed extensive knowledge of mathematics in general and geometry in particular. Prior to that time this kind of knowledge is only very sparsely documented in Greece, and all of it comes to us from Thales. But in the East, in Egypt, Mesopotamia, in India and even, perhaps, in the early Indus valley civilization, as well as in China, we find evidence of extensive insights into these matters. Add to this the many stories which are told concerning his travels in Egypt and more widely. We have to realize, however, that for now there is no solid evidence on which the legends of Pythagoras' travels can be accepted as historical facts. So until some new papyrus is found in Egypt, or a tablet uncovered from ancient Babylon, relating the tale of the Greek *visiting priest* at the Temple, we might as well sit back and enjoy the stories. Some of them simply are too good not to be true!

Pythagoras was born about 570 B.C. at Samos, one of the most fertile Greek islands, just off the coast of Asia Minor. It seems to be general agreement that he died in the Greek city of Metapontium, in southern Italy, probably some time during the first decades of the fifth century B.C., one estimate being approximately 480 B.C. At any rate there are reports that he died at the advanced age of 90 (Fig. 3.4).

Fig. 3.4 Pythagoras as a young man. Inspired by the sculpture *Kapitolinischer Pythagoras*, showing him at an older age. Drawing by the author

Some historians of mathematics think that Pythagoras was a student of Thales. Others feel that the age-gap between them makes this unlikely. But with the – admittedly hypothetical – dates of birth and death we have put down, Pythagoras would have been 25 at the time of Thales' death. This does not preclude him having been a student of Thales, but it is probable that Pythagoras at least also had other teachers, working in the same mathematical environment as Thales. In fact, Samos and Miletus were geographically close.

Iamblichus relates in [37] that Pythagoras *"...went to Pherecydes and to Anaximander, the natural philosopher, and also he visited Thales at Miletus. All of these teachers admired his natural endowments and imparted to him their doctrines. Thales, after teaching him such disciplines as he possessed, exhorted his pupil to sail to Egypt and associate with the Memphian and Diospolitan priests of Jupiter by whom he himself had been instructed, giving the assurance that he would thus become the wisest and most divine of men."*

So according to this source, Pythagoras followed in Thales' footsteps. Not only did he take up his geometry, he also made extensive travels in the known civilized world. In Samos *Polycrates* assumed dictatorial powers, but he was in many ways an enlightened ruler, and at least in the beginning Pythagoras may have had good relations with him.

Polycrates had allied himself with Amasis, the King of Egypt. Polycrates was very successful in the beginning, and he established Samos as a naval power, he build temples, harbors and aqueducts and he encouraged art and science including mathematics. Herodotus relates how Polycrates became worried when he received

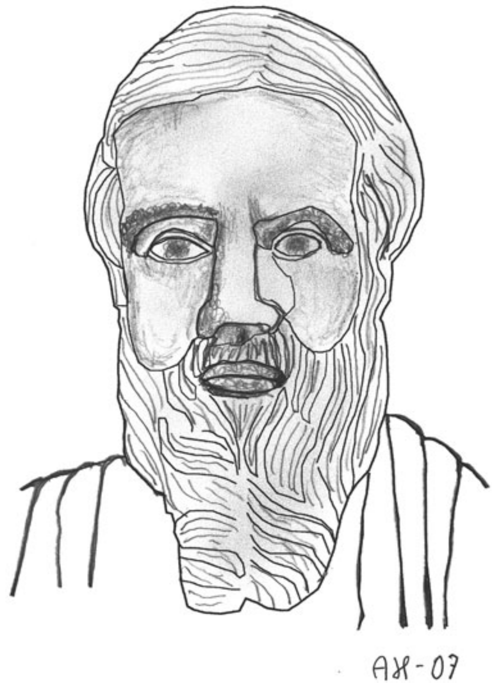


Fig. 3.5 Herodotus of Halicarnassus. *The Father of History.* Drawing by the author

a message from his Egyptian ally, warning him that his good fortune would eventually make the Gods envious, thus bringing some kind of disaster down on him. The advice he gave was for Polycrates to throw away his most valued possession (Fig. 3.5).

The grief this would cause him, should suffice to placate the envious Gods. After thinking about it, Polycrates decided that a precious ring he owned would be a suitable object to loose, and he went out to sea on a boat, where threw his ring into the water. Some days later, however, a local fisherman caught a big fish.

The fish was so extraordinary that the fisherman brought it to Polycrates, expecting to be rewarded lavishly. Polycrates was very pleased, and showed it by inviting the fisherman to his supper, where the fish was to be served. The cook started the preparations and cut the fish open, and in its stomach he found the ring. He brought the ring to Polycrates, who was not exactly overjoyed. When Amasis learned about this, he realized that Polycrates could bring him nothing but bad luck, and cancelled the alliance with him. And in fact, towards the end of his reign Polycrates engaged in some ill-conceived schemes, trying to ally himself with the Persians against the Egyptians. This failed because of mutiny among the men he sent, who with good reason suspected that Polycrates really wanted to get rid of them. He himself was later lured into an ambush by the Persians and suffered a shameful death.

Returning to Pythagoras, he went to Egypt, some say around 535 B.C. Polycrates had supplied him with letters of recommendation, so he could gain access to the Temples there.[1] He visited many temples where he had discussions with the priests. He tried to gain admittance to the Order of the Temples, and finally succeeded when he was admitted into the Temple and Priesthood at Diospolis, near Thebes. Here he stayed for some time, and absorbed their customs and their geometry, as well as their magic and astrology.

But this quiet life was interrupted when there appeared on the scene a Persian King and warlord by the name *Cambyses*. He invaded Egypt in 525 B.C., and capturing Thebes his soldiers came across Pythagoras in the Temple, as Iamblichus relates in [37]. Pythagoras was then taken prisoner by Cambyses, and if this story is true, he must have had some very exiting and interesting years, under Cambyses' rather heavy hand. In the beginning it would not have been too bad, Cambyses himself respected the Egyptians and showed great interest in their traditions and customs. He even had himself designated a *Pharo* under the name *Ramesut*. He also had himself initiated into the priesthood, and if Pythagoras were around this, he might have had something to do with it. In fact, Cambyses' father was *King Cyrus* II or *Cyrus the Great.* He could possibly have met Thales, Pythagoras' mentor, under the following circumstances: According to Herodotus, Thales accompanied King Croesus when he went to war against the Persians under King Cyrus. Croesus lost, and after several dramatic events he was saved from being burned alive on a pyre erected by the victorious Persians. These same events also led him to become a trusted friend and advisor of King Cyrus. This happened in 547 B.C., admittedly late in Thales' life, if not after his death.

Cyrus was one of Persia's great Kings, who went on to capture the marvellous ancient City of Babylon, in 539 B.C. He is the *Cyrus the King* referred to in the Old Testament, who restored the Jews to Palestine and ordered the Temple of Jerusalem to be rebuild. Unfortunately for him, however, he did not rest on his laurels.

Instead, he marched with his troops across *the Araxes*, the river now named *Araks* which flows east to the Caspian Sea. He went against the *Massagetic queen Tomyris*, she ruled over a kingdom in that area (Fig. 3.6). His advisor Croesus was with him, and the crossing of the Araxes was undertaken on his advice. This was a disastrous move. Cyrus had success initially, and managed to trick Tomyris' son, who was her leading general, into a trap in the following manner: Cyrus withdrew from his camp, leaving behind large quantities of delicious prepared food and strong wine. When the Massagetic troops arrived, they thought Cyrus had fled with his men, and started the celebration. Early in the morning Cyrus attacked, and the drunken troops were an easy match. Many were slain, and Tomyris' son captured. When he was brought before Cyrus, he begged him to remove his chains. As Cyrus complied, the prince grabbed a sword from one of the guards and killed himself. When Tomyris got the message from Cyrus, she sent back her answer: *"Bloodthirsty Cyrus! You won in a most dishonorable way! But be sure, I shall satiate your thirst for blood!"* Then she

[1] Some say this, others claim that Pythagoras feared Polycrates, and fled because of him.

Fig. 3.6 Tomyris plunges the head of the slain Cyrus into a sack of blood

attacked, ferociously. The battle was long and hard, but in the end the Persians were defeated and Cyrus slain. And then she kept her promise. This battle took place in 529 B.C.

Then Cyrus' son, Cambyses II, succeeded him on the Persian throne. On his fathers advise, he retained Croesus as an aide and advisor, in spite of the sad outcome of his last service to his father. And Croesus accompanied Cambyses to Egypt. Thus Pythagoras and Cambyses' aide would have some points of contact.

At any rate, the good state of affairs for Pythagoras in Egypt did not last. Cambyses continued his military expansion, and now he met with some very serious, humiliating setbacks and defeats. Without going into details, let us just relate that he turned into a paranoid man, suspicious of everything. When he arrived back from one of his ill-fated expeditions, his troops decimated and starved, having been reduced to cannibalism, he unfortunately came just in time for a big celebration in Memphis. Feeling that the people rejoiced because of his own misfortune, he ordered the leading citizen rounded up and executed. The most repulsive incident occurred when it was explained to him that the celebration was on occasion of the appearance of a very special calf, the latest incarnation of the God Apis. On his orders the calf was brought into his presence. Cambyses, in a fit of senseless rage, grabbed his sword and dealt the Holy Calf a powerful blow, wounding it in the thigh, in front of all the terrified Egyptians. The Holy Calf fell to the ground, and it died some time afterwards from the infected wound.

He also committed various other acts of sacrilege, like several instances of outrageous profanation of temples, killings of priests, he broke up ancient tombs and examined the bodies, burned them in some cases, and so on.

Matters worsened. Cambyses appears to have gone completely mad. According to Herodotus one of the misdeeds he committed was to have his own brother, *Smerdis*,[2] murdered. Smerdis had been a member of his Egyptian expedition, but Cambyses had sent him back to Persia because of jealousness caused by his brother's physical strength. Some time after Smerdis' return, Cambyses had a dream which caused him great worry: He dreamt that a messenger arrived from Persia, telling him that Smerdis was sitting on the royal throne and that his head was touching the sky. Interpreting this to mean that his brother would kill him and seize the throne of Persia, Cambyses sent his most trusted Persian friend *Prexaspes* back to Persia to do away with Smerdis. Prexaspes dutifully did what he had been ordered. And then he informed the people that His Royal Highness the Prince spent all his time in seclusion at the palace, praying for the success of his brother the King during his campaign abroad. Cambyses later rewarded him for his services by murdering his son in front of his very eyes, in order to prove his marksmanship with bow and arrow and ability to hold his liquor.

Now Herodotus relates that Cambyses had left the control of his household with a man who belonged to the caste of *the Magis*, his name was *Patizeites*. Patizeites had a brother, named Smerdis, like the prince. This brother also looked like the murdered prince, and as Patizeites knew of Cambyses' foul deed regarding his brother, he hatched a rather obvious plan: He had his own brother usurp the throne, claiming to be Cambyses' brother![3]

The Magis constituted the hereditary caste of priests among the ancient Persians. They interpreted dreams and performed sacred rituals, being devoted to the Gods. In the New Testament the astrologers who divine the birth of the King of the Jews by the appearance of a star in the East are called *Magis*. The priests of Babylonia are also frequently called *Magis*, and of course the term is preserved today in our word *magic*.

Heralds were sent out proclaiming the change of regent, and one of them happened to encounter Cambyses and his men in Ecbatana in Syria. When brought before the rightful, if incompetent, King, the herald was questioned about the situation. Cambyses suspected that Patizeites had double-crossed him, but the latter had the explanation ready: *"I think, my Lord, that I know what happened. The rebels are the two Magi brothers you left in charge of your household. One of the brothers*

---

[2] According to the Persian sources Cambyses murdered a brother by the name *Bardiya*.

[3] Persian sources give the name of the Magian usurper, or pretender, as *Gaumata*. Thus there is no homonymy in the Persian version of this story, as well as other discrepancies with the account as given by Herodotus. It is generally accepted among historians that Herodotus' version of the story is far from accurate. A political intervention by priests of the temples in the face of a ruler who was obviously incompetent and mentally disturbed, as well as a political rivalry between Medes and Persians with economic and social ramifications, has undoubtedly taken place. But the details are lost today.

is named Smerdis, as you may recall." Cambyses now realized the true meaning of his dream. The Smerdis on the throne was really Smerdis the Magi! The murder of his brother had served no purpose, in fact it had made the prophesy of the dream come true, rather than preventing it from happening. As sanity started to return, he understood the depths to which he had fallen, and he bitterly lamented the abysmal situation in which he found himself. Finally he resolved to march back to Persia at once, to attack the Magi. But as he leaped into the saddle, the cap fell off the sheath of his sword. The exposed blade cut his tight, at the very spot where he had struck the sacred Egyptian Bull of Apis. Cambyses now felt that he was mortally wounded, and asked his men for the name of the town they were in. Being told that the name was Ecbatana, he realized the true meaning of a prophecy from the oracle at Buto (in Egypt, now Tell el-Farein): Namely, that he should die at Ecbatana. He had thought this to be *Median Ecbatana*, his capital city, and that he should therefore die at home of old age. Now he realized that the oracle meant Ecbatana in Syria. At this point sanity fully returned to Cambyses, and he said no more. After 20 days he called the leading Persians together, and explained the situation to them. In tears he bitterly lamented his cruel fate, and the Persians tore their cloths, crying and groaning. Shortly after, gangrene and mortification of the thigh set in, and Cambyses died.

However, his men really did not believe him. They suspected another malicious lie, to set the country against his brother Smerdis.

Thus no obvious course of action seemed to present itself, and about one year of political strife followed in Persia, with the Magi on the throne. Prexaspes originally decided to side with the Magis, out of fear for punishment and also his bad feelings towards the house of Cyrus and Cambyses. Thus he changed his story about having murdering Smerdis the Prince. The Magi rule ended when a young and ambitious nobleman by the name *Darius*, himself of royal descent, headed a successful *coup d'etat*. Prexaspes, repenting his treason to the Persian cause (the Magi were originally a Median caste), confessed his crime to an assembled crowd from the main tower, and then leaped to his death.[4] Darius then assumed power, to become the famous Darius I, *Darius the Great.*

The story of the false Smerdis, the usurpation of power by the Magis and finally the accession of Darius plays an important role in the history of mathematics, at least indirectly. In fact, the Persian version of it, as told to us by Darius himself, forms part of the inscription at Behistun, described in Sect. 2.1, and thus provided the basis for Rawlinson's decipherment of the cuneiform script. This again led to our present insights into the mathematics in Mesopotamia, of the Sumerians, Assyrians and the Babylonians. As already noted, the inscription by Darius himself differs considerably from the tale as told by Herodotus. For more details, see note 25 on page 571 in [28].[5]

---

[4] Still according to Herodotus, the Persian story runs differently. There is no character by the name *Prexaspes* in that version.

[5] Actually Herodotus is a dubious historical source, although an entertaining one. He relates what he was told by the priests in Egyptian temples, and they did not like Cambyses. One of the reasons for this might have been the heavy taxation which Cambyses had subjected them to.

Pythagoras, however, had been brought to Babylon by Cambyses' troops. At least so the story goes. The political situation in the Persian Empire being somewhat murky, he sought refuge in the Temple, where he was once more initiated into the Priesthood. *Iamblichus* writes as follows in [37], in the fourth century A.D.:

"Here the Magi instructed him in their venerable knowledge and he arrived at the summit of arithmetic, music and other disciplines. After 25 years he returned to Samos, being then about 56 years of age."

There are some ancient busts claiming to show what Pythagoras may have looked like. One is a bronze copy of an original believed to be from the fourth century B.C., which is displayed at Villa dei Papiri in Herculaneum, Museo Nazionale, Neapels. Here Pythagoras is shown wearing turban and oriental dress, absolutely compatible with our story. A photo of the bust is shown in [33] and in [58].

Iamblichus has Pythagoras' stay in Egypt to last for 22 years, plus 12 years in Babylon, altogether 34 years abroad. At any rate he spent many years in Egypt and in Babylon, working and learning in the temples.

Cambyses had died in 522 B.C., and Polycrates, the tyrant of Samos, was killed by the Persians about the same time. King Darius I took over in 521 B.C., and after Polycrates death Samos came under his rule. Exactly when Pythagoras returned to Samos is uncertain. Some say that he returned at a time when Polycrates was still alive and in power, others assert that he returned at a time when Samos had fallen under Persian rule. In any case, after the fall of the Magi from power, it would seem to make sense for Pythagoras to leave Babylon, since he presumably had close ties with that group.

Iamblichus reports that Pythagoras formed a school in the city of Samos, called *the semicircle*. He also reports that Pythagoras made a cave outside the city, where he did his teaching, and spent both nights and days doing research in mathematics. But then Iamblichus goes on to tell how Pythagoras attempted to employ the same didactical principles he had learned in the temples of Egypt and Babylon, to teaching the Samians. This did not work too well, they found his teachings too abstract and symbolic. Pythagoras did not like such attitudes any better than some present day college professors do, and decided to leave. At least this is the reason Pythagoras himself is supposed to have given for leaving Samos.

Actually, the Samians were by no means ignorant of geometry. Herodotus relates how they constructed, at the order of Polycrates, an aqueduct for bringing drinking water to the capital city by the same name as the island. They had to dig a tunnel through a mountain, and started to dig at both ends simultaneously. And in fact, they met in the middle of the mountain with remarkable accuracy! The direction of the tunnel had to be found by reasoning with similar triangles. Also a fairly sophisticated use of a diopter had to be employed. *Heron of Alexandria* explains the method in his work *Dioptra*, about 600 years later, around 60 A.D. For details, see [33] or [58]. The engineers, some of them quite possibly being *slaves*, who worked on the tunnel at Samos certainly knew quite sophisticated geometry. But this knowledge was part of their practical work in the field, not necessarily as an object of the *"refined contemplation"* considered worthy of *free men*.

So Pythagoras left for Croton, a Greek city on the coast of southern Italy. Here he formed his school or brotherhood, *The Pythagoreans*. The society consisted of an *inner circle*, whose members were called *mathematikoi*, and an outer circle whose members were known as *the akousmatics*.

The mathematikoi lived permanently with the Society, they had no personal belongings, were vegetarians and practiced celibacy, did not eat *beans*, and did not wear cloths made of animal skin. Presumably this was the way of life Pythagoras had picked up at the temples in Egypt, although *Herodotus* does report on ample supplies of meat and wine for the Egyptian priests.

It should be noted that there are marked similarities between the practices of the Pythagoreans and those associated with *the Orphic Cult*. Orpheus of Thrace was the founder of this cult. He played so divinely on the lyre that all nature stopped to listen. When his wife *Eurydice* died, he went to the nether world, to *Hades*, to bring her back. By the music from his lyre he succeeded in obtaining her release, but on the condition that he would not look at her until they were clear of the world of death. However, he could not bear to refrain from looking, and she had to return to Hades for good.

The akousmatics, however, were allowed to live normal lives. Both men and women were allowed to be Pythagoreans, and there are some reports of women Pythagoreans who became well known mathematicians and philosophers.

There are accounts to the effect that *Pythagoras had a wife*. Her existence would seem to contradict the claimed practice of celibacy, but this particular kind of contradiction should not disturb historians too much. Her name was *Theano*, and she had three daughters with Pythagoras. Together with them she is said to have continued Pythagoras' school after his death. Her most important mathematical work is supposed to have been a treatise on *the Golden Section.* We refer to [40, 63] and to [59]. As far as this author's information goes, this is the first known, or claimed, individual name of a woman mathematician. Pythagoras' three daughters also were Pythagoreans. *Damo* is said to have been entrusted the responsibility for her fathers works, which she refused to sell and therefore had to live in poverty. The two other daughters *Arignote* and *Miyia* were also Pythagoreans, and are credited with several works on a variety of subjects. Other women Pythagoreans were *Themistoclea*, priestess of Apollo at Delphi and said to be Pythagoras' sister, and *Melissa*, thought to have been one of the very first Pythagoreans.

The Pythagoreans were in opposition to the *democratic* movement in Greece. The followers of the philosophical school of the *Sophists* were democrats, while the Pythagoreans believed in *oligarchy*, the rule by a small political elite. Some of the Greek geometers did in fact belong to the democrats. They did not get along too well with the main stream Pythagoreans, who were very influential. Thus for example *Hippasus of Metapontium*, who was a Pythagorean, and nevertheless democrat, made known the findings that not all line segments have a common measure, that there are *incommensurable* line segments. We say more about this below, but the Pythagoreans did not take lightly to this breach of secrecy! In fact, he was severely denounced for having described the *Sphere of the Twelve Pentagons*, in other words the *dodecahedron* and for having revealed *the nature of the non-mensurable* to *the Unworthy*.

To the Pythagoreans the regular pentagon with the inscribed pentagram, the 5-pointed star formed by all the diagonals, was a sacred symbol. There is a story about a Pythagorean who became seriously ill while travelling, far from home. The keeper of the inn where he stayed was a compassionate man, and had his servants nurse him as best they could. The money of our travelling Pythagorean expended, he was reduced to nothing: Seriously ill, at the mercy of foreigners, far from home. Nevertheless the inn-keeper stood by him, providing for him at his own expense. As the unfortunate Pythagorean realized that his *Earthly Goal* for the present incarnation was approaching, he called for his benefactor. Not being able to leave behind any significant earthly values, he told him to paint the symbol of the pentagon with the inscribed pentagram on his door, *but to paint it right, not upside down.* If ever a Pythagorean came this way again, he would generously return the favor. And so the man did, after his foreign guest had passed away. Not that he had much belief in the benefits to be reaped from this undertaking. But years later a rich Pythagorean travelled through the area, saw the pentagon with the inscribed pentagram, and did indeed repay the local good Samaritan generously.

Returning to Hippasus, his treasonous publication may have happened towards the end of Pythagoras' life, maybe after his death. Hippasus was expelled from the Brotherhood, and one version of what happened afterwards is this: The Pythagoreans made a grave monument for him, as he was to be considered dead. Soon afterwards he perished at sea, and this was seen as punishment from the Gods: *He died as a godless person at sea.* Another version of the story is that he was murdered by Pythagoreans, who threw him overboard from a ship at sea. Be this as it may, during this time the opposition to the Pythagoreans grew, Pythagoras himself had to move from Croton to Metapontium. A prominent citizen of Croton by the name *Cylon* is said to have been refused entry into the Pythagorean Brotherhood by Pythagoras, presumably because he was lacking in the spiritual qualities required, and as a result the same Cylon mobilized his followers against Pythagoras and the Pythagoreans. Others report the events differently, but at any rate Pythagoras had to move to Metapontium, not too far from Croton, as the situation became difficult. He died in Metapontium soon afterwards. According to some accounts he was murdered, killed by arson at the house of his daughter Damo.

In Croton the Pythagoreans continued to exist as an organization, but increasingly surrounded by controversy. Finally mobs emanating from the democratic party killed a large number of Pythagoreans when they set fire to the house in which they were assembled, the house of an athlete named *Milo*, a famous wrestler.

As many as 50 or 60 Pythagoreans are said to have been killed at that time. The surviving Pythagoreans fled from Croton, and thus, ironically, the ideas of Pythagoras were spread more widely in the Greek domain. Later still the Pythagoreans reappeared in the area, the last important of them being *Archytas of Tarentum,* 438 –365 B.C. (Fig. 3.7). His best known work is probably an ingenious 3-dimensional construction which solves the problem of Doubling the Cube. We shall explain this in Sect. 3.11.

Again we should reiterate the warning that the story of Pythagoras' life which we have told here is regarded by some as being highly unreliable. Contradicting ones

Fig. 3.7 Archytas of Tarentum. Drawing by the author

are in circulation as well. The indisputable fact, however, is that these stories and legends about him do exist, and have been told for 2,500 years.

3.3 The Geometry of the Pythagoreans

No work by Pythagoras is extant, and in fact the practice of the early Pythagoreans was to ascribe all their findings to the master himself, to Pythagoras. But it is well documented from later sources that the Pythagoreans viewed mathematics as basic to the very fabric of reality, and that certain fundamental doctrines were important to their thinking and teaching. One such doctrine was that *numbers*, that is to say, the natural numbers, formed the basic organizing principle for everything. The motion of the planets could be expressed by ratios of numbers. Musical harmonies could be expressed so as well. The right angle was fixed by ratios like 3:4:5, as a triangle with sides in these proportions is a right triangle.

This takes us to the *geometry* of the Pythagoreans. Several discoveries have traditionally been attributed to the Pythagoreans, but at least some of them are without question of a much earlier origin. We reproduce a list of such discoveries in geometry, together with some comments. See [26] and [60].

1. The Pythagoreans knew that the sum of the angles of a triangle is equal to two right angles. They also knew the generalization to any polygon, namely, that in any n-gon the sum of all the interior angles is equal to $2n-4$ right angles, while the sum of all exterior angles is equal to four right angles.

The last assertion may be viewed as completely obvious, as far as the mathematical realities are concerned. As for the first, that the sum of the angles in a triangle equals two right angles, Egyptian, Babylonian, Chinese and Indian geometers knew well the properties of similar triangles. It is, therefore, hard to believe that the realities behind such properties of triangles were not known before the Pythagoreans. However, the precise formulation as a mathematical proposition, as well as a formal proof may well have been first supplied by them.

2. The Pythagoreans knew that in a right triangle the square on the hypothenuse is equal to the sum of the squares on the two sides containing the right angle.

This theorem, the so-called *Pythagorean Theorem*, was certainly known to the Babylonians at least 1,000 years before Pythagoras. As we have seen, not only did the Babylonians know this, they also knew how to generate *all the so called Pythagorean triples*, namely triples (a, b, c) of integers such that $a^2 + b^2 = c^2$. Whether the Babylonians also knew proofs of the Pythagorean Theorem is more hypothetical. But proofs based on a simple figure combined with some algebraic manipulation could well have been known to the Babylonians, who were superb algebraists.

3. The Pythagoreans knew several types of constructions by straightedge and compass of figures of a given area. They also solved what we would call algebraic problems by *geometric means.*

Again, much of this would be known long before the Pythagoreans. Thus for instance *the Sulva Sutra*, the oldest source of Indian mathematics, contains rules for constructing altars of a given area. Typical assignments would include the following:

(1) *Construct a square altar table, the area of which is twice that of a given square alter table.* Solution[6]: Use the diagonal of the given one as the length of the sides of the new one. We will return to this assignment in Sect. 3.8.

(2) *Given a rectangular altar table. Construct a square one of the same area.* Solution: Let the sides in the rectangular table be a and b, the unknown side of the square be x. Then $x^2 = ab$, thus $a : x = x : b$, in other words, x is the *mean proportional* of a and b. We then draw a half circle of diameter $a + b$, erect a line normal to this diagonal where a is joined to b, and find x as the half-cord. See Fig. 3.8.

By the way, we may also use (2) to solve (1), of course. But the first method is simpler.

Finally we come to a discovery which is universally credited to the Pythagoreans, if not to Pythagoras himself. There are some who think that the discovery was made by a woman mathematician, *Theano*, who was Pythagoras' wife. It is arguably one of the most profound piece of mathematics discovered by the Greek classical school,

[6] The author does not claim that these problems are solved in this manner in the Sulva Sutra. What the typical arguments there would be like is a different matter, which we will not pursue here. In the final analysis, however, the mathematics involved would have to be the same.

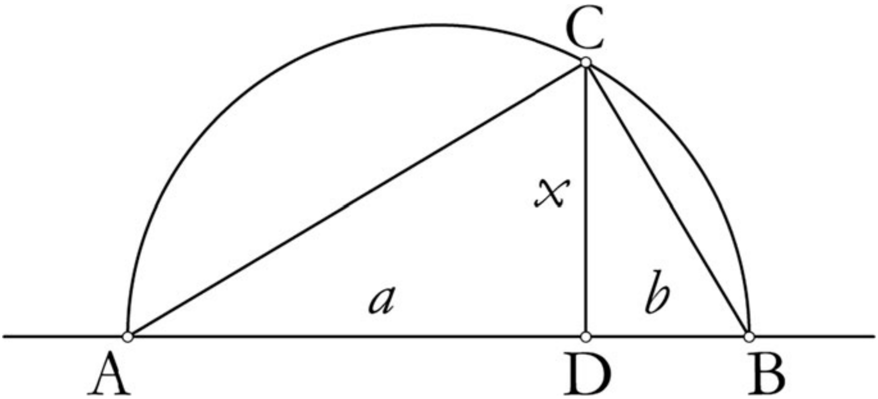


Fig. 3.8 Construction of the mean proportional

and brought the Greeks almost to the point of discovering the system, or *the field,* of real numbers, as we would say in modern language.

But somehow the decisive last step was never taken, and the discovery of the field of real numbers as a powerful extension of the rationals would have to wait for about 2,000 years. Perhaps one of the reasons for this was that the Greeks did not possess any good algebraic notation. Only towards the end of the Hellenistic epoch do we see a movement in this direction, in the work of *Apollonius*. Also, the Greeks were really *true geometers*, and not algebraists. They considered geometry to be a more complete science than algebra, in fact they did their "algebra" in terms of geometry, we would call it *Geometric Algebra*. Perhaps it was this philosophical prejudice which prevented them from taking the last definitive step and discovering the system of real numbers as an extension of the rationals. But even to say that the Greeks worked with rational numbers, is somewhat misleading. To them, what we would understand as the number $\frac{3}{2} = 1.5$ would be the proportion $3 : 2$.

However, when this is said it has to be added that some historians of mathematics seem to have underestimated the sophistication and power of Greek computing abilities. Especially towards the end of the Hellenistic Epoch such abilities to an impressive degree are documented in the work of *Claudius Ptolemy* and others. See Sect. 4.14.

3.4 The Discovery of Irrational Numbers

Presumably the Pythagoreans would early on work from the assumption that given any two line segments a and b, then their *proportion* $a : b$ would always be equal[7] to the proportion between two *numbers*, i.e., in our present language be equal to a

[7] The Greek concept of equality for proportions will be explained below.

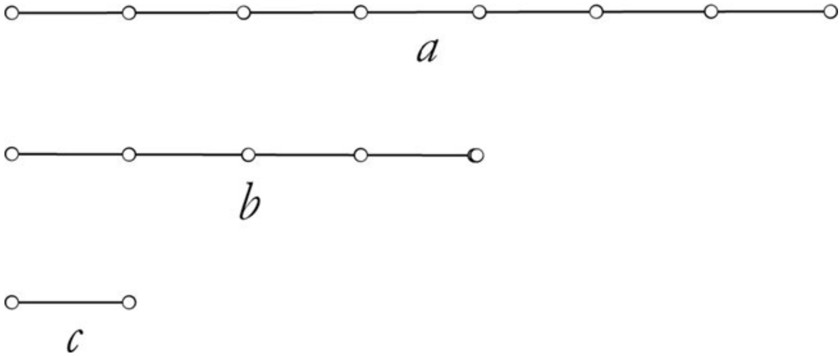


Fig. 3.9 c is the largest common measure of a and b

fraction $\frac{r}{s}$ where r and s are positive integers. Arguably, this would be the position taken by Pythagoras himself, at least originally. Of course at this time many Greek philosophers espoused the *atomistic* view of the physical world. According to this idea, all things are made up of *incredibly many*, but a finite number, of *incredibly small*, but of a definite size, *indivisible atoms.* In fact, this *model* for the physical world became generally accepted all the way up to our own times. Some of the early Pythagoreans applied this idea to geometry and mathematics as well. For *numbers* they had the atom in the number 1, from which all other numbers were built.

In accordance with this general way of thinking, *lines* would consist of small chained *line elements*. In particular two line segments a and b would have *a common measure*: There would exist some line segment c such that c would fit exactly an integral number of times, say r, in a, and exactly an integral number of times, say s, in b: Of course this would be true, at the very worst one would have to take *one of the minuscule line elements*, which would work since the two line segments were made up of whole numbers of such line elements. The line element would *always* constitute a common measure, for any two line segments. Now, for convenience one would let c be the largest such *common measure*. This situation is illustrated in Fig. 3.9.

How would we go about finding the biggest common measure of two given line segments a and b? The procedure is an ancient method, which the Greeks called *antanairesis*, meaning successive subtractions. Literally, given the two line segments a and b, the smallest is subtracted from the biggest. Of the remaining, the smallest is again subtracted from the biggest. This subtraction-procedure is repeated again and again, until the two segments are *equal in length*. Note that if you believe in the *atomistic nature of lines*, then this will occur sooner or later, at the very worst when you are left with two line-atoms, two line elements discussed above. Then a moment of contemplation will convince you that these two equal line segments are indeed the greatest common measure of the original line segments a and b (Fig. 3.9).

This method of *successive subtractions* was very useful in ancient times. It allowed amazingly exact mensuration of an unknown distance, using only a measuring rod without subdivisions, and a good sized compass. It is no accident that the Master Builder so frequently is depicted with the measuring rod and the compass! He would proceed as follows. Let's say that the measuring rod would be, anachronistically, one meter long. First, as carefully as possible he would count the number of times the whole measuring rod could be subtracted from the unknown distance, i.e., find the number of whole meters. Let's say he gets 50. Then he would take the residue, the left over piece, in his compass, and count the number of times *it* could be subtracted from the length of the *measuring rod* itself. Let's say he gets 2, and a new left over piece, a new residue. He now successively repeats the procedure, counting the number of times the new residue can be subtracted from the previous one, and writing down the numbers. Let us say he repeats this 4 more times, getting 1, 1, 4 and 2, at which point there is nothing left, at least as far as he can see: Then, of course, he has to stop. Denoting the length to be measured by L, the measuring rod (here of one meter) by m, the first residue by r_1, the second by r_2, then r_3 and finally, r_4, we obtain

$$\begin{aligned} L &= 50m + r_1 \\ m &= 2r_1 + r_2 \\ r_1 &= r_2 + r_3 \\ r_2 &= r_3 + r_4 \\ r_3 &= 4r_4 + r_5 \\ r_4 &= 2r_5. \end{aligned}$$

To find L in terms of m, we substitute $r_5 = \frac{1}{2}r_4$ from the sixth relation into the fifth relation, obtaining $r_4 = (\frac{1}{4+\frac{1}{2}})r_3$, which substituted into the fourth yields

$$r_3 = \left(\frac{1}{1 + \frac{1}{4+\frac{1}{2}}} \right) r_2$$

and so on, until we finally get

$$L = \left(50 + \cfrac{1}{2 + \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{4+\frac{1}{2}}}}} \right) m.$$

As m is supposed to be one meter, we find after some computing of fractions that the length L is $50\frac{20}{51}$ meters, or 50.39 m, in present days decimal notation. Of course the number is deceptive, as counting the 50 m to begin with could introduce an error of around 5 cm. But using a longer rod, or a longer string of a known length, this measuring error would be reduced.

Now, it is generally thought that the first “irrational number”, discovered by the Pythagoreans, was $\sqrt{2}$. But first of all, the Pythagoreans, as indeed all Greek mathematicians of this time, did not think of this as a *number*. Rather, it was a question about the *proportion* between the lengths of two line segments *not being equal to the proportion of two numbers*, we would say not being a rational number, the fraction of two integers. It is presumed, by some, that the first such pair of line segments found was the diagonal and the side of a square. It is also asserted, frequently, that the so called Pythagorean Theorem should have been essential in realizing this. Others find this questionable. First of all, at the time of Pythagoras proving that two line segments are *incommensurable* would consist in showing that the *process of repeated subtraction applied to these two particular line segments never stops*. Later more sophisticated methods were developed by geometers like Theodorus of Cyrene, (465–398 B.C.), pupil of Pythagoras and teacher of Plato, and by Theaetetus. They are the principal characters in two of Plato’s famous dialogues, one of them dealing with square roots.

At Pythagoras’ time the simplest case to consider would be *the diagonal and the side of the regular pentagon*. This certainly appears surprising, since we would view the regular pentagon as considerably more complicated than a square. But from the point of view of repeated subtraction of the side and the diagonal it is the absolutely simplest figure in existence. A look at Fig. 3.10 will explain this.

Indeed, the diagonal is AC and the side is AB. Now AB = AD, as elementary considerations yield the equality of the angles ∠ABD = ∠ADB. Thus subtracting AB from AC we are left with DC, and the subtraction can only be performed once. In the next step CD is to be subtracted from AB. Now CD = AD’ and AB = AD, thus in this next step we may also only subtract once, and the remainder is D’D. But as CD = CE = ED’, the third step will be to subtract the side of *the inner pentagon from the diagonal of the inner pentagon*! Thus, magnifying the inner pentagon and turning it upside down, we are back to the starting point. Hence the process evidently repeats itself without ever stopping. Thus the incommensurability of diagonal and side of the regular pentagon is proven.

A similar procedure may be carried out for the diagonal and the side of a square, but it is considerably more complicated. And in view of the special relationship the Pythagoreans had to the regular pentagon, it is a very plausible guess that this is how they arrived at the conclusion that *not all line segments are commensurable.*

A final point to be made is this: If we put x = AC:AB, then we obtain

$$x = 1 + \frac{1}{1 + \frac{1}{1+\cdots}} = 1 + \frac{1}{x},$$

which yields the equation

$$x^2 - x - 1 = 0,$$

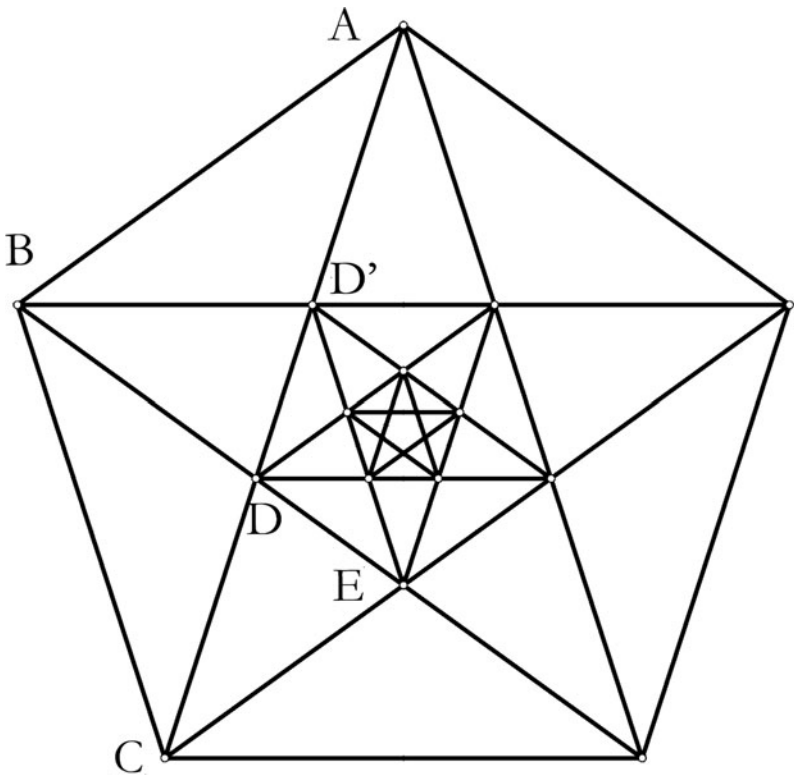


Fig. 3.10 The pentagon and the pentagram

Indeed, this follows in the same manner as the computation carried out on p. 48. Hence $x = \frac{1}{2}(1 + \sqrt{5}) \approx 1.6180$. This number is often referred to as the *Golden Section*.[8]

3.5 Origin of the Classical Problems

There are three problems occupying a special position in Greek geometry, namely the so-called *classical problems*. They are all insoluble in their strictest interpretation. However, they may be solved by various creative procedures and they have generated an enormous amount of mathematics. Their attraction on mathematical amateurs is perhaps paralleled only by the famous *Fermat Conjecture*, which was finally proven not too many years ago by *Andrew Wiles*. The first of these problems

[8] Other names include the Golden Mean, the Golden Number and the Golden Ratio.

Fig. 3.11 Anaxagoras. Drawing by the author

we encounter in the history of mathematics is the problem of *Squaring the Circle.* The problem is the following: Let there be given any circle. Then construct a square with the same area as the one enclosed by the circle, using ruler and compass.

One of the first time we find this problem mentioned, is in connection with the Greek philosopher *Anaxagoras* (Fig. 3.11). Anaxagoras lived at a time when Athens stood at the summit of its power, politically and intellectually.

After Athens and Sparta had won the protracted war against Persian invaders, there followed half a century of peace and prosperity. This was a time of flourishing cultural life in Athens. Many of the expelled Pythagoreans found their way to Athens, and *Socrates* played an important role in the intellectual life of the city state.

Athens had an enlightened leader in *Pericles* for a great part of this time, from about 460 B.C. until he died in the great plague in the year 429 B.C., two years after the peace had been broken and the devastating Peloponnesian war with Sparta had broken out. Unfortunately Pericles must bear a large part of the responsibility for this fratricidal struggle. In fact, he transformed the alliance of the Greek cities against the Persians, *the Delian Alliance*, into an instrument for Athenian dominance. In particular the treasure of the Alliance was moved from Delos to Athens. This worked fine for Athens, but the Spartans and other allies were not pleased. Athens now had more than 3,00,000 inhabitants, one third were slaves and about 40,000 were male citizens enjoying full rights. The city wall also enclosed the port city of *Piraeus*, and their fleet was the dominating power at sea.

Pericles erected the magnificent buildings at Acropolis, and showed great interest in mathematics and philosophy. He belonged to *the democrats*, from the aristocratic wing of the party. He was succeeded by *Cleon* when he died, also a democrat but from the less aristocratic wing.

Pericles' teacher and close friend was *Anaxagoras.* Anaxagoras was born about 500 B.C., in Clazomenae (now Izmir), in Ionia, presently Turkey. He died 428 B.C. in Lampsacus in the Troad, where he had sought refuge for persecution by his enemies in Athens, who continued to press charges for *impiety* against him.

He was more a natural philosopher than a mathematician. Nevertheless he played an important role in Greek geometry, and indeed in the development of mathematics, since he was, apparently, the first to be tied to one of the great problems of antiquity, *the Squaring of the Circle*.

In his teachings, he had denied that the heavenly bodies were divinities. Instead, he explained them as stones torn from the Earth, the Sun being red hot from its motion. The sun was as big as all of the Peloponnes, he asserted, and the moon reflected the light from the sun. The moon was an inhabited world, like the earth, according to Anaxagoras.

These ideas were not accepted, any right-thinking Athenian would be disgusted at such impiety. Consequently Anaxagoras was incarcerated. According to *Plutarchus* Anaxagoras spent the time in prison by attempting to square the circle.

Pericles had to be cautious, since he had many powerful enemies in Athens. But he also stood by his friends, and he finally managed to get Anaxagoras out of prison. But Athens certainly was not a safe place for him any more, and he therefore moved to Lampsacus where he founded his own Academy. Aristotle speaks highly of the reputation he enjoyed there.

The Peloponnesian War broke out in 431 B.C., and two and a half years later Pericles died in the great plague which had started to ravage Athens. One year later Anaxagoras also died.

To the left is shown the latest and largest of the three temples dedicated to Apollo at his birthplace at Delos.

The plague had broken out for full in 427 B.C., the presumed year of Plato's birth. The plague weakened Athens considerably, one fourth of its population is said to have perished. According to the legend, the citizens of Athens sent a delegation to the oracle of Apollo at Delos, to ask for advice on how to deal with their difficulties: A war with Sparta which would be very difficult to win, now that they also had to suffer from this debilitating pest.

The answer delivered by the priestess of Apollo was enigmatic: *The cubic Altar of Apollo should be doubled.* They may also have received other instructions as well, since Athens carried out extensive *purifications* of the island in the year 426 B.C.: Among other things all graves on the island were opened, and the remains which were buried there removed and reburied on the neighboring island of Rheneia. Doubling the cubic altar proved more difficult.

The Greek geometers realized of course that the purest, and most pleasing way to Apollo, would be using compass and straightedge. In other words to perform a geometric construction which for a given cube would render another with volume twice the given:

> *Doubling the Cube, or the Delian Problem.* Given any cube, construct with straightedge and compass the side of another cube, the volume of which is twice that of the given one.

It must have been quite intriguing to the geometers in Athens that this problem proved so hard, since the corresponding assignment for *a square* was so easy. More on that below, in Sect. 3.8. At a later time Eratosthenes from Alexandria wrote a dialogue known today as *Platonicus*. The original is lost. But there are accounts to

Fig. 3.12 Plato, 427–347 B.C. Drawing by the author

the effect that Eratosthenes writes that *Plato* (Fig. 3.12), when consulted about the problem later, voiced the opinion that Apollo had not offered the oracle because he wanted his altar doubled, but that he had intended to censure the Greeks for neglecting mathematics and geometry: By paying more attention to science and philosophy instead of making war, things would start to go better for them.

Greek geometers were fully aware that all circles are similar, as are all squares and cubes. Thus the problems stated above for *any* circle and for *any* cube is equivalent to the same problem stated for *one* circle or for *one* cube: If you can square one circle you can square them all, if you can double one cube, then you can double them all. Not so with the *third problem*, which also circulated in Athens about this time:

> *Trisecting the Angle.* Given any angle, divide it in three equal parts using straightedge and compass.

In this last case the situation is different: There is an infinite number of angles which may be *trisected* using ruler and compass. We show the construction for *a right angle*, that is to say an angle of $\frac{\pi}{2}$ radians or 90°, in Fig. 3.13.

We start with the right $\angle AOB$, and draw a circle with O as center passing through B. Producing BO we find the point C. With C as center draw the circle passing through O. The latter circle intersects the former in D. With D as center draw the circle passing through O, this circle intersects the one about O in E. Then $3 \times \angle AOE = \angle AOB$.

Thus there are angles which may be trisected by compass and straightedge, and there are infinitely many such angles: Namely, we may by continued bisection divide $\angle AOB$ in 2^n equal parts for any n, and the resulting small angle may then be trisected by similarly bisecting $\angle AON$ in 2^n equal parts. Of course these are not all, there are several other kinds of angles which may be trisected by compass and straightedge as well.

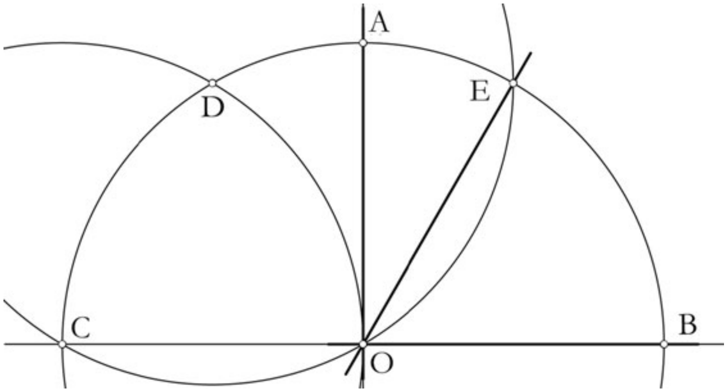


Fig. 3.13 Trisecting a very special angle by straightedge and compass

3.6 Constructions by Compass and Straightedge

Another remark to be made concerning the little construction in Fig. 3.13 is this: The construction illustrates the *legal use of compass and straightedge*. The legal use of compass and straightedge is tied to what later was codified as *Euclid's axioms*. Many complex constructions may be performed under these rules, but the three classical problems are not soluble in this way. This led Greek geometers to introduce other methods, like the use of *conic sections*, also *curves of higher degrees*, even *transcendental curves*, as we would say in modern language: The transition from elementary to *higher geometry* was initiated as a consequence of the struggle with the *classical problems*. The transition is not as unnatural as one might think, since employing conic sections or higher curves is equivalent to solving the problem *by an infinite number of steps* using ruler and straightedge, at each stage in a completely legal manner, according to the rules. We now state these rules.

Legal Use of Compass and Straightedge. A finite set of points is given. A point is constructed if it is a point of intersection between two lines, two circles or a line and a circle as produced according to (1) and (2) below:

(1) The straightedge may be used to draw a line passing through two given or previously constructed points, and to produce it arbitrarily in both directions.
(2) The compass may be used to draw a circle with a given or already constructed point as center, passing through a given or already constructed point.

We note that according to (2), the compass may *not* be used to move a distance. A compass which may only be used in this restricted way, is frequently referred to as a *Euclidian compass*. We may imagine that the compass *collapses immediately when either end is lifted from the paper*.

Using these two procedures is also referred to as *constructing by the Euclidian tools*. By Euclidian tools we may easily perform tasks like *dividing any angle in two equal parts,* drop the normal to a given line from a given point or erect the

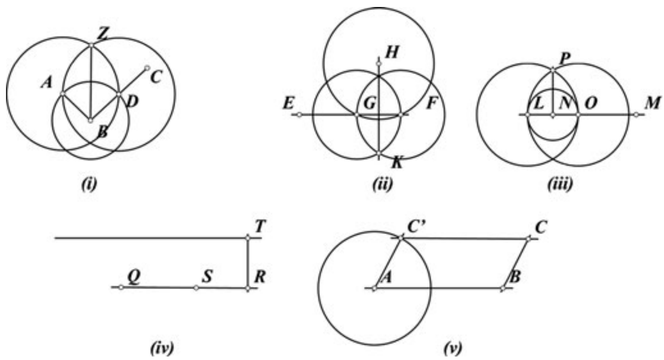


Fig. 3.14 Simple but essential constructions which may be carried out using straightedge and the Euclidian compass

normal at a given point on a given line. This is shown in the three top constructions in Fig. 3.14.

An angle is given by the points A, B and C. We wish to bisect $\angle ABC$. Draw the circle with B as center through A, D is the point of intersection between this circle and the line (possibly produced) BC. Then circles are drawn with A and D as centers, passing through, respectively, D and A. These circles intersect in a point Z such that a line AZ bisects the angle in two equal parts. Next, the line EF is given, as well as the point H outside it. To drop the perpendicular from H to EF, a circle through F is drawn with H as center, intersecting EF in another point G. With F and G as centers, circles are drawn through G and F, respectively, intersecting in K. Then HK is perpendicular to EF, its *foot* is the point of intersection with EF. Finally, we erect the perpendicular to a line LM in the point N. We leave the explanation of this construction to the reader.

In the lower part of the figure, we show how to construct a parallel to a given line QS through a given point T, by first dropping the perpendicular from T to QS (produced), its foot being R, then erecting the perpendicular to RT at T.

We now find a pattern, similar to *proving complex theorems from simpler propositions or axioms*: The construction in (iv) is obtained by appealing to the two previous ones in (ii) and (iii), without having to start from scratch. This becomes even more striking by including construction (v).

Namely, if we allow the compass to be used to draw a circle about a given or constructed point with radius equal to the distance between two other points in the construction, then this is strictly speaking is not allowed according to the rules above. But actually, we may nevertheless do this, since we have the construction (v). Here the points A, B and C are given, and we wish to draw a circle with A as center and radius BC. Proceed as follows: Draw the line BC. Through C construct the parallel to AB, and through A the parallel to BC. They intersect in C'. Now the length of BC equals the length of AC', so draw the circle with center A passing through C'.

The parallel to AB through C is unique since we are in the Euclidian world. The possibility and the uniqueness of the constructions thus hinge on the Fifth Postulate of Euclid. It might be interesting to contemplate what constructions would be like in a non-Euclidian plane.

But to mark off a distance on *the straightedge* is prohibited. By such illegal use of the straightedge one may indeed trisect any angle in three equal parts, as we shall see in Sect. 3.9, and a cube may be doubled, as we shall see in Sect. 4.6. In fact, constructions with compass and a marked straightedge is equivalent to including among the Start Data one single higher curve, namely the *Conchoid of Nicomedes*, which we treat in detail in Sect. 4.6. We also refer to Sect. 17.8.

We now turn to some specifics on the three problems. Even though ideally they should be solved with ruler and straightedge, Greek geometers of course soon realized that this would be very difficult. So they came up with a variety of solutions, ranging from rather simple but effective mechanical schemes, in some cases constructing various kinds of instruments, to very sophisticated geometric constructions like *Archytas'* famous three-dimensional construction for the doubling of the cube, using a cylinder, a cone and a torus. Also employed were a variety of higher algebraic, as well as transcendental, curves in the plane. We shall give some glimpses of these prodigious efforts in the following three sections.

3.7 Squaring the Circle

We have already mentioned that if you can square one circle, then you can square them all. In fact, suppose that a circle of the fixed radius r may be squared, that is to say that we may construct a square of side s such that its area equals that of the circle. The situation is shown in Fig. 3.15.

Here we have a fixed circle, together with a fixed square with side *KQ*, known to have the same area as the area enclosed by the circle. These two being given, we may square *any* circle as follows: We construct a right triangle *VWX*, where the side *VW* is equal to the diameter of the given circle, while *WX* is equal to the side *KQ* of the given square. *VW* and *WX* are the sides containing the right angle. Now consider an arbitrary, new circle, shown in the lower left corner. Mark off *VY* on *VW* equal to its diameter, and let *YZ* be parallel to *WX*, Z falling on *XV*. Then *YZ* is the side of the square of area equal to the that of the new circle.

3.8 Doubling the Cube

We first look at the much simpler problem of *doubling the square* by straightedge and compass. This construction is shown in Fig. 3.16.

Here we have the square ABCD. We now perform the *doubling of the square* in a way very much in the spirit of Greek geometry as follows: Produce the line DC, and mark the point E such that DC = CE. Similarly produce BC and mark F such that $BC = CF$. Then the square *BEFD* will have twice the area of ABCD. Indeed, the former consists of four congruent right triangles while the latter only requires two.

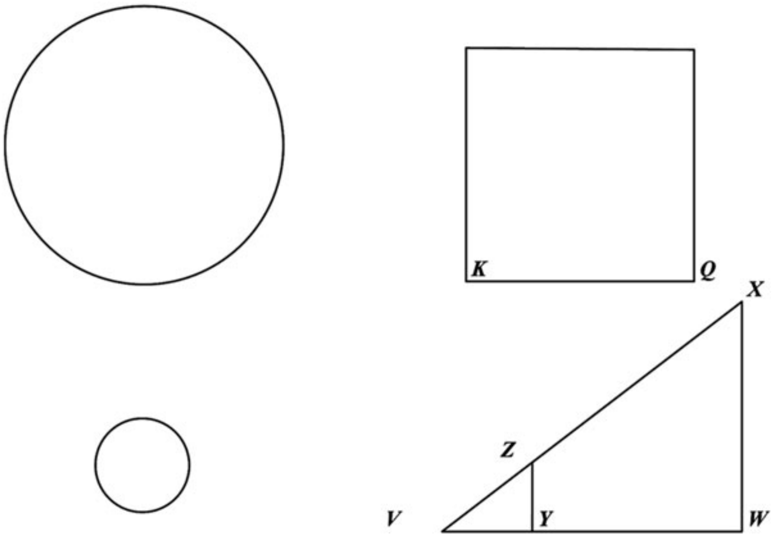
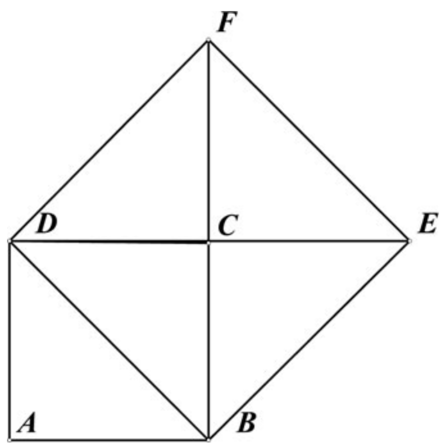


Fig. 3.15 If you can square one circle, you can square them all

Fig. 3.16 Doubling the square by straightedge and compass



But this observation is just the beginning of what led *Hippocrates of Chios* to a most remarkable discovery: Namely, we notice that the triangles *ABD* and *BDF* are similar, thus

$$\mathrm{AB} : \mathrm{BD} = \mathrm{BD} : \mathrm{BF}$$

Thus

$$\mathrm{AB} : \mathrm{BD} = \mathrm{BD} : 2\mathrm{AB}$$

so the side of the double square is the mean proportional between the side and the double side of the given square. Thus putting $\mathrm{AB} = 1$ and using modern notation, we find the side x of the double square by

$$1 : x = x : 2 \text{ or } \frac{1}{x} = \frac{x}{2}$$

so $x = \sqrt{2}$. A construction for doubling the cube which has much of the same flavor, while of course not being possible by straightedge and compass, is attributed to Plato and will be explained in the next section.

It was Hippocrates who realized that the enigma of doubling the cube was but one very special case of a much more general and much more interesting problem: Namely that of *constructing a continued proportionality:*

Construction of a continued proportionality. Let a and b be two line segments. For a given integer n, construct n line segments $x, y, z, \ldots, u, v, w$ such that

$$a : x = x : y = y : z = \cdots = u : v = v : w = w : b$$

x, y z etc. are referred to as the *mean proportionals* of the continued proportionality. A double mean proportionality is one with two mean proportionals, a triple has three, etc.

He saw that doubling a cube of side a is equivalent to constructing *a double continued proportionality* between a and $2a$: To construct x and y such that

$$a : x = x : y = y : 2a$$

We check this with modern notation. We have

$$\frac{a}{x} = \frac{x}{y} = \frac{y}{2a}$$

This gives

$$ay = x^2 \text{ and } 2ax = y^2$$

Squaring the former and substituting y^2 from the latter yields $2a^3x = x^4$, i.e., $x = a\sqrt[3]{2}$.

Recall the following construction of the mean proportional between two line segments $a \geq b$. We refer to Fig. 3.8: First draw a semicircle with diameter AB $= a$, then mark the point D such that AD $= b$.[9] We then have similar triangles ABC and ACD, thus

$$\text{AB:AC} = \text{AC:AD}$$

and so AC $= x$ is the mean proportional.

There is a continuation of this construction to a double continued proportionality, and indeed to any continued proportionality. In fact, from D in Fig. 3.8 we construct a line perpendicular to AC, see Fig. 3.17.

[9] Note that this is a slightly different construction from the one explained when we first encountered Fig. 3.8.

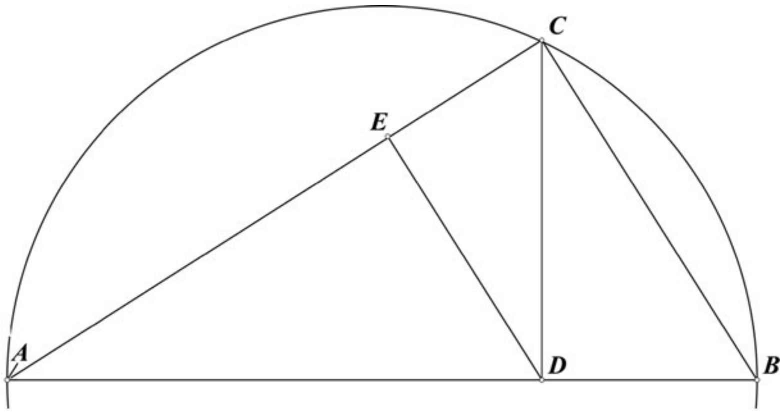


Fig. 3.17 Construction of a double continued proportionality

Letting $\sim$ denote the relation of being similar triangles, we have

$$\Delta \text{ADE} \sim \Delta \text{ACD} \sim \Delta \text{ABC}$$

from which it follows that

$$\text{AB} : \text{AC} = \text{AC} : \text{AD} = \text{AD} : \text{AE}$$

Thus if we wish to construct the double continued proportionality between the line segments $a \geq b$,

$$a : x = x : y = y : b$$

then first draw the semicircle with diameter AB $= a$, and then observe what happens as the point C on the semicircle moves from B to A: In the right ΔABC draw the perpendicular to AB through C, meeting AB in D. Then through D draw the perpendicular to AC, meeting it at E. As now C moves, starting with the degenerate case of C = B where AE $= a$, AE will decrease to 0 when the other degenerate case of C = A is reached. Therefore at some unique location for C on the semicircle, AE $= b$. There we take AC $= x$ and AD $= y$, which solves our problem.

This location for C *cannot be found using straightedge and compass only*, in an allowable manner. But by "cheating" using two of the convenient tools displayed in Fig. 3.18, it becomes simple.

We proceed as shown in Fig. 3.19: First draw the semicircle with diameter AB $= a$. Then mark the points A' and E' on one of the rulers as shown, so that A′E′ $= b$. Now position the rulers as shown in the figure, so that the vertical, unmarked, straightedge meets the marked one in a point on the line AB, where A' coincides with A, and C is found as the point where the marked straightedge crosses the semicircle. E' on the marked straightedge gives us the point E in our figure. We then have the construction from Fig. 3.17.

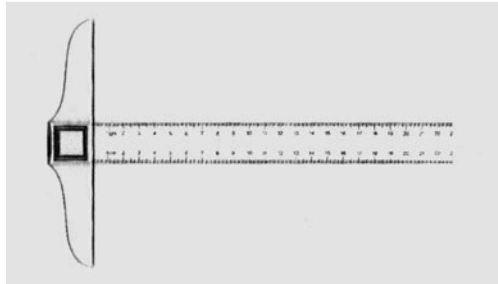


Fig. 3.18 A handy straightedge for finding a double proportionality

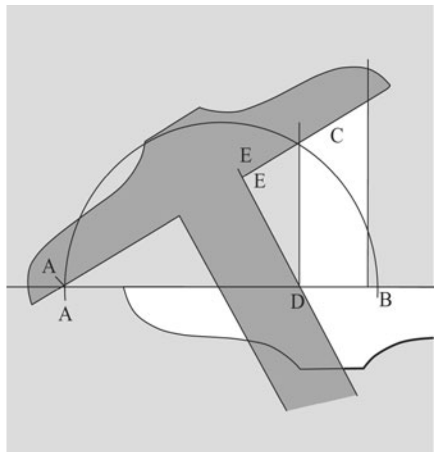


Fig. 3.19 Two rulers, one of them marked, used for finding a double proportionality

This construction is of course completely illegal as a construction with straightedge and compass. In fact, it is even illegal as a version of the already illegal *insertion principle*, which we will explain in the next section. However, in its pure form the insertion principle was much used by Greek geometers, this is also known as a *verging construction.*

3.9 Trisecting Any Angle

The construction of *bisecting any angle* was, as we have seen, very simple. And subdividing a line segment in any number of equal pieces is also a very simple construction. To Greek geometers it must therefore have been a source of frustration and bewilderment that the problem of *dividing any angle into three equal pieces* turned out to be so difficult.

This problem began to attract attention at about the same time as the problem of Doubling the Cube. Some special angles could easily be trisected, as the construction we display in Fig. 3.13.

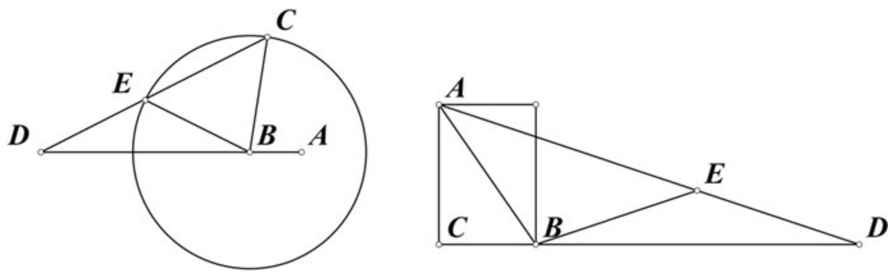


Fig. 3.20 Two verging constructions solving the trisection problem

Greek geometers found solutions to the trisection-problem by solving what they referred to as a *Verging Problem*. We shall not attempt to give a general definition of this concept, but in Fig. 3.20 we present the solution to the trisection problem as being reduced to one variety of such a Verging Problem. Another kind is represented by the famous construction of the regular *7-gon* found by Archimedes, treated in Sect. 4.4. Of course, neither the trisection problem nor the construction of the regular 7-gon are possible by legal use of compass and straightedge.

Now for Fig. 3.20. To the left we have the angle $v = \angle ABC$, we draw the circle about B through C, and then we find the point E on that circle such that the line EC produced meets AB produced in a point D such that the segment DE is equal in length to the radius BC. This is the verging-part of the construction, it is possible by marking off the length BC on the straightedge. Denote the angle at D by u. Then $\angle CEB = 2u = \angle ECB$, thus $v = 3u$. To the right we have the same construction, essentially, but we do not use the circle, nor a marked straightedge, to find the point E such that $AB = BE = ED$. There are simple mechanical devises which may be used, however, based on the construction we have given here.

There are various algebraic curves of degrees higher than 2, so called *Higher Curves*, by means of which the verging problem may be solved. The most famous of these are probably the *Conchoid of Nicomedes*, which we treat in detail in Sect. 4.6.

There is also another famous curve which may be used to trisect any angle, *and* to square the circle as well, in fact it may be used to divide any angle in any number of equal parts and to construct a regular n-gon for *any* number n. A truly marvellous curve! It is the *Quadratrix of Hippias*, treated here in Sect. 4.6 and explained in Fig. 4.31. This is not an *algebraic curve*, however. Like the Archimedean Spiral, it is what we call a *transcendental curve*.

3.10 Plato and the Platonic Solids

Plato was born in 427 B.C. in Athens and died there in 347 B.C. Although he made no original contribution to geometry himself, he has had an immense influence on the subject. In 387 B.C. he founded the *Academy* in Athens, devoted to philosophy and geometry as well as other sciences. Plato had been engaged in the Peloponnesian war as a young man, and he saw his esteemed teacher and friend Socrates condemned and executed. He felt that one reason why the Greek civilization in general, and the one in Athens in particular, was in decline, had to be sought in the disregard of philosophy and geometry. To Plato the problem of *Doubling the Cube*, for example, was a question of developing insights into geometry. Thus it was not a question of finding some practical means for carrying out the physical labor involved, like devising some mechanical instruments or "cheating" with the straightedge. Instead it was a question of understanding the mathematics involved. Therefore Plato would regard highly the doubling-constructions involving higher curves or space-geometric constructions, even if these were of lesser practical value in the actual work of doubling any given cubical altar!

Of course this is exactly how we enjoy this problems today, as well as the one of trisecting any angle or squaring any circle. We understand them in terms of properties of *algebraic numbers*. We return to this in Chap. 17.

To Plato geometry was part of the ideal world, whereas the physical world would only represent imperfect approximations. He ascribed a special significance to the *regular convex polyhedra*, as symbolizing *the four elements Earth, Fire, Air* and *Water.* The fifth one, namely the dodecahedron, stood for the whole *Universe*.

In our modern language a *polyhedron* is a surface enclosing a solid figure composed of (plane) polygons. These are called the *faces* of the polyhedron. The sides of the polygons are called the *edges*, and the corners where the edges meet, are called the *vertices*. Each vertex lies as shown in Fig. 3.21: The vertex is a point P from which the edges, in this case a, b, c and d, emanate. A polyhedral angle at P is a point P in space with some half lines emanating from it.

A polyhedral angle at a point P is said to be *congruent* to a polyhedral angle at the point Q if the angle at P can be moved to Q so that they cover each other completely (Fig. 3.22).

A convex polyhedron is one where a plane containing any face does not cut the other ones. See Fig. 3.23 for an illustration of the property of *convexity*.

We say that a polyhedron is regular if it is *convex* and the faces are *regular polygons of the same kind*, i.e., all are equilateral triangles, all are squares etc. We also

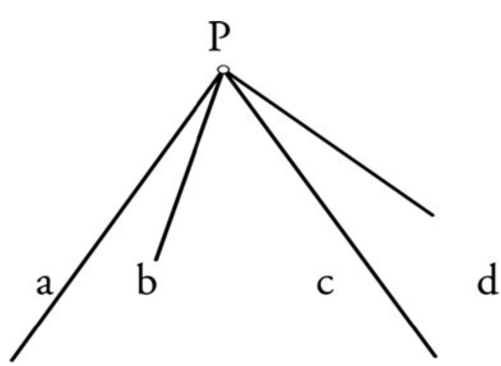


Fig. 3.21 A polyhedral angle

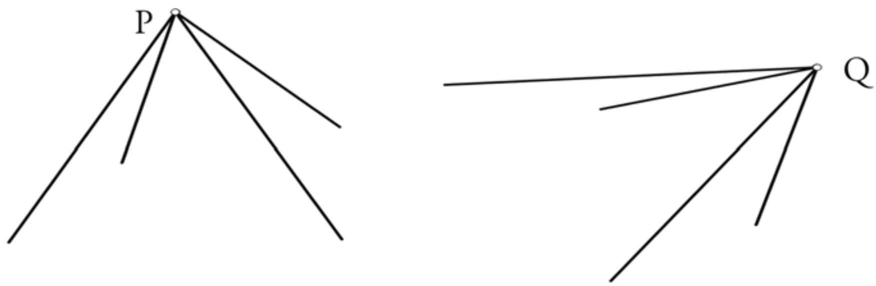


Fig. 3.22 Congruent polyhedral angles

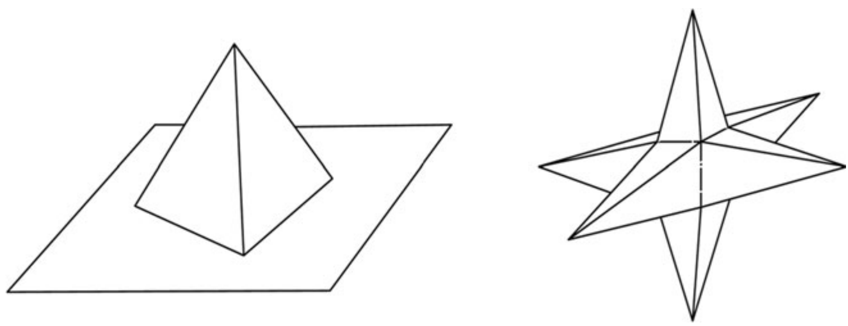


Fig. 3.23 To the left a convex polyhedron. Any plane containing one of the faces, does not cut any other. To the right evidently this property does not hold

require that all polyhedral angles are *congruent*, that is to say that all the configurations of rays at the vertices are the same. We express this by saying that *all vertices are equivalent.*

There are exactly *five such polyhedra*, they are shown in Fig. 3.24.

We can show that these are the only such polyhedra as follows. Let P be a polyhedron of this type, consisting of regular n-gons. Let v be the angle at each vertex. For any convex n-gon, in particular any regular one, the sum of the angles contained by adjacent sides is $(n-2)\pi$. This is easily seen by subdividing it into $n-2$ triangles. Thus $v = \frac{n-2}{n}\pi$. On the other hand the sum of the angles constituting the polyhedral angle must be mv, m being the number of edges meeting at each vertex. Thus we have

$$m(\frac{n-2}{n})\pi < 2\pi$$

and so

$$m(n-2) < 2n$$

For $n = 3$ this leaves the possibilities $m = 3, 4$ or 5, $n = 4$ leaves only $m = 3$, as does $n = 5$. For $n \geq 6$ no value for m is possible. The values for m listed above are indeed realized, and yield the five Platonic Solids.

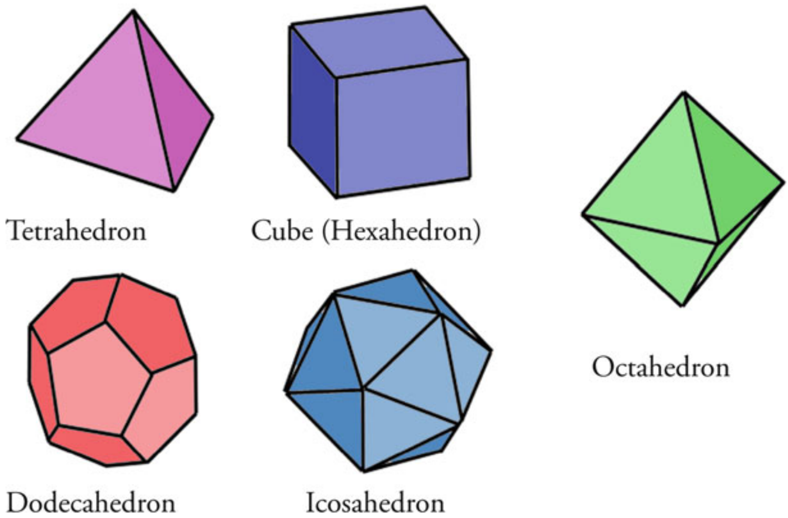


Fig. 3.24 The five platonic polyhedra, or as they are also known, the *Platonic Solids*

3.11 Archytas and Doubling the Cube

Archytas of Tarentum was born 428 B.C. and he died in 365 B.C. in a shipwreck near his home city of Tarentum. Tarentum is located not far from Croton and Metapontium. After the events when the Pythagoreans had been driven out of Italy, things had quieted down to the effect that they had been able to reestablish themselves in the area. He is considered the last great Pythagorean, and in fact Book VIII of Euclid's Elements is generally attributed to him.

He had been a student of another Pythagorean, namely *Philolaus of Tarentum*. Philolaus had studied with some of the expelled Pythagoreans, and he was interested in number magic and mysticism. But he had been allowed to write about the ideas of the Pythagoreans, and the book he wrote is supposed to have been Plato's source of information on the mathematics of Pythagoras and the Pythagoreans.

Archytas made it to the top of Tarentum's politics, he was elected admiral, never lost a battle, and became the ruler of Tarentum with unlimited power. But he is supposed to have been an enlightened ruler, who had a deeply rooted belief in the virtues of philosophy and rationality in politics. He thought that these forces would lead to enlightenment and social justice.

In spite of his political and military work, he also managed to pay attention to mathematics in general and geometry in particular. He lectured extensively, Plato studied under his direction in Tarentum.

Another important Greek geometer who studied under Archytas' direction was *Eudoxus of Cnidus*. Eudoxus had ideas which were precursors to fundamental concepts in our calculus and analysis of today. He probably did the work contained in Euclid's Elements, Book V. See Sect. 4.1.

Archytas' significant contributions to the didactics of mathematics include its division into four subjects: *Arithmetics* constitute the numbers at rest, *Geometry* is the magnitudes at rest, *Music* is the numbers in movement and *Astronomy* is the magnitudes in movement. Later *the mathematical quadrivium* was seen as constituting the *seven free arts,* jointly with a *trivium* which consisted of the subjects *Grammar, Rhetoric and Dialectic.* These ideas were important in didactical practice up to our times.

It is told that Plato once became a prisoner of the notorious tyrant of Syracuse, *Dionysus I,* who ruled with an iron fist, while at the same time writing poems and tragedies. Archytas, who was concerned about the safety of his student and friend, sent a letter to his colleague in Syracuse. In it, he explained to Dionysus that Plato was one of his students and also a dear friend, and that he, Archytas of Tarentum who had never yet lost a single battle, would not like it if his friend should come to harm.[10] This saved Plato's life. A quite significant contribution to philosophy from the admiral in Tarentum.

Archytas solved the problem of doubling the cube by a general construction of the *second continued proportionality* between $a > b$, applied to the case $a = 2b$. His marvellous construction uses an analogy to constructions with straightedge and compass, in the form of finding points in space as intersections of *tori, cylinders and cones.* We show the situation in Fig. 3.25, with the torus, the cylinder and the cone sketched in the first and the second octant, anachronistically including coordinate axes.

We now explain *Archytas' 3-dimensional construction* of the double continued proportionality between $a > b$ in Fig. 3.26. The whole point of the construction is to obtain the right triangles in Fig. 3.17, without using the extended version of the *insertion principle* we employed with our two rulers in Sect. 3.8.

One might say that Archytas' construction appears as a clear cut space-geometric generalization of constructions with straightedge and compass, employing higher dimensional versions of the compass.

In order to describe the construction, we introduce, anachronistically, a Cartesian coordinate system with x, y and z axes. We denote the origin by A. The following description is a slightly edited and commented version of the one given by Archytas himself, as related by Proclus in the *Eudemian Summary*. Of course Archytas did not use terms like *"the xy-plane"* and the like. The situation is visualized in Fig. 3.25, while Fig. 3.26 shows the exact geometry of the construction.

Let $a > b$ be the two given line segments, let Q be a the point on the y-axis such that $AQ = a$. Draw a circle with AQ as diameter in the xy-plane and a semicircle with the same diameter in the first quadrant of the yz-plane. Draw a chord AP of length b to the former circle. On this circle also construct a right cylinder above the xy-plane. The semicircle in the yz-plane is now rotated about the z-axis from Q towards P. While being rotated the semicircle meets the cylinder in a moving point which traces out a curve on the cylinder. In Fig. 3.26 this curve is indicated

[10] Others say that Archytas sent a warship to Syracuse.

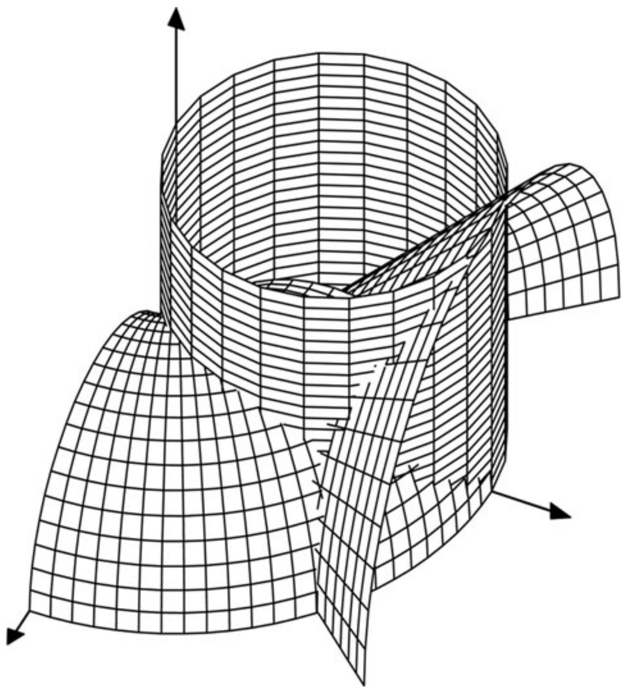


Fig. 3.25 Archytas' setup for the construction of the double continued proportionality, by intersecting a cylinder, a cone and a torus

from Q to A. (In other words, this is the curve of intersection between the cylinder and the torus produced by rotating the circle.) On the other hand, when the prolongation of the chord AP is rotated about the y-axis, then it also meets the cylinder in a moving point, tracing out a curve, which is indicated in the figure from F through the point C pointing towards P. (This is the curve of intersection between the cone and the cylinder.) Evidently these two curves, one sloping upwards from Q and the other sloping downwards from F, will meet in a unique point. (In other words, the three surfaces, the torus, the cylinder and the cone, have exactly one point in common in the first octant.) In Fig. 3.26 this is the point denoted by C. Drop the perpendicular from C to the xy-plane. Denote its foot by D. Now CD of course lies on the cylinder, and thus D lies on the circle in the xy-plane. The moving semicircle through C meets the xy-plane in the point B.

Draw the line PH parallel to the x-axis, it intersects AB in the point G. The line AC meets the circular arc from P to H via R, which P describes as AP is rotated about the y-axis, in a point E. Now EG is perpendicular to the xy-plane: Indeed, it is the intersection of the two planes spanned by ABC and PER, respectively, both of which are perpendicular to the xy-plane. We have now established all points and lines in the figure, and shown their relevant properties. The claim is that we have the double continued proportionality

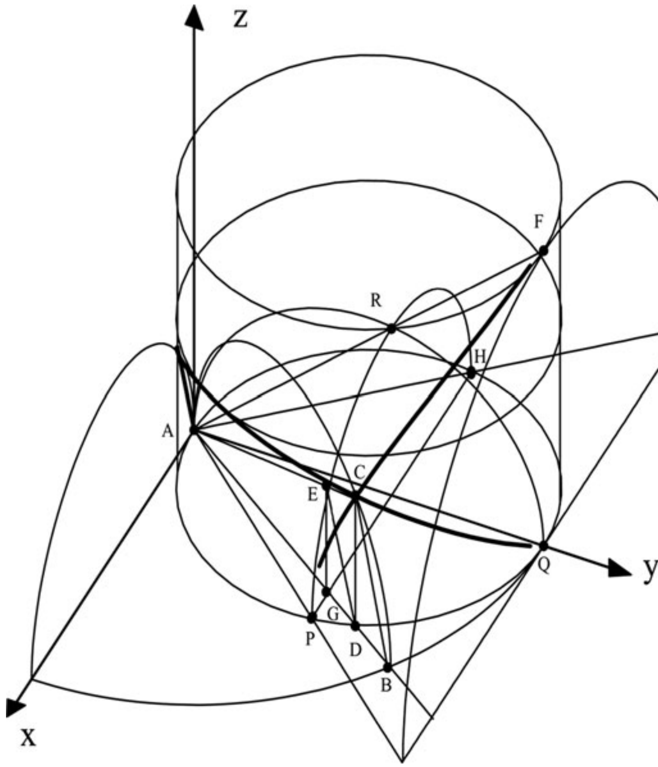


Fig. 3.26 Archytas' construction of the double continued proportionality, by intersecting a cylinder, a cone and a torus. The torus is shown in the third octant only, the cone and the cylinder in the first and the second octants

$$AB : AC = AC : AD = AD : AE$$

which will solve the problem since $AE = AP = b$. From what we already know, it will suffice to show that $\angle AED = \frac{\pi}{4}$, a right angle. In fact, that this suffices was established in the discussion of Fig. 3.17 in Sect. 3.8.

First, from Fig. 3.27 we conclude that $HG : EG = EG : PG$, or in other words, $HG \cdot PG = EG^2$.

But from Fig. 3.28 we find $HG \cdot GP = AG \cdot GD$ since $\Delta DPG \sim \Delta HAG$. Thus we conclude that $AG : EG = EG : GD$.

We now finally use this information on the detail from Archytas' construction shown in Fig. 3.29.

Indeed, we have that $\Delta AGE \sim \Delta EGD$: They have one angle equal, namely the right angle at G, and the sides containing it are pairwise proportional. Hence in particular $\angle EAD = \angle DEG$. But as the corresponding pair of lines AD and EG of these two angles are perpendicular, so must be the case for the other pair. Thus DE is

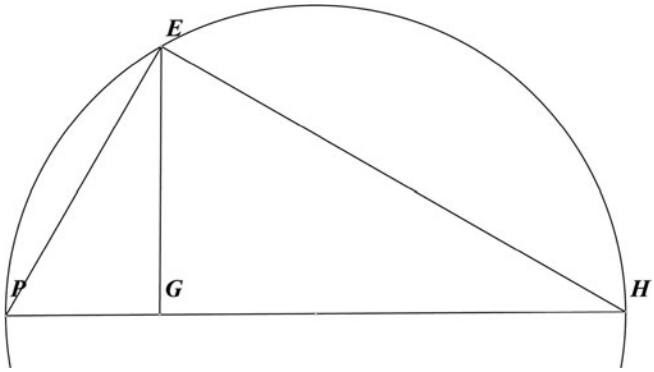
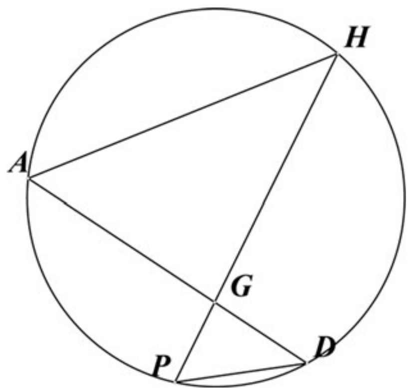


Fig. 3.27 $HG : EG = EG : PG$

Fig. 3.28
$HG \cdot GP = AG{\cdot}GD$



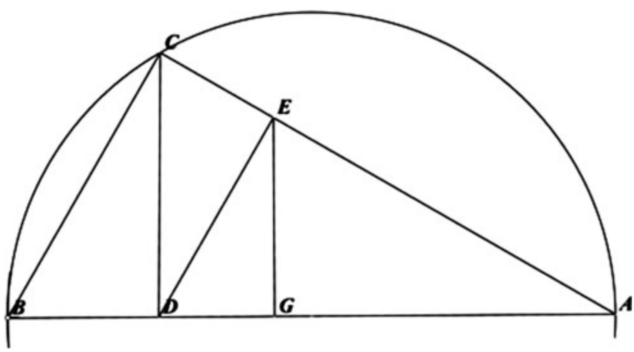


Fig. 3.29 The final argument

perpendicular to *AC*, as claimed. This completes the proof of Archytas' construction of the double continued proportional between $a > b$.

Having completed Archytas' argument, we shall now carry it out by methods which he did not have at his disposal, namely by *algebraic geometry*. Putting the two arguments side by side we are better able to appreciate Archytas' geometric genius, as well as the power and convenience of algebra in geometry. One may even sympathize with those in the beginning of the twentieth century, who resisted the algebraic methods in geometry, feeling that geometry was defaced and destroyed in this way! Plato, incidentally, had similar misgivings about the use of mechanical tools in solving problems like doubling the cube. Comparing Archytas' solution to our crude and illegal use of the two rulers, a procedure very probably well known to Archytas, we may safely conclude that these misgivings were shared by Archytas himself.

The equation of the cylinder in Fig. 3.26 is

$$x^2 + y^2 = ay,$$

the equation of the torus is obtained by putting $r = \sqrt{x^2 + y^2}$, the equation of this surface is then

$$z^2 + r^2 = ar$$

Thus the equation for the torus is

$$x^2 + y^2 + z^2 = a\sqrt{x^2 + y^2}$$

Finally, in (ii) of Fig. 3.30 we see how the cone is produced by rotating the line $y = kx$, where $P = (u, v)$, so that $u^2 + v^2 = av$. Thus $b^2 = av$, and hence

$$k = \frac{v}{u} = \frac{b}{\sqrt{a^2 - b^2}}$$

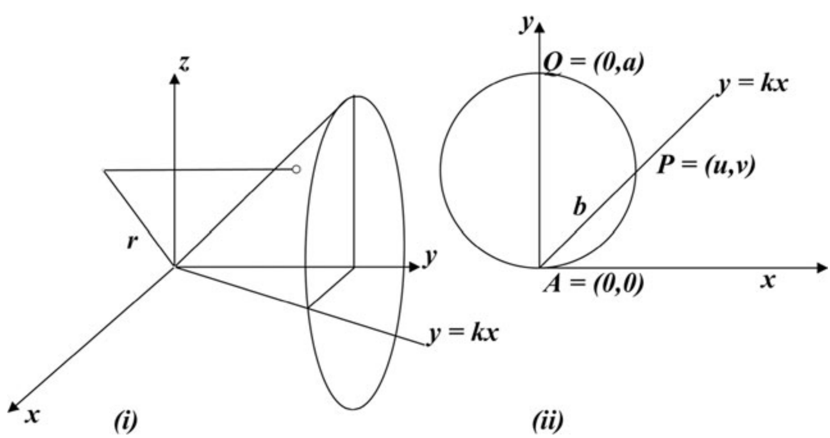


Fig. 3.30 The figure shows how we deduce the equation of the cone in Archytas' construction

When the line in (ii) is rotated about the y-axis, the cone in (i) is generated. With r given by $r^2 = x^2 + z^2$, the equation of the cone becomes

$$y = kr = k\sqrt{x^2 + z^2}$$

i.e.,

$$y^2 = k^2(x^2 + z^2)$$

and when the expression for k, namely $k = \frac{b}{\sqrt{a^2-b^2}}$, is substituted into this equation for the cone, we finally obtain that the cone is given by

$$x^2 + y^2 + z^2 = \frac{a^2}{b^2}y^2$$

where the only constants occurring are a and b. We are now ready to state the

Claim: *With* $\alpha = AC$ *and* $\beta = AD$ *we have*

$$a : \alpha = \alpha : \beta = \beta : b$$

We put $C = (p, q, r)$, so that

$$\alpha = \sqrt{p^2 + q^2 + r^2} \text{ and } \beta = \sqrt{p^2 + q^2}$$

The equation for the torus yields $\alpha^2 = a\beta$, which gives the first proportionality.

From the equation for the cylinder we have $\beta^2 = aq$, while the equation for the cone yields $\alpha = \frac{a}{b}q$, so that

$$b\alpha = \beta^2,$$

which gives the last proportionality.

Exercises

Exercise 3.1 In Sect. 3.4 we carried out the procedure of continued subtractions for the side and the diagonal of a regular pentagon. This yielded the simple continued fraction of the Golden Number. This continued fraction is periodic with period [1], that is to say, the sequence consisting of a single "1" keeps repeating itself. We express this by writing the Golden Number as 1; [1].

Find an algebraic procedure to compute the continued fractions for $\sqrt{n}$ for some values of the positive integer n. Start with $n = 2$, thus essentially carrying out the repeated subtraction for the side and the diagonal of a square.

As we have seen in Sect. 3.8 *Hippocrates of Chios* made significant contributions towards understanding the problem of Squaring the Circle. He had one idea, however, which seemed promising at the time but never led to a solution of the problem.

Today we know that the construction is impossible, but this was not realized until more than 1,000 years after his times.

His idea was to square certain moon-shaped figures. We shall see three of his constructions in the following exercises.

Exercise 3.2 A particular "moon" is bounded by two circular arcs. One is 180° of a circle, the other 90° of an other circle, obviously different from the first one. Show that the area of this moon can be constructed by legal use of ruler and compass, i.e., squared.

Exercise 3.3 Another moon is also bounded by two circular arcs. It is construction is shown in Fig. 3.31:
Let a be a known line segment and construct the trapezoid $ABCB$ with $AB = BC = CD = a$ and $AB = \sqrt{3}a$. Circumscribe a circle about $ABCD$. Construct a circular segment with cord AD which is similar to the three smaller ones with cords AB, BC and CD. These two circular arcs with cord AD define a moon. Show that the area of this moon can be squared by legal use of ruler and compass.

Exercise 3.4 The final moon-construction of Hippocrates which we include here, consists of *a full moon and a crescent moon combined.* We show the situation in Fig. 3.32.
The assignment is to carry out the construction in such a way that both figures *combined* can be squared. Then, if the upper moon could be squared separately, one could square the circle below. Why? Explain that this is not a viable approach to finding a solution to the general problem of squaring the circle.

Exercise 3.5 The method ascribed to Thales for finding the distance to an inaccessible point (see Sect. 3.1, Fig. 3.3) is slightly controversial. The method as described is correct, but it might be impractical since it is necessary to walk away from the point of observation. In fact, Eudemus only states that Thales used the theorem that

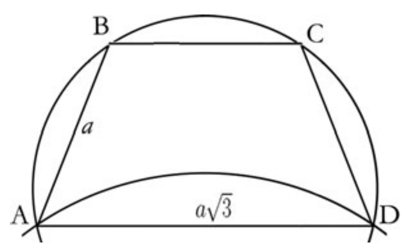


Fig. 3.31 Another moon bounded by arcs

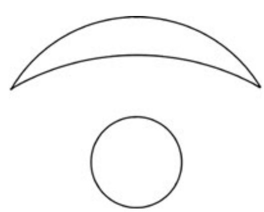


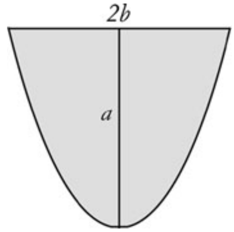
Fig. 3.32 Full moon and a crescent moon combined

two triangles are congruent if they have one side and the two angles at its end points equal, the so called *Angle Side Angle Theorem.* Can you think of an alternative method, using the same theorem but where the observer does not need to walk away from the point of observation?

Exercise 3.6 Thales is credited with finding the height of a pyramid, while visiting Egypt. The method uses a theorem on similar triangles, and uses a measurement of the pyramid's shadow. How would you carry out this task?

Exercise 3.7 Carry out the trisection in equal parts with legal use of straightedge and compass of the angles $u = 45°$ and $v = 27°$

Exercise 3.8 Show that the segment of a parabola shown below can be squared with legal use of ruler and compass, when a and b are known.



Exercise 3.9 Heron of Alexandria, who we tell more about in Sect. 4.10, describes the method used by the workers and engineer Efpalinos when they cut the tunnel through the mountain *Kastron*. He provided a sketch, redrawn by the author from [58] and shown in Fig. 3.33.

The workers had at their disposal a special *diopter*, which was an instrument used in surveying which made it possible to mark off a direction forming a right

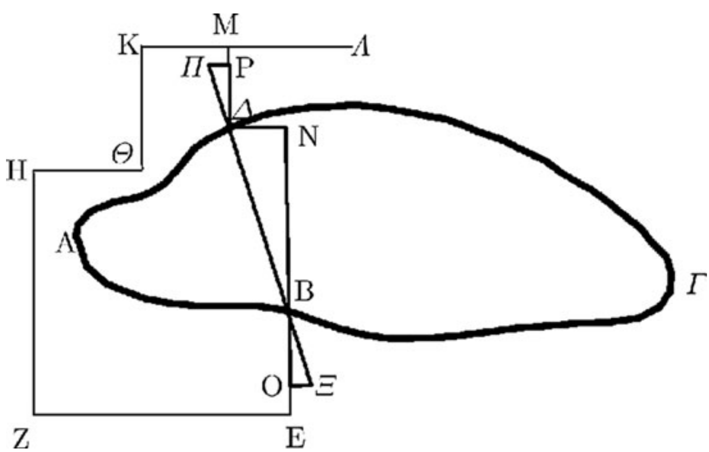


Fig. 3.33 Samos tunnel

angle with a given direction, while passing through some distant point. Moreover, the points of entry and exit of the tunnel were known, but could not be seen from the same point. The sketch shows the lines used, distances between the accessible end points of line segments could be measured. The point N is of course inaccessible. Explain the method used.

Exercise 3.10 Suppose that the figure in Exercise 3.8 represents a half ellipse with half axes a and b. Is this area squarable when a and b are given?

Chapter 4
Geometry in the Hellenistic Era

4.1 Euclid and Euclid's Elements

Alexandria was founded where the Nile meets the Mediterranean by Alexander the Great, in the year 331 B.C. The city became the capital of Egypt, and rapidly developed into one of the richest and most beautiful cities in the world. That is to say, in the world known to the antique.

Alexandria developed into a center for civilization, science, art and culture in general, and remained so for more than three quarters of a millennium. The city possessed many magnificent buildings and awe-inspiring structures. The lighthouse at Faros was counted as one of the worlds seven wonders.

Alexandria was well positioned for trade. It bristled with a lively exchange of valuable goods and commodities between Europe, Asia and Africa. The flourishing city also developed a diverse industrial sector, with products including glass, paper and priceless fabric and cloth. Art and science continued to find fertile soil here, with the most eminent schools of mathematics, astronomy, philology and philosophy.

Euclid – or *Eucleides* which was his real name – was a Greek mathematician who lived around the year 300 B.C. and worked in Alexandria. He should not be confused with another ancient by the same name, a certain *Euclid from Megara*, who was one of the disciples of the philosopher *Socrates,* and appears in Plato's dialogue *Theaetetus*.[1] The latter Euclid has in no way left a comparable legacy to that of the former, disregarding the point of view explained in the footnote.

Euclid collected and systematized the entire body of mathematics known to his time. First and foremost stood *geometry*. Here we must understand that to the Greek mathematical tradition, geometry was in a sense more perfect as a science than computing with numbers. They had no concept of *irrational numbers*, whereas they could "compute" with a quite large class of such numbers via geometry and

[1] However, when Euclid's Elements were reintroduced to Europe towards the end of the Middle Ages, this confusion of the two Euclids did happen. And some historians espouse the theory that *Euclid* was a pseudonym inspired by this dialogue and used by a group of mathematicians working in Alexandria, much in the same way as the name *Nicolas Bourbaki* has been used by a group of French mathematicians in our days.

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_4,

geometric constructions with straightedge and compass. So maybe we can view straightedge and compass as the calculator of the ancient Greeks! And we should realize why the Greeks laid such tremendous importance to the *classical problems*. The problems were all of the same nature: *"What can our calculator do?"*

Euclid based his work on a fundamental idea, which without question was one of the most important ideas in mathematics. It representing a watershed in the understanding of how mathematical insights are gained and secured, and how mathematical activity should be conducted. Essentially taken for granted today by everyone engaged in activities of a mathematical nature, it had emerged through the development of Greek geometry. Philosophers like Plato and Aristotle had also contributed. This fundamental principle is the following:

> *The Hypothetical-Deductive Method* All known geometric facts or theorems should be deduced by agreed upon logical rules of reasoning from a set of initial, self evident truths, called *postulates*.
>
> These postulates should be such that every informed person would agree on their validity, to the extent that they did not require proof. The set of postulates should be kept as small as possible, thus one should endeavor to construct proofs of assertions which, even though self evident, could be deduced from other even more fundamental self evident ones.

In a similar manner Euclid defined the more complicated figures and concepts using fundamental ones like points and lines. He even gave definitions of these, for example asserting that *a point is that which has no parts*, and defining a line as *that which has only length.* Even though his method has stood up through more than twentieth centuries, the postulates themselves have had to be refined and made more precise.

Euclid based geometry on five *axioms* or common notions, and five *postulates*. The former were supposedly more obvious than the latter, nevertheless all were considered as *self evident truths*, not requiring proofs. They could be taken for granted.

Euclid's work was in no way easy to read! When his powerful mentor, King Ptolemy I, asked if there did not exist an easier way to learn geometry than to read all this, Euclid answered:

> – *"No, to the geometry there is no separate road for kings, there is no Royal Road to Geometry."*

Euclid's work *The Elements* has had an enormous influence on mathematics in general and geometry in particular. But it was not confined to geometry alone. It also contains a substantial body of algebra.

Almost up to our time it has been used as textbook, and then been replaced by works which did not always represent improvements. Geometry was, for a long time, synonymous with *Geometry according to Euclid*.

4.2 The Books of Euclid's Elements

We follow Heaths edition of Euclid's Elements [27]. In Book I the foundations for geometry is laid out. Here we find the fundamental definitions and axioms. Their conciseness and precision are remarkable, even today. Of course there has been critical remarks, and as an axiomatic system it has required a considerable amount of work over the more than 2,000 years which have elapsed since these statements were written. But let us enjoy Euclid's terse and precise style! Book I opens with.

4.2.1 Euclid's Definitions

1. A point is that which has no parts.[2]
2. A line is breadthless length.
3. The extremities of a line are points.
4. A straight line is a line which lies evenly with the points on itself.
5. A surface is that which has length and breadth only.
6. The extremities of a surface are lines
7. A plane surface is that which lies evenly with the straight lines on itself.
8. A plane angle is the inclination to one another of two lines in a plane which meet one another and do not lie in a straight line.
9. And when the lines containing the angle are straight, the angle is caller rectilinear.
10. When a straight line set up on a straight line makes the adjacent angles equal to one another, each of the angles is right, and the straight line standing on the other is called the perpendicular to the one on which it stands.
11. An obtuse angle is an angle greater than a right angle.
12. An acute angle is an angle less than a right angle.
13. A boundary is that which is an extremity of anything.
14. A figure is that which is contained by any boundary or boundaries.
15. A circle is a plane figure contained by one line such that all the straight lines falling upon it from one point from those lying within the figure are equal to one another.
16. And the point is called the center of the circle.
17. A diameter of the circle is any straight line drawn through the center and terminated in both directions by the circumference of the circle, and such a straight line also bisects the circle.
18. A semicircle is the figure contained by the diameter and the circumference cut off by it. And the center of the semicircle is the same as that of the circle.

[2] An alternative translation from the Greek original would be: "A point is that which is indivisible into parts."

19. Rectilinear figures are those which are contained by straight lines, trilateral figures being those contained by three, quadrilateral those contained by four, and multilateral those contained by more than four straight lines.
20. Of trilateral figures, an equilateral triangle is that which has its three sides equal, an isosceles triangle that which has two of its sides alone equal, and a scalene triangle that which has its three sides unequal.
21. Further, of trilateral figures, a right-angled triangle is that which has a right angle, an obtuse-angled triangle that which has an obtuse angle, and an acute angled triangle that which has its three angles acute.
22. Of quadrilateral figures, a square is that which is both equilateral and right-angled, an oblong that which is right angled but not equilateral, a rhombus that which is equilateral but not right-angled, a rhomboid that which has its opposite sides and angles equal to one another but is neither equilateral nor right-angled. And let quadrilaterals other than these be called trapezia.
23. Parallel straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

It is interesting to note that the term *line* does not signify only *straight line*, and that the straight lines are of finite length, but may be infinitely produced in either direction. The (curved) lines of Euclid do define angles, and they do have *length*. Today, in an axiomatic treatment of geometry, it is customary to take the terms *point* and *line* as undefined, as well as the relation of *incidence* between points and lines. This will be our approach in Chap. 9. Now Euclid does not specify his *undefined terms*, but he nevertheless makes his definitions using such terms. One is the word *part*, and perhaps *divisible*. Another term would be *length*, and so on. Many eminent mathematicians have worked on the project of understanding Euclid's definitions and postulates. It is fair to say that Euclid's achievement represents the single most fruitful set of ideas in the entire history of mathematics!

David Hilbert has treated Euclid's geometry in his classical work *Grundlagen der Geometrie* [29]. This work is viewed by many as the final word in mathematically securing Euclid's axioms and postulates.

Having thus formulated the basic definitions, in a language and style which summons our admiration even today, Euclid proceeds:

4.2.2 Euclid's Postulates

Let the following be postulated:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and distance.
4. That all right angles are equal to one another.

Fig. 4.1 Euclid, about 325–265 B.C. Drawing by the author, inspired by sculpture at the Oxford University Museum of Natural History, photo by Mark A. Wilson

5. That, if a straight line falling on two straight lines make the interior angles on the same line less than two right angles, the two straight lines, if produced indefinitely meet on that side on which are the angles less than the two right angles.

These are the *five postulates of Euclid* (Fig. 4.1). Among them the *fifth postulate* occupies a special position. First of all, as it stands it appears considerably *less obvious* than the others. The remarkable fact is, of course, that Euclid must have concluded that the assertion was *indemonstrable*, and that he could not come up with some much simpler statement, equivalent to it in the presence of the other four postulates. But the Fifth Postulate of Euclid continued to haunt mathematicians for more than two thousand years, until it was realized that the Fifth Postulate is indeed independent of the other four. With the discovery of non-Euclidian geometry, fully valid geometries revealed themselves, in which the Fifth Postulate is no longer true. We shall return to this subject in Sect. 8.2.

But in the process of attempting to prove the Fifth Postulate, many statements were discovered which in the presence of the other four is equivalent to the Fifth. We shall again follow Heath in [27], and list the most important among these statements.

4.2.3 Alternative Versions of Euclid's Fifth Postulate

1. Through a given point only one[3] parallel can be drawn to a given straight line. Due to *Proclus*. It is commonly known as *Playfair's Axiom*, but was not a new discovery.
2. There exist straight lines everywhere equidistant from one another.
3. There exists a triangle in which the sum of the three angles is equal to the sum of two right angles.
   Due to *Legendre.*
4. Given any figure, there exists a figure similar to it of any size we please. This form is due to Legendre, Wallis and Carnot.
5. Through any point within an angle less than two thirds of a right angle a straight line can always be drawn which meets both sides of the angle.
   Due to Legendre.
6. Given any three points not on a straight line, there exists a circle passing through them.
   Due to Legendre and Bolyai.
7. There exists a triangle, the contents of which is greater than any given area.
   Due to Gauss in a letter to Bolyai in 1799.[4]

Finally Euclid formulated five statements, called *Common Notions* or *Axioms*. These were statements of a general nature, viewed as being universally valid in all fields of human thought. They were the following:

4.2.4 Euclid's Common Notions or Axioms

1. Things which are equal to the same thing are also equal to one another.
2. If equals be added to equals, the wholes are equal.
3. If equals be subtracted from equals, the remainders are equal.
4. Things which coincide with one another are equal to one another.
5. The whole is greater than the part.

[3] Meaning one and only one.

[4] Gauss wrote, according to [27]: "If I could prove that a rectilinear triangle is possible the contents of which is greater than any given area, I am in a position to prove perfectly rigorous the whole of geometry."

On the basis of the Definitions, the Postulates and the Common Notions, Euclid builds the whole of geometry. Now this material was due to many different Greek geometers of course, and several books in the Elements are thought to have been written in the entirety by others than Euclid. The point is that Euclid collected and systematized essentially the complete body of mathematics known at that time.

The last two propositions in Book I, namely I.47 and I.48, treat the "Pythagorean Theorem." Proposition I.47 reads as follows, quoted from [27]:

Pythagoras According to Euclid. *In right angled triangles the square on the side subtending the right angle is equal to the (sum of the) squares on the sides containing the right angle.*

The parenthesis is tacitly assumed in the Elements.

The last proposition of Book I is I.48, which is the converse to the Pythagorean Theorem:

The Converse Pythagoras. *If in a triangle the square on one of the sides be equal to the squares on the remaining two sides, then the angle contained by the remaining sides is right.*

We shall now render Euclid's famous and very elegant proof of the Pythagorean Theorem. We consider Fig. 4.2.

It is not practical to give the proof in the original form, using Euclid's own words. The reason for this is that Euclid proceeds rigorously from the *First Principles*,

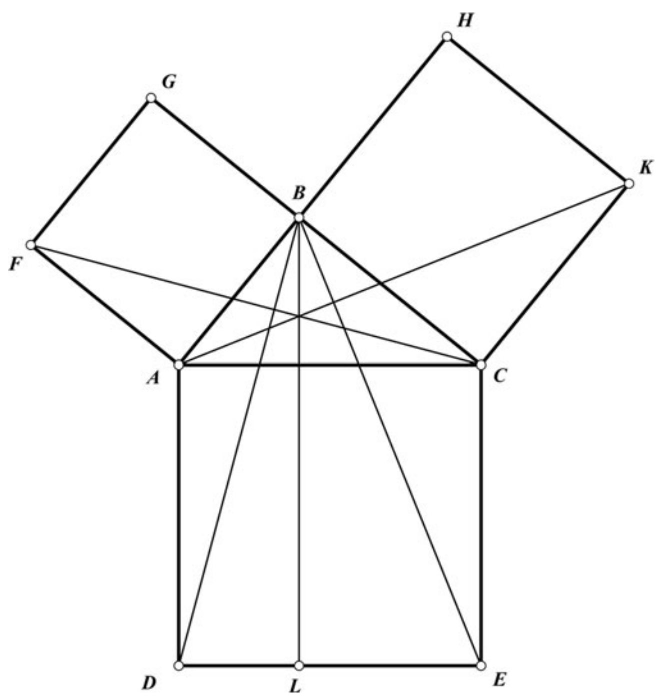


Fig. 4.2 Euclid's illustration to his proof of the Pythagorean Theorem

in other words from the Definitions, Postulates and Common Notions, using also propositions already proven. Thus for example, the fact that the points G, B and C lie on the same straight line requires a proof. In [27] the whole proof fills almost two printed pages, namely pp. 349 and 350 of Volume 1. Here we give a modern version of the geometric contents, in the style we find in textbooks of elementary geometry.

The produced normal from B to the hypothenuse AC divides the square on AC in the two rectangles with base DL and LE and height equal to DA. It will suffice to prove that

(1) The former rectangle has area equal to the square on AB.
(2) The latter rectangle has area equal to the square on BC.

$\triangle ACF$ is congruent with $\triangle ADB$ since the sides are pairwise equal. $\triangle ACF$ is a triangle with base AF and height AB, its area therefore is half of that of the square on AB. $\triangle ADB$ has base AD and height DL, its area therefore is equal to half of the rectangle with base DL and height DA, and (1) is proven. (2) follows analogously.

The geometry continues in Book II, which deals with *Geometric Algebra*. In Greek mathematics Geometric Algebra played a similar, but less prominent, role to the one played by algebra today. Thus for example, the two first propositions are equivalent to the following formulas:

$$(1)\ a(b+c+d+\dots) = ab+ac+ad+\dots$$
$$(2)\quad (a+b)a+(a+b)b = (a+b)^2$$

The first formula must be interpreted as asserting that the area of the big rectangle to the left is equal to the sum of the many small ones to the right. The other formula expresses the sum of the areas of the two rectangles on the left as the area of the square to the right. But in the Elements the formulas were expressed *verbally*, they were formulated as follows, quoted from [27]:

> *Formula 1.* If there be two straight lines, and one of them be cut into any number of segments whatever, the rectangle contained by the two straight lines is equal to the (sum of the) rectangles contained by the uncut straight line and each of the segments.

Again, the parenthesis is tacitly assumed in the Elements.

The next formula is quoted literally:

> *Formula 2.* If a straight line be cut at random, the rectangle contained by the whole and both of the segments is equal to the square on the whole.

This should give the reader a taste of the *verbal nature* of Greek geometry. In fact, the absence of a good algebraic notation and the verbal form of exposition made it very difficult to resume work in geometry once the line of transition from person to person had been interrupted. When the last great geometer had been dead for fifty or a hundred years, it was not easy to continue the work only with written sources! This was no problem in Euclid's times, when the research community in Alexandria and elsewhere consisted of many individuals, and was robust in that it consisted of people at all levels. But towards the end of the Hellenistic Epoch it probably was

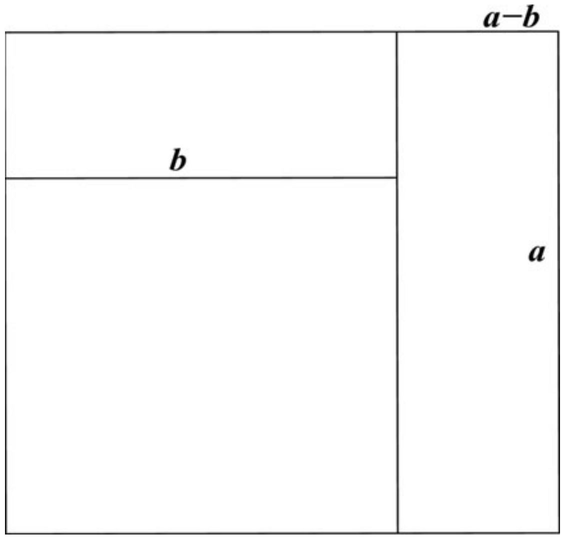


Fig. 4.3 The large square with side a has area a^2, the small one with side b has area b^2. The smaller square is placed at the lower left corner of the bigger one, then the difference gets the shape of the gnomon, or the carpenters square. This gnomon has the breadth $a - b$, and consists of two pieces: One piece has length a, and the other one has length b, a total of $a + b$. The total area is therefore the common breadth, $a - b$, multiplied with the sum of the lengths, $a + b$

a contributing factor to the end of Greek geometry and mathematics. When a new Genius is born, a devoted teacher is essential for it to flourish. A teacher, perhaps, who has had personal contact with the last great master of the subject.

The word *Gnomon* means in some contexts an ancient astronomical instrument. In Greek mathematics it is the *Carpenter's Square*, and using figures of this shape Greek geometers carried out arguments where we would use algebra. In Fig. 4.3 we show the equivalent of the deduction of a familiar formula:

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

In Book II we also find a generalization of the Pythagorean Theorem, in Propositions II.12 and II.13.

Book III deals with properties of circles and circle segments, and ends with the following two propositions:

Two Cords. If in a circle two straight lines cut one another, the rectangle contained by the segments of the one is equal to the rectangle contained by the segments of the other.

A Line Cutting a Circle. If a point be taken outside a circle and from it there fall on the circle two straight lines, and if one of them cut the circle and the other touch it, the rectangle contained by the whole of the straight line which cuts the circle and the straight line intercepted on it outside between the point and the convex circumference will be equal to the square on the tangent.

Book IV contains only problems concerning figures given by straight lines which may be inscribed or circumscribed circles. Book V is probably due to *Eudoxus*, as was mentioned in Sect. 3.11, and contains his theory of proportions between *"magnitudes of the same kind"*. It treats the problem of commensurability in this light.

Two magnitudes are said to be of the *same kind* if they are capable of exceeding one another when suitably multiplied. For two magnitudes of the same kind defines their *ratio*, and the notion of *equality* for ratios is defined in terms of the property of exceeding by multiplication. We shall not go into the details of this here, we have written more about it in [33]. The definition used by Aristotle was to consider two ratios to be equal if the repeated subtraction, the *antanairesis*, had the same pattern in the two cases. Thus to Greek geometers it was possible to say that the ratio between two *line segments* was equal to the ratio between two certain *areas* or *volumes*. A line segment and an area would not be magnitudes of the same kind, but two line segment would. This latter assertion is some times referred to as the *Axiom of Archimedes*: Two line segments are capable of exceeding one another when suitably multiplied.

Book VI applies the theory from V to geometry. Here it is proven that the areas of two triangles or two parallelograms of the same heights are to each other as their bases.

Books VII, VIII and IX deal with what we would call *elementary number theory.* Book VII opens with *the Euclidian Algorithm*, we treat it in detail in Sect. 16.3. The point is to find the *greatest common divisor* of two (positive) integers.

Book VIII treats *continued proportionalities*, such as the *double proportionality* which Hippocrates was led to consider from his attempts of solving the problem of *doubling the cube*: $a : b = b : c = c : d$. This book is believed to be due to *Archytas*.

Proposition 14 in Book IX is equivalent to the fundamental result in number theory, that any integer may be factorized, essentially uniquely, into a product of prime numbers. Here we also find Euclid's famous proof that there are infinitely many prime numbers.

Book X treats *incommensurable magnitudes.* This is where the foundations for the important *Principle of Exhaustion* is established. We find it already in Proposition 1, quoted here from [27]:

> *Euclid X.1.* Two unequal magnitudes being set out, if from the greater there be subtracted a magnitude greater than its half, and from that which is left a magnitude greater than its half, and if this process be repeated continually, there will be left some magnitude which will be less than the lesser of the magnitudes.

This proposition is used to prove Proposition XII.2, namely that the areas of two circles are to one another as the squares on their diameters.

Book XI treats basic solid geometry, in Book XII we find applications of the Principle of Exhaustion, for example proving what we would call the formulas for the volumes of a pyramid and a cone. Finally, the aim of Book XIII is to construct the five regular polyhedra and their circumscribed spheres.

All of this material is due earlier Greek geometers, among them Archytas and Eudoxus.

4.3 The Roman Empire

The Roman State is usually referred to as the *Roman Empire*. It developed from the city-state of Rome, which according to tradition was founded 753 B.C. At the outset it was ruled by kings, and the last of them was *Tarquinius Superbus*, Tarquin the Proud. He was driven out around 510 B.C., having overstepped his powers. The Romans then designed a republican constitution, maintained as a system of traditions and customs, to provide a safeguard against abuse of power. Below we give a simplified account of the Roman governing system.

The people, that is to say the Roman citizen, consisted roughly speaking of two classes. The nobility, the *patricians*, stood against the common people, the *plebeians*. Eventually a balance between the two classes was worked out.

The people could make decisions through the *Popular Assembly*, which consisted of an assembly for the army, a tribal assembly, and the plebeian assembly. The plebeians appointed ten Tribunes[5] of the People each year. They had the power to provide protection against abuse of power. They were declared sacred and inviolable.

The *Senate*, or *Council of Elders*, was an outgrowth of the old council of the king. At the beginning of the third century B.C. it had 300 members, originally they were the heads, Patres or Fathers, of leading families among the patricians. Later came the *drafted plebeian senators* who were called Conscripti, conscripted men, since they were conscripted into service. But the distinction eventually disappeared. The Roman senate was the main governing council of the Republic as well as of the Empire later.

All who had held important offices such as quaestors, praetors or consuls were senators for life. But membership could be stripped if someone was thought to have committed an act against the public morals. Senators who had not held important positions ranked below the others.

The republican era lasted from 510 to 31 B.C. and is usually subdivided into three epochs.

The Old Republic lasted till about 300 B.C., and is often labelled *The Struggle of the Orders*. During this time the arrangement between patricians and plebeians was worked out.

In 387 B.C. Rome was attacked by Gallic forces, and had to surrender after a siege which lasted for months. The Romans did not forget this defeat, and during the following century they conquered the Italian mainland south of the river *Rubicon*,

[5] The Romans were divided into *tribes*, and *Tribunes* were elected positions, originally representatives of the tribes.

now named *Pisatello*. Then during the third century B.C. they defeated the Etruscans in the south and the Gallic tribes in the north. Thus Rome became the master of all of Italy.

The next epoch is often labelled *The Classic Republic*, and lasted from about 300 to 130 B.C. During this time Rome had a stable government, dominated by the senate. Now Rome became the leading Mediterranean power. The conquest of states surrounding the Mediterranean Ocean brought about the wars with Carthage. Carthage was a Phoenician colony originally, the name means *The New City* in the Phoenician language. This Semitic people had an advanced culture, the Greek took their alphabet from them, and according to the Eudemian Summary, it was the Phoenicians who brought the *numbers* to the Greek as well. The rich culture of Carthage was obliterated when the city and its people were annihilated by the Romans at the conclusion of the *Third Punic War*. *Punians* was the Roman term for Phoenicians.

After the end of the Roman Kingdom, the powers of the King were transferred to two *Consuls*, "those who walk together." In case of grave danger the senate would sometimes endow the Consuls with dictatorial power. Ranking below the consuls with chiefly judicial functions were Praetors. Two consuls were elected each year, they served together with veto power over each other's actions.

At the beginning only patricians could be elected to consulships, but in 367 B.C. the plebeians were included.

A consulship was originally considered the crowning achievement and end point of a political career in Rome, but later a former consul would usually serve one or more lucrative terms as Proconsul, or *Governor*, of one of the provinces.

Quaestors were elected officials of the Roman Republic who supervised the treasury and financial affairs of the state, its armies and its officers. Some quaestors were assigned to work in Rome itself, others were assigned to generals or served in the provinces.

The assemblies possessed ultimate legislative and judicial powers in the Roman Republic, and were also responsible for the election of officials such as military commanders.

The final phase of the Republic is the *Century of Civil War*. It lasted from 130 to 31 B.C. This phase opens with the passing of laws which limited how much state land a Roman could own.

Then around 120 B.C. a gigantic migration took place, when the Cimbrians and the Teutons, both Germanic tribes, started their trek towards southern Europe. In 113 B.C. they clashed with Roman troops in what is now Austria. The Romans were overrun, and when the message of the defeat reached Rome, it caused great consternation. However, the Romans had their hands full in the Mediterranean area, and the situation seemed to be under control.

But in 105 B.C. the intruders clashed with and defeated the Romans again. The situation became quite serious for Rome. At the same time *Gaius Marius* became consul for 1 year, he was subsequently reelected and served for six consecutive terms, for the entire period 105–100 B.C. (Fig. 4.4). In the end Marius was

Fig. 4.4 Gaius Marius, 157–86 B.C. Drawing by the author

elected Consul an unprecedented seven times during his career. His final and seventh election was in 86 B.C., at the age of 71, when he died in office.

Marius had a rather modest family background, and in 134 B.C. he served as a soldier under *Scipio Africanus the younger* at Numantia, presently in Spain. Scipio was the Roman general who had carried out the final destruction of Carthage.

The war in Numantia was not going well for the Romans, and when the consul Scipio arrived and undertook much needed reforms in the army Marius enthusiastically embraced it. He distinguished himself convincingly in battle, and when at one occasion Scipio was asked where the Romans after him should find another general like himself, he clapped Marius on the shoulder and said *"Here, perhaps."*

Marius decided to seek public office, and with the aid of the influential *Caecillius Metellus* he became a peoples tribune. Marius' family were dependents of Metellus. But Marius rejected being dependent on him, and their relationship eventually became quite difficult.

When Marius was elected to the position of praetor, he was accused of bribery. But in 109 B.C. Metellus, who had been elected consul, took Marius with him as his legate on his campaign against Jugurtha, who was the king of Numidia in North Africa. Legates were envoys sent by the senate, but they were often used as second in command.

Metellus had to put the army back in order since discipline and combat readiness had deteriorated. After this he won several significant victories, and Jugurtha fled from Numidia. But Metellus was unpopular, and when Marius against Metellus' will travelled to Rome to seek the consulship, he was successful. Now Marius returned as consul in 105 B.C., and took over the command from Metellus. The latter had to return to Rome, having been deprived of his final victory. This was accomplished by Marius, who actually only had to put the finishing touch to the campaign by sending out his quaestor *Sulla*, who we tell more about later, to capture the fleeing Jugurtha in the neighboring Mauritania. In any case Metellus was granted his Triumph in Rome, but this incident sealed a bitter hostility between the two men.

As for the fight against Cimbrians and Teutons, Marius strengthened the army by allowing men without land to become soldiers. Now the Cimbrians and Teutons were finally defeated, they were killed off or committed suicide in and after two huge battles in 102 and 101 B.C. near what is now the French Mediterranean coast.

During the second century B.C. Rome had gradually conquered the Mediterranean area. In the southern part of Gallia (presently Spain) as well as in Africa they consolidated their position. Rome now took control over the entire Hellenistic world. The city of *Pergamon*, a cultural and scientific center rivaling Alexandria, was taken over by Rome, Macedonia was defeated, as was Antiochus III of Syria.

The success came at a price, however. The Roman army had previously had a core consisting of free Italian peasants, who fought mainly during the summer. As long as the wars were fought close to home, the soldiers could still farm their land in between the expeditions. But when the action moved far away, this changed. The economic hardship which resulted, caused many to loose their land, rich landowners bought it up and became even richer.

At the same time the conquests brought taxes from the new provinces to Rome, this was mostly in the form of grain and other agricultural produce. The surplus thus created destroyed the profit for the domestic farmers. And, finally, the conquests also brought large numbers of *slaves* to the center of the Empire, and in addition the Eastern slave markets became available. So the disenfranchised peasantry had no way of finding decent employment on the estates of the great landowners, but had to move to Rome, where they now became a growing impoverished proletariat.

Now the soldiers became more dependent on their generals. The road lay open to a state of affairs where warlords with money could muster loyal troops in their bids for political power. This situation was almost a prerequisite for the civil war which was to commence with full force, as well as for the low quality of some of the emperors who came to more or less short-lived power towards the end of the Roman Empire. The sad story of how the Republic ended will be told in Sect. 4.8.

4.4 Archimedes

One of the very greatest mathematician, scientist and engineer of antiquity was *Archimedes*.

He lived and worked in Syracuse, as a very close associate of King Hiero. Some have speculated that Archimedes might have been one of the Kings relatives. Be that as it may, the King certainly had ample reason for appreciating Archimedes' friendship (Fig. 4.5).

Indeed, not only did Archimedes make inventions and conduct major engineering enterprizes, but he also was the sole reason why Syracuse could hold its own for 4 years against the overwhelming Roman forces. In the so called Second Punic war between Rome and Carthage, Syracuse had sided with Carthage, it had strong historical ties to that Phoenician colony.

Under the leadership of the Roman consul *Marcus Claudius Marcellus*, later given the name of honor "Rome's Sword," Syracuse was under siege for 2 years, from 214 to 212 B.C. (Fig. 4.6).

Fig. 4.5 Archimedes. Drawing by the author

Fig. 4.6 Marcus Claudius Marcellus. Drawing by the author

And the only reason why Syracuse could hold out for so long, and was finally taken only because of treason from within, was the technological superiority which Archimedes' war machines represented. We now quote from Plutarchus [46].

"Marcellus attacked from the sea with sixty galleys, each with five rows of oars, furnished with all sorts of arms and missiles. Eight ships were chained together, with a huge bridge of planks on them. On this bridge there was placed a formidable catapult, an engine to cast stones and arrows to attack the city walls. Marcellus relied on the abundance and magnificence of his preparations and his own previous glory. But all of this seemed but trifles for Archimedes and his machines.

These machines he had designed and contrived, not as matters of any importance, but as mere amusements in geometry. A short time before king Hiero had wanted him to carry out in practice some part of his admirable speculations, and by accommodating the theoretic truth to sensation and ordinary use, bring it more within the appreciation of the people in general. Eudoxus and Archytas had been the first originators of this far-famed and highly praised art of mechanics, which they employed as an elegant illustration of geometrical truths, and as means of sustaining experimentally, to the satisfaction of the senses, conclusions too intricate for proofs by words and diagrams. As, for example, to solve the problem, so often required in constructing geometrical figures, given the two extremes, to find the two mean lines of a proportion,[6] both these mathematicians had recourse to the aid of instruments, adapting to their purpose certain curves and sections of lines. But what of Plato's indignation at it, and his invectives against it as the mere corruption and annihilation of the one good of geometry, which was thus shamefully turning its bach upon the unembodied objects of pure intelligence to recur to sensation, and to ask help (not to be obtained without base supervisions and depravation) from matter. So it was that mechanics came to be separated from geometry, and repudiated and neglected by philosophers, took its place as a military art.

Archimedes, however, in writing to king Hiero, whose friend and close relation he was, had stated that given the force, any given weight might be moved. He had even boasted, we are told, relying on the strength of demonstration, that if there were another earth, by going into it he could remove this.

Hiero was struck with amazement at this, and entreated him to make good this problem by actual experiment, and show some great weight moved by a small engine. Archimedes then had one of the king's three masted cargo ships pulled ashore with many men and great labor. He then loaded the ship so heavily that the crew could not possibly pull it out into the water again.

Archimedes sat some distance away and operated a system of pulleys with his hands, without any great effort. He then drew the ship in a straight line as smoothly and effortlessly as if she had been in the sea.

The king was astonished, and convinced of the power of the art, prevailed upon Archimedes to make him engines accommodated to all the purposes, offensive and defensive, of a siege. These the king never himself made use of, because he spent almost all his life in profound quiet, and the highest affluence. But the apparatus was, in most opportune time, ready at hand for the Syracusans, and with it also the engineer himself."

When the Romans assaulted the walls in two places, then fear and consternation stupefied the Syracusans, since they believed that nothing could resist the attacking forces. But then Archimedes started to use his war machines, and sent masses of stones and arrows against the land forces. The impacts when the cascades of missiles hit was unbearable. The troops were knocked down and fell over each other, breaking up the attack.

[6] This refers undoubtedly to the problem of constructing the double mean proportions, required for doubling the cube.

The ships did not fare better: Huge poles were thrust out from the walls over the ships and sunk some by knocking them down, while others were lifted up into the air by ropes with hooks and pulleys, and then plunged down again from a considerable height. The results were disastrous for the Romans.

When the Roman catapult, mounted on the bridge over eight ships approached, Archimedes had his own catapults hurl several boulders weighing about half a ton each against it. They struck with immense force and noise like thunder, and completely destroyed the catapult.

Lenses and mirrors were employed in setting approaching warships on fire before they could even get close.

The Romans were understandably frightened, and Plutarchus relates that it even went so far that if a Roman vessel had succeeded in approaching the wall, then it sufficed to lower an innocuous piece of rope from the wall in order to frighten it away.

It is not surprising that Marcus Marcellus had ordered his troops to take this man alive, and not to harm him in any way. He should be treated with respect and honors, and brought back to Rome.

One can only speculate on what might have happened to Roman science and mathematics if Marcus Marcellus' plan had succeed. The Romans certainly could have needed someone like Archimedes in Rome. If he had been brought to Rome and founded a Roman Academy there, history might have taken a different course.

But at this moment the following happened[7]: As a Roman soldier stormed into Archimedes' study with sword in hand, the latter sat immersed in geometrical considerations.

He had drawn geometric figures, circles, in the *sand*, as some say. Probably it was *fine dust of glass*, spread out over his drawing-board. This was the scratch paper of antiquity and almost up to our own age, used for writings not to be preserved. Permanent writing material was far too precious to be wasted as scratch paper. Anyway, the soldier demanded Archimedes' name, as he did not know his appearance. But Archimedes had been so engulfed in his geometry, that he was quite unable to say his name. All he could do, was to stutter: *–Please do not touch these* –, as he pointed to his drawing-board. The Roman soldier saw in this remark a lack of due respect for the mighty Roman power, and responded by striking him down with his sword. Thus ended the most remarkable scientific genius of antiquity, perhaps even of the entire human history as we know it.

Marcellus, who was very much saddened and disappointed by the loss of Archimedes, gave orders that he should be buried with full honors. Apologizing to his family, he also carried out Archimedes' wish, stated in his testament, that his most important geometrical theorem be engraved on his tombstone.

Eutocius of Ascalon (see Sect. 4.23) wrote commentaries, among others on Archimedes' work *On the Sphere and the Cylinder*. Based on two of the best Greek

[7] Here we follow one of several competing anecdotes or legends. Another version which is even more pathetic, is that Archimedes was on his way to Marcellus in order to surrender, carrying with him an astronomical instrument (an astrolabe), and a Roman soldier mistook him for a looter.

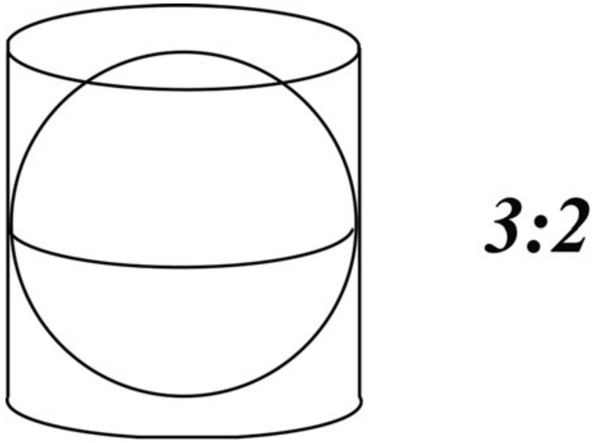


Fig. 4.7 The monument on Archimedes' grave bore this inscription

manuscripts from Archimedes, with Eutocius' commentaries, William of Moerbeke made a translation into Latin around 1270. The original manuscripts have since disappeared. Archimedes regarded the theorem proved here to be his most profound discovery (Fig. 4.7).

That this inscription really stood on Archimedes' grave, is attested to by *Cicero*, in his *"Tusculian Dialogues"*. We tell more about him in Sect. 4.8. He found the grave 140 years after Archimedes' death.

In the year 75 B.C. he was quaestor at Sicily. In "Tusculian Dialogues" he asserts that virtue is more important for human happiness than power and wealth.

"– Only the wise is really happy," Cicero writes. As proof for this claim he invokes the memory of the tyrant Dionysus, who he compares to Plato and Archimedes. He continues as follows[8]:

" – But from Dionysus' own city of Syracuse I will summon up to life a humble man in a modest position, from his dust and drawing-board: Archimedes. How can anyone who has ever had the slightest contact with humanity and scholarship not wish to be like this mathematician, rather than the afore mentioned tyrant?

While I was quaestor in Sicily I sought out his grave, which was unknown to the people of Syracuse. Indeed, they denied that any such thing existed. As it happened, the grave was hidden, on all sides overgrown by thorny shrubs. However, I remembered some simple lines of verse, which I knew were inscribed on his monument. They relate that upon his grave there was set a sphere with a cylinder, modelled in stone.

[8] This quote is based on two translations. Mainly I rely on a translation into Norwegian made for *Viggo Brun*, and used in his excellent book [5]. But I have also supplemented the translation with elements from the one printed in *Michael Grant* [20].

By the Agrientine Gate there is a large number of old graves. And as I looked everywhere, my attention was caught by a small pillar, which did not reach much above the thorny shrubs. On it there were rendered a sphere and a cylinder.

The very noblest of the men from the city were with me, and I said to them at once that this was what I sought. Workers with scythes were summoned, and cleared away the scrubs, opening up access to the place. As now entrance had been made possible, we proceeded to the front of the tombstone. And there the inscription became visible. But the last part of the verse was gone, about half of it.

Thus the noblest of Greek cities, once the most enlightened of them all, would have remained ignorant of the tomb of the most brilliant citizen it had ever produced, had it not been for a man from Arpinum."

Cicero gave orders that the grave should be preserved from then on, but for how long this was done we do not know. The grave was forgotten once more, and although there have been rumors from time to time that the grave has been found, no authoritative photo of the monument seems to exist.

The inscription on Archimedes' grave displayed what he considered his most profound mathematical theorem, concerning a sphere and its circumscribed cylinder: The proportion of the volumes of the circumscribed cylinder to that of the sphere equals the proportion of the surface areas of the same bodies, counting of course top and bottom of the cylinder. This common proportion is $1\frac{1}{2}$. Or, actually, it would have been written as $3 : 2$, since the Greek regarded this as a *ratio*, not a *number* as we do today.

We now turn to the description of some of Archimedes' geometry. A full account is of course impossible to give here, but we refer to Heath's [26] for a more complete treatment.

In 1906 the Danish classical philologist *Johan Ludwig Heiberg* visited Constantinople to study a parchment from the Saint Sepulchi monastery in Jerusalem. This was a so called *palimpsest*, where an original Greek text had been scraped off and replaced by a text of religious contents. Heiberg realized that the text which had been scraped off, contained among other things the priceless book by Archimedes entitled *The Method*, which had been presumed lost. Fortunately the attempted destruction of the Archimedes' works had served to preserve and protect them through the centuries of darkness. For not only did it turn out to be possible to restore Archimedes' original text – or rather, the text copied by a scribe from earlier copies – but the sacral status of the replacement text had protected the parchment from being destroyed. It turned out that the palimpsest contained *On the Sphere and the Cylinder*, almost all of *On Spirals*, and fragments of some other works which are preserved elsewhere. And then it contained *The Method*.

The book is now being studied and restored at the *Walters Art Museum* in Baltimore [1].

In the introduction Archimedes explains how he discovered these theorems using mechanics. He studied certain elements in equilibrium, and concluded from that relations between surface areas or volumes. But he emphasized that he would not consider this *as proofs*, only that *"it is easier, when we have found by this method* *some knowledge about the problem, to find the proof than it would have been to find it without such prior knowledge*.

Archimedes finds the area of a segment of a parabola and the volume of a sphere. We shall treat the latter result.

Archimedes actually leaves it as open whether these methods may be developed into fully valid *proofs*. But he writes: *"I am convinced that this method is not less useful in also proving these statements."* Today it is not difficult to accept this, since we find ideas here which much later developed into modern calculus and definite integration.

In modern language Archimedes considers a sphere of diameter a and volume denoted by V, a cone and a cylinder both with base of radius a and height a with volumes denoted, respectively, by K and S. Archimedes then proves the relation

$$V + K = \frac{1}{2}S.$$

Archimedes proved this equality by first showing that we have the equilibrium displayed in Fig. 4.8.

We assume that the sphere, the cone and the cylinder all have density 1. Let us first assume that the three bodies shown in the figure are in equilibrium. Then the equality of momentum yields that

$$a(V + K) = \frac{a}{2}S,$$

since we may place the total mass of the cylinder in its center of gravity, $\frac{a}{2}$ from the point of suspension. Thus

$$V = \frac{1}{2}S - K$$

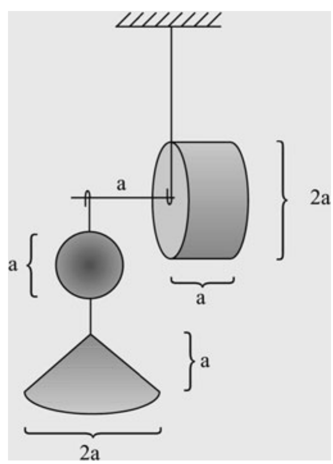


Fig. 4.8 The sphere, the cone and the cylinder at equilibrium

and from this Archimedes could find the volume of a sphere of *diameter* a, since the volume of a cone and a cylinder was known.

This relation yields:

Sphere and Cone. The volume of a sphere is equal to the volume of a cone with base four times as big as the area of a great circle in the sphere and height equal to the radius of the sphere. So the cone has base with radius equal to the diameter of the sphere.

We easily deduce this from the relation $V = \frac{1}{2}S - K$. In fact, we have $S = 3K$, so $V = \frac{3}{2}K - K = \frac{1}{2}K$. If K' is the volume of a cone with *half* the height of K and the same base, then $K' = \frac{1}{2}K$, thus $S = K'$.

In modern notation this relation implies $S = \frac{4}{3}\pi r^3$, where r is the *radius* of the sphere.

To prove the equilibrium of Fig. 4.8, we consider three circular slices, of a very small thickness Δ, as shown in Fig. 4.9. Here x is a number between 0 and a. We now prove that any such configuration is in equilibrium.

The circular slice to the right, the one which has been cut out of the cylinder, has volume equal to $A = \Delta\pi a^2$, and the slice which is cut out of the sphere has volume $B = \Delta\pi y^2$, so that $B = \Delta\pi(ax - x^2)$. Finally the slice cut out from the cone has volume $C = \Delta\pi x^2$. The three slices are in equilibrium if

$$a(B + C) = xA$$

or

$$a(\Delta\pi(ax - x^2) + \Delta\pi x^2) = x\Delta\pi a^2$$

which clearly holds.

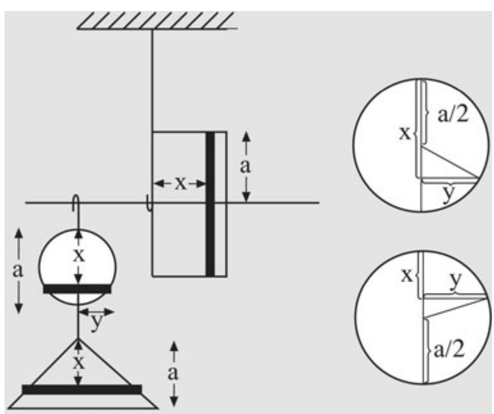


Fig. 4.9 Three circular slices in equilibrium. x is any number between 0 and a, and each slice has a very small thickness Δ. The slice inside the sphere has radius y, the one inside the cone has radius x and the slice inside the cylinder has radius a. To the right we have deduced that $y^2 = ax - x^2$: If $x \le \frac{a}{2}$ we use the lower circle, and get by Pythagoras that $y^2 = (\frac{a}{2})^2 - (\frac{a}{2} - x)^2 = ax - x^2$, and if $x > \frac{a}{2}$ we use the upper one which yields $y^2 = (\frac{a}{2})^2 - (x - \frac{a}{2})^2 = ax - x^2$

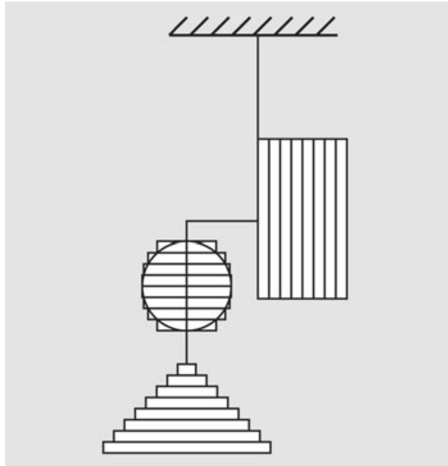


Fig. 4.10 Eight "slice-configurations," which all are in equilibrium, implies equilibrium for the composite configuration

We finally show that this implies the equilibrium of Fig. 4.8. In fact, we now let x grow from 0 to a in steps of Δ. Then the corresponding *"slice-configurations,"* when taken together, will constitute the configuration shown in Fig. 4.10.

By making Δ smaller and smaller, and the number of parts correspondingly bigger, we finally get the equilibrium of the sphere, cylinder and cone as claimed.

From his discovery that the volume of a sphere is equal to the volume of a cone with base of area four times the area of a great circle of the sphere, and with height equal to the radius of the sphere, Archimedes found this result:

> *Area of the Sphere.* The surface area of a sphere is four times the area enclosed by a great circle.

Archimedes gives the following argument for this conclusion: *"In the same way that any circle encloses an area equal to the area of a triangle with base equal to the circumference of the circle and height equal to the radius, it is reasonable to conclude that the volume of a sphere is equal to the volume of a cone with base equal to the surface area of the sphere and height equal to the radius."*

In Fig. 4.11 there are two sketches. To the right there is a regular 12-gon, inscribed in a circle. By adding the areas of the 12 triangles, we find that the area of the 12-gon is the circumference of the 12-gon multiplied by the height of the triangles, divided by 2. If we divide all the central angles in two and proceed to the regular 24-gon, we get the same: The area of the 24-gon is the distance from the sides to the center, multiplied with the circumference and divide by 2. Repeating this, we approach the circle, and the claim follows. In order to get the analogous result for the sphere, we divide the surface of the sphere by circles, as shown to the right. We get a net on the surface and an inscribed (non-regular) polyhedron. The volume of this polyhedron will be approximately equal to the radius multiplied by the surface area of the polyhedron divided by 3, as it is composed of several *pyramids*. Using a finer and finer net of circles, we get the claim in the limit. But apart

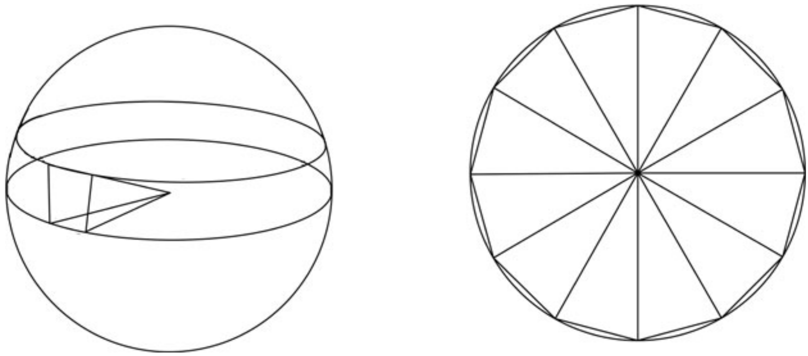


Fig. 4.11 Archimedes' idea for finding the surface area of a sphere

from the anachronistic term *in the limit*, this is only a way of making the claim *plausible*, not of proving it. In fact, the heights involved are not equal, even if they all approach nearer and nearer to the radius.

Archimedes' result on the area of a circle is Proposition 2 in his book *Measurement of a Circle.* This book only contains three propositions, the third is an inequality which is equivalent to the following, in modern notation:

$$\frac{22}{7} = 3\frac{1}{7} > \pi > \frac{223}{71} = 3\frac{10}{71}.$$

He finds this by computing the perimeters of regular 96-gons inscribed in and circumscribed about a circle.

In *On the Sphere and the Cylinder* Archimedes provides full proofs of these claims, and more.

We finally deduce the theorem inscribed on Archimedes' tombstone, using the methods (although not the terminology) which he himself would have employed. We consider a sphere V and a circumscribed cylinder S. The cone with the same base and height as the cylinder will have volume K_1, and we have the relation

$$S = 3K_1$$

We consider the cone K_2 which have base with radius equal to the diameter of the sphere and height equal to the radius of the sphere. Then

$$K_2 = 2K_1$$

and since $K_2 = V$, we have

$$S : V = 3K_1 : 2K_1 = 3 : 2.$$

We now come to the surface areas. We denote the area enclosed by a great circle of the sphere by s. Then the surface area of the sphere is $O = 4s$. The cylinder has a base and a top, the combined area of which is $2s$. The cylinders surface itself has baseline equal to the circumference of the circle and height equal to its diameter. Half of this, with height equal to the radius, has area equal to twice the triangle with the same base and height, in other words $2s$, by Archimedes' remark quoted above. The surface area of the cylinder proper is therefore equal to $4s$, so the surface area counting base and top is $A = 2s + 4s = 6s$. We thus have

$$A : O = 6s : 4s = 3s : 2s = 3 : 2$$

Archimedes described the *semiregular polyhedra*. There are altogether 13 of them (which are not regular), they are referred to as the *Archimedean Polyhedra* or as the *Archimedean Solids*. They are defined similarly to regular polyhedra in Sect. 3.10. We still require that all the sides be regular polygons, but now they are not required to be the of the same kind. More precisely, we say that a polyhedron is semiregular if it is *convex* and the faces are *regular polygons, but not necessarily of the same kind*. For the Platonic Solids we required that all polyhedral angles be congruent. In the Archimedean case this is no longer sufficient, as we shall see below. Instead we impose the condition that the polyhedra be *vertex transitive*, a condition which can only be precisely stated in terms of the *group of symmetries* of the polyhedron, but which may be loosely expressed by saying that *all vertices are equivalent*. We return to this in more detail in Sect. 6.9. For now we shall follow [9] and say that this means that the polyhedron in question looks the same when viewed with any of its vertices directed forward. We shall explore this in more detail below when we turn to the interesting example known as *Miller's Polyhedron*.

In addition to the Archimedean solids, this definition is also satisfied by two infinite families of polyhedra, namely the *n-gonal prisms* and the *n-gonal antiprisms*. The n-gonal prism is a prism with all sides equal, whose base is a regular n-gon, while the antiprism is obtained by twisting the top by an angle $\frac{\pi}{n}$ and filling the gap between the base and the top with equilateral triangles.

They are not included among the Archimedean solids. But two special cases *are regular polyhedra*, namely the 4-gonal prism, that is the *cube*, and the 3-gonal antiprism, the *octahedron*.

In Fig. 4.12 the Archimedean Solids are presented together with the 6-gonal prism and the 6-gonal antiprism. In the top row we have the truncated cube, the truncated tetrahedron, the cuboctahedron and the icosidodecahedron. The second row contains the truncated dodecahedron, the truncated icosahedron, the small rhombcuboctahedron, and the small rhombicosidodecahedron. The third line shows the truncated octahedron, the great rhombicosidodecahedron, the snub cube and the snub dodecahedron. The last line contains the great rhombcuboctahedron, as well as a prism and an antiprism.

The two simplest Archimedean solids are located in the middle of the first column. They are obtained as shown below, by a process which is known as *truncation* from the tetrahedron and the cube, respectively. This process yields new semiregular

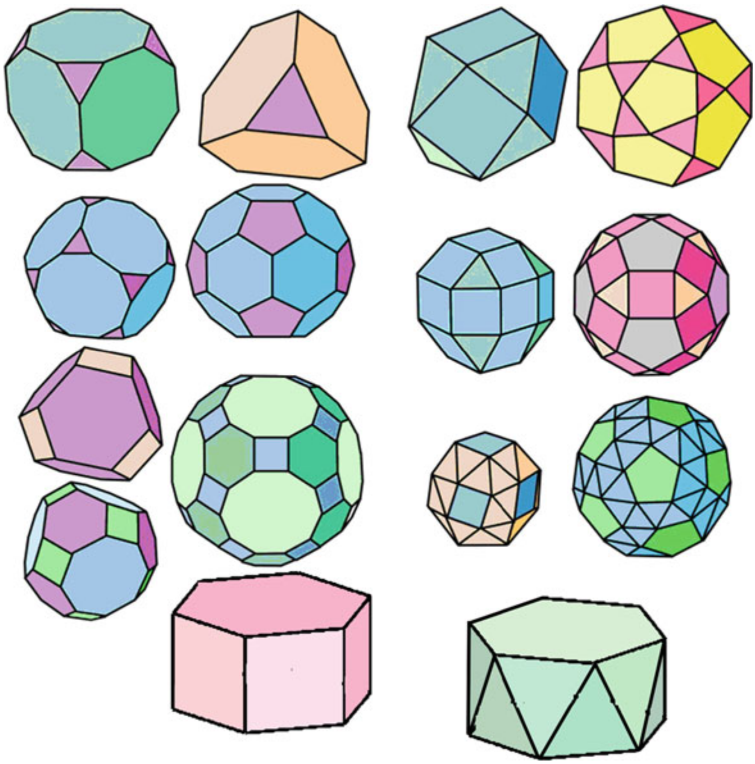


Fig. 4.12 The 13 Archimedean Solids together with a prism and an antiprism

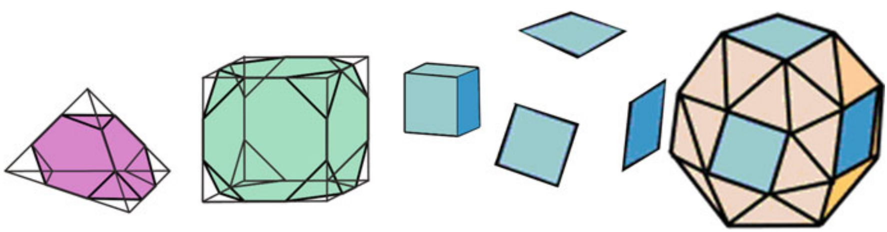


Fig. 4.13 Truncation and snubification

polyhedra when applied to all five Platonic Solids, as well as to the cuboctahedron and the icosidodecahedron. The process of *snubification* of a polyhedron consists, roughly speaking, in cutting loose the sides, lifting them out, twisting them all a certain angle clockwise or counterclockwise, and then filling the gaps with regular triangles, as we see in Fig. 4.13 for the snub cube.

An important invariant of a regular or semiregular polyhedra is a tuple of integers: For example, $(3, 6, 6)$ or equivalently $(6, 6, 3)$ or $(6, 3, 6)$ for the truncated tetrahedron indicates that at all vertices an equilateral triangle and two regular hexagons meet, while the *small rhombicosidodecahedron* has the tuple $(4, 3, 4, 5)$ which indicates that at each vertex a square, an equilateral triangle, a square and one regular pentagon meet in that order up to cyclic rearrangement. The n-gonal prism is a $(4, 4, n)$ while the antiprism is $(3, 3, 3, n)$.

In the third and fourth column of the third line of Archimedean Solids in Fig. 4.12, we find the *snub cube* and the *snub dodecahedron*, respectively. They both have mirror images which are not congruent with themselves, they are so-called *chiral objects*, they come in a left-handed form and right-handed form like our two hands. Thus if we count these mirror images separately we get 15 Archimedean solids instead of the 13 listed here.

The need to strengthen the condition that all polyhedral angle be congruent, is demonstrated by a polyhedron discovered by *J. C. P. Miller.* Like the small rhombcuboctahedron it is given by (3,4,4,4), but it lacks the full symmetry of that polyhedron. Miller's Polyhedron is shown in Fig. 4.14, together with the small rhombcuboctahedron as explained below.

In order to make a model of the small rhombcuboctahedron two bowls are made, pieced together of equilateral triangles and squares with sides of equal lengths. The smaller bowl is turned around and glued together with the larger one in two different ways. Above the result is the small rhombcuboctahedron. Below the larger bowl is rotated slightly, and then the result is Miller's Polyhedron.

Turning the lower bowl destroys much of the symmetry of the small rhombcuboctahedron, for example, only one axis of rotational symmetry from the small rhombcuboctahedron remains. And Miller's Polyhedron is not vertex transitive, which is actually quite clear from the pictures in Fig. 4.14. We shall return to this in Sect. 20.6.

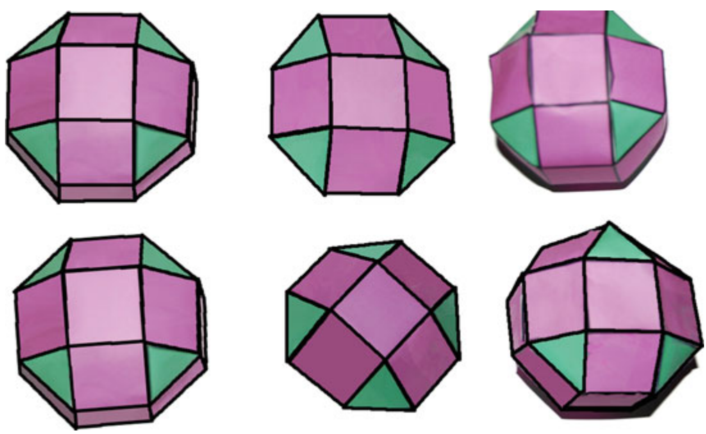


Fig. 4.14 The small rhombcuboctahedron above right, compared to Miller's Polyhedron below right. To the left we see how they are produced by making two bowls and gluing them together in two different ways

It is difficult to believe that Archimedes did not know the polyhedron (re)discovered by Miller. So why did he not include it, to make 14 rather than 13 semiregular polyhedra?

In his book *On Spirals* Archimedes studies spirals of the type which today is known as *Archimedean Spirals.* He defines a spiral as follows:

> *Archimedean Spiral.* If a straight line where a point is kept fixed rotates about that point with even velocity until it comes back to the point of departure, and at the same time a point, starting from the fixed point moves with even velocity outwards along the line, then the point describes a spiral in the plane.

One of the reasons for Archimedes' interest in the spiral was his work on *squaring the circle*.

As we have seen above, Archimedes had drawn certain conclusions from the fact that the area enclosed by a circle is equal to the area of a triangle with base equal to the circumference and height equal to the radius.

Another conclusion to be drawn from this fact is the following: *The problem of squaring the circle is equivalent to the problem of rectifying the circle*. In other words, if we are given a circle with center O through P, then we should try to construct a point T on the normal to OP at O, such that the length of OT is equal to the circumference of the circle. Then all we have to do is to construct a square with area equal to the area of $\triangle OTP$. The latter one was a very well known construction to Greek geometers.

Archimedes performed the rectification of the circle by means of the spiral. In Fig. 4.15, an approximation to such a spiral is drawn. It has been created by letting a line rotate counter clockwise about the point O. Twelve straight lines were drawn through O, spaced at 30°. Along the lines points were marked $1, 2, 3, \ldots, 12$ times a certain distance Δ, then we connected these points by line segments, which taken together approximate the curved line of a spiral.

This curve will intersect the first of the 12 lines at a point P, at this point we construct the tangent to the spiral. The tangent intersects normal to OP in the point T. Now Archimedes proved that *the length of OT is equal to the circumference of the circle about O through P*, in our notation it is equal to $2\pi r$, where r is the length of OP.

Archimedes studied two intriguing figures, the *arbelos* or "shoemaker's knife," and the *salinon* or "salt cellar," bounded by semicircles. They are shown in Fig. 4.16.

We return to some of the many fascinating properties of the arbelos in Exercises 4.3, 4.4 and 4.5.

Archimedes gave a construction of the *regular heptagon*, the regular 7-gon.

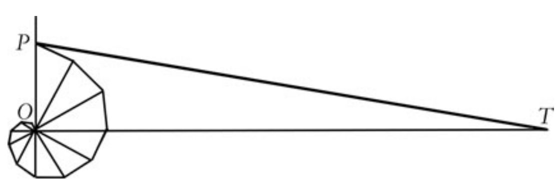


Fig. 4.15 An Archimedean spiral

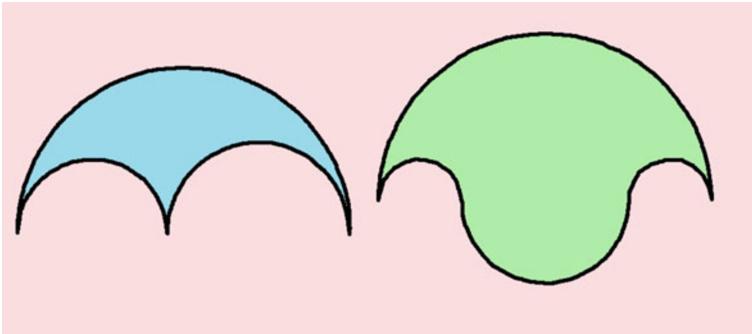


Fig. 4.16 The arbelos to the left, and the salinon to the right

Fig. 4.17 Thabit Ibn Qurra. Drawing by the author

We know this construction through an Arabic translation of an original from Archimedes, which is lost. In fact, according to [58] *Thabit Ibn Qurra*, who wrote the Arabic text, complains about the poor condition of the Greek original, he barely could make out the proofs (Fig. 4.17).

Before we turn to Archimedes' construction, we shall treat the problem in a modern setting. We are going to consider the problem of constructing a regular 7-gon, referred to as a heptagon, using only ruler and compass.

We refer to Fig. 4.18. Since the circumference is subdivided in 7 equal parts, the angles u at the periphery are all $\frac{\pi}{7}$ radians, or $\frac{180}{7}^{\circ}$, being half the measure of their intercepted arcs. We thus find, for instance, that $\angle E'EA = u$, and so on: In Fig. 4.18 the various multiples of u are indicated by arcs, one arc indicating u, two indicating $2u$, etc. Those arcs found as angles at the periphery are fully drawn, those obtained from the angular sum of triangles are dotted.

Denoting the length of the side of the heptagon by s, we find that $AA' = A'E' = E'E = A'B = BE = s$ and also that $AK = s$ by the symmetry. We denote BF by x and FA by y. We then have $AF = FA' = y$ and $FK = FB = x$. The first diagonal is $A'E$, it will not be needed, but the second diagonal AE will be important, its length is denoted by d, so $d = s + x + y$.

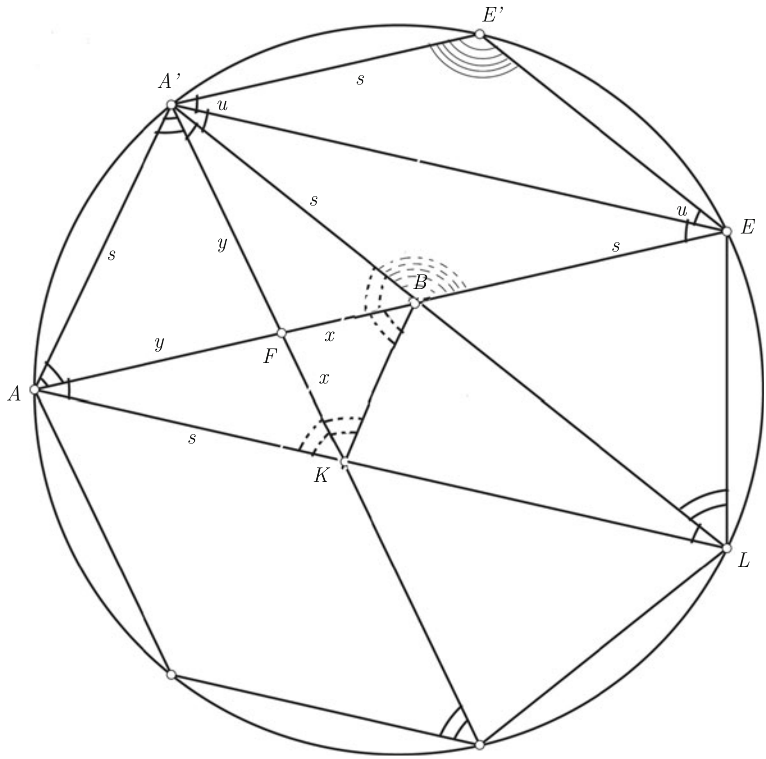


Fig. 4.18 Analysis of the regular heptagon

We now find two similar triangles and one mirror image of a similar triangle shown in Fig. 4.19, and obtain the proportions

$$\frac{y}{x} = \frac{s+x}{y}, \ \frac{y}{s} = \frac{s}{x+y}$$

and we therefore have the equations

$$y^2 = x(s+x), \ y(x+y) = s^2, \text{ and } \quad d = s+x+y$$

thus

$$(d-s-x)^2 = x(s+x), \ (d-s-x)(d-s) = s^2$$

Solving the latter for x and substituting into the former yields, after a straightforward computation,

$$d^4 - sd^3 + s^2d^2 + 2ds^2 - s^4 = 0.$$

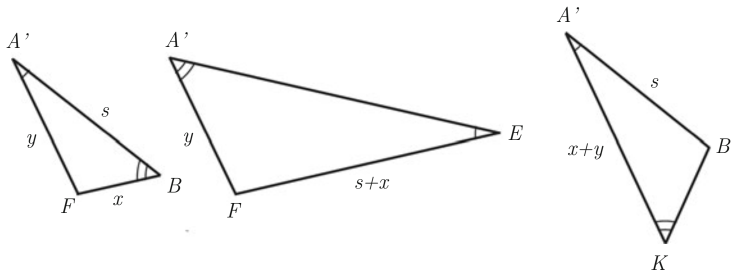


Fig. 4.19 The important triangles

Now $d = s$ satisfies this equation, but then $x = y = 0$ and the original proportions become meaningless. So this is a *false solution* introduced by multiplying an equation by an expression which might be zero. We get rid of it by dividing with $d - s$, and obtain the following relation between s and d:

$$d^3 - 2sd^2 - s^2d + s^3 = 0.$$

Since it is not difficult to scale any construction up or down using ruler and compass, we may normalize the situation by taking $s = 1$. Then our equation becomes

$$X^3 - 2X^2 - X + 1 = 0,$$

where we have replaced d by X. It is not difficult to see, without using a calculator, that this equation has three distinct real roots, one is negative and another is positive but less than 1. In fact, using a calculator we get the roots

$$X_1 \approx -0.8019377358, X_2 \approx 0.5549581321 \text{ and } X_3 \approx 2.246979604$$

The two first roots cannot give the length of a diagonal in the regular heptagon with side 1, and therefore have to be discarded in our analysis. There remains the root X_3.

Proposition 1. *The regular heptagon with side 1 can be constructed with ruler and compass if and only if X_3 can be constructed.*

Proof. If we have managed to construct the regular heptagon with side 1, then we only need to draw the second diagonal. Conversely, suppose we have constructed X_3. We then complete the construction as shown on the illustration in Fig. 4.4.

The distance between A and B is X_3, around A and B are drawn circles of radius 1 (=s), on both sides of the mid-normal to AB are erected normals at the distance $\frac{1}{2}$, these intersect the two circles in points C and D. We now have four points A, C, D and B on the regular heptagon with side 1. Three copies of this construction are easily assembled to the heptagon as shown in Fig 4.20.

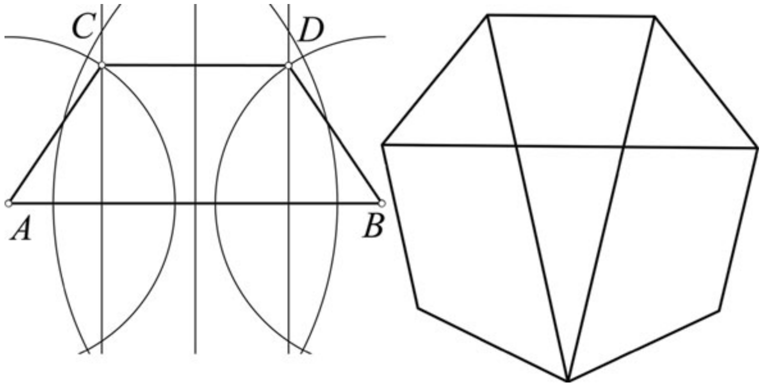


Fig. 4.20 Completing the construction

However, the equation $X^3 - 2X^2 - X + 1 = 0$ may be used to prove that the construction is impossible by *legal use* of compass and straightedge. We return to this in Sect. 17.7. But the relation also explains that the construction is possible by compass and *a marked straightedge*. We return to this in Sect. 4.6. An actual construction using a marked straightedge was found by the great French geometer and algebraist *Viète* (1540–1603). For the details we refer to [25].

We now return to Archimedes' construction. This construction is an example of what the Greek geometers called a *"Neusis"*-construction, it solves a *verging-problem*:

A line is made to slide so that it passes in a certain direction and realizes a specified condition. Typically, this happens with constructions involving a marked straightedge, but here we see a different version. We consider Fig. 4.21.

We see a square $ABCD$. The diagonal AC is fixed, but then the line DE produced is made to rotate about D. The point E is where this line intersects AB produced. Now the line is rotated about D until the following happens: *The areas of* $\triangle DGC$ *and* $\triangle BEH$ *are equal.* We may think of Fig. 4.21 as displaying some kind of mechanical instrument. For example, the lines are metallic rods, $\angle DAB$ and $\angle BCD$ being fixed right angles, so that G may slide up and down along AC, the latter rod being fixed in the diagonal position, at 45°. Denote the distance between B and E by s.

When this arrangement is achieved, then *AE will be the second (the largest) diagonal in the regular heptagon with side* s.

We now put $BE = s$, and $AF = y$, $FB = x$. Denote $\angle GDC$ by v. Since $\angle GDC = \angle HEB$, we have

$$\text{cotan}(v) = \frac{s + x}{y} = \frac{y}{x}$$

and the equality of the areas of $\triangle GDC$ and $\triangle BEH$ gives

$$x(x + y) = cs$$

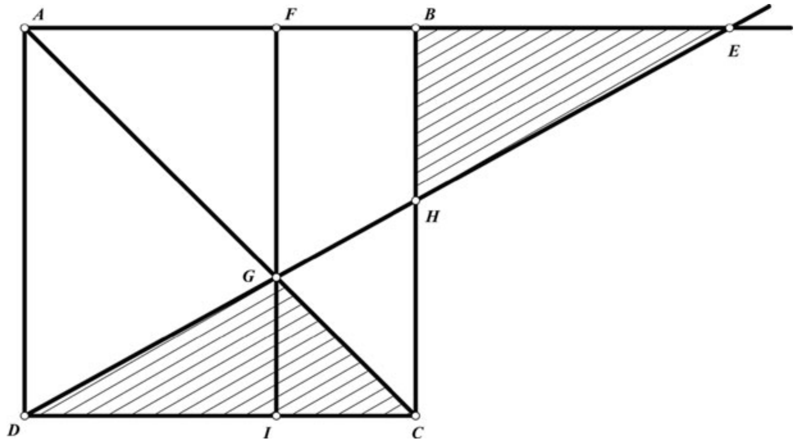


Fig. 4.21 A mechanical instrument for constructing a regular heptagon

where c is the length of BH. Substituting $c = \frac{sy}{s+x}$ into the last equality we get

$$x(x+y) = \frac{sy}{s+x}s = s^2\frac{y}{s+x} = s^2\frac{x}{y}$$

by the first relation, and hence

$$y(x+y) = s^2$$

or

$$\frac{y}{s} = \frac{s}{x+y}.$$

We therefore have shown the same proportions as we had in the analysis of the regular heptagon above. Thus s is the side in the regular heptagon having the second diagonal equal to d. This completes the proof of Archimedes' construction. □

Archimedes is given credit for numerous ingenious inventions. One such invention is the *Archimedean Water Screw.* Pumps of this kind have been in use up to our time in Egypt. At Archimedes' time the pump might be operated by a slave, who would turn the shaft by a treadmill.

4.5 Eratosthenes and Doubling the Cube

Eratosthenes was born in 276 in Cyrene and died in 194 B.C. in Alexandria (Fig. 4.22).

He was the director of the school at Alexandria, and was an outstanding representative for the refined culture which flowered at the Royal Court there. Some see a

Fig. 4.22 Eratosthenes

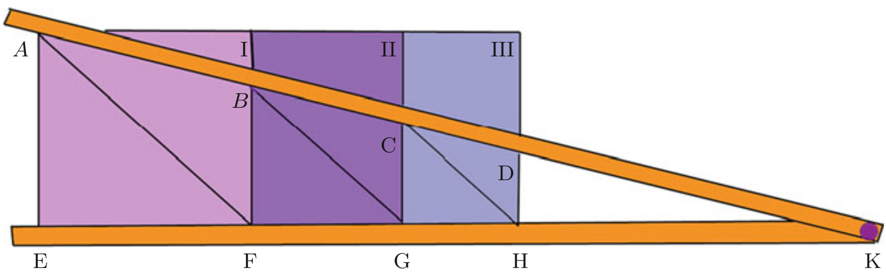


Fig. 4.23 Eratosthenes' instrument for the construction of continued proportionalities

certain disdain for Eratosthenes in some of Archimedes' writings, others disagree in this interpretations. At any rate, while Archimedes and his fellow citizen in Syracuse were fighting off the Romans, their backs against the wall, Eratosthenes and his contemporaries in Alexandria were able to pursue the refined art of dialectics, poetry and rhetoric as well as classical literature and mathematics.

Here we shall only treat a small part of the scientific work of Eratosthenes, one of direct relation to Greek geometry: He invented a mechanical instrument for the construction of the continued mean proportionality, to which the problem of Doubling the Cube had been reduced (Fig. 4.23).

He appears to have been rather pleased with his own invention. Dedicating the device to the king Ptolemy, he had a description of it engraved on a monument which was erected in the kings temple. Eloquently praising the king, a model of the apparatus, in bronze, was put on top of the monument. In van der Waerden's

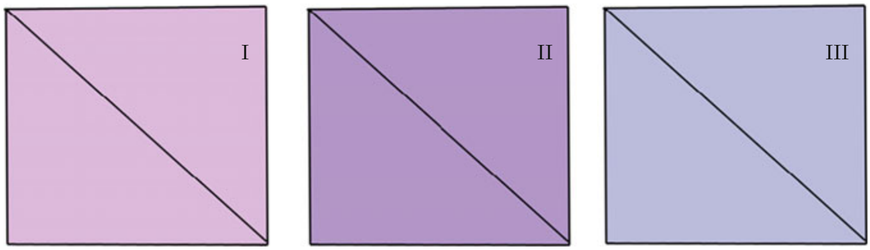


Fig. 4.24 Three of the frames in Eratosthenes' instrument

words, in [58]: *"The subtle complement to the king and his son, which occurs in the epigram, betrays the well-versed courtier."*

A horizontal straightedge was equipped with a grove, in which three or more rectangular frames could slide. On these frames a diagonal had been engraved. In Fig. 4.24 we show three such frames, then a double proportionality may be determined.

The frames were slid along the grove, so that the one marked I would be the outer end, then II and the innermost III. A second straightedge was fixed at K and could be rotated about this point. The frames were slid between the two straightedges as indicated in the figure.

If the points A, B, C and D are on a straight line, then we get four similar triangles, namely

$$\triangle AEK \sim \triangle BFK \sim \triangle CGK \sim \triangle DHK.$$

In addition the diagonals of the three plates connect A and F, B and G and C and H. We therefore have three more similar triangles

$$\triangle AFK \sim \triangle BGK \sim \triangle CHK$$

Altogether these relations yield

$$AE : BF = BF : CG = CG : DH$$

and we find that BF is the first and CG the second mean proportionality between AE and DH.

If we wish to find the mean proportionalities between a and b, we may assume that $a > b$. If not, then we interchange a and b, and interchange the two mean proportionals as well: Indeed, if

$$\frac{a}{x} = \frac{x}{y} = \frac{y}{b}$$

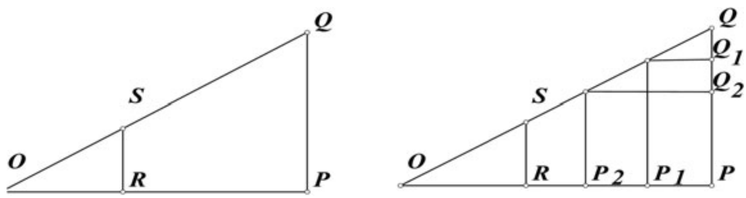


Fig. 4.25 The construction with arbitrary line segments a and b

then

$$\frac{b}{y} = \frac{y}{x} = \frac{x}{a}$$

Assume first that $a = AE$. We mark off b as HD on frame III. Then slide frames II and III and rotate the second straightedge, so that the points A, B, C and D fall along the second straightedge. Then $x = BF$ and $y = CG$ are the wanted mean proportionals.

For the general case, take $\alpha = AE$, in other words the height of the frame. We first find β such that

$$a : \alpha = b : \beta$$

This is done by a right triangle as shown to the left on Fig. 4.25, where we have $OP = a$, $OR = b$ and $PQ = \alpha$, the height of the frames in Eratosthenes' instrument. The parallel to PQ through R intersects OQ in S, we take $RS = \beta$. From the instrument we then find the first and the second continued proportionals ξ and η, respectively. We now look at the figure to the right, where we have marked off $PQ_1 = \xi$ and $PQ_2 = \eta$. We then find the first mean proportional x as OP_1, and the second as $y = OP_2$.

Eratosthenes is said to have become blind in old age and it has been claimed that he committed suicide by starvation.

4.6 Nicomedes and His Conchoid

Essentially nothing is known on Nicomedes' life. We estimate that he was born about 280 B.C., and died approximately 210 B.C. His main work is *On conchoid lines*, where his one and only great discovery is described, namely the *conchoid curve*. We shall return to it in Sect. 15.7, when we have sufficient geometric machinery to complete the study of this very interesting curve (Fig. 4.26).

The curve, or rather any finite number of points on it, is constructed as follows: A line ℓ and a point P at a distance a from ℓ is fixed. Then a circle with center P is drawn, and the circumference subdivided in n equal pieces, say with n some power of 2, in which case the subdivision is always possible with compass and

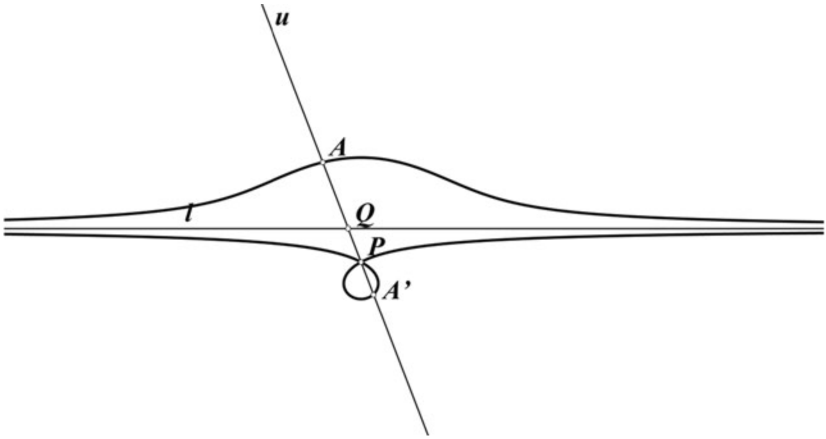


Fig. 4.26 The Conchoid of Nicomedes and how it is defined: Two lengths, or numbers as we would say today a and b are given. In the illustration here $a < b$. A line ℓ and a point P at a distance a from ℓ is fixed. We then draw a line u through P, intersecting ℓ in a point Q. On this line, at the distance b from Q we mark the points A and A'. They trace out the two branches of the conchoid when the line u rotates about P

straightedge.[9] Corresponding to the subdivision, lines are drawn through P. These lines all intersect ℓ, and on each line points are marked at a fixed distance b from the points of intersection with ℓ. These points all lie on the conchoid, and when $b > a$, as is the case in Fig. 4.27, the two sets of points on each line will trace out two branches of the curve.

Note that the conchoid, as with all higher curves considered by Greek geometers, are constructed by compass and straightedge, the Euclidian tools, but *with an infinite number of steps* in the construction. By this kind of constructions, known as *asymptotic Euclidian constructions*, the classical problems may be "solved." That is to say, if we allow a sufficiently large number of steps in the construction, the problems may be solved approximately with any prescribed degree of accuracy. The constructions we refer to below are of this nature. Indeed, by a large but finite number of steps we may approximate the curves used by a chain of small line segments, and using these approximations the classical problems may be solved approximately.

However, Nicomedes used a mechanical instrument to draw this curve, which is described by Heath in [26], Vol. I, p. 239. We follow Heath's explanation:

AB is a ruler, with a slot along the middle, and *CD* is a second ruler at right angle to *AB*. It has a peg P fixed in it. A third ruler *EF* has a slot along the middle which fits the peg P, and has a pointed end at E. We could, anachronistically, think of E being fitted with a pencil, to draw the curve on a piece of paper. Finally P' is a fixed peg on *EF* in a straight line with the slot, P' can move freely along the slot in *AB*.

[9] Of course an exact subdivision is not essential, but mainly serves to make the construction aesthetically appealing.

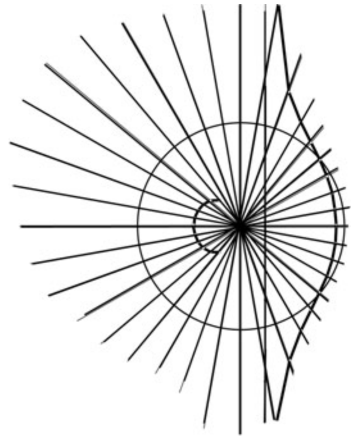


Fig. 4.27 Some points on the Conchoid of Nicomedes with $b = 3 > a = 1$

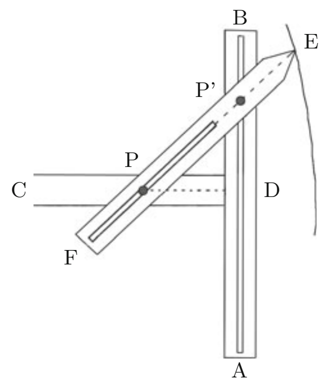


Fig. 4.28 Nicomedes' mechanical device for drawing the conchoid

If now the ruler *EF* moves so that the peg P' traverses the total length of the slot in *AB*, then E will describe a segment of one branch of the conchoid. See Fig. 4.28.

Nicomedes used the conchoid to solve the problem of trisecting an angle and the problem of doubling a cube.

Nicomedes was very proud of his invention, and wrote extensively comparing it to Eratosthenes' mechanical instrument for constructing the double mean proportional, thus in particular, for doubling the cube. He argued that such a mechanical procedure was alien to geometry, and in every way inferior to the method of the conchoid!

Indeed, the technique of the conchoid is more general. It provides a device to solve a large class of *verging problems*. In the present case it is the problem of inserting a line segment of fixed length between two given curves, usually lines or circles, so that the fixed line segment when produced will pass through a given point.

Doubling the cube, as well as trisecting any angle, may be reduced to two such verging problems.

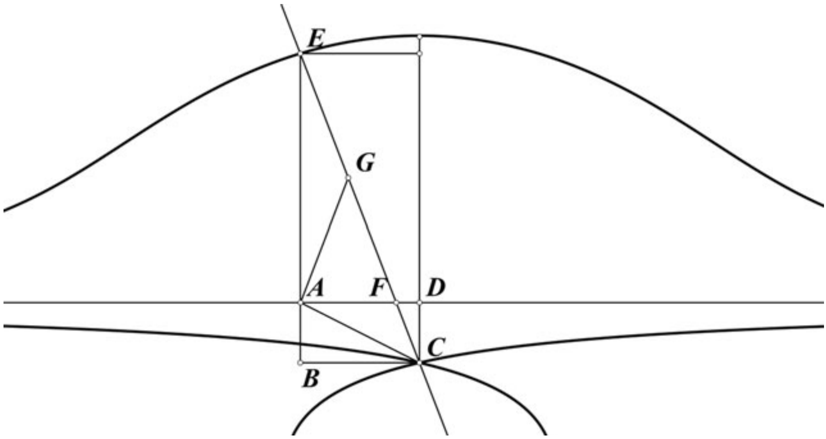


Fig. 4.29 Trisecting any angle using the conchoid

We shall first show this for the *trisection*-problem, and also show how the conchoid is used to solve it.

The construction is rendered in Fig. 4.29, which is pretty much self explanatory. It suffices, of course, to be able to trisect any angle less than a right angle. Such an angle u may be realized as the angle formed by a diagonal and a side in a rectangle $ABCD$, so let $u = \angle ACD$. Now produce DA to the line ℓ, and take C as the point P in the definition of the conchoid. b is taken to be twice the length of the diagonal AC. The corresponding conchoid intersects BA produced in E. The line CE intersects AD in F, and letting G denote the mid point of EF, we find $EG = GF = AG = AC$. Therefore, putting $\angle CEA = v$, we have $\angle AGF = 2v = \angle ACG$, and hence $v + 2v = 3v = \angle ACD = u$.

To double the cube with side a, we set of the line segment AB of length a, and at B erect a line normal to it. We also draw a line forming an angle of 120° with AB at B. See Fig. 4.30. The normal to AB at B is taken as the line ℓ in the definition of the conchoid, and we take $a = b$ in the construction of the conchoid (note that b has a different meaning in Fig. 4.30). We find the point D as shown in the figure, which is such that $\angle DBA = 120°$ and $CD = a$. Put $AC = b$, $BC = c$. Denote $\angle ADB$ by φ. We then have

$$\frac{a}{\sin(30°)} = \frac{c}{\sin(\varphi)}$$

and

$$\frac{a}{\sin(\varphi)} = \frac{a+b}{\sin(120°)}$$

from which we infer that

$$\frac{c}{a} = \sqrt{3}\frac{a}{a+b}$$

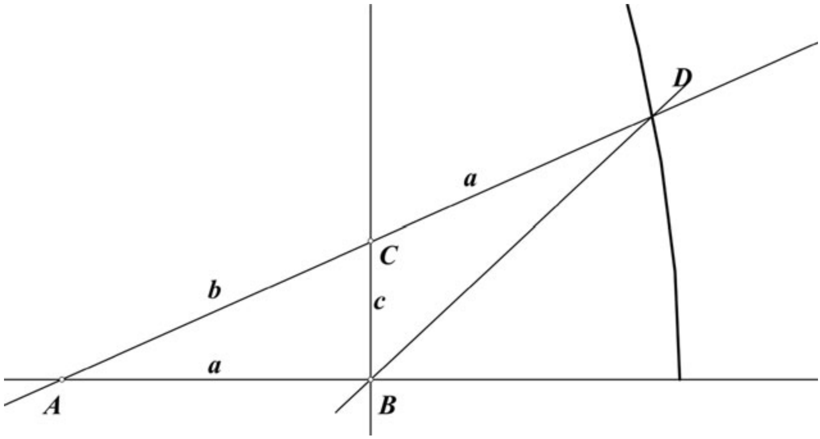


Fig. 4.30 Doubling the cube with side $a = AB$ using the conchoid

This yields

$$\sqrt{3} = \frac{c(a+b)}{a^2}$$

which when squared implies

$$3a^4 = c^2(a+b)^2 = (b^2 - a^2)(a+b)^2$$

Thus

$$3a^4 = b^4 - a^4 + 2ab^3 - 2a^3b$$

hence

$$2a^3(2a+b) = b^3(2a+b)$$

thus

$$2a^3 = b^3$$

Note that the construction obtained in Fig. 4.30 by means of the conchoid, may also be achieved with compass and a straightedge which is capable of moving a distance. To perform the construction, we set off A and B, with the length of AB equal to the side a of our cube. As before we erect the normal n at B and construct another line m forming an angle at B of 120°. Then the distance a is marked by two points C' and D' on the straightedge, which is subsequently slid into a position where C' falls on the normal n and D' falls on m while the straightedge passes through A. This is a typical *Verging Problem*, being solved here by the so called *Insertion Principle*. We may now complete the construction by drawing the line AD, which intersects n at C, such that the length of CD is a.

The construction shown in Fig. 4.29 may also be carried out in a similar manner. But this is not all: It suffices to be able to insert, by means of the straightedge, a *single, fixed, distance*. That is to say, construction by a *marked straightedge*. In fact, suppose we wish to insert a line segment of length a into a construction by means of a straightedge. We have a straightedge on which the distance b is marked off. Then scaling the unfinished construction by the ratio $b : a$, the distance b will correspond to a, which is inserted by the marked straightedge, after which the construction is scaled back by the ratio $a : b$. By this, admittedly cumbersome, procedure we may insert any distance by means of the fixed one.

Nicomedes also solved the third classical problem, the *squaring of the circle*, by means of a curve known as the *Quadratrix of Hippias*.

Hippias of Elis was born about 460 and died around 400 B.C. He travelled from place to place and earned his living from lecturing on a variety of subjects and providing other services. Plato did not think much of him, but regarded him as boastful and arrogant, with a wide but superficial knowledge of the subjects of his times. He was also known as *Hippias the Sophist*, and here we may find an explanation for a possible prejudice against him on Plato's part. But although covering a wide range of subjects, he *did* specialize in mathematics, and Heath gives a rather favorable assessment of him in [26].

The points P on Hippias' curve are defined as follows, with notation as in Fig. 4.31: Construct a square $ABCD$ with side AB. With D as center draw the quarter circular arc AC of radius DA. Let the point E move along this arc at uniform speed, and let at the same time A' move from A to D also at uniform speed, reaching D at the same time as E reaches C. Let P be the point of intersection between the lines DE and $A'B'$. P then traces out a curve, indicated in Fig. 4.31.

We then have

$$\angle EDC : \angle ADC = A'D : AD, \text{ thus } \angle EDC = \frac{A'D}{AD}\frac{\pi}{2}$$

using modern notation.

This curve may also be used to solve the trisection problem. Namely, if we are given an angle $u = \angle EDC$ we may trisect it by trisecting PQ by the point T,

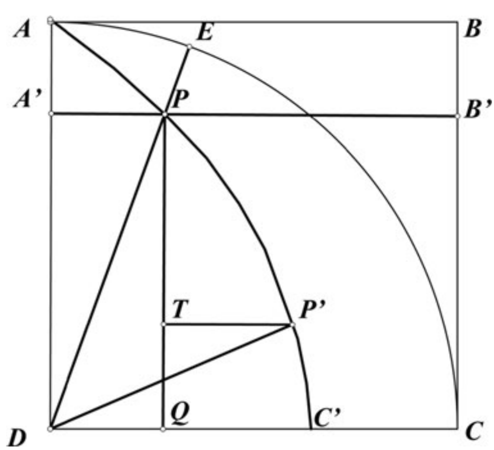


Fig. 4.31 The Quadratrix of Hippias

drawing the line through T parallel with DC, which intersects the quadratrix in the point P', then $\angle P'DC = \frac{1}{3}u$.

Again, we may find as many points on this curve as we want by subdividing the arc AC in n equal pieces, n being a power of 2, say, and dividing the line segment AD in n equal pieces as well.

The reason why Nicomedes was able to use this curve to square the circle, lies in the observation that, with modern notation,

$$\frac{DC}{DC'} = \frac{\pi}{2}.$$

To show this, let $\frac{A'D}{AD} = x$, then $\angle EDC = \frac{\pi}{2}x$, thus $\frac{DQ}{DC} = \frac{DA'}{DC}\frac{DQ}{DA'} = x\text{cotan}(\frac{\pi}{2}x)$. So

$$\frac{DQ}{DC} = \frac{x}{\tan(\frac{\pi}{2}x)}$$

i.e.,

$$\frac{DC}{DQ} = \frac{\tan(\frac{\pi}{2}x)}{x}$$

$Q = C'$ corresponds to $x = 0$, thus we need to use *l'Hôpital's Rule*, which yields

$$\frac{DC}{DC'} = \frac{\frac{\pi}{2}}{1} = \frac{\pi}{2}$$

proving the claim.

To complete the squaring of the circle, we construct the rectangle and the diagonal in Fig. 4.32. We have

$$\frac{DC}{DC'} = \frac{EF}{DC'} = \frac{AE}{AD} = \frac{\pi}{2}$$

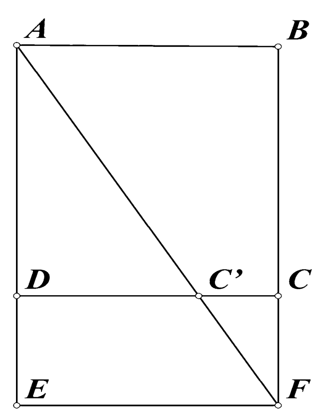


Fig. 4.32 The squaring of the circle according to Nicomedes

thus

$$2AE \cdot AD = \pi AD^2$$

and hence the circle of the given radius AD has the same area as the rectangle of sides $2AE$ and AD.

4.7 Apollonius of Perga and the Conic Sections

Apollonius of Perga is the last of the great Greek geometers. He was born in Perga about 262 and died in Alexandria about 190 B.C. (Fig. 4.33)

Apollonius' most important work is *Conica*, on Conic Sections. This work is in eight books, of which seven are preserved either in the Greek original or in Arabic translation. As a young man he came to Alexandria, where he studied under Euclid's successors. He also stayed at *Pergamon*. Pergamon and Alexandria were the most important scientific and cultural centers in the Hellenistic world.

Before Apollonius Euclid had written a book on conics, which is lost. Also Archimedes and others had studied them, in particular in the work which took place on the *classical problems.*

The reader who is not familiar with the conic sections, that is to say curves in the plane of degree 2, may want to move to Sect. 13.5 now for a review from a modern point of view.

But the full fledged algebraic description found there came much later. The Greek geometers understood them as the curves of intersection between a *cone* and a *plane*. Prior to Apollonius conic sections were understood as the intersections between *three different kinds of cones* and a plane at right angle to one of the *generators* of the cone (see Fig. 4.34). The cones were right circular cones, of the three kinds *right-angled, obtuse-angled* or *acute-angled*. But both Euclid and Archimedes were well aware that conic sections could be produced in other ways.

Fig. 4.33 Apollonius of Perga. Drawing by the author

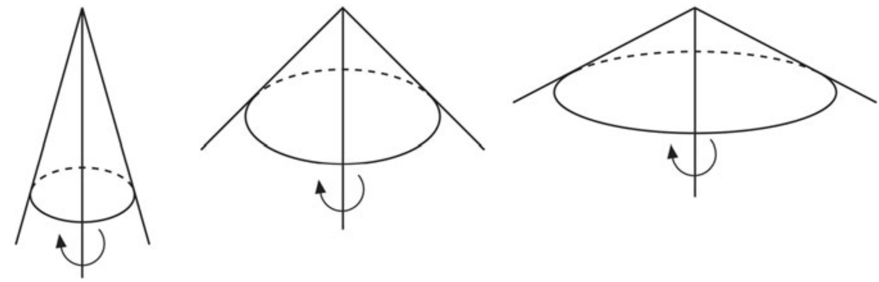


Fig. 4.34 Three different kinds of cones. To the *left* the top angle is less than 90°, in the *middle* it is equal to 90° and to the *right* it is greater than 90°

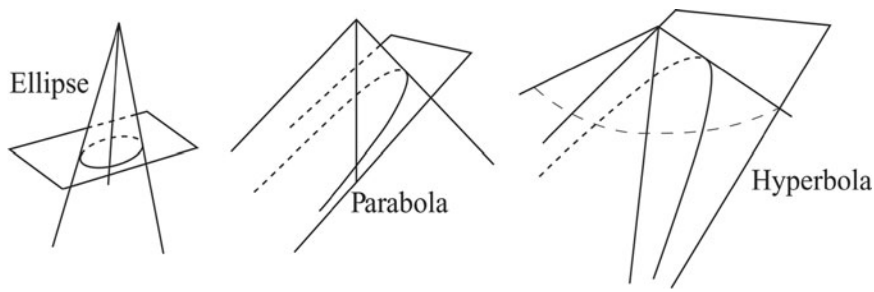


Fig. 4.35 Conic sections prior to Apollonius, cut out in a fixed plane by a varying cone

Thus an *ellipse* would be produced by an acute-angled cone, with an angle *falling short* of a right angle, as the word *ellipse* reflects. A parabola corresponds to a cone with the top angle *equal to* a right angle, while a *hyperbola* is produced by a cone with top angle *greater* than a right angle. See Fig. 4.35. There is a remark to be made at this point, however. Namely, we regard a circle as a special ellipse. Thus it would be produced by an acute angled cone as described above. But the angle would have to be equal to zero, in other words the cone would be a cylinder. And the Greek geometers did indeed consider this case as well.

Apollonius defines a cone as we do today, by rotating a line. In this way we get the *double cone*, and the hyperbola acquires its two branches. See Fig. 4.36.

Also, it was Apollonius who introduced the names ellipse, parabola and hyperbola. These words had been in use in a slightly different way with the Pythagoreans, but the point was the same: The word *ellipse* indicates that *something is left out, it is too little*. The word *parabola* indicates to *liken*, set side by side. The plane cutting the cone runs parallel to the other branch of the cone. And finally, the word *hyperbola* stands for *too much, exaggeration.*

Apollonius' theory of conic sections is a high point of Greek geometry. This theory was essential when *Isaac Newton* very much later deduced the laws of gravity.

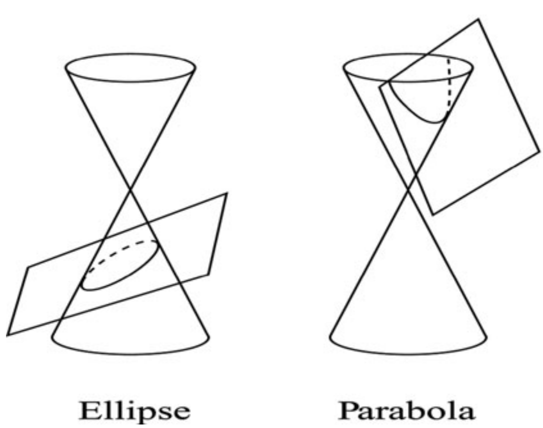


Fig. 4.36 Conic sections with a varying plane and a fixed cone

Apollonius proves a result known as *The Circle of Apollonius*. In modern notation it may be formulated as follows:

The Circle of Apollonius. Let A and B be fixed points and let k be a constant. Then the set of all points P such that $AP : BP = k$ is either a circle, when $k \neq 1$, or a straight line, when $k = 1$.

This is easy to show with modern algebraic techniques. See Fig. 4.37. There we have chosen a coordinate system such that $A = (-1, 0)$ and $B = (1, 0)$. We then have that $AP = kBP$, and by the Pythagorean Theorem we find that

$$AP^2 = (1 + x)^2 + y^2$$

and

$$BP^2 = (1 - x)^2 + y^2.$$

Since now $AP^2 = k^2BP^2$, we get

$$(1 + x)^2 + y^2 = k^2((1 - x)^2 + y^2)$$

The remaining part of the proof is left to the reader.

Another result ascribed to Apollonius is this:

Generalized Pythagorean Theorem. Let ABCD be a parallelogram with sides AB = CD = a, BC = DA = b, and with the diagonals AC = m, BC = n. Then

$$a^2 + b^2 = \frac{1}{2}(m^2 + n^2)$$

We give a modern proof: In Fig. 4.38 we denote AE by x. Let DE $= y$. Then BF $=$ x and FC $= y$, since $\triangle AED$ is congruent with $\triangle BFC$. By the Pythagorean Theorem we have

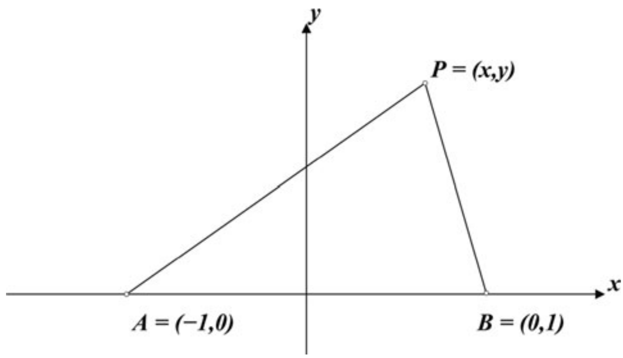


Fig. 4.37 Proof that we get a circle

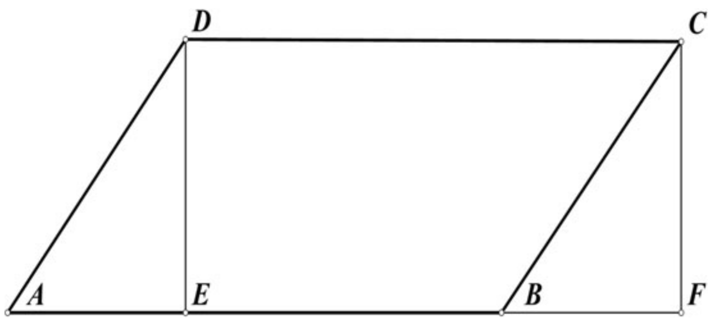


Fig. 4.38 Generalized Pythagorean theorem

$$b^2 = x^2 + y^2$$
$$m^2 = (a + x)^2 + y^2 = a^2 + 2ax + x^2 + y^2$$
$$n^2 = (a - x)^2 + y^2 = a^2 - 2ax + x^2 + y^2$$

Adding the two last equalities and substituting $b^2 = x^2 + y^2$, we get the claim.

Finally Apollonius posed a problem, which in its most general form is very much on the agenda of modern algebraic geometry. Loosely and generally formulated it runs like this:

> *Apollonius' Problem on Tangency.* Given a set of geometric objects, like points, lines or conic sections. Find the objects tangent to the given ones. (Tangency for a point is understood to mean that the curve passes through the point.)

In Apollonius' case there were *three* such fixed geometric objects. For details on the Problem of Apollonius we refer to [26], volume II, page 181, or to [25] page 346.

Apollonius developed geometric algebra a long way towards the complete algebraization of geometry, which came in full much later through the work of Descartes and Fermat.

Fig. 4.39 *René Descartes*, 1596–1650. Drawing by the author

Apollonius worked with an algebraic notation which at least makes him a forerunner for *René Descartes*, an important French philosopher and mathematician who is given credit for having initiated the introduction of algebra into geometry (Fig. 4.39). The name *Cartesian* coordinate system is after him.

4.8 The End of the Republic in Rome

After the defeat of the Germanic intruders, Marius was again elected consul in 100 B.C. This was his sixth consulship.[10]

Marius was now in his fifth consulship, and he campaigned for the sixth very vigorously. Plutarchus writes that not only did he deviate from the state and dignity of his office, but he also gave people a false idea of his own character, by attempting to seem popular and obliging, which in reality was quite contrary to his nature.

Marius was needed in time of war, and this obtained power and dignity for him. But when he sought the first place in civil affairs, then he had to seek the favor of the people, *"never caring to be a good man, so that he were but a great one,"* according to Plutarchus.

Optimates or *The Best of Men, The Good Men,* formed the traditionalist party in Rome. They would limit the power of the popular assembly and extend the power to the senate. Cicero and Sulla, as well as Cato the Younger and Pompey were or became important members of the Optimate party, it later formed the core of the

[10] We now tell this story following Plutarchus [46].

resistance against Julius Caesar's assumption of power. Perhaps what Plutarchus says is that Marcus did not care to be an *optimate*, but merely a "great man."

But now came a serious development which could affect the balance of power between the *assembly* and the *senate*. The peoples tribune *Saturninus* proposed a law for the distribution of land, with a clause that the senate should publicly swear to confirm whatever the assembly should vote. In this power struggle between the assembly and the senate, Marius "cunningly feigned," according to Plutarchus, to be against this provision. The idea was to trick his rival Metellus into refusing to take the oath. But as soon as Metellus had made his declaration, Marius declared that he would abide by the requirement. As a result of this Metellus was banished to Rhodes, where he spent a year "in philosophy," after which he was called back to Rome, to Marius' annoyance.

In the meantime the following happened: Marius, in return for this piece of service, was forced to go along with Saturninus' misdeeds, which now proceeded to the very height of insolence and violence. Without knowing it, Marius became an instrument of mischief beyond endurance, writes Plutarchus. The only course of this was through outrages and massacres to tyranny and the subversion of the government. Marius feared and respected the nobility, but at the same time was eager to court the masses. This led him into a most mean and dishonest action: When some noblemen came to him at night to stir him up against Saturninus, then he secretly let Saturninus in through another door. Then he pretended that he had a health-problem, and ran back and forth between Saturninus and his opponents. Thus he stirred them up against each other.

But soon after this Marius became worried, as he felt that this went too far and Saturninus and his followers now were too powerful. Saturninus was elected tribune for the third time for the year beginning December 10, 100 B.C. Mark Antony was elected without opposition, but another candidate from the optimates, who seemed to have the better chance of success, was beaten to death by the hired agents of Saturninus and his associates while the voting was actually going on.

This outrage infuriated the senate. It met the following day, declared Saturninus and his associates public enemies, and called upon Marius to defend the State. Marius obeyed, Saturninus and his associates were defeated in a battle at the Roman Forum and took refuge on the Capitol. The water supply was then cut off, and they were forced to surrender. Marius had assured them that their lives would be spared. He moved the prisoners to a building which was presumed safe, in order to proceed against them according to the law. But activists supporting the optimate party climbed onto the roof, stripped off the tiles, and stoned Saturninus and many others to death.

When Metellus was recalled from banishment, Marius left for Cappadocia and Galatia. According to Plutarchus the real reason for this journey was to prepare for the final showdown with Mithridates, who Marius thought was preparing for war. And he himself made his best to entice him into challenging the Romans, expecting to be given the Roman command and reap more victory and glory.

However, in 95 B.C. Rome passed a decree that all residents who were not Roman citizens should be expelled from the capital. Then in 91 B.C. Marcus Livius Drusus was elected tribune and proposed a greater division of state lands, the enlargement of the senate, and Roman citizenship for all free men in Italy.

But Drusus was assassinated, and the Italian states then revolted against Rome in the Social War of 91–88 B.C. Marius took command and fought against the rebelling cities with *Lucius Cornelius Sulla Felix*.

Plutarchus relates in [46] that Sulla belonged to a patrician family, and that one of his ancestors, Rufinus, was said to have been consul. In his youth Sulla lived in lodgings, at a low price. This was afterwards impolitely pointed out to him, when people thought him unduly prosperous. For instance, it was told that when he was putting on boastful airs after his campaign in Libya, a certain nobleman said to him: "How can you be an honest man, when your father left you nothing, and yet you are now so rich?"

His personal appearance is indicted in Fig. 4.40, but in addition Plutarchus writes that the gleam of his gray eyes was terribly sharp and powerful, and was rendered even more fearful by the complexion of his face. This was covered with coarse blotches of red, interspersed with white. They say that his surname was given him because of his complexion, and it was in allusion to this that one of the scurrilous jesters at Athens made the verse: "Sulla is a mulberry sprinkled over with coarse flour." In fact, during the later siege of Athens the inhabitants infuriated the Romans in general and Sulla in particular in very insulting ways, from the safety at the city walls. Plutarchus continues, however, that Sulla was fond of good-humored teasing, and in his youth he would converse freely with players and professional jesters, and joined them in all their pleasures. And later when he became supreme master of all, he often got together the most imprudent players and stage-followers of the town, drank and exchanged jests with them without regard to his age and the dignity of his position, and even to the point of neglecting important affairs which required his attention. When he was once at table, it was not in Sulla's nature to admit of anything that was serious, and whereas at other times he was a man of business and wore an austere look, he underwent a complete change as soon as he enjoyed good-fellowship and wine. Then he was gentle and tractable with common singers and dancers, and ready to oblige any one that spoke to him. It seems, writes Plutarchus, to have been a sort of diseased result of this laxity, that he was so prone to amorous pleasures, and yielded without resistance to any temptation of voluptuousness, from which even in his old age he could not refrain. He had a long attachment to Metrobius, a player. He also began a relationship to a common but wealthy woman named Nicopolis. In the end she loved him so much that he was left her heir when she died. He also inherited the property of his step-mother, who loved him as her own son. Thus he became moderately well off.

Sulla was appointed quaestor to Marius in his first consulship, and sailed with him to Libya, in a campaign against Jugurtha. Sulla was sent to neighboring Mauritania in order to eliminate their support for Jugurtha. With the help of

Fig. 4.40 Lucius Cornelius Sulla Felix. Drawing by the author

Bocchus I of Mauretania, Sulla was able to capture Jugurtha and bring the war to a conclusive end. Jugurtha was brought to Rome in chains. Jugurtha was executed by the Romans in 104 B.C., after being paraded through the streets in Marius' Triumph. But some attributed the glory of the success to Sulla, and this annoyed Marius.

Sulla now felt that his reputation justified political activities, and offered himself as a candidate for the city praetorship, but was defeated. However, in the following year he obtained the praetorship, partly because he used money to win support. During his praetorship, when he angrily told *Caesar*, an older relative of the great Caesar, that he would use his own authority against him, Caesar laughed and said: "You are right to consider the authority your own, for you bought it."

After his praetorship, he was sent out to Cappadocia, ostensibly to reinstate Ariobarzanes, but really to check the restless activities of Mithridates. When he returned to Rome, he was appointed consul with Quintus Pompeius.

After the conclusion of the Social War, Mithridates of Pontus began to conquer Rome's eastern provinces and invaded Greece. In 88 B.C., Sulla was elected consul. The senate put Sulla in command of an army which should go against Mithridates. But as we have seen this crossed the plans of Marius, and a short time later *he* won appointment to the command by *the assembly*. Thus the assembly had one general, the senate another.

Now Sulla left Rome and travelled to the army the senate had asked him to lead against Mithridates. Sulla urged his legions to defy the assembly's orders and accept him as their rightful leader. Sulla was successful, the legions stoned the representatives from the assembly. Sulla then commanded six legions to march with him to Rome. This was unforseen by Marius, as no Roman army had ever marched against Rome.

Marius attempted to organize a defence, but his improvised forces were no match for Sulla. Marius was defeated and fled Rome. Sulla and his supporters in the senate passed a death sentence on Marius and some of his allies. A small number of men were executed but not Marius, he narrowly escaped capture and death on several occasions and eventually found safety in Africa.

But many Romans disapproved Sulla's actions. Some who opposed Sulla were actually elected to office in 87 B.C. Gnaeus Octavius was a supporter of Sulla, and Lucius Cornelius Cinna was a supporter of Marius, they were both elected consuls. Sulla was confirmed again as the commander of the campaign against Mithridates, and he took his legions out of Rome and marched east to the war.

While Sulla was on campaign in Greece, fighting broke out between the conservative supporters of Sulla, lead by Octavius, and the supporters of Cinna. Marius and his son now returned from Africa with an army and joined Cinna to oust Octavius. Now Marius entered Rome. Based on the orders of Marius, some of his soldiers went through Rome killing the leading supporters of Sulla, including Octavius. After 5 days Cinna ordered his troops to kill Marius's rampaging soldiers. About 100 prominent Romans had then been killed. The senate passed a law exiling Sulla, and Marius was appointed the new commander in the eastern war. Cinna was chosen for his second Consulship and Marius to his seventh Consulship. But soon afterwards Marius died suddenly at the age of 71.

During the campaign in Greece, Athens was conquered and treated very harshly with great destruction. Some say that this was partly caused by Sulla's anger at the insulting songs from the city walls during the siege. During this time Sulla also seized the library of Apellicon from Teos, who had become an Athenian citizen, and was a famous book collector. In his library were most of the books by Aristotle and Theophrastus. It is said that after the library was carried to Rome, Tyrannio the grammarian arranged most of the works in it, and that Andronicus of Rhodes was furnished by him with copies of them, and published them, and drew up lists. According to Plutarchus, the old Peripatetics did not seem to have had a large or an exact knowledge of the writings of Aristotle and Theophrastus, because the estate of Neleus of Scepsis, to whom Theophrastus bequeathed his books, came into the hands of careless and illiterate people, again according to Plutarchus.

Cinna was elected to two more Consulships afterwards and then died during a mutiny when trying to lead his forces into Greece. The forces of Sulla returned to Italy at Brundisium in 83 B.C., and the sons of Marius died defending Praeneste, a city east of Rome. When he returned to Rome, Sulla started a new reign of terror worse than what had been seen before. Senators and others who had supported Marius were outlawed or executed.

Among them was a young man named *Julius Caesar*, 100–44 B.C. He was a nephew of the wife of the older Marius, and he was married to a daughter of Cinna. He was outlawed, but was later pardoned.

Marius certainly had been a successful Roman general and reformer. His improvements to the structure and organization of the army were effective. But he had broken Roman constitutional tradition, and the days of terror upon his return to Rome is also his responsibility. The controversy with Sulla also contributed to the weakening and ultimate destruction of the Roman political institution.

We have already encountered *Cicero*, in Sect. 4.4. Cicero represents the finest of the Roman intellectual tradition (Fig. 4.41).

Cicero was educated in Rome and Athens, he became recognized as the greatest orator of his times. As a politician his words went unheeded, but in his writings his ideas have been preserved up to the twenty-first century. Cicero is credited with having contributed substantially to the continuation of Greek thinking and Greek science through Rome to our days.

Fig. 4.41 Cicero, 106–43 B.C. Drawing by the author

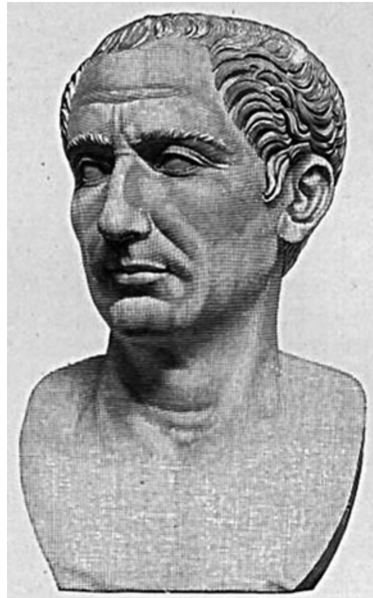


Fig. 4.42 Gaius Julius Caesar, 100–44 B.C.

Cicero had been skeptical to the young and ambitious Julius Caesar, who had started out as a minor political figure in Rome.

Caesar had been made quaestor to the province of Iberia (roughly speaking Spain) in 68 B.C., serving under the praetor *Vestus* who he always honored later. Indeed, when he himself was promoted to praetor of Iberia, he appointed Vestus' son to his old position as quaestor. Returning to Rome, his ambitions and political maneuvering led to the famous comment by *Lutatius Catulus*: *"Caesar no longer attempts to subvert the constitution, he now is starting a frontal attack on it."*

Caesar rose in importance when he was elected to the office of High Priest, *Pontifex Maximus*. This required large sums of money, and put him in debt. His personal finances were brought in order, however, thanks to a term as *propraetor* in Iberia, after which he returned to Rome again (Fig. 4.42).

He now formed an informal alliance known as the *First Triumvirate*, with the successful general *Pompey*, and *Crassus*, called *Dives* (The Wealthy), who had been consul with Pompey in 70 B.C. Crassus had become very rich from buying up land owned by condemned men, and in 71 B.C. he had together with Pompey suppressed the slave revolt led by Spartacus. Caesar's alliance, later fortified by Pompey's marriage to his daughter *Julia*, got him the office of consul. When Caesar together with his powerful allies pushed through controversial and far reaching land reforms, the second consul could not prevent it. To save his life while not sharing responsibility for his partner's actions, he finally locked himself up in his house, where he remained for the rest of his term in office.

After his 1 year term as consul, Caesar received the province of *Gaul*, Gallia, essentially consisting of all of western and northern Europe including Iberia and

Britain, as proconsul for 5 years. This was pushed through by the forceful efforts of his new son in law Pompey, who had filled the Forum with soldiers. Guardians of the constitution like Cicero and *Cato the Younger*, later called *Uticenis* (who died in Utica), protested but were powerless.

Cato was led away to prison on Caesar's orders, but when Caesar saw the negative reactions to this outrage both from the nobility and the people, he had him secretly released. Cato took up the resistance against Caesar, and when alliances later shifted, he found himself allied with Pompey (Pompey's wife had then died in childbirth).

Caesar had his term prolonged by another 5 years, and spent almost 10 years in Gaul, 58–49 B.C. This is where he developed into an extremely successful general, paying his soldiers well. He subdued the barbarians who had grown increasingly active against the Romans.

Among his legates was the younger brother of Cicero, *Quintus Tullius Cicero*. At one point this legate had been surrounded by the enemy, and Caesar sent him a coded message. This is an early example of use of cipher which is documented in the literature. The code consisted in replacing each letter by performing a shift a certain number of places in the alphabet. So for instance, A could be replaced by E, B by F, C by G, and so on. Once you suspect that such a shift-code has been used, it is very easy to break the code by trying all 25 possible shifts.

The rivalry between Caesar and Pompey had been kept in check by their mutual fear of the third man, Crassus. But Crassus was killed in Parthia (Persia).

He had become consul again with Pompey in 55 B.C., after which he got the province of Syria as proconsul, or governor, as the custom was. He was accompanied by his son Publius, who commanded part of the army. When Crassus arrived in Syria he marched directly to the Parthian mainland. The Parthian king Orodes II sent an army with cavalry units against Crassus. The two forces met near the town of Carrhae. The heavily armed and armored horsemen of the Parthians together with a large number of horse archers defeated the Romans, who as always relied on infantry in campaigns on land.

So the campaign did not go well. At one point Publius went out in pursuit of the enemy, and Crassus worried about what had happened. Then some Parthian warriors brought him Publius' head on as spear, scoffingly inquiring where were his parents, and what family he was of, for it was impossible that so brave and valiant a warrior could be the son of so pitiful a coward as Crassus. This gruesome event contributed to breaking down the morals of the Roman soldiers even further. Crassus is said to have made an attempt of restoring their spirit with a forceful speech: "This, O my countrymen, is my own peculiar loss, but the fortune and the glory of Rome is safe and untainted as long as you are safe!"

But the situation was hopeless, and during the retreat the Romans were misled by treacherous guides, and Crassus himself was finally lured into an ambush and killed while trying to negotiate a safe retreat.

Gaius Cassius Longinus was a legate with Crassus, he succeeded in leading about 10,000 survivors to Syria, where he governed as a proquaestor for 2 years. He succeeded in defeating the Parthians. Later still, Cassius participated in the conspiracy to assassinate Caesar in 44 B.C.

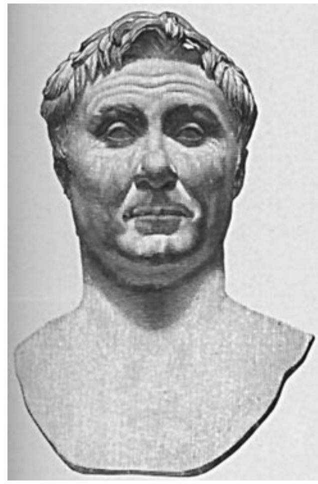


Fig. 4.43 Pompeius Magnus (106–48 B.C.)

But back in Rome anarchy and confusion grew worse. Now Pompey was viewed as the only man who could save the republic. The senate and influential citizen appealed to him not to leave Italy for his provinces which he had been assigned as proconsul, but to stay in Rome and preserve the order. Pompey was now elected sole consul by the senate in 52 B.C., and some say that he took *Scipio* as his co-consul (Fig. 4.43). *Quintus Caecilius Metellus Pius Scipio* was a Scipio by birth, but was adopted into the family Metellus. Thus he belonged to, and came from, distinguished families of Roman generals, and as we have seen, some of them had played a key role in the destruction of Carthage. Scipio was a staunch opponent of Caesar. In 52 B.C. Pompey married Scipio's daughter, Cornelia, the young widow of Crassus' son Publius.

Plutarchus writes the following: "The young lady had other attractions besides those of youth and beauty, for she was highly educated, played well upon the lute, and understood geometry, and had been accustomed to listen with profit to lectures on philosophy. All this too, without in any degree becoming unamiable or pretentious, as some times young women do when they pursue such studies."

By airing his prejudices, not uncommon even today more than 2,000 years later, Plutarchus gives us, almost inadvertently, some very interesting information on the situation concerning women and mathematics in ancient Rome: ... *as some times young women do when they pursue such studies.* Women studying mathematics was not uncommon, it would seem. And the arrogance this knowledge of mathematics ostensibly led to, seems to have annoyed Plutarchus and his peers at the time. This is almost 500 years before Hypatia of Alexandria, who we tell about in Sect. 4.20.

Pompey had his provincial command extended another 5 years. Now even Cicero strongly argued for making Pompey the strong man, the savior of the Republic. But Pompey thus became a close ally of the senate, and there an active group of senators were urging him to head a final showdown with Caesar. Caesar's command would expire in March 49 B.C., but he would not be replaced until 48 B.C. He wanted to run for a second term as consul without having to come in person to Rome. But his opponents were determined to bring his command to an end, and force him to disband his troops and run for consul as a private citizen without an army to support him. Their apprehension would seem quite justified.

After some haggling with the senate however, Caesar made up his mind. He uttered his proverbial *"The dice is cast,"* crossed the Rubicon and marched towards Rome. Pompey panicked, and together with the senate and a large number of nobility he fled into Greece. This was the beginning of a full fledged civil war, which lasted off and on until 31 B.C.

The final battle with Pompey was fought at Pharsalia in Thessaly, in 48 B.C. Here Pompey was completely defeated, and fled to Egypt. Caesar followed him, but arrived too late. In Egypt the political situation was murky. When King Ptolemy Auletes died, he had left his kingdom to his daughter *Cleopatra*, only 17 years old, who was to have ruled jointly with her even younger brother also named Ptolemy, whose wife she was to become according to the customs. A few years later their guardians drove Cleopatra away, and ruled on behalf of the young Ptolemy XIII. The ruling clique consisted of *Pothinius the Eunuch,* who was the head of the royal guard, and *Theodotus of Chios,* a rhetoric master who was the young kings tutor, and finally *Achillas the Egyptian,* who was the commander in chief of the Egyptian army, and technically the regent until the king would come of age. This was the situation as Pompey arrived, defenseless on a single ship, a Seleucian ship not even his own, with his young wife, Scipio's daughter, and their infant son to seek refuge from Caesar. The ruling junta deliberated on how to treat the fallen Roman consul: Granting him asylum would certainly infuriate Caesar, while driving him away would create a potentially dangerous enemy, should the fortunes of war change. It was Theodotus of Chios who gave the deciding argument.

Rhetorically he declared: *"Receive Caesar's enemy Pompey, but at once do away with him! For in doing so we ingratiate ourselves with the former, and have no reason to fear the latter. A dead man cannot bite!"*

When Caesar arrived in Alexandria after the murder of Pompey, Theodotus wanted to ingratiate himself by presenting Caesar with Pompey's severed head. The effort backfired. Caesar turned away in horror, as from a murderer.

It is said that on taking Pompey's signet ring, he wept bitterly. Pompey had been too proud to accept Caesar's mercy. Caesar now made an effort to find those of Pompey's men who were in hiding, wandered about or languished in the dungeons of Egypt.

He later said that the greatest pleasure his victory had given him, had been to save and offer his friendship to those fellow citizens who had stood against him.

In Alexandria Caesar and his men received a somewhat disappointing reception. The indefatigable Pothinius the Eunuch was secretly scheming to destroy Caesar, the Roman soldiers were mistreated with inadequate and low quality food, and Pothinius repeatedly offered his unsolicited advice that Caesar should now leave Egypt to attend to his other important matters, like dealing with Cato and Scipio who were still at large on the African coast with a sizable army. But Caesar stayed on, declaring that he did not need Egyptians as his councillors, and secretly sent for Cleopatra.

Caesar supported Cleopatra in assuming power in Egypt. Then he went after Cato and Scipio. In February 46 B.C., he defeated Scipio's army at *Thapsus*, now Ras Dimas in Tunisia. Uncharacteristically Caesar had Scipio and all his troops killed after they had surrender, but it has been speculated that this happened because Caesar was incapacitated by a fit of epileptics at the time when they surrendered.

Cato then took refuge at Utica a little north of Carthage. When all hope was gone, he fell upon his sword and killed himself, since *"he would only live in a free state, not under the rule of one man"*. Caesar, who arrived too late, is said to have bitterly uttered the words, *"Cato, I must grudge you your death, as you grudged me the honor of saving your life"*.

The picture shows a statue from the Louvre Museum, when he is about to kill himself while reading the Phaedo, a dialogue by Plato with details on the death of Socrates.

Cleopatra must have been a remarkable person. Moralistic historians have judged her harshly, as did many Romans of her own time. But she used all means at her disposal to preserve a measure of independence for her country, in the face of Roman expansion. Caesar almost immediately fell in love with her, and became a loyal ally as well as her lover. Cleopatra named her son with Caesar *Caesarion*, he was born soon after Caesar had left Egypt.

But before that he had helped her defeat her brother, Pothinius having been killed when he was caught plotting to assassinate Caesar. Achillas, who was Pothinius' co-conspirator had escaped to the army, which he commanded against Caesar and Cleopatra.

During this war in Egypt, a *"minor incident"* happened which the soldiers and the generals paid little attention to. Caesar at one point was in the danger of having his communication at sea cut off by the enemy. He averted the danger by setting fire to his ships, the fire spread in the city and the great library was destroyed.

Cato and Scipio had formed an alliance with the king of Numidia, Juba I. But when Caesar attacked, they were defeated at Thapsus in North Africa in 46 B.C.

A large number of Caesar's enemies, including Scipio, wanted to surrender to Caesar. But in an act said to be unusual for Caesar they were all killed.

Upon his return to Rome Caesar led a triumph for the victory over king Juba, whose infant son was carried in the procession. The little African boy who was brought to Rome in this manner, later became king Juba II. He also became one of the most learned historians of his time.

Caesar was made consul a forth time, and went into Spain where he defeated Pompey's sons. Even if they were quite young, they commanded a large army and fought well. After the battle Caesar remarked that he had fought for victory many times before, but this was the first time that he had fought for his life. The triumph Caesar celebrated offended the Romans immensely. For he had not defeated a foreign enemy, he had instead destroyed the children and family of one of the greatest men of Rome.

Nevertheless the Romans elected him *dictator for life*. Although he had made an effort to attain reconciliation with his opponents from the civil war, and spared many lives, the bitterness must have been deeply rooted against him. When he was murdered on March 15, 44 B.C., the plot was led by one of those whose lives he had spared, *Marcus Brutus*.

Brutus and his co-leader Gaius Cassius Longinus, the former legate of Crassus, had thought that the republic would be restored, but the struggle which followed turned out to be about who should succeed Caesar as ruler, as the *new Caesar*.

In spite of the threatening attitudes of Caesar's former associate *Anthony*, the elderly Cicero worked to enlist the senate, the people as well as the provincial governors in an effort to restore the republic.

Cicero wanted to ally himself with a young man who seemed to be a rising star at the political scene: A young relative and heir of Caesar named *Gaius Julius Caesar Octavian*, called *the young Caesar*. Plutarchus tells the story as follows.

It was principally Cicero's hatred of Antony, and as well as his ambitions, which made him ally himself with Octavian. He wanted the support of Caesar's young relative for his own public designs. At first this seemed to work well, for the young Octavian went so far as to call him Father. According to Plutarchus Brutus was highly displeased, and he said that it was manifest that Cicero did not intend liberty for his country, but only a master favorable to himself. Nevertheless Brutus took Cicero's son, who was then studying philosophy at Athens, and gave him a military command and employed him in various ways, with a good result. Cicero's own power at this time was at its greatest height, and he did what he pleased. He drove out Anthony, and sent the two consuls with an army to defeat him, and he persuaded the senate to allow Octavian the ensigns of a praetor.

Antony was defeated, but the consuls were slain. After this the two opposing armies united and came under Octavian's command. The senate now became apprehensive since the young man had become so powerful, and tried to lure his soldiers away, claiming that now as Antony was defeated, the danger was over. *Octavian* became worried, and secretly closed a deal with Cicero that Cicero should help him in a campaign for the consulship, and then Octavian should appoint Cicero to be his partner as the second consul. This was a mistake on the part of Cicero: When Octavian had been elected consul, he dumped Cicero and allied himself with *Marcus Aemilius Lepidus* and Cicero' bitter enemy Mark Antony. Together they formed the *Second Triumvirate* in 43 B.C. It was given a legal cover as a *commission for reorganization of the commonwealth.* Appointed for 5 years, reappointed in 37 B.C. for another 5 years, the commission pursued a harsh policy against all opposition. They started out by putting together a *proscription list* of people who should be outlawed. This meant that any one who had been put on that list, was stripped of his citizenship, of his property and had no protection under the law any more. Anyone could kill him.

One of the first victims was indeed Cicero himself. Mark Anthony of course hated Cicero, and Octavian traded him in return for other concessions from his new partners. Cicero had to flee, together with a few of his friends and servants. In the end he was caught at one of his villas, and when the executioners arrived, his slaves said they did not know where he was. But he was betrayed by *one person* who was with him.

The traitor was later captured by Cicero's friends, who killed him under gruesome torture. Cicero's last words are supposed to have been: *"There is nothing proper about what you are doing, soldier, but do try to kill me properly."* He was decapitated, and his head and hands were displayed on Forum Romanum as Marius and Sulla had done with their murdered enemies before. He was the only victim of the Triumvirate's proscriptions to be treated in this way. People looked at the spectacle in horror, but it has been said that what they saw was not so much Cicero's remains as *Anthony's soul*.

Also killed by the Second Triumvirate was Quintus Tullius Cicero, Marcus Tullius Cicero's younger brother.

It is said that when Octavian, later in life, came upon one of his grandsons reading a book by Cicero and saw the boy trying to conceal it, fearing his grandfather's reaction, he instead took the book from him, read a large part of it and then handed the volume back with the words "My child, this was a learned man, and a lover of his country."

As we shall see shortly, the alliance between Antony and Octavian broke down, and the civil war continued. But after the victory over Antony, when Octavian was consul, he made Cicero's son his colleague in office. Under that consulship the senate removed all statues of Antony. All honors he had been given were annulled, and no member of that family should ever be allowed to bear the name *Marcus*.

How different Cicero had been from the other leaders of his times! He tried to persuade with words, rather than to compel by violence. And he had understood how much Rome stood to loose by abandoning democracy. But in his old age he was misguided by his ambition.

Brutus and Cassius were defeated at Philippi the year after Cicero was murdered, and killed themselves. All hope seemed to be out for the Republic. The alliance between Antony and Octavian was reinforced by the marriage of Antony to Octavian's sister *Octavia*, and Anthony took command of the eastern part of the Empire while Octavian established his control in the West. Lepidus was given North Africa and thus sidelined (Fig. 4.44).

At this point we have to introduce *Cleopatra* into the story once more. Caesar had brought her to Rome, the Romans strongly disapproved of that. She had married a younger brother, remained the mistress of Caesar, and according to some, murdered her husband-brother by poison. After Caesar had been killed, she left and went back to Egypt. Antony now took up residence in Alexandria. Cleopatra's personality soon won her a total dominance over Anthony, at least this is what Plutarchus tells us.

Anthony's liaison with Cleopatra eventually resulted in three children. During this time the library of Alexandria was resupplied with books through an expedition to Pergamon, led by Antony, where the library of Pergamon was raided. At least this is what some sources claim. Unfortunately the library was now in part kept at the *Temple of Serapis*. This had dire consequences later, when zealous Christians decided to burn all pagan literature in Alexandria.

Fig. 4.44 Augustus, as a young man when he was known as Octavianus. Drawing by the author

Thus it is fair to say that Caesar's overthrow of the Republic in Rome not only set in motion the events leading to the decline and fall of the Roman Empire, but also started the process which would eventually lead to the end of Alexandria's Great Library and to the end of the classic civilization in Europe, both still half a millennium into the future.

In Alexandria Cleopatra and Antony lived in profuse and wanton luxury, calling themselves *Isis and Osiris*, claiming to be divinities. Octavian did not like Antony's lifestyle in Alexandria, understandably if for no other reason considering the treatment of his own sister, and finally declared war on Cleopatra. In the sea fight at Actium, Cleopatra and Antony were defeated, and in the end they both committed suicide (Fig. 4.45). This happened in the year 31 B.C., and marks the end of the civil war. The Eastern Provinces surrendered in 29 B.C.

This marks the definite end of the Republic, and the beginning of *the Empire*. By general consent Octavian was called upon to be the ruler, and in 27 B.C. he was endowed with the name of honor *Augustus*, meaning *The Just One*. Formally the republic was restored, but it was only in form. The ideas laid out by Cicero in *De Republica* were apparently realized: A constitutional president of a free people. But in reality Octavian now had become the *Emperor Augustus*.

Fig. 4.45 The battle of Actium, by Lorenzo A. Castro, 1672

4.9 The First Emperors

When Caesar had brought down the Republic, no rational and stable procedure existed for the transfer of power from one ruler to his successor.[11]

The succession turned out to be mainly hereditary or by adoption, often also by assassination, frequently supplemented by large sums of money changing hands. The emperor's guard, the Praetorian Guard, eventually played an important role. *Tiberius Caesar Augustus* or *Tiberius*, was born in 42 B.C. and died in 37 A.D. He was the second Roman Emperor, who ruled from Augustus' death in until 37 A.D.

Tiberius is remembered as one of Rome's greatest generals, whose campaigns in Pannonia, Illyricum, Rhaetia and Germania laid the foundations for the northern frontier. But he is also remembered as a reclusive and gloomy leader, who had never really desired to be Emperor.

After the death of his son Drusus in the year 23 his rule degenerated into a reign of terror, and in 26 Tiberius left Rome and handed the administration over to the heads of the Praetorian guard. Caligula, who was Tiberius' adopted grandson, succeeded him when he died.

Gaius Julius Caesar Augustus Germanicus or just Caligula, as he is known, was born in the year 12 A.D and died in 41. He ruled during 37–41 A.D., when he was killed by members of his own guard, Praetorian guard. He is considered to have

[11] The main sources for this section are [7, 12, 13, 18, 61].

been a *monster*, an insane tyrant. It is told that he appointed his favorite horse to a seat in the senate and attempted to appoint it to the position of consul. But this seems to have been a misunderstanding.

Already in 38 Caligula fell ill, apparently with serious consequences for his mental condition. He started to lead a reckless and boozy personal life, ruthlessly doing away with people who opposed him. He wanted to be worshiped as a good, thus emulating oriental practices.

Caligula is alleged to have had incestuous relationships with his sisters, but there is no credible evidence for this either. In short, the surviving sources are filled with anecdotes of Caligula's cruelty and insanity rather than an actual account of his reign, so making any reconstruction of his time as Emperor is next to impossible. The picture of Caligula is that of a depraved and hedonistic ruler. This image has made Caligula one of the most widely recognizable and detested of all the Roman emperors. The name *Caligula* means "Little Boot."

Tiberius Claudius Caesar Augustus Germanicus, or just Claudius, was born in 10 B.C. and died in 54 A.D. He ruled 41–54, when he died. He became consul with his nephew Caligula in 37 A.D., but was afflicted with a disability, which according to some historians protected him from being killed during the purges of Tiberius' and Caligula's reigns. Indeed, after Caligula's assassination he was he was the only adult male of his family alive.

Thus he was an obvious choice for emperor, and was declared emperor by the Praetorian Guard instead of by the senate, which had been the practice previously. Of course, the guard had then first killed his predecessor Caligula. So the process of succession was already becoming quite unlawful and capricious, and it would get even worse. Claudius was the first emperor who was not elected by the senate. But Claudius turned out to be a competent administrator and a great builder of public works. His reign saw an expansion of the empire, including the conquest of Britain. He completed projects like the Claudian aqueduct, built a new harbor at Ostia and improved Roman jurisprudence in several ways.

He took a personal interest in the law and presided at public trials. However, he was seen as vulnerable throughout his rule, particularly by the nobility. Claudius had been viewed as a less than gifted politician, but was a learned scholar. He is said to have been largely led by his wives, especially the third, Messalina. His fourth wife Agrippina had him poisoned to make her son by a previous marriage, Nero, the new emperor.

4.10 Heron of Alexandria

Heron lived and worked in Alexandria (Fig. 4.46). He was born about 10 A.D. and died around year 75. These dates were confirmed by Neugebauer, who found that Heron refers to a recent eclipse in one of his books, which he could identify as having occurred in Alexandria at 11 p.m. on March 13, in the year 62 A.D.

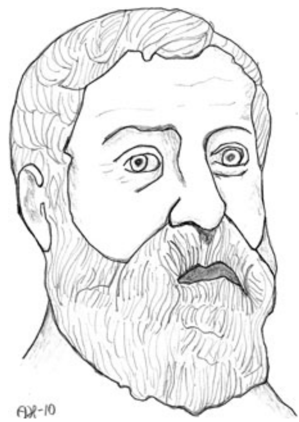


Fig. 4.46 Heron of Alexandria

According to this he lived during a difficult period, at least life was difficult in Rome and probably in the rest of Italy as well. How far the effects of mismanagement under Gaius (Caligula) or Nero extended out into the provinces and affected life in Alexandria, may be another matter. At any rate Heron was born during the last years of the reign of Emperor Augustus, and died a few years after *Vespasian* had restored the order.

But during the reign of Claudius, 41–54 A.D. the Claudian aqueduct, as well as the harbor of Ostia were completed. Great engineering enterprizes, which happened in the middle of Heron's career.

Heron taught at the Alexandrian Academy, and his preserved books are most probably notes for his lectures, either written by himself or by some of his students. Heron must, without question, have been a brilliant teacher as well as an illustrious engineer and applied mathematician. His lecture notes consist to a very large degree of worked examples, such as his explanation in *Dioptra* of how the Samians used geometry to dig their tunnel through the mountain Castro about 600 years before his time, see Sect. 3.2. He also explains the construction of mechanical instruments, machines and gadgets intended as toys or for amusement. Also included are war-machines, a wind-organ as well as a steam powered engine working on the same principle as a *jet engine*! It has occasioned some comments that Heron would include the construction of "mere toys for children" in his lectures. Perhaps this was intended to enliven the exposition of otherwise dull principles of mechanics and physics.

From the point of view of geometry, the following mirror-constructions are of interest. They combine nice geometry with some real fun. We quote from [26]:

– To construct a right-handed mirror, i.e. a mirror which makes the right side right and the left side left instead of the opposite.
– To construct the mirror called the *polytheron*, "with many images."
– To construct a mirror inside the window of a house, so that you can see in it, while you are inside the house, everything that passes in the street.

– To arrange mirrors in a given place so that a person who approaches cannot actually see himself or anyone else but can see any image desired, a *"ghost seer."*

Heron's lecture notes proved so useful, that they were copied and recopied, and used for centuries in Byzantine, Roman and Arab science. As Neugebauer demonstrates definitively in [44], Heron gives expositions of mathematics from the Old Babylonian epoch. He also evidently builds on the mathematics and geometry of Archimedes. Therefore the legacy of Heron is an important segment in the chain tying us, via the Arabs, the Byzantine scholars and the Greek, to the ancient wisdom of the Babylonians and the Sumerians.

The formula for the area of a triangle, known as *Heron's Formula*, very probably was known by Archimedes, and may be even older. It says that if a, b and c are the sides and s is half the circumference, then the area is

$$A = \sqrt{s(s-a)(s-b)(s-c)}$$

A more general formula asserts that if a, b, c and d are the sides of a cyclic quadrilateral, i.e., a quadrilateral which may be circumscribed a circle and s is half the circumference, then the area is given by

$$A = \sqrt{(s-a)(s-b)(s-c)(s-d)}$$

This formula is used by the Indian mathematician *Brahmagupta*, but he does not seem to mention the important assumption that the quadrilateral be cyclic. Thus the formula may well come from another source. There *is* another formula, valid for any quadrilateral, it is given in Exercise 6.3.

In our days, when mathematics is so frequently presented as abstract and heavy with symbols, it is tempting to conclude this section about Heron with his recipe for finding approximate cube roots. No symbolics, just numbers. We quote from [26]. To find the "cube side" of 100:

"*Take the nearest cube numbers to 100 both above and below, these are 125 and 64. Then*

$$125 - 100 = 25,$$

and

$$100 - 64 = 36.$$

Multiply 5 into 36, this gives 180. Add 100, making 280. Divide 180 by 280, this gives $\frac{9}{14}$. *Add this to the side of the smaller cube: this gives* $4\frac{9}{14}$. *This is as nearly as possible the cube root of 100 units."*

Heron probably had other cube-root examples as well, but this is the only one known to us. Several historians of mathematics have tried to reconstruct what Heron's formula for the cube root, and from a careful consideration of what one gets from elementary methods, the following formula emerges: Assume that $a^3 < A < (a+1)^3$, and put $d_1 = A - a^3$, $d_2 = (a+1)^3 - A$. Then

$$\sqrt[3]{A} \approx a + \frac{(a+1)d_1}{(a+1)d_1 + ad_2}$$

But there is a simpler interpretation, which however is less elementary to deduce, namely

$$\sqrt[3]{A} \approx a + \frac{d_1\sqrt{d_2}}{A + d_1\sqrt{d_2}}$$

For $A = 100$ this yields the same number as the first formula, of course, a good approximation to $\sqrt[3]{100}$. But for $A = 90$, the first formula still yields a good approximation, but the second one is quite far off the mark.

Finally a more general formula is conjectured by some investigators, the implication being that Heron, or the person taking notes or later copying them, did not quite understand what was going on. For details we refer to [60]. But in teaching young engineers in Alexandria or elsewhere, Heron's examples are undoubtedly more effective than the formulae of his commentators.

4.11 Nero and the Year of the Four Emperors

The reign of Nero showed how vulnerable the system of government had become, mainly through the capricious system of succession.

Nero has a very bad posthumous reputation. Historians dispute how reliable the horror-stories about him are, but there is no doubt that the sources we have for Nero's life and deeds are quite unfavorably inclined against him.

Nero was a real disaster as emperor, even though historians now judge him less harshly than earlier. The first 5 years were not so bad, mostly thanks to good advisers. But then he had his mother murdered, as well as his wife *Octavia*, the daughter of his predecessor. He also had the son of his predecessor, the rightful heir to the throne so to speak, done away with. The great fire in Rome in 64, is claimed to have been set on his orders.

The fire lasted for 9 days. It was used as a pretext to persecute Christians and Jews in Rome. After his fall in 68 there followed a short turbulent period, but then Vespasian in 69 established a firm rule. This is a brief account of what happened immediately after Nero:

When Nero died in 68, Rome had a succession of short-lived emperors and a year of civil wars, with several "emperors" fighting one another. The first was *Servius Sulpicius Galba,* who was born in the year 3 B.C. and died in 69 A.D. He was emperor from June 8, 68 until he was killed on January 15, 69.

At the death of Nero there followed a period of civil war. In the end *Vespasian* emerged victorious, his full name was Titus Flavius Vespasianus, he lived during 39–81 A.D.

4.12 From Vespasian to Marcus Aurelius

In 66 Vespasian had been put in charge of the war in Judaea. A revolt there had killed the previous governor and routed the governor of Syria when he intervened. Vespasian's elder son Titus went with him. In the end thousands of Jews were killed and many towns destroyed by the Romans.

But Vespasian had acquired a reputation as a competent and forceful commander, and thus he found much support. In 69 he was proclaimed emperor, first by the army in Egypt, and then by his own troops in Judaea.

In the final fighting in Rome the Capitol was destroyed by fire and Vespasian's brother was killed, but Vespasian's forces won.

Vespasian left the war in Judaea to his son Titus and arrived at Rome in 70. He at once devoted his energies to repairing the evils caused by civil war. He restored discipline in the army, which had become utterly demoralized under Vitellius. With the cooperation of the Senate, he put the government and its finances in order.

He renewed old taxes and levied new ones, he increased the tribute of the provinces, and he kept a watchful eye upon the treasury officials. The Latin saying "*Pecunia non olet* or "Money does not smell" is said to have originated when he introduced tax on public toilets. By his own example of simplicity of life he set a good example for the increasingly decadent Roman upper class.

He restored the status and dignity of the senate, but at the same time he made it more dependent on the Emperor. He also reorganized the Praetorian Guard.

In 70, an uprising in Gaul was suppressed by his brother-in-law, and the German frontier thus secured. The Jewish War was won by Rome when Jerusalem was taken. Now the Roman empire had peace for the remaining 9 years of Vespasian's reign, this is called "The peace of Vespasian."

Here we show a Roman coin celebrating Roman victory in Judaea. On the front Vespasian, on the back a Jewish woman in mourning and a bound male Jewish prisoner. To the left behind captured weapons.

In 78, he sent an army to Britain consolidating the Roman presence there. But the following year he died.

Pliny the Elder's *The Natural History* was written during Vespasian's reign, and dedicated to Vespasian's son Titus. Vespasian was noted for mildness and a healthy sense of justice. Much money and effort went into public works and the restoration and beautification of Rome.

A new forum was build, a splendid Temple of Peace, public baths and the vast Colosseum was build in his time.

Vespasian was a plain man, with a strong character, who liked order, prosperity and welfare of his people. He was punctual and regular, carrying out his business early in the morning, and then enjoying a siesta in the afternoon. He liked jokes, even in his last words: *Vae puto, deus fio*, "Alas, I think I'm becoming a god." The Romans had adopted the custom of proclaiming their emperors divine after their death.

Emperor Titus succeeded his father as Emperor in 79. He was an effective and popular emperor. He stopped some ongoing treason trials, and held expensive gladiatorial games.

He completed the Colosseum, where the construction had been started by his father. Titus was emperor during the eruption of Mount Vesuvius in 79 and the consequent destruction of life and property in the cities and resort communities around the Bay of Naples, such as Pompeii and Herculaneum.

In 80, there was a fire in Rome, and Titus spent large amounts of money relieving victims of both the volcano and the fire. He visited Pompeii just after the eruption, and again the following year.

After only 2 years as emperor Titus died of a fever, possibly poisoned by his physician on behalf of his brother Domitian, or due to malaria. He was deified by the Senate and was succeeded by his brother, Domitian. Titus had been a better emperor than his brother, who succeeded him. In fact, he was used as a model by later emperors, especially those known as the Five Good Emperors.

Titus Flavius Domitianus (51–96) became emperor in 81. Here we see a denarius of Domitian. He had received a good education and studied rhetoric and literature, law and administration. He is described as learned and educated.

Domitian proved to be a disastrous administrator, and an inept military commander. The economy deteriorated, so he had to devalue the denarius by mixing an increasing amount of copper into the silver.

The main Roman currency consisted of coins. The aureus (gold), the denarius (silver), the sestertius (bronze), the dupondius (bronze), and the as (copper). These were used from the middle of the second century B.C. until the middle of the third century A.D. Originally the denarius consisted of 4.5 g of silver, but this decreased to a silver contents of 3 g around the late third century. The value originally was 10 asses, the name means "containing ten." In about 141 B.C. it was changed to 16 asses, to reflect the decrease in value of the as. The denarius continued to be the main coin of the empire until it was replaced by the antoninianus in the middle of the third century. The last find of denarii are coins issued by Aurelian between 270–275, in bronze.

To improve the economy, Domitian raised taxes. Discontent soon followed. Domitian invested large sums in the reconstruction and beautification of the city, repairing the damage of the great fire of Rome of 64 and the civil war of 69. Around 50 new buildings were erected and restored, including the Temple of Jupiter in the Capitoline Hill and a palace in the Palatine Hill.

Towards the end of his reign Domitian's regime deteriorated. Jews and Christians were persecuted. The emperor also had senators and other noblemen executed. He disliked aristocrats and stripped the senate of decision making power. Domitian was assassinated in a plot organized by his enemies in the senate.

After Domitian's death in 96, there followed the emperors Nerva, Trajan, Hadrian, Antonius Pius and Marcus Aurelius. These are known as the Five Good Emperors, Nerva was the first of these. He was the last emperor who was Italian both by family and by birth. He had been consul with Vespasian in 71 and with Domitian in 90. When Domitian was assassinated Nerva became emperor.

He released people imprisoned for treason and banned future prosecutions for treason, granted amnesty to many whom Domitian had exiled, restored much confiscated property, and involved the senate in the government again. But he was forced to submit to demands from the Praetorian guard and hand over those responsible for Domitian's death. Nerva adopted Trajan, a commander of the armies on the German frontier, as his son and successor. The Guard Prefect responsible for the mutiny against Nerva was later executed under Trajan. Nerva was struck by a fever and chills and died shortly afterwards, having been emperor only from 96 to 98.

Trajan was born in 53 and died in 117. He was Roman emperor 98–117. Under his rule the Roman empire reached its greatest territorial extent.

Hadrian was born in 76 and died in 138, he was Roman emperor during 117–138. Hadrian's reign saw few major military conflicts. He surrendered Trajan's conquests in Mesopotamia, considering them to be indefensible. There was almost a war with Parthia around 121, but the threat was averted when Hadrian succeeded in negotiating a peace.

However, Hadrian's persecuted the Jewish in Judaea. This led to the massive Jewish uprising 132–135 A.D. led by Bar Kokhba and Rabbi Akiva. Hadrian's army suppressed the revolt and continued the religious persecution of Jews.

Hadrian built fortifications along the empire's borders, such as Hadrian's Wall on Great Britain, and the Danube and Rhine borders were strengthened with a series of mostly wooden fortifications.

Hadrian had a homosexual relation to a Greek youth by the name Antinous. While touring Egypt, Antinous mysteriously drowned in the Nile in 130. Deeply saddened, Hadrian founded the Egyptian city of Antinopolis. Hadrian drew the whole Empire into his mourning, and made Antinous the last new god of antiquity.

Antoninus Pius 86–161 was Roman emperor from 138 to 161. He probably earned the name "Pius" because he compelled the Senate to deify Hadrian.

He became consul in 120, and was next appointed by Hadrian as one of the four proconsuls to administer Italia, then he was proconsul of Asia. Hadrian who adopted him as his son and successor in 138, after the death of his first adopted son Aelius Verus, on the condition that he himself would adopt Marcus Annius Verus, the son of his wife's brother, and Lucius, son of Aelius Verus, who afterwards became emperors Marcus Aurelius and his co-emperor Lucius Aelius Verus.

One of his first acts as emperor was to persuade the senate to grant divine honors to Hadrian, which they had at first refused. He built temples, theaters, and mausoleums, promoted the arts and sciences, and bestowed honors and salaries upon the teachers of rhetoric and philosophy.

His reign was by and large peaceful. Some claim that he had a tendency to sweep problems under the rug, and thus contributed to the pressing troubles that faced not only Marcus Aurelius but also the emperors of the third century.

After the longest reign since Augustus, Antoninus died of fever at Lorium in Etruria, about twelve miles from Rome, on March 7, 161. His body was placed in Hadrian's mausoleum.

Antoninus had married Faustina and they had four children: two sons and two daughters. However, only one daughter was alive at the time of his adoption by

Hadrian. Later, he adopted Marcus Aurelius, who was to be his successor. In 146, Aurelius was recognized as Antoninus's colleague in rule.

Marcus Aurelius was born in 121 and died in 180. He was Roman emperor from 161 to his death. He was the last of the "Five Good Emperors" who governed the Roman Empire from 96 to 180, and is also considered one of the most important stoic philosophers. Marcus Aurelius' work *Meditations*, written on campaign between 170 and 180, is still revered as a literary monument to a government of service and duty and has been praised for its "exquisite accent and its infinite tenderness."

His father's sister was Faustina the Elder, who married the Roman Emperor Antoninus Pius. On the death of Hadrian's first adopted son, Hadrian made it a precondition of making Antoninus his successor that Antoninus would adopt Marcus and Lucius Ceionius Commodus, and arrange for them to be next in the line. The latter was Lucius Aelius' son, 10 years younger than Marcus. He was renamed *Lucius Aurelius Verus*.

In 145, Marcus married Annia Galeria Faustina, who was Antoninus' daughter and his cousin as well. When Antoninus Pius died in 161, Marcus accepted the throne on the condition that he and Verus were made joint emperors. Though formally equal from the constitutional point of view, Verus was subordinate.

The joint succession may have been partly motivated by military experiences, since, during his reign, Marcus Aurelius was almost constantly at war with various peoples outside the empire. A highly authoritative figure was needed to command the troops, yet the emperor himself could not defend both the German and Parthian fronts at the same time. Neither could he simply appoint a general to lead the legions; earlier popular military leaders like Julius Caesar and Vespasian had used the military to overthrow the existing government and install themselves as supreme leaders. Marcus Aurelius solved the problem by sending Verus to command the legions in the east. Verus was authoritative enough to command the full loyalty of the troops, but already powerful enough that he had little incentive to overthrow Marcus. Verus remained loyal until his death on campaign in 169.

This joint emperorship is of a similar nature to the political system of the Roman Republic, with two consuls serving together.

Marcus continued on the path of his predecessors by issuing numerous law reforms, mainly to clear away abuses and anomalies in the civil jurisprudence. In particular, he promoted favorable measures towards categories like slaves, widows and minors; recognition to blood relationships in the field of succession was given. In the criminal law a distinction of class, with different punishments, was made between honestiores and humiliores, The more honorable and The more lowly.

Under Marcus' reign, the status of Christians remained the same since the time of Trajan. They were legally punishable, though in fact rarely persecuted. In 177, a group of Christians was martyrized at Lyon, for example, but the act is mainly attributable to the initiative of the local governor.

In Asia, a revitalized Parthian Empire renewed its assault in 161, defeating two Roman armies and invading Armenia and Syria. Marcus Aurelius sent his joint emperor Verus to command the legions in the east to face this threat. The war ended successfully in 166, although the merit must be mostly ascribed to subordinate generals like Gaius Avidius Cassius. On the return from the campaign, Verus was awarded with a triumph; the parade was unusual because it included the two emperors, their sons and unmarried daughters as a big family celebration. Marcus Aurelius' two infant sons, Commodus and Annius Verus, were elevated to the status of Caesar for the occasion.

Marcus Aurelius was constantly at war, and he died in 180 in the city of Vindobona, modern Vienna. His son and successor Commodus was with him. Marcus Aurelius was immediately deified and his ashes were returned to Rome, and rested in Hadrian's mausoleum until the Visigoth sack of the city in 410. Marcus Aurelius was able to secure the succession for Commodus, whom he had named Caesar in 166 and made co-emperor in 177. This decision put an end to the good series of "adoptive emperors." It was criticized later since Commodus turned out manifestly unfit to be emperor.

4.13 Menelaus of Alexandria

Menelaus of Alexandria was born around 70 and died 130 A.D. Thus his birth coincides with the beginning of the reign of Vespasian, who assumed power in 69 A.D. and opened a good and stable period.

Menelaus spent some time in Rome, where he did astronomical work, and he is mentioned in one of Plutarchus' books. Menelaus wrote extensively, but only one of his works is extant, namely the *Sphaericae*. Here he studies geometry on a sphere, in particular he establishes the properties of spherical triangles in the same way as Euclid treats plane triangles. In this sense one might say that Menelaus' work is a precursor for non-Euclidian geometry.

Menelaus' Theorem says the following, in modern notation (Fig. 4.47):

Theorem 2 (Menelaus). *If a triangle ABC be cut by a line ℓ, which cuts the side (possibly produced) AB in the point D, and similarly BC in the point E and CA in*

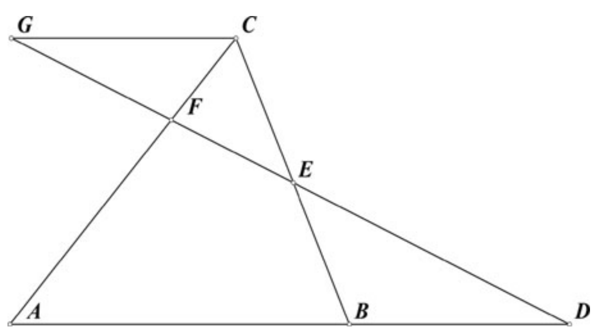


Fig. 4.47 The theorem of Menelaus

F, then the following relation holds

$$\frac{AD}{DB}\frac{BE}{EC}\frac{CF}{FA} = 1$$

where AD, DB, etc., are the lengths of the line segments AD, DB, etc., all considered as positive numbers.

Proof. Let G be the point on ℓ such that $CG//AB$. The $\triangle FCG \sim \triangle FAD$, from which follows

$$\frac{FC}{CG} = \frac{FA}{AD}$$

and $\triangle ECG \sim \triangle EBD$, from which follows

$$\frac{EC}{CG} = \frac{BE}{BD}$$

The former yields

$$\frac{FC}{CG}\frac{AD}{FA} = 1$$

while the latter yields

$$\frac{CG}{EC}\frac{BE}{BD} = 1$$

and when multiplied these relations yield the claim. □

Menelaus also proved statements for *spherical triangles*, analogous to theorems for plane ones, including the theorem given above. For details on Menelaus' theorem for spherical triangles as well as related material, we refer to [26].

4.14 Claudius Ptolemy

Claudius Ptolemy was born about the year 85 A.D., probably in Alexandria but possibly in Hermiou in Upper Egypt. He died around 165, in Alexandria. This would place his death to just before the time when the good period following Nero's death ended in 180 A.D., when the capricious and depraved *Commodus* became emperor.

According to his name he would be of an Egyptian family with Greek background, who had been made Roman citizen. This was a usual practice at this time in the history of the Roman Empire, for provincials who had rendered valuable services. He lived during a stable period of internal tranquility and good government, and his main work appears to have been done during the reign of the Emperor *Titus Antonius Pius*, which lasted from 138 to 161 A.D. Ptolemy's main work represents the definitive state of Greek astronomy. Consisting of 13 books, it bore the title *Mathematical Collection*. Later Pappus wrote an introduction to this work, which came to be called the *Little Astronomy*, Ptolemy's original being referred to as the *Great Collection*. Still later it was translated into the Arabic, the title became something like *"The Greatest,"* or *Al-majisti*.

In turn this ended up as *Almagest*. Thus Ptolemy's Mathematical Collection acquired the name under which it was handed down to posterity.

Together with Euclid's Elements the Almagest of Ptolemy is the scientific text longest in use, up to the Renaissance. The idea of the earth-centered universe on which it is based, made necessary intricate mathematical explanations.

Ptolemy develops extensive trigonometric methods, and in particular introduces the *chord-function*, which is essentially equivalent to our trigonometric functions sin, cos, tan etc. The chord function of the angle v may be defined, anachronistically in modern notation, as $\mathrm{crd}(v) = 120\sin(\frac{v}{2})$.

In Neugebauer's very interesting book [44] the computations of this trigonometric function is described. One such table carries the title *Table of straight lines in the circle*, it is a table of chords. The computations is to the base 60, and uses a symbol for the number *zero*, used in a fully modern way. Neugebauer asserts the following on page 13, which corrects a very common misconception about Greek mathematics:

> *"According to the prevailing doctrine that Greek mathematics is essentially geometry, the historians of mathematics have badly neglected the enormous amount of numerical computations which are readily accessible in works like Ptolemy's "Almagest" or Theon's "Handy Tables." But long before these classics were written, Greek astronomical papyri were covered with computations. While Ptolemy or Theon are today only preserved in Byzantine manuscripts, we do have papyri from the Ptolemaic period* [the last centuries B.C.] *onwards. In these papyri we can find, e.g., the zero sign as it was actually written.*

Ptolemy starts out in Book I, as a preliminary to the Table of Chords, by dividing the circle into 360 equal parts, or *degrees*, and the diameter into 120 equal parts. It then follows that the chord subtending an arc of $v°$ will have length $120\sin(\frac{v}{2})$, and that $\mathrm{crd}(180° - v) = 120\cos(\frac{v}{2})$.

We now follow the explanation provided by Heath in [26].

First, to find the chords subtending arcs of 72 and 36°, i.e., the sides of the regular pentagon and 10-gon (decagon), Propositions 9 and 10 of Book XIII of Euclid's Elements are used. Proposition XIII.9 is equivalent to the formula for s_{10}, the side of a regular decagon inscribed in a circle of radius r,

$$s_{10} = \frac{r}{2}(-1 + \sqrt{5})$$

while Proposition XIII.10 is equivalent to the relation for the side of the inscribed regular pentagon

$$s_5^2 = r^2 + s_{10}^2 \text{ or } s_5 = \frac{r}{2}\sqrt{10 - 2\sqrt{5}}$$

We show these relations in Sect. 17.7. Thus crd(72°) and crd(36°) may be computed, the diameter being 120 one gets

$$\mathrm{crd}(72°) = 30\sqrt{10 - 2\sqrt{5}} \text{ and } \mathrm{crd}(36°) = 30(\sqrt{5} - 1)$$

Ptolemy extracts the square root computing to the base 60, by a method later explained by Theon of Alexandria. The answer is rendered in a mixed notation, with the fractional part to the base 60, as $\text{crd}(72°) = 70^p 32'3''$, which checks with our calculator which gives the answer as ≈ 70.534236 while $70 + \frac{32}{60} + \frac{3}{3{,}600} \approx 70.5341666$.

Ptolemy utilized the immediate observation that the chords subtending v and $180° - v$ form a right triangle with the respective chords containing the right angle and the diameter as the hypothenuse, thus

$$\text{crd}(v)^2 + \text{crd}(180° - v)^2 = 120^2$$

equivalent to the relation $\sin^2 v + \cos^2 v = 1$.

Now $\text{crd}(60°) = 120$, and $\text{crd}(90°) = \sqrt{2 \cdot 60^2} \approx 84^p 51'10''$. To proceed, it is now necessary to have formulas expressing $\text{crd}(\alpha \pm \beta)$ in terms of $\text{crd}(\alpha)$ and $\text{crd}(\beta)$, they are equivalent to the familiar formulas

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$$
$$\sin(\alpha - \beta) = \sin(\alpha)\cos(\beta) - \cos(\alpha)\sin(\beta)$$

Of the two formulas, we shall start by deriving the latter, and this is where the famous *Ptolemy's Theorem* enters the scene. The theorem asserts the following:

Theorem 3 (Ptolemy's Theorem). *In a cyclic quadrilateral the sum of the products of opposite sides is equal to the product of the diagonals.*

Proof. We refer to Fig. 4.48, where the quadrilateral is $ABCD$, and the claim is that

$$AB \cdot DC + BC \cdot DA = AC \cdot BD$$

On AC the point E is marked so that $\angle ABE = \angle DBC$. It then follows that $\triangle EAB \sim \triangle CDB$ and that $\triangle DAB \sim \triangle CEB$. The former relation yields

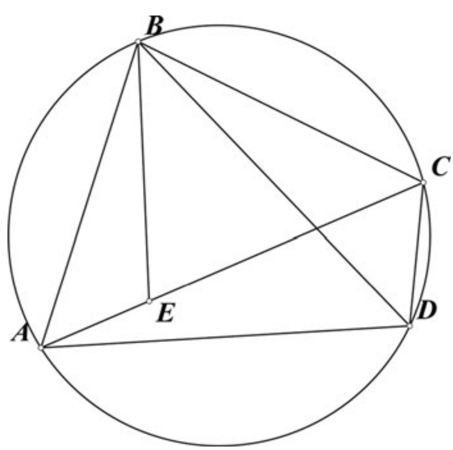


Fig. 4.48 The cyclic quadrilateral is $ABCD$. On AC the point E is marked so that $\angle ABE = \angle DBC$

$$\frac{AB}{AE} = \frac{BD}{DC} \text{ hence } AB \cdot DC = AE \cdot BD$$

while the latter yields

$$\frac{BC}{CE} = \frac{BD}{DA} \text{ hence } BC \cdot DA = CE \cdot BD$$

and adding the two we get

$$AB \cdot DC + BC \cdot DA = AE \cdot BD + CE \cdot BD = AC \cdot BD$$

as claimed. □

We apply the theorem to the cyclic quadrilateral where AD is a *diameter*. Let α be the arc AC, and β be the arc AB. Thus $\text{crd}(\alpha) = AC$, while $\text{crd}(\beta) = AB$. Then the formula of the theorem yields

$$\text{crd}(\alpha - \beta)\text{crd}(180°) + \text{crd}(\beta)\text{crd}(180° - \alpha) = \text{crd}(\alpha)\text{crd}(180° - \beta)$$

or

$$\text{crd}(\alpha - \beta)\text{crd}(180°) = \text{crd}(\alpha)\text{crd}(180° - \beta) - \text{crd}(180° - \alpha)\text{crd}(\beta)$$

To derive a formula for $\text{crd}(\alpha + \beta)$, we use the same cyclic quadrilateral, i.e., we let AD be a diameter. The arc AB is α, but now the arc BC is β. We draw the diameter through B, and get the point E as indicated in Fig. 4.49.

Here $\text{crd}(\alpha + \beta)$ is the length of the chord AC, which is known once we know $CD = \text{crd}(180° - (\alpha + \beta))$. AB as well as BC are known. As $AB = DE$, the latter is known, as is BD, $\triangle ABD$ being right, and $AD = 120$. Thus applying Ptolemy's Theorem to the cyclic quadrilateral $BCDE$ we find the one unknown entity CD, which solves our problem: We get

$$CD \cdot BE + BC \cdot DE = BD \cdot EC$$

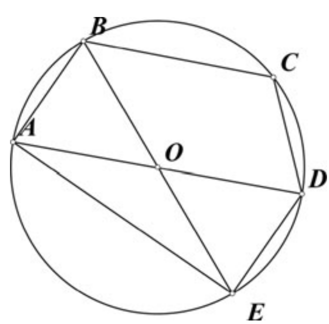


Fig. 4.49 The cyclic quadrilateral *ABCD* where *AD* is a diameter and the diameter through *B* is drawn, yielding the second quadrilateral *BCDE*

thus

$$\begin{aligned}&\mathrm{crd}(180^\circ - (\alpha + \beta))\mathrm{crd}(180^\circ)\\&\quad = \mathrm{crd}(180^\circ - \alpha)\mathrm{crd}(180^\circ - \beta) - \mathrm{crd}(\alpha)\mathrm{crd}(\beta)\end{aligned}$$

Taking $\alpha = \beta$ this last formula yields

$$\mathrm{crd}(\alpha)^2 = \frac{1}{2}\mathrm{crd}(180^\circ)(\mathrm{crd}(180^\circ) - \mathrm{crd}(180^\circ - 2\alpha))$$

equivalent to the familiar $\sin^2 \frac{v}{2} = \frac{1}{2}(1 - \cos v)$.

Thus Ptolemy obtains $\mathrm{crd}(12^\circ) = \mathrm{crd}(72^\circ - 60^\circ)$, and from the last formula, starting with crd(36°) he computes crd(18°) and crd(9°). Then he also captures crd(6°) as well as crd(3°).

To find crd(1°) is more tricky: Ptolemy readily finds $\mathrm{crd}(1\frac{1}{2}^\circ)$ as well as $\mathrm{crd}(\frac{3}{4}^\circ)$. He then determines the value $\mathrm{crd}(1^\circ) \approx 1^p2'15''$ by an ingenious method of interpolation, using the fact which in modern notation says that the function $f(v) = \frac{\sin(v)}{v}$ is monotonously decreasing in the interval $< 0, \frac{\pi}{2} >$.

The method was not new, in fact it is due to an earlier great Greek mathematician, namely *Aristarchus of Samos*, 310–230 B.C. Heath writes about him as follows:

> Historians of mathematics have, as a rule, given too little attention to Aristarchus of Samos. The reason is no doubt that he was an astronomer, and therefore it might be supposed that his work would have no sufficient interest for the mathematician. The Greeks knew better; they called him "Aristarchus the mathematician."

Aristarchus was a precursor for *Copernicus*, in that he was the first to propose a sun-centered universe. He is also remembered for his attempt to determine the sizes and distances of the sun and moon. It is ironical, perhaps, that Ptolemy who so carefully had studied Aristarchus' mathematics, did not know or make use of the work Aristarchus had left behind concerning a sun-centered universe! And that he used Aristarchus' mathematics for the computations when he compiled his tables of chords, partly intended to explain the doctrine of an earth-based universe.

For details we refer to [26].

4.15 The Rule of Sines and the Law of Cosines

As we have seen in the previous section, Ptolemy used equivalent notions to our sine and cosine. In particular he required, and found, procedures equivalent to what we call *the Rule of Sines* and the *Law of Cosines* (Fig. 4.50).[12] Here we present these procedures in the modern form.

$\triangle ABC$ is given. The circumscribed circle has center O and radius $R = OC$. The angles at A, B and C are denoted by the same letters. The $\angle BOE = A$. Thus

[12] A detailed explanation is given in [38].

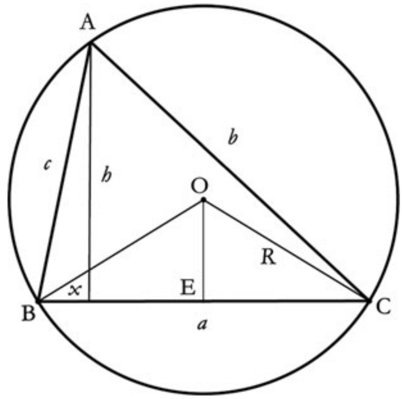


Fig. 4.50 The Rule of Sines and the Law of Cosines

$\frac{a}{2} = R\sin(A)$, so $\frac{a}{\sin(A)} = 2R$, and since the similar relations hold for b and c, the Rule of Sines follows: :

$$\frac{a}{\sin(A)} = \frac{b}{\sin(B)} = \frac{c}{\sin(C)} = 2R$$

To show the Law of Cosines we put $BD = x$. Then we obtain, from the Pythagorean Theorem that

$$x^2 + h^2 = c^2$$

and

$$(a - x)^2 + h^2 = a^2 - 2ax + x^2 + h^2 = b^2$$

Subtracting the former from the latter yields

$$a^2 - 2ax = b^2 - c^2$$

and since $x = c\cos(B)$, this yields

$$a^2 + c^2 - 2ac\cos(B) = b^2$$

which is the Law of Cosines.

4.16 From Commodus to the End of the Crisis of the Third Century

With Commodus the decline of the Roman Empire starts in earnest. Commodus ruled for 30 years, and his entire reign was a nightmare.

He was born in 161, and was killed in 191. His father, emperor Marcus Aurelius, made him Caesar. His two brothers died early, this left Commodus as Marcus Aurelius' sole surviving son.

In 176, Marcus Aurelius granted Commodus the name Augustus, giving him the same status as his own and formally sharing power. He then married before going with his father to the Danube in 178, where Marcus Aurelius died in 180. Thus Commodus became emperor.

Commodus was extremely proud of his physical strength, and ordered many statues to be made showing him dressed as Hercules with a lion's hide and a club. He thought of himself as the reincarnation of Hercules, frequently emulating the legendary hero's feats by appearing in the arena to fight a variety of wild animals. It is said that Commodus was a skilled archer.

He enjoyed gladiatorial combat, and took to the arena himself dressed as a gladiator. Commodus won since his opponents always submitted to the emperor. These public fights would not end in a death, but privately it was his custom to slay his practice opponents.

In 190, part of the city of Rome burned, and Commodus took the opportunity to "re-found" the city of Rome in his own honor, under a new name. The months of the calendar were all renamed in his honor, and the senate was also renamed.

A year later, Commodus was strangled in his bath by a wrestler on orders from Commodus' mistress and cousin Marcia.

When Commodus was assassinated, a successor was proclaimed the next morning. His reign lasted only 86 days, and ended when he was killed by his guard. Then a senator proclaimed himself the new emperor, but he was replaced the same year by *Septimius Severus*. Septimius Severus was emperor 193–211. He was born in present day Libya. His government was a military dictatorship, but he stamped out the moral degeneration and corruption from the reign of Commodus.

When he died in 211, the stability he had provided was gone. He was succeeded by his two sons who were advised by his wife. One of them was killed by his own brother in to 212, the other ruled to 217. Then there was another emperor for 14 months named Macrinus.

Heliogabalus of the Severan dynasty was emperor from 218 to 222 A.D. The view of him is now somewhat controversial, but earlier he has been almost universally viewed as one of the very worst emperors in Roman history. He certainly showed a complete disregard for Roman traditions and sexual taboos, executed people arbitrarily, and so on. He was finally killed by the Praetorian Guard and his body thrown in the Tiber. Then followed his cousin, the emperor Alexander Severus, he ruled until 235.

The reign of Alexander was prosperous until he had to lead the army on several expeditions against the Sassanids in the East and later intruding barbarians in Gaul. His soldiers disliked his lack of military competence and drive, and started to look for another leader. They settled for *Maximinus the Thracian*.

Maximinus the Thracian was emperor only from 235 to 238, and thus led a short lived regime. Nevertheless, in some sense he is an important figure. He is the first *barbarian* to become emperor, and the first of the so-called *barracks emperors* of the third century. He also is the first emperor who never came to Rome. With him the Roman Empire entered *The Crisis of the Third Century.*

As the name suggests Maximinus was born in Thrace. He was reportedly 8 ft. 6 in. (2.59 m) tall and of tremendous strength. He had joined the army during the reign of Septimius Severus, but did not rise to a powerful position until promoted by Alexander Severus. When the troops elected the stern Maximinus, they killed young Alexander and his mother in 235. The Praetorian Guard acclaimed him emperor, and their choice was grudgingly confirmed by the Senate, who did not approve of having a barbarian peasant as emperor.

But Maximinus hated the nobility and was ruthless towards those he suspected of plotting against him.

Maximinus first campaign was against the Alamanni, who were defeated. After this victory he took the title Germanicus Maximus. As this part of the frontier was now under control, he set up his winter quarter at Sirmium in Pannonia, which today is named *Sremska Mitrovica* and lies in Serbia. During the winter he fought the Dacians, the inhabitants of present day Romania and Moldova, and the Sarmatians, a warlike coalition of Persian tribes which had moved from the Caspian plains to Eastern Europe and posed a threat the Romans.

Early in 238 there was a revolt in Africa, the governors there were declared emperors, and the senate supported them. Maximinus assembled his army and marched on Rome.

But there he encountered unexpected difficulties. Early 238 the Praetorian guards in his camp killed him, his son and his chief ministers.

This opens the period known as *The Anarchy*, or *The Crisis of the Third Century.* During 235–284 more than 20 emperors from varying provinces or army units, fought one another in a bloody struggle for power.

In 249 a wave of religious emotion swept the Empire. People flocked to the temples and besieged the gods with preyers. In this frenzy of fear and patriotism the Christians stood apart. They discouraged military service, scorned the pagan goods.

The Christians saw the disintegration of the Empire as the prophesied destruction of Babylon and the coming of Christ. Since the Christians stood for a radically new attitude, in that they considered poor people and slaves as equal to their masters, the influence of the Christians and their bishops became more and more important.

This led to ferocious persecution of the Christians, which however, by the Easter of 251, was practically at an end.

But 6 years later the current emperor, in response to a new crisis, ordered all persons to conform to Roman ceremonials. Christian assemblage was forbidden.

In 284 *Diocletian* became emperor, and we enter the epoch known as *The Later Roman Empire.*

4.17 Diophantus of Alexandria

[60] writes that Diophantus was born around 200 and died in 284. This would be the year when Diocletian became emperor.[13] But there is no historical source linking these two events.

[13] The source for the following account is [60], which we follow closely with some comments.

Essentially nothing is known of his life and there has been much debate regarding the date at which he lived.

Diophantus is some times referred to as the "father of algebra," but as is pointed out by several historians of mathematics this is quite misleading.

His best known work is *Arithmetica*, dealing with the solution of algebraic equations and on the theory of numbers.

Some details of Diophantus's life, which probably are fictitious, come from the Greek Anthology, compiled by Metrodorus around 500 A.D. This collection of puzzles contain one about Diophantus. In the Greek Anthology the problem is number 126 and it is asserted that this inscription stood on Diophantus' tombstone. But as there is a similar problem, namely number 124, it would seem likely that this was a representative of a common type of problems from the time of the Anthology. The actual text is translated as follows:

> *This tomb holds Diophantus. Ah, what great a marvel! the tomb tells scientifically the measure of his life. God granted him to be a boy for the sixth part of his life, and adding a twelfth part to this, he clothed his cheeks with down; He lit him the light of wedlock after a seventh part, and 5 years after his marriage He granted him a son. Alas! late born wretched child: After attaining the measure of half his father's life, chill Fate took him. After consoling his grief by this science of numbers for 4 years he ended his life.*

If we take this verse literally one might speculate if Diophantus, like Eratosthenes is claimed to have done, deliberately ended his life.

Now following [60]:

His boyhood lasted 1/6th of his life.
He married after 1/7th more.
His beard grew after 1/12th more.
And his son was born 5 years later.
The son lived to half his father's age.
And the father died 4 years after the son.

So he married at the age of 26 and had a son who died at the age of 42, 4 years before Diophantus himself died aged 84. Based on this information it might be presumed that he had a life span of 84 years, this being the reason for [60] putting his year of death to 284.

The Arithmetica is a collection of 130 problems giving numerical solutions of determinate equations (those with a unique solution), and indeterminate equations. The method for solving the latter is now known as Diophantine analysis. Only six of the original 13 books were thought to have survived and it was also thought that the others must have been lost quite soon after they were written. There are many Arabic translations, for example by Abu al Wafa, but only material from these six books appeared. Heath writes in [26]:

> The missing books were evidently lost at a very early date. Paul Tannery suggests that Hypatia's commentary extended only to the first six books, and that she left untouched the remaining seven, which, partly as a consequence, were first forgotten and then lost.

Now an Arabic manuscript in the library Astan-i Quds (The Holy Shrine library) in Meshed, Persia has a title claiming it is a translation by Qusta Ibn Luqa, who died in 912, of Books IV to VII of Arithmetica by Diophantus of Alexandria. This discovery was made in 1968. The eminent historian of mathematics *Roshdi Rashed* has compared the four books in this Arabic translation with the known six Greek books and claims that this text is a translation of the lost books of Diophantus. But some are not completely convinced, and a reviewer writes as follows:

> The reviewer, familiar with the Arabic text of this manuscript, does not doubt that this manuscript is the translation from the Greek text written in Alexandria but the great difference between the Greek books of Diophantus's Arithmetic combining questions of algebra with deep questions of the theory of numbers and these books containing only algebraic material make it very probable that this text was written not by Diophantus but by some one of his commentators (perhaps Hypatia?).

For details and references, see [60]. A more extensive treatment of the Arithmetica falls outside the scope of this book, but we have given a modern treatment of one aspect in Sect. 2.7.

4.18 Pappus of Alexandria

As is frequently the case with the ancient mathematicians, there has been some disagreement on his dates. Thus in [26] it is asserted that Pappus lived at the end of the third century A.D.

However, it can be deduced from Pappus' commentary on the Almagest that he observed an eclipse of the sun in Alexandria October 18, 320. For this and other reason one now fixes his date of birth to about 290 A.D., and his year of death to 350.

This puts his birth to the first years of the reign of *Diocletian*, who strengthened and reformed the government of the Roman Empire after a dismal century of civil war and disorder. But already when Pappus was 15 years old, Diocletian and his co-emperor *Maximian* abdicated, after which there followed 18 years of fighting between rival emperors. We give more details on this period in Sect. 4.9. During Pappus' life *Christianity* became the state religion.

Pappus was born in Alexandria, where he lived all his life. Proclus writes that he headed the school in Alexandria, which certainly stands to reason.

Pappus' major work in geometry is entitled *Synagogue* or the *Mathematical Collection*. It is a collection consisting of eight books, probably written around 340. In [26] Heath describes the Mathematical Collection as follows:

> Obviously written with the object of reviving the classical Greek geometry, it covers practically the whole field. It is, however, a handbook or guide to Greek geometry rather than an encyclopedia; it was intended, that is, to be read with the original works (where still extant) rather than to enable them to be dispensed with.... Without pretending to great originality, the whole work shows, on the part of the author, a thorough grasp of all the subjects treated,

> independence of judgement, mastery of technique; the style is terse and clear; in short, Pappus stands out as an accomplished and versatile mathematician, a worthy representative of the classical Greek geometry.

The golden age of Greek geometry had ended with Apollonius of Perga, but for a while there still were competent mathematicians who kept the field alive, although not producing any significant original research. But as time passed even such scholars became rare, and Heron of Alexandria was one of the last great expositors among them. The last one was Pappus of Alexandria.

But Pappus did indeed leave behind original research of considerable interest as well. We return to one of his celebrated theorems in Sect. 13.9.

Pappus gave the following extension of the Theorem of Pythagoras, see [26, vol. 2, p. 369]:

Theorem 4. *Let the triangle $\triangle ABC$ be given, and on the sides AB and AC construct arbitrary parallelograms ABDE and ACFG. Let DE and FG produced meet in the point H, and draw the line HA. Then the sum of the areas of ABDE and ACFG is equal to the area of the parallelogram BCML where L is on DH and BL parallel with AH, and M is on FG and CM parallel with AH.*

Proof. The situation is shown in the left side of Fig. 4.51. To the right, a special case of the situation is displayed which demonstrates that the theorem is a generalization of Pythagoras' Theorem.

To prove the assertion, we note that the area of the parallelogram *ABDE* is equal to the area of the parallelogram *BLHA*, since they have the common base *BA* and equal heights, namely the distance between the parallel lines *BA* and *DE*. But *BLHA* and *BLNK* have the same base *BL* and equal heights, namely the distance between the parallels *BL* and *KH*, thus they have equal areas. Thus *ABDE* and *NKBL* have equal areas. By the same argument *ACFG* and *NKCM* have equal areas, and the claim follows. □

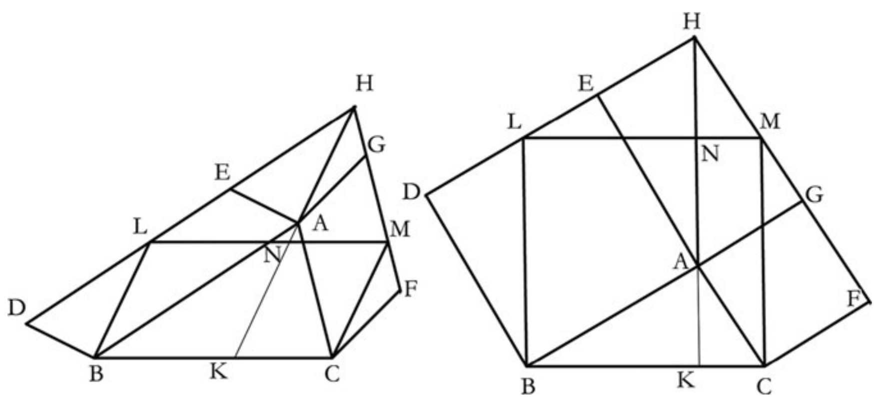


Fig. 4.51 Pappus' generalization of Pythagoras

4.19 The Late Roman Empire

Diocletian was emperor from 284 to 305. He established an autocratic government and laid the ground for the second phase of The Later Roman Empire. His reforms changed the government and helped stabilize the empire, so it could remain basically intact for the next 100 years.

Diocletian at first had to fight a series of wars from to maintain the extended boundaries of the frontiers and stamping out domestic uprisings. By 298, however, his position was secure.

Diocletian believed that the current system of government was unsustainable, and initiated a number of reforms to prevent a return to anarchy. He divided the Empire in two parts, to make it more manageable, he created a new system of succession and removed any the remaining facade of the old Republic. Economic reforms to curb hyperinflation were also put in place.

The republican system had continued to exist in form, while not in reality. Now Diocletian build a new basis for imperial legitimacy in *the state religion*, with himself as semi-divine monarch and high priest. In this context the old tile of *Pontifex Maximus* became more important for the emperor.

Diocletian took the title *Dominus et deus*, Lord and God. He sat on a throne, and was not to be seen in public. If an audience was required, the visitor was required to lie on the ground prostrate and never to look at the emperor.

Diocletian had concluded that the empire was too big for a single emperor. His solution was to split the empire in two, drawing the border just east of Rome.

Diocletian created the following system: A senior emperor should rule in the East and another senior emperor in the West. Each of them would have a junior emperor next to him. The most important title of *Augustus* was reserved for senior emperors, and the junior emperors got the title of Caesar. Diocletian intended that when the senior emperor retired or died, the Caesar would take his place and choose a new junior emperor Caesar, thus solving the problem of succession.

By 292, Diocletian had the system in place and chose the Eastern Empire for himself and gave Maximian the Western Empire. The imperial power was now divided between two people. The two men established separate capitals, neither of which was at Rome. The ancient capital was too far removed from the places where the empire's fate was decided by force of arms. While improving the ability of the two emperors to rule the empire, the division of power further marginalized the senate, which remained in Rome. In 293, Diocletian and Maximian each appointed a Caesar, Galerius and Constantius Chlorus (the Pale), respectively, formally adopting them as their heirs. However, these were not merely successors – each was given authority over roughly a quarter of the Empire. This system of government is known as the *Tetrarchy*.

Now the following happened: In 302, at an imperial sacrifice, the Christians made the sign of the cross to ward off evil demons. Then the priests failed to find the marks on the livers of the sacrificed animals which would signify that the sacrifices had gone well. This was blamed on the presence of unbelieving persons. Diocletian then ordered that all present should offer sacrifices to the gods or be flogged, and the soldiers should also conform or be dismissed.

The year after the four rulers decreed the destruction of all Christian churches and the burning of Christian books and dissolution of Christian congregations. This time, however, the Christians fought back, and even set fire to Diocletian's palace. Diocletian, infuriated, ordered every Christian to be sought out and compelled to worship the Roman Gods by all torture available. Then he resigned.

Gaius Flavius Valerius Aurelius Constantinus or *Constantine the Great* also referred to as Saint Constantine, was a emperor 306–337. He was the son of Constantius Chlorus and his concubine or wife *Helena*, 255–330, who came from a modest background. Constantius was compelled to repudiate her when he became Caesar in 292, but the son Constantine honored her, and she is considered a Roman empress. She was under strong influence of Christianity. In 306 when her son became emperor he gave her an important position. She has later been sainted. She was important for the victory of Christianity in the Roman empire by influencing her son, who fully legalized Christianity in the Edict of Milan in 313. She converted to Christianity, something her son did only on his deathbed when he was baptized (Fig. 4.52).

Constantine went through a series of fights with his rivals, and eventually removed the competition for power. His last colleague and rival was *Licinius*, who was co-emperor from 308 to 324.

In 324 Constantine declared war against him for the last time, and defeated his army at the Battle of Adrianople in 324. Licinius took refuge behind the walls of Byzantium, then his fleet was defeated, and he withdrew to Bithynia, where a final battle resulted in Licinius being interned at Thessalonica. There he was assassinated

Fig. 4.52 The Baptism of Constantine, as imagined by students of Raphael

together with a former co-emperor Sextus Martinianus. Now Constantine was the sole emperor.

The Council of Nicaea in 325 could take place without suppression, but Constantine also continued to support pagan deities.

Constantine rebuilt the old city of Byzantium, and renamed it *Nova Roma*. The city got its own senate and officials as in Rome. The figures of old gods were replaced by Christian symbols.

In 330 Constantine made Nova Roma the capital city of the Empire. After his death in 337, his capital was renamed *Constantinopolis* or Constantinople.

After Constantine's death there followed a struggle between his three sons Constantine II, 337–340, Constantius, 337–361 and Constans, 337–350. The first and the last were slain in battles, and from 351 the sole emperor was Constantius. Constantius had reluctantly proclaimed his cousin Julian Caesar, and sent him to Gaul to meet the threat of the barbarians there. But Julian was too successful, Constantius became worried and ordered him away, the soldiers revolted an proclaimed Julian as Augustus, as emperor. Julian reluctantly accepted, and the death of Constantius in 361 saved the Empire from another round of civil war (Fig. 4.53).

During his brief tenure of sole power 361–363 he attempted to restore the pagan Hellenistic worships, which has earned him the name "Julian the Apostate." He wrote in a letter that Christianity was forced on him as a child by his cousin Constantius, who was a zealous Arian Christian and would tolerate a Pagan relative. It has been said that he reacted against the Christian teaching he had received in his lonely

Fig. 4.53 Julian the Apostate. Last pagan emperor

and miserable childhood. He found solace in the mysticism of his contemporary Neoplatonist philosophers.

Julian reduced the influence of Christian bishops. The land taken by the Church were to be returned to their original owners, and the bishops lost the privilege to travel for free, at expenses of the State.

In 362 Julian proclaimed an edict to guarantee freedom of religion, reverting edicts from 353 and 356 by Constantius which had made Christianity the most influential religion in the Roman Empire. This edict proclaimed that all the religions were equal under the Law, and that the Roman Empire had to return to its policy that the Roman State did not impose any religion on its provinces.

He effectively shored up the barbarian advance in the west, probably significantly delaying the loss of the western part of the empire. He also attempted to check the luxurious practices which had grown up in the court of Constantius.

His campaign in Persia was well conducted and at first successful. But then he was killed, and was succeeded by Jovian, who was with him on the campaign. This caused some surprise, but the election might have been caused by mistaken identity. However, Jovian died a short time later.

Valentian and his brother *Valens* then became joint emperors, Valentian in the west and Valens in the east. Valentian seems to have been the more competent of the two, and when he died in 375 Valens was incapable of dealing with the crisis caused by the Goths, who pressed by the Huns sought protection with the Romans. Valens was defeated and killed in 378. *Gratian*, who had succeeded Valentian in the East now proclaimed *Theodosius* emperor in the West. Gratian was murdered in 383, however, and there followed a sequence of usurpers and murders, Theodosius being busy defeating usurpers and installing his own candidates. This ended with his own illness and death in 395. This year marks the definitive separation of the Roman Empire into the East Roman Empire and the West Roman Empire. In the West *Honorius* became emperor 395–423, in the East first *Arcadius*, 395–408, then *Theodosius II*, 408–450.

4.20 The Murder of Hypatia

We have taken a long step from the time of Euclid. Six hundred years has passed. Alexandria has developed into a magnificent metropolis. Mathematics and geometry, science and the humanities have been nurtured. The greatest minds of the known world have spent time there as students or as visitors. Now, 600 years after Euclid, we find a considerably lesser geometer in Alexandria. Lesser than Euclid, but in no way insignificant.

His name was Theon, Theon of Alexandria. He was active in the fourth century A.D. He wrote commentaries to the unquestioned mathematical masterpiece, *Euclid's Elements.*

Some historians of mathematics slightly contemptuously refer to Theon and others as *The Commentators*. This might imply a certain lack of originality, and a period

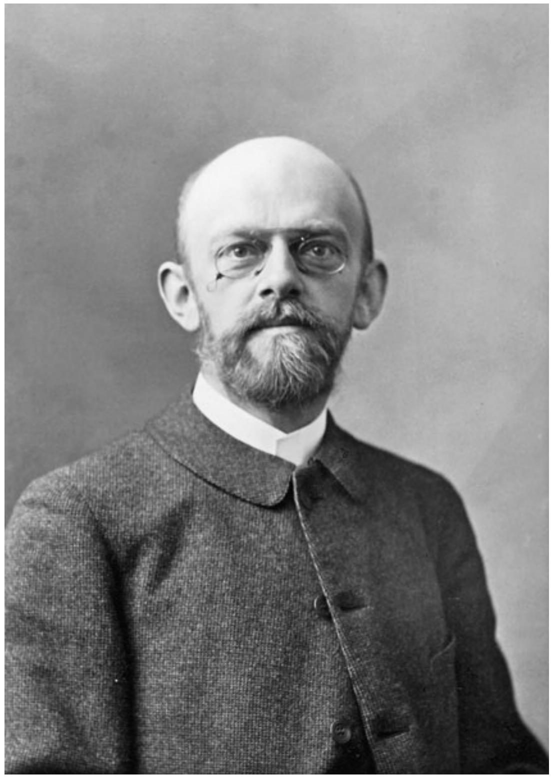


Fig. 4.54 David Hilbert

of decline in Alexandria. In general the mathematical research in Alexandria was indeed in a state of decline at Theon's time. But for Theon that judgment is not an altogether just one.

This is best illustrated by moving even further ahead in time, in fact a staggering leap of sixteenth centuries. *David Hilbert* (1862–1943) is one of the greatest minds of modern mathematics, and as a mathematician he reaches far above Theon, if we may compare people separated by millennia (Fig. 4.54). In 1899 he published his work [29]. Is this fundamentally important work "merely" commentaries?

But of course, Hilbert's commentaries on Euclid are far deeper than Theon's.

David Hilbert was much reverenced in his time, also by the present rulers at the time of his death. But he never concealed his contempt for the Nazis in power. It is told that just before his death, he was asked by one of the leading figures: *"– Now, Herr Professor Hilbert, how is your Institute now that we have gotten rid of all the Jews?"* Hilbert looked at his questioner and answered coldly: *"– The question is easy to answer. My institute does not exist any more."*

At this time the population of Alexandria had grown quite cosmopolitan. Besides Egyptians, it consisted of Greeks and Jews. Many among the Greeks had converted to Christianity.

The Old Testament had been translated into Greek much earlier. Known as *The Septuagint*, an interesting legend relates how it was created. We tell this story by quoting *The Catholic Encyclopedia* [7]:

> The Septuagint Version is first mentioned in a letter of Aristeas to his brother Philocrates. Here, in substance, is what we read of the origin of the version. Ptolemy II Philadelphus, King of Egypt (287–47 B.C.) had recently established a valuable library at Alexandria. He was persuaded by Demetrius of Phalarus, chief librarian, to enrich it with a copy of the sacred books of the Jews. To win the good graces of this people, Ptolemy, by the advice of Aristeas, an officer of the royal guard, an Egyptian by birth and a pagan by religion, emancipated 100,000 slaves in different parts of his kingdom. He then sent delegates, among whom was Aristeas, to Jerusalem, to ask Eleazar, the Jewish high-priest, to provide him with a copy of the Law, and Jews capable of translating it into Greek. The embassy was successful: A richly ornamented copy of the Law was sent to him and 72 Israelites, six from each tribe, were deputed to go to Egypt and carry out the wish of the king. They were received with great honor and during 7 days astonished everyone by the wisdom they displayed in answering 72 questions which they were asked; then they were led into the solitary island of Pharos, where they began their work, translating the Law, helping one another and comparing translations in proportion as they finished them. At the end of 72 days, their work was completed. The translation was read in presence of the Jewish priests, princes, and people assembled at Alexandria, who all recognized and praised its perfect conformity with the Hebrew original. The king was greatly pleased with the work and had it placed in the library.

The Alexandrian Museum, or the *Academy of Alexandria* was a true temple for learning, scientific pursuits and culture. The foremost thinkers, philosophers, mathematicians and scientists of the world lived here. At this unique institution they were still free to carry out their spiritual activities according to their own wishes. The library in Alexandria still contained some of the finest works produced by mankind.

Theon had a daughter, who became one of the greatest names within philosophy, mathematics and other sciences. Hypatia was born around the year 355 A.D. Some put her year of birth at 370 A.D., but M. Dzielska argues persuasively for 355 in [11]. She is the first woman mathematician we know with absolute certainty, although undoubtedly there were others before her. She worked with her father, Theon, on commentaries to the great geographer, astronomer and mathematician *Claudius Ptolemy's* work, and revisions of *Euclid's Elements.* She also wrote commentaries to the works of the great classical mathematicians *Apollonius and Diophantus.*

Much of her work, as well as that of numerous other mathematicians from antiquity, is lost in its original form. It is only known to us through copies, translations, summaries and as rendered by "Commentators."

Hypatia gave lectures and did research, and around the year 400 she became the leading philosopher of the *Neoplatonic Academy in Alexandria*. She also became the Head of the *Museum of Alexandria* and the *Library*. She then had reached the peak within the Alexandrian intellectual élite, as the unquestioned leader of cultural life there.

Hypatia must have been a remarkable person, in more than one way. She was superbly gifted as a scientist and scholar. Eminent thinkers from the entire antique world travelled to Alexandria to hear her speak. She also was a very beautiful woman. Numerous were the offers of marriage she had received from kings and noblemen, but politely turned down: She preferred to devote her life to the pursuit of the eternal truths in philosophy and mathematics.

But Hypatia lived during turbulent times. This period was dominated by the struggle between old ideas which had been forming the core of the antique civilization on one hand, and emerging new ones on the other. The confrontation was hard and merciless. The new religion, *Christianity,* was on the rise. And the old Gods like Enlil and Ishtar, Zeus, Apollo and Athena, Jupiter and Venus, or the most important one in Egypt at this time, *Serapis*,[14] they were *pagan gods*, reprehensible to the Christian zealot. Paganism had to be fought by all means available.

Fanaticism and fundamentalistic narrow-mindedness and bigotry have been present throughout human history. Combined with xenophobia and disdain for the *"aberrant*," as well as with plain and simple *ignorance*, these negative forces of human life have haunted humanity since time immemorial. Such collective tendencies of human nature may well lie dormant under the surface during shorter or longer intervals, and then suddenly flare up under some contemporary pretext, igniting the flames of a Holocaust, of a Witch Hunt, ultimately an apocalyptic conflagration in which civilizations are reduced to smoking ruins.

One of the features which we keep finding time and again, is the need these fundamentalists have for some groups of *scapegoats.* Groups of people who are anathematized as reprehensible enemies, be it for their race, ethnicity, beliefs or sexual orientation. Women are especially vulnerable, above all if they "do not know their proper place." The zealot will frequently stop at nothing, certainly not physical elimination, murder.

Cyril of Alexandria is the name of a man who was elected Patriarch of Alexandria in the year 412 A.D. He was a very partisan warrior for the young Christian Church. An uncompromising guardian of the *true faith*, who not only fought vigorously and without scruples against the unbending *pagans*, but also used all means at his disposal to go after and fight down the abominable *heretics* within the church itself. His detest for Hypatia and all she stood for was intense. She was the epitome of everything he hated.

Cyril regarded the Jews as dangerous enemies. But before he could turn his attention to them, he had to secure his position within his own Church.

Novatianus was one of the early leaders of the Christian Church in Rome. Born about 200 and martyred in 258, he had a high reputation as a learned theologian, but he lost out to a rival named *Cornelius* in the vote for pope. A minority declared itself for Novatianus, who then became the second *antipope* in the history of the Church. His views would in many ways be reprehensible to us today, for example he refused reentry into the Church for those unfortunate Christians who had denounced their faith as a result of persecution. Novatianus and his followers were excommunicated in 251. The schism developed into a sect which spread across the entire Roman

[14] Serapis, or *Osarapis* was the dead Apis worshiped as Osiris. He was the lord of the Nether world, and the Serapis cult incorporated elements of the Greek Gods into the traditional Egyptian ones.

Empire. Novatianus managed to build his own church with his own bishops, and the *Novatians* still had some following in Alexandria at this time. Cyril had to deal with the Novatians in Alexandria, which he did. Their churches were stormed, plundered and burned, the unrepentant killed or driven away, but the ones who converted to the true faith were graciously forgiven.

Having eliminated opposition from Christian *heretics*, Cyril was now free to turn to external enemies. He did so with a vengeance. As far as I know this is the first instance of a large scale pogrom, where practically the entire Jewish population of Alexandria was eliminated, driven away or murdered. Their homes were pillaged and then ignited, their property plundered. The synagogues were burned down. This infamous crime rests on the conscience of concerned Christians of today. Adding to this picture is the unbelievable "explanation" offered by some historians that the Jews had "provoked" the Christians first, by attacking them! It sounds all too hauntingly familiar.

Cyril was now at last strong enough to take on *The Pagans*. His success with the Christian heretics and the Jews, made him confident that he would soon have cleansed this sinful city. And he understood full well that in order to win, he would have to aim for the top. Thus Hypatia had to be eliminated. She had stubbornly refused to become a Christian, and unrepentingly stood by her pagan beliefs, worshiping the sorcerer Pythagoras and his followers with their satanic secret rites. She led young people astray with her talk about the old pagans Socrates and Plato. But she had allies in Alexandria, powerful allies. The allies of Hypatia would have to be immobilized first.

Hypatia's most prominent friend in Alexandria was no other than the Roman Prefect there, an enlightened man named Orestes. He had studied at the Academy, under direction of Hypatia. And he harbored a deep sense of admiration and esteem for his former teacher, now a dear friend and close advisor. Even Cyril could not dare to cross the Roman Prefect. Not yet.

Orestes had kept the peace, he had kept the *Pax Romanum*, the Roman Peace of which the Romans prided themselves. After the final conquest of Egypt, when Cleopatra and Anthony were defeated and utterly destroyed in 31 B.C., the Romans had brought peace and, by and large, prosperity to the region. Certainly order and the rule of law. But now the Empire was in decline. Rome had been hard pressed from many sides, and just to hold the outer provinces together had become a heavy burden. At the death of Theodosius in 395 the Roman Empire was formally split into the East and the West Empire. One of his sons, *Honorius*, had become Emperor in the West, but his power rapidly fell apart, the West Roman Empire had now less than a 100 years left before its final fall. See Sect. 4.9 for more on this. In the East a second son of Theodosius, *Arcadius*, became Emperor. Alexandria and the rest of Egypt had belonged to the eastern part of the Roman Empire, and was ruled from Constantinople after the separation. Arcadius was succeeded by Theodosius II in 408, who originally entertained dreams of reuniting the Roman Empire as a mighty power. But the pressure in the West from aggressive Germanic tribes proved too strong, and the ideas of a reconstructed Empire eventually had to be abandoned.

This was the political situation at the time when Cyril became the leader of the Christian Church in Alexandria.

Orestes had been a good man to have in Alexandria for the emperors in Rome and in Constantinople. He had kept that part of the Empire comparatively quiet. This was an important prerequisite for Theodosius' plans of reuniting the Empire under his own rule.

Cyril must have realized that this was Orestes' weakness. Cyril could muster troops of ardent followers on short notice. And he had organized an echelon of lieutenants, taking their commands from him and carrying them out precisely as ordered.

Orestes soon had serious problems on his hands. Riots flared up all over Alexandria. The mobs chanted their accusations against the ungodly Romans, who showed such respect for the pagan sorcerers, based at the Museum and in the Library where these ideas were preserved. Hypatia was singled out, as having poisoned Orestes' mind against the pious Cyril. Orestes' hands were tied, or so he felt. News of him having sent the Roman legions out onto the streets of Alexandria to put down the disturbances would certainly not have been understandingly received in Constantinople. He had himself become a Christian, out of political convenience more than anything else. Cyril and his followers were not impressed. And as the pogroms were in the making, he had intervened and arrested, at the behest of some of the persecuted Jews, one of Cyril's lieutenants named Hierax. For good measure he had him tortured as well. But Orestes had grossly overplayed his hand, and Hierax had to be freed, emerging as a martyr and hero to Cyril's followers. Not to speak of the unbelievable incident, when he himself was bodily attacked on the street by a mob, led by the monk Ammonius. The same Ammonius had hit him in the head with a stone, causing him to bleed profusely. As the courage of his guards wavered, brave citizen of Alexandria rushed to his assistance, and Ammonius was seized and brought before the enraged Orestes, still bleeding from his wound in the head. Ammonius, far from receiving a fair trial, was sentenced to severe torture, from which he died, thus transformed into another martyr of the Church. Orestes' account with the Emperor in Constantinople was already overdraw. Orestes knew this perfectly well. And he knew all too well who and what where the real targets for this frenzy: The Neoplatonians at the Museum of Alexandria, and above all, their spiritual leader Hypatia. The writings were on the wall.

But Hypatia felt unable to remain silent in the face of the injustices and atrocities committed by Cyril and his people. And she certainly did not wish to flee Alexandria. So she stayed on, apparently not really being able to believe that Cyril would harm her because of her science and philosophy. She attempted to support Orestes, in his feeble efforts to resist Cyril.

Cyril had put one of his monks named *Peter* in charge. They waited for her as she rode home from her lecture at the Museum in her chariot.

Socrates Scolasticus was born in Constantinople towards the end of the fourth century A.D. He was a historian of the early Church, and wrote the fundamental source *Ekklesiastike historia* (Ecclesiastical History). In a time of turmoil and acrimonious disputes, he is generally credited for striving to avoid the animosities and hatred often engendered by theological disputes in these times. Himself not being a priest, he honored clerics and venerated monks, but also urged the study of works by pagan authors. As a historian he is credited with thorough research and with seeking out the primary sources. See [7] for more details. He has related what happened next in Alexandria, which we summarize as follows omitting the graphic details:

> *In front of the Caesarum Church her chariot was stopped, and she was pulled down. They dragged her into the church, where she was killed. They then cut her body in pieces, carried it to a place called Cinaron where her remains were burned.*

Orestes resigned and left Alexandria. The City Council reported her murder to the court in Constantinople, and demanded an investigation. But no serious investigation ever took place. Few witnesses could be found and no evidence against Cyril seemed to exist. But apparently a certain militia controlled by the church was reorganized.

Cyril went on to new victories in forming the dogmas of the Church, which was still united up until 1054. His writings on ethical and theological questions are extensive, and after his death he was sainted by the Church. St. Cyril's day is June 27.[15]

The Museum and the Library were burned down. Much of what remained of the Library was burned already in 392 A.D., when the Christians destroyed the Temple of Serapis, which had also been a center of learning and culture. In any case, the year 415 A.D. marks the beginning of the end of antique civilization, and the end of the beginning of the dark Middle Age.

Part of Theon's mathematical and astronomical work has survived. This include a student edition of Euclid's Elements *The Data* and *The Optics*, which were used by Byzantine scholars in their effort to reconstruct Euclid's work. Also preserved are his commentaries on work by Claudius Ptolemy. Theon also comments on work by the geometer *Menelaus of Alexandria*. Theon's and Hypatia's mathematical and astronomical work also relied on work by another geometer of Alexandria at that time, *Pappus*. Hypatia's mathematical work has been presumed lost. But recent research indicate that we may be able to piece together her contributions. As it probably happened, her work did not get lost, she "just" did not get credit for it. Not an infrequent occurrence in the history of mathematics. Thus efforts to reconstruct her work on Apollonius' *The Conic Sections* indicate that she made substantial contributions. Also, it is now believed that the survival of Diophantus' *Arithmetica* is due in large part to Hypatia's elucidation. Theon of Alexandria, Hypatia's father, was very much engulfed in astrology and Babylonian mysticism, through his strong involvement with Pythagorean doctrine and philosophical thinking. Thus it is not unlikely that some of this material formed part of the ancient insights which flowed to Alexandria from what had been the Babylonian Empire.

For more on these questions we refer to Dzielska [11]. This reference also recounts the events in Alexandria. I have included some of the details from the narrative given there, but omitted others. I admit that my position is a personal one,

---

[15] There are four more saints by the name Cyril, among them the monk who designed the Russian alphabet.

Dzielska views Cyril of Alexandria in a somewhat more favorable light than the present author does.

4.21 Fall of the Roman Empire

From 383 to 395 the Roman Empire was led by *Theodosius*, emperor of the East but engrossed by the duty of upholding the feeble authority of his colleges in the West. The latter were under an increasing pressure from the barbarians, who now threatened to overrun the western part of the empire. When he fell ill and died in 395, his two sons took over the East and the West, respectively, and the separation into the West Roman Empire and the East Roman Empire was formally established.

When the emperor in the West, Theodosius' son Honorius, died in 423 his authority had been seriously eroded. Valentian III ruled for a relatively long time, from 423 to 455 when he was murdered. During his reign the province of Africa was lost, and Atilla invaded Gaul and Italy. He was repelled thanks to the aid of the Christened and half romanized Visigoths.

In 455 *Maximus* was emperor for 3 months, during which time Rome was overrun, plundered and partly burned by the Vandals. From now on the emperors in Rome ruled on the mercy of the *barbarian mercenaries*.

The last emperor was *Romulus Augustus*. His father Orestes, the commanding general of the Roman army, had installed Romulus on the throne after deposing the emperor *Julius Nepos*. Romulus is therefore considered a usurper by some historians, who say that the last emperor was Nepos. Romulus Augustus was a mere child and acted as a figurehead for his father's rule. In any case his reign, or "reign," lasted only 10 months. When tried to assert his authority, or more precisely his father *Orestes* tried to do so, he was ousted as emperor and the father killed. Exit the Emperor, enter *Odoacer the Rugian*, King of Italy. This happened in 476, which we call the year of the fall of the Roman Empire. But the East Empire lived on, however slowly lost its power and glory.

4.22 Byzantium

The East Empire came to be named *Byzantium*, it finally fell to the Turks in 1453.

The most famous and important of all Byzantine emperors, is Flavius Anicius Justinianus, Justinian I, or as he is also known, *Justinian the Great*. He was born in 483 and died in 565. He was a *barbarian* by birth, born in Thrace, according to some sources his name originally was *Petrus Sabbatius*. As a youth he came to Constantinople, and there he received a very good education. That was no accident, his uncle had risen from the ranks of the army to become emperor. His uncle's name was *Justin*, son of a Macedonian peasant. Justin had gone through the ranks in the army, and become head of the Praetorian guard of emperor Anastasius. Anastasius died in 518, and his successor had to be elected. The following story is told about how the election went:

As was frequently the case the new emperor should be elected by a vote in the Praetorian Guard, and Justin was approached by one of the candidates who wanted to bribe the Guard. He left a sizable sum of money with Justin, and asked him to secure his election by distributing the money and asking for the favor. However, Justin was a devious man. He did indeed distribute the money, but told the men that this money came from himself, and more would be forthcoming as soon as he had been elected. Thus he was elected emperor himself, at the mature age of 68.

Gibbon[16] gives a similar account, which is not completely identical to the above but still confirms that essentially something like this happened. We now quote from [17, vol.2, chap. 40]:

> Under the two succeeding reigns, the fortunate peasant [Justin] emerged to wealth and honors [. . .] [He] gradually obtained the rank of tribune, of count, and of general, the dignity of senator, and the command of the guards, who obeyed him as their chief, at the important crisis when the emperor Anastasius was removed from the world. [. . .] [The] eunuch Amantius, who reigned in the palace, had secretly resolved to fix the diadem on the head of the most obsequious of his creatures. A liberal donative, to conciliate the suffrage of the guards, was intrusted for that purpose in the hands of their commander [Justin]. [. . .] But these weighty arguments were treacherously employed by Justin in his own favor; and as no competitor presumed to appear, the Dacian peasant[17] was invested with the purple by the unanimous consent of the soldiers, who knew him to be brave and gentle, of the clergy and people, who believed him to be orthodox, and of the provincials, who yielded a blind and implicit submission to the will of the capital.

However, Justin is supposed to have remained illiterate all his life. He depended on his nephew Flavius Petrus Sabbatius, who he adopted and gave the name *Justinian*. In 527 he succeeded his uncle and adopted father as emperor. In 522 he had married *Theodora*, she was 20 years younger than himself, and according to some sources she was a by profession a courtesan and an actress. Theodora became very influential politically, at first the marriage was a source of scandal. But Theodora proved to be very intelligent, she was a good judge of character and was Justinian's greatest advisor and supporter. Indeed, on his accession to the throne in 527 he made her joint ruler of the empire. This proved to be a wise decision. A strong-willed woman, she showed a notable talent for governance. In the Nika riots of 532, her advice and leadership for a strong response saved the empire.

Much of the unfavorable information from this earliest part of her life comes from the contemporary historian Procopius' *Anecdota* or "Not Published," which understandably enough was published only posthumously. Critics of Procopius have dismissed this work as vitriolic and pornographic, but have been unable to discredit his facts. But Procopius had led an eventful career in the service of his rulers, which in the end left him very disillusioned with them. He was born in Caesarea, presently

[16] Edward Gibbon (1737–1794) was an influential and controversial English historian and Member of Parliament. His most important work is *The History of the Decline and Fall of the Roman Empire*, which was published in six volumes between 1776 and 1788. This work is known for the quality and irony of its prose, its use of primary sources, and its open denigration of organized religion, and is easily available as [17] and [18]. (Source [61]).

[17] Ancient Dacia was part of Thrace.

in Israel, on the sea-coast about halfway between Tel Aviv and Haifa. He received education in the Greek classics, attended law school, possibly at Berytus in modern Beirut, and became a rhetor or lawyer. In 527, the first year of Eastern Roman Emperor Justinian I's reign, he became legal adviser for Belisarius, Justinian's chief military commander.

Theodora became devoted in her Christian belief, but she was an adherent of the Monophysites, who taught that Christ was of one nature. Some argue that her association with Monophysitism was because Justinian had put her in charge of courting the Monophysites' reunion with the Chalcedonian party in the Church, and so while remaining Chalcedonian herself, she was pastorally favorable toward the non-Chalcedonians. The Chalcedonians taught the full humanity and full divinity of Christ. This point of view won out in the end, the Chalcedonian churches now include the Roman Catholic and Eastern Orthodox churches, as well as most Protestants.

Theodora might have been Byzantium's first proponent of abortion. She also advocated the rights of married women to commit adultery, putting them on an equal footing to their men, and the rights of women to be socially serviced, helping to advance protection of various kinds for them. She showed genuine concern for prostitutes and other downtrodden people. Both Justinian and Theodora are now sainted.

In 535 Justinian opened an offensive westwards, with the intention of restoring the old Roman Empire. His generals *Belisarius* and *Narses* conquered Italy as well as a large portion of the North African coast, as well as southern Spain. Again the shores of the Mediterranean were essentially Roman land, ant the Mediterranean was once more *Mare Nostrum*. The largest extension of the Byzantine Empire was reached in 565, after which it started to fall apart. But by the middle of the ninth century it still included Sardinia, Sicily and southern Italy.

The conquest by Justinian did not carry with it a full revitalization of the culture and mathematics of the old days, however. In fact Justinian was the one who closed the Academy in Athens, as being *pagan*.

4.23 Preservation of a Heritage

For a while antique mathematics and philosophy lingered on, in Alexandria and elsewhere. Longest, perhaps, lasted the Academy of Constantinople, where many works were preserved. *Proclus* (410–485) headed a Neoplatonic Academy in Athens, where he wrote *Commentaries on the First Book of Euclid*. This work is, as already explained, our main source for the history of early Greek mathematics, as so many of the originals went lost by the conflagration in Alexandria, and for other reasons as well.

Ammonius is reported to have been a student of Proclus. It would be consistent with this to put his year of birth to around 450 A.D. He wrote commentaries on

Aristotle, and Ammonius was appointed head of the Alexandrian school, which still existed in his time.

Ammonius had two students in Alexandria, who both contributed significantly to the preservation of the classical heritage. *Eutocius of Ascalon* was born around 480 A.D. and died about the year 540. Ascalon, now named Ashqelon in Israel, was the city of *Herod the Great*. It had an old history when it was conquered by Alexander the Great in 332 B.C., then it became a Roman city in 104 B.C. and finally was destroyed completely during the Crusades. Excavations at the cite reveal what a magnificent city this was.

Eutocius wrote commentaries on Archimedes and Apollonius. As we have seen in Sect. 4.4, his commentaries on Archimedes' work *On the Sphere and the Cylinder* served to preserve this important work for posterity.

Another student of Ammonius was *Simplicius of Cilicia*. Cilicia is located in southern Anatolia in present day Turkey. After completing his studies under Ammonius of Alexandria, he went to Athens where the Academy of Plato was still in existence. There he studied under the Neoplatonian Damacius, who had become head of the Academy around 520 A.D.

Simplicius is given credit for preserving numerous classical works for posterity, through his comments and writings. Another important teacher of mathematics around this time was the architect and mechanical engineer *Isidorus of Miletus*, who directed the building of the Hagia Sophia (*Holy Wisdom*) in 537.

The picture shows Hagia Sophia as it stands today. The cathedral was converted to a mosque after the fall of Constantinople, and the minarets were then added. Today the Hagia Sophia is a museum.

In 529 A.D. the Academy in Athens was closed by the (East Roman) Emperor Justinian as being pagan. Damacius, who was still the director of the Academy, together with Simplicius and others had to flee, and were well received by the Persian King Khrosrow I, an enlightened patron of philosophy and culture. Although the exiles were able to return to Athens under a peace agreement worked out between Justinian and Khrosrow in 532 A.D., their freedom of expression was now severely constrained.

Exercises

Exercise 4.1 Referring to Fig. 4.15, recall that Archimedes considered the tangent line to the spiral at the point of the first rotation, and its intersection T with the turning line at the 270 degree rotation. Then the length of OT is exactly equal to the circumference of the circle about O of radius OP. Archimedes used this result to prove that the problems of the *rectification of the circle*, and the quadrature of the circle, are equivalent. Recall from the text how this is proved, assuming his result on the distance OT. Then use modern calculus to prove that the length of OT is equal to the circumference of the circle with radius OP.

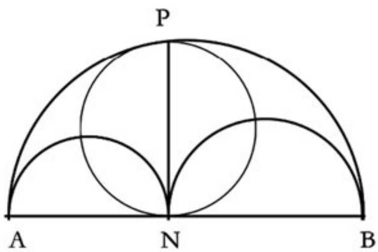
Archimedes indicates that the same idea might be applied to show that the volume of a sphere is equal to the volume of a circular cone (or a pyramid) with the area of the base equal to the surface area of the sphere and height equal to the radius. He does not appear to be completely confident that this can actually be built out to a valid proof. What do you think?

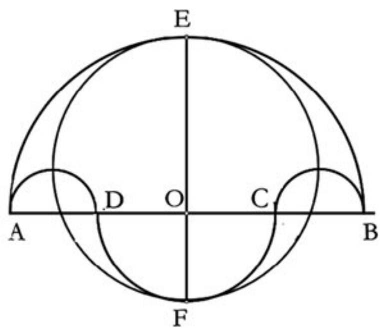
Exercise 4.2 *([14, p. 407]) The following statement is known as* the Axiom of Archimedes:

> *When we are given two magnitudes of the same kind, then we can find a multiple of the smaller which exceeds the larger.*

Use this axiom to prove the statement Euclid X.1 given at the end of Sect. 4.2. Then use Euclid X.1 to prove that the difference in area between a circle and a circumscribed regular polygon can be made smaller than any preassigned (small) area.

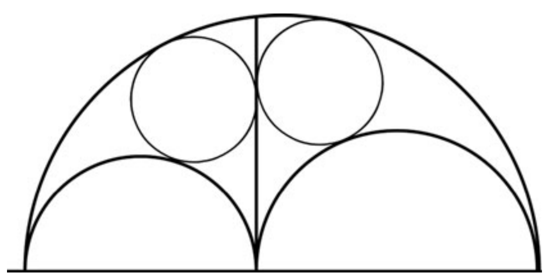
Exercise 4.3 Verify Archimedes' result on the area of the arbelos and the salinon:





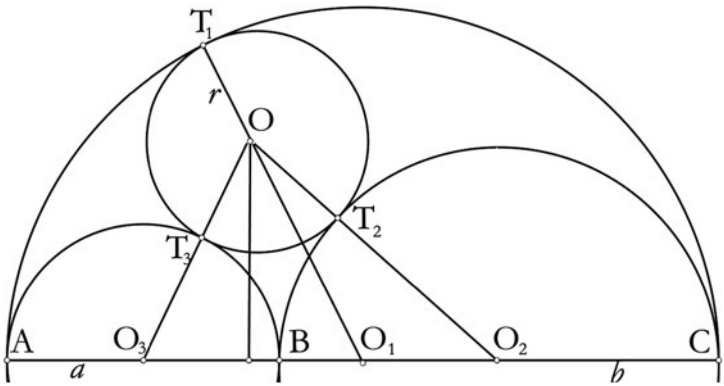
He proved that the area of the arbelos is equal to the area of the inscribed circle with diameter NP shown to the left. He also showed that the area of the salinon equals the area of the circle of diameter EF shown to the right.

Exercise 4.4 Verify Archimedes' result on the two tangent circles:



He proved that the two circles shown above have equal diameters. Prove this by computing the diameters of both.

Exercise 4.5 Archimedes computed the diameter of a circle tangent to all three semicircles defining the arbelos:



The arbelos is defined by the semicircle of diameter $AC = 2(a + b)$ with center O_1, the semicircle of diameter $AB = 2a$ about O_3, and the semicircle of diameter $BC = 2b$ about O_2 The circle of center O and radius r is tangent to all three semicircles. Show that then

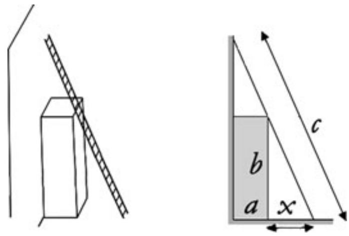
$$r = \frac{a^2b + ab^2}{a^2 + ab + b^2}.$$

Exercise 4.6 A planet of radius r has an atmosphere extending out a distance h from the surface. What is the volume occupied by the atmosphere?

Use the statement about the volume of a sphere by Archimedes to find a (modern) formula similar to the Egyptian "formula" for the volume of a frustum of a right pyramid directly in terms of the surface areas of the planet and its upper atmospheric limit.

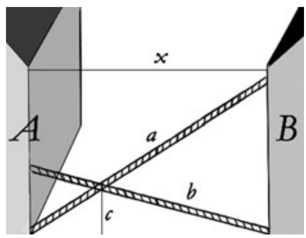
Exercise 4.7 The the simplest example from the class of problems known as *Ladder Problems* is the problem from the Old Babylonian Epoch given in Exercise 2.2. A considerably more difficult one is the classical *Ladder Box Problem*. The first documented occurrence of this problem is in work by Pappus, who attributes it to Apollonius.[18] In this case a ladder of length c is erected against a wall such that it rests on a box of height b and side a as shown below. In this form it is called *The Ladder Box Problem:*

[18] We follow an exposition given in [30].



The problem is the following: Given line segments a, b and c, find x. In the classical spirit, this means to construct x using ruler and compass. At this point you can find an equation which determines x uniquely, and when you have finished reading this book, you will be able to determine whether the construction is possible or not.

Exercise 4.8 *The Crossed Ladders Problem* deals with two ladders of lengths a and b, respectively, being erected as shown below.



The problem is to find the distance x between the two houses A and B, when the height c above the ground of the crossing-point is known.

Exercise 4.9 Let α, β and γ be three angles such that $\alpha + \beta + \gamma = \pi$. Show that then

$$\sin^2(\alpha) = \sin^2(\beta) + \sin^2(\gamma) - 2\sin(\beta)\sin(\gamma)\cos(\alpha).$$

Exercise 4.10 In Theorem 2 of Sect. 4.13, we saw that if $\triangle ABC$ is cut by a line which cuts the side (possibly produced) AB in the point D, BC in E and CA in F, then

$$\frac{AD}{DB}\frac{BE}{EC}\frac{CF}{FA} = 1$$

where AD, DB, etc., are the lengths of the line segments. Let the positive direction be from left to right. Show that with this convention the relation should be written as

$$\frac{AD}{DB}\frac{BE}{EC}\frac{CF}{FA} = -1$$

Prove that this condition is necessary and sufficient for D, E and F to be collinear, that is, to lie on the same line.

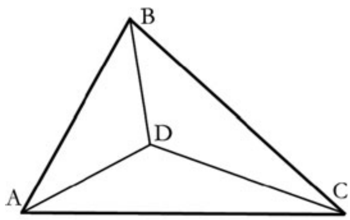
Exercise 4.11 Let ABC be a triangle, and form the midnormals n_1, n_2 and n_3 on the sides AB, BC and CA, respectively. Show that they meet in one point, the center of the circumscribed circle or the *circumcircle* of $\triangle ABC$. Its radius is referred to as the *circumradius*, the center as the *circumcenter*.

Exercise 4.12 Show that the circumradius may be expressed as

$$R = \frac{abc}{4S}$$

where S is the area of $\triangle ABC$ and a, b, c denote the length of the sides.

Exercise 4.13 In Proposition 4 of Book IV of the Elements, Euclid inscribes a circle inside an arbitrary triangle ABC.



He does so by showing that the bisectors of all three angles $A = 2\alpha = \angle CAB$, $B = 2\beta = \angle ABC$ and $C = 2\gamma = \angle BCA$ meet at a point which is equidistant from the three sides of the triangle. r is then the radius of the inscribed triangle, or the *inradius*. Use modern trigonometry to show that the inradius is $r = 4R\sin(\alpha)$ $\sin(\beta)\sin(\gamma)$, where R is the circumradius of the triangle.

Exercise 4.14 Let a be the length of the side opposite to the vertex A of the triangle in Exercise 4.13, define b, c similarly. Let $s = \frac{a+b+c}{2}$. Show the alternative formula for the inradius

$$r = \sqrt{\frac{(s-a)(s-b)(s-c)}{s}}$$

Exercise 4.15 With notations as in Exercise 4.14, prove "Heron's Formula" for the area S of $\triangle ABC$:

$$S = \sqrt{s(s-a)(s-b)(s-c)}$$

There are more problems of this type at the end of Chap. 6.

Chapter 5
Arabic Mathematics and Geometry

The Arabs made important contributions to science and culture. In particular to mathematics and geometry, the subject of this book. Some authors refer to *Islamic* mathematics, or *Hindu* mathematics, for that matter. Here we shall simply refer to "Arabic mathematics", as it was created within the Arabic culture and written in the Arabic language.[1]

5.1 The Arab Expansion

Traditionally *Arabia* designates the peninsula between the Red Sea and the Persian Gulf. From prehistoric times the people living there, the *arabs*, was divided in many different groups, from small nomadic tribes to impressive city states. The arabs were dominated by powerful neighbors, and at the beginning of the seventh century, the 600s, Arabia was divided an powerless. But this was to change dramatically.

We now quote from the biography of Muhammad in [61]:

> *Muhammad ibn Abd Allah was born around 570 in Mecca and died June 8, 632 in Medina. He was the founder of Islam and is regarded by Muslims as the last messenger and prophet of God, or Allah. Muslims do not believe that he was the creator of a new religion, but the restorer of the original, uncorrupted monotheistic faith of Adam, Abraham and others. They see him as the last and the greatest in a series of prophets of Islam.*
>
> *Sources on Muhammad's life concur that he was born ca. 570 in the city of Mecca in Arabia. He was orphaned at a young age and was brought up by his uncle, later worked mostly as a merchant, and was married by age 26. At some point, discontented with life in Mecca, he retreated to a cave in the surrounding mountains for meditation and reflection. According to Islamic tradition, it was here at age 40, in the month of Ramadan, where he received his first revelation from God. Three years after this event, Muhammad started preaching these revelations publicly, proclaiming that* God is One*, that complete surrender to Him is the only religion acceptable to God, and that he was a prophet and messenger of God, in the same vein as Adam, Noah, Abraham, Moses, David, Jesus, and other prophets.*
>
> *Muhammad gained few followers early on, and was largely met with hostility from the tribes of Mecca; he was treated harshly and so were his followers. To escape persecution,*

[1] The main sources for this chapter are [34, 38, 49, 60, 61].

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_5,

Muhammad and his followers migrated to Yathrib (Medina) in the year 622. This historic event, the Hijra, marks the beginning of the Islamic calendar. In Medina, Muhammad managed to unite the conflicting tribes, and after 8 years of fighting with the Meccan tribes, his followers, who by then had grown to ten thousand, conquered Mecca. In 632, on returning to Medina, Muhammad fell ill and died. By the time of his death, most of Arabia had converted to Islam.

The revelations, which Muhammad reported receiving till his death, form the verses of the Quran, regarded by Muslims as the word of God, *around which the religion is based. Besides the Quran, Muhammad's life (sira) and traditions (sunnah) are also upheld by Muslims.*

The succession of Muhammad carried with it controversies, religious as well as political. The strongest position was that the *Calif* should be selected among the best men of the Prophet's own family. The first four Caliphs held court in Medina. The first three were Abu Bekr (632–634), Omar Ibn al-Khattab (634–644) and Othman Ibn Affan (644–656). Al-Khattab made great conquests, see the map on Fig. 5.1. He was succeeded by *Othman*, who belonged to a powerful family in Mecca, the *Umayyads*. He appointed members of his family to important positions, among them his relative *Muawija*, who became governor in Syria based in i Damascus. Although these civil servants attended to their duties irreproachably enough, the dissatisfaction with Othman grew among the populace. When he was murdered in 656, he was succeeded by *Ali* (656–661), Muhammad's cousin and son-in-law, married to Muhammad's daughter *Fatimah*. Ali's supporters maintained that the Prophet himself had designated Ali as his successor, and that consequently the three previous choices were illegal. But Ali's time as Calif was marred by the bitter controversy with the omayyedes. In the end Ali was murdered, and his *Hussein* tried to succeed

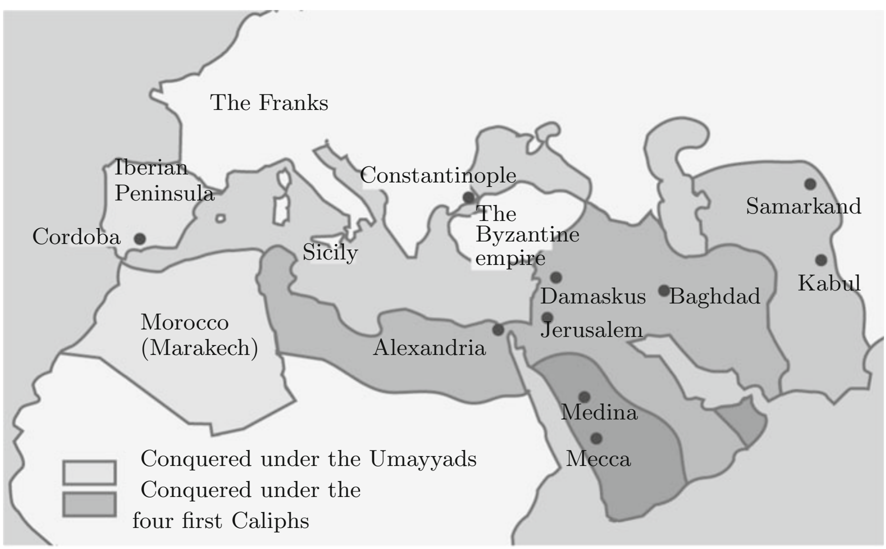


Fig. 5.1 Map showing the conquests under Muhammad's successors

him. He was, however, unable to prevail over Muawija in Damascus, who became the new Caliph. Thus the Umayyads were back in power, and the capital was moved to Damascus. But Hussein was still powerful, and in the ensuing struggle he was finally killed in the old city of *Hira*, between Kufa, now Najaf, and Kerbala. The tale of Hussein's martyrdom together with his brother *al-Abbas* is still important in Shia Islam, the largest minority denomination (10–20%) after Sunni Islam. He is buried at the Imam Hussein-mosque in Kerbala.

After Muhammad's death in 632 the expansion continued, now directed against the Persian and the Byzantine empires. Damascus was conquered in 636, Jerusalem in 638. In 616 Alexandria had been conquered by the Persians. The Neoplatonic Greek scholars who had to flee when the academy in Athens was closed, had been well received in Persia, and the old Persian empire now stood at the top of its might and flourished culturally. Zoroastrianism was the state religion, opposing Islam as well as Christianity. But already in 630 muslim Arabs started their conquest of Persia, and in 640 Alexandria was taken by the forces of Calif Omar, led by the general *Amru*. Alexandria was taken after a siege which lasted for 14 months, during which time no help was forthcoming from the emperor in Constantinople. He had, apparently, enough with his own problems.

It is told that general Amru sent a message to the Caliph reporting that the conquered city had 4,000 palaces, 4,000 baths, 12,000 dealers in fresh oil, 12,000 gardeners, 40,000 Jews who pay tribute, 400 theaters or places of amusement. In 645 a Byzantine fleet recaptured the city, however it fell again the following year, and this time for good. According to this the city had again a sizable Jewish population, following the pogroms of 415. But the report contains so round and regular numbers that one probably should view it with skepticism.

Another story is the following: In Alexandria lived and worked a most learned man by the name *John the Grammarian* at the time when Amru captured the city. He requested from the general permission to take over all the books in the famous *library*. Amru forwarded the request to the Caliph Omar, who returned the following answer: "If these books write the same as the Koran, then they are superfluous and should therefore be burned. But if they write something different from the Koran, then they contain heresy, and certainly should be burned in this case as well." Consequently the books were handed over to the 4,000 baths as fuel for heating the water. This kept the boilers heated for 6 months!

As this dubious story is repeated in some western texts on the history of mathematics, it is worthwhile to make a rough estimate of the facts entering into it. First of all, the Arabs were very competent merchants, and were without doubt well aware of the market value of the priceless books in the Alexandrian library. There are accounts of spoils of war being sold to cover the expenses of the campaign, and not very long after these events Arab scholars travelled extensively to collect ancient literature. For that reason alone the story is blatantly far fetched. In addition to this, we can estimate how much fuel which is necessary to heat all these boilers: If heating one boiler for one day requires five sacs of firewood (by present day standard), each weighing 20 kg, then one boiler needs 100 kg of firewood for one day. For 180 days it requires 18,000 kg. All 4,000 baths therefore will require 72,000,000 kg. If one book weighs, on the average, $\frac{1}{4}$ kg, then this load of wood corresponds to a staggering 288 million books. This is assuming that the books, which would have been on parchment, would yield the same amount of energy as wood. See Exercise 5.1 for an alternative computation.

As for John the Grammarian, there did indeed live a learned Byzantine by that name in Alexandria, but about 200 years earlier. He is supposed to have been the first who suggested an experiment to check the claim by Aristotle that a heavy body will fall faster than a lighter one.

The Umayyads continued to rule until 750, and conquered the territories shown on Fig. 5.1. This dynasty was deposed in 750, when a descendant of Mohammad's uncle *Abbas* became Caliph. The Abbasids moved the capital to *Baghdad*, which at this time bore the name *Dares-selam*, "The house of Peace."

Now the Umayyads where persecuted, and many killed. But one managed to escape to Spain. His name was *Abdurrahman.*

The fifth Umayyad Caliph is Harun al-Raschid, who ruled 786–809. He is the Caliph of *The Thousand and One Nights.* Even if his rule stands as a high point according to Western opinion, the base for decline and later fall was laid. His son, *Mamun* who ruled 809–833, played a great role as a patron of science and culture. Under his rule the ancient Greek texts were translated into the Arabic. But political decline was not far off. One might say that the old history from the Roman empire repeated itself, but this time in Baghdad. The refined and cultured citizen of Baghdad were no longer fit for military service and war. Consequently the ruling Abbasids became more and more dependent on mercenary soldiers, mostly from the *Turks*. Under the Caliph *Muqatir* (908–932) the commander of his guard, who was referred to as *Amir al-umara*, or the *Amir of the Amirs*, ruler of the rulers, was in reality the one who held absolute power. The worldly power in Baghdad thus lay in the hands the *Bujides*, a powerful clan of Turks.

Arabs continued to hold their leading position until the middle of the eleventh century, when the Seljuq Empire was established. This was a medieval Sunni Muslim empire established by the Qynyq branch of Oghuz Turks, their advance marked the beginning of Turkic power in the Middle East. The empire was founded by Tugrul Beg in 1037.

The Turk chieftain *Seljuq* had migrated together with his tribal allies in the 10th century from Central Asia into present day Uzbekistan. The first mention of if the Turks is in a Chinese text from around 500. His two grandsons *Toghril Beg* and *Tachyr Beg* advanced further into the area, and founded an empire which was to include Mesopotamia, Syria, Palestine and a large part of Persia. Around 1040 they had conquered the entire northeastern part of Persia. Toghril Beg proclaimed himself sultan at Nishapur in 1038 made his entry in Baghdad in 1055.

Thus was a time of upheaval, with political and religious strife. He now held power in Baghdad and was to have married the daughter of the Caliph. But he died before the marriage had taken place. His brother Tachyr had died already in 1059. The two next Seljuq sultans were Tachyr's son and grandson, they made further conquests and gathered a large empire.

In Baghdad the Abbasid caliphs were powerless against the Turks, who however found it convenient to maintain the Caliph in Baghdad as the spiritual head.

The Byzantine empire also eventually lost most of its power in Asia Minor to the Turks. At this time *Konya* or *Iconium* in Anatolia became the capital of a powerful Seljuq sultan, but from the end of the thirteenth century his empire came under pressure from new groups of Turks who pressed forward from the east. One of these groups was led by a chieftain named *Osman*, and his son *Orhan* founded a new empire which later was to expand greatly. This is what we know as the *Turk Osman* or *Ottoman* Empire, which existed until general *Mustafa Kemal,* later given the name of honor *Atatürk,* seized power and founded modern Turkey in 1922.

When the Mongols invaded Baghdad in 1258, the last Caliph of the Abbasid dynasty was killed. But some of the Abbasids managed to escape to Egypt, where the Sultan proclaimed one of them to be the Caliph. Thus this dynasty continued for some more time, but in 1517 the Ottoman Sultan *Selim 1.* conquered Egypt, and then he brought the Abbasid Caliph to Constantinople, and there he compelled his prisoner to transfer the dignity of the Caliphate to himself. So from then on the Turk Sultans also carried the title of Caliph.

As mentioned above, the Umayyad *Abdurrahman* escaped to Spain when the Abbasids seized power in Damascus. At that time Spain was a remote Arab province, which had in reality been independent of the Caliph in Damascus. But now, as this Caliph had been ousted, Abdurrahman succeeded in being proclaimed the *Emir* in Cordoba. At the time of his death in 787 he had gathered practically all of Muslim Spain under his rule. But it was *Abdurrahman 3.* who finally decided that the time had come to assume the title of Caliph, *Ruler of the Faithful.* The word *Emir* just means "Ruler." Cordoba was now a cultural center in the Arab world.

Abdurrahman 3. was succeeded by his son *Hakam 2.* (961–976), who upheld the kingdom his father had gathered. He worked energetic to enlighten the populace, and make the capital Cordoba a true center for learning and science. It has been said about him that he is the most learned monarch who ever ruled. He collected a library of 4,00,000 books, and it was said that he had personally read all of them. Cordoba remained the center of science and culture until 1031, when the last Caliph of this dynasty was murdered. Then Sevilla became the most important islamic center in Spain.

Another dynasty of Caliphs descended from *Fatima*, Muhammad's daughter. These Caliphs ruled in North Africa during 909–1171, thereafter in Egypt, where they founded the city of *Cairo*. Later they also gained partial control in Syria.

5.2 Arab Science and Culture

The Arabs continued Greek and Persian science and philosophy. The first conquests provided access to this cultural heritage, which they preserved for the future and developed and expanded. The ancient classical works were translated into Arabic, Arab merchants and travellers came into contact with India and China, and their

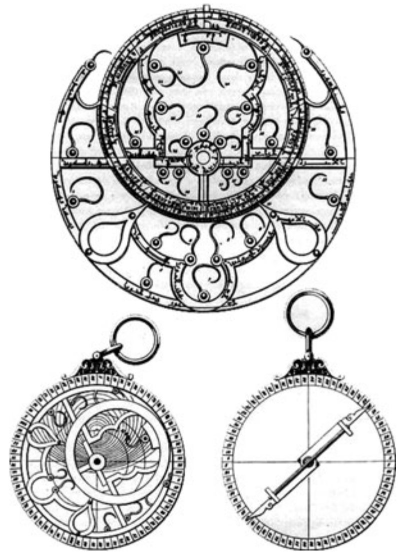


Fig. 5.2 An Arab astrolabium from Persia, 1208

scientists were highly regarded and respected. Without this base in Arabic science and culture, the reawakening of Europe would have been very much more difficult than it was, if at all possible.

Arabic astronomy was foremost in the world. Arab astronomers constructed and used astrolabes with great skill.

The astrolabe was invented in Hellas and is commonly attributed to Hipparchus (Fig. 5.2).[2] The astrolabe was a gadget for working out several different kinds of problems in spherical astronomy. Theon of Alexandria wrote a detailed treatise on the astrolabe, and some believe that Ptolemy used an astrolabe to make the astronomical observations in his treatise on astrology, Tetrabiblos.

Brass astrolabes were developed in the Islamic world, chiefly as an aid to navigation and to finding the direction of Mecca. The first person credited with building the astrolabe in the Islamic world is the eighth century Persian mathematician al-Fazari. The mathematical background was established by the Arab astronomer al-Battani. In the Islamic world, astrolabes were used to find the times of sunrise and the rising of fixed stars, to help schedule morning prayers. In the tenth century, al-Sufi first described over 1,000 different uses of an astrolabe, in areas as diverse as astronomy, astrology, horoscopes, navigation, surveying, timekeeping, prayer.

Arzachel (al-Zarqali) of al-Andalus constructed the first universal astrolabe instrument which, unlike its predecessors, did not depend on the latitude of the observer, and could be used from anywhere on the Earth. This instrument became known in Europe as the "Saphaea." The astrolabe was introduced to other parts of Europe via Islamic Spain in the eleventh century.

[2] Following [61] on the astrolabe.

The Arabic mathematicians advanced far ahead in the development of the numbers, computation and algebra. Our present day decimal system with our numerals comes via the Arabs from India and China. The concept and symbol for *zero* was also passed on to us by the Arabs. The words *algebra* and *algorithm* was not exactly *invented* by the Arabs, but comes from them, nevertheless. Indeed, they are derived from the great Arab mathematician al-Khwarizmi, which we tell more about in Sect. 5.4.

They made original discoveries in algebra and *number theory*. They studied prime numbers and algebraic equations, and continued the Greek tradition of geometry. Their insights into *ratios* enabled them to develop a precise theory for notes, chords, harmony and discords, and to make well tempered musical instruments.

Baghdad is located in the middle of the fertile Mesopotamia, and not far from the ruins of ancient Babylon. Here old trade routes come together, and the cultural roots to the past still existed at this time. Now a cultural and scientific flourishing took place in this region, so rich in tradition, where mathematics, science and culture had made so tremendous advances before.

It is a characteristic feature of all mathematics of high quality that it proceeds beyond the immediate needs dictated by practical matters of the time. Thus for instance, the ancient Pythagoreans saw mathematics as a form of spiritual cleansing, necessary for the soul in order to be able to break out of the eternal circulation of repeated reincarnations. In Arabic mathematics, by some referred to as *Islamic mathematics*, we find a similar spiritual component. Typically the mathematician would commence his writing with a pious praise of God, of *Allah*. Then during the work Gods help was called on and invoked.

But mathematics was pursued by believers of other religions as well, Zoroastrians, Jews, Christians and others. However, it is all written in the *Arabic language.* For that reason we shall call it *Arabic mathematics*, rather than Islamic or Arab mathematics.

The Arabs are, as we have seen, blamed by some for having burned the Library and the Museum in Alexandria. But this is undoubtedly a great exaggeration. It is likely that very little remained of the Library or the Ademy at the date of the conquest, and whatever there was would have been carefully collected by the economy-minded Arabs. On the contrary the Arabs should be credited with having saved ancient mathematical texts for posterity. Several fundamental texts, among them Euclid's Elements, were actually reintroduced into the Christian world as translations from the Arabic to Latin, at a much later time.

Indeed, as the Arabs expanded into southern Spain, westerners have generally viewed this as a grave threat to the civilized world. But the Arabian Muslims founded Academies for mathematics, science and medicine there. When Europe later began its reawakening, Christian scholars travelled south, and disguised as Arabs managed to attend these centers of culture and learning.

5.3 The Founder of the House of Wisdom in Baghdad

Harun al Rashid became the fifth Caliph of the Abbasid dynasty in 786. His capital city was Baghdad, where he ruled over an empire from the Mediterranean to India. At his court culture and science flourished. When he died in 809 he was succeeded by his brother *al Mamun* after a short struggle. Al-Mamun continued in his father's tradition and founded an academy, *the House of Wisdom* in Baghdad. Here ancient Greek works were collected and translated. Thus he started the first major library since the one in Alexandria. He also built observatories in which Arab astronomers made significant contributions to astronomical knowledge.

Al Rashid had good contact with *Charles the Great,* Carolus Magnus or Charlemagne, who ruled over the Franks. In 800 Carolus was crowned *Imperator Augustus* by Pope Leo III in an attempt to revive the Roman Empire. He died in 814, however. The year after this crowning, al-Rashid sent him gifts, including an elephant and a water clock, both of which attracted considerable attention at his court.

Even as civilization flourished in Baghdad, life was harsh, and Harun al-Rashid died in 809 during a campaign to quell an uprising in Tus.

Harun had two sons, the eldest being *al-Amin* and the younger *al-Mamun*. In the power-struggle between these two, al-Amin was defeated in a battle in 813 and killed. Al-Mamun then became Caliph. He continued his fathers commitment to science and culture, and it was he who established the *House of Wisdom*, Bayt al-Hikma, in Baghdad. This academy continued the tradition after the academies in Athens and Alexandria, it existed for 200 years and became very important for mathematics.

Al-Mamun continued extending the library of Baghdad, it eventually was to play a similar role to the one in Alexandria. He also build astronomical observatories, where arabic astronomers could continue to extend the knowledge of earlier times on the stars and the planets.

Old manuscripts were collected to the library from several libraries throughout the Middle East, where the scholars from Athens and Alexandria had sought refuge from persecution. Among these manuscripts were several classical Greek books, they were now translated into the Arabic. At the House of Wisdom scientific texts from India, and presumably also China, were studied and translated. It is also fair to assume that this rich scientific environment also absorbed the impulses which still were present from the times when the rich Mesopotamian mathematics flourished.

As is pointed out by many historians of science, in this Islamic culture secular knowledge was not perceived as being alien to pious faith, but rather conceived as one of the paths leading to sacred wisdom.

According to some historians of mathematics, the definitive version of the history of Arabic mathematics has yet to be written. In fact, it is speculated that a significant number of ancient Arabic mathematical manuscripts still lie unstudied in collections unknown to researchers. Even though a growing number have been studied by Arabic speaking scholars and translated from the Arabic, much remains to be done. Political conflicts also contribute to the difficulties.

5.4 Abu Ja'far Muhammad Ibn Musa Al Khwarizmi

The first known Arabic geometry text is a section of al-Khwarizmi's algebra book mentioned in the previous section. In it, he gives the approximation $\pi \approx 3\frac{1}{7}$, the simple fraction used by Archimedes. But he also mentions the less accurate $\pi \approx \sqrt{10}$. He gives a third and quite accurate approximation $\pi \approx \frac{62{,}832}{20{,}000} = 3.1416$, which we first know from the Indian mathematician and astronomer *Aryabhata the Elder*, 476–550 A.D.

Al-Khwarizmi and his colleagues worked at the House of Wisdom in Baghdad. They started a magnificent scientific tradition which lasted well into the fifteenth century, when the tradition was continued by the Europeans.

Today many historians of mathematics realize that the Arabs have not been given their due credit for the significant contributions they made to mathematics. The Arabs had been seen merely as preservers, commentators and "messengers", who delivered ancient Greek mathematics to the proper heirs so to speak, namely the Europeans. But today the general feeling is that this view is unjustified, since Arabic mathematicians made very significant and original contributions. It is no accident that the word *algebra* is derived from the title of one of *al-Khwarizmi's* fundamental books, *Al-kitab al-muhtasar fi hisab al-jabr wa-l-al-muqabala*, abbreviated to *Hisab al-jabr wa-l-al-muqabala*. This is the first book to be written on *algebra as such*. The title means something like *The condensed book on arithmetic by "al-jabr" and "al-muqabala",* the two Arabic words meaning, respectively, "setting together" and "balancing." The first word is the origin of our *algebra.* It is told that in southern Spain barbers used to be called *algebraists*, presumably because their duties included performing simple surgical procedures such as reducing a fracture.

He also computes the area of a rhombus when the two diagonals are given, by *"multiplying one by half of the other."* A rhombus is a quadrilateral where all sides are of equal length (Fig. 5.3). Thus the diagonals will subdivide it into four congruent or mirror images right triangles where the hypotenuse is a side of the rhombus and the two other legs are half of the two diagonals, respectively.

The area of a triangle where the sides are given as 13, 14 and 15 is computed as follows: With our notations we denote the triangle by $\triangle ABC$, as shown in Fig. 5.4.

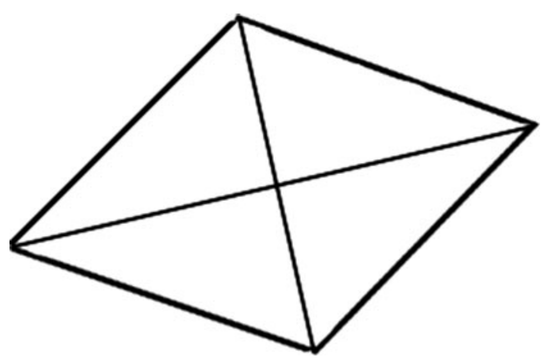
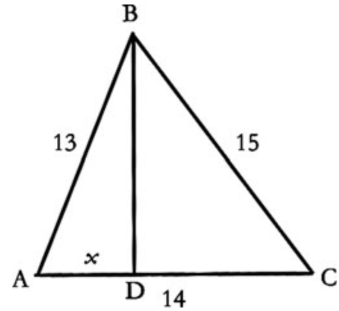


Fig. 5.3 A rhombus

Fig. 5.4 A triangle



Drop the normal to AC through B. Its foot is D, let the length of BD be h and the length of AD be x. By the Pythagorean theorem we have $13^2 = x^2 + h^2$ and $15^2 = (14 - x)^2 + h^2$. Substituting the former in the latter yields $15^2 = 14^2 - 28x + 13^2$ thus $13^2 - (15 + 14)(15 - 14) = 28x$, i.e., $28x = 140$ so $x = 5$. Thus $h^2 = 13^2 - 5^2 = 144$ so $h = 12$. Thus we finally have the area as $\frac{1}{2} \cdot 12 \cdot 14 = 84$.

Al Khwarizmi presents his algebra in the form of explicit *geometric algebra*, advancing the science significantly from the grand achievements of the Greek. The following assertion, quoted in [60], 263–77, is a relevant point of view:

> Al-Khwarizmi's algebra is regarded as the foundation and cornerstone of the sciences. In a sense, al-Khwarizmi is more entitled to be called "the father of algebra" than Diophantus because al-Khwarizmi is the first to teach algebra in an elementary form and for its own sake, Diophantus is primarily concerned with the theory of numbers.

Al-Khwarizmi's ancestors came from the province of Khwarizm, to the south of the Aral Sea. Now this region forms part of Uzbekistan, Turkmenistan and Persia, it has a rich and interesting history, and has been known under many different names. One of the names is *Khorasan*, the present name of the largest province of Persia. See the map on Fig. 5.5.

He was one of the first scholars to work at the House of Wisdom. According to Katz [38], al Khwarizmi was also active as *astrologer*, and he cast the Caliph's horoscope, assuring him of a long life, eh should, according to the horoscope, live for another 50 years! However, he died after just ten days.

Al-Khwarizmi was born in Baghdad, in 780 (Fig. 5.6). This was 3 years before the founder of the House of Wisdom, Harun al-Rashid, assumed the position of Caliph. Al Khwarizmi died around 850.

In his work at The House of Wisdom al-Khwarizmi translated several old Greek mathematical texts. Above all he engaged in highly original and path-breaking research in geometry, algebra and astronomy. Al-Khwarizmi's most famous work bears the title *Al-kitab al-muhtasar fi hisab al-jabr wa-l-al-muqabala*, often abbreviated to *Hisab al-jabr wa-l-al-muqabala*. The full title may be translated as *The Complete Book on Calculation by Completion and Balancing*.

This was the first algebra book in history, in fact the very title gave rise to the name of this mathematical discipline, as the word *al-jabr* turned into our *algebra*. Following the explanation given in [38], Al-jabr actually stands for *putting together*

Fig. 5.5 Map of the region important for Arabic mathematics

or *restoring*, as when a quantity is subtracted for one side of an equality and added to the other. Moreover, al-muqabala stands for comparing by reducing two quantities by subtracting the same positive number from both. According to some accounts the term *algebraist* was used in southern Spain for a barber, as barbers undertook simple surgery, including the procedure of reducing broken bones.

Mathematically we may illustrate the employment of putting together and balancing in the following solution of an equation, phrased in our modern notation:

The step from

$$8x + 5 = 9 - 2x$$

to

$$10x + 5 = 9$$

is *al-jabr*, while the step to

$$10x = 4$$

is *al-muqabala.*

Al-Khwarizmi has also given rise to another household word of today, namely the term *algorithm*. He describes the decimal (base ten) numeral system, and gives

Fig. 5.6 Al Khwarizmi. Drawing by the author

practical descriptions of the arithmetic procedures, proving their correctness. He gave the proofs in a geometric form.

Some think that he build on Book 2 of Euclid's Elements, while others doubt that he knew this Greek text. According to [55, p. 69] or [38, p. 244], his proofs are not in the Greek tradition, but rather in the Babylonian. On the other hand, as noted in [60], in al-Khwarizmi's youth al Hajjaj had translated Euclid's Elements into Arabic, and al Hajjaj also worked at the House of Wisdom. But Al-Khwarizmi does not use definitions, axioms, postulates, and has no demonstrations of Euclidean type.[3]

[3] One of the first translators of the Harranian school of mathematic and astronomy is known by his arabized name as al-Hajjaj ibn Yusuf ibn Matar (786–833). He is credited with having made the first translation of the Elements and one of the first of Ptolemy's astronomical work. He must not be confused with the earlier al Hajjaj ibn Yusuf ath Thaqafi, (661–714), who was the governor of

As argued in [38], al-Khwarizmi's descriptions of the algebraic procedures closely follow the patterns used by the ancient Babylonian scribes.

Indeed, the procedures of al-Khwarizmi are similar to geometric procedures described on the Babylonian tablets. He also provides detailed classifications of the problems he treats, and according to Rashed in [49], the book represents the culmination of earlier work, as well as the start of something radically new.

The quantities he works with are of the following three types: squares, roots and numbers. Or, as we would say: x^2, *x where x is an unknown quantity, and numbers.*[4] One of the problems may be formulated as follows: *A square is equal to forty roots, deducted by four squares.* By *al-jabr* this is transformed into: *Five squares is equal to forty roots.* Thus the problem is transformed into the first problem on his list: *Squares equal to roots.* other words, a certain number of squares is equal to a certain number of roots.

1. $ax^2 = bx$.

In a similar manner he transforms all problems with squares, roots and numbers into one of the problems on the list which is completed below, using our modern algebraic notation.

2. Squares equal to a number, $ax^2 = c$.
3. Roots equal to a number, $bx = c$.
4. Squares and roots equal to a number, $ax^2 + bx = c$.
5. Squares and number equal to roots, $ax^2 + c = bx$.
6. Roots and number equal to squares, $bx + c = ax^2$.

The important feature is that al-Khwarizmi does not just seek the solutions of specific equations, instead he makes the *class of quadratic equations as such* an object of study for his science. He proceeds in this systematic study, and observes that by division or by multiplication with numbers, he may reduce to the case when $a = 1$, as we would express it. Al-Khwarizmi thus gives a systematic procedure by which any problem (equation) involving squares, roots and numbers, may be reduced to one of the *canonical forms* listed below:

(1) Square equal to roots, $x^2 = bx$.
(2) Square equal to a number, $x^2 = c$.
(3) Root equal to a number, $x = c$.
(4) Square and roots equal to a number, $x^2 + bx = c$.
(5) Square and number equal to roots, $x^2 + c = bx$.
(6) Square equal to roots and a number, $x^2 = cx + c$.

Al-Khwarizmis solutions of the first three equations is quite straightforward, and follows the lines established by the Babylonians and the Greeks. For (3) there

Iraq during the reigns of abd al Malik ibn Marwan and al-Walid I of the Umayyad dynasty, and was an able though apparently rather ruthless general and military man.

[4] See Rashed [49, p. 10] for more on the basic concepts in al-Khwarizmi's algebra-book.

remains nothing to do, (1) yields $x = b$ of course, and for (2) a square root has to be extracted.

Al-Khwarizmi solved problems of type (4), (5) and (6) by *geometric algebra*. Following [49] we now reproduce this description.We first look at his treatment of an example of problems of type (4):

$$x^2 + 10x = 39:$$

> What is the square which when increased by 10 of its own roots becomes 39? The rule for this problem is that you divide the roots in two halves. In this problem that is 5, which when multiplied with itself makes 25. Add this to the 39, then you have 64. The root of this is 8, you subtract the half of the roots, namely 5, and 3 remains. This is the root of the square you seek, its square is 9, which is the answer.

This is a completely algebraic procedure, but phrased in words rather than in algebraic notation, se we would do today. We would perform this by *completing the square* as follows:

$$\begin{gathered} x^2 + 10x = 39 \\ x^2 + 10x + 5^2 = 39 + 25 = 64 \\ (x + 5)^2 = 64 \\ x + 5 = 8 \\ x = 8 - 5 = 3 \\ x^2 = 3^2 = 9. \end{gathered}$$

Al-Khwarizmi gives the following geometric proof for his algebraic solution[5]:

On Fig. 5.7 we have shaded the area which corresponds to the original square and the ten roots. This L-shaped area has the form of a *gnomon*. The Gnomon is the part of a sundial which casts the shadow. The gnomon is an ancient geometric figure known from Babylonia and China, it is said to have been introduced in Greece by Anaximander from Miletus, the student of Thales who became his successor as leader of the Milesian School.

The gnomon may be completed to a square by adding a square in the manner indicated: The sides of the added square must be half the *number* of the roots to be added. Even if he proceeds in a geometric manner, the procedure is really algebraic: In principle there is no difference between numbers, lengths or areas.

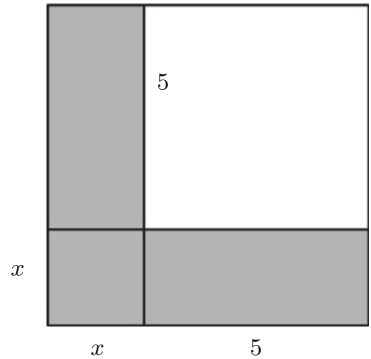
He explains that the word *root* should not be understood as the side of a square, but rather as[6] *anything composed of units which can be multiplied with itself, or any number greater than unity multiplied by itself or that which is found to be diminished below unity when multiplied by itself.*

---

[5] According to [38]. Reference [60] gives a somewhat different explanation, but based on the same principles. According to [49, p. 13], al-Khwarizmi some times gives more than one explanation.

[6] Following [38, p. 246].

Fig. 5.7 Geometric solution of a quadratic equation



With al-Khwarizmi algebra has taken a significant step from the implicit geometric algebra practiced by the Greek and before them the Babylonians towards "algebra" in a modern meaning.

See the recent book [15] for details on Babylonian origins of Greek geometry.

With modern notation al-Khwarizmi solves the equation

$$x^2 + bx = c \text{ as } x = \sqrt{\left(\frac{b}{2}\right)^2 + c} - \frac{b}{2},$$

where the formula is expressed in words, and the proof of correctness is presented in the form of a geometric figure.

Al-Khwarizmi's solution of equations of type (5) is interesting, here he takes a big step forward from Babylonian and Greek mathematics. With our symbols the task is to solve an equation of the type

$$x^2 + c = bx,$$

and he gives it in the form which we would write as follows:

$$x = \frac{b}{2} \pm \sqrt{\left(\frac{b}{2}\right)^2 - c}.$$

Of course no such formula appears in al-Khwarizmi's texts, but he writes that[7] we get *two solutions by subtracting or adding half of the number of roots (i.e.,* $\frac{b}{2}$*) to the root of what you get when this number is multiplies with itself (i.e.,* $(\frac{b}{2})^2$*), and subtracted the number which is to be added to the square (i.e.,* c*). If half the number of roots multiplied with itself is less then the number which is to be added to the square, then the problem has no solution. But if the product is equal to the number,*

[7] Following [38].

then the answer is half of the number of roots, without the need to add or subtract anything.

Al-Khwarizmi also wrote a book where he introduced what we today call the *indo-arabic numeral system*, that is the use of the digits 1, 2, 3, 4, 5, 6, 7, 8, 9 and 0, representing integers to the base 10.

5.5 Ibn Qurra and al-Battani

Ibn Qurra was born in Harran in 826 and died in Baghdad in 901. His native language, the Syriac, was an Eastern Aramaic language once spoken in much of Mesopotamia. Like the Arabs, Syriac speaking scholars have been viewed as merely custodians Greek science and culture, but they made in fact significant original contributions. Thabit Ibn Qurra made important mathematical discoveries such in algebra and geometry as well as astronomy. In astronomy Thabit was one of the first reformers of the Ptolemaic system, and in mechanics he was a founder of statics.

As a youth he so impressed a visitor from the Baghdad House of Wisdom that he was brought there as a student. There he studied mathematics as well as medicine, and since he knew Greek very well, he participated in the translation of the ancient texts. In Sect. 4.4, we have seen the verging-construction of the regular heptagon, which we know through the translation of Ibn Qurra. He also gives a generalization of the Pythagorean theorem, which is first known from Pappus, Sect. 4.18. This result was used by the astronomer al-Battani, 850–929, who came from the same region as Ibn Qurra. Early Arabic trigonometry worked with the *cord* of an angle like Ptolemy, or with the more practical *half cord* of the *double* angle, in other words essentially with what is known today as the *sine*.

Al-Battani was also born in Harran, in 850.[8] Al-Battani made his remarkably accurate astronomical observations. He has been described as a famous observer and geometer. In his astronomical work he gave his own observations of the sun, moon, and the planets, more accurately than what is found in Ptolemy's *Almagest*.

Al-Battani's most important book is *Kitab al-Zij*. It begins with the necessary mathematical tools, such as the sexagesimal numeral system and the trigonometric functions. Then he gives the date from his own observations, and subsequently goes on to treat different astronomical problems in line with the Almagest. The motions of the sun, moon and five planets are discussed, the theory given is that of Ptolemy.

After giving results to allow data given for one era to be converted to another era, al-Battani then explains how his tables are to be read. Chapters 49–55 cover problems in astrology, while Chap. 56 discusses the construction of a sundial and the final chapter discusses the construction of a number of astronomical instruments.

Al-Battani used trigonometrical methods, which represents something new. For example, he gives trigonometric formulae for right angled triangles such as (in modern notation):

[8] We follow [60].

$$b\sin(A) = a\sin(B).$$

with a the length of the side opposite the angle A, b the length of the side opposite the angle B.

Thus, in his trigonometry al-Battani worked with what we call the *cosine*, the sine of the *complementary angle.* Al-Battani's work became important to later European astronomers like Copernicus, Brahe, Kepler and Galileo. In fact, Kepler got the idea to the *Law of Cosines* from al-Battani. In our modern terminology the laws of sine and cosine was explained in Sect. 4.15.

5.6 Muhammad Abu al Wafa al-Buzjani

Muhammad Abu'l-Wafa Al Buzjani was born 940 in Buzjan in the Khorasan region (of present day Persia), and died 998 in Baghdad. In 945 Ahmad Buyeh conquered Baghdad, which had been the capital of the Abbasid dynasty since Abbas ibn Abd al Muttalib, descendent of Muhammad's youngest uncle assumed power in 750 and overthrew the Umayyad caliphs from all but Al Andalus where an Umayyad dynasty remained in power. The most important capital in the Arab world then was moved from Damascus to Baghdad.

The new Buyid Islamic dynasty ruled in western Persia and Iraq from 945 to 1055. Adud ad-Dawlah ruled from 949 to 983. He supported science, especially astronomy and mathematics, and in 959 Abul Wafa joined his group of mathematicians working at his court.

When Sharaf ad-Dawlah succeeded his father in 983, he continued to support mathematics and astronomy. An astronomical observatory was built in the palace garden and opened in 988, its director became *al Quhi*, see Sect. 5.7. At the observatory he was to collaborate with Abul Wafa, among others. But when the caliph Sharaf ad Dawlah died a year later, unrest broke out and the observatory was closed.

Abul Wafa translated and wrote commentaries on the works of Euclid, Diophantus and al Khwarizmi. One of his books has a title which may be translated to *Book on what is necessary from the science of arithmetic for scribes and businessmen*.

At this time two kinds of arithmetic texts circulated, one with the Indian numeric symbols and the other with the numbers expressed in words and the calculations done mentally using finger reckoning. Abul Wafas text was of the latter type although Abul Wafa certainly was an expert on the Indian numerals, but finger reckoning was the system used in business at that time. The work is quite comprehensive, from the point of view of geometry it is interesting to note that Part III of altogether seven parts is concerned with *Mensuration*, finding area of figures, volume of solids and distances.

It should also be noted that in Part II negative numbers are used, which according to [60] seems to be the only place where negative numbers have been found in medieval Arabic mathematics.

Another book by Abu al Wafa of geometrical interest, is a book on the geometric constructions which are necessary for a craftsman. It deals with design and testing of drafting instruments, construction of right angles, approximate angle trisections, constructions of parabolas, and many other geometric constructions. When possible he tries to solve the problems with ruler and compass, and when this is not possible he uses approximate methods. In addition there is a collection of problems which he solves using a ruler and fixed compass, a so called *rusty compass*, where the angle between the legs is fixed. Such constructions give more exact drawings than when the opening of the compass varies.

Abul Wafa was first to use the trigonometric tan-function, and he compiled tables of sines and tangents, as part of his work on the orbit of the Moon. He also introduced the sec and cosec.

Abul Wafa devised a new method of calculating sine tables. His trigonometric tables are more accurate the ones by Ptolemy. His other works include a simplified version of Ptolemy's Almagest, the observational data in it was used by many later astronomers.

5.7 Abu Sahl Wijan bin Rustam al Quhi

Al Quhi was born in 940 in Tabaristan, now Mazanderan in Persia, and he died around 1000.

He was leading in a revival of Greek higher geometry in the Arabic world, according to Berggren in [3] perhaps the most accomplished of the many mathematicians patronized by the Buyid rulers of that time. The problems he studied typically led to quadratic or cubic equations. Nasir al Din al Tusi, which we treat in Sect. 5.11, described one of the problems considered by al Quhi. Following the translation in [3], he writes:

I say that Abu Sahl Wijan bin Rustam al Quhi has an epistle called "Filling a Lacuna in the Second Discourse of Archimedes' Book", and in it he said:

Here are three constructions from one domain. One of them is to construct a segment of a sphere equal to a segment of a sphere and similar to a segment of another sphere. The second is to construct a segment of a sphere whose surface is equal to the surface of a segment of a sphere and [which is] similar to a segment of another sphere. And the third is to construct a segment of a sphere equal to a segment of a sphere and whose surface [is equal to] the surface of a segment of another sphere.

Archimedes showed the first two, but he neglected the third, and no one after him appended it to the two of them.

Then he [Abu Sahl al Quhi] presented it. And its proof is as follows.

Al-Quhi's solution uses results from Euclid's Elements, Apollonius's Conics and Archimedes' On the sphere and cylinder. We shall give a modern deduction of his results below.

On Fig. 5.8, the surface S of the segment is indicated in orange to the left, to the right the segment is viewed directly from the side. An enlarged "infinitesimal" piece $\mathrm{d}S$ of the surface is indicated in red. With notations as on the figure, we then have

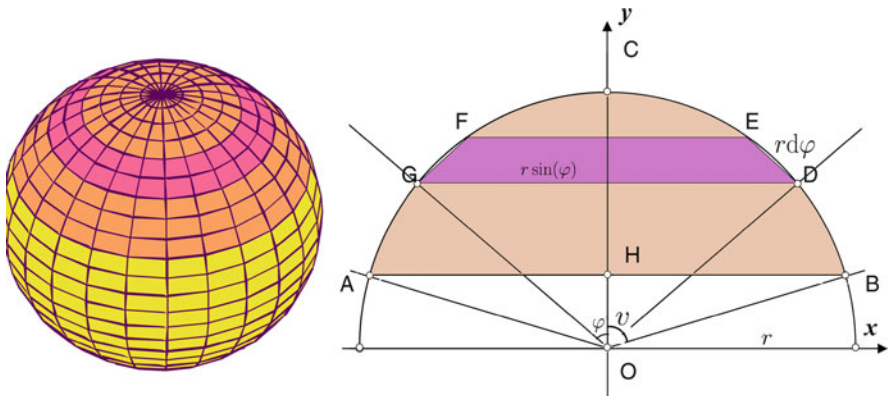


Fig. 5.8 Area of a segment of a sphere

$$dS = (2\pi r \sin(\varphi))r d\varphi = 2\pi r^2 \sin(\varphi) d\varphi$$

and thus

$$S = 2\pi r^2 \int_0^v \sin(\varphi) d\varphi = 2\pi r^2(1 - \cos(v)) = 2\pi r h$$

where $h = HC$ is the *height* of the spherical segment.

We may proceed to compute the volume of the segment in a similar way, but we get a somewhat simpler computation by using the observation of Archimedes referred to in Exercise 4.1. In fact, Archimedes writes that the volume of a sphere must be equal to that of a cone where the area of the base is equal to the surface area of the sphere, and the height is equal to the radius of the sphere. We may arrive at this idea by approximating the volume of the sphere by a collection of pyramids, say pyramids with bases equal to the small rectangles where the sides are the cords of the arcs appearing in the grid on the sphere shown left on Fig. 5.8. When the mesh is infinitely refined, we approach the volume of the sphere itself. If we restrict to the mesh appearing on the orange piece of the sphere, we similarly conclude that the volume of the conical piece of the sphere with base equal to the area of the sphere segment and top at the center of the sphere will have volume $W_1 = \frac{1}{3} r S$. Denoting by W_2 the volume of the cone with top at the center of the sphere and the circular base with diameter AB, we get the volume V of the spherical segment as

$$\begin{aligned} V = W_1 - W_2 &= \frac{1}{3}(2\pi r^2 h) - \frac{1}{3} r \cos(v) \pi (r \sin(v))^2 \\ &= \frac{2\pi}{3} r^2 h - \frac{\pi}{3} r \cos(v)(r^2 - r^2 \cos^2(v)) \\ &= \frac{\pi}{3}\left\{2r^2 h - (r-h)(r^2 - (r-h)^2)\right\} \\ &= \frac{\pi}{3}\left\{2r^2 h - (r-h)(2rh - h^2)\right\} = \frac{\pi}{3}(3rh^2 - h^3) \end{aligned}$$

Now there are given two segments of spheres, of radius r_1 and r_2, heights h_1 and h_2, areas S_1 and S_2 and volumes V_1 and V_2, respectively. The conditions which the unknown segment must satisfy are

$$S = S_1, \quad V = V_2,$$

in other words

$$2\pi rh = 2\pi r_1 h_1, \quad \frac{\pi}{3}(3h^2 r - h^3) = \frac{\pi}{3}(3h_2^2 r_2 - h_2^3)$$

i.e.,

$$rh = r_1 h_1, \quad 3h^2 r - h^3 = 3h_2^2 r_2 - h_2^3.$$

Using the former, the latter yields

$$h^3 + (3h_2^2 r_2 - h_2^3) = 3(r_1 h_1)h$$

which is a problem of the type *a cube and a number equal to sides*, and multiplying this by r^3 we get

$$(hr)^3 + (3h_2^2 r_2 - h_2^3)r^3 = 3(r_1 h_1)hr^3$$

which yields

$$r^3 + \frac{r_1^3 h_1^3}{3h_2^2 r_2 - h_2^3} = \frac{3r_1^2 h_1^2}{3h_2^2 r_2 - h_2^3} r^2$$

which is a *cube and a number equal to squares.*

In both cases a solution of the cubic equation is found by intersecting a parabola and a simple hyperbola. Indeed, we first consider, in modern terms, a general equation of the type *cube and number equal to sides.* Dividing by the coefficient of X^3 we may write this, without loss of generality, as

$$X^3 + b = aX,$$

or, since we may assume that $X \neq 0$,

$$X^2 + b\frac{1}{X} = a.$$

Thus we find a solution by intersecting the parabola

$$X^2 + bY = a$$

with the hyperbola

$$XY = 1.$$

As for *cube and number equal to squares*, we similarly consider

$$X^3 + b = aX^2,$$

or

$$X^2 + b\frac{1}{X} = aX,$$

where a solution is found by intersecting the parabola

$$X^2 + bY = aX$$

with the same hyperbola as above. Thus writing

$$b = 3h_2^2 r_2 - h_2^3$$

and

$$a = 3r_1 h_1$$

we get the equation

$$X^3 - aX + b = 0$$

and letting $r_1 = 3, h_1 = 2, r_2 = 2$ and $h_2 = 1$ this yields $a = 18$ and $b = 5$ and thus the equation

$$X^3 - 18X + 5 = 0.$$

Using MAPLE we find that the equation has three real solutions, namely $x_1 \approx -4.38$, $x_2 \approx 0.28$ and $x_3 \approx 4.10$. Here only x_3 corresponds to a geometric solution yielding $r \approx 4.10$ and thus $h = \frac{r_1 h_1}{x_2} \approx \frac{6}{4.1} \approx 1.46$.

Using the first version of the method, we have to find the intersection of the two conic sections

$$x^2 + by - a = 0 \text{ and } xy - 1 = 0$$

obtaining the graphs which confirms what we stated above (Fig. 5.9).

5.8 Yusuf al Mutaman ibn Hud and his Library

Yusuf al Mutaman ibn Hud was king of Saragossa from 1081 to 1085. He wrote an extensive book on geometry, relying on classical Greek and more recent Arabic sources, which he had in his large library. But in 1110 the Banu Hud dynasty was driven out of Saragossa by the Almoravids, an invading Berber dynasty from the Sahara. They took refuge in the fortress of Rueda de Jalón in Aragon, bringing the library along. This library came to play an important role for the translation of Greek and Arabic works which later took place in Toledo.

Ibn Hud did significant mathematical work himself. Thus the so called *Ceva's Theorem*, which we treat in Exercise 5.2, is credited to the Italian mathematician

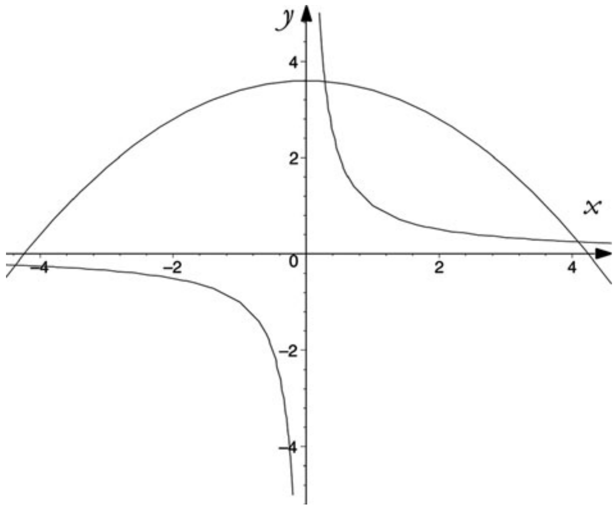


Fig. 5.9 Solution by intersection of two conics

Giovanni Ceva, 1647–1734. But this beautiful theorem was actually discovered by al-Mutaman ibn Hud.

5.9 Omar al-Khayyam

Ghiyath al-Din Abu'l-Fath Omar Ibn Ibrahim al-Nisaburi al-Khayyam, or Omar Khayyam as he is known in Western literature, was born in 1048 and died 1131 in Nishapur, Persia.

Al-Khayyam lived and worked during difficult times (Fig. 5.10). The Seljuk ruler *Toghril Beg* entered Baghdad in 1055, when al-Khayyam was a young boy. Toghril Beg was the second ruler of the Seljuk dynasty, he established the Seljuk Sultanate after having conquered the Capital of Baghdad from the Buyid Dynasty. The Abbassid Caliphs became mere figureheads. He then proceeded to use the caliphate's armies against the Byzantine Empire and the Fatimid Caliphate and thus consolidate his rule.

In these times of political and religious strife al-Khayyam, who in addition to mathematics also was brilliantly active in poetry and philosophy, complained bitterly. When he was a student, he wrote the following according to his biography in [19]:

"I was unable to devote myself to the learning of this algebra and the continued concentration upon it, because of obstacles in the vagaries of time which hindered me; for we have been deprived of all the people of knowledge save for a group, small in number, with many troubles, whose concern in life is to snatch the opportunity, when time is asleep, to devote themselves meanwhile to the investigation and

Fig. 5.10 Omar al Khayyam. Drawing by the author

perfection of a science; for the majority of people who imitate philosophers confuse the true with the false, and they do nothing but deceive and pretend knowledge, and they do not use what they know of the sciences except for base and material purposes; and if they see a certain person seeking for the right and preferring the truth, doing his best to refute the false and untrue and leaving aside hypocrisy and deceit, they make a fool of him and mock him."

Mathematically he is best known for his work on cubic equations, "Treatise on Demonstration of Problems of al-Jabr and al-Muqabala".

The words in the title are explained in Sect. 5.4. This book was written in Samarkand in Uzbekistan, 1070 he moved to that city. There he was supported by Abu Tahir, a prominent jurist of Samarkand.

Al-Khayyam is known both as an eminent astronomer and as a brilliant mathematician. Toghril Beg's capital was Esfahan, where his grandson *Malik-Shah* ruled from 1073. Malik-Shah and his vizier Nizam al-Mulk invited al Khayyam to come to Esfahan in order to direct the founding of an Observatory. Al Khayyam was the leader of the scientists working there for 18 years, during this time al Khayyam could work as mathematician, astronomer, poet and philosopher. With his group he produced astronomical tables and played a key role in a calendar reform which took effect in 1079. Gibbon writes that this calendar surpassed the Julian, and approached the accuracy of the Gregorian one.

In fact, [60] reports that al Khayyam measured the length of the year as 365.24219858156 days, a very accurate computation. The length of a year changes in the sixth decimal place over a lifetime, at the end of the nineteenth century was 365.242196 days and today it is 365.242190.

In 1092 political events ended Khayyam's period of peaceful existence. Malik-Shah died in November of that year, a month after his vizier Nizam al-Mulk had been murdered on the road from Esfahan to Baghdad by the movement called the Assassins. Malik-Shah's second wife took over as ruler for 2 years but she was less favorable to the Observatory and al Khayyam's calendar reform. Al Khayyam also came under attack for his philosophical ideas, which were in line with those of Aristotle.

Khayyam remained at the Court and tried to continue his work. He also wrote a book where he described former rulers in Persia as men of great honor who had supported public works, science and scholarship, obviously to call the attention of the present rulers to these worthy representatives of the earlier ruling classes.

Malik-Shah's third son Sanjar, who was governor of Khorasan, became the overall ruler of the Seljuq empire in 1118. Sometime after this al Khayyam left Esfahan and travelled to Merv (now Mary, Turkmenistan) which Sanjar had made the capital of the Seljuq empire. Sanjar created an important center of learning in Merv where al Khayyam continued his work for some years. He died in 1131, in Nishapur. His mausoleum still exists there, in fact our picture here is of a statue in the Mausoleum.

His *Treatise on Demonstration* gives algorithms for the solution of cubic equations, analogous to the ones al-Khwarizmi had given for the three types of quadratic equations.

But al-Khayyam is compelled to conclude that *neither we nor anyone else working with algebra have been able to do this. Perhaps someone coming after us will be able to succeed.*

However he did gibe a complete classification of cubic equations, with geometric solutions, found by intersecting suitable conic curves.[9]

As al-Khwarizmi had done before him, al-Khayyam of course only worked with positive numbers. Consequently, he would treat separately each case of the different cubic equations with at least one real solution. Excluding the case where the third power of the unknown entity does not occur, there are altogether 14 different types, which he collected in three groups.

Below we list these types, taking as our point of departure the usual modern form of a cubic equation

$$ax^3 + bx^2 + cx + d = 0.$$

As we assume that the third power of the unknown actually does occur, we have $a = 1$. Moreover, as we are only allowed to work with positive real numbers, the

[9] We now follow [38].

terms with negative coefficients will have to be moved to the right hand side of the equal sign.

The first group consists of the equations consisting of two terms only. However, if $d = 0$ then some power of x may be cancelled, and as an equation like, say $x^3 + 1 = 0$ can have no positive real solution, the first group consists only of the equation

$$x^3 = d.$$

Then there is a group of six equations, each with three terms:

$$\begin{aligned}
x^3 + cx &= d \\
x^3 + d &= cx \\
x^3 &= cx + d \\
x^3 + bx^2 &= d \\
x^3 + d &= bx^2 \\
x^3 &= bx^2 + d.
\end{aligned}$$

The third and final group consist of seven equations, each with four terms:

$$\begin{aligned}
x^3 + bx^2 + cx &= d \\
x^3 + bx^2 + d &= cx \\
x^3 + cx + d &= bx^2 \\
x^3 &= bx^2 + cx + d \\
x^3 + bx^2 &= cx + d \\
x^3 + cx &= bx^2 + d \\
x^3 + d &= bx^2 + cx.
\end{aligned}$$

Of course al-Khayyam never used a notation resembling this, but it is simpler to understand than the one employed by al Khayyam. Thus for instance, the first equation he formulated as the problem of

> A cube equal to a number,

while the second one was given as

> A cube and sides equal to a number,

and the second to last in the second group being

> A cube and a number equal to squares.

In all 14 cases al Khayyam shows how solutions may be found by conic sections. The first equation listed amounts to taking the square root of d, as a construction with ruler and compass this is impossible already for $d = 2$ as we shall see in Sect. 17.5. However, Greek geometers soon realized that the problem is soluble by intersecting conic sections, rather than just lines and circles. We show the solution of this problem in the spirit of al Quhi from Fig. 5.9 on Fig. 5.11. The solution of the

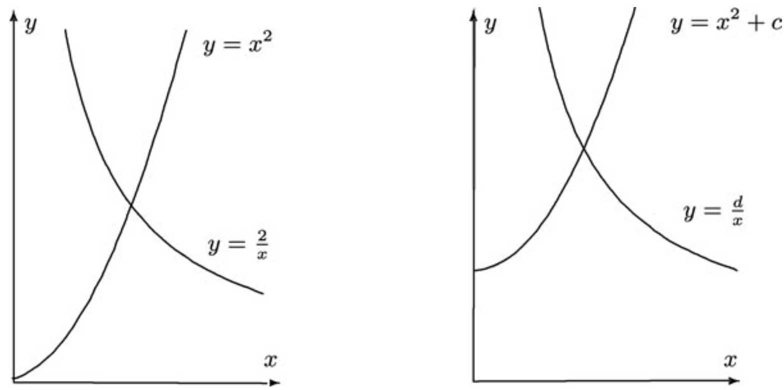


Fig. 5.11 Solution by conics

first equation listed, "A cube equal to a number", by a parabola and a hyperbola is shown to the left. Right the second equation, "A cube and sides equal to a number", is shown, also as the intersection of a parabola and a hyperbola.

As for the second equation on the list above,

$$x^3 + cx = d,$$

it will have only one real root, which may be found in the same spirit from the point of intersection of the two conics

$$y = \frac{d}{x} \text{ and } y = c + x^2.$$

We obtain the graphs to the right on Fig. 5.11.

However, al Khayyam proceeds differently. For example, he finds the solution of the second equation by intersecting a *circle* and a parabola, perhaps more in line with the Greek tradition of ruler and compass:

$$\left(x - \frac{d}{2c}\right)^2 + y^2 = \left(\frac{d}{2c}\right)^2 \text{ and } x^2 = \sqrt{c}\,y.$$

We see that this yields the solution by substituting $y = \frac{x^2}{\sqrt{c}}$ in the first equation:

$$\begin{aligned}&\left(x - \frac{d}{2c}\right)^2 + \left(\frac{x^2}{\sqrt{c}}\right)^2 - \left(\frac{d}{2c}\right)^2 \\ &= x^2 - \frac{dx}{c} + \left(\frac{d}{2c}\right)^2 + \frac{x^4}{c} - \left(\frac{d}{2c}\right)^2 \\ &= \frac{x}{c}(x^3 + cx - d)\end{aligned}$$

Using our algebraic notation, it is not difficult to see that all the equations listed above may be solved by intersecting conics in this way. Indeed, consider the general cubic equation $x^3 + bx^2 + cx + d = 0$. If $d = 0$, it is reduced to a quadratic equation, and is therefore obviously solvable by conic sections. So we may assume that $d \neq 0$. Then $x = 0$ is not a root, thus the equation is equivalent to $x^2+bx+c+\frac{d}{x} = 0$, which may be solved as the intersection of the parabola $y = x^2+bx+c$ and the hyperbola $y = -\frac{d}{x}$. But this general equation may also be solved by intersecting a circle and a parabola (Fig. 5.12). In fact, we are given the equation

$$x^3 + bx^2 + cx + d = 0.$$

As we easily see, the substitution

$$z = x + \frac{b}{3}$$

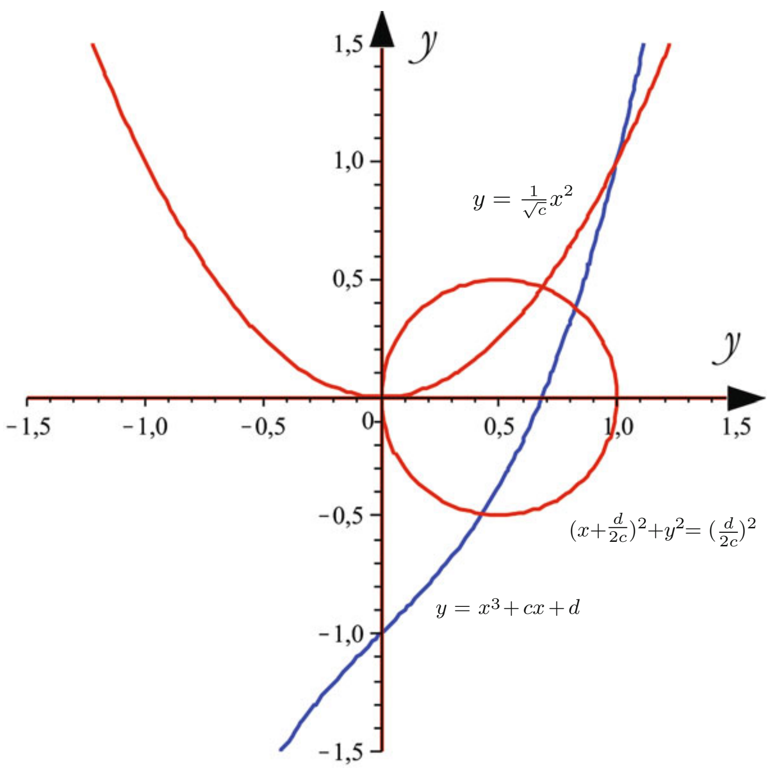


Fig. 5.12 The general cubic equation $x^3 + cx + d = 0$, solved by intersecting a circle and a parabola, in red. Here $c = 1, d = -1$. We also plot the corresponding curve given by $y = x^3 + cx + d$, in blue

will transform the equation into

$$z^3 + \bar{c}z + \bar{d} = 0$$

with new coefficients. Hence we may assume, without loss of generality, that $b = 0$.

A simple computation shows that the equation

$$x^3 + cx + d = 0$$

is satisfied for x, provided (x, y) is a point of intersection between the circle and the parabola given below:

$$\left(x + \frac{d}{2c}\right)^2 + y^2 = \left(\frac{d}{2c}\right)^2$$

and

$$y = \frac{1}{\sqrt{c}}x^2.$$

Here we have assumed that $c, d \neq 0$.

His work was continued by other Arabic mathematicians, in particular by Sharaf al-Din al-Tusi, who further developed the algebra required to find, much later still, the *formula* for the solutions to the general cubic equation.

Al-Khayyam refers to work, now lost, which uses the *binomial coefficients* organized in the pattern which we know today as the *triangle of Pascal*. This pattern had been explored by *al-Karajial-Karaji,* an eminent Arabic algebraist who lived from 953 to about 1029. According to [60] Al-Karaji can be *regarded as the first mathematician who freed algebra from geometrical operations and replace them with the type of operations which are at the core of algebra today.*

He also wrote commentaries on Euclid, med with the title "Explanation of the difficulties in Euclid's postulates". Here he discusses Eudoksos' theory for *ratios*, as it is explained in Book 10 of Euclid's Elements. Al-Khayyam treats these ratios as *numbers*, a major step towards an introduction of *real numbers*, in our modern language. Thus, he is able to understand the ratio of the diagonal to a side as a number, namely the irrational number which we denote by $\sqrt{2}$, or the ratio between the diameter and the circumference of a circle, which we understand as the transcendental number π. Thus one may say that al Khayyam in a mathematically precise manner introduced the *positive real numbers*, long before this was completed in Europe through the work of Richard Dedekind.[10] The Greek never considered such ratios as numbers, and even if al-Khayyam treats them as numbers, he never claims that they *really are* numbers, although he does raise the question.

[10] Julius Wilhelm Richard Dedekind's, 1831–1916, major contribution was a redefinition of irrational numbers in terms of Dedekind cuts. He introduced the notion of an ideal which is fundamental to ring theory. Source: [60].

Another important contribution to the development of our modern concept of numbers due to al-Khayyam, is his proof that the two different definitions of ratios, or *proportions*, which we find in Greek mathematics are equivalent. Eudoxus' definition is equivalent to the one attributed to Aristotle, namely that two magnitudes of the same kind have the same ratio as two other magnitudes of the same kind if the two sets have the same *antanairesis*. This means that the, possibly infinite, process of repeated subtractions performed for one of the pairs is the same as that of the other pair. See Sect. 4.2.4.

In his comments to Euclid, al-Khayyam also attempts to prove the *Fifth Postulate*. He defines two lines to be parallel if they everywhere have the same distance. Of course this attempt was unsuccessful, as we know today that there exist geometries where the Fifth Postulate is not true, while the remaining postulates hold, the so-called *non Euclidian geometries*, see Sect. 8.2.

Al-Khayyami was in all respects an eminent representative of Arabic intellectual and cultural life. As poet he has made a lasting impression, and as philosopher he belonged to the same direction as his important European contemporary thinkers like *Pierre Abélard*.

Peter Abélard or Petrus Abaelardus or Abailard, was a medieval French scholastic philosopher, theologian and logician. We refer to [61] and [34] for more on the sad story of his biography.

Like Abélard he got in trouble with the prevalent orthodox religious thinking of his time. Al-Khayyami is some times called *Persia's Voltaire*, perhaps it would be more appropriate to refer to Voltaire as *the al-Khayyam of France*, although arguably this might be an overestimation of Voltaire. François-Marie ("Voltaire") Arouet (1694–1778), was a French writer and philosopher, an outspoken supporter of social reform despite strict censorship laws. He criticized the dogmas of the Church and the French institutions and political system. We refer to [61] for more about this important philosopher.

5.10 Sharaf al-Din

Sharaf al-Din al-Tusi was born 1135 in the province Tus northwest in Persia, and died in 1213. It is not known whether he came from the city of *Tus* in the province by the same name, or from the city of Nishapur, near Tus in northern Persia. This is the city which al-Khayyam came from and where he studied. So these two mathematicians came from the same intellectual environment.

Sharaf al-Din is reported to have been teaching in Damaskus around 1165. The Seljukian Turks had conquered this city in 1154 and made it their capital. In Damaskus Sharaf al-Din taught from Euclid's and Ptolemy's works, until he moved to Aleppo, after Damaskus the largest city in the area.

Aleppo had 50 years earlier endured the siege of the Crusaders, and Sharaf al-Din taught mathematics, astronomy and astrology to an audience coming from a population with strong infusion of Jews and Muslims. From Aleppo Sharaf al-Din moved on to the city of *Mosul* north west, in present Irak, by the river Tigris.

Mosul flourished during this time, under the *Zangid* dynasty. Here Sharaf al-Din got a student who became quite famous, namely *Kamal al-Din Ibn Yunus*. He again became the teacher of the great *Nasir al-Din al-Tusi*, see Sect. 5.11.

Sharaf al-Din became quite famous, and students from all over the Middle East came to his lectures in great numbers.

Presumably Sharaf al-Din still remained in Mosul at the time when another former resident of that city returned, at the head of his army: That was the Kurd *Salah al-Din Yusuf*, known in the West under the name “Saladin”. Saladin took Damaskus i 1174, and about this time Sharaf al-Din left Mosul. He now remained in Baghdad for the rest of his life, teaching and writing his mathematical works.

Sharaf al-Din improved al-Khayyam’s methods for treatment of cubic equations. Like al-Khayyam he classifies them in groups, but with a different arrangement than the one used by al-Khayyam. He had good reasons for this change: Indeed, he wanted to analyze the conditions under which the equations had one, two or three solutions, meaning of course real and positive ones. In achieving this he penetrates much deeper into the theory than al-Khayyam had done before him. Indeed, Rashed argues in [49] that Sharaf al-Din’s algebra points forward to an algebra which studies *curves in terms of their equations*, and thus represents the opening up of a new field in mathematics, namely the field of *Algebraic Geometry*.

Sharaf al-Din has 25 types of equations of degree at most 3.

The first group consists of 12 types, and is formed by the equations which may be reduced to equations of degree 2, and the equation $x^3 = d$. The next group consists of eight types which all have at least one positive solution, while the third group, consisting of five types, are those which for some values of the coefficients have, and for other values do not have, positive solutions.

For solving the equations in this group he gives the same type of methods as the ones employed by al-Quhi, treated here in Sect. 5.7 and al Khayyam, which we treat in Sect 5.9, namely by intersecting conic curves. But Sharaf al-Din is very careful in proving that the two conics do indeed intersect one another.

The five equations in his third group of are treated in a new and original way. As we have done before, we use our modern notation, and may then describe the five equations by giving the following standard forms, where all coefficients are positive numbers:

$$\begin{gathered} x^3 + d = bx^2 \\ x^3 + d = cx \\ x^3 + bx^2 + d = cx \\ x^3 + cx + d = bx^2 \\ x^3 + d = bx^2 + cx. \end{gathered}$$

We illustrate the method of Sharaf al-Din by treating two of the equations on this list starting with the first, namely

$$x^3 + d = bx^2.$$

He first transforms this equation into

$$x^2(b - x) = d.$$

The equation will then have a solution if there exists a number x such that the expression to the left attains the value d, otherwise not. If $x = 0$, the value of the expression is 0, and the same is true if $x = b$. Between these two values of x the expression $x^2(b - x)$ will first increase, and thereafter decrease to 0 again. The problem therefore is to determine the maximal value which this expression can attain as x increases from 0 to b. If this maximal value is greater than or equal to d, he takes it as evident that there must exist a value $x = x_1$ between 0 and b such that $x_1^2(b - x_1) = d$. Then $x = x_1$ will be a solution of the equation. With our present days standard we would say that this assertion requires a *proof*, and here it is decisive that what we call the *function* $y = f(x) = x^2(x - b)$, is *continuous* between $x = 0$ and $x = b$.

A formal discussion of the concept of continuity and its properties is beyond the scope of this book, although the subject is not all that difficult. However, somewhat popularized and simplified we may say if a curve defined over a closed connected interval $[a, b]$ by the function $y = f(x)$ may be drawn without lifting the pencil from the paper, then the function $y = f(x)$ is continuous on $[a, b]$. See the Fig. 5.13, where we show the graph of a continuous function $y = f(x)$ on the closed interval $[a, b]$ to the left, and of a function $y = g(x)$ which is not continuous on $[a, b]$ to the right, it has a *discontinuity* at $x = c$.

With this heuristic "definition" of continuity, it is evident that if d is a number between $f(a)$ and $f(b)$, then there is a value $c \in [a, b]$ such that $f(c) = d$. For the function $g(x)$ this is equally evidently not the case. Of course, with the formal definition of continuity this is a theorem which requires a proof, and the proof is not entirely obvious.

Sharaf al-Din now says that the maximal value of the expression $x^2(b - x)$ for positive values of x is attained for $x = x_0 = \frac{2}{3}b$. This is absolutely right! How he arrived at this result, is a mystery. Some believe that he simply *guessed,* building on a result in Euclid's Elements, where the analogous problem for $x(b - x)$ is solved: This expression attains its maximum for positive x at $x = \frac{1}{2}b$. Others believe that he may have carefully studied Archimedes' famous book *On the Sphere*

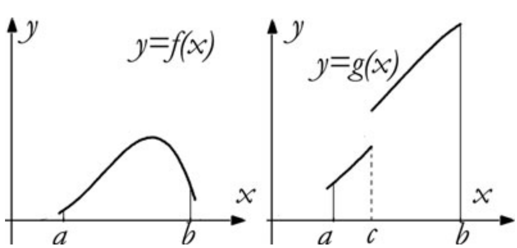


Fig. 5.13 Intuitive continuity

and the Cylinder, where a similar problem is considered. Another possibility is that Sharaf al-Din may have carried out a program which to a large degree has anticipated central elements of modern mathematics, by a procedure analogous to modern derivation. Indeed, the derivative vanishes where an expression stops growing and start decreasing again, and the derivative of $y = x^2(b - x)$ is $y' = 2xb - 3x^2$, vanishing for $x = x_0 = \frac{2}{3}b$.

Be that as it may, when we substitute $x = x_0 = \frac{2}{3}b$ in $x^2(b - x)$ we get the maximal value of y as $y_{max} = \frac{4}{27}b^3$, and Sharaf al-Din gives a completely correct geometric proof for that this is the maximal value of $y = x^2(b - x)$ for positive x.

As we can see from Fig. 5.14, there are normally two values of c, namely x_1 and x_2 which are solutions of the equation $x^2(b - x) = d$, provided that

$$d \leq \frac{4}{27}b^3,$$

and when we have equality then $x_1 = x_2 = x_0$.

We next tret the second equation on the list,

$$x^3 + d = cx$$

which he writes as

$$x(c - x^2) = d.$$

Sharaf al-Din first notes that if x is a (positive) root, then $(c - x^2) \geq 0$, thus $x^2 \leq c$, i.e., $x \leq \sqrt{c}$. As before he finds that the expression $y = x(c - x^2)$ attains its maximum for $x = x_0 = \sqrt{\frac{c}{3}}$. This yields the maximal value for y as $y_{max} = \frac{2c}{3}\sqrt{\frac{c}{3}}$, and thus the condition for the existence of a (positive) root is that

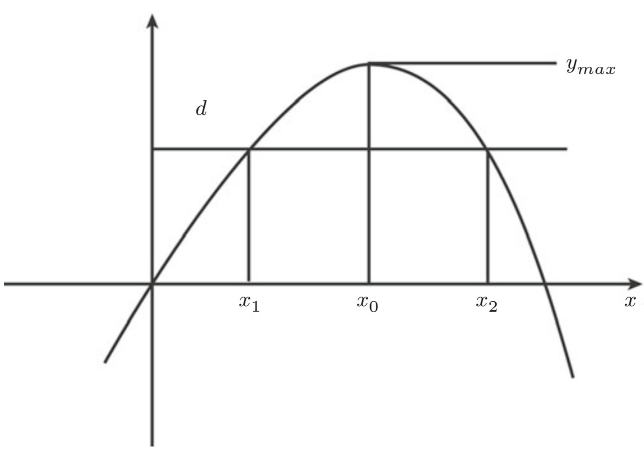


Fig. 5.14 The two positive solutions x_1 and x_2

$$d \leq \frac{2c}{3}\sqrt{\frac{c}{3}}$$

or equivalent

$$d^2 \leq \frac{4c^3}{27},$$

a condition which may be written as

$$0 \leq \left(\frac{c}{3}\right)^3 - \left(\frac{d}{2}\right)^2.$$

We see that in this case as well there are normally two such solutions.

These two examples from Sharaf al-Din's list explained above, give criteria for equations of the specified type to have more than one solution. The two criteria have the appearance of being quite different: The equation

$$x^3 + d = bx^2$$

will have more than one solution if and only if

$$d \leq \frac{4}{27}b^3,$$

while the for the equation

$$x^3 + d = cx$$

the criterion is

$$0 \leq \left(\frac{c}{3}\right)^3 - \left(\frac{d}{2}\right)^2.$$

However, both are special cases of a general criterion, which acquired its final form much later. We now move ahead several hundred years to explain this criterion.

We start out by looking at the general quadratic equation, from a modern point of view. In particular we no longer require that the coefficients, nor the solutions, be *positive* numbers:

$$ax^2 + bx + c = 0 \text{ where } a \neq 0.$$

The general solution for this equation of this equation is given by a formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Thus there are two solutions,

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \text{ and } x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

When $b^2 - 4ac < 0$, the equation has no (real) solution. Of course, today we consider expressions involving square roots of negative numbers as genuine numbers, but it took a long time before mathematicians and other users of mathematics could really compute with such *imaginary* numbers with confidence. Indeed, it was to a large degree precisely the study of the roots of algebraic equations which finally led to this breakthrough. At first it was realized that when such a solution did not exist, it was still possible to find true and useful information as an end result of computations involving *imagined* solutions, represented as a "sum" of a *true number* and an *imagined number*, the latter being a square root of a negative number. Today of course we work with composite numbers, the *complex numbers*, consisting of a *real part* and an *imaginary part*.

In this spirit we now compute with the solutions x_1 and x_2, whether they are *imagined* or *real*. We obtain the following:

$$x_1 - x_2 = \frac{\sqrt{b^2 - 4ac}}{a}$$

and thus

$$(x_1 - x_2)^2 = \frac{b^2 - 4ac}{a^2}.$$

This is of course a real number, which we denote by $D_2(a, b, c) = D_2$, referring to it as *the discriminant* of the equation. The reason for this language is that $D_2(a, b, c)$ makes it possible to distinguish or *discriminate* between the cases when the equation has (real) solutions and when it does not: A quadratic equation $ax^2 + bx + c = 0$ has (real) solutions if and only if $D_2(a, b, c) \geq 0$. It is practical to divide by a, so $D_2 = D_2(1, b, c)$.

We now move on to cubic equations,

$$x^3 + bx^2 + cx + d = 0.$$

This equation also has a formula for the solutions in terms of the coefficients b, c, d, square roots, cube roots and the arithmetical operations. In fact, we have a formula named after *Girolamo Cardano*. He was an Italian medical doctor and mathematician who wrote an important book known by its abbreviated entitle *Ars Magna*. This was the first Latin treatise devoted solely to algebra. Here he gives the methods for solving the cubic and quartic equations the first of which he had learnt from another Italian mathematician named Tartaglia, the second being due to his student *Lodovico Ferrari*.

Ferrari, by the way, was born in Bologna and as a young boy he was employed as a servant by Cardano. Cardano, realizing that the youth was very gifted and had taught himself to read and write, began teaching him mathematics. Ferrari rapidly moved form the position of being a servant to that of an assistant to Cardano, and eventually found the solution of quartic equations expressed by radicals and arithmetical operations in the coefficients of the equation. While such a "formula", in geometric terms or as rhetorical algebra, was at least very close to being known for the cubic equations by the Arabic algebraists, such as Sharaf al-Din, the corresponding result for quartic equations apparently was a completely new discovery. Then the search for the corresponding result for quintic equations went on for more than 200 years, until finally the Norwegian mathematician *Niels Henrik Abel* proved it impossible for all degrees ≥ 5. We give more detail on this in [34].

We now define the discriminant for a cubic equation by its roots analogously to what we did for quadratic equations, indeed we put

$$D_3(b, c, d) = (x_1 - x_2)^2 (x_1 - x_3)^2 (x_2 - x_3)^2,$$

and similarly also define the discriminant

$$D_n(b_1, b_2, \ldots, b_n) = \prod_{1 \leq i < j \leq n} (x_i - x_j)^2$$

for an equation of degree n,

$$x^n + b_1 x^{n-1} + \cdots + b_{n-1} x + b_n = 0.$$

We may show in general that D_n can be expressed in terms of the coefficients for the equation, analogously to what we had for $n = 2$. In particular we find that

$$D_3 = -27d^2 + 18dcb + b^2c^2 - 4b^3d - 4c^3.$$

By the way we defined D_3 it is clear that if the three solutions x_1, a_2 and x_3 are real, then $D_3 > 0$, and it is not too difficult to show the converse. Moreover, there are coinciding roots if and only if $D_3 = 0$.

Now we may observe that in fact, Sharaf al-Din found this criterion for the equations listed in his third group.

5.11 Nasir al-Din al-Tusi

Muhammad ibn Muhammad ibn al-Hasan al-Tusi was born in the province of Tus, north west in Persia, in 1201 and he died 1274 near Baghdad (Fig. 5.15). He lived during turbulent times, when Mongols overran the Muslim world with great cruelty. Nasir's life and work was very much influenced by these events. We follow the biographies in [60, 61].

His father was a jurist at the school of the Imamites, but he died when Nasir was quite young. Here he received a religious education, studying Arabic, the Quran, Hadith, Shi'a jurisprudence, but also logic, philosophy, mathematics, medicine and astronomy. It is reported that he had an uncle who attended to his secular upbringing, which as we noted included algebra and geometry, astronomy, physics as well as logics.

Fig. 5.15 Nasir al-Din al-Tusi. Drawing by the author

His family belonged to the so called *Twelver School*, a main branch of Shia Islam. As his father had wished he took learning and scholarship very seriously and travelled extensively to attend the lectures of renowned scholars. At a young age he moved to Nishapur to study philosophy and mathematics under two learned scholars, *Farid al-Din Damad* and *Muhammad Hasib*, respectively. He also met Farid al-Din al-Attar, a legendary Sufi master who was later killed by the Mongol invaders.

In Mawsil he studied mathematics and astronomy with Kamal al-Din Yunus. According to [61] the mysticism taught by Sufi masters at the time did not appeal to him, so he wrote a book with his own ideas on the subject, entitled "Awsaf al-ashraf", *The Attributes of the Illustrious*.

In 1214 the Mongols under the leadership of Dsjengis-Khan began their invasion of this area, and by 1220 they reached Tus, causing great destruction. However, before it came to this Nasir had completed his studies in Nishapur, to the West of for Tus. Here he had studied mathematics under *Kamal al-Din Ibn Yunus*, a former student of Sharaf al-Din al-Tusi. Nasir rapidly acquired a reputation as an outstanding man among the scholars in the area.

When the mongols invaded the area, he fled to join the so called "Assassins", a branch of the Ismailis who practised an intellectual form of extremist Shi'ism, and were devoted to resistance of foreign invaders. They controlled the castle of Alamut in the Elburz Mountains, and other similar impregnable forts in the mountains. When invited by the Ismaili ruler Nasir ad-Din 'Abd ar-Rahim to join the service of the Assassins, al-Tusi accepted and became a highly regarded member. Whether he would have been able to leave, had he wished to, is according to [60], not entirely clear. Be that as it may, al-Tusi did some of his best work while moving round the different strongholds, and during this period he wrote important works on logic, philosophy, mathematics and astronomy. The first of these works, Akhlaq-i nasiri, was written in 1232. It was a work on ethics which al-Tusi dedicated to the Ismaili ruler Nasir ad-Din'Abd ar-Rahim.

In 1256 Nasir stayed at the castle of Alamut, when it was attacked and conquered by the Mongols under their leader *Hulagu Kahn*, one of the exceedingly numerous grandsons of the great conqueror Dsjengis-Khan. Some say that Nasir betrayed the Assassins[11] but others strongly dispute this claim.

Hulagu treated Nasir with respect, although Hulagu himself cannot have been much interested in mathematics, judging from the later events in Baghdad. But he *was* interested in prophecies, divination and astrology, and he must have felt that a learned scholar like Nasir could be of assistance with such information. In any case the nature of Nasirs cooperation with Hulagu is a matter of controversy, and probably falls outside the scope of this book.

However, through careful maneuvering Nasir more and more acquired Hulagu's confidence, and finally became one of the most trusted officials of the ruler.

It seems reasonably clear that Nasir represented a civilizing force within the Mongol regime at the time. When Nasir proposed to build an astronomical observatory, the idea was well received by Hulagu. He had just moved his capital to Maragheh in present East Azarbaijan province, Iran. Aserbajdsjan. Here the observatory was constructed, it still exists and is a vigourous center of research.

The observatory was build by the Persians and Chinese astronomers. Many of the fie instruments it was equipped with, were constructed by Nasir personally. The observatory also got an exquisite library with the most outstanding scientific literature at the time. Undoubtedly these books had been taken form conquered and destroyed libraries, probably in Baghdad and other unfortunate cultural centers. At any rate, this made the observatory into an academy in the old Greco-Alexandrian tradition.

Nasir worked out astronomical tables, based on observations over 12 years. The tables were first written in Persian, and then later translated into the Arabic.

Nasir worked with the Ptolemaic model, and this model caused increasing problems as observations became better and better. The astronomers could not see that their observations were in agreement with Ptolemy's explanation! Eventually they had to modify Ptolemy's model, in rather artificial ways which were difficult to justify. Nasir al-Tusi came up with the most significant modification, before *Nicolaus Copernicus* scrapped Ptolemy' model altogether and proposed a fully scientific heliocentric description of the Solar System, thus initiating the Scientific Revolution.

One of the tools devised to save the Ptolemaic point of view, was to describe the movement along a straight line as the sum of two circular movements, a so called *Tusi-pair*. The same ideas are used in Copernicus' work.

Nasir created *trigonometry* as an independent mathematical discipline, and not just a tool for astronomical applications. In his *Treatise on the quadrilateral*, Nasir

[11] See [60]. In fact, the Assassins are quite controversial, the name itself is defamatory: It literally means "users of hashish", evidently not devised by their friends. The harsh procedures employed in their resistance struggle in turn gave name to political murder.

gave the first known exposition of plane and spherical trigonometry. This work also contains *Rule of Sines*, see Sect. 4.15.

Another mathematical contribution was Nasir's manuscript, dated 1265, concerning the calculation of nth roots of an integer. This work is probably an exposition of material coming from al-Karaji's school. In the manuscript Nasir determines the coefficients of the expansion of a binomial to any power giving the binomial formula and the "Pascal" triangle for binomial coefficients.

He also did important work in logic. He introduced symbols for the *implications*, where we now use $\Rightarrow$, "if–then", and a symbol for "either–or", we use $\vee$ today.

Exercises

Exercise 5.1 Assume that each public bath in Alexandria needed to heat 5 cubic meters of water from 15 to 75° C each day. Assume that the caloric value of each book in the library was, on the average, $\frac{1}{2}$ kWh, and that the efficiency of the stove and heating systems used was 60%. 1 kWh corresponds to 857 kcal. How many books did the library of ancient Alexandria contain, assuming that the stories about the number of public baths and the burning books is accurate?

Exercise 5.2 (Ceva's Theorem) The following result is credited to the Italian mathematician *Giovanni Ceva*, 1647–1734, but was actually discovered by al-Mutaman ibn Hud. It says the following:

Given a triangle $\triangle ABC$, and points D, E, and F which lie on the lines BC, CA and AB respectively, then the lines AD, BE and CF pass through a common point (are concurrent) if and only if $\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = 1$. Prove this theorem.

A line segment joining a vertex of a triangle with a point on the opposite side is referred to as a *Cevian line* or just a *Cevian*.

Chapter 6
The Geometry of Yesterday and Today

6.1 The Dark Middle Ages

The Roman Empire had become the carrier of a common civilization of philosophy, learning and mathematics throughout *"The Known World"*, the *The Oikumene.* Strangely, the Romans themselves were not particularly interested in it. It has been said that the only contribution the Romans ever made to mathematics is due to *Cicero*, when he rediscovered Archimedes' grave. But the fall of the Roman Empire marks the end of this common civilization, at least as a web encompassing the entire "known world". When the Roman Empire ceased to exist, this cultural web shrunk, and went into a kind of hibernation. In *Constantinople*, among the Arabs with the Caliphs at Baghdad and elsewhere, and to some extent on Sicily the seeds of culture and learning were preserved, as well as among individual thinkers, many of them monks in the monasteries, within the Christian Church.

The Church itself was very powerful, and may well be viewed as a successor to the Roman Empire. To the constantly changing kingdoms, countries and alliances, the Pope in Rome with his administration and wide network was the single stable institution. More than a millennium after the fall of the West Roman Empire, the Pope in Rome crowned the "Holy Roman Emperors" of Europe.

The legacy of the Roman Empire lasted a long time, in many ways it is still with us today. In fact, some of the most dangerous conflicts in Europe and the Middle East may be traced back to events during the final 500 years of the empire. And we may speculate what the situation in the world would have been today, if the Romans then had listened to Cicero and rejected Caesar.

The *Dark Ages* commences with the fall of the West Roman Empire and lasts until the middle of the eleventh century. During this time practically no mathematical activity took place in Europe. The exceptions to this claim are so meagre as only to strengthen the assertion.

Boethius was born around 480 and died 524 (Fig. 6.1). He belonged to a distinguished family, which counted several important senators from the old days of the Empire. He is said to have been *"the last of the Romans which Cato and Cicero could have acknowledged for their countrymen"*. He spent his boyhood in Rome when Odoacer was monarch. His father, *Flavius Manlius Boetius* had been consul in

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_6,

Fig. 6.1 Boethius teaching his students. From a 1385 Italian manuscript of the Consolation of Philosophy

487, and probably died soon after that year. According to Boetius himself, when he lost his parents *"men of the highest rank"* took him under their charge. He received an excellent education, as such scholarship still existed in Rome. He soon became known as a promising and able young man. But these were turbulent times, and Odoacer was by no means firmly seated on the throne of Italy. A dangerous rival was the Ostrogoth *Theodoric*, with his army of formidable warriors. A vicious enemy as well as an unreliable ally, he had received an extraordinary education for a barbarian: At the age of 7 he had been sent to the court in Constantinople as a hostage, and he remained there for 10 years. Returning to his father he established himself as an able leader, who greatly increased his father's domain. After he had succeeded his father as king, he invaded Italy in 489, and pushed Odoacer further and further back until he was able to lure him into a trap in the form of a grandiose banquet at the palace of Laurentum. There Theodoric killed Odoacer with his own hands. Now Theodoric became the new king, and his eyes fell on Boetius who became a favorite with the new ruler.

In 510 Boetius became consul, and in 522 his two sons, who were still young, became consuls together. Boetius now stood on the top of his career, when he was placed between his two sons in the Circus and received the ovations from the people.

But the fall from the summit was near. Intrigues at the palace resulted in charges of treason. He was supposed to have written letters to the Emperor in Constantinople, in order to restore the Empire in Rome. He unequivocally denied the accusations, but of course the resentment against the barbarians in power was deep among the old Roman families. So such accusations would not be altogether preposterous. At any rate he was brought before the king, and found guilty. He was thrown in jail, where he wrote his most famous work *De Consolatione Philosophiae*. He was executed in 524. The king is said to have later very much repented his rashness in putting Boetius to this mistreatment, so much so that he passed away soon after 524.

Boetius' mathematical works became standard texts throughout the Middle Ages. But his *Geometry* consist only of the propositions in Book I and parts of Books III and IV of Euclid's Elements. It also contains some elementary applications to mensuration. His *Arithmetic* is based on the work by *Nicomachus*, four centuries earlier.

Nevertheless, he appeared at a time when contempt for intellectual pursuits was widespread, and his fame and influence increased after his death. A central theme in his philosophical writing was to reconcile the ideas of Plato with those of Aristotle, and his grand idea was to revive the spirit of his countrymen by filling them with the thoughts of the ancient Greek writers.

Now Theodoric had been an Arian, that is to say a follower of the heretic *Arius*. He was the origin of the first great heresy-struggle in the Christian Church. His teachings, dealing with the relation between the *Father* and the *Son*, had been repudiated at the council of Nicaea in 325, but the controversy raged on.

In fact Arius himself would have been readmitted into the Church in Constantinople, from which he had been excommunicated, had he not died suddenly while walking with a friend one evening.

Since Boetius could be viewed as having lost his life while trying to depose the heretic ruler Theodoric, by aiding and abetting the pious orthodox emperor in Constantinople, his star rose even higher. So now Boetius was canonized as *Saint Severinus*, precisely on the merit of the false accusations once made against him.

Bede was the most learned Englishman of his age. He was born in 673 in Northumberland and died in 735. He is considered the father of English history, but his writings were on a broad range of subjects, making a total of about 40 different treatises.

These essentially amount to an entire encyclopedia. He also included some mathematics, a calendar and a treatment of finger reckoning. But there is no geometry, and Bede who was brought up in a monastery from childhood and remained there his entire life, was first and foremost a pious religious thinker.

Bede became known as Venerable Bede soon after his death, but this title probably comes from an error in Latin by a medieval scribe who meant to write about the venerable works of Bede.

The eighth century has been called *The Age of Bede.* Bede must be considered an important scientific figure, he wrote several major works (Fig. 6.2). *On Time* is an introduction to computing the date of Easter. *On the Reckoning of Time* is a longer work on the same subject. This book became important in clerical scientific education later.

The Reckoning of Time included an account of ancient and medieval view of the cosmos, with an explanation of how the spherical earth influenced the changing

Fig. 6.2 Depiction of Bede from the Nuremberg Chronicle, 1493

length of daylight, of how the seasonal motion of the Sun and Moon influenced the changing appearance of the New Moon at evening twilight, and the relation between the Tides and the daily motion of the moon. He also made a new calculation of the age of the world, but this got him in trouble. He was accused of heresy, and had to recant.

Alcuin, who gave his full name in Latin as *Flaccus Albinus*, was another learned Englishman. He was born in Yorkshire in 735 and died in 804. He was educated in Yorkshire under the direction of *Archbishop Egbert*, whom he succeeded as director of the seminary. The word of the learned Englishman reached the Emperor *Charlemagne*, and Alcuin was called upon to instruct the Emperor and his family in the subjects of *rhetoric, logic, mathematics* as well as *divinity*, the study of Christianity. He assisted Charlemagne in building up a seminary, or university, in Tours, which had a strong influence on higher education in France. But Alquin forbade the reading of the classical poets.

Gerbert was born about 950 in Auvergne in France and died in 1003 as *Pope Sylvester II*. It is generally assumed that he came from a rather poor background, but as a young boy he entered the Benedictine cloister of St. Gerald at Aurillac, where he received a good education, showing himself as exceptionally gifted.

In 967 Count Borrell of Barcelona visited the monastery, and the abbot asked the count to take Gerbert back to Spain with him so that the promising student could study mathematics there. The count did so, and put the young man under the protection of the bishop of Vic in Catalunya, where there was a cathedral school.

Here there was extensive contact with the Arab culture and civilization of al-Andalus to the south. Al-Andalus was much more advanced that Christian Europe, the library in the Islamic capital of Cordoba was overwhelming by contemporary

European standards. The libraries of the cathedral of Vic and the nearby monastery of Ripoll were among the best in Europe.

Here he came into contact with learned Arabian Islamic scholars. His studies earned him profound insights, by European standards at this time, into the subjects of *mathematics, astronomy* and *music*. Then in 970 he came to Rome, where the unusual sagacity of this young man caught the attention of the Pope himself. The Pope recommended him to the Emperor, *Otto the Great*, and for a while he served as the tutor and advisor of the latter. Later he went to Reims, where he had a group of students, and wrote on geometry as well as arithmetic. Gerbert was active in politics as well, worldly and ecclesiastical. In 999 this led to his election to the elevated office as Pope. As Pope he is not considered to be particularly outstanding, while as a mathematician and astronomer he is regarded the foremost for this period, around the previous turn of millennium. This is the more remarkable considering the rather meager nature of his scientific findings.

In his work on geometry, Gerbert solves the following problem, considered to be very difficult: In a right triangle the area and the hypothenuse are given. Find the remaining two sides. Of course this problem would have been handled easily by the Babylonians, and certainly as well by the Greek. It is a trivial consequence of the Pythagorean theorem with the use of some algebra, as the Babylonians would done it, or by geometric algebra, as the Greek would have proceeded. With our modern notation we proceed as follows: Let the two sides in question be of lengths x and y, the known diagonal be d and the area be A. Then we have

$$x^2 + y^2 = d^2$$
$$xy = 2A$$

We then find

$$(x + y)^2 = d^2 + 4A$$
$$(x - y)^2 = d^2 - 4A$$

from which x and y are easily found.

Gerbert also expresses the area of an equilateral triangle of side s as

$$A = \frac{s}{2}\left(s - \frac{s}{7}\right)$$

which corresponds to $\sqrt{3} \approx 1.714$, not a very good approximation by any means.

Gerbert is credited by some with being the first to have used the Indian-Arabic numerals in Europe. However, this is not a unanimously accepted view among historians of mathematics. Such were the superstitions and general ignorance in Europe at this time, that rumors of him having sold himself to Satan started to gain acceptance after his death.

6.2 Geometry Reawakening: A New Dawn in Europe

The classical mathematical books had been lost to Europe, but were preserved by the Arabs. In the twelfth century they were reintroduced into Europe, being translated from Arabic to Latin. Among the first was *Euclid's Elements*.

The translation of Euclid's Elements from the Arabic to Latin was done by the English monk *Adelard of Bath*. He visited Spain between 1126 and 1129, and travelled widely in Greece, Syria and Egypt.

The Jewish mathematician *Abraham bar Hiyya* played an important role during this time. He is also known under the name *Savasoda*. He lived and worked in Barcelona, being born there in 1070, and died in Provence in France in 1136.

He has written a book entitled *Treatise on Measurement and Calculation*, covering a broad range of subjects. It is the first book in Europe covering Arab algebra, containing the complete solution of the general quadratic equation. Abraham was familiar with the works of Greek geometers such as Euclid, Theodosius, Apollonius and Heron.

Gherardo of Cremona was active in the twelfth century, and oversaw the work of a group of translators who worked with material obtained from *Toledo*, which had been captured by the Christians in 1085. They translated more than 90 works from Arabic into Latin.

As we have already told in Sect. 4.22, Sicily was part of the Byzantine Empire until the middle of the ninth century. Then for about 50 years it was captured by the Arabs, so recaptured by the Byzantine Empire, and then captured by the Normans. The period of Norman rule was very good for science and culture. Sicily was then a melting pot for the Greek, Arab and Latin cultures, the contact with Constantinople and Baghdad was good and Greek and Arab manuscripts were translated into Latin.

Merchants from Italian and Spanish cities established ties with the East, so in this way as well the undeveloped and barbaric Europe was little by little gaining access to the cultural heritage of which it had been so long deprived.

In conclusion it should be emphasized that the Arabs not only served as preservers and messengers of the ancient civilization, they developed their science vigorously and originally in their own right.

6.3 Elementary Geometry and Higher Geometry

Following Euclid, it is has been customary to distinguish between *Elementary Geometry and Higher Geometry.* According to this tradition, elementary geometry deals with configurations built up from *points and lines*, as well as *circles*. Higher geometry is concerned with *general conic sections* as well as curves of *higher degrees*, or even *transcendental curves*. The line between these two parts has not always been sharply drawn, and the distinction is today rendered obsolete, at least within the main stream of research in geometry. It has, however, to some extent survived in some didactical treatments of the subject.

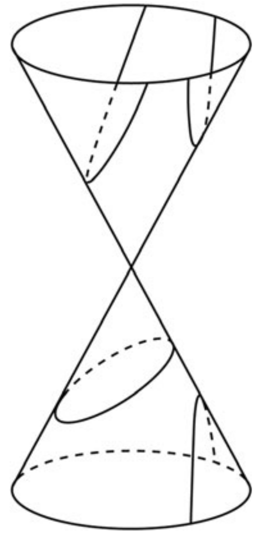


Fig. 6.3 In this figure we have cut a fixed cone with a plane in several different ways, and thus obtained one from each of the classes of conic sections: We see an ellipses, a parabola and a hyperbola

From a historical point of view, however, it is important to keep these two faces of geometry in mind. Indeed, let us consider the three famous classical problems, namely *Squaring the Circle, Doubling the Cube* and *Trisecting an Angle*. In their original and enigmatic form, these three constructions should be performed by straightedge and compass only, used in the *legal fashion* as explained in Sect. 3.6. That is to say, the problems should be solved with tools from *Elementary Geometry*, with the *Euclidian Tools.* As we shall prove in Chap. 17, in this form they are insoluble, all three of them. However, with tools from *higher geometry* they do have solutions, some of them very beautiful and deep ones, as we have seen in Chaps. 3 and 4.

Conic Sections is a class of curves in the plane, which are obtainable by *intersecting a circular cone with a plane.* This is indicated in Fig. 6.3.

Of course a *circle* is a conic section, but traditionally it was regarded as *"more elementary"* than the others. This is undoubtedly due to the fact that a circle may be produced by the Euclidian tool *compass*, while a general ellipse, for example, may not.

Up until the fifteenth century the two parts of geometry, elementary and higher, were pursued by methods which were basically the same, by the so-called *synthetic* methods. A circle may be drawn with a compass, a line using a straightedge. Points are found as intersections of lines, lines and circles or two circles. As for *higher curves* or conic sections, they could not be drawn in one piece by straightedge and compass, but were given in terms of definite constructions or procedures, as the *loci of points satisfying certain defining properties*, allowing the construction, by straightedge and compass, of a finite but arbitrarily large number of points on them. In many cases, as we have seen in the previous chapters, this made possible the manufacture of certain mechanical tools, which like a straightedge or compass made

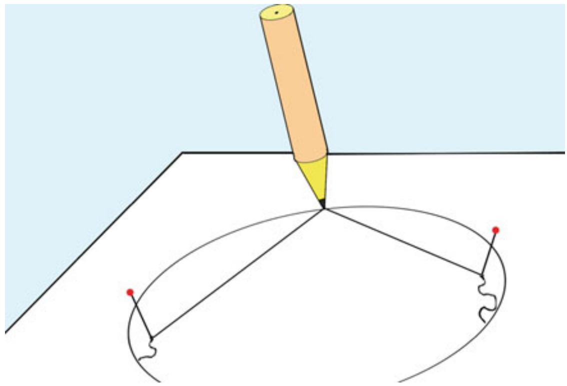


Fig. 6.4 We may draw an ellipse by means of two needles, a piece of sting and a pencil as shown here

it possible to produce the curve in question. The simplest case is the string-rule for producing an ellipse, explained in Fig. 6.4.

In the fifteenth century the methods from Greek geometry are augmented by powerful new techniques through work by *Roberval, Torricelli, Pascal, Desargues* and others. Roberval studied tangency for curves, considering a curve as being the locus of a point whose movement is the result of two separate movements. The vector sum of the two corresponding velocities would then define the tangent direction at any given point. We understand his approach today in terms of curves on parametric form. Torricelli had similar ideas, and the two mathematicians locked horns in priority disputes over their work, a disease endemic to the entire field of mathematics, past and present.

Descartes introduced *algebra* into geometry in a decisive manner. One might perhaps, with some justification, say that algebra was *reintroduced* into geometry, since the *Babylonians* had a well developed geometry based on their superb mastery of algebra. To a somewhat lesser extent this may also be said of the ancient Egyptians. And of course Greek geometers had used geometric algebra, above all Euclid and Apollonius. However, these observations do not belittle the importance of Descartes' contribution. Babylonian mathematics had become completely forgotten, and its rediscovery is an accomplishment of the twentieth century.

Moreover, Descartes' credit for the algebraization of geometry should be shared with another great French mathematician, *Pierre de Fermat*. His work was outlined in a letter to Roberval in 1636, it was then already 7 years old. His work on these matters was not published until after his death.

The line which we will follow is that of *higher geometry*, which has become the main stream of modern geometry.

But it certainly should be emphasized that the synthetic methods persist in *Axiomatic Geometry*. This field is tied to interesting questions in combinatorics and general algebraic systems. However, it falls outside the scope of the present book.

Another area which split off is that of *General Topology*, again giving rise to *Algebraic Topology*. This will also not be treated here.

Fig. 6.5 Gérard Desargues. Drawing by the author

6.4 Desargues and the Two Pascals

The French engineer and architect *Gérard.* Desargues lived in Lyon, which at that time was the second most important city of France (Fig. 6.5).

He belonged to a very wealthy family. He participated in the campaign against the city of la Rochelle as an engineer, spent some time in Paris and later in life retired to his estate in Condrieu.

During his time in Paris he pursued his interest in geometry in an environment including great mathematicians like *Descartes, Étienne Pascal* and his son, the young *Blaise Pascal* (1623–1662).

Desargues published a treatise on conic sections in Paris in 1639, without question his most important work. The title was *"Rough draft for an essay on the results of taking plane sections of a cone"* (Brouillon project d'une atteinte aux evenemens des rencontres du Cone avec un Plan). The book is short and densely written. He gives a unified treatment of *conic sections*.

A small number of copies was printed in Paris in 1639. The work had a very limited circulation, according to some it was mostly ignored, although it is said by others to have influenced Desargues' student, the young Blaise Pascal (Fig. 6.6). At any rate, all copies disappeared, and it was not until 1845 when the important French Geometer *Chasles* found a copy of it by one of Desargues' students, that the significance of Desargues' work was recognized. Then, in 1951 an original copy of the book resurfaced, so today the record has been straightened out as far as this part of Desargues' work is concerned.

Desargues' famous *Theorem of Perspective* was first published in 1648, in an appendix to work on perspective by Abraham Bosse on Desargue's method.

Fig. 6.6 Blaise Pascal

This theorem asserts that if two triangles are such that the three lines through corresponding vertices pass through a common point, then the three points of intersection of the (prolongations of) corresponding sides will lie on a common line. This important result will be treated in detail in Sect. 13.1.

Blaise Pascal's father Étienne Pascal also made contributions to mathematics in general and geometry in particular, see Sect. 15.7. But his son was by far a more important mathematician. Already at the age of 16 he wrote a treatise on conic sections, where he proved the theorems which we treat in Sects. 13.9 and 15.10. His contribution to projective geometry and conic sections is of fundamental importance. He also wrote on the *cycloids*, using methods which essentially amount to integration. And he laid the foundation for probability theory, corresponding with Fermat. His work was not confined to mathematics alone, indeed he profoundly influenced physics by bold and controversial ideas, like the possibility of a *vacuum*.

Then he turned to religious problems, and finally spent the last 8 years of his life in a monastery.

6.5 Descartes and Analytic Geometry

René du Perron Descartes (1596–1650) was a French mathematician and philosopher. In 1637 he published a book which contained an appendix, containing some path-breaking ideas.

The appendix had the title *La Géométrie,* while the book itself was entitled *"Discours de la méthode pour bien conduire sa raison et cherche la vérité dans les sciences"*. In translation it should be something like "A discussion of the method for correct reasoning and seeking the truth in science". We showed his picture in Sect. 4.7.

Descartes main contribution to science is contained in this appendix. It consists in about 100 pages, is divided into three parts, and presents the important ideas it contains in a rather obscure style. Some say that this was intentional on the part of Descartes.

Today Descartes' main idea would strike us as so obvious that it is hard to see how this would not be self evident to anyone doing geometry. The set of real numbers are identified with the points on a line. A point in the plane, correspondingly, is represented by a pair (x, y) of real numbers. The set of points $(x, 0)$ then form the x-axis, and the y-axis is the set of all points $(0, y)$. x and y are called the coordinates of the point $P = (x, y)$, x being referred to as the *abscissa* and y as the *ordinate* of P.

Now a line is defined by a relation, where a, b and c are fixed real numbers

$$ax + by + c = 0,$$

which holds if and only if the point (x, y) is on the line, and similarly any planar curve is given by a corresponding equation: Thus for example,

$$(x - 1)^2 + (y - 5)^2 - 25 = 0$$

represents a circle of radius 5 about the point (1,5).

In this way geometric considerations are translated into algebraic or analytic ones, depending on what kinds of equations we are dealing with: A line or a circle are algebraic curves, being defined by polynomial equations, while something like

$$y = \sin(x) \text{ or } y = e^x$$

are *analytic* curves which are not algebraic, we call such curves *transcendental curves*.

Of course geometry in 3-space is done similarly, a point now being $P = (x, y, z)$.

6.6 Newton and Leibniz

Subsequent translators and commentators have contributed to a more lucid exposition than that of Descartes, and later mathematicians like the eminent *Gottfried Wilhelm Leibniz* and *Isaac Newton*, eventually provided insights which finally resulted in the main framework of what is known as *Analytic Geometry* to all students of introductory calculus.

At the age around 25 Newton had to leave the University and go home because of the outbreak of a plague in Cambridge. There he worked on his own and did revolutionizing work in mathematics and other natural sciences. In particular he laid the foundations for differential and integral calculus, many years before it was independently discovery by Leibniz. His method of *fluxions* was the base, and he found analytical methods unifying diverse earlier techniques for finding areas, tangents, the lengths of curves and the maxima and minima of functions.

A deep observation by Newton, which we today take for granted in our Calculus, is that the two procedures of finding tangents and of finding area, are in some sense reverse to one another: The process of derivation is reverse to integration, in our language of today.

Newton wrote this mathematics up in 1671 but it was not published until much later. This resulted in a bitter fight over priority with Leibniz, who found his version independently of Newton but at a later time. But Leibniz also developed a better notation than Newton did, and it is Leibniz' version of calculus which is taught at colleges and universities today (Fig. 6.7).

Newton was the first who continued the study of conic sections by classifying *cubic curves* in the plane. He described 72 cases, missing six. The *Tridents* constitute one of these classes, these curves have equations of the form $xy = Ax^3 + Bx^2 + Cx + D$ (Fig. 6.8).

6.7 Geometry in the Eighteenth Century

In the eighteenth century calculus as we know it today took shape, amidst strong controversies.

Michel Rolle, a noted French mathematician, found Calculus totally inferior to Geometry with the rigorous standard of proofs inherited from Euclid. He went so far as to assert that Calculus was merely *"a collection of ingenious fallacies"*. He nevertheless made important contributions to the subject himself, and in later years became more approving of the new ideas of Calculus. He made two contributions which we shall mention here. First, *Rolle's Theorem* asserts that if a function $y = f(x)$ is continuous in the closed interval $[a, b]$ with $f(a) = f(b)$ and $f(x)$ is differentiable in the interior $< a, b >$, then there exists an x_0 in $< a, b >$ such that $f'(x_0) = 0$ (Fig. 6.9). He is also credited with having introduced the notation $\sqrt[n]{a}$ for the nth root of a.

Fig. 6.7 To the left Gottfried Wilhelm Leibniz, painting by Bernhard Christoph Francke, Braunschweig, Herzog-Anton-Ulrich-Museum, from 1700. To the right Isaac Newton at 46 in Godfrey Kneller's portrait from 1689

The Scottish mathematician *Collin Maclaurin* (Fig. 6.10) is best known by students of Calculus for his series-expansion, but also made contributions to Geometry. Indeed, he proposed an ingenious construction for *trisecting an arbitrary angle in three equal parts* by means of a curve of degree 3, today known as *the Trisectrix of Maclaurin*.

Maclaurin also published a theorem which is known today as *Cramer's Rule*. *Gabriel Cramer* found this result independently a little later than Maclaurin, and got all the credit for it. Probably this is due to a better notation in his treatment than the one employed by Maclaurin.

For the benefit of readers who are not familiar with linear algebra, we recall Cramer's theorem, which is important for geometry. This will be used later on. Actually Cramer published the result in a treatise on the geometry of lines, precisely a special case of the way we shall also benefit from the theorem here.

We start out with the simplest case, namely a system of two linear equations:

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 &= b_1 \end{aligned}$$

In this case Cramer's Theorem has the following simple form:

Theorem 5. *The system above has a unique solution if and only if*

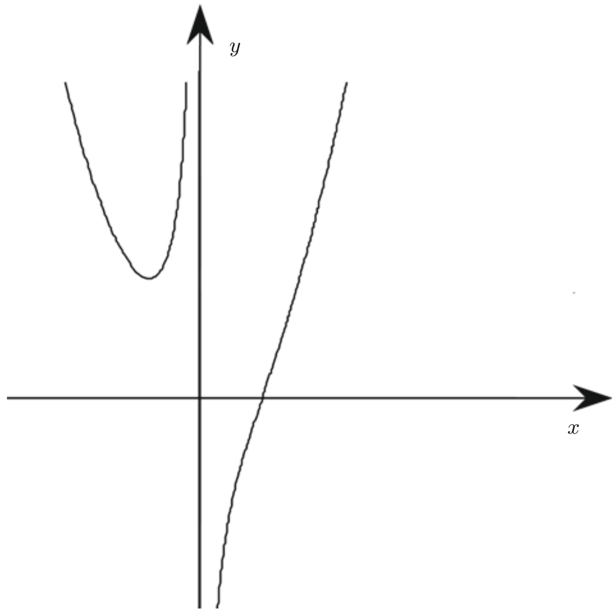


Fig. 6.8 The Trident with $A = 1$, $B = C = 0$, $D = -1$

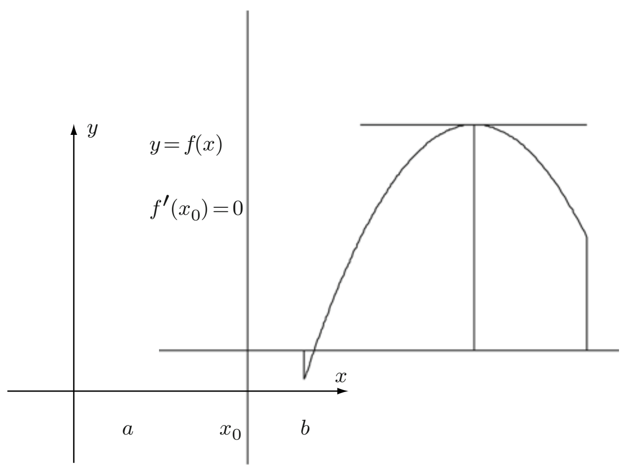


Fig. 6.9 Rolle's theorem

$$a_{1,1}a_{2,2} - a_{1,2}a_{2,1} \neq 0$$

If so, then the solution is

$$x_1 = \frac{b_1a_{2,2} - b_2a_{1,2}}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}}, x_2 = \frac{a_{1,1}b_2 - a_{2,1}b_1}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}}$$

Fig. 6.10 Collin Maclaurin

If $a_{1,1}a_{2,2} - a_{1,2}a_{2,1} = 0$ *then the system may have infinitely many solutions, or none. If* $b_1 = b_2 = 0$*, then there are infinitely many solutions in this case.*

We shall not give a separate proof for this special case. The conventional proof becomes quite messy even in this very simple special case. Instead, we give a proof valid in complete generality, which is very simple and conceptual. It requires no sophisticated abstract reasoning, and it lays to rest the common misunderstanding that a proof of Cramer's Theorem requires the use of the inverse matrix. Strangely it does not seem to appear in the standard textbooks, but we have used it in [32].

The number $\Delta = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$ is called *the determinant* of the *matrix* of the system of equations. The number is denoted by

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix}$$

In general any such $n \times n$ arrangement of numbers, which is called an $n \times n$ *matrix*, may be assigned a number called its *determinant*:

$$\det \left\{ \begin{matrix} a_{1,1} \ \dots \ a_{1,n} \\ \dots \\ a_{n,1} \ \dots \ a_{n,n} \end{matrix} \right\} = \begin{vmatrix} a_{1,1} \ \dots \ a_{1,n} \\ \dots \\ a_{n,1} \ \dots \ a_{n,n} \end{vmatrix}$$

An important feature of the determinant is its behavior under elementary row operations in the matrix:

(1) If all the numbers on one of the rows are multiplied by the same number r, then the determinant gets multiplied by that number.
(2) The determinant changes sign when two rows are interchanged.
(3) If a row is multiplied by some number and added to another row, the determinant is unchanged.

In addition we need another rule:

Rule of the diagonal. If all the numbers under the diagonal are zero, then the determinant is the product of all the numbers on the diagonal.

Elementary row operations may be performed in any $m \times n$ matrix

$$A = \begin{Bmatrix} a_{1,1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,n} \\ & \dots & \\ a_{m,1} & \dots & a_{m,n} \end{Bmatrix}$$

By a finite number of the above operations (1), (2) and (3), where the numbers r used in (1) are all non-zero, any matrix may be brought on the form shown in Fig. 6.11. Such operations are called *elementary row operations*. We then say the matrix A has been brought on *reduced row echelon form*.

The procedure is the following: If the matrix A consists of only zeroes, it is already on the required form. Otherwise, we select the first column in A which does not consist of all zeroes, by interchanging rows we may assume that this non zero number lies in the first row. Dividing the first row by that number, we may assume that it is a 1, usually referred to as a "leading 1".

We then modify the matrix according to Rule 3, subtracting suitable multiples of the first row from the second, third and so on. We thus finally get zeroes under the leading 1 of the first row.

Looking away from the first row, we then treat the remaining part of the matrix in the same way, getting the next box, also containing a leading 1. Subtracting a multiple of the second row from the first, we ensure that there is a zero directly above it. Repeating this a finite number of times, we get a result of the type shown in Fig. 6.11.

Now consider a general system of equations

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \ldots + a_{2,n}x_n &= b_2 \\ \\ \dots \\ a_{m,1}x_1 + a_{m,2}x_2 + \ldots + a_{m,n}x_n &= b_m \end{aligned}$$

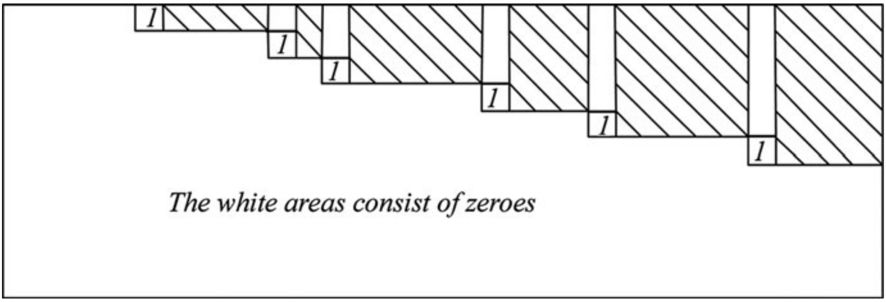


Fig. 6.11 A matrix on reduced row echelon form. The white area consists of only zeroes, and the shaded area may contain any numbers, depending on the original matrix A, of course. The so called leading 1's are shown, their positions are called the pivot positions

It is clear that if we perform elementary row operations in the matrix

$$T = \left\{ \begin{matrix} a_{1,1} & \dots & a_{1,n} & b_1 \\ a_{2,1} & \dots & a_{2,n} & b_2 \\ & \dots & & \\ a_{m,1} & \dots & a_{m,n} & b_m \end{matrix} \right\}$$

then we get a system of equations which is *equivalent to the original one.* Thus we may bring the above matrix on reduced row echelon form, and then it is much simpler to analyze the system. This method is referred to as *Gaussian elimination.* We use this important principle to prove the following:

Theorem 6 (Cramer's theorem). *Given a system of n linear equations in n unknowns*

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,n}x_n &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,n}x_n &= b_2 \\ &\dots \\ a_{n,1}x_1 + a_{n,2}x_2 + \dots + a_{n,n}x_n &= b_n \end{aligned}$$

and put

$$A = \left\{ \begin{matrix} a_{1,1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,n} \\ & \dots & \\ a_{n,1} & \dots & a_{n,n} \end{matrix} \right\}$$

Moreover, let A_i be the matrix obtained by replacing column number i in A with the $b_1, b_2, \dots, b_n$. Let $a = \det(A)$ and $a_i = \det(A_i)$. Then

(i) The system has a unique solution if and only if $a \neq 0$.
(ii) In that case, the solution is

$$x_1 = \frac{a_1}{a}, x_2 = \frac{a_2}{a}, \dots, x_n = \frac{a_n}{a}$$

Proof. (i) Denote the reduced row echelon form of

$$T = \begin{Bmatrix} a_{1,1} & \dots & a_{1,n} & b_1 \\ a_{2,1} & \dots & a_{2,n} & b_2 \\ & \dots & & \\ a_{n,1} & \dots & a_{n,n} & b_n \end{Bmatrix}$$

by

$$\overline{T} = \begin{Bmatrix} a'_{1,1} & \dots & a'_{1,n} & b'_1 \\ a'_{2,1} & \dots & a'_{2,n} & b'_2 \\ & \dots & & \\ a'_{n,1} & \dots & a'_{n,n} & b'_n \end{Bmatrix} = \begin{Bmatrix} & b'_1 \\ & b'_2 \\ A' & \vdots \\ & b'_n \end{Bmatrix}$$

The system has a unique solution if and only if $\overline{T}$ is equal to

$$T' = \begin{Bmatrix} 1 & 0 & \dots & 0 & b'_1 \\ 0 & 1 & \dots & 0 & b'_2 \\ & & \dots & & \\ 0 & 0 & \dots & 1 & b'_n \end{Bmatrix}$$

since otherwise there would be an equation saying $0 = b'_n$, either being a contradiction or an empty condition leading to an infinite number of solutions.

But it follows from the rules for the determinant quoted above that $\overline{T}$ is of this form if and only if $a \neq 0$. We leave this simple verification as Exercise 6.1.

(ii) Each elementary row operation in T gives the same operation in A and all the A_i's. Thus all $\frac{a_i}{a}$ are unchanged at each step which transforms T into T'.

So to check the formulas we may assume that the system is such that T is on reduced row echelon form. This is left as Exercise 6.1 □

In general the number ρ of 1's in the reduced row echelon form of a matrix A is called its *rank*.

6.8 Some Features of Modern Geometry

In modern times the higher geometry has developed rapidly, and split up into several different subfields. The difference between them may be in terms of the geometric contents, or it may be in terms of the methods employed. One of the features they share, is the central position occupied by geometric objects of *higher dimensions*, in addition to classical and reasonably familiar *curves* and *surfaces*. To the left Fig. 6.12 we show the familiar *hyperbola*, given as the set of all points in the plane $\mathbb{R}^2$ satisfying the equation $x^2 - y^2 = 1$. To the right we look at a somewhat less familiar case, namely a hyperbola where the equation has a *cross term*, where the product xy occurs in the equation. This corresponds to the curve being *tilted* relative to the coordinate system.

Algebraic surfaces in the Euclidian space, in $\mathbb{R}^3$, are given by equations in x, y and z. In Fig. 6.13 we show the surface defined by the equation

$$x^2 + y^2 - z^2 - 1 = 0,$$

which belongs to the family of *conic surfaces* in $\mathbb{R}^3$, that is to say surfaces given in $\mathbb{R}^3$ by an equation of degree 2.

Another geometric object, familiar to some readers, is shown in Fig. 6.13. The surface given by the equation $x^2+y^2-z^2-1 = 0$, which is a rotational hyperboloid. We have omitted the coordinate axes here.

The higher dimensional geometric objects are referred to by names like *manifolds*, *varieties* or *schemes*, and are the objects of study within the *Differential Geometry*, the *Analytic Geometry* or the *Algebraic Geometry*. The difference between these three branches of geometry basically lies in the *functions*, in several variables, used to describe the geometric objects in question. We illustrate the

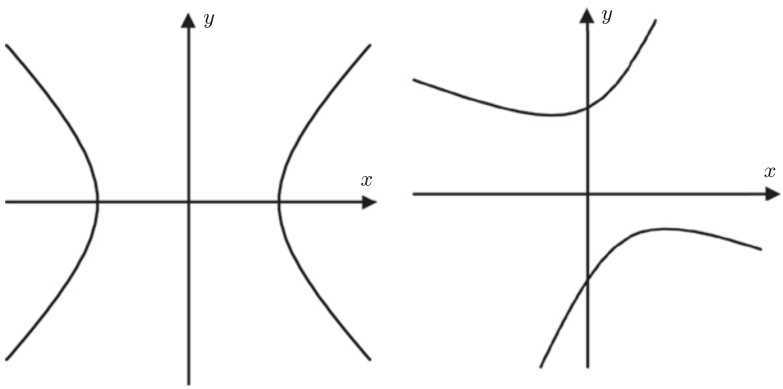


Fig. 6.12 A familiar curve to the *left*, a somewhat less familiar one to the *right*

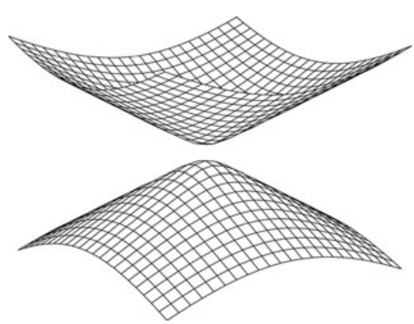


Fig. 6.13 A surface

point by loosely explaining what an $\mathbb{R}$-variety or as we also say, a *real algebraic variety*, is.

The simplest case is that of an *affine algebraic variety*. This is the zero locus in $\mathbb{R}^n$ of a finite set of polynomials in n variables, $f_i(X_1, \ldots, X_n), i = 1, \ldots, r$. For the complex numbers instead of the reals, affine algebraic $\mathbb{C}$-varieties are defined similarly as subsets of $\mathbb{C}^n$. We talk about affine algebraic varieties over $\mathbb{R}$ or over $\mathbb{C}$. For such varieties we have a good concept of *dimension*. It will take us too far to provide a detailed discussion, but the dimension then is an integer ≥ 0, dimension equal to zero corresponding to points, dimension 1 to curves and dimension 2 to surfaces.

In general and roughly speaking, a real algebraic variety X is glued together by pieces which are affine varieties, $V_\alpha, \alpha \in A$, A denoting some indexing set. So the pieces V are as given below:

$$V = \{(a_1, \ldots, a_n) \in \mathbb{R}^n | \ f_i(a_1, \ldots, a_n) = 0, i = 1, \ldots, r\}$$

There are more technical points in the definition, but the main idea is captured by the above. The individual pieces referred to above are called *affine open subsets* of X. The local dimension of X at a point x is the dimension of a sufficiently small affine piece of X containing x, and if the local dimension is constant throughout X, this number is called the *dimension* of the algebraic variety X.

In *Differential Geometry* the objects under study are defined similarly, but the functions are no longer polynomials but functions in n variables for which *all partial derivatives*, mixed and to any order, exist.

In *Analytic Geometry* this condition is replaced by the condition that all the f_i's possess power series expansions.

Note that any variety or manifold over $\mathbb{C}$ also is one over $\mathbb{R}$, but the dimension as an $\mathbb{R}$-variety is twice that as a $\mathbb{C}$-variety.

We leave the matter of explaining what varieties are at this point. But general as it is, we are only at a timid beginning. The theory of *Grothendieck schemes* is even more general, the definition of these objects will not be given here. But with this theory we have come full circle. The basic reference for this is [22]. Now it is no longer a matter of having introduced algebra into geometry, but perhaps of introducing geometry into algebra: The theory of Grothendieck schemes represents a single, powerful common generalization of *algebra and geometry*. And then, beyond the schemes of Grothendieck there is *non-commutative* algebraic geometry...

The obvious question of *why we want to study such objects* may be given several answers. Basically, one answer would be that we study them for the same reason as the Greeks studied circles and straight lines: They encountered them when dealing with the tasks undertaken in their science and technology. So do we, with these objects. In fact, to take a comparatively recent development: Physicists consider *models* for the physical universe where the space in which we live is not 3-dimensional, but of a considerably higher dimension. Only three of these dimensions are noticeable in our life, the remaining ones being *curled up* in *very small* cylinders, or something like that. The diameter of these cylinders would be so small as to fit well inside the nucleus of the atoms. Thus a *point* would actually be a *very small* but higher dimensional, *subspace* of the space in which we live. The idea is not as preposterous as one might think at first encounter: A point is a *zero-dimensional* subspace of the usual Euclidian space, in our conventional way of thinking, a line is a one-dimensional subspace and a plane a two-dimensional one. *"Fattening"* the usual space by some extra, curled up, dimensions, would fatten up the points, lines and planes correspondingly. A point in our usual space would then have to be understood as the manifestation, the section, of a higher-dimensional *"point-manifold"*.

Modern geometry studies geometric objects which are very difficult to visualize. Even capturing them as some sort of subspaces of simple spaces like $\mathbb{R}^n$, $\mathbb{C}^n$, or the corresponding projective spaces which will be explained in Chap. 12 may not be possible. But it should be noted that if we have r polynomials $f_1(X_1, \dots, X_n), \dots, f_r(X_1, \dots, X_n)$ with real coefficients, then we obtain an algebraic variety consisting of only one affine piece, namely the one given by the zero-locus of the given polynomials in $\mathbb{R}^n$. Such algebraic varieties are, as we have explained, called *affine* algebraic varieties, and as we shall explain in Chap. 12, we may similarly define *projective algebraic varieties* as well, contained in projective spaces.

An important phenomenon studied extensively in modern geometry is that of *singular points*. If X is an $\mathbb{R}$-variety in the sense loosely explained above, then it turns out that for most points $x \in X$ there is a small neighborhood of x which may be identified with the unit ball in $\mathbb{R}^n$, by modifications without breaking or merging. These points are called *smooth points*. Points which are not smooth are called *singular points*. A simple example of a singular point is the origin in the plot of the curve in $\mathbb{R}^2$ with equation

$$y^3 + y^2 - x^2 = 0,$$

shown to the left in Fig. 6.14.

To the left the origin is a singular point, since any small neighborhood would look like two crossed line segments, which cannot be deformed into a single line segment without breaking or merging.

Another example a singularity is provided by the *vertex of a cone*, shown to the right in Fig. 6.14. The vertex of a cone is a singular point. If the vertex is the origin, and the axis is along the x-axis, the equation for it as a subset of affine 3-space is $y^2 + z^2 - R^2x^2 = 0$, where R is a constant. Here we see that around any other point of the cone than the vertex, we may take a small disk which can be identified with the unit disc around the origin in the plane. For the vertex, however, this is not possible.

We also have surfaces where there is an entire curve of singularities, as in the example shown in Fig. 6.15.

Today physicists study some interesting objects called *black holes* . Black holes started out as being highly hypothetical, but are becoming more and more part of what we know to be the reality we live in. A black hole is created as the density of a collapsing star becomes so immense that the generated gravitational force near its

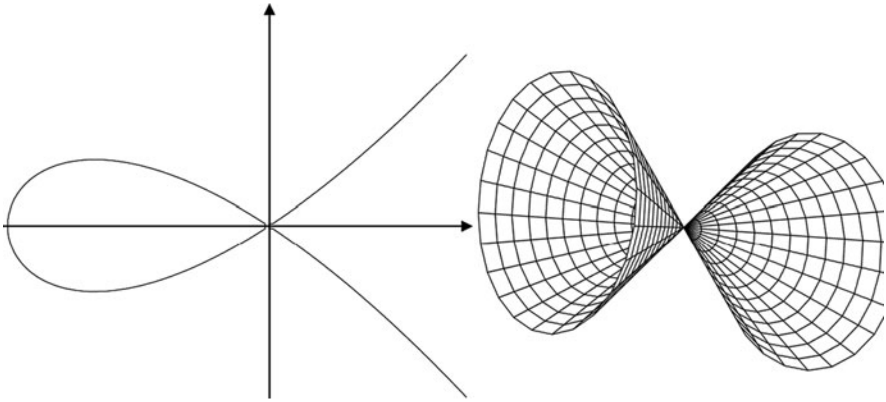
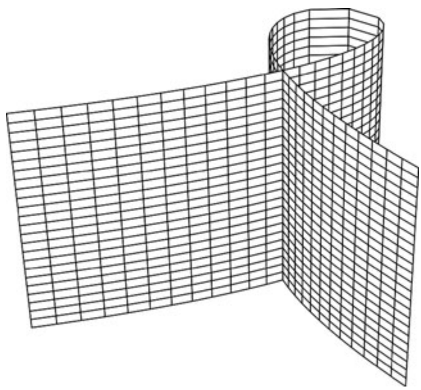


Fig. 6.14 A singular curve to the *left*, a cone to the *right*

Fig. 6.15 The cylinder over the curve, shown in Fig. 6.14, with a singular point at the origin yields a surface, with an entire curve of singular points



surface makes it impossible for light to escape: The object then becomes invisible, and anything coming within a certain boundary, including light, will inevitably be sucked into it. Black holes may be understood as singularities of space itself. Matter has been merged with geometry to become part of the very fabric of space.

Another modern application of geometry is the so-called *Catastrophe Theory.* We treat this theme in some detail in Sect. 7.2. Catastrophe Theory was very much on the public agenda some years ago, but the interest in it has since abated somewhat. However, the field still exists and very much also still holds the promise of becoming an effective tool in predicting dramatic changes in important systems we depend on. New insights have already been gained, one such insight being a renewed awareness of how misleading an old *dogma* of natural science really is: Namely the assertion that *Nature non facit saltus*: Nature does not make jumps, natural phenomena always change continuously. At the micro-level this dogma has been out for some time, contradicting as it does the very fundament of quantum theory. But a chilling realization is that at the macroscopic level nature may indeed perform jumps as well. Thus in particular, the *global warming* we are now experiencing, or so some claim, might not just smoothly move our global climate from where we have been to some new slightly different state, but the climate may be thrown suddenly into a new mode. And not only that: Even if we succeed in reducing the greenhouse effect again, the climate may remain in the new state for a long, long time. We shall return to these ideas in the light of the geometry of Catastrophe Theory in Sect. 7.2.

The concept of dimension is more interesting and more subtle than our definition above would suggest. In fact, there are geometric objects where the original definition of dimension, due to Felix Hausdorff, yields a number which is not an integer. Such objects, having a *fractional dimension*, are referred to as *fractals*. The *Theory of fractals* has received considerable attention during recent years, and one might say captured the limelight from Catastrophe Theory. Fractals are, according to some of the proponents of this field, to be found everywhere in nature. A shoreline could be viewed as, or *modelled as* a fractal: Looking at it from high above, it would give one appearance, but as the observer descends and moves closer to the surface, its features change. In the end the observer is looking at small pebbles and grains of sand, and the "curve" which was the shoreline is no longer that, instead it has dissolved into something between a curve and a strip of surface: It suggests a dimension less than 2 but more than 1. Some fractal theorists assert that the typical shoreline would have a dimension of about 1.2. Similarly *clouds* would be objects with an estimated dimension of more than 2 but less than 3, most being deemed to be of dimension around 2.3. Computing or estimating dimensions in this way is somewhat controversial. Certain estimates on the degree of *"self similarity"* of the object under examination have to be made, and these estimates are certainly not above discussion. We give details in Chap. 18.

6.9 Archimedean Polyhedra and Tessellations

In Sect. 3.10 we have seen the five regular, or Platonic, polyhedra. In Sect. 4.4 we have given a preliminary treatment of the semiregular, or *Archimedean*, polyhedra. We now return to these fascinating objects, and also explore the rich world of *tilings* or *tessellations*. They are closely related to the polyhedra, as we shall see. An important source for the treatment of polyhedra in this section is the very readable and interesting book [9], which we highly recommend.

For a general polyhedron we associate a tuple of integers to each vertex, generalizing what we did in Sect. 4.4. Thus the vertex v of the general polyhedron P is given the tuple $(m_1, m_2, \ldots, m_n)$ if n (not necessarily regular) polygons of orders $m_1, m_2, \ldots, m_n$ meet at v. If all polyhedral angles are congruent this invariant is very useful, as it is then the same at all vertices.

From now on we assume that the polyhedron P has sides which are regular polygons, that it is convex and that all the polyhedral angles are congruent. As we have seen from Miller's Polyhedron this does not imply that P is semiregular. We also include, and treat simultaneously, the limiting case that all the polygons lie in that same plane, i.e., that we have a tessellation of the (Euclidian) plane. The condition that the tessellation be *vertex transitive* is kept as for the Archimedean polyhedra.

We now deduce conditions which the invariant $(m_1, m_2, \ldots, m_n)$ must satisfy.

The angular sum of a regular m-gon is $(m-2)\pi$, thus the angle contained by adjacent sides is $\varphi_m = (1 - \frac{2}{m})\pi$. Assume that a semi-regular convex polyhedron is of type $(m_1, m_2, \ldots, m_n)$, i.e., that at each vertex n regular polygons of orders $m_1, m_2, \ldots, m_n$ meet. This sequence is of course cyclic, e.g., $(m_2, m_3, \ldots, m_1)$ gives the same polyhedron. We then have the *Polyhedral Angle Inequality*, for short just the polyhedral inequality:

$$\varphi_{m_1} + \varphi_{m_2} + \cdots + \varphi_{m_n} \leq 2\pi,$$

with equality if and only if the "polyhedron" is actually a plane tessellation. By the expression for φ_m above, the Polyhedral Angle Inequality is equivalent to

$$\frac{1}{m_1} + \frac{1}{m_2} + \cdots + \frac{1}{m_n} \geq \frac{n}{2} - 1$$

with equality if and only if we have a tessellation. We also have the obvious condition that

$$m_i \geq 3$$

for all $i = 1, 2, \ldots, n$.

Moreover, since

$$\varphi_{m_i} = \left(1 - \frac{2}{m_i}\right)\pi \geq \frac{\pi}{3}$$

we get

$$\varphi_{m_1} + \varphi_{m_2} + \cdots + \varphi_{m_n} \geq n\frac{\pi}{3}$$

and so $n\frac{\pi}{3} \leq 2\pi$, hence

$$3 \leq n \leq 6$$

the left inequality being obvious.

In Sect. 3.10 we found the five possibilities for regular polyhedra. To find the *regular tessellations*, that is the tessellations where all polygons are regular of the same kind, with n-gons, we get the following: Let φ be the angle at each vertex. The sum of the angles contained by adjacent sides is $(n-2)\pi$, thus $\varphi = \frac{n-2}{n}\pi$. On the other hand the sum of the angles constituting the polyhedral angle must be $m\varphi$, m being the number of edges meeting at each vertex. Thus we have for a plane regular tessellation

$$m\left(\frac{n-2}{n}\right)\pi = 2\pi$$

and so

$$m(n-2) = 2n.$$

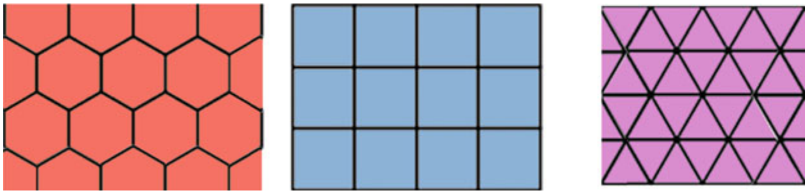


Fig. 6.16 The regular tessellations

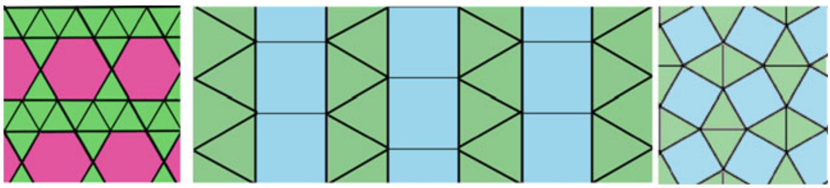


Fig. 6.17 The Archimedean tessellation (3,3,3,3,6) *left*, (3,3,3,4,4) in the *middle* and (3,3,4,3,4) *right*

For $n = 3$ this gives $m = 6$, $n = 4$ gives $m = 4$ and $n = 5$ is impossible while $n = 6$ gives $m = 3$. We then have the tessellations shown in Fig. 6.16, obviously all vertex transitive.

We are now going to list the possibilities for $m_1, \ldots, m_n$ yielding Archimedean polyhedra, when $n = 3, 4, 5, 6$. For $n = 6$ we have $\frac{n}{2} - 1 = 2$, thus the polyhedral inequality holds only for $m_1 = \cdots = m_6 = 3$, which yields equality and thus a tessellation, which we see to the right in Fig. 6.16. This also shows that if $m_1 = \cdots = m_5 = 3$ then the only possibility for m_6 is 3, thus we are finished with $n = 6$.

For $n = 5$ we find the following tuples $(m_1, \ldots, m_5)$, arranged in increasing lexicographic order, which satisfy the polyhedral inequality:

m_1	m_2	m_3	m_4	m_5	$\frac{1}{m_1} + \cdots + \frac{1}{m_5}$	Yields
3	3	3	3	3	$\frac{5}{3} > \frac{n}{2} - 1 = \frac{3}{2}$	Icosahedron
3	3	3	3	4	$\frac{19}{12} > \frac{n}{2} - 1 = \frac{3}{2}$	Snub cube
3	3	3	3	5	$\frac{23}{15} > \frac{n}{2} - 1 = \frac{3}{2}$	Snub dodecahedron
3	3	3	3	6	$\frac{3}{2} = \frac{n}{2} - 1 = \frac{3}{2}$	Archimedean tessellation
3	3	3	4	4	$\frac{3}{2} = \frac{n}{2} - 1 = \frac{3}{2}$	Archimedean tessellation

For all but the last one changing the order just gives a cyclic permutation (rearrangement), and therefore the same polyhedron or tessellation. As for the last one, a rearrangement yields (3,3,4,3,4), which also yields an Archimedean tessellation. The polyhedra are shown on Fig. 4.12, the tessellations we have found are shown in Fig. 6.17 with (3,3,3,3,6) to the left, (3,3,3,4,4) in the middle and (3,3,4,3,4) to the right.

They are all "vertex transitive" in the sense that for any two corners there exists a transformation (a symmetry) of the tessellation onto itself such that the transformed

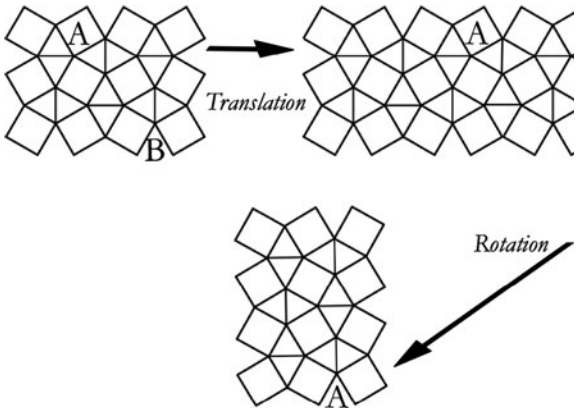


Fig. 6.18 A translation followed by a rotation provides the symmetry transformation which carries A to B in such a way that the tessellation is unchanged

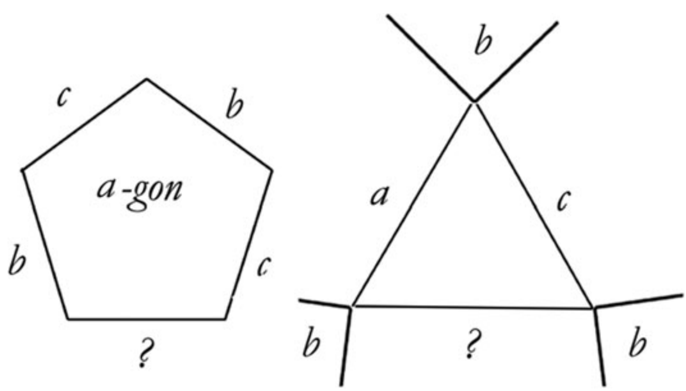
tessellation looks exactly the same as it did before. A completely satisfactory treatment from a formal point of view would require a little bit more group theory than we are prepared to go into here, but the idea is illuminated by the illustration on Fig. 6.18.

Following [9] we now prove an important lemma, first shown by Kepler:

Lemma 1. *A polyhedron or tessellation where all vertices are surrounded in the same way cannot have polyhedral angles of the following types:*

(i) (a, b, c) *where a is odd and* $b \neq c$.
(ii) $(3, a, b, c)$ *where* $a \neq c$.

Proof. (i) Consider an a-gon in the polyhedron or tessellation, then b-gons and c-gons will alternate around it as shown to the left below.



Evidently this is impossible unless $b = c$, a being odd. As for (ii), there is always a b-gon opposite the 3- gon. Thus we have the situation to the right, which is impossible unless $a = c$. □

We next find the numbers $3 \leq m_1 \leq \cdots \leq m_4$ which satisfy the polyhedral inequality, and list the corresponding tuples $(m_1, \ldots, m_4)$ with $m_1 \leq m_2 \leq m_3 \leq m_4$ in lexicographical order. After that we have to find the rearrangements which will yield polyhedra or tessellations.

First of all, the case $(3, 3, 3, n)$ corresponds to an n-gonal antiprism, the snub cube or the octahedron. Thus we need only consider tuples lexicographically $\geq (3, 3, 4, 4)$. Moreover, according to part (ii) of the lemma, a legal (unordered) triple of integers $3, 3, m_3, m_4$ in the range we are now examining must have $3 < m_3 = m_4$. We start the list with those three values which satisfy the polyhedral inequality. Next, we find three more cases starting with 3,4,4. Finally there is just one case starting with 4,4,4.

m_1	m_2	m_3	m_4	$\frac{1}{m_1} + \cdots + \frac{1}{m_4}$
3	3	4	4	$\frac{2}{3} + \frac{1}{2} > \frac{n}{2} - 1 = 1$
3	3	5	5	$\frac{2}{3} + \frac{2}{5} > \frac{n}{2} - 1 = 1$
3	3	6	6	$\frac{2}{3} + \frac{2}{6} = \frac{n}{2} - 1 = 1$
3	4	4	4	$\frac{1}{3} + \frac{3}{4} > \frac{n}{2} - 1 = 1$
3	4	4	5	$\frac{1}{3} + \frac{2}{4} + \frac{1}{5} > \frac{n}{2} - 1 = 1$
3	4	4	6	$\frac{1}{3} + \frac{2}{4} + \frac{1}{6} = \frac{n}{2} - 1 = 1$
4	4	4	4	$\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{n}{2} - 1 = 1$

We next look for rearrangements to produce polyhedra or tessellations. $(3, 3, n, n)$ is not possible for $n \geq 4$ by part (ii) of the lemma, but (3,4,3,4) yields the cuboctahedron. Similarly (3,5,3,5) yields the icosidodecahedron, (3,6,3,6) yields an Archimedean tessellation, shown in Fig. 6.19. (3,4,4,4) gives the small rhombicuboctahedron. The unordered integers $3, 4, 4, n$ are not possible for $n > 6$, but (3,4,5,4) gives the small rhombicosidodecahedron and (3,4,6,4) gives an Archimedean tessellations shown in Fig. 6.19.

We finally turn to the case $n = 3$. (3,3,3) gives the tetrahedron. By part (i) of the lemma $n_1 = 3$ implies that $m_2 = m_3 = m$. Then $\frac{1}{3} + \frac{2}{m} \geq \frac{3}{2} - 1 = \frac{1}{2}$, thus $m \leq 12$. Also the case of m being odd is excluded by the same lemma. We thus get the first new possibilities:

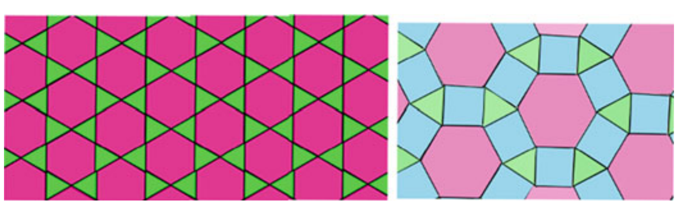


Fig. 6.19 The Archimedean tessellation (3,6,3,6) *left* and (3,4,6,4) *right*

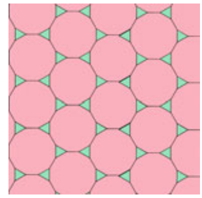


Fig. 6.20 The Archimedean tessellation (3,12,12)

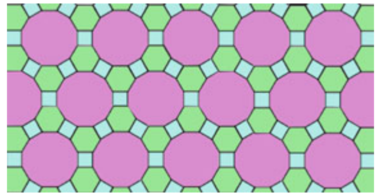


Fig. 6.21 The Archimedean tessellation (4,6,12)

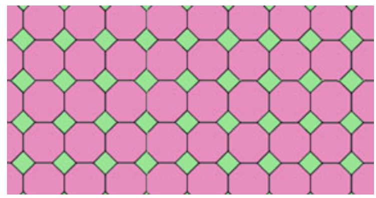


Fig. 6.22 The Archimedean tessellation (4,8,8)

m_1	m_2	m_3	$\frac{1}{m_1}+\frac{1}{m_2}+\frac{1}{m_3}$
3	4	4	$\frac{1}{3}+\frac{2}{4}>\frac{n}{2}-1=\frac{1}{2}$
3	6	6	$\frac{1}{3}+\frac{2}{6}>\frac{n}{2}-1=\frac{1}{2}$
3	8	8	$\frac{1}{3}+\frac{2}{8}>\frac{n}{2}-1=\frac{1}{2}$
3	10	10	$\frac{1}{3}+\frac{2}{10}>\frac{n}{2}-1=\frac{1}{2}$
3	12	12	$\frac{1}{3}+\frac{2}{12}=\frac{n}{2}-1=\frac{1}{2}$

Here (3,4,4) is a triangular prism, (3,6,6) gives the truncated tetrahedron, (3,8,8) the truncated cube, (3,10,10) the truncated dodecahedron and finally (3,12,12) an Archimedean tessellation, shown in Fig. 6.20.

Next, take $m_1 = 4$. Here $(4, 4, n)$ gives the n-gonal prism. For the next possibility we must have m_2 even, by part (i) of the lemma. Thus next possibility is $(4, 6, 6)$, which yields the truncated octahedron. As (4,6,7) is impossible, the next in line is (4,6,8), which corresponds to the great rhombicuboctahedron. (4,6,9) is impossible, so the next, i.e., (4,6,10), it corresponds to the great rhombicosidodecahedron. For $(4, 6, n)$ the polyhedral inequality $\frac{1}{4}+\frac{1}{6}+\frac{1}{m} \geq \frac{1}{2}$ gives $m \leq 12$. $m = 11$ is impossible, and $m = 12$ gives the Archimedean tessellation shown in Fig. 6.21.

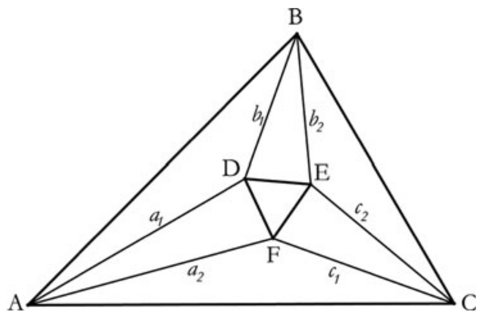
The next possibility is (4,8,8), which yields the Archimedean tessellation shown in Fig. 6.22.

We next examine the case of $m_1 = 5$. We now note that any tuple (m_1, m_2, m_3) where there are two different odd numbers, is excluded by the lemma. The first new case to consider is (5,5,5), which gives the dodecahedron. The next case is (5,6,6), which gives the truncated icosahedron. Then comes (5,6,8), which however does not satisfy the polyhedral inequality. The next case is (6,6,6), which yields e regular tessellation, shown in Fig. 6.16. The next case would be (7,7,7), which does not satisfy the polyhedral inequality. This exhausts all possibilities.

Exercises

Exercise 6.1 In the proof of Theorem 6 of Sect. 6.7, verify the assertion that $\overline{T} = T'$ if and only if $a \neq 0$. Then verify Cramer's Theorem when the matric T is on reduced row echelon form.

Exercise 6.2 (Morley's Miracle) The following result is quite striking.



It was discovered in 1899 by *Frank Morley,* professor at Haverford College. Today it is referred to as *Morley's Miracle:* Trisect the angles of any triangle, and take the three points of intersection between corresponding trisectors as shown in the figure above. These three points always form an equilateral triangle. Use the Rule of Sines and the Law of Cosines to prove this theorem. Search the internet for different proofs. How many can you find?

Exercise 6.3 Let *ABCD* be any quadrilateral with sides a, b, c and d. Let the angle φ be half the sum of any pair of opposite angles. Show the "Generalized Brahmagupta formula" for the area T of *ABCD*:

$$T = \sqrt{(s-a)(s-b)(s-c)(s-d) - abcd\cos^2(\varphi)}$$

where again s is half the sum of all sides.

Exercise 6.4 (Orthocenter) An altitude of a triangle is a straight line through a vertex and perpendicular to the opposite side or its extension. Show that the three altitudes of any triangle meet in one point. This point is called the *orthocenter* of the triangle.

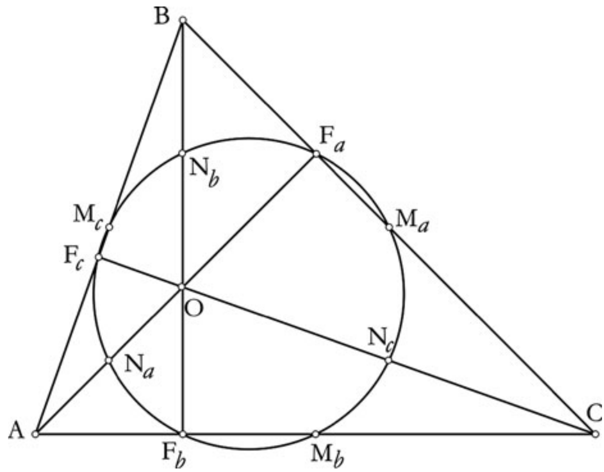
Exercise 6.5 (Barycenter) A median of a triangle is a straight line through a vertex which bisects the opposite side. Show that the three medians of any triangle meet in one point. This point is called the *centroid* or the *barycenter* of the triangle.

Exercise 6.6 (Angle Bisectors) Show that the three lines through the vertices of any triangle $\triangle ABC$ which bisect the corresponding angles meet in one point. This is the center of the inscribed circle in $\triangle ABC$, referred to as the *incenter* of $\triangle ABC$.

Exercise 6.7 (Gergonne Point) *Let D, E and F be the points where the inscribed circle touches the sides of $\triangle ABC$. Prove that the lines AD, BE and CF intersect in one point G. G is referred to as the* Gergonne Point*, named after Joseph Diaz Gergonne (1771–1859).*

Exercise 6.8 (Symmedians) The reflection of the medians in the angle bisectors are called the *symmedians* of the triangle. Show that the symmedians meet in one point. This point is called the *Symmedian Point* or the *Lemoine Point*, named after *Emile Michel Hyacinthe Lemoine,* 1840–1912.

Exercise 6.9 (Nine Point Circle) The following nine points lie on the same circle: The three midpoints M_a, M_b, M_c of the sides a, b, c, the three foot points of the altitudes F_a, F_b, F_c and the three mid points N_a, N_b, N_c of the segments of the altitudes from the vertex to the orthocenter O of the triangle.



Chapter 7
Geometry and the Real World

7.1 Mathematics and Predicting Catastrophes

Mathematics is important in understanding nature. By mathematics we create *models*, which provide explanations for the phenomena we observe. If a mathematical model yields a result which is in contradiction to the observations we make, then the theory will have to be scrapped, no matter how beautiful the mathematics in it should be. But can mathematics guide us in finding new knowledge about nature itself? In other words, not just arrange the knowledge we already have in a nice and orderly model, but actually *predict observations* which we have not yet made? The answer is, of course, affirmative. In fact this phenomenon is the reason for the usefulness of working with models in the first place.

A classic example is the discovery of the planet Neptune. In 1843 *John Adams* used certain irregularities in the orbits of the known planets to predict the existence of an unknown planet outside the orbit of the planet Uranus. He even gave the exact coordinates where the telescope should be aimed to find this new planet. But his computations were not taken seriously, and it was not until 1846 when *Urbaine le Verrier* arrived at the same result that the astronomers bothered to look. And there it was!

Later the planet Pluto has been found in a similar manner. Some years ago new claims were made concerning a further planet, far beyond the orbit of Pluto. Some even claim that the Sun may have a dark companion-star, tentatively named *"Nemesis"*. This ominous name is due to speculations concerning occurrences of periodic *mass extinctions*, documented in the fissile record.

During one such mass extinction, the *Cretaceous-Tertiary mass extinction*, the dinosaurs died out in a, geologically speaking, very short time-interval. This happened 65 million years ago. The leading theory explaining that catastrophic event, is that the Earth was hit by an astroid of about 10 km in diameter. An impact crater, presumably being the result of this killer-astroid, is located partly off the coast of Yucatan. Other theories are also being offered, but as of today the astroid-theory seems to be the most plausible and best supported by evidence.

At any rate, the connection with the hypothetical "Nemesis" is that it would have a very eccentric orbit around the Sun, and therefore at intervals of about 60 million

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_7,

years pass through a belt of comets or asteroids. Deflecting some of these objects, the inner solar system would, at such intervals, be exposed to showers of comets or asteroids.

The Nemesis-theory has been loosing ground, however. Today there seems be general consensus among astronomers that the culprit is the planet *Jupiter*. In fact, Jupiter influences the orbits of the asteroids inhabiting the astroid belt beyond Mars. The influence happens in a way which is perfectly determined by the mathematics of planetary orbits, but which is nevertheless of such complexity that the phenomenon appears *chaotic*. Moreover, small changes from a stable situation will lead to abrupt and large changes in the orbital structures. These *chaotic instabilities* caused by the proximity of this giant planet occasionally sends an astroid off course and into the inner solar system. Such an event could happen tomorrow, or in one thousand years, or in 65 million years. It could be a very unpleasant surprise for us Earthlings!

But as it has happened many times in the past, we already have numerous asteroids with odd orbits, some of them coming near Earth from time to time.

An impact of an astroid of a diameter of 10 km would lead to gigantic tsunamis, flooding large areas, dust would be ejected into the air blocking sunlight, and leading to a global winter lasting for several years. Impact by an astroid with a diameter of 1 km would also lead to a catastrophe on a global scale. But even impact of an object with a diameter of 100 m could, under unfortunate circumstances, devastate a large modern city. To get an impression of the force involved here, it can be recalled that on June 30, 1908, 1,000 km^2 of Siberian pine forest was flattened by the blast caused by an impacting astroid. The estimate is that the explosion, which happened above ground, was equivalent to the detonation of a 10 megaton hydrogen bomb. The astroid is estimated to have had a diameter of merely 70 m.

This scenario is taken seriously. Enough so as to having the watch for near-Earth asteroids being one of the activities of NASA. But the program is to map all asteroids in near-Earth orbit – NEOs as they are known – of more than 1 km diameter. The mapping moves along very slowly, one estimate being that it will take 20 years to have mapped 90% of all NEOs of this size.

These insights contributed to alerting the public to the possibility of a *nuclear winter* which could follow an all-out nuclear war, and so has served a usefully purpose already. In relatively recent times a volcanic eruption is known to have caused conditions with a taste of this kind of calamity. It happened in 1815, when the volcano *Gunung Tanbore* at the island of Sumbawa in present day Indonesia erupted in the period between April 7 and April 12, the largest volcanic eruption known in historic times. More than 30 km^3 of dust, ashes and stone was flung into the atmosphere, the plume extending 44 km up in the air. An estimated 90,000 people died in the area, and that as well as the following year the summer was extremely bad over the entire northern hemisphere. Famine resulted in Europe, and in New England 1816 was known as *"the year without summer"*.

Climatic change is not only threatening in the form of the "nuclear winter" scenario. The insidious process known as *global warming*, caused by the increased emission of CO_2 as well as other *greenhouse gases* might be leading us into a situation where the ice is starting to melt in the polar regions, where the oceans are heating up, storing immense additional quantities of energy, giving rise to a variety of troubling and threatening scenarios.

Thus it not surprising that the prediction of volcanic eruptions and earthquakes is an important task for modern science. Again this is a question of having a model for the geological dynamics of the planet Earth, which is good enough to not only explain past events, but which is also capable of yielding reliable predictions. Such models will certainly have to rely heavily on sophisticated mathematics, including the mathematics which goes into the field some call Catastrophe Theory. But names for the same mathematical phenomena may vary, in the present case we find terms like *Bifurcation Theory* or *Stability Theory*.

Even if the chilling prospect of Earth being hit by an astroid captures our imagination, the climatic change due to global warming may constitute a greater danger. The nature of this process is different from the sudden impact, which might be prevented entirely by invoking the sophisticated technology, heroic individuals and the combined resources of a united Humankind. Less heroic is the task of arguing to convince politicians, dependent on contributions from various business groups, that measures limiting the emissions of CO_2 are needed to avoid a more or less drastic climatic change on our planet. But at least the problem and the process may be understood better by means of concepts from Catastrophe Theory. We shall give some indications of this in Sect. 7.2, and provide a more complete mathematical treatment in Chap. 19.

7.2 Catastrophe Theory

In the early 1960s the distinguished mathematician *René Thom* created the foundation of this field. Among other significant contributors are *V.I. Arnold*, *B. Malgrange*, *John Mather* and *Christopher Zeeman*.

In this book we may only touch upon this interesting field, but readers who would like to learn more are referred to the book by *P.T. Saunders* [53].

An instructive introduction to Catastrophe Theory is provided by the famous example of Zeeman dealing with *aggression* in dogs as a function of the emotions *rage* and *fear*. The example is treated in his classic Scientific American article from 1976 [65], and also explained in [53]. The basic assumption on which this example builds is that aggression in dogs is a result of the two emotions *rage* and *fear*. Rage and fear is estimated on a numerical scale by observing by how much the mouth is open (*rage*) and by how much the ears are laid back (*fear*). The aggression displayed by the animal's behavior is then estimated on a numerical scale as well, and an important observation is made: In some situations the same animal may display different levels of aggression when being under the same amount of fear and rage: The aggression produced depends, in some cases, on the previous history.

Zeeman uses the paradigm of Catastrophe Theory to model this situation. He postulates the variable z for *Aggression* being such that the point $(x, y, z) \in \mathbb{R}^3$, where x and y are the numerical estimates for *Rage* and *Fear*, respectively, lies

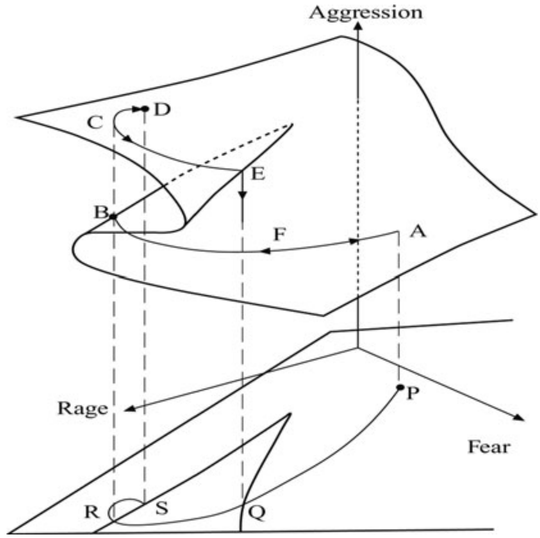


Fig. 7.1 The Cusp Catastrophe: Aggression as a result of rage and fear

on a certain fixed surface in $\mathbb{R}^3$, see Fig. 7.1. Surfaces of this shape may be given by polynomial equations of degree 3, this is explained in Sect. 19.1. There it is also explained that in the (x, y)-pane there is a wedge-shaped area, inside a certain curve known as a *semi-cubic parabola*, see Sect. 13.3, such that for a point in this area there are three points on this surface directly above it, and for $P = (x, y)$ outside the area there is a unique point on the surface above P. Thus for these values of rage and fear, the dog's behavior is perfectly predictable: The level of aggression is uniquely determined by its rage and fear.

However, for a point P *inside* the wedge shaped area there are in principle three points above P. But only two of them correspond to actual levels of aggression. This may sound strange and appear unmotivated, but it follows from the mathematics of Catastrophe Theory that this is the situation as far as the theory is concerned. Again, this will all be explained in Chap. 19. So with this model for aggression in dogs, there are two possible *modes* corresponding to rage-fear levels inside the wedge. *The middle sheet of the fold of the surface is removed.*

Now suppose that the dog starts out at the rage-fear point P, and a sequence of events moves it along the curve indicated, from P through Q and R to S. At the point Q it passes into the critical area, but nothing much happens, aggression remains about the same. The path on the surface has passed from A, through F, and proceeds to B. This corresponds to the rage-fear point R. This is where the *"Catastrophe"* occurs, in the present case, this is where the apparently quiet dog (who has, however, quietly more and more bared its teeth) suddenly tries to bite you! Aggression jumps, as it has to according to our model, to the level given by the point C. Zeeman calls this event *an attack catastrophe.* You now have to control the dog, inducing fright or reducing its rage by taking various measures. At first this does not help much. Moving back through S the dog still stays at a high level

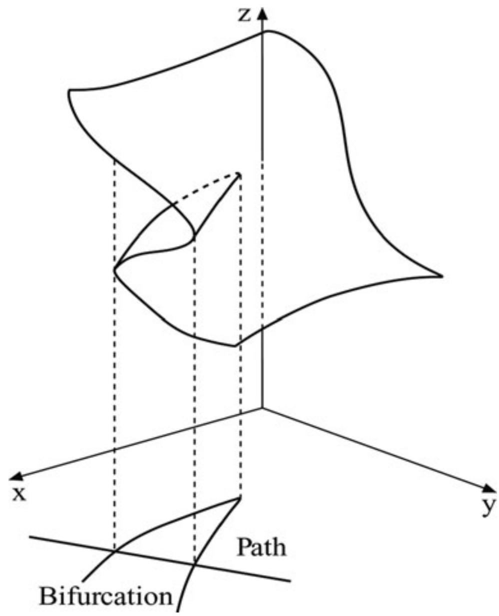


Fig. 7.2 The paradigm of the Cusp Catastrophe: z is average global temperature, x is the level of greenhouse gases in the atmosphere and y the area covered by carbon dioxide-consuming plants. As we cut down the forests and produce carbon dioxide, we move along the indicated path, from right to left

of aggression. If the path is traversed form S and back to P, then the point on the surface follows a different path, namely from D, through C, then onwards to E, where the dog suddenly ceases being aggressive, the point jumping down to F. Zeeman calls this *a flight catastrophe.* The dog's behavior now follows the original path back to A.

The singularity of the wedge shaped curve, the semi-cubic parabola, is known as a *cusp*. Correspondingly this situation is referred to as a *Cusp Catastrophe.*

Some feel that Catastrophe Theory has been somewhat oversold during the last two or three decades. Today a common view is that while the theory does not provide explanations for everything, it may provide a point of view which elucidates some of the issues on our agenda. Take for example the issue of *Global Warming*. We represent the mean global temperature, on some scale, by the variable z, we let x be the level of *greenhouse gases*, and let y denote the *area covered by carbon dioxide-consuming plants* on the Earth. Could it be that the situation may be modelled by a cusp catastrophe? And if so, could it be that we are indeed moving along the path indicated, from right to left, presently cruising smugly inside the bifurcation set? This disturbing scenario is illustrated in Fig. 7.2.

7.3 Geometric Shapes in Nature

It is a fascinating aspect of studying geometry to search for and to find geometric shapes in nature. As a source of inspiration to geometers this activity goes back to the very origin of the subject. And closely related to it is the creation of decorative

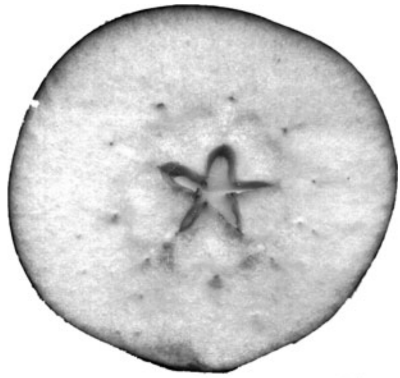


Fig. 7.3 A section through an apple, revealing a pentagram and a regular pentagon around it

Fig. 7.4 A section through a Kiwi-fruit to the *left*, section through an orange to the *right*

patterns for pottery or mosaics, often reproducing patterns observed in the natural world.

One of the simplest ways to encounter *the Golden Section* is to cut an apple along "the equator". There the kernels are arranged in a chamber, the cross section of which has an outline which is *a pentagram*. We show an example of this in Fig. 7.3.

The pentagon and the pentagram, as well as the Golden Section were important to Pythagoras and the Pythagoreans, as explained in Sect. 3.2. And how the Golden Section determines the pentagram – and conversely – was explained in Sect. 3.4.

By cutting a Kiwi-fruit, we get a different geometric image. It seems very plausible that such natural works of art have played a part in the beginnings of geometry and mathematics.

In Fig. 7.4 we finally cut through an *orange*, and find the assignment of subdividing the circumference of a circle in equal parts. In addition to many other things the reader may find. What do *you* see in this picture?

Fig. 7.5 A computer generated, stylizer representation of a tree to the *left*, a more natural-looking, though still computer generated, rendering of a tree to the *right*

In Fig. 7.5 we show a computer generated, stylizer representation of a tree. The author made this image by means of a simple algorithm implemented in the system MATHEMATICA, see [23]. The image shows an appealing regularity, but looks as much like some type of *fern* than a tree.

In the rightmost part of Fig. 7.5 the algorithm has been modified in some obvious ways, mimicking the presumed randomness in the details as branches develop: Thickness increases with age, direction and length, to some extent randomly. Here the randomness and the developments are determined by a set of input parameters to the program. Also, foliage is now included. The result is a much more natural-looking tree. But with other input parameters, the same program will yield an object looking more like a pineapple than a tree. And, by the way, the original stylized image above also comes from the general program by setting all input parameters equal to zero.

This has brought us to the realm of computer generated images, we return to this theme when we come to the mathematical theory of *Fractal Geometry*. For now, we briefly indicate how fractals may constitute a realistic way for describing certain natural phenomena.

7.4 Fractal Structures in Nature

For a long time the Euclidian geometry was the basis for understanding space. But this traditional way of describing real-life phenomena is increasingly coming under question: Are *straight line segments or smooth curves* really suitable tools in describing the real world as we experience it? For example, take a *shoreline*. As viewed from above, from an airliner, at an altitude of say 30,000 ft., it may look like a smooth curve.

As the plane starts to descend, this smooth curve undergoes a transition: It starts to dissolve. If we observe the same stretch of it from various distances, what looked once as a smooth and straight curve will become more like an irregularly wiggling one. The situation is very much different from what we get if we take a photo at the first elevation, and later enlarge a portion of it.

Coming yet closer, some of the wiggles may turn out to be, actually, small *islands*. Coming even closer, the coastal line may reveal itself as being composed of large rocks, then of somewhat smaller rocks.

Assuming a totally calm sea, we may proceed with a closer and closer examination. Being now out of the airplane, we start to examine the shoreline with a microscope, and find that the *shoreline* simply does not exist, as a *curve*.

Is it therefore impossible to describe, to find *a model for* a shoreline using geometry? Well, at least we may find a better model than the naive one we started out with, using pieces of smooth curves.

Pieces of smooth curves have a property which makes them quite unsuitable for describing a shoreline. Namely, if we consider a geometric object build up from such pieces, then zooming in on a point of it the pieces of smooth curves adjacent to the point will look more and more like *pieces of straight lines*. This is practically the definition of smooth curves, as we shall see in Sect. 13.3.

Euclidian Geometry, which we have studied in Sect. 4.1, is build on the basic elements of *points* and *lines*, as well as *circles*. That is for the planar case, in space we also consider planes and spheres of course. For giving a description of the physical world, however, these elements fall short of even yielding anything approximately correct. As our example with the shoreline shows, reality does not fit into this framework. Plato explained the situation by proclaiming *geometry* to be a *perfect realm of ideas*, while real life suffered from all kinds of imperfections. This was, and still is, a very fruitful point of view, which has inspired geometers and mathematicians throughout our history. But today scientists have turned the tables, viewing *reality* as a perfect and enigmatic realm, which we may never be able to comprehend fully. We have to content ourselves by constructing *mathematical models* for reality. These models make it possible to predict the end results of various sequences of events which we may chose to set in motion. Or we may be able to predict with some reasonable accuracy, the outcome of some natural processes. But the models, the mathematical systems, are imperfect. Not in the sense of being erroneous, of having faulty mathematics build into them. Rather, it is the question of only being able to include assumptions of limited validity into any mathematical model. And no mathematics, as advanced as it may be, can extract more from a system of assumptions than what is already implicit in the agreed-upon *first principles*.

Now the *shoreline* which we investigated have very little in common with a *straight line*. *Lines and planes* are completely unsuitable to even approximately describe shorelines, mountains, hillsides or forests. Nevertheless they do share an important property: They have the property known as *self similarity*. We may phrase this roughly as follows:

Self Similar Objects: An informal definition. An object is said to be self similar, if any part of it is equal to some strictly smaller part when the latter is magnified up to the size of the former.

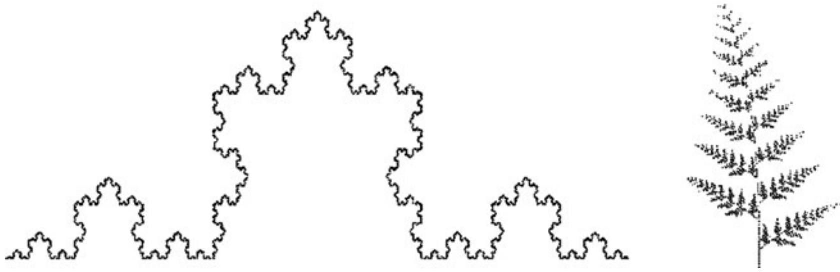


Fig. 7.6 The celebrated von Koch Snowflake Curve to *left*, a computer generated fern to the *right*

Lines and planes certainly have this property. Circles, however, do not. The shoreline do have the property in an approximate sense within certain limits. As always in the real world there are limits to any description we choose to give!

Objects which are *approximately self similar* are ubiquitous in nature. *Fractal Geometry* deals with mathematical objects of this kind, offering fascinating mathematical models for objects encountered in the real world. We shall return to this in Chap. 18, but conclude this section with pictures of self similar objects from nature and from mathematics. It is not easy to tell one type from the other. First we show a variety of the *von Koch snow-flake curve*, which will be explained in more detail in Chap. 18. Here we just observe the appearance, which amply justifies its name.

The von Koch Snowflake Curve is a mathematical object, but it resembles shapes which we encounter in the real world. We cut out two segments, one containing the other. Then when the smaller piece is enlarged up to the size of the bigger one, it will coincide with the bigger piece. The curve is a *self similar* object.

The rendering of the Snowflake Curve shown in Fig. 7.6 has been created by a computer program. This curve was constructed by the Swedish mathematician *Helge von Koch*, with the aim of giving an example of a function $y = f(x)$ defined on an interval, say $I = [0, 1]$, which is continuous everywhere in the interval where it is defined, but such that the *derivative* does not exist anywhere. For such a function the resulting graph is a curve, in this case the von Koch Snowflake Curve, but a rather *pathological curve*: It is continuous but has no tangents, it is of finite extension but of infinite length. As if this were not enough, its dimension is not a whole number! We shall return to this fascinating mathematical object in Sect. 18.2, where we compute the dimension to be approximately 1.262.

Part II
Introduction to Geometry

Chapter 8
Axiomatic Geometry

8.1 The Postulates of Euclid and Hilbert's Explanation

While representing a true watershed in the development of mathematics, in their original formulations the postulates of Euclid for Planar Geometry are not easy to understand. In fact, according to present day standards of rigor, they need to be made more precise as well as to be supplemented by additional statements.

The ideas Hilbert developed in *Grundlagen der Geometrie* [29] have had a profound influence on subsequent mathematical thinking.

Euclid's system of five postulates for planar geometry rests on altogether 23 statements which he call *Definitions*. Euclid also states five so called *Common Notions.* The original wording is by no means inferior, we have presented Euclid's own words, in Heaths vivid translation, in Sect. 4.2. In the statements below, "line" means *straight line*, contrary to what is the case in Euclid's own formulations. We also state the Fifth Postulate in the version due to Proclus.

Postulate 8.1 *Through two different points there passes one and only one line.*

Postulate 8.2 *If two points on a line are in a plane, then the line lies in the plane.*

Postulate 8.3 *All right angles are equal.*

Postulate 8.4 *Given two points in a plane. Then there may be drawn a circle with the first point as center, passing through the second point.*

Postulate 8.5 *Given a straight line α and a point P outside it. Then there is one and only one line β passing through P which does not intersect α.*

It must be noted here that the terms entering into these definitions have really not been defined in a satisfactory way from our point of view today. So far we have only Euclid's original "definitions", quoted in Sect. 4.2:

> A point is that which has no parts. A line is breadthless length. A straight line is a line which lies evenly with the points on itself.

Such "definitions" are not satisfactory by our mathematical standards today, but are certainly of historical interest.

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_8,

Hilbert solved this problem by taking as starting point three *undefined terms*, namely *point, line* and *plane.* Among these he introduced altogether *six* undefined relations, namely *being on, being in, being between, being congruent, being parallel, being continuous*. And finally, Euclid's five *Axioms* (common notions) and five *postulates*, were replaced by a collection of *21 statements.* This system has since been known as *Hilbert's Axioms*. Since Hilbert's work, other axiomatic systems for planar geometry have been devised. At this point, what is important for us in this exposition is not the details of Hilbert's Axioms, but rather their existence: That essentially Euclid's postulates can be made to work in a rigorous modern axiomatic setting.

A system based on Hilbert's axioms, but without the *Parallel postulate*, is frequently referred to as *Neutral geometry*. The study of neutral geometry is interesting, since it consists of all the theorems whose proofs do not require the Parallel postulate. Thus, for example, it follows that even though the parallel to a line ℓ through a point P outside it is not uniquely determined, there is always a unique *normal* from P to ℓ. In neutral geometry there always exist a line through P parallel to ℓ, but it is not necessarily unique.

In neutral geometry there is an absolute angular measure, namely the *radian*. By contrast, in order to introduce *distance* or *measure of length*, it is necessary to *choose* some line segment AB and declare it to be of length 1. It is a theorem in neutral geometry that the angular sum of any triangle is less than or equal to π radians.

We refer to [21] for details on neutral geometry.

Under the assumptions of Hilbert's axioms, he proves that the Euclidian plane may be identified with $\mathbb{R}^2$, and with the usual definition of distance. Today we may short-circuit the axiomatic approach to Euclidian Geometry altogether, by defining Euclidian n-space simply as $\mathbb{R}^n$, with the distance between the points $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$ given by

$$d(P, Q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}$$

The interplay between algebraic properties of $\mathbb{R}$ and geometric assertions is very interesting. The point is that by starting from an initial, weak, set of axioms one can show that a plane subject to these axioms may be parameterized as pairs of elements (x, y), x and y taking their values in some general algebraic system, of which the real numbers would be a special example. Then as axioms are added, each one will be equivalent to some algebraic property of the algebraic system providing the parametrization. In the end, when the axioms are complete, the algebraic system is uniquely determined as $\mathbb{R}$.

We do keep Euclid's postulates in mind, in the version due to Hilbert, when we next discuss the emergence of *non-Euclidian* geometry.

8.2 Non-Euclidian Geometry

The fifth and last statement among Euclid's postulates is the celebrated *Euclid's Fifth Postulate*, also referred to as the *Parallel Postulate.* Euclid himself must somehow have been unhappy with it. While apparently never doubting its truth, he made, according to what we know, numerous attempts of proving it as a consequence of the other four. Not only did Euclid do that, but the pursuit of this elusive goal dominated the lives and careers of many geometers for the next two millennia.

Today we know why it was so difficult. A very important discovery was made to the effect that this postulate cannot be deduced from the other ones. It is *independent* of them. Thus we may construct geometries in which the Fifth Postulate and its consequences are not valid, but where otherwise everything functions as in the Euclidian plane.

This discovery was probably first made by the great German mathematician *Carl Friedrich Gauss*. There had been others before him, coming close to the discovery as they relentlessly worked on finding a contradiction from assuming the converse of the Fifth Postulate. But Gauss never published his discovery. In a letter from 1829 to the German mathematician *Friedrich Wilhelm Bessel* (1784–1846), Gauss writes the following, quoting from the translation in [6]: *"There is another topic, one which for me is almost 40 years old, that I have thought about from time to time in isolated free hours, I mean the first principles of geometry [. . .] and my conviction that we cannot completely establish geometry a priory has become stronger. [. . .] Perhaps this will never [be published] in my lifetime, since I fear the cry of the Boetians [. . .]"*

Euclidian Geometry had for a long time been taken for granted through out the middle ages. Its truth was considered to be absolute. *God had created the World, complete with its Euclidian Geometry.* And these ideas had been adopted by the great philosopher *Immanuel Kant* (1724–1804). He asserts in his *Critique of Pure Reason* from 1781, that *"the concept of [Euclidean] space is by no means of empirical origin, but is an inevitable necessity of thought."*

Strange as this appears to us now, the struggle to free scientific thought from preconceived constraining dogmas had been dead serious. In 1600, about 200 years earlier, the scientist and philosopher *Giordano Bruno* was burned alive for espousing the opinion that Earth moved in a circular orbit around the Sun, thus not being the Center of the Universe. No one would run such risks any more, of course. But Gauss did not wish to spend his time involved in controversy.

It is told that Gauss had his assistants climb various high mountaintops in Germany, measuring large triangles with the mountaintops as their corners.

He wanted to check the sum of the angles in these large triangles, to see if it really added up to 180°. Why would a knowledgeable mathematician do such a thing?

Some claim that already Euclid had realized that his Fifth Postulate was equivalent to the assertion that the sum of the angles in any triangle is twice a right angle. But Heath ascribes this version of the Fifth Postulate to Legendre, from Gauss' own times.

The Hungarian mathematician *Jáons Bolyai* and the Russian *Nikolai Ivanovitch Lobachevsky* (1793–1856), independent of each other found the existence of non-Euclidiangeometry somewhat later.

And these two are credited with the discovery. Bolyai's system of axioms for *hyperbolic Geometry* was published in 1832.

The Fifth Postulate may be replaced with either of the two following, yielding geometries as mathematically valid as the Euclidian one:

Postulate 8.6 *Given a line α and a point P outside it. Then all lines β through P will intersect α.*

However, it should be pointed out that we may not just replace the Parallel postulate in *Hilbert's* system by Postulate 8.6, since parallels always exist in neutral geometry, as pointed out in the previous section. Other modifications of the axioms have to be carried out as well.

Postulate 8.7 *Given a line α and a point P outside it. Then there are at least two lines β_1 and β_2 through P which do not intersect α.*

If we use the last Postulate, we obtain the so-called *hyperbolic geometry.* The first yields the *elliptic geometry.* We return to the issue of how to realize these geometries. But before we can do so, we first have to explain how we understand the concepts of *axioms, systems of axioms and models for such systems* in a precise mathematical setting. This requires some preliminaries on *Set Theory*.

8.3 Logic and Intuitive Set Theory

Logic and Set Theory are mathematical fields with a high level of precision, as they represent the very foundation for the edifice of mathematical theory. And if the foundation is not sound, then the total body of knowledge cannot be considered secure.

It therefore was considered deeply troubling when *contradictory assertions* could be deduced by the same methods of proofs which were, unquestioningly, used to prove *the theorems*. Some of these contradictions were quite technical, but a very simple one was found in 1902 by the mathematician and philosopher *Bertrand Russell*. We shall treat this *Russell's Paradox* in Sect. 8.4.

We shall now give a brief summary of some basic notions from logic and set theory. In practice we do not need the intricacies of these theories, however. Indeed, it will suffice to understand the concept of a *statement* as assertions like *"$2 + 2 = 4$"*, *$2 + 2 = 3$"* or, say *"The moon is made of cheese"*. Our statements may be either *true*, in which case they will be assigned the *"truth-value" T*, or *false*, in which case they are assigned the truth-value F. No other alternatives exist.

For a given statement P, we let $\neg P$ denote *the negation* of P: If P is the statement $n = m$ then $\neg P$ is $n \neq m$. Further, for two statements P and Q, we let $P \wedge Q$ and $P \vee Q$ denote the statements *"P and Q"* and *"P or Q"*, respectively. Finally,

Fig. 8.1 Truth table for some composite statements

P	Q	$P \wedge Q$	$P \vee Q$	$P \implies Q$	$P \iff Q$	$\neg P$
T	T	T	T	T	T	F
T	F	F	T	F	F	F
F	T	F	T	T	F	T
F	F	F	F	T	T	T

let $P \implies Q$ denote the statement *"P implies Q"*, and we let $P \iff Q$ denote the statement *"P implies Q and Q implies P"*, thus that P and Q are equivalent.

The composite statements introduced above may be viewed as *boolean functions*, which means functions where the variables P and Q only may take the values T or F, and the functions themselves also may take only the values T or F. Tables like the one given in Fig. 8.1 define such functions.

We also recall the following *set notations*: If A is a set and a is an element in it, we express this by writing $a \in A$. Moreover, $A \subset B$ signifies that A is a proper subset of B, while $A \subseteq B$ means that A is a subset, possibly equal to, B.[1] $A \cup B$ denotes *union* of A and B, i.e., all elements which lie either in A or in B, or both. $A \cap B$ denotes *the intersection* of A and B, thus all elements which both lie in A and in B.

The set of all elements a which satisfy a statement $P(a)$ we write as

$$\{a|\ P(a)\},$$

or expressed in words: *The set of all a such that the statement $P(a)$ is true*.

8.4 Axioms, Axiomatic Theories and Models

An *axiomatic theory* consists of a set *undefined terms*, and a system of *axioms* which these terms satisfy. Throughout modern mathematics one encounters a number of such axiomatic theories, all mathematical disciplines are in one way or another build on such a foundation.

But the foundation under the mathematical edifice was often put in place long after the construction of the building started, indeed frequently long after it had been completed. And furthermore, through the ongoing research on the foundations of mathematics, new levels under the building is being added all the time. Thus the final result is elusive, if even attainable. In fact, the process itself is probably more important than its "goal". Not only does mathematics reach out towards the outer limits of our universe of thinking, it also pierces deep into the microcosms of its foundations.

[1] Some authors use the symbol $\subset$ to mean $\subseteq$, so it is a good idea to check the notation before drawing conclusions.

Fig. 8.2 Georg Cantor

At certain points in the history of mathematics it really looked as if the whole edifice might collapse, or at least would have to be thoroughly rebuild.

Mathematics had been build on the so-called *intuitive* – non-axiomatic – set theory. Through the efforts of the brilliant number theorist and set-theorist *Georg Cantor* (1845–1918) this foundation appeared to be safe and secure (Fig. 8.2). It therefore came as a considerable shock when several apparently grave inconsistencies surfaced. The first one was discovered in 1897 by the Italian mathematician *C. Bural-Forti*, and 2 years later Cantor himself found a similar paradox. Annoying as these paradoxes were, they dealt with rather exotic constructions known as *transfinite numbers*. These concepts lay, at the time, at the outer fringes of mathematical knowledge. The paradoxes were, therefore, not as threatening as possible inconsistencies right within the central body of mathematical knowledge would have been. Then, in 1902, the real bombshell struck: Bertrand Russell discovered a paradox which only depended on the basic concept of *a set* and *an element being a member of a set*. Proofs based on similar reasoning had been accepted as fully valid throughout mathematics.

At the time of this discovery, *Friedrich Ludwig Gottlob Frege* (1848–1925) had just completed a prodigious work in two volumes, on the foundations of arithmetic based on Cantor's Set-Theory. He is quoted as having asserted that *"Every good mathematician is at least half a philosopher, and every good philosopher is at least half a mathematician"*. At the end of the last volume Frege was now obliged to acknowledge, in a note added in print, that due to a communication from Russell he now finds himself in the undesirable position of seeing the very base under his efforts giving way. We shall now explain the famous *Russell's Paradox*.

Let Ω denote the set of *all existing objects*, the *universal set which contains absolutely everything in the universe*. Since this set contains absolutely everything, it must contain itself, thus is *an element in itself*:

$$\Omega \in \Omega$$

Ω has of course many elements which themselves are sets. Some of these are elements in themselves, such as Ω, while others – the most – do not have this somewhat unusual property. We denote *the set of all objects having the property that it is not an element in itself* by Δ:

$$\Delta = \{\Gamma \mid \Gamma \notin \Gamma\}$$

The question now becomes: *Is it true or false that* $\Delta \in \Delta$*?*

Assume that $\Delta \in \Delta$. Then Δ does not satisfy its defining condition, thus Δ cannot be an element in Δ, i.e., $\Delta \notin \Delta$. Assume on the other hand that $\Delta \notin \Delta$. Then *the defining condition is satisfied*. Hence we do get $\Delta \in \Delta$. So both alternatives are impossible. This is clearly a self-contradictory result.

The solution to this and other mathematical debacles stemming from deficiencies in the set-theoretical foundation of mathematics, was to introduce *Axiomatic Set-Theory.* The axioms for set theory prescribe in detail exactly which sets can exist. Thus for instance, the set Ω introduced above *does not exist*, it is "too big".

The Axiomatic Set-Theory takes as its starting point the undefined terms *Set* and *Element*, as well as a relation $\in$ which may exist between an element and a set. The theory does not take any position on "what the sets and elements really are", the interplay between the terms as prescribed by the axioms being the issue of concern.

We shall not give the complete system of axioms for set theory, but confine ourselves to some selected, and even somewhat simplified axioms to give a flavor of the theory. The axioms behind our treatment here are due to *Ernst Zermelo*, *Adolf Abraham Halevi Fraenkel* and *Albert Thoralf Skolem*.

As usual we use the simplified notation $\alpha \notin A$ instead of $\neg(\alpha \in A)$. Also, we shall use the term *"statement"* as explained in Sect. 8.3. In addition to the symbols we have explained above, we also use the so-called *quantifiers*, $\forall$ and $\exists$: When writing $(\forall\, a \in A)(P(a))$ what we mean to say is this: *"For all elements a in the set A the statement P(a) is true"*. Furthermore, the symbols $(\exists\, a \in A)(P(a))$ expresses that there exists an element a in the set A such that the statement $P(a)$ is true. We are now ready to state the first axiom in the *Zermelo-Fraenkel-Skolem* system for Axiomatic set theory:

Axiom 8.1 (ZFS 1) *Given two sets A and B for which the following statement is true:*

$$(\forall\alpha)(\alpha \in A \Longleftrightarrow \alpha \in B)$$

Then $A = B$.

This is certainly a property which is required in such a theory. It asserts that the membership-relation $\in$ and the rules of mathematical logic determine the sets uniquely. The next axiom assures that whenever a set is given, and a statement involving elements in the given set is formulated, then that statement determines a unique *subset* of the given set:

Axiom 8.2 (ZFS 2) *Given a set Ω, and a statement $P(\alpha)$ about elements α. Then there exists a set Δ such that the following statement is true:*

$$(\forall\alpha)(\alpha \in \Delta \iff (\alpha \in \Omega) \wedge P(\alpha)).$$

Almost as in the *intuitive set theory* we write:

$$\Delta = \{\alpha \in \Omega |\ P(\alpha)\}.$$

Already these two simple axioms suffice to salvage some of the most elementary rules from the *intuitive* set theory. But note that there is a big difference between the axiom above on the one hand, and to postulate the existence of the following set on the other:

$$\Delta' = \{\alpha |\ P(\alpha)\}.$$

Assuming this would lead to a contradiction in the same manner as for Russell's Paradox in the intuitive (more precisely: Cantor's) setting.

So far so good. But we wish to make sure that there does exist sets at all! Therefore we add the

Axiom 8.3 *There exists one and only one set $\emptyset$ such that the statement*

$$\alpha \in \emptyset$$

is false for all elements α.

The existence of the empty set may of course not be taken for granted, it has to be secured by a separate axiom in the *Zermelo-Fraenkel-Skolem* System. Actually as it turns out, the assertion will follow from some stronger existence axioms. We omit these considerations here. But if we choose to include the empty-set axiom in this form, then we need only to postulate *the existence*, as *the uniqueness* follows from the axioms we have already introduced. This proof is left to the reader.

We now should see what the guiding principle in building the system of axioms is: As much as possible of the intuition should be saved, and put on firm and secure ground. A further step in this process comes with the next axiom:

Axiom 8.4 (ZFS 3) *Given two sets A and B. Then there exists a set Γ such that $A \in \Gamma$ and $B \in \Gamma$.*

It follows from this, of course, that any *set* also appears as an element in some other set.

We would of course like to have the usual constructions of union, intersection and complement. For this we need the

Axiom 8.5 (ZFS 4) *Given two sets A and B. Then there exists one and only one set, which is denoted by $A \cup B$, such that the following statement is true:*

$$(\forall\alpha)(\alpha \in A \cup B \iff (\alpha \in A) \vee (\alpha \in B)).$$

Using these axiom we may prove the existence of a fairly large collection of sets. As an example we shall see how *the complement* of one set in another set is constructed:

Theorem 7. *Given two sets A and B. Then there exists a uniquely determined set $B - A$ such that the following statement is true:*

$$(\forall\alpha)(\alpha \in B - A \iff \neg(\alpha \in A) \wedge \alpha \in B)).$$

Proof. We use ZFS 2 with $\Omega = B$ and $P(\alpha) = \neg(\alpha \in A)$. □

We have not postulated the existence of *the intersection* of two sets. This may also be deduced from the axioms:

Theorem 8. *Given two sets A and B. Then there exists one and only one set, which is denoted by $A \cap B$, such that the following statement is true:*

$$(\forall\alpha)(\alpha \in A \cap B \iff (\alpha \in A) \wedge (\alpha \in B)).$$

Proof. We use ZFS 2 with $\Omega = A$ and $P(\alpha) = (\alpha \in B)$ □

Having reached this stage, it is not difficult to define – by abstract use of the axioms – what we mean by the relation $A \subset B$ between to sets. We also easily give meaning to the usual notation

$$A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$$

for a *finite set*.

In the further development of the system we come to an axiom *ZFS 5*, which guarantees the existence of *the set of all subsets* $\mathcal{P}(A)$ of an arbitrary set A. Using this and the earlier axioms one may prove the existence of *the product* of two sets A and B:

$$A \times B = \{(\alpha, \beta) |\ (\alpha \in A) \wedge (\beta \in B)\}.$$

A further axiom *ZFS 6* guarantees the existence of sets of *infinitely many elements*: This axiom postulates the existence of a set A such that $\emptyset \in A$ and such that if $\alpha \in A$, then we have $\alpha \cup \{\alpha\} \in A$.

The last axiom which we shall explain here, is the seventh one, *ZF 7*. It is called *The Axiom of Choice*, and may be formulated as follows in plain language: *If Ω is a set whose elements are sets, then we may form a set Δ by choosing one element from each set which occur as an element in Ω.*

This apparently innocuous statement is actually somewhat controversial. The point is that the notion of *a set* is still quite far-reaching, and some mathematicians feel that to just *choose with no specification of constructive procedure* from a *set of sets*, is a rather sweeping acceptance of what kinds of sets we will allow to exist. But in many cases where one proves a result using the axiom of choice, it is actually possible to proceed by more constructive and thus, safer, means.

The author learned set theory from Skolem in the fall of 1962, attending the very last lectures he gave on this subject as *Professor Emeritus* in Oslo, Norway. Skolem was very skeptical to the Axiom of Choice, and spent several lectures deducing seemingly absurd consequences of this axiom. The implied question being: *"Now, do we really want this?"* In the end he summed up his view as follows: *"In mathematics, we cannot and indeed, should not, prohibit anyone from the study of any odd system of axioms. However, we must be allowed to question the meaning of results obtained from such a set of assumptions."*

The two last axioms are of a more technical nature, and will not be treated here. The reader may consult [54] for this and further reading.

8.5 General Theory of Axiomatic Systems

Set theory lies at the base for all mathematics, at least in theory. Thus in all other axiomatic theories one may take this as starting point. The undefined terms could then be elements in certain sets, and relations between these. Other axiomatic theories may deal with more general classes of objects than sets, this is the case with *Category Theory,* for instance. Such theories fall outside the scope of the this book.

When it is possible to give a concrete interpretation of the undefined terms, in such a way that all the axioms become true statements, then we have constructed a *model* for the axiomatic theory. If this is possible, then clearly the system of axioms do not contain any contradiction. This statement has to be qualified somewhat, however. Suppose that the model is constructed in terms of the real numbers. This will be the case for all the models we are going to consider here. Then we may conclude from the existence of the model that if there is no contradiction among the axioms leading up to the construction of the reals, then there is no contradiction in our axiomatic system under consideration. As the former statement is taken for granted, we may say that our new system is free of contradictions.

This is *the fine print of math. We should realize it is there, but not spend our lives worrying about it.* In this case we say that the system is *consistent*. If no axiom can be deduced from the remaining ones, then we say that the system is *independent.*

Finally, we say that an axiomatic system is *complete* if any statement involving the undefine terms either may be proven or may be disproved, i.e., the *negation* of the statement may be proven, by means of the axioms.

We cannot leave the subject of general axiomatic theories without mentioning a result which is considered deeply troubling by some, but others have found ways to live with: In 1931 the brilliant mathematician and logician *Kurt Gödel* showed that *if an axiomatic system is complete, then it cannot be proven consistent by means from inside the system itself.* Thus if we want to be certain that there are no contradictions in our system, then we have to live with some *undecidable* questions. This insight certainly came as a shock to Hilbert and others, who had wanted to treat all of mathematics as a huge axiomatic system, complete and consistent, thus providing a firm and absolutely secure basis for our knowledge. Unfortunately – or fortunately – the reality was not that simple.

Exercises

Exercise 8.1 Recall the original Definitions, Postulates and Common Notions of Euclid, as stated in Sect. 4.2. Several of Euclid's propositions state that certain constructions are possible, using the five postulates. In fact, Euclid's very first proposition is this:

Proposition 1. On a given finite straight line to construct an equilateral triangle.

Letting AB be the (straight) line segment, Euclid proves this by applying Postulate 3 to a circle of radius AB and center A, and a circle of radius BA and center B. However, it is generally recognized that the proof is incomplete as stated in Book 1 of the Elements.

Complete the proof by using the following statement, perhaps tacitly assumed by Euclid in some form or another:

Principle of Continuity. If a (not necessarily straight) line belongs entirely to a plane [in this case the circle about A], which is divided in two parts [in this case the inside and outside of the circle about B], and if the line has at least one point in common with each part, then it must also meet the boundary between the parts. (Or, as Euclid might perhaps have written, the "line which is the common extremity of the two parts").

Exercise 8.2 Use the postulates together with Proposition 1 to prove Euclid's next proposition:

Proposition 2. To place at a given point (as an extremity) a straight (finite) line equal to a given straight (finite) line.

Show that this proposition extends the power of the compass in performing constructions: The so-called *Euclidian compass* can only draw a circle with a given or already constructed point as center, passing through a given or already constructed point. Explain that this proposition allows the geometer to draw a circle with a known point as center and with radius equal to the distance between two other known points.

Chapter 9
Axiomatic Projective Geometry

9.1 Plane Projective Geometry

The axiomatic treatment of *plane projective geometry* has at its starting point three *undefined terms: point, line and incidence*. We are given one *set* $\mathcal{P}$, which we call *the set of points*, and another set $\mathcal{L}$ which we call *the set of lines*. Further, there is given a *relation* between elements from $\mathcal{P}$ and elements from $\mathcal{L}$ which is denoted by I, and referred to as *incidence*. If $PI\alpha$ holds for $P \in \mathcal{P}$ and $\alpha \in \mathcal{L}$, then we say that *the point P is incident with the line* α.

Of course we will tie our intuition to the implementation of the abstract setting where α is a "real" line, P a "real" point and that $PI\alpha$ means that P lies on the line α in the sense that $P \in \alpha$. But as we will see later, it is both possible and indeed some times desirable, to give *other* implementations of these terms, such that the axioms which we shall formulate below hold also in these cases.

Furthermore, we write αIP as meaning the same as $PI\alpha$. It simplifies the exposition, and is in close agreement with our intuition and usual language to extend the meaning of the relation I in this manner.

Thus our point of departure is to have abstractly given two sets, $\mathcal{P}$ of "points" and $\mathcal{L}$ of "lines", and a relation I which may hold between elements $\alpha \in \mathcal{L}$ and $P \in \mathcal{P}$, and which we write as $PI\alpha$ or as αIP.

For reasons of practicality only, we will not use this unfamiliar and a little awkward relation I, but instead of "$PI\alpha$" or *P is incident by* α we simply say that "*the point P lies on the line* α". Similarly we replace the statement αIP or α *is incident with* P by "*the line* α *passes through the point* P". We also retain, based on the stringent meaning prescribed above, such formulations as: *"The two lines* α *and* β *meet in the point* P*"*.

This yields a more natural language in line with our usual geometric intuition. As made precise in this way, "informal" statements will not be less "mathematically precise" than the more formal statements using the relation I as well as logical and set-theoretical notation. Since we have given the intuitive geometric language a *new meaning,* and tied it to our new stringent axiomatic context, there is nothing preventing us from continuing to use the familiar words in our further work. Indeed, one of the reasons why the experiment with the so-called *"New Math"* in Elementary

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_9,

School turned out to be a disappointment may have been the belief that a high level of formalism is a prerequisite for mathematical precision. But that is not always the case, and high symbolic complexity may hide rather than enhance the mathematical ideas. This is especially relevant for the teaching of *geometry*.

Projective Plane Geometry is built on the following system of axioms. As we shall see later there are other, equivalent axiom-systems as well.

Axiom 9.1 *Let P and Q be two different points. Then there exists one and only one line α such that P and Q lie on α.*

Axiom 9.2 *Two different lines meet in one and only one point.*

This axiom shows that we are no longer dealing with *Euclidian geometry*: Tow lines will always meet. Intuitively we may adapt the familiar notions from Euclidian plane geometry by adding points at infinity, and then prescribing that parallel lines meet at infinity. Then we still have an axiom which expresses something not alien to our experience.

Three different points are said to be *collinear* if they lie on the same line.

The next axiom is needed to get more than a single line in our projective plane:

Axiom 9.3 *There are at least three non-collinear points.*

We need an axiom to the effect that lines may not merely be pairs of points.

Axiom 9.4 *On every line there are at least three points.*

The Seven Point Plane consists of seven points and seven lines. Each line has three points on it. Three of the lines are drawn as the sides in a triangle, with three of the points appearing as the vertices. On each line the third point is depicted between the two "vertices". The seventh point in the geometry is depicted in the center of the triangle, with three lines passing through it, as shown. We now have six lines, and the seventh is shown as a circle, in the picture it is inscribed in the triangle. Note that there are no other points than the ones marked, thus the lines and the circle are drawn only for the purpose of illustration.

A model for this axiomatic system is depicted in Fig. 9.1. This model is forced on us as the smallest geometry which satisfy the four axioms above. In fact, there must exist a line, and on it there must be at least three points. Then, there are at least three non-collinear points, thus a point outside the line. We then have the "triangle", as well as three points on the " base-line". But by the system of axioms, the other two lines must have an extra point, yielding at least three on each of them. We now have six points but only three lines. But the "middle point" on each of these three lines and the point outside it, together determine a line, yielding three more, so we have six. But so far these new three lines have only two points. If we let them all pass through a seventh point in the middle, we have added new points in the most economical manner. So we do that, wishing as we do, to keep the number of points minimal. The system demands that two points on different "sides" of the triangle, together determine a line. This is accomplished, in the minimal way, by letting that

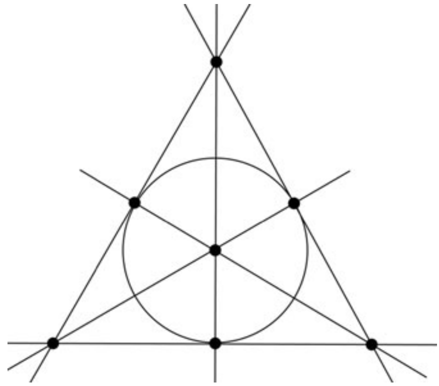


Fig. 9.1 The seven point plane

line pass through all the "middle points". Now all of the axioms are satisfied, and there is no need to add more points or more lines.

We have thus shown that *this axiom system is consistent* or *that it contains no contradiction*. Indeed, if it contained contradictions, then it could not have a model. We make the following

Definition 1 (Projective Plane). A model for the above system of axioms is called a projective plane, or a plane projective geometry.

We now pose the question of whether this axiom-system is *minimal*. Conceivably, for example, the fourth axiom might be deducible from the first three. We now prove that this is not possible.

Theorem 9. *The first three axioms do not imply the fourth.*

Proof. It suffices to exhibit a model in which the first three axioms hold, but not the fourth. Such a model is depicted in Fig. 9.2. □

In a similar way we easily prove that none of the four axioms follow by the remaining three. Thus we have the

Theorem 10. *The four axioms for plane projective geometry are independent.*

We shall now state and prove the first simple theorems in projective plane geometry. We start by making a definition.

Definition 2. An Arc of Four is a set of four points such that any choice of three among them is not collinear.

We may prove the following:

Theorem 11. *There exists at least one Arc of Four.*

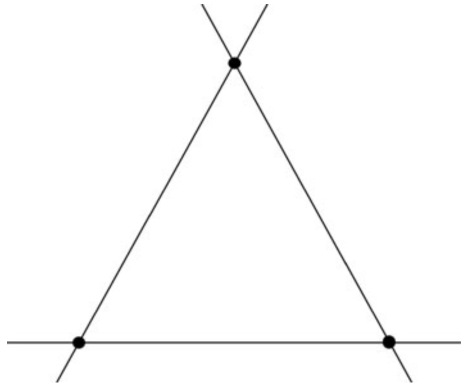


Fig. 9.2 This model, which shows that the fourth axiom does not follow from the first three, consists of only three points, not on the same line. Evidently the three first axioms are satisfied in this model, while the fourth does not hold

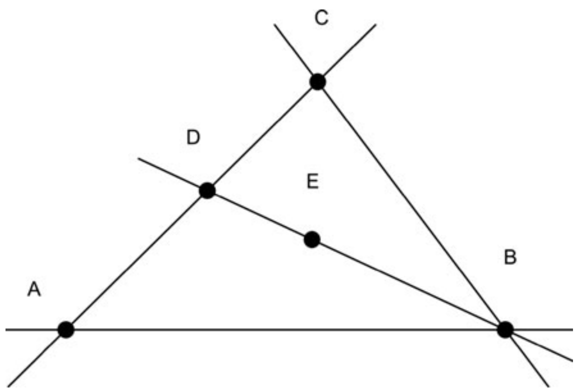


Fig. 9.3 The construction of an Arc of Four

Proof. The construction in the proof is shown in Fig. 9.3.

By Axiom 9.3 there exist three non-collinear points A, B and C. By Axiom 9.1 we have uniquely determined lines AB, AC and BC. By Axiom 9.4 the line AC contains one more point, say D. For the same reason,the line BD must contain one more point, which we denote by E. We now claim that $\{A, B, C, E\}$ will constitute an Arc of Four. Namely, by construction A, B, C are not collinear. If A, B, E were collinear then the point D would be on AB, such that C would be on AB, contrary to the above. If A, C, E were collinear, then E and D would coincide because of the uniqueness of the point of intersection of two different lines, Axiom 9.2. It remains to check B, C, E: If these were collinear, then C and D would coincide by Axiom 9.2 as in the previous case. Thus the claim is proven. □

The four axioms for the Projective Plane have a remarkable property: If we interchange the words "point" and "line", then we get *true* statements, in the sense that they may be deduced from the given four axioms. Thus, for instance, Axiom 9.1 is transformed into Axiom 9.2, while Axiom 9.2 is transformed into Axiom 9.1 by interchanging "point" and "line". The Axioms 9.3 and 9.4 are transformed into the statements of the following two theorems:

Theorem 12. *There are at least three lines not passing through the same point.*

Proof. Choose the points A, B, C, by Axiom 9.3. Then the lines AB, AC and BC will not pass through the same point. □

Theorem 13. *For all points there exist at least three lines passing through it.*

Proof. Let A be an arbitrary point, and choose two new points B, C such that A, B, C are not collinear. This is possible, as otherwise all points in the geometry would lie on the same line, so that Axiom 9.4 would not hold. We now choose a third point on the line BC, call it D. Then the three lines AB, AC, AD satisfy the claim. □

This simple observation is of fundamental and far reaching significance: It lies at the foundation of the so-called *Principle of Duality*.

We shall say that a statement is a *valid theorem in projective plane geometry*, if it may be deduced from the axiom-system for the projective plane given above.

Then we have the following result:

Theorem 14 (Principle of Duality). *We get a new valid theorem in projective plane geometry whenever we interchange the words "point" and "line" but retain incidence, in a valid theorem for plane projective geometry.*

We have proved a valid theorem in projective plane geometry above, namely the existence of an Arc of Four, Theorem 11. This now implies the

Corollary 1. *There are at least four lines so that no selection of three among them all pass through the same point.*

Proof of the Principle of duality: Let P be a valid theorem in plane projective geometry. Then

The axioms 9.1, 9.2, 9.3 and 9.4 $\Longrightarrow$ P

In other words, our system of axioms implies the statement P. But this means that P is a true statement *in all models for the system of axioms*, i.e., P will be true no matter how the words "point", "line" and "incidence" are interpreted, as long as the statements provided by the four axioms are valid for the interpretation.

We now form a model by interchanging the words "point" and "line", retaining the meaning of incidence. Abstract as it is, this still will be a model, since we have proven that all four axioms hold with this interchange of "points" and "line". In particular the statement P is true for this model, i.e., the dual statement is true. This completes the proof. □

We may replace the Axioms 9.3 and 9.4 by the statements in Theorem 11 and Corollary 1. To prove this, it remains to show that these four statements conversely imply the Axioms 9.1, 9.2, 9.3 and 9.4. This system of axioms would have the advantage that the Principle of duality now would be completely obvious, as the axiom-system would be *self dual*. The disadvantage is that such an axiom-system would appear less intuitively evident and less natural.

9.2 An Unsolved Geometric Problem

We shall say that a projective plane is *finite* if the set $\mathcal{P}$ of points is a finite set. Omitting the set of lines from our notation, we will from now on say *"the projective plane* $\mathcal{P}$*",* letting the set of lines be understood.

Definition 3. Given a finite projective plane $\mathcal{P}$. Let M denote the maximal number of points on any line. Then we say that the geometry $\mathcal{P}$ is of order $m = M - 1$.

Thus for example, the 7-point geometry which we have encountered already, is of order 2. Not only that, but the number of points on *all* lines is $M = 3$ in this case.

Indeed, one may prove in general that the number of points on all lines is equal to the maximal number M. Moreover, the number of lines through any point is equal to M as well.

We have, as noted, constructed one geometry of order 2. It also is very easy to construct geometries of order p, where p is a prime number. And it is not difficult to construct geometries of order p^r, where p is a prime and r a natural number > 0. But the following conjecture has turned out to be very difficult.

Conjecture 1. All finite geometries have order equal to p^r, where p is a prime number and r a natural number > 0.

It may appear surprising that this should be so difficult to decide. But this is the way it frequently turns out in mathematics: A problem may be very simple both to state and to understand, but extremely difficult to solve.

Thus the problem is the following: *Assume that m is a natural number which is not a power of a prime. Show that then there is no plane projective geometry of order m.*

As far as I know the best result towards a solution of this conjecture was obtained in 1949 by *R. H. Bruck* and *H. J. Ryser*: They showed the following:

Theorem 15 (Bruck–Ryser). *If the number m is not a power of a prime, and is such that division by 4 leaves a remainder of 1 or 2, and furthermore m cannot be written as a sum of two squares, then m is not the order of some plane projective geometry.*

This remarkable result excludes an infinite number of cases. The first of them are the numbers $6, 14, 21$ and 22. Take for instance 6: Division by 4 yields the remainder 2, and as $6 = 1 + 5 = 2 + 4 = 3 + 3$ are the only ways in which one may write 6 as the sum of two natural numbers, we see that 6 cannot be written as the sum of two squares. Thus the result of Theorem 15 shows that 6 is not the order in a plane projective geometry.

The first case which remains open after the Bruck–Ryser Theorem is therefore the case $m = 10$. Division by 4 does give the remainder 2, but $10 = 1 + 9$, so that it is the sum of two squares and the last part of the test fails. So the first case to check comes down to the following problem:

Problem 1. Does there exist a plane projective geometry in which all lines have 11 points?

Other open cases after the theorem stated above are $m = 12, 15, 18, 20$. Of course there are infinitely many such open cases.

It may appear surprising that the case $m = 10$ should be difficult to decide. But we may rapidly see why one cannot just sit down with paper and pencil and check out all cases which may occur. Namely, as mentioned above there are $M = 11$ lines passing through each point. Since two lines will have exactly one point in common, the total number of points in such a projective plane would be

$$11 \times 10 + 1 = 111 :$$

Indeed, choose a point say P_0. Through this point passes 11 lines. Each of them have 10 points in addition to P_0. By Axiom 9.1 all points in the plane will appear on one of these lines, so in addition to P_0 there are altogether 11×10 points. Including P_0 this gives a total 111 of points, as claimed.

We now wish to check all possibilities for prescribing subsets of a set containing 111 elements, or "points", yielding a set of subsets, called "lines", such that our four axioms are satisfied.

A digression at this stage: The alert reader may now be concerned, that we have moved to a concrete interpretation of the undefined term "line". But this is always possible: We may harmlessly identify a "line" ℓ with the *subset* $\mathcal{P}_\ell$ *of* $\mathcal{P}$ *consisting of all* $P \in \mathcal{P}$ *such that* $P I \ell$. Namely, as is easily seen we have that for two lines α and β we have $\alpha = \beta$ if and only if P is incident with α exactly when P is incident with β. The formal proof is left to the reader, the key being that if $\alpha \neq \beta$, then they have exactly one point in common.

If we make this identification, then the incidence relation is interpreted as $\in$, the membership relation between an element and a set.

Thus a "line" will be a subset of $\mathcal{P}$ containing 11 points. Altogether there are a total of

$$\binom{111}{11} = 473239787751081$$

such subsets. Now the number of "lines" in a Plane Projective Geometry is equal to the number of "points", thus in this case 111. Hence the candidates for projective planes of order 10 will be every selection of 111 among these 473239787751081 subsets, altogether a staggering $\binom{473239787751081}{111}$ possibilities to be checked for compliance with our four axioms. The reader is recommended to take a few minutes to compute this number, say by MAPLE or a similar system. Just do not try to do it by hand or calculator.

This is certainly a completely insurmountable task. Even though we do not have to work through absolutely all the possible selections, and even though the checking can be made a lot smarter than just working through all possibilities, even this initial case presents us with a real challenge.

Thus it attracted a great deal of attention when a group of mathematicians and computer scientists at Concordia University in Montreal showed that $m = 10$ cannot be the order of any plane projective geometry. Their proof is dependent on massive computer usage, but also involves intricate mathematical considerations. For details we refer to [4] and to the interesting article by *B.A. Cipra* in Science [8].

The history of this problem is fascinating, we give a short account of the beginnings below, and refer to the paper [39] for more details.

In a paper from 1782, Euler asks for an arrangement of 36 officers of 6 ranks and from 6 regiments in a square formation of size 6 by 6. Each vertical and each horizontal column is to contain exactly one officer of each rank and exactly one from each regiment. Euler denoted the 6 regiments by Latin letters *a,b,c,d,e,f*, and the 6 ranks by Greek letters $\alpha, \beta, \gamma, \delta, \epsilon$ and ζ. The problem thus is of arranging the 36 combinations of two letters, one Latin and one Greek, in a square in such a way that every row and every column contains the six Latin and the six Greek letters. Such an arrangement has been called a *Graeco-Latin square*, and Euler started by considering the problem of placing the Latin letters in a square where no letter is missing from any row or any column, he called this a *Latin square*. See [4] for more details.

Another problem which contains the same mathematics is found in [10]. Here we are in the difficult negotiations leading up to Algerian independence from France. We have six representatives from Algeria, six representatives from France, and six mediators. We want to schedule six tours of Paris to let the negotiating parties get to know each other as follows. Each representative from Algeria goes on all six tours with a French representative and a mediator in such a way that the Algerian representative tours with each of the six mediators and no mediator has to go on the same tour twice. The same conditions must also hold for French representatives.

In the simpler case with only three representatives, A_1, A_2, A_3 from Algerie, French representatives F_1, F_2, F_3, tours t_1, t_2, t_3, and mediators m_1, m_2, m_3 we get one of the possible plans as follows:

	F_1	F_2	F_3
A_1	$t_1 m_1$	$t_2 m_2$	$t_3 m_3$
A_2	$t_2 m_3$	$t_3 m_1$	$t_1 m_2$
A_3	$t_3 m_2$	$t_1 m_3$	$t_2 m_1$

The table says that A_1 in row 1 and F_1 in column 1 go on tour t_1 with mediator m_1, A_2 and F_1 on tour t_2 with mediator m_3, etc. All representatives go on all tours with all mediators, and no mediator goes on the same tour twice.

So here we have two Latin squares of order 6. In this case they are *orthogonal Latin squares*: In fact, we say that a pair of Latin squares $A = (a_{i,j})$ and $B = (b_{i,j})$ are orthogonal if all the ordered pairs $(a_{i,j}, b_{i,j})$ are distinct.

Euler found no solution to this problem, and he conjectured that no solution exists if the order of the square is of the form $n \equiv 2\ (mod\ 4)$. This is the famous Euler's conjecture. The first case $n = 2$ is trivially impossible. In 1901 it was verified that

Euler's conjecture holds for $n = 6$. But in 1960 it was disproved in combined work by R.C. Boise, S.S. Shrikhande and E.T. Parker. We refer to [4].

The following result is also due to Bose, see Theorem 8.4.12 in [4]:

Theorem 16. *We may construct a projective plane of order* $n \geq 3$ *if and only if there exist a set of* $n - 1$ *mutually orthogonal Latin squares of order* n.

In particular this reconfirms that there is no projective plane of order 6.

9.3 The Real Projective Plane

We start by recalling some standard analytic geometry. When we introduce coordinates, a point is represented by a *pair* of real numbers, (x, y). A line is represented by its equation,

$$AX + BY + C = 0,$$

where A, B, C are real constants and X, Y are the *variables*. The line given by the equation, then, is the set of all points (x, y) in the plane represented as $\mathbb{R}^2$ consisting of all points (x, y) such that

$$Ax + By + C = 0.$$

This gives us the tool for transforming geometric considerations from the usual Euclidian plane into algebraic computations.

To be precise, we have here constructed a *model for the Euclidian plane,* in an axiomatic setting. We realize, however, that this will not be a model for the axiomatic system defining plane projective geometry. Indeed, all the axioms hold, *except for* Axiom 9.2.

We now show how $\mathbb{R}^2$ may be extended to the *real projective plane*, such that this axiom also holds. We do that by adding some points, which we will call *the points at infinity.*

The resulting model will be the so-called *real projective plane* $\mathbb{P}^2(\mathbb{R})$.

At first the definition may look strange and unnatural:

Definition 4. The set $\mathcal{P}$ of points in the real projective plane $\mathbb{P}^2(\mathbb{R})$ is the set of all lines through the origin $(0, 0, 0) \in \mathbb{R}^3$.

The set $\mathcal{L}$ of lines is the set of all planes in $\mathbb{R}^3$ which pass through $(0, 0, 0)$.

We say that the projective point P, that is to say, the line $a \subset \mathbb{R}^3$, is incident with the projective line L, i.e., the plane $p \subset \mathbb{R}^3$, if $a \subset p$. We write then, as in the formal setting, $PI\alpha$.

We next verify that all four axioms are satisfied.

Verification of Axiom 9.1: Let P and Q be two different "points" in *the real, projective plane*, as defined above. Thus P and Q are two different lines through

$(0, 0, 0)$ in $\mathbb{R}^3$. Two such lines span a unique plane through $(0, 0, 0)$. This plane is by our definition a "line" in $\mathbb{P}^2(\mathbb{R})$, which we denote by a in Fig. 9.4.

Verification of Axiom 9.2: Let a and b be two different "lines" in the real, projective plane, thus two planes through $(0, 0, 0)$ in $\mathbb{R}^3$. These two planes will intersect in a unique line p in $\mathbb{R}^3$ passing through the origin. This line is, according to our definition, a "point" in the real, projective plane $\mathbb{P}^2(\mathbb{R})$. We change its name, denoting it by P, and thus have that the two lines a and b intersect in the point P. Moreover, P is uniquely determined. The situation is illustrated in Fig. 9.5.

Verification of Axiom 9.3: Let p, q and r be three non-coplanar lines (that is to say, lines not in the same plane) passing through $(0, 0, 0)$ in $\mathbb{R}^3$, see Fig. 9.6. These

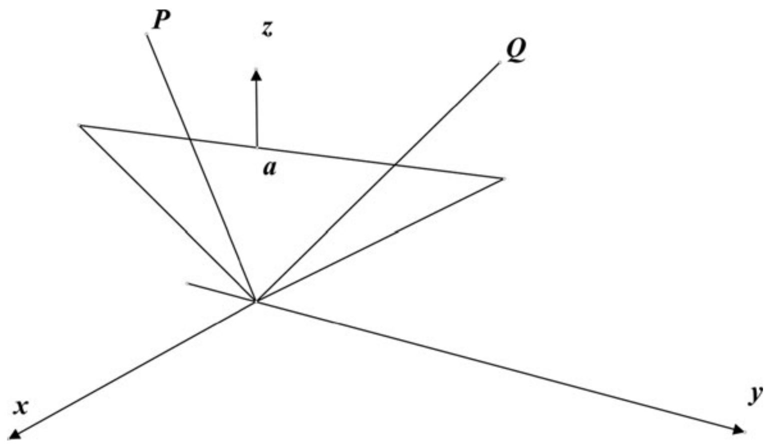


Fig. 9.4 Two points $P \neq Q$ in $\mathbb{P}^2(\mathbb{R})$ determine the line a in $\mathbb{P}^2(\mathbb{R})$

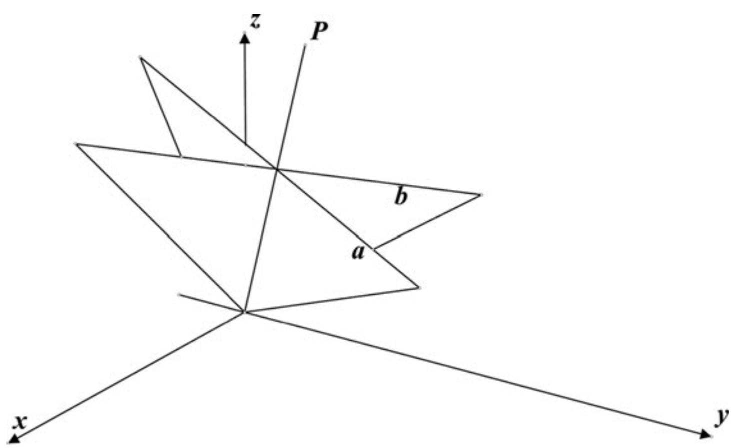


Fig. 9.5 The two lines $a \neq b$ in $\mathbb{P}^2(\mathbb{R})$ determine a unique point P in $\mathbb{P}^2(\mathbb{R})$

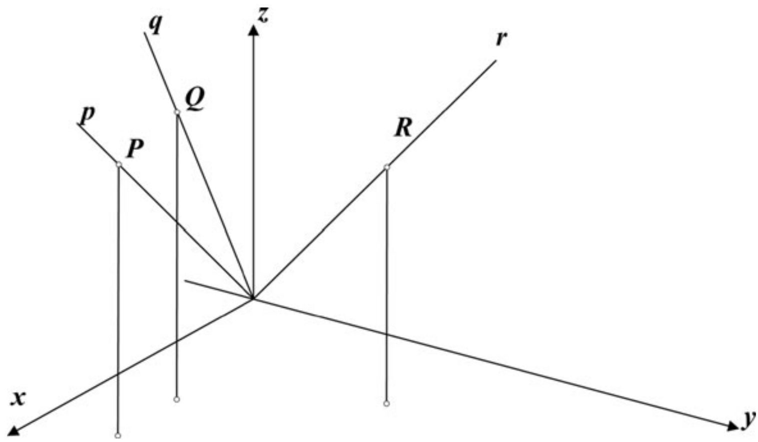


Fig. 9.6 There are three non-collinear points in $\mathbb{P}^2(\mathbb{R})$

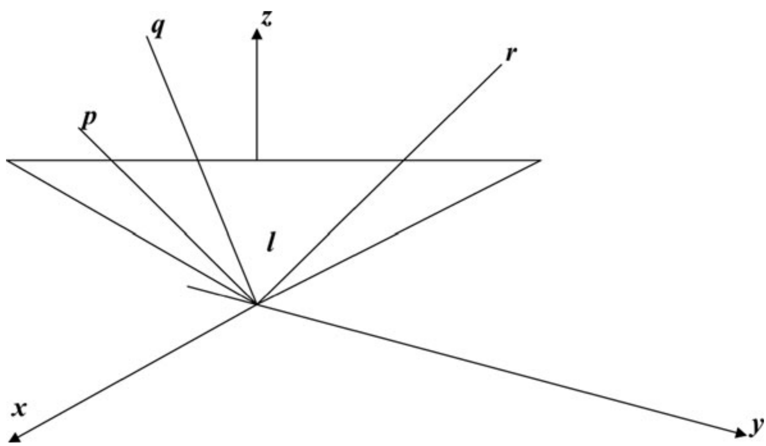


Fig. 9.7 There are at least three points on every line in $\mathbb{P}^2(\mathbb{R})$

lines then represent points, which we rename P, Q, R, which are not on the same line in the model $\mathbb{P}^2(\mathbb{R})$.

Verification of Axiom 9.4: Let ℓ be a line in the model $\mathbb{P}^2(\mathbb{R})$. According to our definition, this is a plane passing through the origin in $\mathbb{R}^3$. Clearly there are at least three different lines p, q, r in that plane, passing through the origin. The situation is shown in Fig. 9.7.

We shall now compare the points in the model $\mathbb{P}^2(\mathbb{R})$ to the usual points of $\mathbb{R}^2$. In fact, we prove that the points in $\mathbb{R}^2$ may be identified with *certain* points in $\mathbb{P}^2(\mathbb{R})$.

We let the point $(x, y) \in \mathbb{R}^2$ correspond to the line which passes through the points $(0, 0, 0)$ and $(x, y, 1)$ in $\mathbb{R}^3$. This line is uniquely determined by the point $(x, y) \in \mathbb{R}^2$, and will be denoted by $a(x, y)$. When we let the point (x, y) run

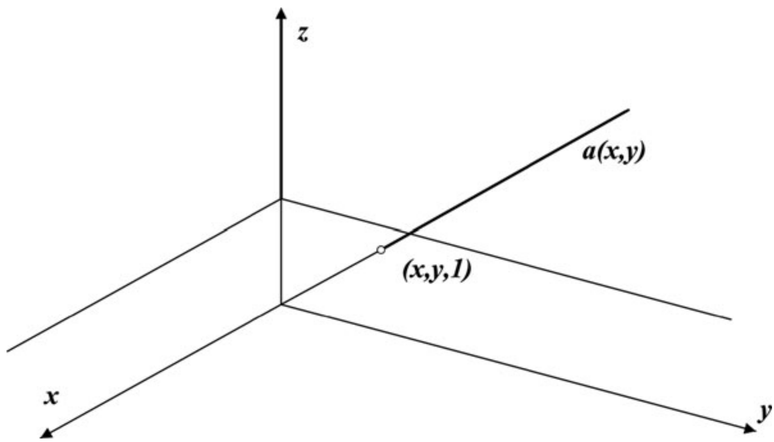


Fig. 9.8 $\mathbb{R}^2$ is augmented by a projective line of points at infinity to form the projective plane $\mathbb{P}^2(\mathbb{R})$: A line through the origin $(0, 0, 0) \in \mathbb{R}^3$ which does not lie in the xy-plane corresponds to the point $(x, y) \in \mathbb{R}^2$, where the coordinates are given by the line meeting the plane $z = 1$ in the point $(x, y, 1)$. The projective line at infinity is then the projective line which corresponds to the plane $z = 0$, the xy-plane

through $\mathbb{R}^2$ then the corresponding line $a(x, y)$ will run through the set of lines through the origin in $\mathbb{R}^3$ which are not contained in the xy-plane. We denote the set of these lines, considered as points in $\mathbb{P}^2(\mathbb{R})$, by $\mathbb{P}^2(\mathbb{R})_0$. We sum up what we have proved as the

Proposition 2. *The mapping which sends the point* (x, y) *to the line* $a(x, y)$ *identifies* $\mathbb{R}^2$ *with* $\mathbb{P}^2(\mathbb{R})_0$.

Denote the complement of the set $\mathbb{P}^2(\mathbb{R})_0$ in $\mathbb{P}^2(\mathbb{R})$ by $\mathbb{P}^2(\mathbb{R})_\infty$. This is the set of points in the model $\mathbb{P}^2(\mathbb{R})$ given by lines which lie in the xy-plane, and those are the "new" points which have been added to $\mathbb{R}^2$. They are thus the *points at infinity*. If we now let (x, y) move outward towards infinity in $\mathbb{R}^2$, then the corresponding line will approach the xy-plane more and more, but never actually quite reach it. The *limiting positions* for these lines will therefore be the lines in the xy-plane, which we view as points at infinity. See Fig. 9.8.

We take a closer look at the points in $\mathbb{P}^2(\mathbb{R})_\infty$, the set of all lines through the origin and contained in the xy-plane. The basic observation is the following: The *xy-plane is simply a line in the model* $\mathbb{P}^2(\mathbb{R})$, whose points, which we have denoted by $\mathbb{P}^2(\mathbb{R})_\infty$, all lie at infinity. The xy-plane is therefore nothing but *the line at infinity in* $\mathbb{P}^2(\mathbb{R})$. We also denote this line by L_∞. Adding this line to $\mathbb{R}^2$ represents a what is called *a compactification* of $\mathbb{R}^2$.

It falls outside the scope to fully explain the term *compactification*, but heuristically the idea is to enlarge a space by adding a boundary. The simplest compactification of the plane $\mathbb{R}^2$ is to add one point, thereby obtaining a sphere, as indicated in Fig. 9.15.

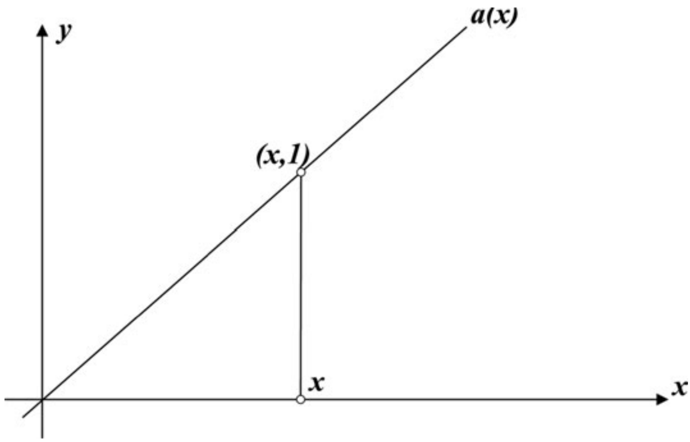


Fig. 9.9 $\mathbb{R}$ is augmented by a point at infinity to form the projective line $\mathbb{P}^1(\mathbb{R})$

The set of points which lie on a fixed line ℓ in $\mathbb{P}^2(\mathbb{R})$ correspond to lines through the origin $(0, 0, 0) \in \mathbb{R}^3$, contained in the plane p which corresponds to the line ℓ. We may identify the plane p with $\mathbb{R}^2$, and then the points on the projective line ℓ correspond to the lines in $\mathbb{R}^2$ passing through the origin $(0, 0) \in \mathbb{R}^2$. Thus we have a completely analogous situation to the points in $\mathbb{P}^2(\mathbb{R})$, except for the dimension being reduced from 3 to 2. Indeed, in the previous situation we had a *projective plane*, whereas we now are dealing with a *projective line*.

Definition 5. The set of lines through $(0, 0)$ in $\mathbb{R}^2$ is called the points on the real, projective line, and is denoted by $\mathbb{P}^1(\mathbb{R})$.

Again, in the same way as we saw for $\mathbb{P}^2(\mathbb{R})$, $\mathbb{P}^1(\mathbb{R})$ may be viewed as $\mathbb{R}$ augmented by a point at infinity, ∞. In Fig. 9.9 this point at infinity corresponds to the x-axis.

The real, projective plane $\mathbb{P}^2(\mathbb{R})$ may be viewed as a *compactification* of $\mathbb{R}^2$, obtained by adding a *boundary* to the surface $\mathbb{R}^2$ consisting of a *real, projective line* at infinity. This real, projective line in turn consists of $\mathbb{R}$, to which there is added one point at infinity, the *one point compactification* of the reals $\mathbb{R}$. That, of course, may be identified with *a circle*, as illustrated in Fig. 9.10.

The lines through $(0, 0) \in \mathbb{R}^2$ may be identified with the points of the semicircle shown in Fig. 9.10, with the exception of the two points on the x-axis, which corresponds to the two diametrically opposite points A and B. When these two points are identified, then the semicircle is joined to a circle, and the set of lines through $(0, 0) \in \mathbb{R}^2$ is identified with it. Thus the real projective line $\mathbb{P}^1(\mathbb{R})$ may be identified with a circle.

The projective plane $\mathbb{P}^2(\mathbb{R})$ may be given a similar interpretation. This is illustrated in Fig. 9.11.

Here we represent a line through $(0, 0, 0) \in \mathbb{R}^3$ by its point of intersection with the northern hemisphere of a fixed sphere with center at the origin. Under this corre-

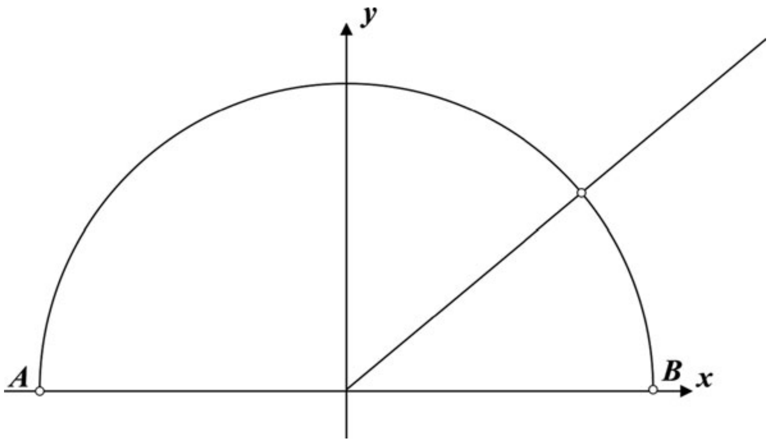


Fig. 9.10 The projective line $\mathbb{P}^1(\mathbb{R})$ may be identified with a circle, as it consists of $\mathbb{R}$ to which is added a single point at infinity

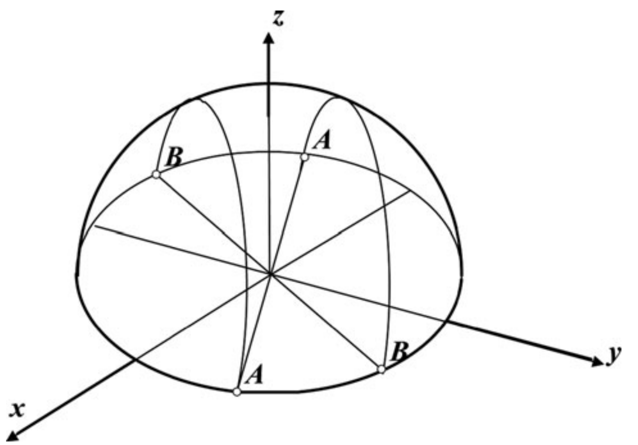


Fig. 9.11 The real projective plane $\mathbb{P}^2(\mathbb{R})$ is the surface obtained from the upper hemisphere, with diametrically opposite points on the equator identified. A highway on the real projective plane will have some curious properties, as explained in the text

spondence the points at infinity, that is to say, the ones corresponding to lines in the xy-plane, correspond to two points, diametrically opposite on the equator. Thus we have to identify diametrically opposite equatorial points, and in doing so we obtain a representation of the surface $\mathbb{P}^2(\mathbb{R})$.

We shall return to the procedure for identifying points later, this technique requires some firm and stringent foundations which we have not yet developed. At this point one may view the identification as an informal explanation.

The surface $\mathbb{P}^2(\mathbb{R})$ has some very interesting properties. Thus for instance, it is a surface with only one side! This is seen as follows. In Fig. 9.11 we may imagine that

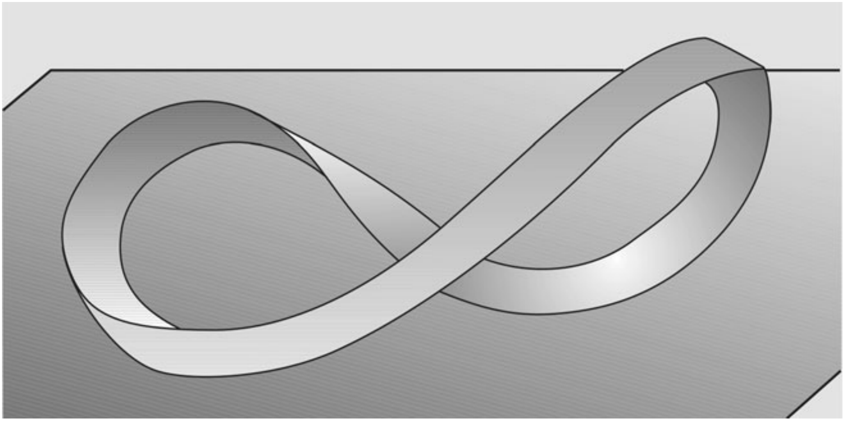


Fig. 9.12 The highway on the real projective plane is nothing but the *Möbius strip*

a highway is constructed, as indicated by the strip drawn over the upper hemisphere shown there. We may now drive off on this highway, approaching the equator. We then actually approach the points at infinity, judged from our initial position. As we move along, however, the perception of where infinity is located changes, so as to always be off in the distance. The line at infinity would be like the rainbow, always moving away as we attempt to chase after it. We will eventually return to the starting point. We will then be driving on the other side of the road, in a somewhat disturbing way: The car will be upside down, *under the pavement.* This is best understood by cutting out the highway from $\mathbb{P}^2(\mathbb{R})$, and examining it more closely. Indeed, we get a strip from the upper hemisphere, fairly approximated by a rectangular strip of paper. Then the diagonally opposite points at the short sides will have to be identified, as they were diametrically opposite equatorial points on the hemisphere. Thus we twist and glue the strip, obtaining the *Möbius strip* shown in Fig. 9.12.

It follows as well, from the considerations above, that the points in the usual Euclidian plane $\mathbb{R}^2$ may be represented by the points of the northern hemisphere, where the equator is not included. Thus the points at infinity, which are added to the Euclidian plane to give the real projective plane, are precisely the points obtained by identifying diametrically opposite points on the equator. This is the projective line of points at infinity. Now any projective line may be identified with a circle, as we have seen. Thus, whenever we identify diametrically opposite points on a circle, we get another circle as result.

When we identify the Euclidian plane $\mathbb{R}^2$ with the upper hemisphere, then the lines are represented as the circles of intersection between the northern hemisphere of the sphere and planes through the origin $(0, 0, 0)$. That is to say, the lines correspond to the northern pieces of *great circles* on the sphere. We thus have a model for Euclidian geometry in which the points are the points on the northern hemisphere, and the lines are the northern parts of great circles. It is not difficult to verify that Euclid's postulates are satisfied by arguing directly with these definitions. However,

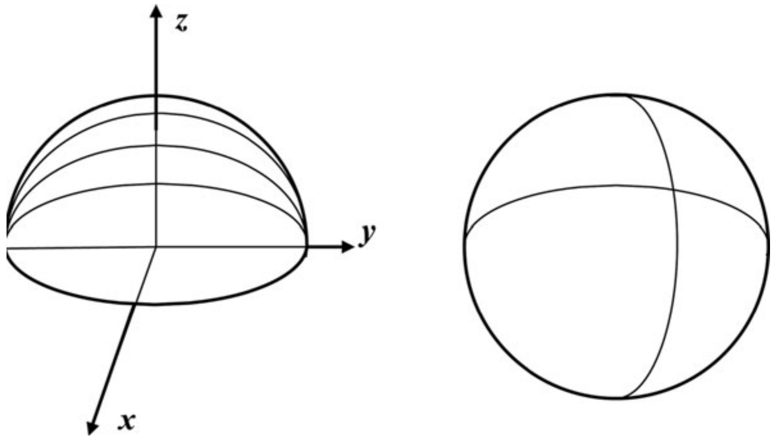
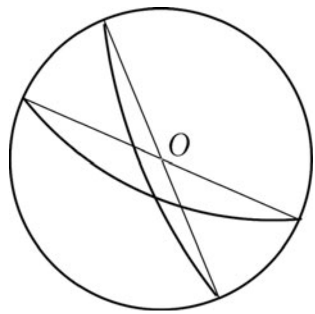


Fig. 9.13 Two models for the Euclidian plane

Fig. 9.14 The model for the Euclidian plane in which the *lines* are *circle arcs*



we of course know this already by the way this last model was constructed from $\mathbb{R}^2$, where the axioms do hold.

We may take this one step further, and *project* this model onto the xy-plane, parallel with the z-axis. This yields a third model for the Euclidian plane, in which the points are the points in the interior of a fixed circle, and the lines are half ellipses, with the longest axis along a diameter of the fixed circle. The points on the boundary of the fixed circle is not included. Figure 9.13 illustrates the situation.

A variation of this model is obtained by letting the lines be circle arcs, with a diameter of the fixed circle as a cord. It is shown in Fig. 9.14.

Finally, there is another way of representing the points of the Euclidian plane: It is to use *all* the points of the entire sphere, *except* for the *North Pole.* Taking out the North Pole leaves us with a *punctured sphere.*

The correspondence between the points in the Euclidian plane $\mathbb{R}^2$ and the points on the punctured sphere is given by *projection with center at the North Pole, N:* A point A on the punctured sphere is mapped to the point $pr_N(A) = B$ obtained as the point of intersection between the xy-plane and the line passing through the points N and A. See Fig. 9.15.

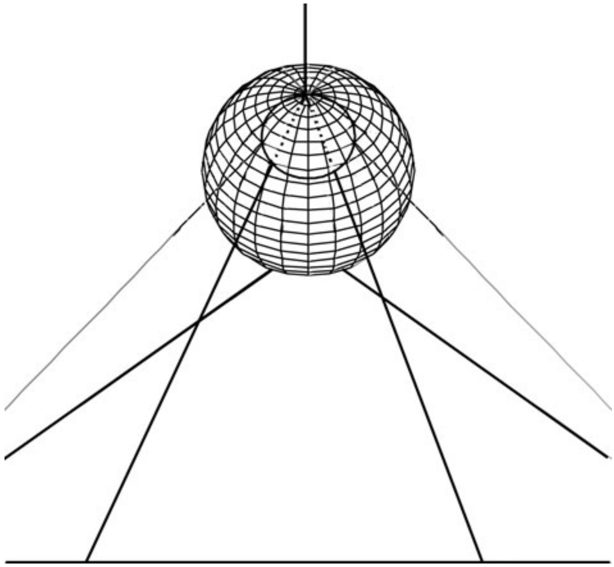


Fig. 9.15 The punctured sphere projected onto the plane

The xy-plane is tangent to the sphere at the origin $(0, 0, 0)$, the South Pole. The line ℓ will correspond to a small circle as indicated, namely to the circle of intersection between the sphere and the plane determined by the line ℓ and the North Pole, the center of projection.

Thus the entire sphere, including the North Pole, may be viewed as another kind of *compactification* of the Euclidian plane $\mathbb{R}^2$ than the real projective plane $\mathbb{P}^2(\mathbb{R})$ which we have already seen. Here we only add *one single point* at infinity. We say that the sphere is the *one point compactification* of $\mathbb{R}^2$.

This is not out of line with the framework of projective geometry. In fact the sphere is nothing but *the projective line over the complex numbers,* which we denote by $\mathbb{P}^1(\mathbb{C})$.

Exercises

Exercise 9.1 Check that the Euler Conjecture is false for $m = 2$.

Exercise 9.2 Let $\mathcal{P}$ be a projective plane of order m. Use $\mathcal{P}$ to construct a set of $m - 1$ mutually orthogonal Latin squares.

Exercise 9.3 Show how the process from Exercise 9.2 may be reversed to construct a set of $m - 1$ mutually orthogonal Latin squares from a projective plane $\mathcal{P}$ of order m.

Exercise 9.4 In the case of the Seven Point Plane there is only one matrix, of order 2. Use the method from the two previous exercises to label the model given in Fig. 9.1 and compute the corresponding Latin square.

Exercise 9.5 Draw a picture with a model for a finite projective plane of order 3, and compute the two mutually orthogonal Latin squares.

Chapter 10
Models for Non-Euclidian Geometry

10.1 Three Types of Geometry

The absolutely most fundamental model in geometry is of course the real Euclidian plane $\mathbb{R}^2$. But there are many others, even alternative ones for Euclidian geometry. Indeed, in the last chapter we saw how we get several models for the Euclidian plane. One of them was given by representing the points as the points in the interior of a fixed circle, and letting the lines be circle arcs with the cord along a diameter of the fixed circle.

We now show how we may realize the two other possibilities, in which the Fifth Postulate of Euclid does not hold. The models we are going to describe do closely resemble the above-mentioned models for Euclidian geometry, but the subtle changes from it make a lot of difference.

The three possibilities were named by Felix Klein, in 1871, as *hyperbolic* geometry, *parabolic* geometry and as *elliptic geometry.* Here parabolic geometry stands for the usual Euclidian plane, whereas the other two will be described in the following sections. The reason for these names comes from Klein's program of classifying geometry according to the groups of transformations leaving its geometric properties unchanged – *invariant.* A further discussion along these lines does, however, fall outside the scope of this book.

10.2 Hyperbolic Geometry

The *hyperbolic plane* is characterized by satisfying all the axioms and postulates of the Euclidian plane, except for the Fifth Postulate, which is replaced by the assertion that

Postulate 10.1 (Hyperbolic plane) *Given a line and a point outside it. Then there are at least two lines through the point which do not meet the line.*

Hyperbolic geometry is a special case of neutral geometry. We obtain a model for the hyperbolic plane by letting the set of points $\mathcal{P}$ be the set of points inside a

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_10,

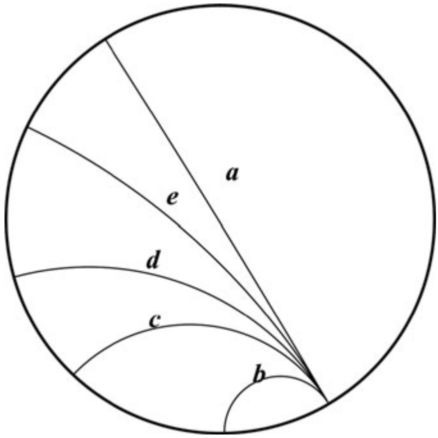


Fig. 10.1 The points in this model for the hyperbolic plane are the points inside a fixed *circle*, and the *lines* are the *circular arcs* which are perpendicular to the fixed *circle*. We show five parallel lines a, b, c, d and e, meeting at "a point at infinity", loosely speaking since this point at the boundary is not part of the model: The lines are parallel, so they do not meet

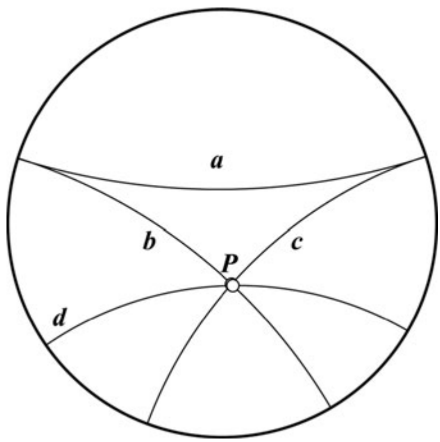


Fig. 10.2 The two bounding parallels to a through P

fixed circle, and the set of lines $\mathfrak{L}$ be circular arcs inside the fixed circle, which are all perpendicular to the fixed circle. The situation is shown in Fig. 10.1.

A point P is incident with such a line if it lies on the circular arc, so we keep the usual concept of incidence. This will be so for all the models discussed in this chapter.

In Fig. 10.2 we consider a line a and a point P which does not lie on a – which is not incident with a. We see the two lines b and c through P which do not meet a, two lines through P parallel with a. The two lines shown are extremal cases as far as parallels to a through P is concerned: Between them we find an infinite number of parallels through P to a, one of them shown and denoted by d. b and c are referred to as *bounding parallels*.

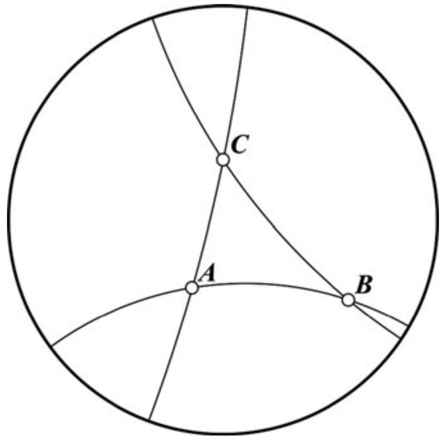


Fig. 10.3 A hyperbolic triangle has angular sum less than two times a right angle

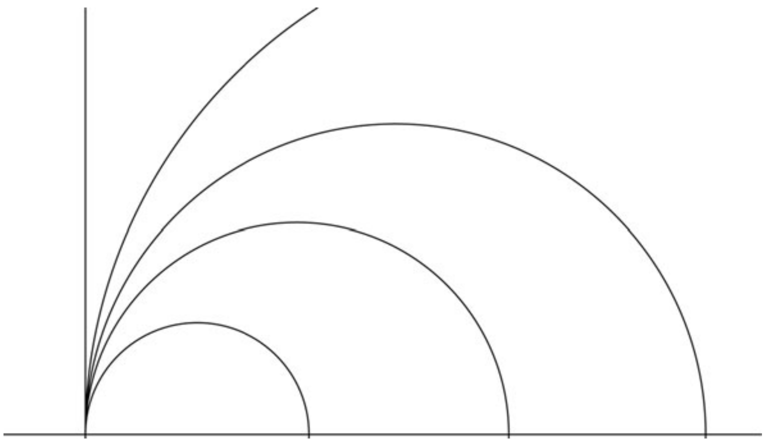


Fig. 10.4 The upper half plane model for the hyperbolic plane

This model is due to *Henri Poincaré*, and it shows some of its beauty by accurately representing the *angle between lines.* Indeed, a distance – or *metric* – may be defined in the model, but it is quite different from the usual distance between points in the plane. On the other hand, the angle which may be defined between lines in hyperbolic geometry, is the one measured between the circular arcs representing them in this model. We express this by saying that Poincaré's model is *conformal*. From this, we see that the angular sum of a hyperbolic triangle is less than two right angles. This is illustrated in Fig. 10.3. We return to the concept of *distance* in this model in Sect. 10.4.

If we keep the center of the circle fixed and let the radius tend to infinity, then this model will approach the usual Euclidian plane. Thus the Euclidian plane is a limiting case of the hyperbolic plane. On the other hand, if we let the radius tend to infinity but keep a point on the circumference fixed, then we get *the upper half-plane model* for the hyperbolic plane, shown in Fig. 10.4.

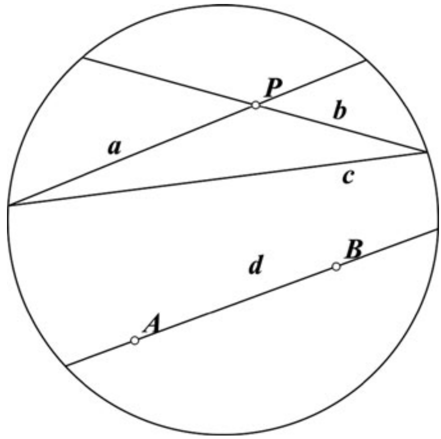


Fig. 10.5 Klein's model for the hyperbolic plane, with two points A and B which determine a unique line d, as well as a point P outside the line c, with the bounding parallels a and b as they are represented in this model

Another model for the hyperbolic plane has been devised by Klein. He also uses the points in the interior of a fixed circle, but defines the "lines" as being all possible *cords* to this fixed circle. This model is illustrated in Fig. 10.5.

Klein's model does not have the appealing property of Poincaré's, in that the angle between lines is faithfully represented. In fact, the angle between the cords are the usual ones from the Euclidian plane, thus using them we would get the angular sum of a triangle equal to two right angles, while as asserted above the angular sum of a hyperbolic triangle is less than this. But there are two significant features of Klein's model. First of all, it is manifestly a model for hyperbolic geometry developed entirely within Euclidian geometry. Thus consistency of Euclidian geometry *implies the consistency of hyperbolic geometry.* This was an important observation at the time when Klein developed the model. It proves that hyperbolic geometry is at least as consistent as the Euclidian counterpart. The latter being universally accepted, this dispels any objections to hyperbolic geometry from a mathematical point of view.

The second nice feature of this model, is that we may give an elegant description of the *distance* between two points. It uses a fundamental and very interesting concept from *projective geometry*. The concept we use is that of the *cross ratio of four collinear points*. We consider two points P and Q in Klein's model, and mark points S and T *outside* the model, namely the points of intersection of the line through P and Q in the ordinary Euclidian plane with the fixed circle (Fig. 10.6).

In general the *cross ratio* of four such points, on the same line, is defined as the fraction

$$[T : S; P : Q] = \frac{\frac{TP}{PS}}{\frac{TQ}{QS}}.$$

Here, for instance, TP is the distance from T to P, normally with a sign, so we endow the line with an orientation, a positive direction. If the cross ratio is -1, then the ratio is called a *harmonic ratio*, if it is 1 then we speak of an *antiharmonic ratio*.

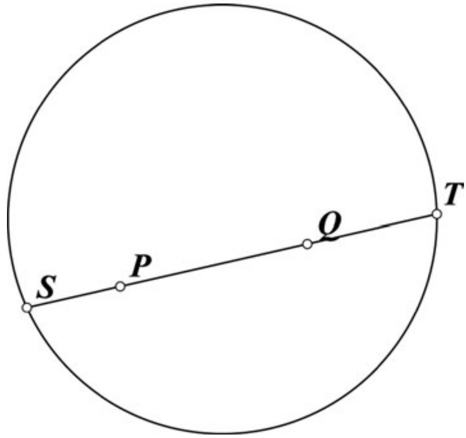


Fig. 10.6 We compute the distance between the points P and Q in Klein's model for the hyperbolic plane as the logarithm of the cross ratio of the four points indicated, as described in the text

In the present situation we shall only consider absolute values, so the cross ratio is always non-negative. Klein now defined the *distance* between P and Q as the absolute value of the logarithm of the cross ratio:

$$d(P, Q) = \left|\log\left(\frac{TP}{PS}\Big/\frac{TQ}{QS}\right)\right|.$$

If one of the points P or Q approaches the rim of the fixed circle while the other stays put, then $d(P, Q) \longrightarrow \infty$.

The general concept of distance will be discussed in Sect. 10.4.

Finally, we recall from Sect. 4.1 the following alternative version of the Parallel postulate, due to Gauss:

There exists a triangle, the contents of which is greater than any given area.

Thus in a hyperbolic plane there is an upper bound to the areas of triangles.

10.3 Elliptic Geometry

The version of the Fifth Postulate which defines *elliptic geometry* is the following:

Postulate 10.2 (Elliptic plane) *Given a line ℓ and a point P outside it. Then all lines through P meet ℓ.*

We may not introduce this axiom into the system of Hilbert, minus the parallel postulate: Elliptic geometry is not a special case of neutral geometry. But by a suitable modification of some of Hilbert's other axioms, we get a firm axiomatic base for elliptic geometry as well.

We have already constructed a model for this geometry, namely $\mathbb{P}^2(\mathbb{R})$, the real projective plane. As we have seen before, we may identify $\mathbb{P}^2(\mathbb{R})$ with the points

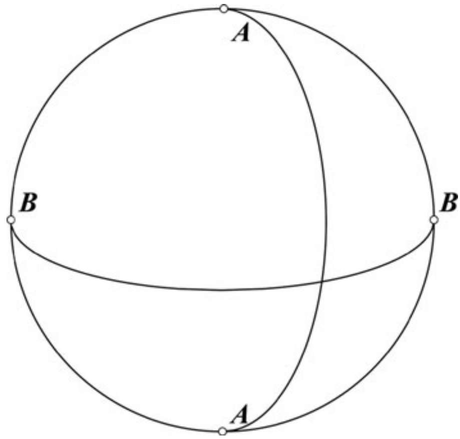


Fig. 10.7 Model for the elliptic plane. Diametrically opposite points on the circumference are identified. The lines are half ellipses, end points identified, where the longest axis coincides with a diameter in the fixed circle

inside a fixed circle, with the points on the circumference included this time, but with diametrically opposite points identified. The set of *lines* in this model is the set of half ellipses, end points identified, where the longest axis coincides with a diameter in the fixed circle. See Fig. 10.7. This model for elliptic geometry is attributed to Klein.

As was the case for Poincaré's model for hyperbolic geometry, Klein's model for elliptic geometry is conformal. Using this, we see that in the elliptic plane the angular sum of any triangle is greater than two right angles.

We do not go into the details of distance and angular measure in the hyperbolic and the elliptic plane. Instead we refer to Greenbergs interesting book [21]. But by building on Hilbert's axioms with the appropriate versions of the Parallel Postulate,[1] we get geometries with all features from the Euclidian case, except of course Euclid's Fifth Postulate. In particular there is distance and angular measure. In the Euclidian and hyperbolic cases there is an absolute measure of angles, namely the *radian*. Analogously, in elliptic geometry there is an absolute measure of *length*, in other words of distance. All lines in the elliptic plane have the same finite length. Of course right angles exist in all three versions of geometry.

10.4 Euclidian and Non-Euclidian Geometry in Space

We may construct models for Euclidian, hyperbolic and elliptic *space* by letting the points be all points in the interior of a fixed sphere in the Euclidian and hyperbolic case, and in the elliptic case with the addition of the points at the surface with diametrically opposite points identified. The lines are defined as for the corresponding planes.

[1] And, as already noted, some additional modifications in the elliptic case.

The concept of a metric is quite general, and it is one of the most fruitful abstractions undertaken in the development of modern mathematics. So we shall take a few moments to explain it.

We start out with some set S. We make no assumptions about the set S whatsoever, the abstract concept of a metric may be defined in complete generality. If it is to be of any use, however, we need to restrict our attention somewhat, but this is a subject which we shall not pursue here.

A metric defined in S is a *real valued function*, $d(P, Q) \in \mathbb{R}$ of two variables $P, Q \in S$, such that the following is true:

Distance is always non-negative: For all $P, Q \in S$,

$$d(P, Q) \geq 0.$$

If two points are different, then the distance between them is never zero. In an equivalent formulation:

$$d(P, Q) = 0 \Longleftrightarrow P = Q.$$

The distance from me to you is of course the same as the distance from you to me:

$$d(P, Q) = d(Q, P).$$

The final *axiom for a metric* is usually referred to as the *triangular condition* – the distance travelled from point P to point R is never shortened by passing by some third point Q:

$$d(P, R) \leq d(P, Q) + d(Q, R).$$

In the Euclidian plane $\mathbb{R}^2$ the metric is the usual distance between two points. Of course one readily sees that the three axioms above are satisfied in this case. More formally, we define the metric by writing the function explicitly in terms of the coordinates of the points involved. Denoting the metric in the Euclidian plane by ρ_2, we have

$$\rho_2(P, Q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2},$$

where $P = (x_1, y_1)$ and $Q = (x_2, y_2)$. It is an easy exercise in applying the Pythagorean theorem to verify that this yields the usual distance, namely the length of the line segment joining the two points.

Quite analogously, the metric in the Euclidian 3-space $\mathbb{R}^3$ is the usual distance, the length of the line segment joining the two points. This metric is given by the formula

$$\rho_3(P, Q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2},$$

where now

$$P = (x_1, y_1, z_1) \text{ and } Q = (x_2, y_2, z_2).$$

Again, this is the usual distance. The proof is not quite as simple as in the planar case, here we need to appeal to Pythagoras *twice*.

In general a metric space is defined as a set S, in which there is given a metric as explained above. The set $\mathbb{R}^n$, which consists of all n-tuples $(x_1, x_2, \ldots, x_n)$ where $x_1, \ldots, x_n$ are all real numbers, can also be made into a metric space by defining a metric by

$$\rho_n((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

It is easy to see that the first two axioms for a metric are satisfied. The third is a little bit more tricky to verify, and is left as a challenge to the reader. ρ_n is referred to as the *Euclidian distance* in $\mathbb{R}^n$.

The *abstract notion of a metric space*, which we have explained above, admits many *"spaces"* which we intuitively would reject as both strange and unusual. Thus for instance, we may take *any set S*, say the set of all US senators, and make it into a *metric space* by defining the distance σ as follows:

$$\sigma(P, Q) = \begin{cases} 0 \text{ if } P = Q \\ 1 \text{ if P } \neq Q. \end{cases}$$

The metric thus defined clearly satisfies the three axioms. We refer to this metric as the *discrete metric* on the set S. This metric space clearly provides no real information beyond the description of the set S. To pursue this digression one step further, one might try to introduce politics in this, and instead define the distance between two senators from the same party to be 0, and to be 1 for senators from different parties. But this does not yield a metric space. Rather, we obtain what is known as a *pseudo-metric* space.

In the model for Euclidian space, the points are the points in the interior of a fixed sphere. We may introduce a metric in this set, given by a somewhat complicated formula which will not be given here. But with this metric the interior of the fixed sphere becomes a metric space, and if we let the radius of the fixed sphere tend to infinity, then the metric we defined for the interior of the sphere will approach the usual distance in Euclidian 3-space.

If we imagine that beings in this universe inside the fixed sphere starts at the center, with a spaceship 100 m long, and travels along towards the boundary at a fixed speed, then an observer outside the sphere would find that 1. The spaceship would shrink in size as the boundary was approached, and 2. The speed would appear to decrease towards zero as the boundary was approached. To the beings inside the universe, however, the size of the spaceship as well as its speed would appear unchanged. Thus they would have no way of knowing that they actually lived in a *bounded universe*. In fact, from their point of view their universe would appear unbounded. Only for the outside observer would the boundedness of this universe be manifest. In this model the lines are halves of ellipses, with the longest axis along a diameter of the space, analogous to what we have for the planar model. The *planes* will be halves of ellipsoids, as shown in the illustration given in Fig. 10.8.

There is a close connection between the metric and the curves designated as *lines*. Indeed, the line-segment joining the two points P and Q is precisely the path

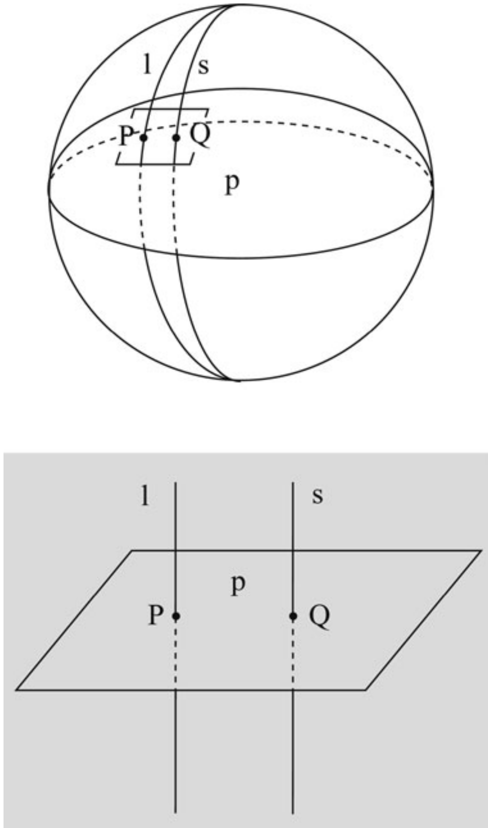


Fig. 10.8 The Euclidian universe inside a fixed sphere. The half ellipsoid with two of its axes along two diameters of the sphere, is a "plane" in this model and is denoted by p. Two "lines" l and s perpendicular to the plane p are shown, intersecting the plane p in the two points P and Q indicated. They also illustrate the Fifth Postulate of Euclid: l and s are parallel, being both perpendicular to the same p, and the line s is the unique parallel to l which passes through the point Q. Below we have cut out a small piece of this space, locally it looks like the normal Euclidian space

yielding the shortest route from P to Q. As *light* will travel the shortest route from one point to another, the lines are therefore nothing but all the possible *light rays* we may have in the universe we are studying. Another name for a such curve is *a geodesic curve* or just *a geodesic*.

The explicit formula for the metric is complicated in this and most other serious cases, and is therefore usually omitted. We may do so because of the following important phenomenon: Suppose that our set S is a subset of $\mathbb{R}^3$ (or $\mathbb{R}^n$, for that matter). It then turns out that to specify a metric is equivalent to prescribing how a measuring rod of unit length will shrink or expand as it is moved around within the space. In the present case this means, moved towards the surface of the sphere, the boundary of the universe. Having a formula for this "shrinkage" makes it possible to reconstruct the metric, and indeed as it turns out, this formula for the shrinkage of a small measuring rod is a more convenient tool for the description of the geometry of a metric space than the distance-formula itself. The reader should note that this explanation is of an informal nature, thus mathematically deficient in many respects. We will, however, return to it in the following section.

We now turn our attention to the *hyperbolic universe*. Again, the points of this universe are the points inside a fixed sphere, of radius R, say.

In this set we may introduce a metric by specifying that the size of a measuring rod of unit length, located at a point $P = (x, y, z)$ of distance $r = \sqrt{x^2 + y^2 + z^2}$ from the center, is equal to

$$\delta_P = 1 - \frac{r^2}{R^2}.$$

We see that as long as the measuring rod is located at the origin, deemed as the center of the universe by the outside observer, then our "universe-dwellers" inside the sphere and the outside observer will see it as being of the same length, namely 1.

But as the universe-dwellers take the rod and move away from the center, a disagreement developers: While it retains the length of 1 as far as the universe-dweller is concerned, the outside observer will see it as shrinking. Correspondingly, even if the insider has kept up a decision to move at a constant speed, the outsider will perceive the speed as diminishing when the distance from the center increases.

It can be shown that the shortest route between two points P and Q in this universe is attained by following a circular arc which passes through P and Q, and in addition is perpendicular to the surface of the sphere.

This is the "lines" we have already described for Klein's model for the hyperbolic plane. So these are the geodesic curves, the "straight lines", in the hyperbolic universe. The situation is illustrated in Fig. 10.9.

We similarly obtain a model for the *elliptic universe* (Fig. 10.10). Here we include the points on the surface of the sphere, but with diametrically opposite points identified. The "lines" described in the Euclidian case are kept in this model for the elliptic universe, except that we add to the line the one point obtained by identifying the two points where the half ellipse meets the surface of the sphere.

Again, this elliptic 3-space is closely related to the elliptic plane, which has been explained earlier.

We may introduce a metric in this space as well, and the "lines" of the model then become the geodesics.

10.5 Riemannian Geometry

Among the concepts and properties which Euclid did not state, but implicitly took for granted, we find the notion of *distance* or as we say today, the metric discussed in the preceding section. Tied to this concept is the notion of *transformations* preserving the metric, as well as transformations preserving *shape* or *shape and size*

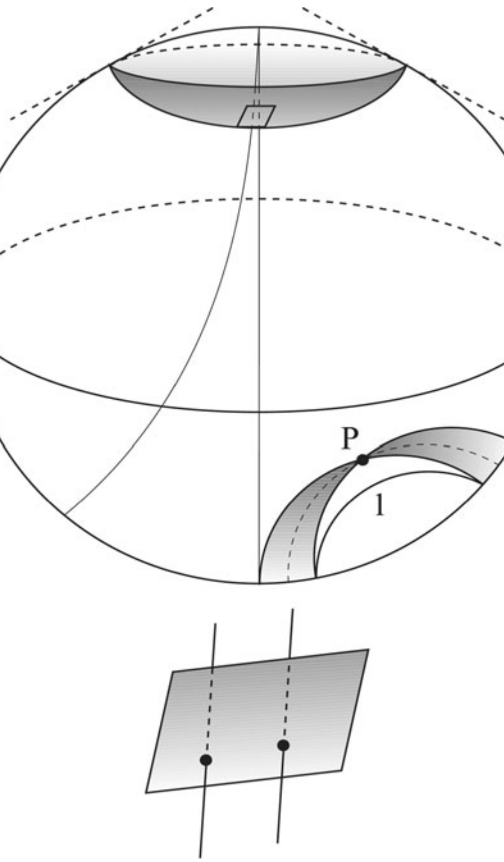


Fig. 10.9 The hyperbolic universe inside a fixed sphere. Lines are circular arcs, perpendicular to the surface of the sphere. Two lines perpendicular to the shaded hyperbolic plane are shown, one of them happens to be a diameter to the sphere, a special case of a hyperbolic line. As in the Euclidian case above we have shown a small piece of the space around where the two lines pass through the plane. In this universe there are infinitely many parallels to a given line ℓ through a point P outside it, the situation is shown in the lower right corner of our universe

of geometric objects: Any figure may be moved without altering its shape, form or size.

Making this precise, turning these ideas into mathematics, has been on the agenda of a number of renowned mathematicians.

Another towering mathematician, *Georg Riemann*, presented his probationary lecture at the University of Göttingen in 1854 (Fig. 10.11). Here he outlines a new course in the understanding of space and geometry.

Riemann's very fundamental idea has already been informally hinted at in the previous section, in the form of the *shrinking measuring rod*. His idea was to express the *distance between two infinitesimally close points*. We first illustrate this idea by

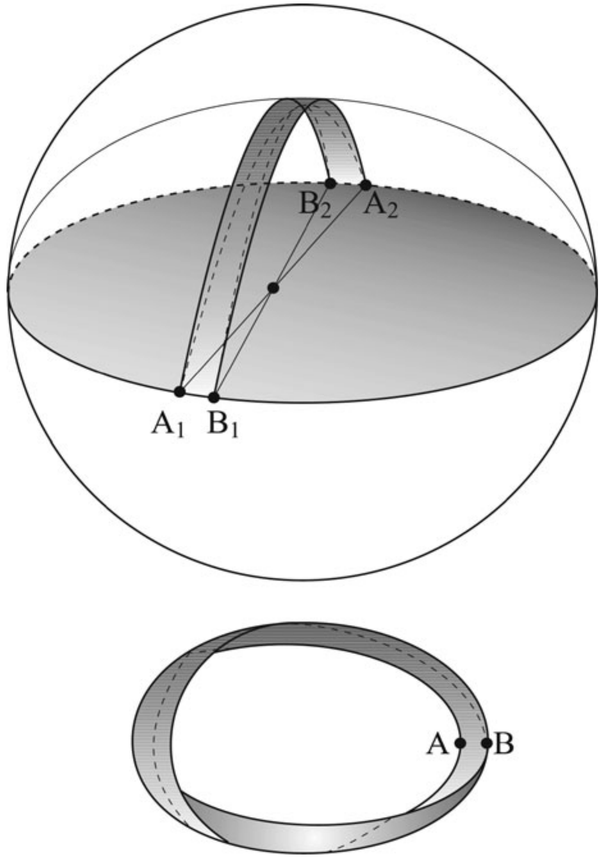


Fig. 10.10 The elliptic universe inside a fixed sphere. Here points on the surface of the sphere are in the space, diametrically opposite ones being identified. Thus the lines are closed curves, of a finite length. Here A_1 is identified with A_2, B_1 with B_2, and we get the indicated "highway". Its edges are not straight lines. But we can see two straight lines in the illustration, namely half ellipses joining A_1 to A_2 and B_1 to B_2. The situation is shown below the sphere

working with the space $\mathbb{R}^3$. Here the usual distance between two points

$$P = (a_1, b_1, c_1) \text{ and } Q = (a_2, b_2, c_2)$$

is give by

$$\rho_3(P, Q) = \sqrt{(a_2 - a_1)^2 + (b_2 - b_1)^2 + (c_2 - c_1)^2}$$

as we have seen in the previous section. For our purpose now it is more practical to write this definition as follows:

$$\rho_3(P, Q)^2 = (a_2 - a_1)^2 + (b_2 - b_1)^2 + (c_2 - c_1)^2.$$

Fig. 10.11 Georg Friedrich Bernhard Riemann, picture from 1868

If we let the two points approach each other, to become very close, this relation may be written as

$$(\Delta s)^2 = (\Delta x_1)^2 + (\Delta x_2)^2 + (\Delta x_3)^2.$$

As we pass to the infinitesimal, we get the expression

$$\mathrm{d}s^2 = \mathrm{d}x_1^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2.$$

This relation is very useful when we wish to calculate the length of a curve-segment. Suppose a curve-segment C is given on parametric form as

$$x_1 = \varphi(t), x_2 = \chi(t), x_3 = \psi(t), t \in [t_1, t_2].$$

Then the length of an infinitesimal piece of the curve is given by

$$ds = \left(\sqrt{\frac{\mathrm{d}\varphi}{\mathrm{d}t}(t)^2 + \frac{\mathrm{d}\chi}{\mathrm{d}t}(t)^2 + \frac{\mathrm{d}\psi}{\mathrm{d}t}(t)^2}\right)\mathrm{d}t.$$

and hence the length of the curve-segment is given by the integral formula, well known from calculus to some of the readers:

$$L = \int_{t_1}^{t_2} \left(\sqrt{\frac{\mathrm{d}\varphi}{\mathrm{d}t}(t)^2 + \frac{\mathrm{d}\chi}{\mathrm{d}t}(t)^2 + \frac{\mathrm{d}\psi}{\mathrm{d}t}(t)^2}\right)\mathrm{d}t.$$

Frequently a formula as above for ds is referred to as a metric. A general Riemannian metric on $\mathbb{R}^3$ is of the form

$$\begin{aligned}\mathrm{d}s^2 &= g_{1,1}\mathrm{d}x_1^2 + g_{1,2}\mathrm{d}x_1\mathrm{d}x_2 + g_{1,3}\mathrm{d}x_1\mathrm{d}x_3 \\ &\quad + g_{2,1}\mathrm{d}x_2\mathrm{d}x_1 + g_{2,2}\mathrm{d}x_2^2 + g_{2,3}\mathrm{d}x_2\mathrm{d}x_3 \\ &\quad + g_{3,1}\mathrm{d}x_3\mathrm{d}x_1 + g_{2,3}\mathrm{d}x_2\mathrm{d}x_3 + g_{3,3}\mathrm{d}x_3^2,\end{aligned}$$

where the $g_{i,j}$ are functions of x_1, x_2, x_3 in general. As $\mathrm{d}x_i\mathrm{d}x_j = \mathrm{d}x_j\mathrm{d}x_i$, we may adjust the $g_{i,j}$'s such that $g_{i,j} = g_{j,i}$ for all i and j.

For the reader with some knowledge of linear algebra we note the convenient organization of these functions in a *symmetric* 3×3 *matrix*:

$$\mathbf{g} = \begin{bmatrix} g_{1,1} & g_{1,2} & g_{1,3} \\ g_{2,1} & g_{2,2} & g_{2,3} \\ g_{3,1} & g_{3,2} & g_{3,3} \end{bmatrix}.$$

Here $g_{i,j} = g_{j,i}$. The metric may expressed as

$$ds^2 = [\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3]\mathbf{g}\begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \\ \mathrm{d}x_3 \end{bmatrix}.$$

In Riemannian geometry it was assumed that the determinant of **g** be non zero, and that $\mathrm{d}s^2 > 0$. This last condition is, however, abolished when these ideas are used in Einstein's General Relativity, which we come to later.

In physics and applied mathematics the matrix **g** is usually referred to as the *(covariant) metric tensor.* We shall not use this language here, the concept of covariant and contravariant tensors is with some justification perceived as rather murky and obscure when it is first encountered. However, solid knowledge of linear algebra makes it crystal clear and perfectly obvious.

The generalization of these ideas to $\mathbb{R}^n$ is rather immediate, we write

$$ds^2 = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} g_{i,j}dx_i dx_j,$$

where as before one assumes that $g_{i,j} = g_{j,i}$ for all i and j.

Frequently the space under consideration is not simply some subset of $\mathbb{R}^n$, but rather is one which may be *patched together* of small pieces of Euclidian n-space $\mathbb{R}^n$. Thus for instance, a spheric surface can be viewed as being pieced together of small patches from $\mathbb{R}^2$, as may a toric surface, or more generally a "torus with n holes in it". Such a surface is called *a compact Riemann surface of genus n*. This is illustrated in Fig. 10.12.

Each patch from Euclidian n-space comes with the coordinates $x_1, \ldots, x_n$, which are now only valid within each patch. When two patches overlap, it is necessary to have a set of *transition functions* between them. That is to say, in the intersection between the patch U with coordinate functions $x_1, \ldots, x_n$ and V with $y_1, \ldots, y_n$,

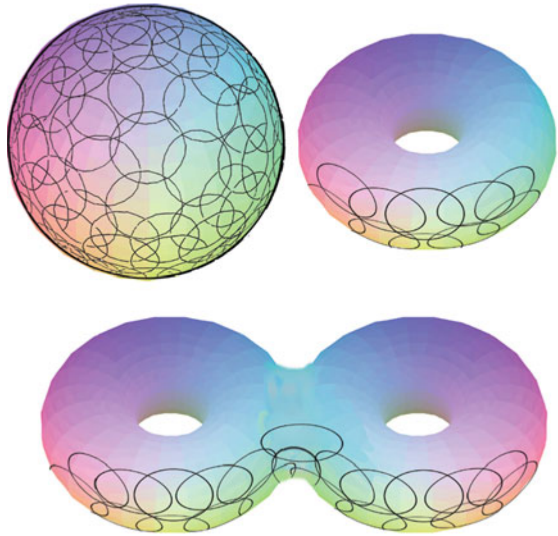


Fig. 10.12 Some surfaces as pieced together by patches from $\mathbb{R}^2$: A sphere, a torus (a compact Riemann surface of genus 1) and a compact Riemann surface of genus 2

we have

$$\begin{aligned} x_1 &= \tau_1(y_1, \ldots, y_n) \\ x_2 &= \tau_2(y_1, \ldots, y_n) \\ &\cdots \\ x_i &= \tau_i(y_1, \ldots, y_n) \\ &\cdots \\ x_n &= \tau_n(y_1, \ldots, y_n). \end{aligned}$$

The mathematics involved in getting all the details of this right is too complicated to explain here, but the metric of this patched together-space is given on each patch as above, the **g** only being valid within its patch. Then the transition to an overlapping patch is expressed in terms of the transition functions.

In terms of the metric tensor **g** we may express the so-called *Riemannian curvature* of the space. We shall not be specific in this exposition, only mention that while Euclidian space has curvature zero, elliptic space will have a positive curvature while hyperbolic space has negative curvature.

A remarkable application of these ideas may be found in *Einstein's theory of general relativity.* As we have seen, the *straight lines* must be the *geodesic curves* in the space, that is to say the path travelled by *light.* Thus light is intimately tied to the geometry of space, and one remarkable consequence is the mathematical necessity of regarding the speed of light as a *universal constant*, determined by the geometry of space itself. This had indeed been observed for some time prior to Einstein's theory, and had been regarded as a puzzling paradox.

Einstein also realized that the concept of *universality of time* had to be abandoned. Instead, time is incorporated into the geometry of space, as a fourth dimension.

The 4-dimensional space-time universe is conceived as a subspace of $\mathbb{C}^4$, 4-dimensional complex space. The four real dimensions in it are arranged with the three spatial axes along the first three real axes in the three first complex dimensions, and the fourth, the time-axis, along the complex axis of the fourth. The basic transformations from one coordinate system to another are *rotations* in this 4-dimensional complex space.

We will not pursue these ideas here, as they require more technical mathematics than the scope of the present book warrants.

Chapter 11
Making Things Precise

11.1 Relations and Their Uses

In Sect. 9.3 we saw how a model for the projective plane may be constructed by taking the northern hemisphere of a spherical surface, including the equator, and then *identifying* diametrically opposite points on the equator.

For a construction of this type to make sense mathematically, we need a stringent base for being able to identify points in this way. We simply *cannot* identify points according to any odd rule we may think of. We need the following important concept:

Definition 6 (Equivalence relation). Let $\mathcal{M}$ be a set, where there is given a relation $m \sim n$, which may be satisfied between two elements m and n from $\mathcal{M}$. We say that $\sim$ is an *equivalence relation* if the following conditions are satisfied:

$$\begin{array}{ll} (R) & \text{For all } m \in \mathcal{M} \text{ we have that } m \sim m \\ (S) & m \sim n \Longrightarrow n \sim m \\ (T) & m \sim n \text{ and } n \sim r \Longrightarrow m \sim r \end{array}$$

In this definition (R) is referred to as *reflexivity*, (S) as *symmetry* and (T) is called *the condition of transitivity*.

These three important properties express the essence of the identification process: The relation $\sim$ will subdivide the set $\mathcal{M}$ into a collection of disjoint *classes of mutually equivalent* elements. We have the following

Proposition 3. *Let $\sim$ be an equivalence relation in $\mathcal{M}$. We put*

$$[m] = \{r \in \mathcal{M} |\ r \sim m\},$$

in other words, the set $[m]$ consists of all elements r in $\mathcal{M}$ such that the relation $r \sim m$ holds true. For two arbitrary elements m and $n \in \mathcal{M}$ we then have that

$$[m] = [n] \text{ or } [m] \cap [n] = \emptyset,$$

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_11,

that is to say, two sets in the subdivision which are distinct are disjoint. Finally, we have that

$$[m] = [n] \text{ if and only if } m \sim n.$$

Definition 7. $[m]$ is referred to as the *equivalence class* of the element m.

Proof. We first prove the last biimplication.

Proof for $\Longrightarrow$: Assume that [m] = [n]. Since $n \in [n]$ because of (S), we get $n \in [m]$, thus $n \sim m$ by the definition of $[m]$ stated in the proposition. But by (S) this gives $m \sim n$.

Proof for $\Longleftarrow$*:* Assume that $m \sim n$. We shall prove that $[m] = [n]$. For this we have to prove the two inclusions $[m] \subseteq [n]$ and $[n] \subseteq [m]$. From our assumption that $m \sim n$ we know by (S) that we also have $n \sim m$, so it suffices to prove the first inclusion, the other then following simply by interchanging the roles of m and n. So we prove that

$$m \sim n \Longrightarrow [m] \subseteq [n].$$

So let $r \in [m]$, that is to say, assume that $r \sim m$. Out assumption that $m \sim n$ then gives $r \sim n$, on account of (T). Thus we have $r \in [n]$, as claimed.

We next prove the first part of the proposition, namely that $[m]$ and $[n]$ either coincide or else have no common element. Assume that $[m]$ and $[n]$ are different subsets, but nevertheless have a common element, say r:

$$r \in [m] \cap [n].$$

We shall prove that this leads to a contradiction: By assumption we first of all have that $r \sim m$ and $r \sim n$. But by (S) we then get that $m \sim r$, which together with $r \sim n$ gives $m \sim n$ because of (T). But above we have already shown the biimplication in the assertion of the proposition, so this gives

$$[m] = [n]$$

which is a contradiction. □

11.2 Identification of Points

Whenever we have a set with an equivalence relation $\sim$, we may subdivide the set $\mathcal{M}$ in a *set of equivalence classes*. This new set of equivalence classes is denoted by $\mathcal{M}/\sim$:

$$(\mathcal{M}/\sim) = \{[m] \mid m \in \mathcal{M}\}.$$

The set $\mathcal{M}/\sim$ is thus obtained by *identifying equivalent points* in $\mathcal{M}$. We have an important surjective mapping

$$\mathcal{M} \longrightarrow \mathcal{M}/\sim$$
$$m \mapsto [m]$$

which *carries out the identification of equivalent points in* $\mathcal{M}$.

We note that the classes need not have the same number of elements, and indeed, some of the classes could be infinite, others finite and some could even consist of a single element. We consider a few examples:

1. The relation of being equal, $=$, is an equivalence relation, the three conditions of course being obvious in this case.
2. The relations of ordering $\geq$ and $\leq$ defined for real numbers are not equivalence relations. They do satisfy (R) and (T), but not (S).
3. The so-called strong relations of ordering, $<$ and $>$ only obey (T), and they therefore are also not equivalence relations.
4. For any set Ω we can define the set of all subsets, denoted $\mathcal{P}^{\Omega}$. In this set we have the various inclusion-relations $\subseteq$, $\supseteq$, $\subset$ and $\supset$. They are not equivalence relations either, but behave as the relations of ordering.
5. In the set $\mathcal{P}^{\Omega}$ we put $P \sim Q$ if P has the same number of elements as Q, which may be ∞. This is an equivalence relation on $\mathcal{P}^{\Omega}$.
6. We consider the set of all smooth surfaces in $\mathbb{R}^3$. If S and T are two such surfaces, we put $S \sim T$ if S can be deformed into T smoothly without breaking anything. This is an informal description of an important equivalence relation known as *topological equivalence.*
7. We say that two geometric figures in the pane $\mathbb{R}^2$ or in the space $\mathbb{R}^3$ are *congruent* if one may be placed on the other by a translation followed by a rotation. This is an equivalence relation.[1]
8. We say that two geometric figures in $\mathbb{R}^2$ or in $\mathbb{R}^3$ are *similar* if one may be placed on the other by a translation followed by a rotation and a shrinkage or an enlargement. This also is an equivalence relation.

We now draw the conclusion which was announced as part of our motivation for introducing the machinery of equivalence relations (Fig. 11.1). Namely, we show how we may construct the set of points which went into the model for the projective plane, by identifying the appropriate points on the northern hemisphere, equator included:

Proposition 4. *Let $\mathcal{M}$ be the set of points on the northern hemisphere, including the equator. We define the relation $\sim$ by setting $P \sim Q$ if and only if*

$P = Q$ or P and Q are diametrically opposite points on the equator.

This is an equivalence relation.

Proof. Except for the case when some point is on the equator, this is nothing but the relation $=$, and the three conditions are clear outside the equator. Moreover, whether P is on the equator or not, clearly $P \sim P$, so (R) holds.

[1] According to this definitions a "mirror image" F' of a figure F, more precisely a reflection of F in a line for $\mathbb{R}^2$ or in a plane for $\mathbb{R}^3$, is not congruent to F. Some authors regard a figure in the plane and its mirror image as being congruent, however. One might say that a figure in $\mathbb{R}^2$ and its mirror image are congruent as figures in the larger space $\mathbb{R}^3$, but not as figures in $\mathbb{R}^2$.

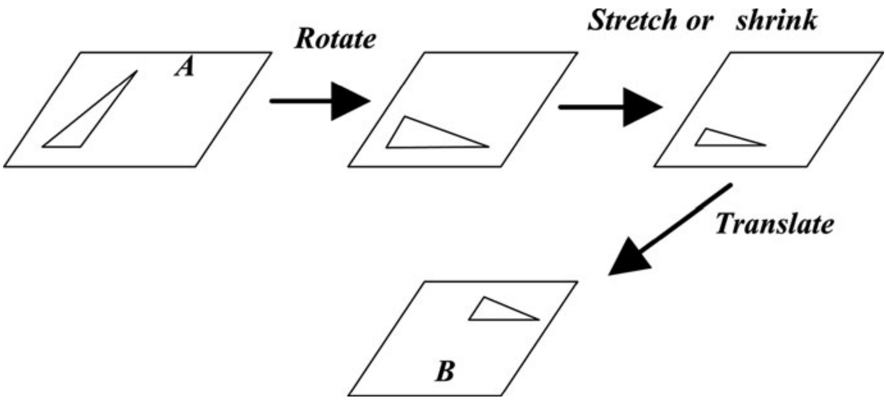


Fig. 11.1 The equivalence relation of similarity applied to two triangles: The triangle A is similar to B

If P is on the equator and $P \sim Q$, then Q must be on the equator, and it is either equal to P or diametrically opposite to P. In either case it follows that $Q \sim P$, thus (S) is proven.

Finally we show (T), assume $P \sim Q$ and $Q \sim R$. If P is not on the equator, then $P = Q$, and so $P \sim R$. So assume that P is on the equator. If $P = Q$, we are finished as before. If not, then P and Q are diametrically opposite points. Then either $R = Q$ or $R = P$, in both cases it follows that $P \sim R$. □

11.3 Our Number System

One of the most amazing applications of the concept of equivalence relations, is the systematic and mathematically rigorous construction of our system of numbers, starting from the natural numbers

$$\mathbb{N} = \{1, 2, 3, \dots\},$$

which we take as intuitively given. Actually, they may be constructed too, from the Zermelo–Fraenkel–Skolem axioms for the Set-theory. But we take the natural numbers as given intuitively, as has been the case from time immemorial.

We proceed one step at the time: First the set of natural numbers $\mathbb{N}$ is extended to the *set of integers* $\mathbb{Z}$. Thereafter $\mathbb{Z}$ is extended to the *rational numbers* $\mathbb{Q}$, which again is extended to the *real numbers* $\mathbb{R}$. Finally we briefly introduce the complex numbers $\mathbb{C}$.

Passing from the natural numbers, the set of which is denoted by $\mathbb{N}$, to the integers $\mathbb{Z}$ was historically accomplished in two steps: First, the discovery of the integer *zero* was more important and more difficult than most of us realize today.

Secondly, the concept of *negative numbers* was historically very difficult, and was achieved in comparatively recent times. We shall, however, not retrace this historical path. Instead, we follow another line, and will introduce the integers as the result of *an attempt to solve an impossible problem*. Ludicrous as it may seem, there are few more fruitful endeavors throughout the development leading to present day mathematics than this: Attempting to solve an impossible problem.

Let $\mathcal{M} = \mathbb{N} \times \mathbb{N}$, and write

$$(a, b) \sim (c, d) \text{ whenever } a + d = b + c.$$

It is a fairly easy exercise to verify that this is an equivalence relation on the set $\mathcal{M}$. Now define

$$\mathbb{Z} = \mathcal{M}/\sim .$$

We introduce operations called *addition* and *multiplication* in the set $\mathcal{M}$ by putting

$$(a, b) + (c, d) = (a + c, b + d) \text{ and } (a, b) \cdot (c, d) = (ac + bd, ad + bc).$$

These, apparently strange, definitions are motivated by the idea that the pair (a, b) of natural numbers should correspond to the *integer* (not yet defined) $a - b$. We now see the motivation for the definition of the equivalence relation $\sim$ given above: $(a, b) \sim (c, d)$ if $a - b = c - d$, or phrased in terms of natural numbers: $a + d = b + c$. The relation $\sim$ defined above has a very important property which goes beyond that of merely being an equivalence relation: It is a *congruence relation* for the two operations addition and multiplication defined above: That is to say, the following two implications hold:

$$(a, b) \sim (c, d) \Longrightarrow (a, b) + (e, f) \sim (c, d) + (e, f) \text{ for all } (e, f) \in \mathcal{M},$$

$$(a, b) \sim (c, d) \Longrightarrow (a, b) \cdot (e, f) \sim (c, d) \cdot (e, f) \text{ for all } (e, f) \in \mathcal{M}.$$

These two properties are easily verified, as is the important consequence that

$$[(a, b) + (c, d)] = [(a', b') + (c', d')] \text{ and } [(a, b) \cdot (c, d)] = [(a', b') \cdot (c', d')]$$

whenever

$$(a, b) \sim (a', b') \text{ and } (c, d) \sim (c', d').$$

Thus addition and multiplication as defined in $\mathcal{M}$ induce addition and multiplication in $\mathbb{Z}$. We finally note that $[(a, b)] = [a + n, b + n]$ for all natural numbers n, and in particular that adding $[n, n]$ to any $[(a, b)]$ produces no change: Thus we denote this element (the same for all choices of n) by 0. Moreover, the element $[(a + n, n)]$ is independent of n, thus this element, this *integer*, is identified with the *natural number* a. We have obtained an injective mapping,[2] or as we say an *embedding*

[2] A mapping $\varphi : A \longrightarrow B$ is said to be injective if $a \neq b \Rightarrow \varphi(a) \neq \varphi(b)$.

$$\varphi : \mathbb{N} \hookrightarrow \mathbb{Z},$$

which satisfies the important conditions

$$\varphi(a+b) = \varphi(a) + \varphi(b) \text{ and } \varphi(ab) = \varphi(a) \cdot \varphi(b).$$

In language from abstract algebra we say that φ is an injective *homomorphism of the semi-ring* $\mathbb{N}$ *into the ring* $\mathbb{Z}$. We shall not pursue these algebraic concepts much further here, except to note that if a, b are natural numbers, then the equation

$$a + X = b,$$

which does not always have a solution in $\mathbb{N}$, does indeed always have a unique solution in $\mathbb{Z}$, namely the integer $[(b, a)]$. Identifying, as we always do, a with $\varphi(a)$, we then have that $[(b, a)] = b - a$.

We next pass to *the Rational Numbers.* This time we put

$$\mathcal{M} = \{(a, b) | \; a, b \in \mathbb{Z}, b \neq 0\}$$

as we now have the integers $\mathbb{Z}$ at our disposal. Here we define

$$(a, b) \sim (c, d) \text{ whenever } ad = bc$$

which again is easily seen to be an equivalence relation, enabling us to define the quotient-set as above:

$$\mathbb{Q} = \mathcal{M} / \sim$$

Again, analogously to the above, we define the operations addition and multiplication in $\mathcal{M}$ by

$$(a, b) + (c, d) = (ad + bc, bd) \text{ and } (a, b) \cdot (c, d) = (ac, bd),$$

and again, $\sim$ is a congruence relation for the two operations addition and multiplication, i.e. the two implications

$$(a, b) \sim (c, d) \Longrightarrow (a, b) + (e, f) \sim (c, d) + (e, f) \text{ for all } (e, f) \in \mathcal{M},$$
$$(a, b) \sim (c, d) \Longrightarrow (a, b) \cdot (e, f) \sim (c, d) \cdot (e, f) \text{ for all } (e, f) \in \mathcal{M}$$

hold. Thus the addition and multiplication defined in $\mathcal{M}$ induce addition and multiplication in $\mathbb{Q}$. We define an injective mapping

$$\varphi : \mathbb{Z} \hookrightarrow \mathbb{Q}$$

by putting $\varphi(t) = [(t, 1)]$, and verify right away that $\varphi(s + t) = \varphi(s) + \varphi(t)$ and $\varphi(st) = \varphi(s) \cdot \varphi(t)$. By means of φ we may identify $\mathbb{Z}$ with at subset of $\mathbb{Q}$, in such a way that addition and multiplication in $\mathbb{Q}$ reduces to the one we already know for

$\mathbb{Z}$ on this subset. Finally, the equation

$$aX = b,$$

which is not always solvable in $\mathbb{Z}$ when b and $a \neq 0$ are integers, is now solvable in $\mathbb{Q}$, for $a \neq 0, b \in \mathbb{Q}$.

To treat the *real numbers* we let $\mathcal{F}$ be the set of all sequences of rational numbers,

$$\mathcal{F} = \{\{x_n\}, n = 1, 2, \ldots, x_n \in \mathbb{Q}\}$$

We consider the subset $\mathcal{M} \subset \mathcal{F}$ of all the so-called *Cauchy sequences*: $\mathcal{M}$ is the set of all sequences $x = \{x_n, n = 1, 2, \ldots\}$ which satisfy the condition

$$\forall \epsilon > 0 \exists n_0 \in \mathbb{N} \text{ such that } m, n \geq n_0 \Longrightarrow |x_m - x_n| < \epsilon$$

or in plain language, more readable but less precise: The absolute difference between any two members of the sequence can be made arbitrarily small, once the indices are chosen sufficiently big. An example of such a sequence would be

$$x_1 = 3, x_2 = 3.1, x_3 = 3.14, x_4 = 3.141, x_5 = 3.1415, \ldots,$$

x_n being π to the first n digits. The informal idea is that for any Cauchy sequence x all the x_n are in $\mathbb{Q}$, they are rational numbers, while the sequence *converges to* a real number. Moreover, all real numbers are obtainable in this manner. One of several approaches to *defining* the real numbers is to use this idea. We introduce an equivalence relation in the set $\mathcal{M}$ by

$$x \sim y \Leftrightarrow (\forall \epsilon > 0 \exists n_0 \in \mathbb{N} \text{ such that } n \geq n_0 \Longrightarrow |x_n - y_n| < \epsilon).$$

Informally, the absolute difference $|x_n - y_n|$ is arbitrarily small for all sufficiently big values of n. It is easy to see that this is indeed an equivalence relation.

We now proceed by defining addition and multiplication in the set $\mathcal{M}$ in the obvious manner, namely

$$x + y = z \text{ where } z_n = x_n + y_n \text{ and } x \cdot y = z \text{ where } z_n = x_n y_n.$$

It actually requires a proof that the operations so defined yield new Cauchy sequences. It also must be proven that $\sim$ is a congruence relation for these operations. We omit these verifications, not too difficult but of some mild complexity. In the end we define $\mathbb{R} = \mathcal{M}/\sim$. $\mathbb{R}$ is an extension of $\mathbb{Q}$ by the injective mapping

$$\varphi : \mathbb{Q} \hookrightarrow \mathbb{R} \text{ given by } \varphi(r) = [x] \text{ where } x_n = r \text{ for all n.}$$

The extension of $\mathbb{Q}$ to $\mathbb{R}$ makes it possible to solve a general problem, insoluble in general within $\mathbb{Q}$: In fact, given a Cauchy sequence $x = \{x_n, n = 1, 2, \dots\}$. Then the limit $\lim_{n\to\infty} x_n$ always exists. We shall not pursue this theme here, however.

We finally come to *the complex numbers.* We define the set of complex numbers by $\mathbb{C} = \mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, and define addition and multiplication in $\mathbb{C}$ by

$$(a, b) + (c, d) = (a + c, b + d), (a, b) \cdot (c, d) = (ac - bd, ad + bc).$$

It is a straightforward while uninspiring exercise to check that these operations satisfy the usual properties of addition and multiplication. Moreover, $\mathbb{C}$ becomes an extension of $\mathbb{R}$ by the injective mapping

$$\varphi : \mathbb{R} \hookrightarrow \mathbb{C} \text{ given by } \varphi(a) = (a, 0).$$

We identify $\mathbb{R}$ with this subset of $\mathbb{C}$. The whole point of this construction is contained in the following simple computation:

$$(0, 1) \cdot (0, 1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) = (-1, 0).$$

Writing $(0, 1) = i$, we have

$$i^2 = -1,$$

having identified $\mathbb{R}$ with a subset of $\mathbb{C}$ as described above. We also note that

$$(a, b) = a + ib,$$

where $i = \sqrt{-1}$, which is the usual form of a complex number. The extension of $\mathbb{R}$ to $\mathbb{C}$ makes it possible to solve the following problem, insoluble in $\mathbb{R}$: The equation $X^2 + 1 = 0$ has a solution, namely i.

In fact, the set of complex numbers represents a much stronger enlargement of $\mathbb{R}$. We have the following beautifully theorem, known as *the Fundamental Theorem of Algebra:*

Theorem 17. *A polynomial*

$$X^n + a_1 X^{n-1} + \dots + a_{n-1} X + a_n = P(X)$$

where the coefficients $a_1, \dots, a_n$ *are complex numbers, may be factored into a product of (possibly repeated) factors of the type* $X - x$, *where* $x \in \mathbb{C}$. *In particular, all equations of the type*

$$X^n + a_1 X^{n-1} + \dots + a_{n-1} X + a_n = 0$$

where the coefficients $a_1, \dots, a_n$ *are complex numbers, have solutions in* $\mathbb{C}$.

11.4 Complex Numbers and Trigonometry

There is a beautiful geometric interpretation of the complex numbers.[3] *Caspar Wessel*, 1745–1818, was a Norwegian surveyor working in Denmark. In 1782 he was assigned to the task of conducting a trigonometrical survey of the duchy of Oldenburg. Oldenburg had been Danish since 1667 but in 1773 was exchanged by the Danish king for Holstein–Gottorp. Wessel worked on the survey till 1785, developing new and sophisticated mathematical methods. He explained this in a report he wrote in 1787, and this report contains Wessel's geometric interpretation of complex numbers.

In 1796 Wessel had completed the triangulation of Denmark, and wrote his only mathematical paper. He was allowed to present it to a meeting of the Royal Danish Academy in 1797. In fact, the year before the Academy had started to allow non-members to present papers. Wessel's paper was the first such paper to be accepted. It was published, in Danish, in 1799.

Today this geometric interpretation is called the Argand diagram, but Wessel's work came first and was rediscovered by Argand in 1806. It was again rediscovered in 1831, by Gauss.

Wessel's paper was not noticed by the mathematical community until 1895. Then the Norwegian mathematician *Sophus Lie* republished it, and a French translation was prepared by the Danish mathematician *Hieronymous Georg Zeuthen*, 1839–1920. It was published in 1897, but an English translation was not published until 1999, 200 years after it had first appeared.

After this historical introduction we turn to the mathematics.

The set $\mathbb{C}$ with the operations we have introduced in the previous section is referred to as *the field of complex numbers.*

For $z = x + iy$ with x and y real we write $\bar{z} = x - iy$. We refer to $\bar{z}$ as the complex conjugate of z. As is easily seen,

$$z\bar{z} = x^2 + y^2.$$

If $z \neq 0 = (0, 0)$, then of course $x^2 + y^2 \neq 0$. Thus in this case we get

$$z\frac{\bar{z}}{x^2 + y^2} = 1,$$

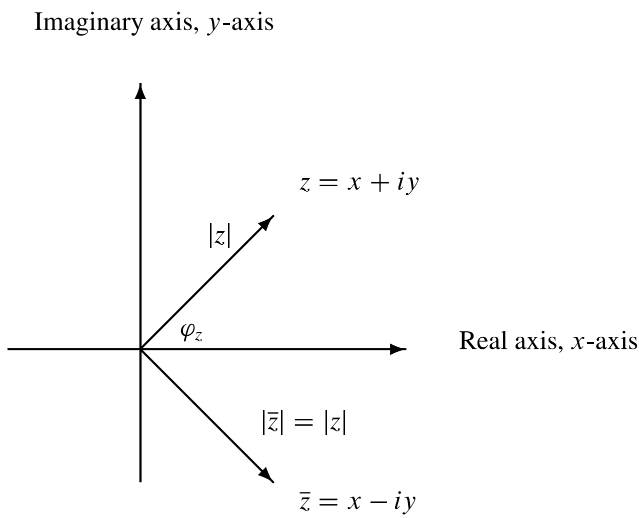
and thus we have shown that all complex numbers different from zero have a multiplicative inverse.

We next define the symbols $\Re(z)$ and $\Im(z)$, $\mathrm{Arg}(z)$ and the absolute value $|z|$: We put

$$|z| = \sqrt{x^2 + y^2},$$

[3] The source for the following historical account is mainly [34] and above all [60].

which we illustrate as follows:



With notation as in the illustration we put

$$\mathrm{Arg}(z) = \varphi_z, r = |z|, x = \Re(z), y = \Im(z).$$

We may now write complex numbers on trigonometric form. With notations as above we get

$$z = r(\cos(\varphi) + i\,\sin(\varphi)),$$

and we are ready to state a result of great importance for computing with complex numbers. This is the key to Caspar Wessels breakthrough in surveying:

Proposition 5. *Let*

$$\begin{aligned} z_1 &= r_1(\cos(\varphi_1) + i\,\sin(\varphi_1)),\\ z_2 &= r(\cos(\varphi_2) + i\,\sin(\varphi_2)). \end{aligned}$$

Then

$$z_1 z_2 = r_1 r_2(\cos(\varphi_1 + \varphi_2) + i\,\sin(\varphi_1 + \varphi_2)).$$

Proof. By the definition of multiplication of complex numbers (Fig. 11.2), we find that $z_1 z_2$ is equal to

$$r_1 r_2(\cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\sin(\varphi_2) + i(\cos(\varphi_1)\sin(\varphi_2) + \sin(\varphi_1)\cos(\varphi_2))).$$

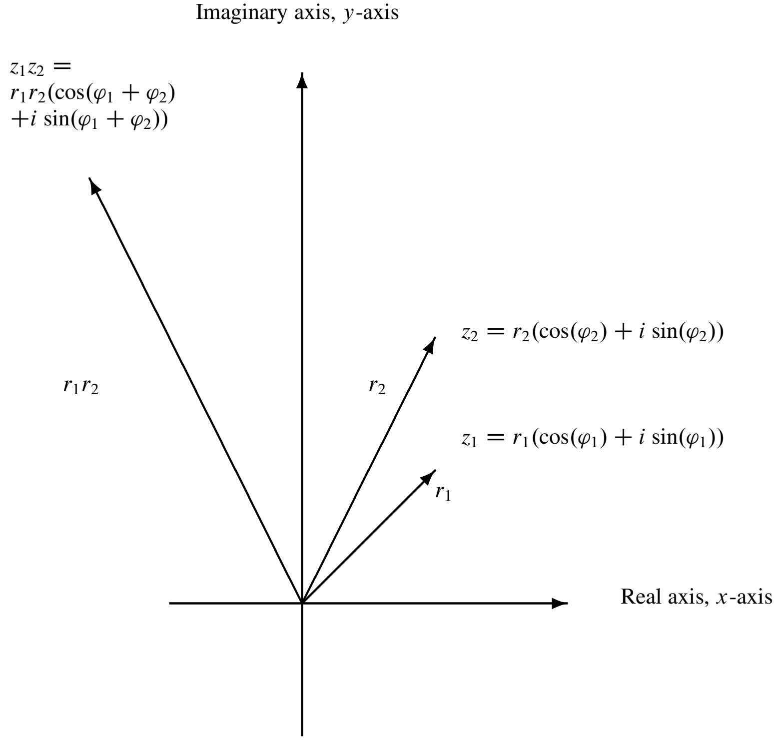


Fig. 11.2 Multiplying complex numbers

The claim now follows by the formulas for sine and cosine of a sum of two angles.
□

Repeated use of this formula yields

Corollary 2 (De Moivre's Formula).

$$(\cos(\varphi) + i\,\sin(\varphi))^n = \cos(n\varphi) + i\,\sin(n\varphi)$$

We finally arrive at a remarkable formula, which was discovered by *Leonard Euler* (1705–1783). The basis for Euler's Formula is the fundamental observation that we may extend all real functions which may be developed as a power series in the argument, to a similar function of a complex variable. Indeed, if

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n + \cdots$$

where x is a real variable, and the series converges within a certain interval about 0, then we may extend the real function $f(x)$ to a complex function $f(z)$ by simply writing

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n + \cdots$$

where now z is a *complex* variable. This complex power series will converge for all z in a disc about the origin. In particular the following series converge for all values of z, extending the corresponding real functions to complex arguments:

$$\begin{aligned}
e^z &= 1 + z + \frac{1}{2}z^2 + \frac{1}{3!}z^3 + \frac{1}{4!}z^4 + \cdots + \frac{1}{n!}z^n + \cdots \\
\sin(z) &= z - \frac{1}{3!}z^3 + \cdots + (-1)^m \frac{1}{(2m+1)!}z^{2m+1} + \cdots \\
\cos(z) &= 1 - \frac{1}{2}z^2 + \frac{1}{4!}z^4 - \cdots + \frac{1}{(2m)!}z^{2m} + \cdots
\end{aligned}$$

In principle we may prove the formula

$$e^{(x_1+x_2)} = e^{x_1} e^{x_2}$$

for real values of x_1 and x_2 by multiplying together, rearranging and simplify the two power series for e^{x_1} and e^{x_2}. We know that this will work, since the formula is known to us for real x_1 and x_2 by other arguments. But then the very same multiplication of power series will also work for complex arguments. Indeed, the rules for multiplying power series extend *mutatis mutandi* from the real case to the complex case.[4]

In particular we get for $z = x + iy$ that

$$e^z = e^{x+iy} = e^x e^{iy}.$$

It is quite remarkable that when we substitute $z = iy$ in the power series for e^z, and rearrange the terms, then we get the power series for $\cos(y)$ and $\sin(y)$ entering into the expression as follows[5]:

$$\begin{aligned}
e^{iy} &= 1 - \frac{1}{2}y^2 + \frac{1}{4!}y^4 - \cdots + \frac{1}{(2m)!}y^{2m} + \cdots \\
&\quad + i(y - \frac{1}{3!}y^3 + \cdots + (-1)^m \frac{1}{(2m+1)!}y^{2m+1} + \cdots) \\
&= \cos(y) + i\sin(y).
\end{aligned}$$

Letting $y = \pi$ we get *Euler's Formula:*

$$e^{i\pi} + 1 = 0.$$

[4] This needs to be proved, of course, but that falls outside our scope.

[5] Again the rules for manipulating power series fall outside our scope here.

Exercises

Exercise 11.1 Check that in Sect. 11.4, Proposition 5 is only used in proving Corollary 2. In particular the formula $e^{x+iy} = e^x e^{iy}$ can be used to prove the formulas for sine and cosine of a sum of two angles, and to prove Corollary 2. Work out the details in this argument.

Exercise 11.2 Use Corollary 2 of Sect. 11.4 to prove the following formulas

$$\cos(2\varphi) = \cos^2(\varphi) - \sin^2(\varphi), \sin(2\varphi) = 2\sin(\varphi)\cos(f)$$

$$\cos(3\varphi) = 4\cos^3(\varphi) - 3\cos(\varphi), \sin(3\varphi) = 4\sin(\varphi)\cos^2(\varphi) - \sin(\varphi)$$

Also prove the recursion formulas

$$\cos((n+1)\varphi) = \cos(\varphi)\cos(n\varphi) - \sin(\varphi)\sin(n\varphi)$$
$$\sin((n+1)\varphi) = \cos(\varphi)\sin(n\varphi) + \sin(\varphi)\cos(n\varphi)$$

Then show that $\sin(3\varphi) = 4\sin(\varphi)\sin(\frac{\pi}{3} + \varphi)\sin(\frac{\pi}{3} - \varphi)$.

Chapter 12
Projective Space

Projective space is not merely a dry and theoretical invention. On the contrary, it is a living reality. It would be impossible to create perspective and depth in a painting without an understanding of projective space. When we are painting or making pictures of physical objects in space, like buildings, roads or other objects, we need to project the 3-dimensional space with all it contains onto a 2-dimensional canvas, piece of paper, or photographic film.

Then two parallel lines in nature will be represented in the picture as two lines *intersecting* at a point located at *the line of perspectivity*. We might say that the Euclidian 3-space is projected onto the *projective plane*.

An example is shown in Fig. 12.1, a photo from a side street at Berkeley in the mid 1980s. In this otherwise totally uninteresting picture where nothing happens, there is a conspicuous collection of *parallel lines*: The power lines, the features on the facades of the buildings on the right hand side of the street, then the lines giving the outlines of the cars parked along the sidewalks, the street itself, the sidewalks. All these lines are projected, by the process of taking the photo, onto the photographic film. And the projections intersect at a point around the middle of the left hand edge of the picture. That is the projection of a point at infinity of the 3-dimensional space. So even if we cannot see this "point at infinity" of the usual 3-space, we capture its projection onto the film!

12.1 Coordinates in the Projective Plane

We shall now explain how it is possible to introduce *coordinates* for the projective plane, in a similar manner to what we have for the Euclidian plane $\mathbb{R}^2$.

Recall that the points in $\mathbb{P}^2(\mathbb{R})$ are the *lines* through (0,0,0) in $\mathbb{R}^3$. Such a line α is uniquely determined by a *vector* $(a, b, c) \neq (0, 0, 0)$. This vector gives the direction of the line, and as the line passes through the origin, it is given as follows:

$$\alpha = \{(x, y, z) \mid x = at, y = bt, z = ct, \quad \text{where } t \in \mathbb{R}\}.$$

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_12,

Fig. 12.1 A side street at Berkeley, in the mid 1980s. We notice the parallel lines in space projected onto a bunch of lines intersecting in a point near the middle of the left edge of the picture. Photo by the author

We refer to this as *the line α on parametric form.* Thus to every value of the parameter t there corresponds a uniquely determined point $P(t) \in \alpha$, and conversely, to every $P \in \alpha$ there corresponds a uniquely determined value of $t, t = t_P$, such that

$$P = P(t_P).$$

We may also describe some other space curves in the same manner. Then the curve C is given by

$$C = \{(x, y, z) \mid x = f(t), y = g(t), z = g(t), \text{ where } t \in \mathbb{R}\}$$

which defines the curve C in $\mathbb{R}^3$. For example, a line which *does not* pass through the origin, will have the following parametric form:

$$\alpha = \{(x, y, z) \mid x = x_0 + at, y = y_0 + bt, z = z_0 + ct, t \in \mathbb{R}\}$$

where (x_0, y_0, z_0) is a point on the line which we may chose as we like. Thus clearly one and same line may be given on different parametric forms. Choosing another point than (x_0, y_0, z_0) gives another parametric form, and the vector defining the direction of the line may be replaced by any non-zero multiple.

We return to the lines through the origin. As mentioned above, if we change the vector (a, b, c) to some (a', b', c') which is proportional to the original one, then the new parametric form yields the same line. Thus it is only *the ratio* between the numbers a, b and c which is important. We let $(a : b : c)$ denote *the set of all such vectors* proportional to (a, b, c):

$$(a : b : c) = \{(a', b', c') \big| \, a' = ra, b' = rb, c' = rc, \text{ where } r \neq 0\}$$

Proceeding formally, we define a relation in the set

$$\mathcal{M} = \mathbb{R}^3 - (0, 0, 0) :$$

by putting
$(a, b, c) \sim (a', b', c')$ whenever there exists $r \neq 0$ such that $a'=ra, b' = rb$, $c' = rc$.
Thus for instance

$$(1 : 2 : 3) = (2 : 4 : 6) = (\pi : 2\pi : 3\pi).$$

This relation is, as is immediately checked, an equivalence relation, and $(a : b : c)$ is the equivalence class defined by the element (a, b, c). These equivalence classes are in bijective correspondence with the lines through the origin, i.e., with the set of points of $\mathbb{P}^2(\mathbb{R})$.

Further, it is clear that if $a \neq 0$, then we may assume that $a = 1$: We just choose $r = \frac{1}{a}$ so that $(a, b, c) = r(1, \frac{b}{a}, \frac{c}{a})$.

The same applies to the other coordinates as well, so if a, b and c are $\neq 0$, then

$$(a : b : c) = \left(1 : \frac{b}{a} : \frac{c}{a}\right) = \left(\frac{a}{b} : 1 : \frac{c}{b}\right) = \left(\frac{a}{c} : \frac{b}{c} : 1\right).$$

We now define the *projective coordinates* of a point $\alpha \in \mathbb{P}^2(\mathbb{R})$:

Definition 8. If the point $\alpha \in \mathbb{P}^2(\mathbb{R})$ is given as the line in $\mathbb{R}^3$ defined on parametric form as

$$\alpha = \{(x, y, z) \, | x = at, y = bt, z = ct, t \in \mathbb{R}\}$$

then we denote this point by $(a : b : c)$. The ratio $a : b : c$ is referred to as the projective coordinates of the point α, this name is also applied to the tuple (a, b, c), which is only determined up to a constant multiple.

The relation between the projective and the usual (*"affine"*) coordinates of a point in $\mathbb{R}^2$ is given by the following

Proposition 6. *Let as before $\mathbb{R}^2$ be identified with the points in $\mathbb{P}^2(\mathbb{R})$ which correspond to lines which are not contained in the xy-plane. Under the identification performed in Proposition 2 we get that*

$$(x : y : 1) = (x, y).$$

Proof. We may choose $(a, b, c) = (x, y, 1)$ in the parametric form of the line which corresponds to the point. $(x, y, 1)$ is then the point of intersection between this line and the plane given by $z = 1$. The claim now follows from Proposition 2. □

12.2 Projective n-Space

In this section we shall define the *n-dimensional* projective space $\mathbb{P}^n(\mathbb{R})$. We then consider the set

$$\mathcal{M} = \mathbb{R}^{n+1} - \{(0, \dots, 0)\}$$

and as for the case $n = 2$ define a relation by

$$(a_1, a_2, \dots, a_{n+1}) \sim (b_1, b_2, \dots, b_{n+1})$$

whenever there exist $r \in \mathbb{R}$ such that

$$a_i = rb_i \quad \text{for all } i = 1, 2, \dots, n+1.$$

We easily verify that this is an equivalence relation.

Definition 9. The set $\mathcal{M}/\sim$ is denoted by $\mathbb{P}^n(\mathbb{R})$ and referred to as the projective n-space over the reals $\mathbb{R}$.

As in the case $n = 2$ we use the following notation for the equivalence classes:

$$[(a_1, a_2, \dots, a_{n+1})] = (a_1 : a_2 : \dots : a_{n+1}).$$

Whenever $a_{n+1} \neq 0$, we may assume that $a_{n+1} = 1$, and with this assumption the other coordinates are uniquely determined. Thus we may identify the sets

$$\{(a_1 : a_2 : \dots : a_{n+1}) |\; a_{n+1} \neq 0\}$$

and

$$\mathbb{R}^n.$$

The set

$$\{(a_1 : a_2 : \dots : a_{n+1}) |\; a_{n+1} = 0\}$$

is referred to as *the points at infinity* in $\mathbb{P}^n(\mathbb{R})$. This subset may in turn be identified with $\mathbb{P}^{n-1}(\mathbb{R})$ in the obvious manner by ignoring the last coordinate, which is zero. We obtain the following description of $\mathbb{P}^n(\mathbb{R})$:

$$\mathbb{P}^n(\mathbb{R}) = \mathbb{R}^n \cup \mathbb{P}^{n-1}(\mathbb{R})$$

and we say that $\mathbb{P}^n(\mathbb{R})$ *is obtained by adjoining to* $\mathbb{R}^n$ *a space* $\mathbb{P}^{n-1}(\mathbb{R})$ *of points at infinity.*

By dividing up $\mathbb{P}^{n-1}(\mathbb{R})$ similarly, and repeating the process all the way down to $\mathbb{P}^0(\mathbb{R})$, we get

$$\mathbb{P}^n(\mathbb{R}) = \mathbb{R}^n \cup \mathbb{R}^{n-1} \cup \cdots \cup \mathbb{R} \cup \mathbb{P}^0(\mathbb{R}).$$

But $\mathbb{P}^0(\mathbb{R})$ consist of only one point: In fact, if a and b are real non-zero numbers, then

$$(a) \sim (b)$$

since

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

Thus $\mathbb{P}^0(\mathbb{R}) = \{pt\}$. This yields

$$\mathbb{P}^n(\mathbb{R}) = \mathbb{R}^n \cup \mathbb{R}^{n-1} \cup \cdots \cup \mathbb{R} \cup \{pt\}$$

where union is disjoint, and $\{pt\}$ denotes a set which only consist of the single point pt. This union is some times referred to as the cell decomposition of $\mathbb{P}^n(\mathbb{R})$.

12.3 Affine and Projective Coordinate Systems

In the previous paragraph we saw that a point in $\mathbb{P}^n(\mathbb{R})$ is given by $n+1$ projective coordinates: $P = (a_1 : \ldots : a_n : a_{n+1})$. Before we proceed, we shall switch to another notation for the coordinates: Instead of writing $P = (a_1 : a_2 : \ldots : a_{n+1})$ we put $P = (a_0 : a_1 : \ldots : a_n)$, where we have relabelled the last coordinate and put it up front: $a_0 = a_{n+1}$. This is customary in the most recent literature.

In this notation $\mathbb{R}^n$ is the subset of $\mathbb{P}^n(\mathbb{R})$ consisting of the points $(a_1, \ldots, a_n) = (1 : a_1 : \ldots : a_n)$.

For the remainder of the paragraph we only treat the case $n = 2$, thus the affine and the projective *plane*. However, everything we show may easily be proven in the general case as well.

Let $L = b_0X_0 + b_1X_1 + b_2X_2$. An expression of this type is referred to as *a linear form* in X_0, X_1 and X_2. We then put

$$V_+(L) = \{(a_0 : a_1 : a_2) \in \mathbb{P}^2(\mathbb{R}) \mid b_0a_0 + b_1a_1 + b_2a_2 = 0\}$$

and

$$D_+(L) = \{(a_0 : a_1 : a_2) \in \mathbb{P}^2(\mathbb{R}) \mid b_0a_0 + b_1a_1 + b_2a_2 \neq 0\}$$

Some books omit the $+$ in this terminology. However, we will keep it and reserve V and D without the $+$ for the sets

$$V(L) = \{(a_0, a_1, a_2) \in \mathbb{R}^3 \mid b_0a_0 + b_1a_1 + b_2a_2 = 0\}$$

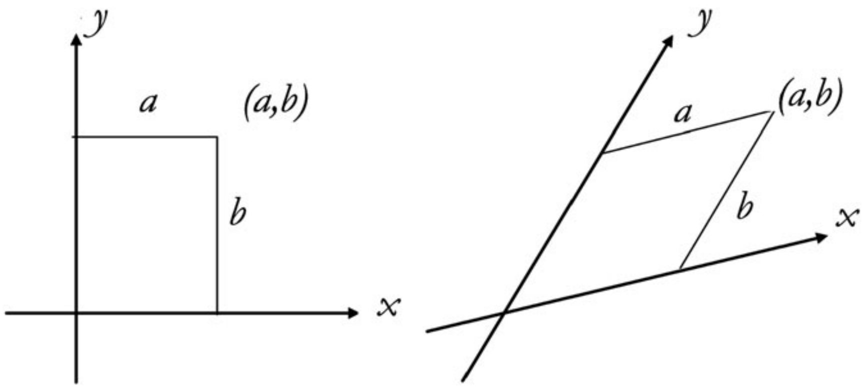


Fig. 12.2 A Cartesian coordinate system to the *left*, and a more general affine one to the *right*

and

$$D(L) = \{(a_0, a_1, a_2) \in \mathbb{R}^3 \big| \; b_0a_0 + b_1a_1 + b_2a_2 \neq 0\} .$$

By switching to a new projective coordinate system, $D_+(L)$ may be identified in a natural way with $\mathbb{R}^2$, while $V_+(L)$ is identified with the projective line $\mathbb{P}^1(\mathbb{R}) = \mathbb{R} \cup \{pt\}$. We shall now explain this in some detail.

An *affine* coordinate system in $\mathbb{R}^2$ is a skew coordinate system. That is to say, the axes are not necessarily orthogonal to one another, and the scales may be different on the x-axis and the y-axis. The difference between a Cartesian coordinate system and an affine one is shown in Fig. 12.2. The affine coordinate system do not necessarily have the x- and y-axes oriented counter clockwise, as is the case with the Cartesian system.

The relation between an old affine coordinate system with coordinates denoted by (x, y), and a new one $(\overline{x}, \overline{y})$ is given by

$$\overline{x} = \alpha + ax + by$$
$$\overline{y} = \beta + cx + dy$$

where a, b, c, d, α and β are real numbers, such that

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \neq 0$$

The $\overline{x}$-axis is given, in the old coordinate system, by the equation

$$\beta + cx + dy = 0$$

while the $\overline{y}$-axis is given by

$$\alpha + ax + by = 0.$$

The coordinates of the old origin in the new system is (α, β).

The condition $ad - bc = 1$ will ensure that if the old system is a Cartesian one, then so is the new.

It is clear that a curve in $\mathbb{R}^2$ given by an equation of degree m in the coordinates x, y will be given by an equation of the same degree m in the coordinates $\overline{x}, \overline{y}$. Indeed, we get

$$x = \frac{d(\overline{x} - \alpha) - b(\overline{y} - \beta)}{ad - bc}, \quad y = \frac{-c(\overline{x} - \alpha) + a(\overline{y} - \beta)}{ad - bc}$$

Rather than to view this as moving from one coordinate system to another, we may regard it as describing a mapping from the plane to itself, a so called affine transformation:

$$\mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

$$(x, y) \mapsto (\overline{x}, \overline{y})$$

where

$$\overline{x} = \alpha + ax + by$$

$$\overline{y} = \beta + cx + dy$$

and a, b, c, d, α and β are real numbers such that

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \neq 0.$$

Some of the geometric properties of curves or other figures are preserved by the affine transformations: That is to say, such properties as incidence, collinearity, passing through a fixed point or being a straight line, are preserved. Among properties *not* preserved we mention *the property of being a circle*, while being an ellipse *is preserved*. The property for two lines to form a right angle is not preserved, while the property for a line to *bisect* the angle formed by two given lines in equal parts, is preserved. The property for a point to divide a line segment in equal parts is an affine property, while the property of dividing the segment in a part of preassigned length is not. Finally, the property for a line to be *tangent* to a given curve is preserved by an affine transformation.

Definition 10. The properties preserved by the affine transformations are referred to as *affine* properties.

The identity transformation is affine, as it is given by $\alpha = \beta = 0$ and the identity matric where $a = d = 1$, $b = c = 0$. The composition of two affine transformations is again an affine transformation, being given by adding the vectors (α, β) and taking the matrix-product of the two matrices, and finally the inverse of an affine transformation is therefore given by $-(\alpha, \beta)$ and the inverse matrix. Thus it is also affine. We express this by saying that *the affine transformations form a transformation-group*.

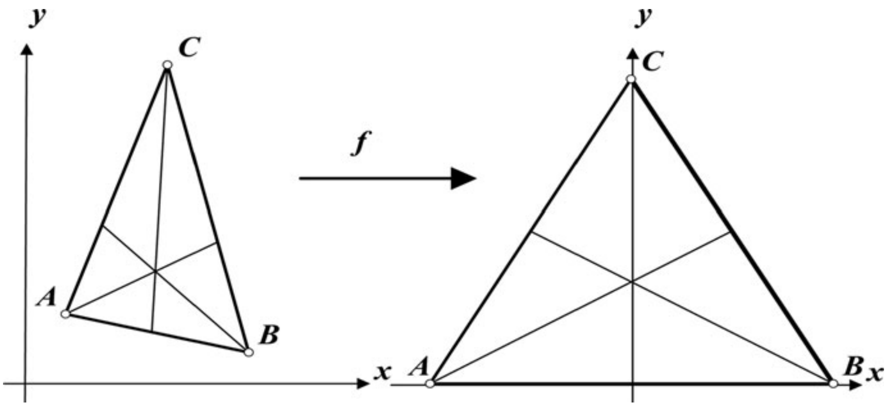


Fig. 12.3 Simplifying the geometry by a change of coordinate system

One may well take another, but equivalent, point of view and say that the affine properties are the ones which are preserved under affine changes of coordinate systems.

This represents a great advantage when we wish to find simple proofs for theorems in affine geometry. For example, suppose we wish to prove the following result:

The medians of a triangle intersect in one point. *Let there be given a triangle ABC. Draw the medians, i.e., the line joining A to the mid point of BC, the line joining B to the mid point of AC and the line joining C to the mid point of AB. Then these three lines meet in a single point.*

A proof is provided by considering the *center of gravity* of the triangle: Since the triangle will balance on a "knifes edge" along the medians, the center of gravity lies on the three median lines, hence they intersect in that point.

Simple as this proof is, it does presuppose quite a bit of calculus. A purely algebraic proof runs as follows: We may choose an affine coordinate system such that $A = (-1, 0)$, $B = (1, 0)$ and $C = (0, \sqrt{3})$. This is always possible by considerations which will be explained below. Since the property of the triangle we wish to prove may be checked by algebra only, and is independent of the coordinate system, we have reduced the question to proving the property for a triangle where the vertices have these coordinates in a *Cartesian coordinate system*. But that is an *equilateral* triangle with sides equal to 2. For an equilateral triangle the claim is obvious by symmetry. The situation is illustrated in Fig. 12.3.

For the projective plane we may proceed in the same way. We introduce a new coordinate system by

$$\overline{x}_0 = \alpha_{0,0}x_0 + \alpha_{0,1}x_1 + \alpha_{0,2}x_2$$

$$\overline{x}_1 = \alpha_{1,0}x_0 + \alpha_{1,1}x_1 + \alpha_{1,2}x_2$$

$$\overline{x}_2 = \alpha_{2,0}x_0 + \alpha_{2,1}x_1 + \alpha_{2,2}x_2$$

where we assume that the determinant is $\neq 0$:

$$A = \begin{vmatrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{vmatrix} \neq 0.$$

As in the affine case, we may view this as describing a transformation

$$\mathbb{P}^2(\mathbb{R}) \longrightarrow \mathbb{P}^2(\mathbb{R})$$

$$(a_0 : a_1 : a_2) \mapsto (\overline{x}_0 : \overline{x}_1 : \overline{x}_2)$$

given as above. This is referred to as a *projective transformation*, and as in the affine case these transformations form a *transformation-group*: The identity mapping is a projective transformation, the composition of two projective transformations is again a projective transformation, and the inverse of a projective transformation is again a projective transformation.

Definition 11. A projective property is a property which is preserved by all projective transformations.

We may also express this by saying that the projective properties are independent of choice of projective coordinate system.

However the points at infinity are not preserved. Being at infinity is not a projective property: This is definitely dependent on the coordinate system. In the original coordinate system the points at infinity are the points in the subset $V_+(x_0)$, as we have identified the affine plane $\mathbb{R}^2$ with $D_+(x_0)$. In the new coordinate system the points at infinity will be the points in the subset $V_+(\overline{x}_0) = V_+(\alpha_{0,0}x_0 + \alpha_{0,1}x_1 + \alpha_{0,2}x_1)$, and the subset $D_+(\overline{x}_0) = D_+(\alpha_{0,0}x_0 + \alpha_{0,1}x_1 + \alpha_{0,2}x_2)$ is now identified with $\mathbb{R}^2$.

The simplifying new coordinate system which we introduced in affine space $\mathbb{R}^2$ to prove the property of the medians, comes from the following proposition:

Proposition 7. *Given four points P_1, P_2, P_3 and $P_4 \in \mathbb{P}^2(\mathbb{R})$, such that no three of them are collinear, i.e., such that the four points constitute an arc of four. Then there exists a projective coordinate system in $\mathbb{P}^2(\mathbb{R})$ such that*

$$P_1 = (1 : 0 : 0), P_2 = (0 : 1 : 0), P_3 = (0 : 0 : 1), P_4 = (1 : 1 : 1)$$

Before we proceed to the proof, we state a more general result, this time in terms of *projective transformations*:

Corollary 3. *Given four points P_1, P_2, P_3 and $P_4 \in \mathbb{P}^2(\mathbb{R})$ as in the proposition, as well as another set of four points P_1', P_2', P_3' and $P_4' \in \mathbb{P}^2(\mathbb{R})$ with the same property. Then there exists a projective transformation G of $\mathbb{P}^2(\mathbb{R})$ onto itself, mapping P_i to P_i' for i = 1,2,3, and 4.*

The proposition implies the corollary, since it implies the existence of a projective transformation F mapping P_1 to $(1:0:0)$, P_2 to $(0:1:0)$, etc., and a projective transformation F' mapping P'_1 to $(1:0:0)$, P'_2 to $(0:1:0)$, etc. Take G equal to the composition $(F')^{-1} \circ F$.

Proof (of the proposition). We give the proof below. But this proof may well be skipped, at least at a first reading. It is of a purely algebraic nature, and belongs to the field of linear algebra more than to projective geometry. But the result itself is very useful, and saves us from a lot of tedious reasoning. By going through this, we do away with these technical matters once and for all.

We may write the transition from one coordinate system to another as a matrix multiplication as follows:

$$\left\{\begin{matrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{matrix}\right\} \cdot \left\{\begin{matrix} x_0 \\ x_1 \\ x_2 \end{matrix}\right\} = \left\{\begin{matrix} \overline{x}_0 \\ \overline{x}_1 \\ \overline{x}_2 \end{matrix}\right\}.$$

Let $P_i = (a_{i,0} : a_{i,1} : a_{i,2})$ for $i = 1, 2, 3, 4$. The aim is to find a matrix $\{\alpha_{i,j}\}$, with determinant $\neq 0$, such that the following four conditions are satisfied for suitable choices of r, s, t, u, all $\neq 0$:

$$\left\{\begin{matrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{matrix}\right\} \cdot \left\{\begin{matrix} a_{1,0} \\ a_{1,1} \\ a_{1,2} \end{matrix}\right\} = \left\{\begin{matrix} r \\ 0 \\ 0 \end{matrix}\right\}$$

$$\left\{\begin{matrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{matrix}\right\} \cdot \left\{\begin{matrix} a_{2,0} \\ a_{2,1} \\ a_{2,2} \end{matrix}\right\} = \left\{\begin{matrix} 0 \\ s \\ 0 \end{matrix}\right\}$$

$$\left\{\begin{matrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{matrix}\right\} \cdot \left\{\begin{matrix} a_{3,0} \\ a_{3,1} \\ a_{3,2} \end{matrix}\right\} = \left\{\begin{matrix} 0 \\ 0 \\ t \end{matrix}\right\}$$

$$\left\{\begin{matrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{matrix}\right\} \cdot \left\{\begin{matrix} a_{4,0} \\ a_{4,1} \\ a_{4,2} \end{matrix}\right\} = \left\{\begin{matrix} u \\ u \\ u \end{matrix}\right\}.$$

Replacing the matrix and the parameters r, s, t and u by their respective inverses, we find that it suffices to find a matrix $\{\beta_{i,j}\}$ such that

$$\left\{\begin{matrix} \beta_{0,0} & \beta_{0,1} & \beta_{0,2} \\ \beta_{1,0} & \beta_{1,1} & \beta_{1,2} \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} \end{matrix}\right\} \cdot \left\{\begin{matrix} 1 \\ 0 \\ 0 \end{matrix}\right\} = \left\{\begin{matrix} ra_{1,0} \\ ra_{1,1} \\ ra_{1,2} \end{matrix}\right\}$$

$$\begin{Bmatrix} \beta_{0,0} & \beta_{0,1} & \beta_{0,2} \\ \beta_{1,0} & \beta_{1,1} & \beta_{1,2} \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} \end{Bmatrix} \cdot \begin{Bmatrix} 0 \\ 1 \\ 0 \end{Bmatrix} = \begin{Bmatrix} sa_{2,0} \\ sa_{2,1} \\ sa_{2,2} \end{Bmatrix}$$

$$\begin{Bmatrix} \beta_{0,0} & \beta_{0,1} & \beta_{0,2} \\ \beta_{1,0} & \beta_{1,1} & \beta_{1,2} \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} \end{Bmatrix} \cdot \begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix} = \begin{Bmatrix} ta_{3,0} \\ ta_{3,1} \\ ta_{3,2} \end{Bmatrix}$$

$$\begin{Bmatrix} \beta_{0,0} & \beta_{0,1} & \beta_{0,2} \\ \beta_{1,0} & \beta_{1,1} & \beta_{1,2} \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} \end{Bmatrix} \cdot \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix} = \begin{Bmatrix} ua_{4,0} \\ ua_{4,1} \\ ua_{4,2} \end{Bmatrix}.$$

Clearly the first three of these conditions will be satisfied, for all r, s and t, if we put

$$\begin{Bmatrix} \beta_{0,0} & \beta_{0,1} & \beta_{0,2} \\ \beta_{1,0} & \beta_{1,1} & \beta_{1,2} \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} \end{Bmatrix} = \begin{Bmatrix} ra_{1,0} & sa_{2,0} & ta_{3,0} \\ ra_{1,1} & sa_{2,1} & ta_{3,1} \\ ra_{1,2} & sa_{2,2} & ta_{3,2} \end{Bmatrix}.$$

Indeed, multiplication of this matrix by the column vector

$$\begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}$$

to the right yields the first column of the matrix. In the same way the two other column vectors with the 1 in the middle and at the bottom, respectively, yield the second and third column of the matrix. To get a matrix which satisfies the final fourth condition, it therefore will suffice to determine r, s and t such that

$$\begin{Bmatrix} ra_{1,0} & sa_{2,0} & ta_{3,0} \\ ra_{1,1} & sa_{2,1} & ta_{3,1} \\ ra_{1,2} & sa_{2,2} & ta_{3,2} \end{Bmatrix} \cdot \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix} = \begin{Bmatrix} a_{4,0} \\ a_{4,1} \\ a_{4,2} \end{Bmatrix}.$$

This amounts to an equation for r, s and t which may be written as

$$\begin{Bmatrix} a_{1,0} & a_{2,0} & a_{3,0} \\ a_{1,1} & a_{2,1} & a_{3,1} \\ a_{1,2} & a_{2,2} & a_{3,2} \end{Bmatrix} \cdot \begin{Bmatrix} r \\ s \\ t \end{Bmatrix} = \begin{Bmatrix} a_{4,0} \\ a_{4,1} \\ a_{4,2} \end{Bmatrix}.$$

So we have a system of equations with three equations and three unknowns, with determinant $\neq 0$, since the points P_1, P_2 and P_3 are not collinear. But in Sect. 6.7 we showed Cramer's Theorem which yields that such a system has a unique solution, (r, s, t). This completes the proof. □

Chapter 13
Geometry in the Affine and the Projective Plane

In this chapter we shall, among other things, prove the classical theorems of *Desargues, Pappus and Pascal*. These theorems are valid in the projective plane $\mathbb{P}^2(\mathbb{R})$, and we shall give simple algebraic proofs, which fully take advantage of the strength inherent in *Analytic Planar Geometry.*

13.1 The Theorem of Desargues

The most important result in Desargues' book, which we told about in Sect. 6.4, is a typical example of what we call an *incidence theorem*. It is frequently referred to as *Desargues' Perspective Theorem* and says the following:

Theorem 18 (Desargues). *Let two triangles ABC and $A'B'C'$ be given in $\mathbb{P}^2(\mathbb{R})$, such that $A \neq A'$, $B \neq B'$ and $C \neq C'$. Then if the lines through corresponding vertices pass through the same point, the intersections of (the prolongation of) corresponding sides will intersect in points lying on the same line.*

Proof. We introduce notation as in Fig. 13.1.

We now preform a very common trick, which immediately reduces the proof to a very simple and elementary fact which is quite well known from high school math. We remember that the question of whether three points are collinear or not certainly is independent of the chosen projective coordinate system in $\mathbb{P}^2(\mathbb{R})$. Now choose a coordinate system in $\mathbb{P}^2(\mathbb{R})$ such that the points P and Q *lie on the line at infinity.* We then have to prove that the point R also lies at the line at infinity. This will prove the claim, since when they all lie at infinity, they are collinear, all being on the same line, namely the one at infinity.

This will signify that the lines AB and $A'B'$ are parallel, and that the same holds for AC and $A'C'$, both intersecting in points at infinity. We have to prove that BC and $B'C'$ also are parallel. But this is a well known fact from elementary planar geometry. The situation is illustrated in Fig. 13.2.

Here two of the pairs of triangles with top O are similar. But then so is the third pair, and the claim follows. □

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_13,

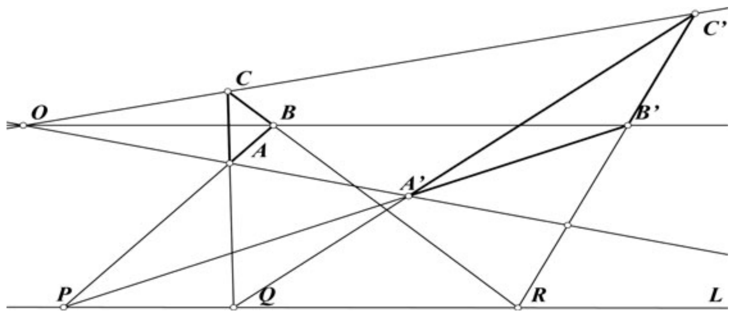


Fig. 13.1 Desargues' theorem

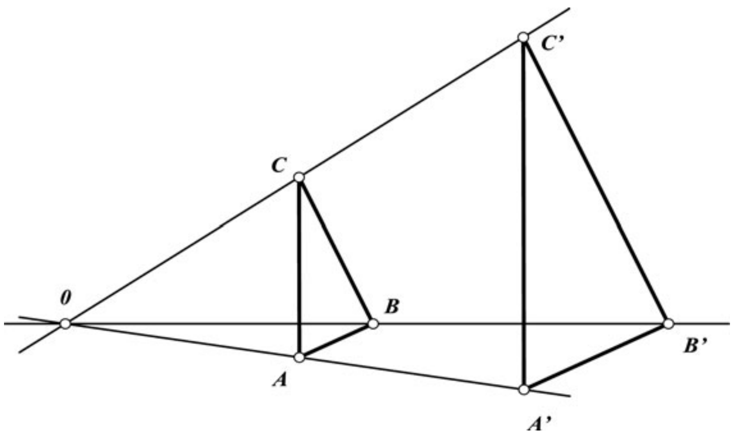


Fig. 13.2 Desargues' theorem, after a good choice of coordinate system in $\mathbb{P}^2(\mathbb{R})$

This theorem has the following

Corollary 4. *Let two triangles ABC and $A'B'C'$ be given in $\mathbb{P}^2(\mathbb{R})$, and assume that $A \neq A'$, $B \neq B'$ and $C \neq C'$. If the intersections of (the prolongations of) corresponding sides lie on the same line, then the lines through corresponding vertices all pass through the same point.*

Proof. This is an immediate consequence of the *Principle of Duality* for $\mathbb{P}^2(\mathbb{R})$, which we shall prove in the next section. Here we find that the dual result is actually a *converse* to the assertion of the theorem. The figure in the proof above is actually *self dual*. We leave the verification of this beautiful fact to the reader. □

13.2 Duality for $\mathbb{P}^2(\mathbb{R})$

For $\mathbb{P}^2(\mathbb{R})$ we have an *extended principle of duality.* Recall that the usual principle of duality in axiomatic projective geometry says the following: Any statement about points, lines and incidence, which may be deduced from the system of axioms, is transformed into another statement which also may be deduced from the axioms if the words *point* and *line* are interchanged. We refer to the latter statement as the *dual* of the former. But here we are not dealing with the axiomatic system, but rather with a special *model* for it. The collection of statements covered by the original principle of duality only consists of those which may be deduced from the axiomatic system given in Chap. 9, Sect. 9.1. And the model $\mathbb{P}^2(\mathbb{R})$ has many more true statements than that, in fact the Theorem of Desargues, which we have just proved in $\mathbb{P}^2(\mathbb{R})$ may not be deduced from that system of axioms. Indeed, this statement is usually taken as one of the additional axioms needed to finally arrive at the full axiomatic description of $\mathbb{P}^2(\mathbb{R})$, in the spirit of Hilbert.

For $\mathbb{P}^2(\mathbb{R})$ we have the following *stronger* principle of duality:

Theorem 19 (Duality for $\mathbb{P}^2(\mathbb{R})$). *If P denotes a true statement about $\mathbb{P}^2(\mathbb{R})$ dealing with points, lines and incidence, then the dual statement $P^\vee$ is also a true statement.*

Proof. The statement P may be translated into a collection of relations between the projective coordinates of the points involved and the coefficients of the equations of the lines involved. The relations will all be of the form

$$A_0\alpha_0 + A_1\alpha_1 + A_2\alpha_2 = 0,$$

where $A = (A_0 : A_1 : A_2)$ is a point and α_0, α_1 and α_2 are the coefficients in the equation for a line in $\mathbb{P}^2(\mathbb{R})$, so the line is given by

$$\alpha_0 X_0 + \alpha_1 X_1 + \alpha_2 X_2 = 0.$$

If we denote this line by ℓ, then the statement $A\ I\ \ell$, or $A \in \ell$, is equivalent to the relation above. For every point in $\mathbb{P}^2(\mathbb{R})$ we now let correspond a *line* in $\mathbb{P}^2(\mathbb{R})$ given by the equation whose coefficients are the coordinates of the point, and to every line we let correspond the point whose projective coordinates are the coefficients of the equation giving the line.

We then have the following beautiful and simple situation:

> The collection of algebraic relations between the coordinates of points in $\mathbb{P}^2(\mathbb{R})$ and coefficients of lines in $\mathbb{P}^2(\mathbb{R})$ which expresses the truth of the statement P is the same as the collection of relations which expresses the truth of $P^\vee$

This completes the proof. Note that this proof remains valid even if the collections of lines and/or points are infinite. □

13.3 Naive Definition and First Examples of Affine Plane Curves

For most purposes the following definition of an affine algebraic curve in $\mathbb{R}^2$ will suffice: It is a subset of $\mathbb{R}^2$ given as the set of points (x, y) which satisfy an equation

$$f(x, y) = 0,$$

where $f(X, Y)$ is a polynomial with real coefficients in the variables, or as we should rather say, *in the transcendentals* X and Y.

This definition suffices for most purposes, but it becomes insufficient when the need arises to consider *curves occurring with a certain multiplicity*. Thus for instance, the x-axis is given by the equation $y = 0$. But in some considerations it is convenient to consider the x-axis with *multiplicity 2*. This geometric object would then have the equation $y^2 = 0$.

We will return to this below, for now we content ourselves with the naive definition given above.

13.4 Straight Lines

In general straight lines are given by linear equations in x and y,

$$Ax + By + C = 0,$$

where A and B are not both zero. The curve given by this equation will intersect the x-axis in the point $(-\frac{C}{A}, 0)$ provided that $A \neq 0$. The curve given by $By = C$ is a line parallel with the x-axis, as in this case $B \neq 0$. Similarly, if $B \neq 0$ the line intersects the y-axis in the point $(0, -\frac{C}{B})$.

Give two distinct points in $\mathbb{R}^2$, (x_1, y_1) and (x_2, y_2). Then there is a unique line ℓ passing through them, and in High School we are usually taught some more or less cumbersome ways of finding the equation of this line. Knowing some *linear algebra* greatly facilitates this task, we now explain how.

We seek A, B and C such that ℓ is given by the equation

$$Ax + By + C = 0.$$

We then must have

$$Ax_1 + By_1 + C = 0$$

and

$$Ax_2 + By_2 + C = 0.$$

Here the numbers A, B and C are not all zero. Now we may regard the three relations as equations in the *unknowns* A, B, and C, with coefficients x, y and 1 for the first equation, $x_1, y_1, 1$ for the second and $x_2, y_2, 1$ for the third. As $A = B = c = 0$ also is a solution, the system does not have a unique solution, thus by Cramer's Theorem from Sect. 6.7 we find that *the determinant of the system* is zero. Thus the equation of the line passing through the points (x_1, y_1) and (x_2, y_2) may be written as a determinant as follows:

$$\begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0.$$

13.5 Conic Sections in the Affine Plane $\mathbb{R}^2$

A *Conic Section* is, as we know, a curve in the plane $\mathbb{R}^2$ of degree 2. The general form of the equation of such a curve is usually written as

$$q(x, y) = Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0.$$

One may wonder why some of the coefficients are written as 2 times some other constant. The answer to this is that writing the equation in this form is a great convenience later, in that many expressions we need will take a simpler form.

All such curves may be obtained by cutting a cone with a plane. This will be shown in Sect. 15.5.

We classify the conic sections as the ellipses, the parabolas and the hyperbolas. In addition to these, we have some *degenerate cases*, in that an ellipse may shrink to a point, a hyperbola may degenerate to two lines and a parabola collapse to one or degenerate into two parallel lines. The precise definition runs as follows:

Definition 12. The equation $q(x, y) = Ax^2+2Bxy+Cy^2+2Dx+2Ey+F = 0$ defines a degenerate conic section if the corresponding polynomial factorizes as

$$q(x, y) = Ax^2+2Bxy+Cy^2+2Dx+2Ey+F = (A_1x+B_1y+C_1)(A_2x+B_2y+C_2)$$

We remind the reader that in the non-degenerate case, the equation can be brought on one of the three so-called *canonical forms* shown in Figs. 13.3, 13.4 and 13.5.

In fact, we consider changes of coordinate system in $\mathbb{R}^2$ from x, y to x', y' of the form

$$x' = (x - a)\cos(v) + (y - b)\sin(v)$$

$$y' = -(x - a)\sin(v) + (y - b)\cos(v).$$

This corresponds to a new coordinate system with origin in (a, b) and rotated an angle v, see Fig. 13.6.

Such a shift may also be regarded as a transformation of the plane onto itself, known as a *rigid motion*. Throughout this chapter, by the term *a new coordinate system in* $\mathbb{R}^2$ we mean a new coordinate system of this type.

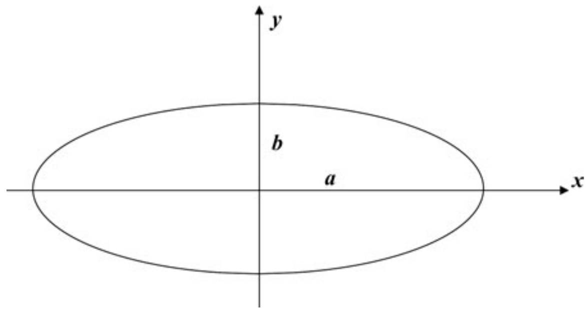


Fig. 13.3 The ellipse given by $(\frac{x}{a})^2 + (\frac{y}{b})^2 = 1$

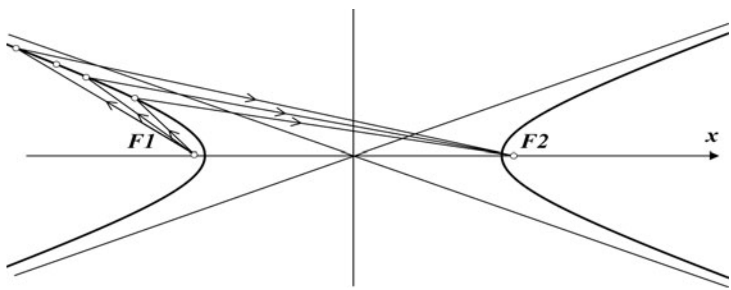


Fig. 13.4 The hyperbola given by $(\frac{x}{a})^2 - (\frac{y}{b})^2 = 1$. The asymptotes are also shown, those are the lines given by $(\frac{x}{a})^2 - (\frac{y}{b})^2 = 0$

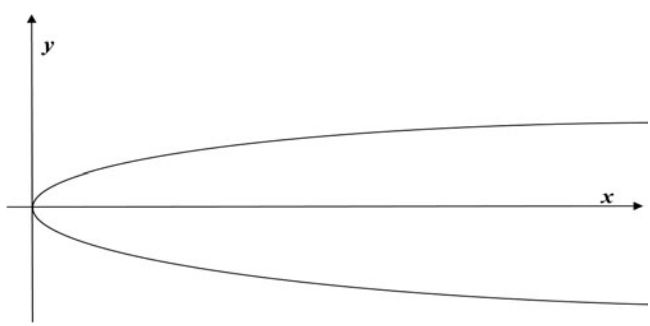


Fig. 13.5 The parabola given by $px = y^2$

We divide the non-degenerate conic sections into three classes, namely *ellipses, parabolas* and *hyperbolas.* After a change of coordinate system, any ellipse may be described by an equation of the type

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

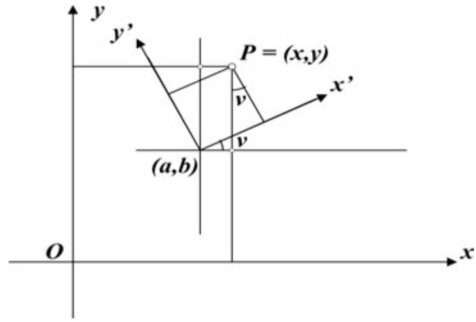


Fig. 13.6 A new coordinate system

A hyperbola is a conic section which may be transformed into the form

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = \pm 1,$$

after a change of coordinate system (and here the $\pm$ is really redundant).

Finally a parabola is a conic section which may be brought on the form

$$px = y^2.$$

In Fig. 13.3 we have indicated the *half-axes* of the ellipse. Their lengths are, respectively, a and b, and if they are equal then evidently we have a circle of radius $a = b$. We have not indicated the *focal points* of the ellipse, but their locations are at the longest axis of the ellipse, at a distance c from the center, i.e., the point of intersection of the two axes, where

$$a^2 = b^2 + c^2.$$

The focal points have two interesting properties. First of all, we have the

Proposition 8. *The ellipse is the locus of all points such that the sum of their distances from F_1 and F_2 is constant, namely $2a$ where a is the length of the longest half axis.*

Proof. Assume first that the ellipse is given by the equation

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

Denote a point on the ellipse by $P = (x, y)$. We show that the sum of the distances to the focal points is $2a$. We remark that there is an angle φ such that

$$x = a\cos(\varphi), y = b\sin(\varphi),$$

and in this way we obtain a parametric description of the ellipse. Indeed, if the point is on the ellipse, then the point $(\frac{x}{a}, \frac{y}{b})$ is on a circle of radius 1 and center at the origin, and conversely. The two focal points are $F_1 = (-c, 0)$ and $F_2 = (c, 0)$, thus the distances from P to these points are, respectively

$$PF_1 = \sqrt{(x+c)^2 + y^2} = \sqrt{(a\cos(\varphi) + c)^2 + b^2\sin^2(\varphi)}$$

and

$$PF_2 = \sqrt{(x-c)^2 + y^2} = \sqrt{(a\cos(\varphi) - c)^2 + b^2\sin^2(\varphi)}$$

which after a short computation yield

$$PF_1 = a + c\cos(\varphi) \text{ and } PF_2 = a - c\cos(\varphi).$$

Hence the claim follows.

Next, let F_1 and F_2 be two points, and let the distance between them be $2c$. Let $r > 2c$, and put $a = \frac{r}{2}$. Choose a coordinate system with origin at the mid point between F_1 and F_2, and x-axis along F_1F_2 directed towards F_2. Let $P = (x, y)$, then the condition that $F_1P + PF_2 = r$ is expressed as

$$\sqrt{(c+x)^2 + y^2} + \sqrt{(c-x)^2 + y^2} = 2a$$

see Fig. 13.7.

We now perform a standard trick, with which the ancients were well acquainted: We simply multiply both sides of the equality above by *the difference* of the two root expressions:

$$(\sqrt{(c+x)^2 + y^2} + \sqrt{(c-x)^2 + y^2})(\sqrt{(c+x)^2 + y^2} - \sqrt{(c-x)^2 + y^2})$$
$$= 2a(\sqrt{(c+x)^2 + y^2} - \sqrt{(c-x)^2 + y^2})$$

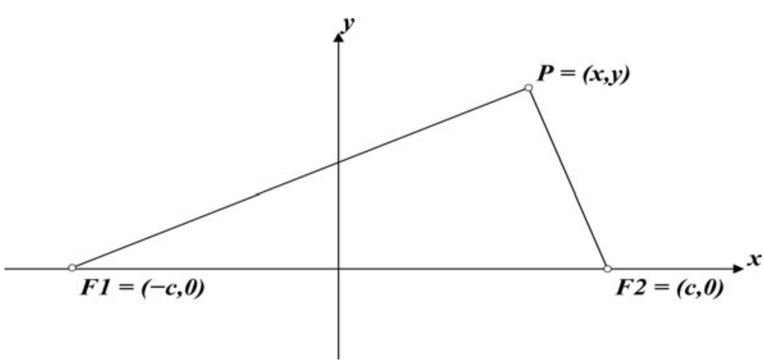


Fig. 13.7 The points F_1 and F_2 and the point $P = (x, y)$

which yields

$$((c+x)^2+y^2)-((c-x)^2+y^2)=2a(\sqrt{(c+x)^2+y^2}-\sqrt{(c-x)^2+y^2})$$

or

$$4cx=2a(\sqrt{(c+x)^2+y^2}-\sqrt{(c-x)^2+y^2})$$

thus

$$\sqrt{(c+x)^2+y^2}-\sqrt{(c-x)^2+y^2}=2\frac{c}{a}x.$$

Adding this to the original equality yields

$$\sqrt{(c+x)^2+y^2}=a+\frac{c}{a}x$$

which when squared yields

$$(c+x)^2+y^2=\left(a+\frac{c}{a}x\right)^2$$

and after a short computation this becomes

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

so letting $b=\sqrt{a^2-c^2}$ we get the equation of the ellipse on its usual form. □

Remark 1. Note that the parametric form of the ellipse given in the proof above is not the one obtained through introducing polar coordinates for the ellipse, but rather the one deduced from polar coordinates for the circle. Polar coordinates for the ellipse yields more complicated expressions. In fact, the angle φ is the so-called *eccentric angle of the point* $P=(x,y)$.

It also should be pointed out that strictly speaking the first part of the proof, involving the parametric form, is redundant: Indeed, the last part of the argument can be made to work both ways. We leave this analysis to the reader.

Using Proposition 8 we may draw an ellipse with given focal points and given half axis a by means of two needles, a piece of string and a pencil as shown in Fig. 6.4 in Chap. 6.

By means of Proposition 8 it is also easy to explain the reason for calling the two points F_1 and F_2 the *focal points*. The reader may contemplate the illustration in Fig. 13.8.

Ellipses form an important class of curves. As we know, the planets move around the Sun in orbits which are approximate ellipses, with the Sun at one of the focal points. Many comets have other types of orbits, parabolas or hyperbolas. Those constitute the two other classes of non-degenerate conic sections.

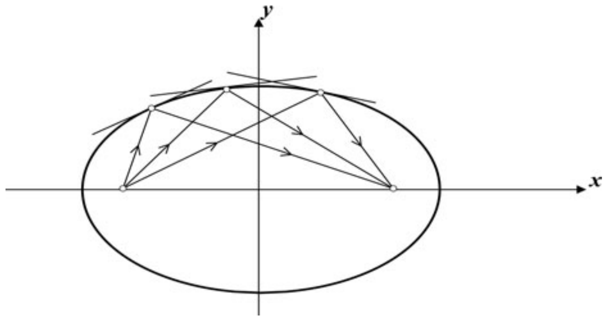


Fig. 13.8 Assume that the ellipse is a mirror. Then an observer at one focal point sees the other in any direction. Similarly, a source of light at one of the points will emit light rays which are reflected in rays which are focused at the other one. This is the second interesting property of the focal points of an ellipse

With a and b as before we define the *eccentricity* e by

$$e^2 = 1 - \left(\frac{b}{a}\right)^2 = \left(\frac{c}{a}\right)^2.$$

Thus if the ellipse is a circle then $e = 0$.

Parabolas may, by a change of coordinates in $\mathbb{R}^2$, be given by an equation of the form

$$px = y^2.$$

This class of conic sections can be obtained by deforming an ellipse in the following way: One of the focal points is kept fixed, the other is moved towards infinity. The remaining focal point will then be the focusing point for a ray of incoming signals, light or radio waves, which is parallel with the axis. We will see, by a closer analysis which we omit here, that under this stretching-procedure the eccentricity e will approach 1, as the largest half-axis tends to infinity faster than the smaller half axis. Thus parabolas are of eccentricity 1. Parabolas are given by a similar geometric property to the one which determines the ellipses, by a *focal point* F and a line L referred to as the *directrix* of the parabola. We have the following:

Proposition 9. *The parabola is the locus of all points such that their distances from a fixed point F and a fixed line L are equal.*

Proof. We choose a coordinate system so that the line L has equation $x + c = 0$ and $F = (c, 0)$. Then the condition becomes

$$\sqrt{(x-c)^2 + y^2} = x + c$$

which is equivalent to

$$y^2 = 4cx.$$

□

Hyperbolas form a class of conic sections which after a change of coordinate system in $\mathbb{R}^2$ are given by

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

The eccentricity e is given by

$$e^2 = 1 + \left(\frac{b}{a}\right)^2.$$

In particular $e > 1$.

Moreover, there are two focal points as indicated in Fig. 13.9, where the distance c from the origin is given by

$$c^2 = a^2 + b^2.$$

For hyperbolas we have the

Proposition 10. *The hyperbola is the locus of all points such that the difference of their distances from two fixed points F_1 and F_2 is constant.*

The proof is omitted. It is very similar to the last part of the proof of Proposition 8, using the last part of the remark after the proof.

Finally, the reflection property for the focal points of the hyperbola is illustrated in Fig. 13.9.

The equation

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 0,$$

yields two lines, the so-called asymptotes of the hyperbola whose equation it resembles. As x and y become very large, the two lines will be indistinguishable from the hyperbola itself: The curve approaches these lines as the point moves towards infinity. In precise terms, the two asymptotes are the tangents of the curve at its two points at infinity. We return to this phenomenon later in this chapter.

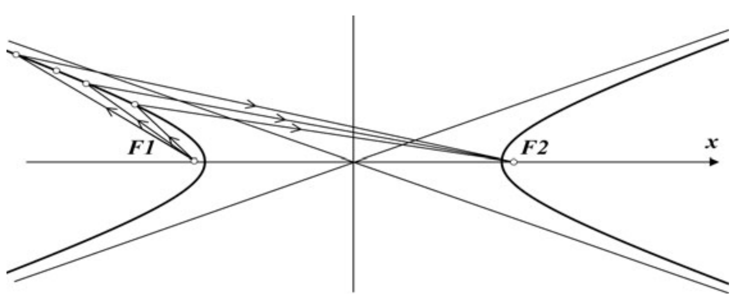


Fig. 13.9 The hyperbola have similar reflection properties to that of the ellipse

We next consider the following problem, which is a continuation of what we did regarding the equation of a line passing through two given, distinct points.

Consider 5 distinct points in $\mathbb{R}^2$, $P_i = (x_i, y_i)$, $i = 1, 2, \ldots, 5$. If the points are in *"sufficiently general position"*, then there is a unique, non-degenerate conic section, in other words a non-degenerate curve of degree 2, passing through them. Here we may immediately make precise the requirement *"in sufficiently general position."* It is the condition that no three of them be collinear. If three are collinear, but not four, then there is a unique *degenerate* conic section passing through them, namely the union of two lines, and if four are collinear but not all five then there are four degenerate conics through them, if all five are collinear then there are infinitely many degenerate conics passing through them, all having the line containing them as a component.

To find the equation of the unique conic section passing through the sufficiently general points, we may proceed in an analogous manner to what we did for lines. Indeed, the equation is

$$\begin{vmatrix} x^2 & xy & y^2 & x & y & 1 \\ x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 & 1 \\ x_2^2 & x_2y_2 & y_2^2 & x_2 & y_2 & 1 \\ x_3^2 & x_3y_3 & y_3^2 & x_3 & y_3 & 1 \\ x_4^2 & x_4y_4 & y_4^2 & x_4 & y_4 & 1 \\ x_5^2 & x_5y_5 & y_5^2 & x_5 & y_5 & 1 \end{vmatrix} = 0.$$

If uniqueness fails, the coefficients of this equation are all zero. If the conic section is degenerate, the equation factors as a product of two linear polynomials.

13.6 Constructing Points on Conic Sections by Compass and Straightedge

Constructions involving *conic sections* are *asymptotic Euclidian constructions.* In other words, using compass and straightedge in the legal way, we may construct as many points on any conic section as we wish, and thus a construction involving general conic sections may be completed to any prescribed degree of accuracy using compass and straightedge.

Proposition 8 gives a method for constructing as many points as we wish on an ellipse by the *Euclidian tools*. The construction is given in Fig. 13.10. The longest axis is PQ.

A point A on the ellipse satisfies $F_1A + AF_2 = r = PQ$. We proceed as follows: We subdivide the line segment F_1F_2 into n pieces, not necessarily of equal length. Call a point in the subdivision S. We then draw a circle with center F_1 and radius PS, and another circle with center F_2 and radius SQ. The two points of intersection

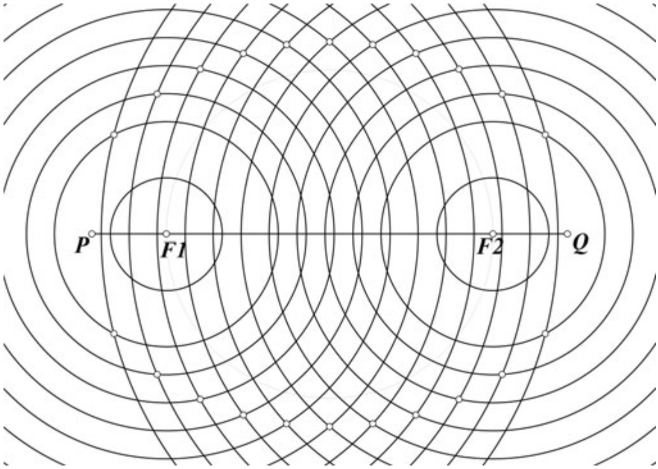
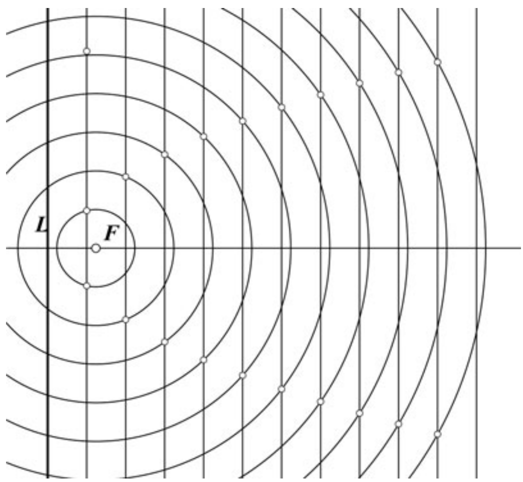


Fig. 13.10 The construction of points on an ellipse with given focal points F_1 and F_2, and given largest axis $r = 2a = PQ$

Fig. 13.11 The construction of points on a parabola with a given focal point F and directrix L



of these two circles then lie on the ellipse with largest axis r and focal points F_1 and F_2.

Similarly we may construct points on a parabola as shown in Fig. 13.11.

A point on the parabola is the intersection between a circle with center at F and radius ρ, and lines parallel with L, at a distance equal to ρ from L. Varying ρ we thus get as many points as we wish on the parabola.

Finally we show how to construct points on a hyperbola. We start out by marking the two focal points F_1 and F_2. About F_1 and F_2 we draw families of circles of radius Δ, 2Δ, $3\Delta, \dots, n\Delta$ for some small fixed distance Δ. We then obtain

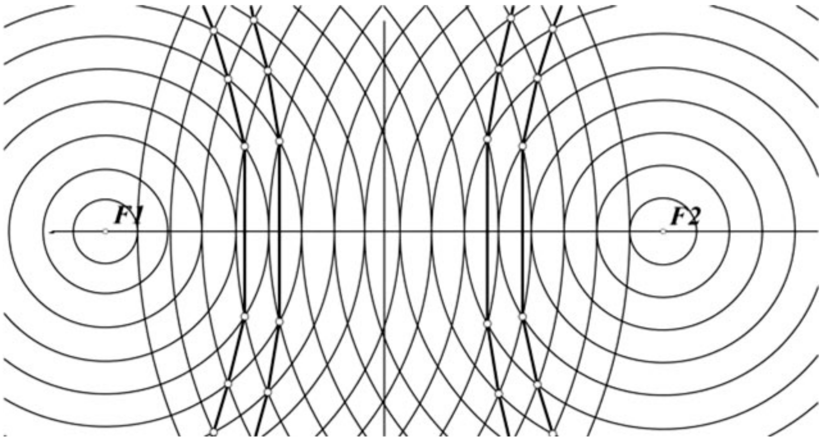


Fig. 13.12 The construction of points on a hyperbola with given focal points F_1 and F_2 and given difference between the distances from a point on the hyperbola to the two focal points

points on a family of hyperbolas, corresponding to fixed differences in distance to F_1 and F_2. This is shown in Fig. 13.12.

This principle is utilized in the navigational system GPS, the letters stand for *Global Positioning System.* The system works in 3-dimensions, but the same principle is used in the older, and now essentially outdated, LORAN-system which is 2-dimensional. Then we have several radio beacons, and a ship with a receiver and appropriate equipment is able to compute the difference between the distances to any two transmitters it is receiving. Thus the navigator may pinpoint the ship's position to several such hyperbolas, three signals will leave four possible locations, and in normal circumstances three of them may be ruled out from other information. GPS works with satellites located at F_1 and F_2, and we now get the position as being on a *surface* from receiving two satellites. The surface is obtained by rotating the hyperbola about the line F_1F_2, such a surface is a special case of what we call a *hyperboloid surface.*

With three satellites we get the possible positions on certain curves, to get points we need a minimum of four satellites. Then in principle there are eight possibilities. An added difficulty is that we are working with approximations. For these and other reasons we need to receive a larger number of satellites to get a good position.

13.7 Further Properties of Conic Sections

In the seventeenth century a conic section was defined as the locus of points with a certain fundamental relation to a fixed point and a fixed line. Due to Johan de Witt and John Wallis this definition may in some sense be understood as a precursor to the full algebraization of the subject. It is, however, used less frequently today. It has the advantage of being very geometric in nature, and to elucidate the difference between the three classes of conic sections, while at the same time providing a unified treatment. Here we give it in the form of a proposition:

Proposition 11. *Given a fixed point F and a line ℓ. Let $0 \leq e$ be a real number. Let $\mathcal{C}$ denote the set of point in $\mathbb{R}^2$ such that the ratio of the distance from F to the distance from ℓ is constant and equal to e. If $e = 0$*[1] *then the curve is a circle, if $0 < e < 1$ then it is an ellipse, for $e = 1$ it is a parabola and for $e > 1$ it is a hyperbola. All conic sections can be obtained in this way.*

Proof. We shall only sketch the proof here, but return to it in Sect. 13.8. For now, the case of $e = 0$ will give us some trouble, and right now we shall only indicate how this is dealt with. So assume that $e > 0$. We may choose a coordinate system such that F is the origin and such that ℓ is parallel with the y-axis, and intersects the x-axis at a distance p from the origin, where p may be positive, zero or negative. Then the condition is

$$\frac{\sqrt{x^2 + y^2}}{x + p} = e.$$

This yields after a short computation

$$(1 - e^2)x^2 - 2pe^2x + y^2 = e^2p^2,$$

and from this everything follows *except* for the case $e = 0$. But this case is troubling, since it would appear that the assertion is false as stated in this case. This point will be explained in Sect. 13.8. □

We return to the concept of *degeneration* for conic sections.

Theorem 20. *The equation*

$$q(x, y) = Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$$

yields a non-degenerate conic section if and only if the following determinantal criterion is satisfied:

$$\begin{vmatrix} A & B & D \\ B & C & E \\ D & E & F \end{vmatrix} \neq 0.$$

Proof. We need the notion of a *non-singular* point of a plane curve, we return to a refined treatment of this important concept in Sect. 14.3:

[1] $e = 0$ is strictly speaking not covered by the proposition as it is stated. See remark at the end of the proof. But this is the form in which this result is often stated.

Definition 13. Let Z be a plane curve given by the equation

$$f(x, y) = 0.$$

Let (x_0, y_0) be a point on the curve such that the two partial derivatives do not both vanish,

$$\left(\frac{\partial f}{\partial x}(x_0, y_0), \frac{\partial f}{\partial y}(x_0, y_0)\right) \neq (0, 0).$$

Such a point is called a non-singular point on the curve. At all non-singular points we define the tangent line[2] by the equation

$$\frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) = 0.$$

A point which is not non-singular is called a singular point.

Remark 2. In general the property for a point on a curve to be singular or non-singular is independent of the choice of affine coordinate system, which we shall prove as Proposition 19. A simpler case, which is immediate to verify, is that when checking singularity or non-singularity for a point P on some curve, we may assume $P = (0, 0)$ without loss of generality, modifying the equation appropriately, of course.

Now let $\mathcal{C}$ be the conic section given by the equation above, and let (x_0, y_0) be a point on it. Then

$$\frac{\partial q}{\partial x}(x_0, y_0) = 2Ax_0 + 2By_0 + 2D, \frac{\partial q}{\partial y}(x_0, y_0) = 2Bx_0 + 2Cy_0 + 2E.$$

Thus we observe that the point $(x_0, y_0) \in \mathcal{C}$ is a non-singular point if and only if it does not lie on both of the lines given by

$$Ax + By + D = 0, Bx + Cy + E = 0.$$

Since the equation of $\mathcal{C}$ may be written as

$$\begin{aligned} &Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F \\ &\quad = x(Ax + By + D) + y(Bx + Cy + E) + Dx + Ey + F, \end{aligned}$$

these considerations provide the proof of the

Lemma 2. *The point* $(x_0, y_0) \in \mathbb{R}^2$ *is a singular point of the curve given by the equation*

[2] This concept will be explained in more detail in Sect. 14.4.

$$q(x,y) = Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$$

if and only if

$$\begin{aligned} Ax_0 + By_0 + D &= 0 \\ Bx_0 + Cy_0 + E &= 0 \\ Dx_0 + Ey_0 + F &= 0 \end{aligned}$$

We next prove the following proposition:

Proposition 12. *A non-degenerate conic section in* $\mathbb{R}^2$ *does not have any singular points. Moreover, if* $\mathcal{C}$ *is a degenerate conic section in* $\mathbb{R}^2$ *without singular points, then* $\mathcal{C}$ *consists of two distinct parallel lines.*

Proof. Assume that (x_0, y_0) were a singular point. We may, without loss of generality, assume that $P = (0, 0)$, so the equation becomes

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey = 0.$$

The origin being a singular point, however, we have $D, = E = 0$, thus the equation is

$$Ax^2 + 2Bxy + Cy^2 = 0.$$

The zero locus of this equation in $\mathbb{R}^2$, in other words the set of all points $(a, b) \in \mathbb{R}^2$ such that $Axa^2 + 2Bab + Cb^2 = 0$, may consist of the origin alone, which is a degenerate conic section. Alternatively, the equation may be written as

$$(a_1x + b_1y)(a_2x + b_2y) = 0$$

which either represents a double line, or two lines through the origin, both cases being degenerate conics.

Next, assume that $\mathcal{C}$ is degenerate but without singular points. We may assume that $(0, 0) \in \mathcal{C}$. Thus the case of $\mathcal{C}$ being a single point, i.e., the origin, is excluded as the equation would then be $Ax^2 + 2Bxy + Cy^2 = 0$, so $(0, 0)$ would be singular. Hence $\mathcal{C}$ consists of two (possibly coinciding) lines. Then the equation factors as

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey = (ax + by)(cx + dy + e).$$

If the two lines given by $ax + by = 0$ and $cx + dy + e = 0$ coincide, then we may write the equation as $(ax + by)^2 = 0$, all of whose points are singular. If the lines are distinct but intersect, the point of intersection will have to be a singular point as is easily checked. Thus the lines are parallel, and the claim is proven. □

We may now prove Theorem 20. Assume first that

$$\begin{vmatrix} A & B & D \\ B & C & E \\ D & E & F \end{vmatrix} \neq 0.$$

Then by Lemma 2, $\mathcal{C}$ has no singular points: Indeed, by Cramer's Theorem 5 in Sect. 6.7 the only solution to the system of equations would be $(x_0, y_0, z_0) = (0, 0, 0)$. Thus by Proposition 12 it is non-degenerate, unless it consists of two parallel lines, i.e.,

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey = (ax + by)(ax + by + e)$$

where $e \neq 0$. Thus $A = a^2, B = ab, C = b^2, D = \frac{1}{2}ae, E = \frac{1}{2}be$ and $F = 0$. But with these values the determinant is zero.

Conversely, assume that $\mathcal{C}$ is non-degenerate. Evidently we may assume $(0, 0) \in \mathcal{C}$ without loss of generality, so $F = 0$. If

$$\begin{vmatrix} A & B & D \\ B & C & E \\ D & E & F \end{vmatrix} = 0$$

then by Cramer's Theorem 5 in Sect. 6.7, there are real numbers x_0, y_0 and z_0 not all zero such that

$$\begin{aligned} Ax_0 + By_0 + Dz_0 &= 0 \\ Bx_0 + Cy_0 + Ez_0 &= 0 \\ Dx_0 + Ey_0 + Fz_0 &= 0. \end{aligned}$$

If $z_0 \neq 0$, we may assume $z_0 = 1$, dividing through the equations. Then (x_0, y_0) is a singular point on $\mathcal{C}$, by Lemma 2. This contradicts $\mathcal{C}$ being non-degenerate, by Proposition 12.

If $z_0 = 0$ in all triples (x_0, y_0, z_0) satisfying the system of equations above, then we choose one solution $(x_0, y_0, 0) \neq (0, 0, 0)$, and note that $(tx_0, ty_0, 0)$ also is a solution for all $t \in \mathbb{R}$. Since we have

$$\begin{aligned} Atx_0 + Bty_0 &= 0 \\ Btx_0 + Cty_0 &= 0 \\ Dtx_0 + Ety_0 &= 0 \end{aligned}$$

we find that $(tx_0, ty_0) \in \mathcal{C}$ for all $t \in \mathbb{R}$ since

$$\begin{aligned} &tx_0(Atx_0 + Bty_0) + ty_0(Btx_0 + Cty_0) + 2(Dtx_0 + Ety_0) \\ &\quad = q(tx_0, ty_0) = 0. \end{aligned}$$

But then $\mathcal{C}$ contains the line joining $(0, 0)$ and (x_0, y_0), contradicting the assumption of it being non-degenerate. □

We conclude this section by discussing tangents to non-degenerate conic sections in $\mathbb{R}^2$, as well as the related concepts of *pole and polar line.*

So let $P = (x_0, y_0)$ be a point on the non-singular conic section given by the equation

$$q(x, y) = Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0.$$

Then the tangent at P is given by

$$(Ax_0 + By_0 + D)(x - x_0) + (Bx_0 + Cy_0 + E)(y - y_0) = 0,$$

which after a short calculation, using that $q(x_0, y_0) = 0$, takes the following beautiful form:

$$Ax_0x + B(y_0x + x_0y) + Cy_0y + D(x + x_0) + E(y + y_0) + F = 0.$$

This equation is also important when the point is not on C. Indeed, we have the following

Proposition 13. *Let $P = (x_0, y_0)$ be a point and let $\mathcal{C}$ be the conic section given by the equation*

$$q(x, y) = Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0.$$

There are two tangent lines to $\mathcal{C}$ passing through P, coinciding if P is on C. Let the two points of tangency be Q_1 and Q_2. Then the line p passing through Q_1 and Q_2 is given by the equation

$$Ax_0x + B(y_0x + x_0y) + Cy_0y + D(x + x_0) + E(y + y_0) + F = 0.$$

Proof. The situation is shown in Fig. 13.13.

Let $Q_1 = (x_1, y_1)$ and $Q_2 = (x_2, y_2)$, then the two tangents in question will have equations

$$Ax_1x + B(y_1x + x_1y) + Cy_1y + D(x + x_1) + E(y + y_1) + F = 0,$$

$$Ax_2x + B(y_2x + x_2y) + Cy_2y + D(x + x_2) + E(y + y_2) + F = 0.$$

These lines pass through $P = (x_0, y_0)$, thus

$$Ax_1x_0 + B(y_1x_0 + x_1y_0) + Cy_1y_0 + D(x_0 + x_1) + E(y_0 + y_1) + F = 0,$$

$$Ax_2x_0 + B(y_2x_0 + x_2y_0) + Cy_2y_0 + D(x_0 + x_2) + E(y_0 + y_2) + F = 0.$$

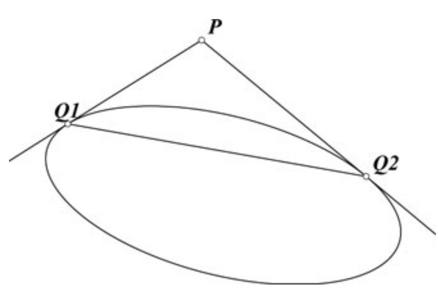


Fig. 13.13 The line joining the two points of tangency

But this demonstrates that the line whose equation is given in the assertion of the proposition, does indeed pass through the two points Q_1 and Q_2. Hence the claim follows. □

Definition 14. The point P and the line p in Proposition 13 are called the pole and the polar line corresponding to each other.

Remark 3. If we take a point inside an ellipse, then there will be no real points of tangency, even though we get a well defined polar line using the equation. But if we compute the *complex* points of tangency, we find that corresponding coordinates are complex conjugates, and we get a real line joining them.

Moreover, if we choose the center of the unit circle, say, then the "line" given by the formula is just $0 = 1$, which has no points on it. The explanation for this is that the polar of the center is the *line at infinity.* Thus we see here both the need for computing complex points as well as for considering points at infinity.

The latter is the subject of the next section.

13.8 Conic Sections in the Projective Plane

We shall now consider the *projective closure* of the conic sections in $\mathbb{R}^2$. We substitute

$$x = \frac{X_1}{X_0}, y = \frac{X_2}{X_0}$$

into the equation for a conic section $\mathcal{C}$,

$$q(x, y) = Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0,$$

which after clearing denominators, yields the following equation

$$\begin{aligned} &Q(X_0, X_1, X_2) \\ &\quad = AX_1^2 + 2BX_1X_2 + CX_2^2 + 2DX_0X_1 + 2EX_0X_2 + FX_0^2 = 0. \end{aligned}$$

This is a *homogeneous equation*. Recall that a polynomial $F(X_0, X_1, X_2)$ is said to be homogeneous if *all monomials which occur in it are of the same degree*. If $F(X_0, X_1, X_2)$ is a homogeneous polynomial of degree 2, then as is immediately verified

$$F(ta_0, ta_1, ta_2) = t^2 F(a_0, a_1, a_2)$$

for all real numbers t, a_0, a_1, a_2. Hence we get the equation for a curve in $\mathbb{P}^2(\mathbb{R})$, which when intersected with $D_+(X_0) = \mathbb{R}^2$ gives back the original curve. We call this curve in $\mathbb{P}^2(\mathbb{R})$ the *projective closure* of $\mathcal{C}$.

We first wish to determine its points at infinity. Those are the points $C \cap V_+(X_0)$. The point $(u : v : 0)$ is in $\mathcal{C}$ if

$$Au^2 + 2Buv + Cv^2 = 0,$$

and we immediately get the

Proposition 14. *(1) $\mathcal{C}$ has no real points at infinity if $B^2 - AC < 0$.*
(2) $\mathcal{C}$ has one real point at infinity if $B^2 - AC = 0$.
(3) $\mathcal{C}$ has two points at infinity if $B^2 - AC > 0$.

Thus (1) corresponds to a possibly degenerate ellipse, (2) to a possibly degenerate parabola and (3) to a possibly degenerate hyperbola.

In general, whenever

$$F(X_0, X_1, X_2) = 0$$

is a homogeneous equation, it gives a curve in $\mathbb{P}^2(\mathbb{R})$. Let $P = (a_0, a_1, a_2)$ be a point on it. In Sect. 15.4 we show that the equation

$$\frac{\partial F}{\partial X_0}(a_0, a_1, a_2)X_0 + \frac{\partial F}{\partial X_1}(a_0, a_1, a_2)X_1 + \frac{\partial F}{\partial X_2}(a_0, a_1, a_2)X_2 = 0$$

yields the tangent line at P.

Definition 15. If the partial derivatives involved in the equation for the tangent line all vanish at some point on the curve, then the point is said to be a singular point. If they do not all vanish, the point is called non-singular.

The equation for the tangent to the conic section in $\mathbb{P}^2(\mathbb{R})$ given by the equation $Q(X_0, X_1, X_2) = 0$ at the point $P = (x_0, x_1, x_2)$ is

$$\begin{aligned}(Ax_1 + Bx_2 + Dx_0)X_1 + (Bx_1 + Cx_2 + Ex_0)X_2 \\ + (Dx_1 + Ex_2 + Fx_0)X_0 = 0\end{aligned}$$

or written on a more appealing form

$$\begin{aligned}Ax_1X_1 + B(x_1X_2 + x_2X_1) + Cx_2X_2 + D(x_0X_1 + x_1X_0) \\ + E(x_0X_2 + x_2X_0) + Fx_0X_0 = 0\end{aligned}$$

This is similar to what we found in the affine case.

If the point P is singular, then its projective coordinates constitute a non-trivial solution of the homogeneous system of equations which we encountered in the previous section:

$$\begin{aligned}Au + Bv + Dw &= 0\\ Bu + Cv + Ew &= 0\\ Du + Ev + Fw &= 0.\end{aligned}$$

In $\mathbb{P}^2(\mathbb{R})$ we have the best way of understanding the relation between *pole and polar*. With the conic section given by

$$AX_1^2 + 2BX_1X_2 + CX_2^2 + 2DX_0X_1 + 2EX_0X_2 + FX_0^2 = 0$$

and a point $(x_0 : x_1 : x_2)$, the polar is given by the equation

$$Ax_1X_1 + B(x_1X_2 + x_2X_1) + Cx_2X_2 + D(x_0X_1 + x_1X_0) \\ + E(x_0X_2 + x_2X_0) + Fx_0X_0 = 0.$$

Now we may compute the polar line of a circle with respect to the center. Let the circle be given by

$$x^2 + y^2 = R^2.$$

The projective closure is

$$X_1^2 + X_2^2 - R^2X_0^2 = 0,$$

and the point is $P = (1 : 0 : 0)$. The polar line is given by

$$-R^2X_0 = 0.$$

Thus we get the line at infinity, as claimed earlier.

We are now also in the position to clarify the remaining aspects of the definition from Proposition 11. We chose a coordinate system such that F is the origin and such that ℓ is parallel with the y-axis, and intersects the x-axis at a distance p from the origin, where p may be positive, zero or negative. Then the condition defining the curve is

$$\frac{\sqrt{x^2 + y^2}}{x + p} = e,$$

which yields the equation

$$(1 - e^2)x^2 - 2pe^2x + y^2 = e^2p^2.$$

The projective closure of this curve in $\mathbb{P}^2(\mathbb{R})$ is given by the equation

$$(1 - e^2)X_1^2 - 2pe^2X_0X_1 + X_2^2 - e^2p^2X_0^2 = 0,$$

and the origin becomes $(1 : 0 : 0)$. The polar of our conic section with respect to the point $(x_0 : x_1 : x_2)$ is given by

$$(1 - e^2)x_1X_1 - pe^2x_0X_1 - pe^2x_1X_0 + x_2X_2 - e^2p^2x_0X_0 = 0.$$

With $x_0 = 1, x_1 = x_2 = 0$, this becomes the line

$$pe^2X_1 + e^2p^2X_0 = 0,$$

or

$$X_1 + pX_0 = 0,$$

and in $\mathbb{R}^2$ this is the line $x = -p$, thus the directrix. If furthermore we assume $e = 0$, then the second equation gives a line, whereas the first one does not. As indicated in Sect. 13.7, we let e tend to 0 and p tend to infinity to capture this situation. Thus $\frac{1}{p} \to 0$ and the line becomes $X_0 = 0$. This line is the line at infinity.

We sum up our findings:

Proposition 15. *The polar of a focal point of a conic section is a directrix.*

We note, however, that the concepts of eccentricity, focal point and directrix depend on the representation of the curve in $\mathbb{R}^2$. Thus for instance, the circle given by $x^2 + y^2 = 1$ has the projective closure in $\mathbb{P}^2(\mathbb{R})$

$$X_1^2 + X_2^2 - X_0^2 = 0,$$

which when restricted to $D_+(X_1)$ yields a hyperbola.

13.9 The Theorems of Pappus and Pascal

Let $\mathcal{C}$ be a conic section in $\mathbb{P}^2(\mathbb{R})$, and let A, B, C as well as A', B', C' be points on $\mathcal{C}$. We assume $\mathcal{C}$ to be non-degenerate, or degenerate as two separate lines ℓ_1 and ℓ_2, called the *components* of the degenerate conic section. In the former case we may choose the 6 points freely on $\mathcal{C}$, but in the latter case the points may be chosen according to the following restriction: No point may be chosen at the intersection of the two components, and the first 3 must lie on ℓ_1 and the last 3 on ℓ_2. See Fig. 13.14, where the notation to be used in the sequel is introduced.

The points P_A, P_B and P_C are the points of intersection between, respectively, the lines CB' and $C'B$, the lines AC' and $A'C$, and between AB' and $A'B$. The subscript marking the point is the letter missing in the designation of the lines.

We have already encountered *Pappus of Alexandria* in Sect. 4.18. The following result is referred to as *Pappus' Theorem* in the degenerate case and as *Pascal's Theorem* (named after Blaise Pascal) in the non-degenerate case:

Theorem 21. *The points P_A, P_B and P_C are collinear.*

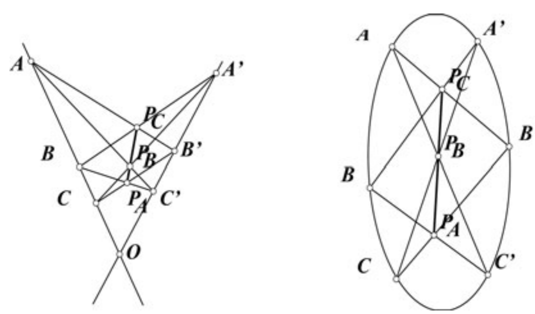


Fig. 13.14 Illustration to Pappus' and Pascal's theorems, with notation

Proof. Assume first that the conic section is non-degenerate. By Proposition 7 we may normalize the coordinate system by selecting four points in $\mathbb{P}^2(\mathbb{R})$, no three of them being collinear, and then introduce a new projective coordinate system in $\mathbb{P}^2(\mathbb{R})$ in which these four points have projective coordinates (0:0:1), (1:0:0), (1:1:1), and (0:1:0). As straight lines and collinearity are of course independent of the chosen coordinate system, we may choose the four points so as to obtain a very simple equation for the conic section, and then hopefully be able to prove the result easily. This is the strategy.

We choose as our points, in this order, C', A, B', and a point on the tangent line to the conic section at the point C'. In the projective coordinate system which this gives rise to, C' is the point at infinity on the y-axis and the tangent to $\mathcal{C}$ at C' is the line $V_+(X_0)$, the line at infinity. The point A is the origin. $B' = (1, 1)$, and this is all our information on the points.

We can now draw several conclusions. First of all, we have a non-degenerate conic section in $\mathbb{R}^2$ which has exactly one point at infinity, thus it is a parabola. Moreover, that point at infinity is the infinite point on the y-axis, and the curve passes through the origin, thus the parabola is of the type $y = px^2$. It also passes through the point $(1, 1)$, hence $p = 1$. Thus the conic section $\mathcal{C}$ restricted to $\mathbb{R}^2$ is given by the equation

$$y = x^2.$$

The situation which we have arrived at is illustrated in Fig. 13.15.

We now put $C = (c, c^2)$, $B = (b, b^2)$ and $A' = (a, a^2)$. Here $c, b < 0$, of course, while $a > 0$.

It now is an elementary exercise to check that the points P_A, P_B, P_C are as follows:

$$P_A = (b, b + bc - c),\ P_B = (0, -ac),\ P_C = \left(\frac{ab}{a+b-1}, \frac{ab}{a+b-1}\right).$$

Indeed, for P_A clearly the first coordinate is b. The line CB' has the equation

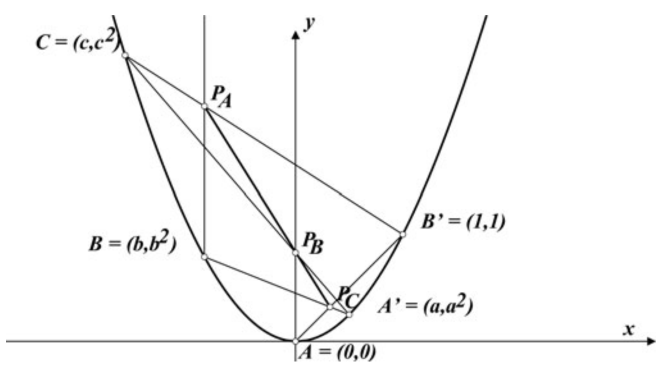


Fig. 13.15 The normalized situation

$$y = (c + 1)x - c,$$

thus the claimed second coordinate follows. P_B lies on the y-axis, and the equation for the line $A'C$ is

$$y = (a + c)x - ac,$$

and the second coordinate follows as claimed. Finally, P_C is the intersection between the two lines

$$y = x \text{ and } y = (a + b)x - ab.$$

We thus get

$$x(a + b - 1) = ab,$$

and we get P_C as claimed. To prove that these three points are collinear, we compute the determinant

$$\begin{vmatrix} 1 & b & b + bc - c \\ 1 & 0 & -ac \\ 1 & \frac{ab}{a+b-1} & \frac{ab}{a+b-1} \end{vmatrix} = 0,$$

and the proof is complete in the non-degenerate case.

We next turn to the degenerate case, which we deduce from the non-degenerate case by a classical method known as *degeneration*. We fit a family of hyperbolas having the two lines as asymptotes into the picture, as shown in Fig. 13.16. The algebraic details on how this is done should be clear by now: We may assume that the lines intersect at the origin, and that their equations may be written as

$$\left(\frac{y}{b}\right)^2 - \left(\frac{x}{a}\right)^2 = 0.$$

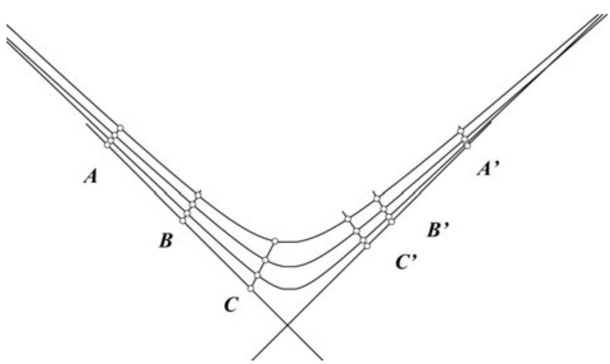


Fig. 13.16 Degenerating a family of hyperbolas to the given lines

We then consider the family of hyperbolas

$$\left(\frac{y}{b}\right)^2 - \left(\frac{x}{a}\right)^2 = \epsilon,$$

where ϵ is a positive number which we let tend to zero. Clearly, then, the family of hyperbolas will tend to the limiting position of the two lines, their common asymptotes. Now we may chose 6 points on each hyperbola, denote them by $A(\epsilon), \dots, C'(\epsilon)$, in such a way that $A(\epsilon)$ tends to A as $\epsilon \longrightarrow 0$, etc. Since we have collinearity for the three points in question for each value of ϵ, the same must hold in the limit. This completes the proof. □

Chapter 14
Algebraic Curves of Higher Degrees in the Affine Plane $\mathbb{R}^2$

14.1 Curves of Degree 3 and 4 in $\mathbb{R}^2$

Apart from *lines*, the simplest class of curves in the plane $\mathbb{R}^2$ is the class of *conic sections*, which we first treated in Sect. 4.7. The parabola is one of the three types of conic sections. It has the equation $y = x^2$, if necessary after a suitable change of coordinate system in $\mathbb{R}^2$. It has the graph shown in Fig. 14.1.

The simplest curve of *higher degree*, by which we mean degree higher than 2, is the curve known as the *cubic parabola*. A typical representative of this class of curves has the equation

$$y = x^3,$$

and its graph is shown in Fig. 14.2.

The next step in complexity is a curve which may be brought on the form

$$y^2 = x^3$$

It is called a *semi-cubic parabola.* It has the graph displayed to the left in Fig. 14.3. To the right we see a *degenerate cubic curve*, given by the equation

$$x^3 + x^2y + xy^2 + y^3 + x^2 - y^2 = 0.$$

We say that a curve is degenerate if it decomposes into the union of two curves of lower degrees.

For a cubic curve this would signify the curve being the union of a conic section and a line, or of three lines (some possibly coinciding). An example of such a curve is shown in to the right in Fig. 14.3. In addition to the concept of a degenerate curve, and the related process of *degeneration of a family of curves* is important. The simplest example of a degenerate cubic curve would be the y-axis with multiplicity 3. Its equation is $x^3 = 0$. We have not yet made the notion of curves with multiplicity precise, this comes in Sect. 14.2. But we may already at this point consider a family of semi-cubic parabolas, *degenerating to* the triple y-axis. Namely, consider the curves depending on the parameter t, as $t \to 0$:

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_14,

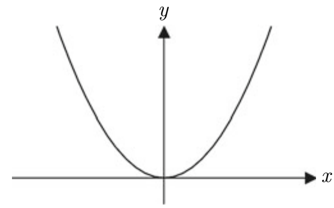


Fig. 14.1 The parabola

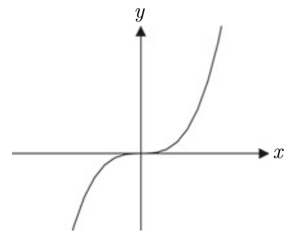


Fig. 14.2 The cubic parabola

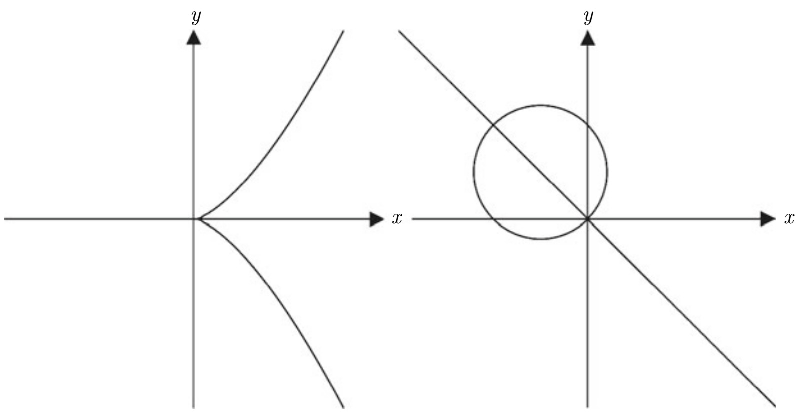


Fig. 14.3 The semi-cubic parabola to the *left*, and a degenerate cubic curve to the *right*

$$ty^2 = x^3.$$

We show some members of this family in Fig. 14.4. The values of t in the plots are $t = 10, 4, 1, 0.1$.

We note also that when $t \to \infty$, then the limit is the x-axis with multiplicity 2, since this degeneration is equivalent to letting u tend to 0 for the family given by the equation

$$y^2 = ux^3.$$

We have used the term degeneration loosely, without a formal definition. The idea we intend to convey by this, is to have one curve, say the semi-cubic parabola $y^2 = x^3$, be a member of a family of curves depending on a parameter, all but a finite number of which are of the same type. Then the exceptional members

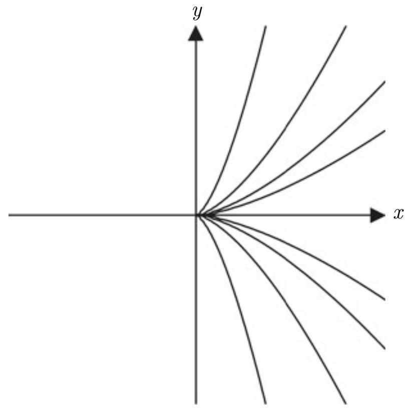


Fig. 14.4 Degeneration of $ty^2 = x^3$ to the triple y-axis

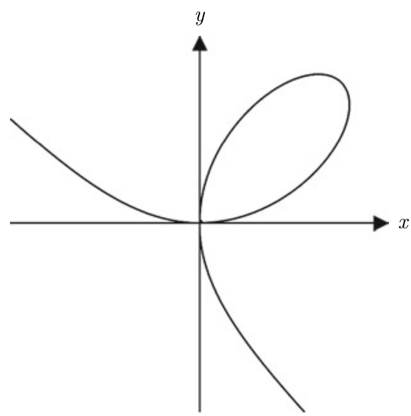


Fig. 14.5 The Folium of Descartes, given by the equation $x^3 + y^3 = 3xy$

are understood as degenerate cases. This is, of course, the way we may view two intersecting lines as a degenerate hyperbola, or a double line as a degenerate hyperbola or a degenerate parabola, and so on.

Two more types of non-degenerate curves of degree three exist, up to a *projective change of coordinate system*. We will explain this *projective equivalence* for curves in $\mathbb{P}^2(\mathbb{R})$ (and in $\mathbb{R}^2$) later, in Sect. 15.5. The simple *affine equivalence* for two curves means that one may be obtained from the other by a suitable *affine transformation*. This kind of equivalence is more complicated than the projective equivalence, there are more equivalence classes of affine cubic curves under this affine equivalence. But from our point of view, the projective equivalence is more interesting than the affine one.

The first of the remaining classes of cubic curves is represented by the *Folium of Descartes*. It was given as an example by Descartes in an argument over how to properly define tangent lines to curves. The Folium is shown in Fig. 14.5.

We also give another curve, belonging to the same class as the Folium under projective equivalence, but to a separate class under the affine equivalence. It looks

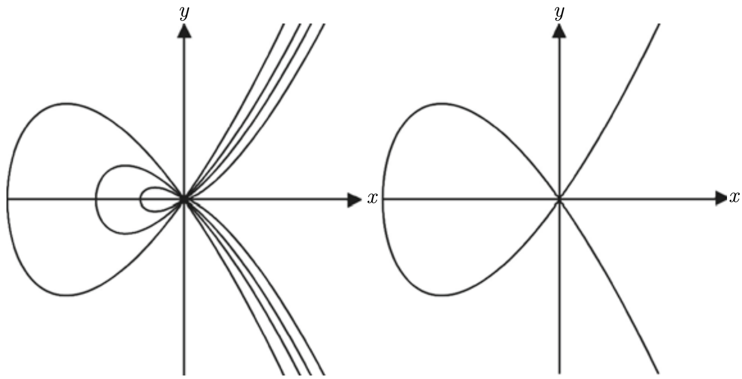


Fig. 14.6 The usual nodal cubic to the *right*, the degeneration to the *left*

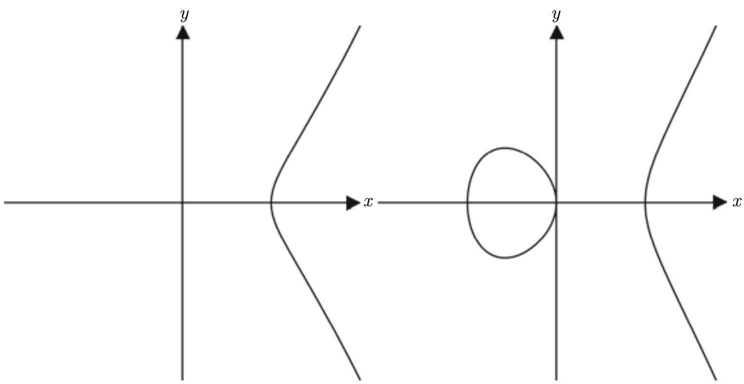


Fig. 14.7 To the left an unusual "nodal cubic". To the right an elliptic curve

somewhat similar to the semi-cubic parabola. In fact, the latter may be obtained by deforming the former.

This is the simplest and most used example of a *nodal cubic curve* in $\mathbb{R}^2$. It is shown to the right in Fig. 14.6. The deformation referred to is obtained from the family

$$y^2 - x^3 - tx^2 = 0,$$

and to the left in Fig. 14.6 we see some of the corresponding plots, for $t = 0, 0.5, 2$.

We also plot the curve given by

$$y^2 - x^3 + x^2 = 0$$

in to the left in Fig. 14.7. Actually, the origin is on the curve, but that point appears to be isolated from the main part of it. But there are complex points, invisible in $\mathbb{R}^2$, which establish the connection. To the right we see the elliptic cubic given by $y^2 - x^3 + x = 0$. The third class of cubic curves in $\mathbb{R}^2$ under the projective equivalence, is the class of *elliptic cubic curves*. But by all means: They are not ellipses! Ellipses are conics, elliptical curves are called so for an involved reason, ultimately coming from the computation of the length of a curve segment of an ellipse. Figure 14.7 shows an elliptic cubic curve in $\mathbb{R}^2$.

Elliptic curves constitute an important class of curves. But an extensive investigation of this theme falls outside the scope of the present book. We only mention two points. First of all, elliptic curves are tied closely to the *torus surfaces*. But this link comes from the inclusion of *complex points* on a curve. This will be explained in Sect. 14.2. The second point is that we need to investigate the behavior of the curves at infinity, also to be explained in Sect. 14.2. In fact, adding the points at infinity, the points of an elliptic cubic curve as the one we give here, form a very interesting *Abelian group!* But this theory also falls outside the scope of the present book.

A variation of the Folium of Descartes is provided by the *Trisectrix of Maclaurin*. As the name indicates, this is a curve which may be used to trisect an angle. The curve is given by $x^3 + xy^2 + y^2 - 3x^2 = 0$. Two lines are drawn through a point on it, located in the first quadrant. One line passes through the origin, the other through the point $(2, 0)$. We show the situation in Fig. 14.8. The curve intersects the x-axis at the origin O and in the point $Q = (3, 0)$. We mark the point $P = (2, 0)$, and let $u = \angle QPP'$ be the angle which is to be trisected. It is not difficult to use the results which we prove in Sect. 17.4 to show that then the angle $v = \angle QOP'$ is one third of u.

Another curve looks like a *clover leaf*. It has equation

$$(x^2 + y^2)^2 + 3x^2y - y^3 = 0$$

and is shown in to the left in Fig. 14.9. We can also get an airplane wing as an algebraic curve, it has the equation $x^4 + x^2y^2 - 2x^2y - xy^2 + y^2 = 0$, shown to the right in Fig. 14.9.

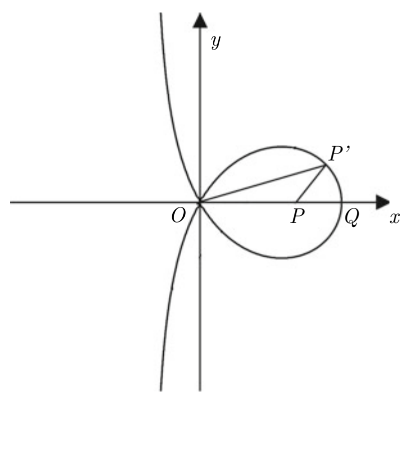


Fig. 14.8 The Trisectrix of Maclaurin

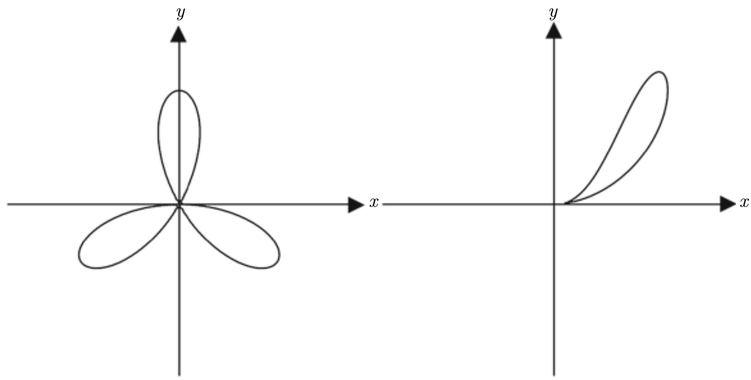


Fig. 14.9 The Clover Leaf curve to the *left*, the airplane wing to the *right*

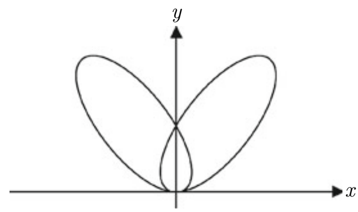


Fig. 14.10 A curve with two singular points

We finally show, in Fig. 14.10, a curve of degree four with two points where it crosses itself, that is to say with two *singular points*. It is given by the equation $2x^4 - 3x^2y + y^4 - 2y^3 + y^2 = 0$. We now have a sufficient base of *examples* to appreciate some more general theory. We already mentioned the need to incorporate complex points in connection with elliptic cubic curves above. We now take a closer look.

14.2 Affine Algebraic Curves

Let the curve C be given by the equation $f(x, y) = 0$. We then study the set of pairs (u, v) of *complex numbers* such that $f(u, v) = 0$. So we consider the zero locus of $f(x, y) = 0$, the set of all $(a, b) \in \mathbb{C}^2$ such that $f(a, b) = 0$. We may denote this set by $C(\mathbb{C})$, and the curve considered as a subset of $\mathbb{R}^2$ we may denote by $C(\mathbb{R})$. If we identify $\mathbb{C}^2$ with $\mathbb{R}^4$, this locus is identified with a *surface* defined by two equations. Namely, writing

$$u = x_1 + ix_2, v = x_3 + ix_4$$

and

$$f(u, v) = f_1(x_1, x_2, x_3, x_4) + if_2(x_1, x_2, x_3, x_4)$$

then f_1 and f_2 are polynomials with real coefficients in four variables, and the set of all complex points on the curve is given as

$$C(\mathbb{C}) = \left\{ (a_1, a_2, a_3, a_4) \in \mathbb{R}^4 \;\middle|\; \begin{matrix} f_1(a_1, a_2, a_3, a_4) = 0 \\ f_2(a_1, a_2, a_3, a_4) = 0 \end{matrix} \right\}.$$

This is a surface in 4-space, in $\mathbb{R}^4$, defined by two polynomials. In many situations we really need to include all complex points of a curve, although we usually still confine ourselves to sketch the real points only. And even if the complex points form a surface in $\mathbb{R}^4$, it is important to keep in mind that we really are studying a *curve* in the plane, and not a surface in four space. Indeed, of we switch to regard our object under study as a surface in $\mathbb{R}^4$, then *it will also have complex points*, thus yielding a *fourfold in* $\mathbb{R}^8$, and so on. Thus we have to remember that we are studying complex points on a curve in the plane, rather than the real points of a surface in 4-space.

The further important extension is to include the *points at infinity* of a curve. This is a somewhat more technical matter, which we come to in Chap. 15, where we study *projective curves*. But first we give more details on the affine case.

We are given a curve in the plane $\mathbb{R}^2$ as the set of zeroes of the equation

$$f(x, y) = 0$$

where $f(x, y)$ is a polynomial in the variables x and y:

$$f(x, y) = a_{0,0} + xa_{1,0} + ya_{0,1} + x^2a_{2,0} + xya_{1,1} + y^2a_{0,2} + \dots$$

$$\dots + x^d a_{d,0} + x^{d-1}ya_{d-1,1} + \dots + y^d a_{0,d}$$

Some, but not all, of the coefficients may be zero. The largest integer d such that not all $a_{d-i,i}$ are zero is the *degree* of the polynomial, and this is by definition the *degree of the curve*. But here we have a problem, best elucidated by an example.

The equation

$$y = 0$$

defines the x-axis. But so does the equation

$$y^2 = 0$$

at least as a *point-set*. But algebraically we need to distinguish between these two cases. The former equation defines the x-axis as a line, whereas the latter defines *a double line* along the x-axis: Informally speaking, it defines twice the x-axis.

The situation becomes even more difficult when we consider complicated polynomials. Thus for example we may consider the curve defined by the equation

$$(y^2 - x^3 - x^2)(y^2 - x^2) = 0.$$

When we are given the equation on this partly factored form, it is not difficult to see what we get: It is the nodal cubic curve displayed in Fig. 14.6 together with the two lines defined by $y = \pm x$. But suppose that we are given the following equation, on expanded form

$$3\,y^2x^4 - 3\,y^4x^2 + y^6 - x^7 + 2\,x^5y^2 - x^3y^4 - x^6 = 0$$

then it is not so easy to understand the situation. Using some PC-program to plot this curve, we should get the same picture as above. But this result is quite deceptive. Indeed, if we *factor* the left hand side of the equation, say again by some PC-program, we find that the equation becomes

$$(x^3 + x^2 - y^2)(x + y)^2(x - y)^2 = 0$$

which certainly defines the same point set, but reveals that this time the two lines occurring should *be counted with multiplicity 2*.

Recall that an *irreducible polynomial* in x and y is a polynomial $p(x, y)$ which may not be factored as a product of two polynomials, both non-constants. Thus for instance $p(x, y) = x^3 + x^2 - y^2$ is irreducible, as is $r(x, y) = x + y$ and $s(x, y) = x - y$. A special case of an important theorem is the following:

Theorem 22 (Unique factorization of polynomials). *Any polynomial in x and y with real (respectively complex,) coefficients, may be factored as a product of powers of irreducible polynomials with real (respectively complex), coefficients. These irreducible polynomials are unique except for possibly being proportional by constant factors.*

We make the following definition:

Definition 16 (The factorization in irreducible polynomials). The irreducible factorization of $f(x, y)$ is defined as an expression

$$f(x, y) = p_1(x, y)^{n_1} \cdots p_r(x, y)^{n_r}$$

where n_i are positive integers and all $p_i(x, y)$ are irreducible and no two are proportional by constant factor.

This factorization is unique up to constant factors, by the theorem.

Remark 4. Theorem 22 also holds for a polynomial in any number of variables, 1 up to any N. Definition 16 is also unchanged in the general case.

A polynomial may be irreducible as a polynomial with real coefficients, but *reducible* when considered as a polynomial with *complex* coefficients. This is the case for the polynomial

$$g(x, y) = x^2 + y^2,$$

which may not be factored as a polynomial with real coefficients, while

$$x^2 + y^2 = (x + iy)(x - iy)$$

The curve given by this polynomial has another interesting feature: As a curve in $\mathbb{R}^2$ it consists only of the origin, while it consists of two (complex) *lines* in $\mathbb{C}^2$, with equations $y = \pm ix$. They have only one real point on them, namely their point of intersection which is the origin. We would consider this as a degenerate case, say as a member of a family of circles, where the radius has shrunk to zero.

We shall mainly be concerned with *real curves* in this book, but as we see the picture may be quite different when we pass to the complex case. We are now ready to make the following definition:

Definition 17 (Real Affine Curve). A real affine plane curve C is the set of points $(a, b) \in \mathbb{R}^2$ which are zeroes of a polynomial $f(x, y)$ with real coefficients. The irreducible polynomials $p_i(x, y)$ occurring in the irreducible factorization of $f(x, y)$ referred to in Definition 16 define subsets C_i of C called the irreducible components of C. The exponent n_i of $p_i(x, y)$ in the factorization of $f(x, y)$ is called the multiplicity of the irreducible component.

In other words, C_i occurs with multiplicity n_i in C.

Remark 5. This definition suffices for our purpose in this book, but it should not be concealed that it does represent a simplification. Indeed, according to the definition the "real affine curve" defined by $x^2 + y^2 = 0$ is the same as the one defined by $x^2 + 2y^2 = 0$. For a variety of reasons this is undesirable. One solution is to simply *define* a curve in $\mathbb{R}^2$ as being an equivalence class of polynomials, two polynomials being regarded as equivalent if one is a non-zero constant multiple of the other. This is mathematically sound, but only applies to a special geometric situation, where one geometric object, here the curve, is contained in another geometric object of one dimension higher, here the plane, and is defined by one "equation". In any case this point of view belongs to a somewhat more advanced treatment of algebraic geometry than the one offered in the present book.

Another solution is to include all the complex points on a real curve. This would be a logical next step, having digested the present treatment of curves.

Finally we may consider more general "points", which takes us way beyond the scope of this book and into Grothendieck's theory of Schemes [22].

By a change of variables, which corresponds to a change of coordinate system,

$$\bar{x} = \alpha_{1,1}x + \alpha_{1,2}y$$

$$\bar{y} = \alpha_{2,1}x + \alpha_{2,2}y$$

the curve given by $f(x, y) = 0$ is expressed by the equation $\bar{f}(\bar{x}, \bar{y}) = 0$, where $\bar{f}(\bar{x}, \bar{y})$ is obtained by substituting the expressions obtained by solving for x and y,

$$x = \beta_{1,1}\bar{x} + \beta_{1,2}\bar{y}$$

$$y = \beta_{2,1}\bar{x} + \beta_{2,2}\bar{y}$$

into $f(x, y)$.

There are curves in the affine plane $\mathbb{R}^2$ which are not affine algebraic, but nevertheless form an important subject in geometry. We have encountered some of them: The Archimedean spiral and the quadratrix of Hippias. They both were invented to solve some of the Classical Problems. They are not defined by a polynomial equation. Some other simple examples are the curves defined by $y = \sin(x)$ or by $y = e^x$. This class of curves is called *the Transcendental Curves*. In this book we will confine the general theory to treating the algebraic curves, that is to say the ones defined by a polynomial equation.

14.3 Singularities and Multiplicities

We now return to some general concepts introduced in Sect. 13.7, where we needed it to understand the degeneracy of conic sections. Consider an algebraic affine curve K with equation

$$f(x, y) = 0.$$

Furthermore, let (a, b) be a point on the curve, i.e., $f(a, b) = 0$. We note that the following definition relies heavily on the *equation* of the curve, not just the curve as a subset of $\mathbb{R}^2$:

Definition 18. (a, b) is said to be a smooth, or a non-singular, point on K if

$$\left(\frac{\partial f}{\partial x}(a, b), \frac{\partial f}{\partial y}(a, b)\right) \neq (0, 0).$$

Otherwise (a, b) is called a singular point on K. A curve all of whose points are non-singular is referred to as a non-singular curve.

The vector $\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$ is referred to as the Jacobian vector (for short, the *Jacobian*) of the polynomial $f(x, y)$. Thus by definition a singular point is a point on the curve at which the Jacobian evaluates to the zero vector.

In Sect. 13.7 we saw that a non-degenerate conic section is a non-singular curve. We look at the situation in more detail by the examples below.

Examples (1) We first look at some simple *conic sections*, and start out with a circle of radius $R > 0$, which has the equation

$$x^2 + y^2 = R^2.$$

Here $f(x, y) = x^2 + y^2 - R^2$, and

$$\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = (2x, 2y) .$$

Evidently no point outside the origin can be a singular point of the circle, and as $R > 0$, every point on the circle is therefore smooth. We note that the same proof shows that an ellipse on standard form,

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

is smooth everywhere as well.

A (non-degenerate) hyperbola on standard form, which is given as

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1$$

similarly has Jacobian

$$\left(\frac{2}{a^2}x, -\frac{2}{b^2}y\right)$$

which also does not vanish outside the origin, showing that a hyperbola is smooth.

A degenerate hyperbola is one which has collapsed to the asymptotes, hence a curve with equation

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 0.$$

This curve has the same Jacobian as in the non-degenerate case, but now the origin actually lies on the curve, which therefore has the origin as its only singular point. Of course this degenerate hyperbola consists of two irreducible components which are lines intersecting at the origin, and that point is singular.

Our final conic section is the parabola with equation

$$ay - x^2 = 0$$

where $a \neq 0$. The Jacobian is $(-2x, a)$, so the only possibility of getting the zero vector at a point would be to have $x = 0$ and $a = 0$. For $a \neq 0$ we therefore have no singular points. If $a = 0$, then the equation yields the y-axis with multiplicity 2, and we see that then *all points on the curve are singular*.

(2) We next turn to the *nodal cubic curve* with equation

$$y^2 - x^3 - x^2 = 0$$

which is plotted in Fig. 14.6. The Jacobian is

$$(-3x^2 - 2x, 2y)$$

and thus (x, y) is a singular point if and only if the two additional equations below are satisfied:

$$-3x^2 - 2x = 0$$

$$2y = 0.$$

Thus $(x, y) = (0, 0)$ or $(x, y) = (-\frac{2}{3}, 0)$, and only the former lies on the curve, so the only singular point is $(0, 0)$.

(3) If $f(x, y)$ is any polynomial, then all points on the curve given by $f(x, y)^n = 0$ for n an integer greater than 1, will have all its points singular. This follows at once, since the Jacobian is

$$\left(nf(x, y)^{n-1}\frac{\partial f}{\partial x}, nf(x, y)^{n-1}\frac{\partial f}{\partial y}\right).$$

At this point it is highly recommended that the reader examines the curves plotted in Sect. 14.1, and determines their singular points.

14.4 Tangency

Now let (a, b) be a smooth point on the curve K. Then we find the equation for the tangent line at that point as follows. We first consider *the parametric form* for any line through (a, b) with direction given by the vector (u, v):

$$L = \left\{(x, y) \,\middle|\, \begin{array}{l} x = a + ut \\ y = b + vt \end{array} \text{ where } t \in \mathbb{R}\right\}.$$

This line will have the point (a, b) in common with K. We wish to determine other points of intersection. To do so we substitute the expressions for x and y in the parametric form for L into the equation for K, and get

$$f(a + ut, b + vt) = 0.$$

Expanding the left hand side in a Taylor series we obtain

$$\begin{aligned} f(a, b) &+ t\left(u\frac{\partial f}{\partial x}(a, b) + v\frac{\partial f}{\partial y}(a, b)\right) \\ &+ t^2\left(u^2\frac{\partial^2 f}{\partial x^2}(a, b) + 2uv\frac{\partial^2 f}{\partial x \partial y}(a, b) + v^2\frac{\partial^2 f}{\partial y^2}(a, b)\right) + \ldots = 0 \end{aligned}$$

which since $f(a, b) = 0$ gives

$$\begin{aligned} t &\left(u\frac{\partial f}{\partial x}(a, b) + v\frac{\partial f}{\partial y}(a, b)\right) \\ &+ t^2\left(u^2\frac{\partial^2 f}{\partial x^2}(a, b) + 2uv\frac{\partial^2 f}{\partial x \partial y}(a, b) + v^2\frac{\partial^2 f}{\partial y^2}(a, b)\right) + \ldots = 0 \end{aligned}$$

The points of intersection between the curve and the line are found by solving this equation for t. Of course we have $t = 0$ as one solution, and we see that this solution will occur with multiplicity 1 if and only if

$$u\frac{\partial f}{\partial x}(a,b) + v\frac{\partial f}{\partial y}(a,b) \neq 0.$$

Such values of u, v exist if and only if (a, b) is a smooth point on the curve. In that case there is exactly one line which *does not intersect the curve with multiplicity 1*, namely the line corresponding to u and v such that

$$u\frac{\partial f}{\partial x}(a,b) + v\frac{\partial f}{\partial y}(a,b) = 0.$$

By substituting

$$ut = x - a$$

$$vt = y - b$$

in this equation, we get *the equation for the tangent* to the curve at the point (a, b)

$$(x-a)\frac{\partial f}{\partial x}(a,b) + (y-b)\frac{\partial f}{\partial x}(a,b) = 0.$$

We next turn to the question of what happens at *a singular point*. So let $P = (a, b)$ be a singular point on the curve K. Since the situation is more complicated than in the case when P is smooth, we introduce new variables by

$$\overline{x} = x - a, \overline{y} = y - b$$

In other words, we shift the variables so that the new origin falls in P, $P = (0, 0)$. We then find a new polynomial g such that

$$f(x, y) = g(\overline{x}, \overline{y})$$

by substituting $x = \overline{x} + a$ and $y = \overline{y} + b$ into $f(x, y)$. The curve is also given by the equation

$$g(\overline{x}, \overline{y}) = 0.$$

Since the origin is a point on the curve given by $g(\overline{x}, \overline{y}) = 0$, it is clear that the polynomial $g(\overline{x}, \overline{y})$ has no constant term. We now collect the terms of $g(\overline{x}, \overline{y})$ which are of lowest total degree, and denote the sum of those terms by $h(\overline{x}, \overline{y})$.

Thus for example, if

$$g(\overline{x}, \overline{y}) = 2\overline{x}\overline{y}^2 - 5\overline{x}^2\overline{y} + 10\overline{x}^9\overline{y}^2 + 15\overline{x}^2\overline{y}^{12},$$

then

$$h(\overline{x}, \overline{y}) = 2\overline{x}\overline{y}^2 - 5\overline{x}^2\overline{y}.$$

This piece of a polynomial consisting of all terms of lowest total degree is called the *initial part* of the polynomial. Of course the sum of the degrees of the two variables $\overline{x}$ and $\overline{y}$ is the same for all terms occurring in $h(\overline{x}, \overline{y})$. If the point $P = (a, b)$ is smooth, then the Taylor expansion around the point (a, b) immediately shows that the polynomial $h(\overline{x}, \overline{y})$ is nothing but

$$\frac{\partial g}{\partial x}(0,0)\overline{x} + \frac{\partial g}{\partial y}(0,0)\overline{y} = \frac{\partial f}{\partial x}(a,b)(x-a) + \frac{\partial f}{\partial y}(a,b)(y-b).$$

Thus the concept introduced below generalizes the tangent at a smooth point, to a concept which applies to *singular points as well.*

With notations as above the polynomial $h(\overline{x}, \overline{y})$ defines a curve which is a finite union of lines through the point $(0, 0)$. In terms of x and y, the equation

$$h(x-a, y-b) = 0$$

defines a finite union of lines through $P = (a, b)$, some of them occurring with multiplicity > 1. Indeed, we have

$$h(\overline{x}, \overline{y}) = a_0\overline{x}^m + a_1\overline{x}^{m-1}\overline{y} + \cdots + a_i\overline{x}^{m-i}\overline{y}^i + \cdots + a_m\overline{y}^m$$

where not all a_i vanish. If (α_0, β_0) satisfies $h(\alpha_0, \beta_0) = 0$, then we also have $h(s\alpha_0, s\beta_0) = 0$ for all real numbers s, as one immediately verifies since all the monomials of h are of the same total degree m.

These lines are called the *lines of tangency* at the point $P = (a, b)$. If P happens to be smooth, then there is only one line, occurring with multiplicity 1.

Definition 19. The curve given by $h(x-a, y-b) = 0$ is referred to as *the (affine) tangent cone* of K at P.

Any line through $P = (a, b)$ may, as we have seen, be written on parametric form as

$$x - a = ut, y - b = vt$$

and its intersections with the curve is determined by the equation

$$f(a + ut, b + vt) = g(ut, vt) = 0.$$

The multiplicity of the root $t = 0$ in this equation is referred to as *the multiplicity of intersection* between the curve and the line at the point $P = (a, b)$.

A moments reflection will convince the reader of an important fact: *All lines through* $P = (a, b)$ *which do not coincide with one of the lines of tangency, intersect the curve with multiplicity equal to the number* m. This number m is of course only dependent upon the polynomial $f(x, y)$ and the point $P = (a, b)$.

In fact, we may assume that $P = (0, 0)$. An arbitrary line through $(0, 0)$ has the parametric form

$$L = \left\{ (x, y) \,\middle|\, \begin{matrix} x = ut \\ y = vt \end{matrix} \text{ where } t \in \mathbb{R} \right\}.$$

To find all points of intersection between this line and the curve K, we substitute the expressions for x and y into $f(x, y)$ and get

$$f(a + ut, b + vt) = 0.$$

This gives

$$h(ut, vt) + R(ut, vt) = 0.$$

where $R(x, y)$ denotes $f(x, y) - h(x, y)$. Thus the points of intersection are given by the roots of the equation

$$t^m (h(u, v) + t\varphi(t)) = 0.$$

One of the roots is $t = 0$, and this solution will occur with multiplicity $\geq m$, where equality holds if and only if

$$h(u, v) \neq 0.$$

thus if and only if L is not one of the lines of tangency.

We conclude with the

Definition 20 (Multiplicity of a point on a curve). The number m referred to above is called the multiplicity of the point P at K.

We thus have the observation

Proposition 16. *A point on an affine algebraic curve is smooth if and only if it has multiplicity 1.*

Exercises

Exercise 14.1 (a) Show that the parabola as well as the cubic parabola are non-singular curves, while the semi-cubic parabola has one singularity, namely at the origin. What is the multiplicity of this point?
(b) Show that all three curves have the x-axis as the only line of tangency at the origin, and find the multiplicity of this line of tangency for all three curves.
(c) Find rational parameterizations of these curves.

Exercise 14.2 (a) Explain the curve to the right in Fig. 14.3 by factoring the polynomial $P(x, y) = x^3 + x^2 y + x y^2 + y^3 + x^2 - y^2 = 0$ in its irreducible factors.
(b) Prove that the origin is a singular point of this curve. Is it the only singularity?

Give reason for your answer.
(c) Use MAPLE or a similar system to investigate the family of curves given by

$$tx^3 + x^2y + xy^2 + y^3 + x^2 - y^2 = 0$$

as t varies. On the basis of this, formulate a conjecture on how this family behaves, and try to prove it.

Exercise 14.3 Let the curve C be given by the equation

$$(x-1)(x-2)y^2 - 4x^2 = 0.$$

(a) Find the singular points of the curve, and their multiplicities.
(b) Find all the tangential lines at the singular points.

Exercise 14.4 Let the curve C be given by the equation

$$x^6 - x^2y^6 - y^{10} = 0.$$

(a) Find the singular points of the curve, and their multiplicities.
(b) Find all the tangential lines at the singular points.

Exercise 14.5 Find all the singularities and their multiplicities, as well as all tangential lines, to the following curves which have been treated in the text:

(a) The Folium of Descartes $x^3 + y^3 - 3xy = 0$.
(b) The Nodal Cubic $y^2 - x^3 - x^2 = 0$.
(c) The Trisectrix of Maclaurin $x^3 + xy^2 + y^2 - 3x^2 = 0$.
(d) The Clover Leaf $(x^2 + y^2)^2 + 3x^2y - y^3 = 0$.
(e) The Airplane Wing $x^4 + x^2y^2 - 2x^2y - xy^2 + y^2 = 0$.
(f) The Interlacing Ovals $2x^4 - 3x^2y + y^4 - 3y^3 + y^2 = 0$.
(g) The Four Clover Leaf $(x^2 + y^2)^4 + xy(x^2 - y^2) = 0$.

Find rational parameterizations of the curves in (a), (b) and (c).

Exercise 14.6 Find the singular points and their multiplicities as well as the tangential lines for the curves

(a) $y^3 - yx^2 - x^2 = 0$
(b) $x(x^2 + y^2) + x^2 - y^2 = 0$
(c) $x^4 + x^2y^2 - 2x^2y - x^2y + y^2 = 0$
(d) $xy^2 - y - x^3 = 0$

Find rational parameterizations of the curves (a) and (b).

Chapter 15
Higher Geometry in the Projective Plane

15.1 Projective Curves

We define curves in the *projective plane* $\mathbb{P}^2(\mathbb{R})$ analogously to curves in the affine plane $\mathbb{R}^2$. The difference is that we cannot use ordinary polynomials in two variables, but have to work with *homogeneous polynomials in three variables* instead. We have seen this in Sect. 13.8, for conics.

Thus the polynomial

$$X_0 + 5X_0X_1^2$$

is not homogeneous, since one monomial which occurs is X_0, and another is $5X_0X_2^2$. They are of degrees 1 and 3, respectively. On the other hand, the polynomial

$$X_0^3 + 5X_0X_1^2$$

is homogeneous, the two monomials which occur are both of degree 3.

Now assume that we have a homogeneous polynomial with real coefficients

$$F(X_0, X_1, X_2) = \Sigma_{I\in S} c_I X_0^{i_0} X_1^{i_1} X_2^{i_2}$$

where $I = (i_0, i_1, i_2)$, $d = i_0 + i_1 + i_2$, and the symbol $\Sigma_{I\in S}$ means that we have a sum where I runs through a finite subset S of triples of non-negative integers, when no confusion is possible we usually write just Σ_I. c_I is a real number, called the *coefficient* of the *monomial* $X_0^{i_0} X_1^{i_1} X_2^{i_2}$. Let $(a_0, a_1, a_2) \in \mathbb{R}^3$. Then we have

$$\begin{aligned} F(ta_0, ta_1, ta_2) &= \Sigma_I c_I (ta_0)^{i_0} (ta_1)^{i_1} (ta_2)^{i_2} \\ &= t^d (\Sigma_I c_I a_0^{i_0} a_1^{i_1} a_2^{i_2}) = t^d F(a_0, a_1, a_2) \end{aligned}$$

since $d = i_0 + i_1 + i_2$. Thus we find that whenever $t \neq 0$, then

$$F(ta_0, ta_1, ta_2) = 0 \text{ if and only if } F(a_0, a_1, a_2) = 0.$$

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_15,

It follows that the zero locus for a homogeneous polynomial in X_0, X_1 and X_2 is well defined in $\mathbb{P}^2(\mathbb{R})$. Moreover, we also note the

Theorem 23. *In the irreducible factorization of a homogeneous polynomial, given by Remark 4, all the irreducible polynomials occurring are also homogeneous.*

Proof. The proof is by induction on $d = \deg(F)$. For $d = 1$ the claim is immediate. Suppose that the claim is true for all homogeneous polynomials of degree $< d$, and let F be homogeneous of degree d. If F is irreducible, there is nothing to prove. Otherwise we may write

$$F = F_1 F_2$$

where F_1 and F_2 are polynomials of degrees $< d$. We may write

$$F_i = H_i + G_i, \text{ for } i = 1, 2$$

where H_i is the homogeneous piece of highest degree of F_i. Thus

$$F = H_1 H_2 + G_1 H_2 + G_2 H_1 + G_1 G_2 = H_1 H_2 + G$$

but since F and $H_1 H_2$ are homogeneous of the same degree, and G, if it were non zero would be of degree $< d$, it follows that

$$F = H_1 H_2$$

and the claim follows by induction. □

Definition 21 (Projective Algebraic Curve). A plane projective curve $C \subset \mathbb{P}^2(\mathbb{R})$ is the zero locus of a homogeneous polynomial F in X_0, X_1 and X_2, with coefficients from $\mathbb{R}$:

$$C = \left\{(a_0 : a_1 : a_2) \in \mathbb{P}^2(\mathbb{R}) \middle| \ F(a_0, a_1, a_2) = 0\right\}$$

The irreducible components of C, as well as their multiplicities, are defined analogously to the affine case by means of Theorem 23.

15.2 Projective Closure and Affine Restriction

Given an affine curve $K \subset \mathbb{R}^2$, with equation $f(x, y) = 0$. In the same way as we did for curves of degree 2, we may define the *projective closure* $C \subset \mathbb{P}^2(\mathbb{R})$ of K. It is defined by the equation $F(X_0, X_1, X_2) = 0$ where $F(X_0, X_1, X_2)$ is constructed by putting $x = \frac{X_1}{X_0}$ and $y = \frac{X_2}{X_0}$ and substituting this in $f(x, y)$, and writing the result as

$$f\left(\frac{X_1}{X_0}, \frac{X_2}{X_0}\right) = \frac{F(X_0, X_1, X_2)}{{X_0}^m}.$$

where X_0 does not divide the numerator. Here $F(X_0, X_1, X_2)$ is a homogeneous polynomial with real coefficients, uniquely determined by $f(x, y)$ as follows: If

$$f(x, y) = \sum_{I=(i_1,i_2)\in\Phi} a_I x^{i_1} y^{i_2}$$

where Φ denotes a finite set of tuples of non-negative integers (i_1, i_2), then the degree of K is $d = \max\{i_1 + i_2 | (i_1, i_2) \in \Phi\}$, and the projective closure is given by the equation

$$F(X_0, X_1, X_2) = \sum_{I=(i_1,i_2)\in\Phi} a_I X_0^{d-i_1-i_2} X_1^{i_1} X_2^{i_2} = 0.$$

d is the degree of the original affine curve K as well as of its projective closure C.

Definition 22. The homogeneous polynomial $F(X_0, X_1, X_2)$ as defined above is denoted by $f^h(X_0, X_1, X_2)$, and referred to as the homogenization of the (non-homogeneous) polynomial $f(x, y)$.

The key to understanding the relation between an affine curve and its projective closure lies in the simple and beautiful relation

$$f(a,b) = f^h(1, a, b)$$

which holds for all a and b.

Thus if K is the affine curve defined by $f(x, y) = 0$, then the projective closure C of K is defined by the equation $f^h(X_0, X_1, X_2) = 0$. Conversely, if we are given a projective curve C by the equation $F(X_0, X_1, X_2) = 0$, then we may define its *affine restriction* to $D_+(X_0)$ as identified with $\mathbb{R}^2$ as the curve given by the equation $F(1, x, y) = 0$. But this affine restriction is not always defined: Namely, if $F(X_0, X_1, X_2) = X_0^d$, then C is the line $L_\infty = V_+(X_0)$, the line at infinity, with multiplicity d. Of course the affine restriction of this curve to $D_0(X_0)$ is given by the equation $1 = 0$, so we might say that the *affine restriction of this curve to $D_+(X_0)$ is empty.* On the other hand, if we chose to take the affine restriction to $D_+(X_1)$ instead, and put $x = \frac{X_0}{X_1}$ and $y = \frac{X_2}{X_1}$, then the affine restriction is the curve given by $x^d = 0$, in other words the y-axis counted with multiplicity d.

So the concepts of projective closure and affine restriction are not independent of the coordinate system. The change to another projective coordinate system in $\mathbb{P}^2(\mathbb{R})$ has been described in Sect. 12.3. The equations defining the new coordinate system may also be used to define a bijective mapping of $\mathbb{P}^2(\mathbb{R})$ onto itself, known as *a projective transformation.* This was also explained in Sect. 12.3, and will not be repeated here.

Even though the concepts of projective closure and affine restriction do depend on the coordinate system, they are very useful in the investigation of properties and concepts which *are* coordinate independent. Normally we perform the projective closure by letting $V_+(X_0)$ contain the added points at infinity, and identify the affine plane $\mathbb{R}^2$ with $D_+(X_0)$. When an alternative procedure is used, this will be explicitly stated. Also, if $V_+(aX_0 + bX_1 + cX_2)$ is a projective line in $\mathbb{P}^2(\mathbb{R})$, then we may identify $D_+(aX_0 + bX_1 + cX_2)$ with $\mathbb{R}^2$ and carry out affine restrictions to $\mathbb{R}^2$ by restricting to $D_+(aX_0 + bX_1 + cX_2)$. Again, if this non-standard procedure is used we shall explicitly state so. The most convenient method is to choose a new projective coordinate system by putting

$$\overline{X}_0 = aX_0 + bX_1 + cX_2$$

and choosing linear forms $a_1X_0 + b_1X_1 + c_1X_2$ and $a_2X_0 + b_2X_1 + c_2X_2$ such that the determinant of the coefficients of the three forms is non-zero, so letting

$$\overline{X}_1 = a_1X_0 + b_1X_1 + c_1X_2 \text{ and } \overline{X}_2 = a_2X_0 + b_2X_1 + c_2X_2$$

we get a new projective coordinate system. Then the affine restriction is carried out in the standard fashion with respect to it.

Using an affine restriction we are able to study local properties of a curve, like questions of tangency or singularity, with greater precision. Taking projective closure we obtain information on how the curve behaves very far away from the origin, *at infinity*, information crucial to a global understanding of the affine curve itself. An example of this which we shall return to later is the determination of all the *asymptotes* of a curve in $\mathbb{R}^2$. We conclude this section on projective closure and affine restriction with the

Proposition 17. *(1) Let K be an affine curve in $\mathbb{R}^2$, and let C be its projective closure. Then the affine restriction of C is equal to K.*
(2) Let C be a projective curve, and let K be its affine restriction. If C is just a multiple of $V_+(X_0)$ then K is empty.[1] *Otherwise K is an affine curve, and its projective closure C' consist of all irreducible components of C, with the same multiplicity as before, except possibly for the component $V_+(X_0)$, which is removed when passing from C to C'.*

Proof. To prove (1), let K be given by

$$f(x, y) = \sum_{I=(i_1,i_2)\in\Phi} a_I x^{i_1} y^{i_2}.$$

Then the projective closure is given by

$$F(X_0, X_1, X_2) = \sum_{I=(i_1,i_2)\in\Phi} a_I X_0^{d-i_1-i_2} X_1^{i_1} X_2^{i_2} = 0.$$

[1] Or, as we shall say here, K does not exist.

Substituting $X_0 = 1$, $X_1 = x$ and $X_2 = y$ clearly gives us back $f(x, y)$, and (1) is proven.

As for (2), assume that C is given by the homogeneous polynomial

$$F(X_0, X_1, X_2) = X_0^r \left(\sum_{I=(i_1,i_2)\in\Phi} a_I X_0^{d-i_1-i_2} X_1^{i_1} X_2^{i_2} \right)$$

where the polynomial inside the parenthesis is not divisible by X_0. Denoting the latter by $G(X_0, X_1, X_2)$, we find that

$$F(1, x, y) = G(1, x, y)$$

and the affine restriction of C is defined by $G(1, x, y) = 0$. So the projective closure C' of the affine restriction is defined by $G(X_0, X_1, X_2)$. This completes the proof. □

15.3 Smooth and Singular Points on Affine and Projective Curves

Let C be given by the equation

$$F(X_0, X_1, X_2) = 0.$$

Moreover, let $P = (a_0 : a_1 : a_2)$ be a point on C.

Definition 23. We say that the point P is a smooth point on C if

$$\left(\frac{\partial F}{\partial X_0}(a_0, a_1, a_2), \frac{\partial F}{\partial X_1}(a_0, a_1, a_2), \frac{\partial F}{\partial X_2}(a_0, a_1, a_2) \right) \neq (0, 0, 0)$$

Whenever this condition is not satisfied, the point is referred to as a *singular* point. Correspondingly, a smooth point is also referred to as a *non-singular* point.[2]

Earlier we defined the term *smooth point* for affine curves $K \subset \mathbb{R}^2$. Even if this previous definition is similar to the one we have given here, we need to show that they do not contradict one another. Namely, when we form the *projective closure* of the affine curve K, we obtain a *projective curve* $C \subset \mathbb{P}^2(\mathbb{R})$. A point $p \in K$ should then be smooth as a point of the affine curve K if and only if it is smooth as a point on the projective curve C.

[2] In more advanced texts on algebraic geometry, the terms "smooth" and "non-singular" have slightly different meanings.

This problem is disposed of by means of the following proposition:

Proposition 18. *With notations as in as in Sect. 15.2 we have*

$$\left(\frac{\partial f}{\partial x}\right)^h (X_0, X_1, X_2) = \frac{\partial f^h}{\partial X_1}(X_0, X_1, X_2)$$

and

$$\left(\frac{\partial f}{\partial y}\right)^h (X_0, X_1, X_2) = \frac{\partial f^h}{\partial X_2}(X_0, X_1, X_2)$$

Proof. We put

$$f(x, y) = \sum_{I=(i_1,i_2)\in\Phi} a_I x^{i_1} y^{i_2}$$

then $F = f^h$ is given by

$$F(X_0, X_1, X_2) = \sum_{I=(i_1,i_2)\in\Phi} a_I X_0^{d-i_1-i_2} X_1^{i_1} X_2^{i_2} = 0.$$

The verification of the claim is immediate from this. □

Corollary 5. *Let K be an affine curve, and let C be the projective closure of K, where $V_+(X_0)$ is the points at infinity. Then (a, b) is a smooth point on the affine curve K if and only if $(1 : a : b)$ is a smooth point on the projective curve C.*

Proof. We apply the relation

$$g(a, b) = g^h(1, a, b)$$

to the partial derivatives. □

The second important observation concerning smooth or singular points is contained in the

Proposition 19. *The concept of smooth point on a projective curve is independent of the projective coordinate system.*

Proof. We may write the transition from one coordinate system to another as a matrix multiplication as follows:

$$\begin{Bmatrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{Bmatrix} \cdot \begin{Bmatrix} Y_0 \\ Y_1 \\ Y_2 \end{Bmatrix} = \begin{Bmatrix} X_0 \\ X_1 \\ X_2 \end{Bmatrix}$$

where the matrix has determinant $\neq 0$,

$$\begin{vmatrix} \alpha_{0,0} & \alpha_{0,1} & \alpha_{0,2} \\ \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} \end{vmatrix} \neq 0.$$

Clearly

$$\frac{\partial X_i}{\partial Y_j} = \alpha_{i,j}.$$

Moreover, if the curve C is given in the original coordinate system as

$$F(X_0, X_1, X_2) = 0$$

then it will be given in the new coordinate system by

$$G(Y_0, Y_1, Y_2) = 0$$

where

$$\begin{aligned} &G(Y_0, Y_1, Y_2) \\ &\quad = F(\alpha_{0,0}Y_0 + \alpha_{0,1}Y_1 + \alpha_{0,2}Y_2, \alpha_{1,0}Y_0 + \alpha_{1,1}Y_1 + \alpha_{1,2}Y_2, \alpha_{2,0}Y_0 + \alpha_{2,1}Y_1 \\ &\qquad + \alpha_{2,2}Y_2) \end{aligned}$$

Now let the point P be expressed as $(a_0 : a_1 : a_2)$ and $(b_0 : b_1 : b_2)$ in the two coordinate systems. Then we get by the chain rule

$$\begin{aligned} \frac{\partial G}{\partial Y_0}(b_0, b_1, b_2) &= \alpha_{0,0}\frac{\partial F}{\partial X_0}(a_0, a_1, a_2) + \alpha_{1,0}\frac{\partial F}{\partial X_1}(a_0, a_1, a_2) \\ &\quad + \alpha_{2,0}\frac{\partial F}{\partial X_2}(a_0, a_1, a_2) \\ \frac{\partial G}{\partial Y_1}(b_0, b_1, b_2) &= \alpha_{0,1}\frac{\partial F}{\partial X_0}(a_0, a_1, a_2) + \alpha_{1,1}\frac{\partial F}{\partial X_1}(a_0, a_1, a_2) \\ &\quad + \alpha_{2,1}\frac{\partial F}{\partial X_2}(a_0, a_1, a_2) \\ \frac{\partial G}{\partial Y_2}(b_0, b_1, b_2) &= \alpha_{0,2}\frac{\partial F}{\partial X_0}(a_0, a_1, a_2) + \alpha_{1,2}\frac{\partial F}{\partial X_1}(a_0, a_1, a_2) \\ &\quad + \alpha_{2,2}\frac{\partial F}{\partial X_2}(a_0, a_1, a_2) \end{aligned}$$

Since the determinant of the matrix of the α's is non-zero, it follows by *Cramer's Theorem* 5 in Sect. 6.7 that the vector of the evaluated partials to the left will not all vanish if and only if the vector of the evaluated partials to the right do not all vanish. Thus smoothness or singularity for a point on C is independent of the coordinate system in which the corresponding condition is expressed. □

15.4 The Tangent to a Projective Curve

Before we deduce the equation for *the tangent line to a projective curve,* we need to make some comments on *lines and other curves on parametric form* in $\mathbb{P}^2(\mathbb{R})$. We first consider the case of lines. A line $L \subset \mathbb{P}^2(\mathbb{R})$ which passes through the points $(a_0 : a_1 : a_2)$ and $(b_0 : b_1 : b_2)$ may be expressed as follows, on parametric form:

$$L = \left\{ (X_0 : X_1 : X_2) \;\middle|\; \begin{array}{l} X_0 = ua_0 + vb_0 \\ X_1 = ua_1 + vb_1 \\ X_2 = ua_2 + vb_2 \end{array} \right\}.$$

Here u and v are two real parameters which yield all the points on the line L, but as we see, it is only the ratio $(u : v)$ which distinguish between the points. In particular we have that $(u : v) = (1 : 0)$ yields the point $(a_0 : a_1 : a_2)$, while $(u : v) = (0 : 1)$ yields $(b_0 : b_1 : b_2)$.

More generally we may consider a curve in $\mathbb{P}^2(\mathbb{R})$ given on parametric form:

$$C = \left\{ (X_0 : X_1 : X_2) \;\middle|\; \begin{array}{l} X_0 = \xi_0(u, v) \\ X_1 = \xi_1(u, v) \\ X_2 = \xi_2(u, v) \end{array} \right\},$$

Here we assume that the polynomials $\xi_0(u, v)$, $\xi_1(u, v)$ and $\xi_2(u, v)$ are *homogeneous of the same degree* in the variables u and v. The class of curves which may be so described do not contain all projective curves in $\mathbb{P}^2(\mathbb{R})$, there are curves which are *not* parameterizable by polynomials. But it does include all lines in $\mathbb{P}^2(\mathbb{R})$.

If we choose

$$\xi_0(u, v) = u^2, \xi_1(u, v) = uv \text{ and } \xi_2(u, v) = v^2$$

we get a curve of degree 2, in other words *a projective conic section*: It has the equation

$$X_1^2 - X_0 X_2 = 0.$$

If we choose $\xi_0(u, v)$, $\xi_1(u, v)$ and $\xi_2(u, v)$ as general homogeneous polynomials of degree 2, then we get general projective curves of degree 2 in $\mathbb{P}^2(\mathbb{R})$: The class of projective curves of degree 2 consists of parameterizable ones. But even if we let $\xi_0(u, v)$, $\xi_1(u, v)$ and $\xi_2(u, v)$ be general homogeneous polynomials of degree 3, we only obtain a special class of degree 3 projective curves in $\mathbb{P}^2(\mathbb{R})$, namely the *rational cubics* in $\mathbb{P}^2(\mathbb{R})$.

In general, the curves are parameterizable as described above by homogeneous polynomials of the same degree d are referred to as *the rational degree d-curves in* $\mathbb{P}^2(\mathbb{R})$. Here an explanation should be interjected: This number d, the common degree of the polynomials $\xi_0(u, v)$, $\xi_1(u, v)$ and $\xi_2(u, v)$, turns out to be the degree of the equation

$$F(X_0, X_1, X_2) = 0$$

which expresses the relation between the polynomials $\xi_0(u,v)$, $\xi_1(u,v)$ and $\xi_2(u,v)$.

We now come to the concept of *tangent line of a general projective curve* in $\mathbb{P}^2(\mathbb{R})$. We consider a point $P = (a_0 : a_1 : a_2)$ on the curve C defined by the homogeneous polynomial $F(X_0, X_1, X_2)$,

$$F(X_0, X_1, X_2) = 0.$$

As we did in the affine case, we consider the collection of all lines passing through P, as we saw above these lines are all given on parametric form as

$$L = \left\{ (X_0 : X_1 : X_2) \,\middle|\, \begin{array}{l} X_0 = ua_0 + vb_0 \\ X_1 = ua_1 + vb_1 \\ X_2 = ua_2 + vb_2 \end{array} \right\}$$

where $P = (a_0 : a_1 : a_2)$ is the fixed point on C, and $Q = (b_0 : b_1 : b_2)$ is another point $\neq P$ in $\mathbb{P}^2(\mathbb{R})$ and u and v are the parameters describing the line L passing through P and Q, the points on L corresponding to the ratio $u : v$. The point P corresponds to $u : v = 1 : 0$, while Q corresponds to $u : v = 0 : 1$. We wish to examine the points of intersection of the line L with C, as well as the multiplicities with which they occur. We then have to find all u and v which satisfy the equation

$$F(ua_0 + vb_0, ua_1 + vb_1, ua_2 + vb_2) = 0.$$

But our objective now is *not* to find all the other points of intersection between L and C. Instead, we are interested in examining *how the line intersects the curve in the point P,* in other words we wish to study the solution $(u, v) = (1, 0)$ of the equation, and since only the ratios count, this amounts to studying the solution $t = 0$ of the equation

$$\varphi(t) = F(a_0 + tb_0, a_1 + tb_1, a_2 + tb_2) = 0.$$

Since $P \in C$, $t = 0$ certainly is a solution. As in the affine case the multiplicity of the solution $t = 0$ is referred to as the *multiplicity with which the line L intersects C at P.*

Expanding $\varphi(t)$ in a Taylor series around $t = 0$ we actually get a polynomial of degree d, the degree of the curve C. We get

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \frac{1}{2}\varphi''(0)t^2 + \cdots + \frac{1}{i!}\varphi^{(i)}(0)t^i + \cdots + \frac{1}{d!}\varphi^{(d)}(0)t^d.$$

Here $\varphi(0) = 0$, and using the general Chain Rule we obtain

$$\varphi'(t) = b_0 \frac{\partial}{\partial X_0} F(a_0 + tb_0, a_1 + tb_1, a_2 + tb_2)$$

$$
\begin{aligned}
&+ b_1 \frac{\partial}{\partial X_1} F(a_0 + tb_0, a_1 + tb_1, a_2 + tb_2)\\
&+ b_2 \frac{\partial}{\partial X_2} F(a_0 + tb_0, a_1 + tb_1, a_2 + tb_2)\\
= &\left(\left(b_0 \frac{\partial}{\partial X_1} + b_1 \frac{\partial}{\partial X_1} + b_2 \frac{\partial}{\partial X_2}\right) F\right)(a_0 + tb_0, a_1 + tb_1, a_2 + tb_2)
\end{aligned}
$$

and hence

$$\varphi'(0) = \left(\left(b_0 \frac{\partial}{\partial X_1} + b_1 \frac{\partial}{\partial X_1} + b_2 \frac{\partial}{\partial X_2}\right) F\right)(a_0, a_1, a_2).$$

Taking the derivative of $\varphi'(t)$ and using the Chain Rule again, we similarly get the expression

$$\varphi''(0) = \left(\left(b_0 \frac{\partial}{\partial X_0} + b_1 \frac{\partial}{\partial X_1} + b_2 \frac{\partial}{\partial X_2}\right)^2 F\right)(a_0, a_1, a_2)$$

The expression

$$\left(b_0 \frac{\partial}{\partial X_0} + b_1 \frac{\partial}{\partial X_1} + b_2 \frac{\partial}{\partial X_2}\right)^2 F$$

is short term for the more elaborate

$$
\begin{aligned}
&b_0^2 \frac{\partial^2 F}{\partial X_0^2} + b_1^2 \frac{\partial^2 F}{\partial X_1^2} + b_2^2 \frac{\partial^2 F}{\partial X_2^2}\\
&\quad + 2b_0 b_1 \frac{\partial^2 F}{\partial X_0 \partial X_1} + 2b_0 b_2 \frac{\partial^2 F}{\partial X_0 \partial X_2} + 2b_1 b_2 \frac{\partial^2 F}{\partial X_1 \partial X_2}
\end{aligned}
$$

The point is that we have the general formula

$$\varphi^{(m)}(0) = \left(\left(b_0 \frac{\partial}{\partial X_0} + b_1 \frac{\partial}{\partial X_1} + b_2 \frac{\partial}{\partial X_2}\right)^i F\right)(a_0, a_1, a_2)$$

where the expression

$$\left(b_0 \frac{\partial}{\partial X_0} + b_1 \frac{\partial}{\partial X_1} + b_2 \frac{\partial}{\partial X_2}\right)^m F$$

has a similar meaning as in the case $i = 2$: We multiply out the polynomial in D_0,D_1 and D_2,

$$\Delta_m(D_0, D_1, D_2) = (b_0 D_0 + b_1 D_1 + b_2 D_2)^m$$

and then replace the monomials

$$D_0^{j_0} D_1^{j_1} D_2^{j_2}$$

by

$$\frac{\partial^{j_0+j_1+j_2} F}{\partial X_0^{j_0} \partial X_1^{j_1} \partial X_2^{j_2}}.$$

It is a reasonably straightforward exercise to prove this by induction on the exponent i. We thus have the following formula:

$$\begin{aligned}\varphi(t) &= F(a_0, a_1, a_2) + (D_{(b_0,b_1,b_2)} F)(a_0, a_1, a_2)t \\ &\quad + \frac{1}{2}(D^2_{(b_0,b_1,b_2)} F)(a_0, a_1, a_2)t^2 + \cdots + \frac{1}{i!}(D^i_{(b_0,b_1,b_2)} F)(a_0, a_1, a_2)(0)t^i \\ &\quad + \cdots + \frac{1}{d!}(D^d_{(b_0,b_1,b_2)} F)(a_0, a_1, a_2)(0)t^d\end{aligned}$$

where

$$D_{(b_0,b_1,b_2)} = b_0 \frac{\partial}{\partial X_0} + b_1 \frac{\partial}{\partial X_1} + b_2 \frac{\partial}{\partial X_2}$$

Actually we can give a precise formula for $(D^m_{(b_0,b_1,b_2)} F)(a_0, a_1, a_2)$. In fact, there is a generalization of the familiar *binomial formula*

$$(D_0 + D_1)^m = \sum \frac{m!}{i_0! i_1!} D_0^{i_0} D_1^{i_1}$$

where the sum runs over all non-negative i_0, i_1 such that $i_0 + i_1 = m$, to the case of any number of indeterminates $D_0, \ldots, D_r$. Indeed, we have the formula

$$(D_0 + D_1 + \cdots + D_r)^m = \sum \frac{m!}{i_0! i_1! \cdots i_r!} D_0^{i_0} D_1^{i_1} \cdots D_r^{i_r}$$

where the sum runs over all non-negative $i_0, i_1, \ldots, i_r$ such that $i_0 + \cdots + i_r = m$. We may prove this formula by induction by first noting that it holds for $m = 0$ or 1. Then assuming it for $m - 1$ we need only verify the multiplication

$$\begin{aligned}&\left(\sum_{i_1+\cdots+i_r=m-1} \frac{(m-1)!}{i_0! i_1! \cdots i_r!} D_0^{i_0} D_1^{i_1} \cdots D_r^{i_r} \right) (D_0 + D_1 + \ldots D_r) \\ &= \sum_{i_1+\cdots+i_r=m} \frac{m!}{i_0! i_1! \cdots i_r!} D_0^{i_0} D_1^{i_1} \cdots D_r^{i_r}\end{aligned}$$

which we leave to the reader.

Using this *Multinomial Formula* we obtain the important identity, valid for any number of variables but stated here only for three:

$$\begin{aligned}&(D^m_{(b_0,b_1,b_2)}F)(a_0,a_1,a_2)\\&\quad=\sum b_0^{i_0}b_1^{i_1}b_2^{i_2}\frac{m!}{i_0!i_1!i_2!}\left(\frac{\partial^m}{\partial X_0^{i_0}\partial X_1^{i_1}\partial X_2^{i_2}}F\right)(a_0,a_1,a_2)\end{aligned}$$

where the sum runs over all non-negative i_0, i_1, i_2 such that $i_0 + i_1 + i_2 = m$.

We have $F(a_0, a_1, a_2) = 0$, so the constant term of $\varphi(y)$ is zero for all choices of (b_0, b_1, b_2). It may happen that the coefficient of t vanishes as well, for all choices of (b_0, b_1, b_2), and so on, up to a certain t^m. We make the following definition:

Definition 24 (Multiplicity of Points on Projective Curves). The point $P = (a_0 : a_1 : a_2)$ on the projective curve C given by $F(X_0, X_1, X_2) = 0$ is said to be of multiplicity m if for all $n < m$ and all i_0, i_1, i_2

$$\left(\frac{\partial^n F}{\partial X_0^{i_0}\partial X_1^{i_1}\partial X_2^{i_2}}\right)(a_0,a_1,a_2)=0,$$

while for at least one choice of i_0, i_1, i_2

$$\left(\frac{\partial^m F}{\partial X_0^{i_0}\partial X_1^{i_1}\partial X_2^{i_2}}\right)(a_0,a_1,a_2)\neq 0.$$

This definition is independent of the projective coordinate system, the proof is straightforward but a little complicated. We omit it here. Clearly we have the following result:

Proposition 20. *The point $P = (a_0 : a_1 : a_2)$ on the projective curve C is of multiplicity 1 if and only if it is smooth.*

We also note the following:

Proposition 21. *The point $P = (a_0 : a_1 : a_2)$ on the projective curve C given by $F(X_0, X_1, X_2) = 0$ is of multiplicity m if and only if for all $n < m$ and all $(b_0, b_1, b_2) \notin C$*

$$\sum_{i_0+i_1+i_2=n} b_0^{i_0}b_1^{i_1}b_2^{i_2}\frac{n!}{i_0!i_1!i_2!}\left(\frac{\partial^n F}{\partial X_0^{i_0}\partial X_1^{i_1}\partial X_2^{i_2}}\right)(a_0,a_1,a_2)=0.$$

while for at least one tuple $(b_0, b_1, b_2) \notin C$

$$\sum_{i_0+i_1+i_2=m} b_0^{i_0}b_1^{i_1}b_2^{i_2}\frac{m!}{i_0!i_1!i_2!}\left(\frac{\partial^m F}{\partial X_0^{i_0}\partial X_1^{i_1}\partial X_2^{i_2}}\right)(a_0,a_1,a_2)\neq 0.$$

This last proposition shows that any line L through the point P will intersect C at P with multiplicity at least equal to m, the multiplicity of the point P on C. And there exists at least one line through P which intersects C at P with multiplicity m.

We need the following result:

Proposition 22. *Let $F(X_0, X_1, X_2)$ be a homogeneous polynomial of degree d. Then the following identity holds:*

$$X_0 \frac{\partial F}{\partial X_0} + X_1 \frac{\partial F}{\partial X_1} + X_2 \frac{\partial F}{\partial X_2} = dF(X_0, X_1, X_2)$$

Proof. We consider the following identity which holds for the variables $X_0, \ldots, X_n$ and t

$$F(tX_0, tX_1, \ldots, tX_n) = t^d F(X_0, X_1, \ldots, X_n).$$

This identity is verified by substituting tX_i for X_i in the polynomial, and observing that by definition all the monomials which occur are of the same degree d.

We now compute the derivative with respect to t. The right hand side yields

$$dt^{d-1} F(X_0, X_1, \ldots, X_n)$$

while the chain rule applied to the left hand side yields

$$\frac{\partial F}{\partial X_0}(tX_0, tX_1, tX_2)X_0 + \frac{\partial F}{\partial X_1}(tX_0, tX_1, tX_2)X_1 + \frac{\partial F}{\partial X_2}(tX_0, tX_1, tX_2)X_2.$$

These two polynomials are equal, and putting $t = 1$ we get the claimed formula.

□

Now assume that $P = (a_0 : a_1 : a_2) \in C$ has multiplicity m on C. We define the curve $T_{C,P}$ by the equation

$$\begin{aligned} &H_{C,P}(X_0, X_1, X_2) \\ &= \sum_{i_0+i_1+i_2=m} X_0^{i_0} X_1^{i_1} X_2^{i_2} \frac{m!}{i_0! i_1! i_2!} \left(\frac{\partial^m F}{\partial X_0^{i_0} \partial X_1^{i_1} \partial X_2^{i_2}} \right) (a_0, a_1, a_2) = 0. \end{aligned}$$

This equation actually does define a curve of degree m in $\mathbb{P}^2(\mathbb{R})$, as it is non-zero and homogeneous of degree m. We have the following important result:

Theorem 24. *$T_{C,P}$ is the union of a finite number of lines through P. The irreducible components of $T_{C,P}$ are exactly the lines through P which intersect C with multiplicity $> m$.*

Proof. We start out by noticing that $P \in T_{C,P}$. Indeed, we have the identity

$$\begin{aligned} &d(d-1)\ldots(d-m+1)F(X_0, X_1, X_2) \\ &= \sum_{i_0+i_1+i_2=m} X_0^{i_0} X_1^{i_1} X_2^{i_2} \frac{m!}{i_0! i_1! i_2!} \left(\frac{\partial^m F}{\partial X_0^{i_0} \partial X_1^{i_1} \partial X_2^{i_2}} \right) = 0 \end{aligned}$$

as is easily seen by repeated application of Proposition 22, first to F, then to the first order partial derivatives, so to the second order ones and so on, up to the partials of order m.

Since by definition of the multiplicity m there exists a point $(b_0 : b_1 : b_2)$ such that $H_{C,P}(b_0, b_1, b_2) \neq 0$, $T_{C,P}$ is a curve. Moreover, if $H_{C,P}(b_0, b_1, b_2) = 0$ then the line through P and $Q = (b_0, b_1, b_2)$ intersects C at P with multiplicity $> m$, since the corresponding $\varphi(t)$ has $t = 0$ as a root occurring with multiplicity $> m$. Thus any point $Q' = (b_0', b_1', b_2')$ on that line will also satisfy $H_{C,P}(b_0', b_1', b_2') = 0$. Thus the curve $T_{C,P}$ consists of lines passing through P. □

Definition 25. The curve $T_{C,P}$ is referred to as the projective tangent cone to C at P.

We have earlier defined the concept of multiplicity and tangent cone in the affine case. These concepts are completely compatible under affine restriction and under projective closure:

Proposition 23. *Let K be an affine curve, and $p = (a, b) \in K$. Let C be the projective closure, and put $P = (1 : a : b) \in C$. Then the point p, as a point on an affine curve, is of the same multiplicity as the point P on the projective curve C. Moreover the affine restriction of $T_{C,P}$ is equal to the affine tangent cone of K at p, as given in Definition 19.*

Proof. We may change the projective coordinate system to one in which $P = (1 : 0 : 0)$, this corresponds to a change of affine coordinate system to one in which $(a, b) = (0, 0)$. In this case the claim is easily checked. □

The irreducible components of the curve $T_{C,P}$ are referred to as the *lines of tangency* to C at P. If the point $P \in C$ is smooth, then $m = 1$ and there is only one line of tangency, which we refer to as the *tangent line to C at P,* and denote as before by $T_{C,P}$. The equation is

$$X_0 \frac{\partial F}{\partial X_0}(a_0, a_1, a_2) + X_1 \frac{\partial F}{\partial X_1}(a_0, a_1, a_2) + X_2 \frac{\partial F}{\partial X_2}(a_0, a_1, a_2) = 0.$$

We finally compute an example. Consider the projective curve given by

$$F(X_0, X_1, X_2) = X_0 X_2^2 - X_1^3 - X_0 X_1^2 = 0$$

which is the projective closure of the affine curve defined by

$$y^2 - x^3 - x^2 = 0,$$

in other words, the nodal cubic curve. To find the projective tangent cone at the point $(1 : 0 : 0)$, it is most convenient to pass to the affine restriction. Then we immediately see that the affine tangent cone is defined by

$$y^2 - x^2 = 0,$$

which has the projective closure given by

$$X_2^2 - X_1^2 = 0.$$

This is the fastest way to proceed. But we could also use the definition directly. Then we compute

$$\frac{\partial F}{\partial X_0} = X_2^2 - X_1^2, \frac{\partial F}{\partial X_1} = 3X_1^2 - 2X_0 X_1, \frac{\partial F}{\partial X_2} = 2X_0 X_2$$

They all evaluate to zero at $(1 : 0 : 0)$, so this point is singular (as we know from the affine restriction). We differentiate again, and obtain

$$\frac{\partial^2 F}{\partial X_0^2} = 0, \frac{\partial^2 F}{\partial X_1^2} = 6X_1 - 2X_0, \frac{\partial^2 F}{\partial X_2^2} = 2X_0$$
$$\frac{\partial^2 F}{\partial X_0 \partial X_1} = -2X_1, \frac{\partial^2 F}{\partial X_0 \partial X_2} = 2X_2, \frac{\partial^2 F}{\partial X_1 \partial X_2} = 0$$

Evaluating at P, we get

$$\frac{\partial^2 F}{\partial X_0^2} = 0, \frac{\partial^2 F}{\partial X_1^2} = -2, \frac{\partial^2 F}{\partial X_2^2} = 2$$
$$\frac{\partial^2 F}{\partial X_0 \partial X_1} = 0, \frac{\partial^2 F}{\partial X_0 \partial X_2} = 0, \frac{\partial^2 F}{\partial X_1 \partial X_2} = 0$$

and thus according to our formula the equation for $T_{C,P}$ is

$$\frac{2!}{0!2!0!}(-2)X_1^2 + \frac{2!}{0!0!2!}2X_2^2 = 0,$$

confirming what we found above.

15.5 Projective Equivalence

We have now arrived at a point where it is appropriate to attempt using the techniques introduced to create some system, or order, in the vast *menagerie* of different algebraic curves, affine or projective, which exist in $\mathbb{R}^2$ and in $\mathbb{P}^2(\mathbb{R})$.

For this we introduce the notion of *projectively equivalent plane curves*. We make the following

Definition 26. Two irreducible projective curves C and C' are called projectively equivalent if C is mapped to C' by a projective transformation. Two irreducible affine curves K and K' are said to be projectively equivalent if their projective closures are projectively equivalent, and an affine curve K is projectively equivalent to its projective closure C.

Remark Frequently we replace the sentence C *is mapped to* C' *by a projective transformation* by C *becomes* C' *by a projective change of coordinate system.* This way of expressing the equivalence is perfectly legitimate when we think of the curves as given by explicit equations.

Thus for example the affine version of the nodal cubic

$$y^2 - x^3 - x^2 = 0.$$

is projectively equivalent to the projective version defined by

$$X_0X_2^2 - X_1^3 - X_0X_1^2 = 0.$$

However, note that *we do not assert* that its affine tangent cone at the origin

$$y^2 - x^2 = 0$$

is projectively equivalent to the projective tangent cone at $(1 : 0 : 0)$,

$$X_2^2 - X_1^2 = 0$$

as we use this terminology for irreducible curves only.

We first study non-degenerate conic sections in light of the above definition. In Theorem 20 we proved the following:

The equation

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$$

yields a non-degenerate conic section if and only if

$$\begin{vmatrix} A & B & D \\ B & C & E \\ D & E & F \end{vmatrix} \neq 0.$$

A non-degenerate conic section is irreducible, as it would otherwise be two lines or one line with multiplicity 2, both cases being degenerate conics. The projective closure is given by the equation

$$AX_1^2 + 2BX_1X_2 + CX_2^2 + 2DX_0X_1 + 2EX_0X_2 + FX_0^2 = 0.$$

This equation, with the condition that the above determinant be non-zero, yields a general non-degenerate conic section: Any non-degenerate conic section may be expressed in this way. We have carried out an investigation of conic sections in $\mathbb{P}^2(\mathbb{R})$ in Sect. 13.8. Now we are going to make a suitable choice of projective coordinate system which gives this general equation a very simple form. For this we use Corollary 3. We proceed as follows:

First of all, since the conic section is assumed non-degenerate, any four points Q_1, Q_2 Q_3 and Q_4 on it form *an arc of four*: No three points among them are collinear. By the corollary referred to above there is a projective transformation G of $\mathbb{P}^2(\mathbb{R})$ onto itself mapping Q_1 to $(1 : 0 : 0)$, Q_2 to $(1 : 1 : 0)$, Q_3 to $(1 : 0 : 1)$ and Q_4 to $(1 : 1 : 1)$. Thus after a projective transformation we may assume that these four points lie on the projective conic section C. Taking the affine restriction of the situation, we get the points $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$ in $\mathbb{R}^2$, they lie on the affine conic section given by

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0.$$

Thus we obtain

$$F = 0, A + 2D = 0, C + 2E = 0 \text{ and } A + 2B + C + 2D + 2E = 0.$$

Thus $B = 0$ and the equation becomes

$$Ax^2 - Ax + Cy^2 - Cy = 0.$$

So completing the square we obtain the equation as

$$\lambda_1 x^2 + \lambda_2 y^2 + \lambda_3 = 0$$

and going back to $\mathbb{P}^2(\mathbb{R})$ by the projective closure we get

$$\lambda_3 X_0^2 + \lambda_1 X_1^2 + \lambda_2 X_2^2 = 0.$$

Since the conic section is non-degenerate, λ_1, λ_2 and λ_3 are all non-zero. Since we are dealing with real theory, we also may exclude the case that λ_1, λ_2 and λ_3 have the same sign. Thus, again after a projective transformation if needed, we may assume that $\lambda_3 < 0$, and $\lambda_1, \lambda_2 > 0$. Dividing by $|\lambda_3|$ the equation becomes

$$\mu_1 X_1^2 + \mu_2 X_2^2 - X_0^2 = 0$$

which after the projective transformation

$$(a_0 : a_1 : a_2) \mapsto (a_0 : \sqrt{\mu_1} a_1 : \sqrt{\mu_2} a_2)$$

becomes

$$X_1^2 + X_2^2 - X_0^2 = 0.$$

We have completed the proof of the following:

Theorem 25. *Up to projective equivalence there is only one non-degenerate conic section in* $\mathbb{P}^2(\mathbb{R})$.

There is a very close connection between *projective transformations* on one hand and actual *projections* on the other.

In fact, it is fair to say that these two concepts are practically equivalent. We shall now explain this.

We identify the affine plane $\mathbb{R}^2$ with the plane $z = 1$, as shown in Fig. 15.1. Then, as was explained in Sect. 9.3, the line through the origin O and the point $(x, y, 1)$ identifies the point (x, y) with the point $(1 : x : y)$ which corresponds to it under the identification of $D_+(X_0)$ with $\mathbb{R}^2$, and the lines contained in the xy-plane then correspond to the points at infinity, in other words to the points in $V_+(X_0)$.

The projection mapping with center at the origin O from the plane $z = 1$ to a general plane pl, subject only to the condition that $O \notin pl$, may be described as follows: A point $(x, y) \in \mathbb{R}^2$ is identified with $p = (x, y, 1)$, and the line ℓ through O and p is produced to intersect the plane pl in the point q. This point q is then the projection of p from the center O to the plane pl. To describe this mapping in terms of coordinates, we need to find a suitable description of the plane pl. Any plane in $\mathbb{R}^3$, the xyz-space, which does not pass through the origin O is uniquely determined by three linearly independent vectors,

$$\begin{aligned} a_0 &= (a_0, b_0, c_0) \\ a_1 &= (a_1, b_1, c_1) \\ a_2 &= (a_2, b_2, c_2). \end{aligned}$$

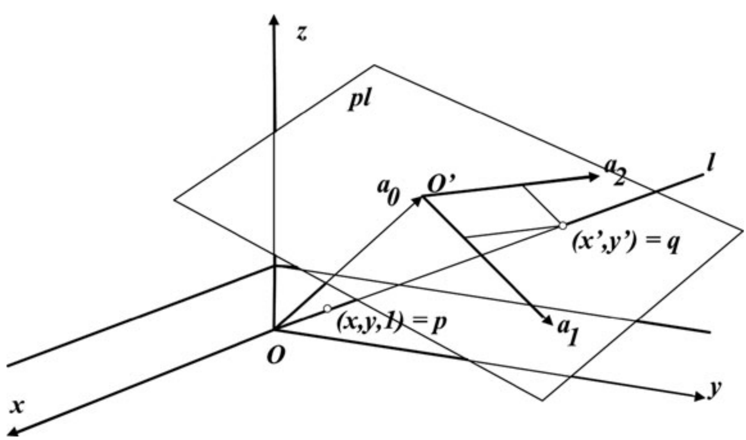


Fig. 15.1 A projective transformation from the plane $z = 1$ to the plane defined by the three vectors α_0, α_1 and α_2 as explained in the text

The first vector a_0 determines a point O' in pl, which we will use as the origin in a coordinate system to be introduced in pl, see Fig. 15.1. Further, to specify the plane we need the two vectors a_1 and a_2, which together with a_0 form a basis for the vector space $\mathbb{R}^3$. On (double) parametric form the plane pl is then given as

$$pl = \left\{ (X,Y,Z) \in \mathbb{R}^3 \;\middle|\; \begin{array}{l} X = a_0 + ua_1 + va_2 \\ Y = b_0 + ub_1 + vb_2 \\ Z = c_0 + uc_1 + vc_2 \end{array} \text{ where } (u,v) \in \mathbb{R}^2 \right\}$$

With this description of pl we let the origin O' of a coordinate system in pl be the end point of a_0, in other words the point (a_0, b_0, c_0). The vector a_1 is affixed to O', and determines the x'-axis, similarly a_2 determines the y'-axis. Their lengths are taken to be unit lengths along the respective axes, thus the point q is represented by the coordinates (x', y') which are the values of u and v defining q in the parametric form of pl.

On the other hand the parametric form of ℓ is

$$\ell = \left\{ (X,Y,Z) \in \mathbb{R}^3 \;\middle|\; \begin{array}{l} X = xt \\ Y = yt \\ Z = t \end{array} \text{ where } t \in \mathbb{R} \right\}.$$

Thus we have

$$\begin{aligned} tx &= a_0 + x'a_1 + y'a_2 \\ ty &= b_0 + x'b_1 + y'b_2 \\ t &= c_0 + x'c_1 + y'c_2 \end{aligned}$$

Thus if we identify (x, y) with $(1 : x : y) = (x_0 : x_1 : x_2)$ where, as usual, $z = x_0$, and similarly identify (x', y') with $(1 : x' : y') = (\overline{x}_0 : \overline{x}_1 : \overline{x}_2)$, then we get

$$\begin{aligned} x_0 &= c_0\overline{x}_0 + c_1\overline{x}_1 + c_2\overline{x}_2 \\ x_1 &= a_0\overline{x}_0 + a_1\overline{x}_1 + a_2\overline{x}_2 \\ x_2 &= b_0\overline{x}_0 + b_1\overline{x}_1 + b_2\overline{x}_0 \end{aligned}$$

from which we conclude that the projection is really a projective transformation.

We now see that Theorem 25 immediately implies the following result, which shows that Conic Sections as defined by Apollonius are precisely the curves in $\mathbb{R}^2$ of degree 2:

Corollary 6. *The (non degenerate) curves of degree two in $\mathbb{R}^2$ are precisely those which can be obtained as the intersection between a fixed circular cone and a varying plane.*

Proof. By the proof of Theorem 25 any non degenerate conic C is a projection from the center $(0, 0, 0)$ onto some plane, of the circle $x^2 + y^2 = 1, z = 1$ in the plane $z = 1$. Thus C is the intersection between that plane and the cone generated by the circle with vertex at the origin. Conversely any such intersection curve is projectively equivalent to the circle and is therefore a conic.

We conclude this section with an examination of two classes of curves studied in Sect. 14.1. A semi-cubic parabola may, by a change of affine coordinate system, be brought on the form

$$x^3 - y^2 = 0.$$

The usual projective closure of this curve is given by

$$X_0 X_2^2 - X_1^3 = 0$$

in $\mathbb{P}^2(\mathbb{R})$. Taking the affine restriction to $D_+(x_2)$ and letting $\overline{y} = \frac{X_0}{X_2}$, $\overline{x} = \frac{X_1}{X_2}$, we get the equation

$$\overline{y} - \overline{x}^3 = 0$$

which is a cubic parabola. Thus cubic and semi-cubic parabolas are projectively equivalent.

15.6 Asymptotes

We may now give a simple treatment of a subject which often appears rather mysterious. An *asymptote* to a given curve is defined as a line such that the distance from a point on the curve to the line tends to zero as the point on the curve moves further and further away from the origin.

This definition renders it quite mysterious how to actually compute all asymptotes to a given curve. Another drawback is that it defines the concept in terms of *distance*, thus the concept defined in this way is not an algebraic one.

The following definition is equivalent to the one given above for algebraic affine curves in $\mathbb{R}^2$:

Definition 27. Let K be the affine curve defined by

$$f(x, y) = 0.$$

Let C be the projective closure in $\mathbb{P}^2(\mathbb{R})$ obtained by letting $x = \frac{X_1}{X_0}$, $y = \frac{X_2}{X_0}$ as usual. Let $P_1, \dots, P_m$ be the points at infinity of C, and let $L_1, \dots, L_r$ be all lines in $\mathbb{P}^2(\mathbb{R})$ different from $V_+(X_0)$ and appearing as a line of tangency to C at one of the points $P_1, \dots, P_m$. Let $\ell_1, \dots, \ell_r$ be the affine restrictions of $L_1, \dots, L_r$. Then $\ell_1, \dots, \ell_r$ are all the asymptotes of K in $\mathbb{R}^2$.

The Trisectrix of Maclaurin has equation $x^3 + xy^2 + y^2 - 3x^2 = 0$. It is treated in Sect. 14.1, and its appearance makes one wonder if it might have a vertical asymptote, crossing the x-axis somewhere to the left of the origin. We shall now check this.

The projective closure of the trisectrix is given by the equation $X_1^3 + X_1X_2^2 + X_0X_2^2 - 3X_0X_1^2 = 0$. We find the points at infinity by substituting $X_0 = 0$ into this equation, we get $X_1^3 + X_1X_2^2 = 0$. This yields one real point, given by $X_1 = 0$, and two complex points determined by $X_1^2 + X_2^2 = 0$, which do not concern us as we are dealing with the real points only. Thus the one (real) point at infinity is $(0 : 0 : 1)$. We now take the affine restriction to $D_+(X_2)$ by putting $x' = \frac{X_0}{X_2}$ and $y' = \frac{X_1}{X_2}$. This affine restriction is given by $y'^3 + y' + x' - 3x'y'^2 = 0$. Hence the origin is a smooth point, the tangent there is given by $x' + y' = 0$. Going back to the projective plane, this line has the equation $X_0 + X_1 = 0$, and taking the affine restriction to the original affine xy-plane, we get the equation $x = -1$: This, then, is the asymptote of the curve, affirming our suspicion that such a line might exist.

An even simpler example, but an important one, is to verify the asymptotes of a general hyperbola. Assume it is given on standard form, as

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

To show is that the asymptotes are given by

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 0.$$

We leave this verification as an exercise.

15.7 General Conchoids

In Sect. 4.6 we explained the construction of the Conchoid of Nicomedes. Now we treat his great invention in more detail. First we deduce the equation of the conchoid. We choose the coordinate system with origin at the fixed point P, y-axis parallel to ℓ and x-axis normal to ℓ as shown in Fig. 15.2. A line through P has the equation $y = tx$, and a point (x, y) on it, at distance b from its intersection with ℓ must satisfy

$$y = tx$$
$$(x - a)^2 + (y - ta)^2 = b^2$$

which when $t = \frac{y}{x}$ from the former is substituted in the latter, yields

$$(x - a)^2x^2 + y^2(x - a)^2 = b^2x^2$$

or

$$(x - a)^2(x^2 + y^2) = b^2x^2$$

The line ℓ is an asymptote for the conchoid. This discovery is attributed to Nicomedes himself. We check the result with our method for finding all asymptotes.

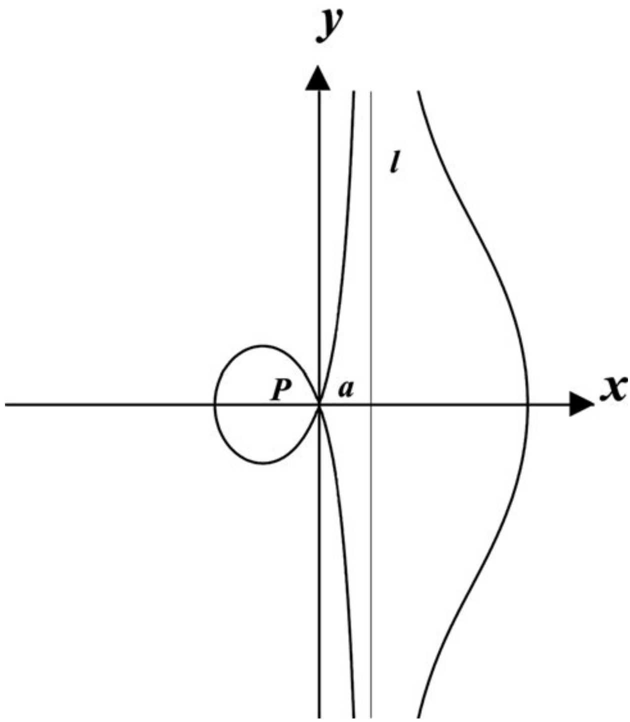


Fig. 15.2 The Conchoid of Nicomedes with $b = 3 > a = 1$

The projective closure of the conchoid is given by

$$(X_1 - aX_0)^2(X_1^2 + X_2^2) - b^2X_0^2X_1^2 = 0.$$

The points at infinity are determined by

$$X_1^2(X_1^2 + X_2^2) = 0$$

thus as we only consider real points, the only point at infinity is $(0 : 0 : 1)$. We now take the affine restriction to $D_+(X_2)$ by letting $x' = \frac{X_0}{X_2}$ and $y' = \frac{X_1}{X_2}$. The equation in the $x'y'$-plane becomes

$$(y' - ax')^2(y'^2 + 1) - b^2x'^2y'^2 = 0.$$

The homogeneous part of lowest degree of this polynomial is $H(x', y') = (y' - ax')^2$, so the tangent cone at the point $(0, 0)$ is the line $y' = ax'$, with multiplicity 2. Taking the projective closure again we get the projective line $X_1 - aX_0 = 0$, and its affine restriction to our original affine plane $\mathbb{R}^2 = D_+(X_0)$ is $x = a$. Thus we have proved Nicomedes' theorem on the asymptote of the conchoid.

The Conchoid of Nicomedes is a special case of a general class of curves. We make the following

Definition 28. Given an affine curve K and a fixed point P. Consider the collection of all lines through P. The conchoid of K for the pole P and constant b is then the locus of all points Q such that Q lies on one of these lines at a distance b from its intersection with K.

A conchoid of *a circle for a fixed point on it* is called a *Limaçon of Pascal*, the first part of the name picked by Étienne Pascal, the name meaning *snail* in French. When b is equal to the diameter d of the circle, the curve is called the *cardioid*, in other words the *heart curve*, and if the constant b is equal to the radius of the circle, we get a curve which may be used to trisect an angle in equal parts, often referred to as a *trisectrix* (but not to be confused with the Trisectrix of Maclaurin, treated earlier). We shall now analyze the different cases. Depending on the relation between b and d we get three versions of the Limaçon, one being shown in Fig. 15.3, the two others in Fig. 15.4.

These curves are simple to describe in polar coordinates, as

$$r = b + d\cos(\varphi)$$

where d is the diameter of the circle and b is the constant, see Fig. 15.3.

A simple computation yields

$$x^2 + y^2 - dx = db\cos(\varphi) + b^2$$

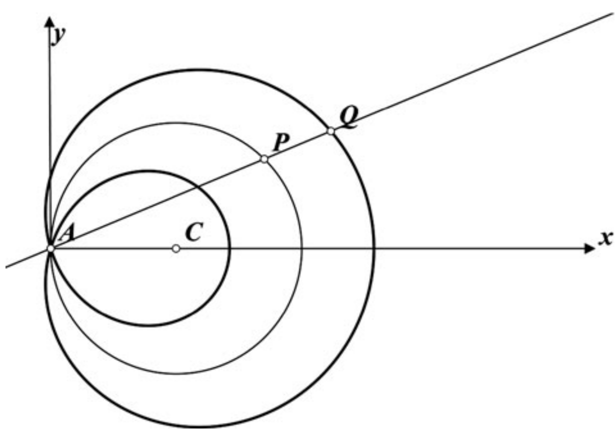


Fig. 15.3 The Conchoid of a circle, called a Limaçon. The circle about C with radius AC is fixed. The line ℓ rotates about A, and the point Q is on ℓ at the fixed distance from the circle $b = PQ$. Of course $d = 2AC$, here $d > b$

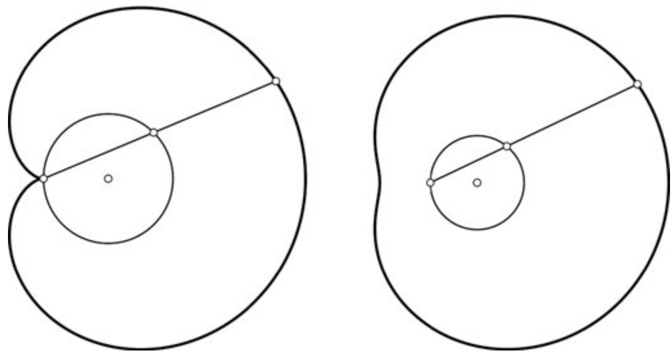


Fig. 15.4 The two other versions of the limaçon. To the left $d = b$, to the right $d < b$

and hence

$$(x^2 + y^2 - dx)^2 = d^2b^2\cos^2(\varphi) + 2db^3\cos(\varphi) + b^4 = b^2(x^2 + y^2)$$

Thus we obtain the somewhat less transparent form in usual xy-coordinates

$$(x^2 + y^2 - dx)^2 - b^2(x^2 + y^2) = 0.$$

In Fig. 15.4 we see some limaçons.

15.8 The Dual Curve

In projective algebraic geometry the *principle of duality* acquires a very precise meaning. For $\mathbb{P}^2(\mathbb{R})$ we have the following:

Every line in $\mathbb{P}^2(\mathbb{R})$ is given by an equation

$$a_0X_0 + a_1X_1 + a_2X_2 = 0.$$

If we multiply each coefficient by a common non-zero constant, then we get the same equation. Thus the line may be associated with a uniquely determined point of another copy of the projective plane,

$$L^\vee = (a_0 : a_1 : a_2) \in \mathbb{P}^2(\mathbb{R}).$$

Conversely, to any point $P \in \mathbb{P}^2(\mathbb{R})$ we may associate a *line* $P^\vee \subset \mathbb{P}^2(\mathbb{R})$. The correspondence $(\ )^\vee$ preserves *incidence*, as already explained in Sect. 13.2

We now extend this to *projective, algebraic curves.* We get the following concept of duality: For any projective curve $C \subset \mathbb{P}^2(\mathbb{R})$, consider the subset

$$C^\vee = \left\{(L)^\vee \,|L \text{ is a line of tangency to } C\right\}.$$

We denote this set by $C^\vee$, and refer to it as the *dual curve of* C. Indeed, it turns out that this subset of $\mathbb{P}^2(\mathbb{R})$ is actually a projective curve, in $\mathbb{P}^2(\mathbb{R})$, except for the case when C is a projective line, in which case $C^\vee$ consists of just one point.

Assume that C has the equation $F(X_0, X_1, X_2) = 0$. The equation for the dual curve is then expressed in terms of the indeterminates Y_0, Y_1 and Y_2 when we eliminate X_0, X_1, X_2 in the system

$$\begin{aligned}\frac{\partial F}{\partial X_0}(X_0, X_1, X_2) &= Y_0\\ \frac{\partial F}{\partial X_1}(X_0, X_1, X_2) &= Y_1\\ \frac{\partial F}{\partial X_2}(X_0, X_1, X_2) &= Y_2\\ F(X_0, X_1, X_2) &= 0.\end{aligned}$$

Here we have six variables X_0, X_1, X_2 and Y_0, Y_1, Y_2, and four relations among them. In general we may then *eliminate* any three of them, and obtain *one relation* between the remaining variables. We now eliminate X_0, X_1 and X_2. This will give as result a single equation

$$G(Y_0, Y_1, Y_2) = 0$$

which defines the dual curve $C^\vee$.

We may think of what we are doing here in the following way: We write

$$\begin{aligned}\frac{\partial F}{\partial X_0}(X_0, X_1, X_2) &= Y_0(X_0, X_1, X_2)\\ \frac{\partial F}{\partial X_1}(X_0, X_1, X_2) &= Y_1(X_0, X_1, X_2)\\ \frac{\partial F}{\partial X_2}(X_0, X_1, X_2) &= Y_2(X_0, X_1, X_2).\end{aligned}$$

We then *solve this system of equations* for X_0, X_1 and X_2:

$$\begin{aligned}X_0 &= X_0(Y_0, Y_1, Y_2)\\ X_1 &= X_1(Y_0, Y_1, Y_2)\\ X_2 &= X_2(Y_0, Y_1, Y_2)\end{aligned}$$

and then get

$$G(Y_0, Y_1, Y_2) = F(X_0(Y_0, Y_1, Y_2), X_1(Y_0, Y_1, Y_2), X_2(Y_0, Y_1, Y_2)).$$

Of course, usually we are not able to find X_0, X_1 and X_2 as homogeneous polynomials in the Y's, not even as single valued functions. To put these considerations on a mathematically sound basis, it was necessary to develop the machinery of *elimination theory*. But we shall bypass this, and work for a while with such fictitious entities as the X_0, X_1 and X_2 as functions of Y_0, Y_1 and Y_2. In the end they are gone, and only the $G(Y_0, Y_1, Y_2)$, which does exist thanks to elimination theory, remains. But in some happy cases the X_0, X_1 and X_2 *do exist* as homogeneous polynomials in Y_0, Y_1 and Y_2, and then they simplify the situation considerably.

Since questions of tangency are independent of projective coordinate system, the same is true for questions of duality.

We shall use this important observation in proving the following

Theorem 26. *The dual curve of a non-degenerate conic section in* $\mathbb{P}^2(\mathbb{R})$ *is again a non-degenerate conic section.*

Proof. We showed in Theorem 25 that any non-degenerate conic section is projectively equivalent to the one given by

$$X_1^2 + X_2^2 - X_0^2 = 0$$

thus we may assume that $F(X_0, X_1, X_2) = X_1^2 + X_2^2 - X_0^2$. Then

$$\frac{\partial F}{\partial X_0} = -2X_0$$
$$\frac{\partial F}{\partial X_1} = 2X_1$$
$$\frac{\partial F}{\partial X_2} = 2X_2.$$

So we have to eliminate X_0, X_1, X_2 in

$$-2X_0 = Y_0$$
$$2X_1 = Y_1$$
$$2X_2 = Y_2$$
$$X_1^2 + X_2^2 - X_0^2 = 0.$$

In this case we may solve for X_0, X_1 and X_2, which yields

$$Y_1^2 + Y_2^2 - Y_0^2 = 0$$

which is a non-degenerate conic section. □

As we see, the equation is the same as the one we started with, only the indeterminates have different names. Just looking at this one example, one might be tempted to draw the conclusion that $C = C^\vee$. But this is far from true, even for general conic sections. The point is that the property of *having a non-degenerate conic section as dual* is independent of the coordinate system, while the property of *having a dual which is defined by a fixed homogeneous polynomial* certainly very much depends on the coordinate system. However, we have the following important theorem, which is true in much greater generality than the version we give here:

Theorem 27. *Let C be a curve given by an irreducible homogeneous polynomial in $\mathbb{P}^2(\mathbb{R})$. Then if we dualize the dual curve of C, we get C back: $C^{\vee\vee} = C$*

Proof. We use the simplified form given above, avoiding elimination theory. Then G is really nothing but the original F, but expressed in terms of the variables Y_0, Y_1, Y_2 instead of X_0, X_1, X_2. We therefore only have to prove that if we put

$$\overline{X}_0 = \frac{\partial F}{\partial Y_0}$$
$$\overline{X}_1 = \frac{\partial F}{\partial Y_1}$$
$$\overline{X}_2 = \frac{\partial F}{\partial Y_2}$$

then

$$\overline{X}_0 = \alpha X_0, \overline{X}_1 = \alpha X_1 \text{ and } \overline{X}_2 = \alpha X_2$$

for some real number $\alpha \neq 0$. This is done as follows: By Proposition 22, we have that

$$X_0 Y_0 + X_1 Y_1 + X_2 Y_2 = dF(X_0, X_1, X_2)$$

where d is the degree of F in X_0, X_1, X_2. Hence

$$\overline{X}_i = \frac{1}{d}\frac{\partial}{\partial Y_i}(X_0 Y_0 + X_1 Y_1 + X_2 Y_2) = X_i$$

for $i = 0$, 1 and 2. Thus the claim follows. □

15.9 The Dual of Pappus' Theorem

The degenerate case of Theorem 21 is known as *Pappus' theorem*. It is shown to the left in Fig. 15.5.

We shall now show how we may obtain a new, apparently completely different, theorem simply by dualizing Pappus' theorem.

To the left we have a point O, with two lines ℓ and ℓ' passing through it. To the right there is a *line* o, with two points L and L' on it. To the left we have three points, all different and different from O, on each of the lines. They are labelled A, B and C for the points on ℓ, and A', B' and C' for the points on ℓ'. Dually, to the right there are lines a, b and c through L, and a', b' and c' through L', all of them different from o and from one another.

Now we draw the lines AB' and BA', and mark their point of intersection. We do the same for the pairs AC' and $A'C$, BC' and $B'C$. Then, as Pappus' Theorem tells us, these three points are *collinear*, they lie on one line, labelled p.

Dually, to the right, we take the point of intersection between a and b', and the point of intersection between a' and b. We draw the line through these two points.

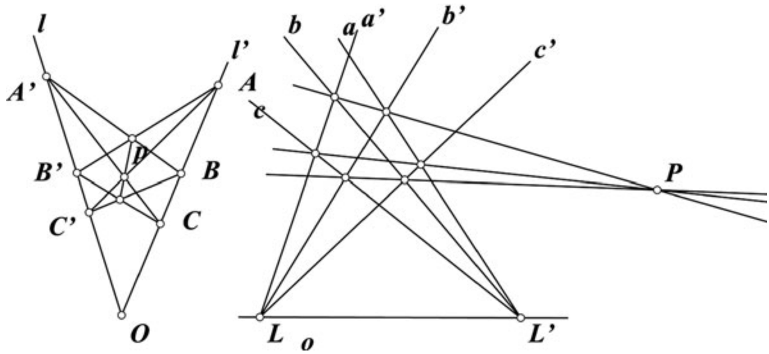


Fig. 15.5 Pappus' theorem, to the left, and its dual to the right

Similarly we find the point of intersection of a and c', as well as the point of intersection of a' and c. We draw the line between these two points as well. Finally, we find the point of intersection of b and c', as well as the point of intersection of b' and c. And, for the third time, draw the line between these two points. We now have drawn altogether three lines. The theorem is that *these three lines pass through a common point*, labelled P in the right part of the figure.

15.10 Pascal's Mysterium Hexagrammicum

In Fig. 15.6 the ellipse to the left illustrates Pascal's Theorem. Six points A, B, C and A', B', C' are given on the non-degenerate conic section. Then we draw lines connecting each point to two of the other points, we have the lines AB', AC', $A'B$, $A'C$ and BC', $B'C$. We form three pairs of these lines by pairing the ones labelled by the same letters, primed or unprimed, the three pairs determine three points of intersection labelled D, E and F. Then by the theorem these three points lie on one line, labelled p.

To the right we dualize the situation. The dual of a non-degenerate conic section is again a non-degenerate conic section. Choosing points on the conic section corresponds to selecting tangents on the dual conic section, they are labelled a, b, c and a', b', c. To draw lines connecting each point to two of the other points, correspond dually to taking the points of intersection between one tangent with two of the other ones. We then have the points ab', ac', $a'b$, $a'c$ and bc', $b'c$. Taking the points of intersection between similarly labelled lines dually corresponds to drawing lines between similarly labelled points. In Pascal's Theorem the points of intersection lie on the same line p, which dually corresponds to the lines in the dual situation passing through the same point P.

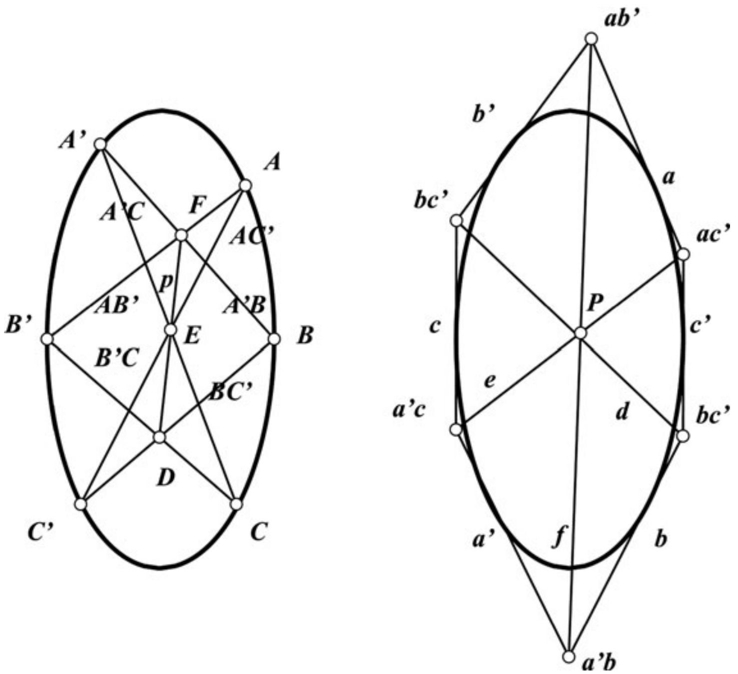


Fig. 15.6 Pascal's *Mysterium Hexagrammicum*

We have proved the following:

Theorem 28. *Let there be given six tangent lines to a non-degenerate conic section. In the resulting circumscribed hexagon connect diametrically opposite corners. The resulting lines then pass through a common point.*

Exercises

Exercise 15.1 (a) Find the projective closures of all the curves listed in Exercises 14.1.
(b) Find the asymptotes, if any, of all the above affine curves.
(c) Find the equations of the curves dual to the projective curves in (a).
(d) Find all points of intersection between the projective curves above and the line $X_0 + X_1 + X_2$, if necessary by using MAPLE or an advanced calculator. Also compute the intersection multiplicity at each of these points.

Exercise 15.2 The same problem as in Exercise 15.1 for the curves in 14.2.

Exercise 15.3 The same problem as in Exercise 15.1 for the curves in 14.3.

Exercise 15.4 The same problem as in Exercise 15.1 for the curves in 14.4.

Exercise 15.5 The same problem as in Exercise 15.1 for the curves in 14.5.

Exercise 15.6 The same problem as in Exercise 15.1 for the curves in 14.6.

Exercise 15.7 Let K be one of the curves in 14.1. For each of these curves compute the conchoid of K for the pole $P = (\alpha, \beta)$ and constant b. Recall that this is the locus of all points Q such that Q lies on a line through P at a distance b from its intersection with K.

Exercise 15.8 Let K be one of the curves in 14.5. For each of these curves compute the conchoid of K for the pole $P = (0, 0)$ and constant 1.

Chapter 16
Sharpening the Sword of Algebra

16.1 On Rational Polynomials

A *rational polynomial in the variable x* is an expression

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0, a_n \neq 0,$$

where $a_n, \ldots, a_0$ are rational numbers. Similarly we define integral, real or complex polynomials. The degree of $f(x)$ is the integer n. For the *zero polynomial*, the one where all coefficients are zero, it is convenient to define the degree as $-\infty$.

We add polynomials by adding the coefficients of the same powers of x, and multiplication is carried out by the formula

$$(a_m x^m + a_{m-1}x^{m-1} + \ldots + a_0) \cdot (b_n x^n + b_{n-1}x^{n-1} + \ldots + b_0)$$
$$= c_N x^N + c_{N-1}x^{N-1} + \ldots + c_0.$$

where $N = m + n$, and $c_N = a_m b_n$, $c_{N-1} = a_m b_{n-1} + a_{m-1}b_m$, and so on, down to $c_0 = a_0 b_0$.

The concepts of addition, subtraction and multiplication of polynomials are probably reasonably well known to the readers already, but less familiar, perhaps, is the concept of *polynomial division*.

Suppose that we wish to divide the polynomial $x^5+2x^4+x^2+4x+2$ by x^2+1. This will give a quotient $q(x)$ *as well as a remainder* $r(x)$. This is very similar to the situation for *division of integers*. We proceed as follows:

$$x^5 + 2x^4 + x^2 + 4x + 2 : x^2 + 1 = x^3 + 2x^2 - x - 1$$

$$\begin{array}{l} -(x^5 + x^3) \\ \hline 2x^4 - x^3 + x^2 + 4x + 2 \\ -(2x^4 + 2x^2) \\ \hline -x^3 - x^2 + 4x + 2 \end{array}$$

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_16,

$$\begin{array}{l} -(-x^3 - x) \\ \hline -x^2 + 5x + 2 \\ -(-x^2 - 1) \\ \hline 5x + 3 \end{array}$$

So here the quotient is

$$q(x) = x^3 + 2x^2 - x - 1,$$

and the remainder is

$$r(x) = 5x + 3.$$

We may also express this by writing

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

In this form most students of calculus will need polynomial division as a tool for computing integrals of rational functions.

In the general case we proceed exactly as in the example above: Suppose that we have polynomials $a(x)$ and $b(x)$, both with rational coefficients. Then we may find polynomials $q(x)$ and $r(x)$, where $r(x)$ is either the zero polynomial or is of degree $<$ the degree of $b(x)$, such that

$$\frac{a(x)}{b(x)} = q(x) + \frac{r(x)}{b(x)}.$$

If $a(x)$ is of degree *less than* the degree of $b(x)$, we take $q(x) = 0$, and $a(x) = r(x)$. If, on the other hand, $a(x)$ is of degree $\geq$ the degree of $b(x)$, we find $q(x)$ as in the example: Assume that

$$a(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0, a_n \neq 0,$$

and

$$b(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_0, b_m \neq 0.$$

where we assume $n \geq m$. Then the first term (*the leading term*) of $q(x)$ is $\frac{a_n}{b_m} x^{n-m}$. We multiply this term by $b(x)$, the result is subtracted from $a(x)$ and we obtain a new polynomial $a_1(x)$ which is of degree $< n$. Then the process is repeated with $a_1(x)$ instead of $a(x)$, this yields the next term of $q(x)$, and so on, until we get a polynomial $a_h(x)$ which is of degree $< m$. Then it is no longer possible to repeat the procedure, and we have found $q(x)$, while $r(x)$ is the polynomial $a_h(x)$. We have shown the following:

Proposition 24. *Let there be given two polynomials $a(x)$ and $b(x)$, both with rational coefficients. Then there exists two other rational polynomials $q(x)$ and $r(x)$,* *where $r(x)$ is either the zero polynomial or is of degree < the degree of $b(x)$,*[1] *such that*

$$\frac{a(x)}{b(x)} = q(x) + \frac{r(x)}{b(x)}.$$

We shall need a result about *rational roots of polynomial equations*. This is also useful in other situations, for instance in computing integrals involving rational functions.

Proposition 25. *Let*

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 = 0, \text{ where } a_n, a_0 \neq 0,$$

be an equation with integral coefficients,

$$a_n, \ldots, a_0 \in \mathbb{Z}.$$

Let $x = \frac{r}{s}$ be a rational root, where r and s are mutually prime integers, i.e., the expression for x cannot be simplified. Then there exist integers a and b such that

$$a_0 = ra, a_n = sb.$$

In particular if $a_n = 1$, which we express by saying that the polynomial is monic, then $s = \pm 1$, and every rational root of the equation must be an integer dividing the constant term a_0.

Proof. It suffices to show that there exist an integer a such that $a_0 = ra$: Indeed, $y = \frac{s}{r}$ is a root of the equation

$$a_0 y^n + \cdots + a_n = 0,$$

and the first part of the claim applied to this situation yields the second part of the claim.

We get

$$a_n(\frac{r}{s})^n + a_{n-1}(\frac{r}{s})^{n-1} + \cdots + a_0 = 0,$$

thus

$$a_n r^n + a_{n-1} r^{n-1} s + \cdots + a_0 s^n = 0.$$

This gives

$$a_0 s^n = r(a_n r^{n-1} + \cdots + a_1 s^{n-1}) = rA,$$

where A is an integer. We therefore obtain

[1] Strictly speaking the first part of the sentence is redundant, as we have defined the zero polynomial to be of degree $-\infty$.

$$a_0 = \frac{rA}{s^n}.$$

So the rational expression $\frac{rA}{s^n}$ is, therefore, *an integer*, which means that any prime number dividing s must divide rA. But no prime factor of s divides r, so that it has to divide A. Repeating the argument, if necessary, we find that all prime factors of s^n divide A, and raised to the power to which they occur in s^n, so

$$a = \frac{A}{s^n}$$

has to be an integer, and the claim follows. □

Remark 6. In this proof we have taken certain fundamental properties of the integers for granted. These are properties which we normally use without thinking twice about it, but nevertheless the terms involved need to be defined, and the facts need proofs. They are as follows:

1. A prime number is a positive integer with no divisors except 1 and itself. The first prime numbers are $2, 3, 5, 7, 11, 13, \ldots$
2. If a prime number p divides a product of two integers ab, then p either divides a or it divides b.
3. Every positive integer n may be written uniquely as

$$n = p_1^{i_1} p_2^{i_2} \cdots p_r^{i_r},$$

where $p_1 < \cdots < p_r$ are prime numbers.

We shall not pursue the issue here, but refer instead to any introductory text on algebra or number theory. The material outlined here was important to Euclid, and forms part of Books VII, VIII and IX of his Elements.

16.2 The Minimal Polynomial

If a real number α satisfy an equation $p(\alpha) = 0$, where $p(x)$ is a polynomial with rational coefficients, then we frequently need to find *all* rational polynomials which have α as a root. It turns out that the answer is very simple and elegant.

Let $p_0(x)$ denote a polynomial with coefficients from $\mathbb{Q}$. We assume that $p_0(\alpha) = 0$, and that $p_0(x)$ is not the zero polynomial. Assume also that it is of minimal degree among the polynomials which have α as a root. We may assume that *the coefficient of the highest power of x which occurs in $p_0(x)$ is equal to 1*, this may be accomplished by dividing the polynomial by this coefficient. The following important result implies that $p_0(x)$ is uniquely determined, and we call $p_0(x)$ the *minimal polynomial of α over $\mathbb{Q}$*.

Theorem 29. *If $p(x)$ is a rational polynomial such that $p(\alpha) = 0$, and $p_0(x)$ denotes the minimal polynomial of* α *over* $\mathbb{Q}$, *then $p(x)$ is equal to $p_0(x)h(x)$ for some rational polynomial $h(x)$.*

Proof. By Theorem 24 we find

$$p(x) = q(x)p_0(x) + r(x),$$

where the polynomial $r(x)$ is of degree < the degree $p_0(x)$. If $r(x)$ is not the zero polynomial, then $r(\alpha) = 0$ will contradict that $p_0(x)$ is of minimal degree among the non-zero polynomials which have α as root. Thus $r(x)$ must be the zero polynomial, and the claim follows. □

In particular this theorem implies that $p_0(x)$ is an irreducible polynomial, that is to say that it cannot be written as a product of other non-constant polynomials with rational coefficients: In fact,

$$p_0(x) = a(x)b(x)$$

implies that

$a(x)$ or $b(x)$ are constant polynomials.

Indeed, if $a(\alpha)b(\alpha) = 0$, then at least one of the factors vanish, say $a(\alpha) = 0$. But then

$$a(x) = p_0(x)q(x),$$

so that

$$p_0(x) = a(x)b(x) = p_0(x)q(x)b(x).$$

Abbreviating this expression we get

$$q(x)b(x) = 1,$$

thus in particular we obtain that both $b(x)$ and $q(x)$ are constant polynomials.

Remark 7. Above we have used the fact that in a polynomial expression we may always perform abbreviations. This follows from the observation that the product of two non-zero polynomials with coefficients from $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ cannot be the zero polynomial. In fact, if

$$a(x) = a_m x^n + a_{m-1}x^{m-1} + \cdots + a_0$$

and

$$b(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_0$$

where a_m and b_n are $\neq 0$, then $a_m b_n \neq 0$, thus the product $a(x)b(x)$ is not the zero polynomial.

16.3 The Euclidian Algorithm

Euclid's Algorithm is a method for finding *the greatest common divisor* of two integers or of two polynomials. We first illustrate the method for two integers.

So let there be given two integers a and b, which we may assume to be > 0. Division of a by b yields a quotient $q \geq 0$ and a remainder r such that $0 \leq r < b$:

$$a = bq + r, 0 \leq r < b.$$

If $r = 0$, then $a = bq$, so that the greatest common divisor of a and b is b itself. If on the other hand $r > 0$, then we repeat the division as follows:

$$b = rq_1 + r_1, 0 \leq r_1 < r.$$

If now $r_1 = 0$, then the greatest common divisor of a and b is r: Clearly r divides both a and b. For we have $a = bq_1 + r, b = rq_1$, so that $a = r(qq_1 + 1)$. On the other hand if d is a common divisor of a and b, then d will have to divide r, since $r = a - bq$.

If $r_1 > 0$ we repeat the process. Sooner or later we get a remainder $r_i = 0$: Indeed, we have that

$$r > r_1 > r_2 > \ldots \geq 0,$$

and clearly such a sequence of strictly decreasing non-negative integers cannot be infinite, so at some point $r_i = 0$. We get the following, where we have put $r = r_0$ and $q = q_0$:

$$a = bq_0 + r_0, 0 < r_0 < b$$

$$b = r_0q_1 + r_1, 0 < r_1 < r_0$$

$$r_0 = r_1q_2 + r_2, 0 < r_2 < r_1$$

$$\vdots$$

$$r_{i-3} = r_{i-2}q_{i-1} + r_{i-1}, 0 < r_{i-1} < r_{i-2}$$

$$r_{i-2} = r_{i-1}q_i$$

and we have $r_i = 0$ in the last step, after $i + 1$ steps. This sequence of divisions, which ends when the remainder becomes 0, is referred to as *the Euclidian Algorithm.* The point is the following result:

Theorem 30. *The greatest common divisor for the positive integers a and b is the number r_{i-1}, the last non-vanishing remainder in the Euclidian Algorithm for the numbers a and b.*

Proof. By the last line in the algorithm r_{i-1} divides r_{i-2}, which by the next to the last line implies that r_{i-1} divides r_{i-3}, and so on. In the end we find that r_{i-1} divides b, which by the first line finally yields that r_{i-1} divides a. So r_{i-1} is a common divisor of a and b.

Conversely assume that d is a divisor of a and b. By the first line d divides r_0, so by the second line it also divides r_1,and so on. The next to the last line shows that d divides r_{i-1}, and thus r_{i-1} is the greatest common divisor of a and b. □

Remark 8. When we speak of the greatest common divisor of the positive integers a and b, the naive interpretation is to mean the largest integer c which divides both a and b. But this turns out to be the same as to require that the positive integer c shall divide both a and b, and that any integer d which divides a and b, should also divide c. It is easily seen that these two definitions of *greatest common divisor* for positive integers are equivalent, and we have used the latter form in the above proof.

It is also this definition which is used for greatest common divisor of two *polynomials* $a(x)$ and $b(x)$:

Definition 29. Let a and b be either positive integers, or polynomials with coefficients from $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ in the variable x, different from the zero polynomial. Then d is called a divisor in a if there exists q, a positive integer or a polynomial as the case may be, such that $a = dq$. c is called the greatest common divisor of a and b if c is a divisor in a and b, and if every divisor in a and b also is a divisor in c. We put

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

Note that c is positive when a and b are positive integers. We shall require that $c(x)$ is a monic polynomial if $a = a(x)$ and $b = b(x)$ are polynomials in x: This means that the coefficient of the highest power of x which occurs in $c(x)$ is 1.

The lines constituting the Euclidian Algorithm contain much useful information. For example we may do the following: Putting $A_0 = 1$ and $B_0 = -q_0$, we get from the first line

$$r_0 = aA_0 + bB_0,$$

which by line 2 yields

$$r_1 = aA_1 + bB_1,$$

where $A_1 = -A_0q_1$, $B_1 = 1 - B_0q_1$. In the same way we get from line 3 that

$$r_2 = aA_2 + bB_2,$$

where $A_2 = A_0 - A_1q_2$, $B_2 = B_0 - B_1q_2$. From the line numbered i we find that

$$r_{i-1} = r_{i-3} - r_{i-2}q_{i-1} = aA_{i-3} + bB_{i-3} - (aA_{i-2} + bB_{i-2})q_{i-1},$$

which gives

$$r_{i-1} = aA_{i-1} + bB_{i-1}$$

where

$$A_{i-1} = A_{i-3} - A_{i-2}q_{i-1}, B_{i-1} = B_{i-3} - B_{i-2}q_{i-1}.$$

These computations are valid both if a and b are positive integers, and when they are polynomials different from the zero polynomial. In particular we have proved the following:

Theorem 31. *Let a and b be either positive integers or polynomials in x different from the zero polynomial, with coefficients from $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$. Then there exist A and B, integers (not necessarily positive) or polynomials of the type in question, such that*

$$(a, b) = aA + bB.$$

The identity in Theorem 31 is referred to as *Bézout's Identity*. Etienne Bézout, 1730–1783, was a French mathematician who among other things is known for some excellent textbooks. He gave the first satisfactory proof of the assertion that two projective curves in $\mathbb{P}^2(\mathbb{R})$ or in $\mathbb{P}^2(\mathbb{C})$ of degrees m and n, without any common components, intersect in exactly mn points, real or complex, and counted with multiplicity. This fact was asserted by Maclaurin, but is credited to Bézout as *Bézout's Theorem.*

16.4 Number Fields and Field Extensions

We have encountered numbers with various different properties: The set of the *natural* numbers, *the integers*, or *the rational numbers*. The *complex numbers* represent the most extensive algebraic system of objects which we may refer to as *numbers* without stretching the concept. But we may go on, if we are willing to abandon some of the fundamental properties: The *Hamiltonian Quaternions* resemble the complex numbers in many ways, but multiplication is no longer commutative: In general $ab \neq ba$. The *Cayley Numbers* or the *octonians* are even more weird, not only is multiplication non-commutative, it is *non-associative* as well: In general $a(bc) \neq (ab)c$. As $\mathbb{C}$ is based on a multiplication introduced in $\mathbb{R}^2$, so the quaternions is based on $\mathbb{R}^4$ and the octonians on $\mathbb{R}^8$. There are also "number systems" in which there are only *finitely many integers*, in that we may have

$$1 + 1 + \cdots + 1 + 1 = 0$$

if we add 1 to itself sufficiently many times. The smallest positive N such that adding N 1's together yields 0 must be a prime number (an ordinary prime number, of course), it is called the *characteristic* of the "number system". Actually computers internally work with a system where $N = 2$, and if N is very, very large, we might perhaps consider living with a number system like that.

Here we shall stay within the framework of the *the rational numbers*. However, as we shall see the rationals encompass a rich and interesting structure. Many questions, appearing simple, concerning *integers* remain unanswered despite intensive research. Other problems, which have occupied mathematicians since antiquity, have been answered only in modern times. The answers have become possible only by invoking the finer structure provided by *the system of real numbers*. This applies to the so called *classical problems*, namely *the trisection of an angle, the doubling of a cube* and *the squaring of a circle*. These geometric problems may be answered once and for all using the material we are about to explain. It is not easy, but that has to be expected. After all, the sharpest brains humanity has produced tried in vain to find the solution for more than 2,000 years!

We start out with the following fundamental definition.

Definition 30. A (real) number field K is a set of real numbers which contains 0 and 1 and which is such that if $r, s \in K$ then $r \pm s \in K, rs \in K$, and $\frac{r}{s} \in K$ if $s \neq 0$.

Clearly, if K is a number field, then $K \supset \mathbb{Q}$. It is also clear that $\mathbb{R}$ itself is a number field. It is the study of the number fields between these two extremes which yields the insights into the *classical problems*.

Definition 31. If $K \subset L$ are two number fields, then we refer to L as an extension of K.

In the first two paragraphs of this chapter we have treated polynomials with coefficients from $\mathbb{Q}$ and $\mathbb{R}$. But everything we did there applies equally well to polynomials over K where K is any number field, that is to say polynomials whose coefficients all lie in K. In particular we have the following:

Proposition 26. *Let there be given two polynomials $a(x)$ and $b(x)$ over the number field K, i.e., with coefficients from k. We assume that $b(x)$ is not the zero polynomial. Then there exists two other polynomials over K $q(x)$ and $r(x)$, where $r(x)$ either is the zero polynomial or is of degree < the degree of $b(x)$, such that*

$$\frac{a(x)}{b(x)} = q(x) + \frac{r(x)}{b(x)}.$$

Assume that the real number α satisfies an equation $p(\alpha) = 0$, where $p(x)$ is a polynomial over K. As we did for $K = \mathbb{Q}$ we consider the non-zero polynomial $p_0(x)$ with coefficients from K which have α as a root, and which is of minimal degree among the non-zero polynomials which have this property. As before we also specify that the highest power of x which occurs in $p_0(x)$ has coefficient 1, i.e., that the polynomial be *monic*. We refer to $p_0(x)$ as *the minimal polynomial of α over K*. As before we have the

Theorem 32. *If $p(x)$ is a polynomial over K such that $p(\alpha) = 0$, and $p_0(x)$ denotes the minimal polynomial of α over K, then $p(x) = h(x)p_0(x)$ for some polynomial $h(x)$ over K.*

Let K be a number field, and let α be a real number. If there exists a polynomial $p(x)$ over K, different from the zero polynomial, such that $p(\alpha) = 0$, then we say that α is *algebraic over K*. Otherwise we say that α is *transcendental* over K. If

$K = \mathbb{Q}$ then we say only that α is an algebraic or a transcendental number. Clearly all numbers in K are algebraic over K, as we may take simply $p_0(x) = x - \alpha$. Moreover, $\sqrt{2}$ is algebraic, being root in the equation $x^2 - 2 = 0$. On the other hand it is known that the number e – base for the natural logarithms – and the number π both are transcendental. The proofs are absolutely non-trivial. We return to this subject in Sect. 17.6.

Now let α be a number and let K be a number field. We let $K[\alpha]$ denote all polynomial expressions in the number α with coefficients from K:

$$K[\alpha] = \{\beta \,|\beta = f(\alpha) = a_n\alpha^n + \cdots + a_1\alpha + a_0, \text{ where } a_n, \ldots, a_0 \in K\}.$$

Assume first that α is algebraic over K. If the minimal polynomial $p(x) = b_m x^m + \cdots + b_0$ of α over K is of degree m, then we may assume that $n < m$ in the definition of $K[\alpha]$ above: Indeed, we have that

$$f(x) = q(x)p(x) + r(x), \text{ where } \deg(r(x)) < m,$$

so that if necessary we may replace $f(x)$ by $r(x)$.

We shall prove that

Proposition 27. *If α is algebraic, then the set $K[\alpha]$ actually is a number field.*

Proof. The proof is in no way obvious: We must show that if we have a polynomial $f(x)$ over K such that $f(\alpha) \neq 0$, then there exists another polynomial $g(x)$, also over K, such that $f(\alpha)g(\alpha) = 1$. We show this as follows: We claim that $(p_0(x), f(x)) = 1$. Indeed, if $p_0(x)$ and $f(x)$ should have a common factor of degree > 0, then this factor would have to be a constant multiple of $p_0(x)$ itself, since $p_0(x)$ an irreducible polynomial by Theorem 32. But then $f(\alpha) = 0$, against the assumption. Thus $(p_0(x), f(x)) = 1$, and by Theorem 31 there exist polynomials $A(x)$ and $B(x)$ such that

$$1 = p_0(x)A(x) + f(x)B(x),$$

which gives

$$1 = p_0(\alpha)A(\alpha) + f(\alpha)B(\alpha) = f(\alpha)B(\alpha).$$

Thus we may take $g(x) = B(x)$, and the claim is proven. □

Now let α be any number, algebraic or transcendental over K. We consider the set $K(\alpha)$ of all rational expressions in α:

$$K(\alpha) = \left\{\beta \,\middle|\, \beta = \frac{a_m\alpha^m + \ldots + a_0}{b_n\alpha^n + \ldots + b_0}, b_n\alpha^n + \ldots + b_0 \neq 0\right\}.$$

Clearly this is a number field, and it is the smallest number field which contains K and α. We have the following:

Proposition 28. $K(\alpha) = K[\alpha]$ *if and only if* α *is algebraic over* K.

Proof. If α is algebraic over K then $K[\alpha]$ is already a number field, as we have shown in Proposition 27. Thus $K[\alpha] = K(\alpha)$. If conversely $K[\alpha] = K(\alpha)$, then in particular $\frac{1}{\alpha} \in K[\alpha]$, thus there exists $a_0, \dots, a_m \in K$ such that $\frac{1}{\alpha} = a_m\alpha^m + \cdots + a_1\alpha + a_0$. But this implies that α is a root in a polynomial equation,

$$a_m x^{m+1} + \cdots a_1 x^2 + a_0 x - 1 = 0,$$

so that α must be algebraic over K. This completes the proof. □

We finally consider some simple examples. First let $\alpha = \sqrt{2}$. Then the minimal polynomial is $p_0(X) = X^2 - 2$. We express the situation as follows:

$$\left.\begin{array}{c}\mathbb{Q}(\sqrt{2})\\ | \\ \mathbb{Q}\end{array}\right]\ X^2-2.$$

Next consider the extension $\mathbb{Q}(\sqrt[4]{2})$ of $\mathbb{Q}$. Since $\sqrt{2} = (\sqrt[4]{2})^2 \in \mathbb{Q}(\sqrt[4]{2})$, we have $\mathbb{Q}(\sqrt[4]{2}) \supset \mathbb{Q}(\sqrt{2})$. We thus have a small *"tower"* of extensions which looks like this:

$$\left.\begin{array}{cl}\mathbb{Q}(\sqrt[4]{2}) & \\ | &]\ X^2-\sqrt{2} \\ \mathbb{Q}(\sqrt{2}) & \\ | &]\ X^2-2 \\ \mathbb{Q} & \end{array}\right]\ X^4-2.$$

More generally we have

$$\left.\begin{array}{cl}\mathbb{Q}(\sqrt[mn]{2}) & \\ | &]\ X^n-\sqrt[m]{2} \\ \mathbb{Q}(\sqrt[m]{2}) & \\ | &]\ X^m-2 \\ \mathbb{Q} & \end{array}\right]\ X^{mn}-2.$$

These examples show that the minimal polynomial changes when the base number field changes: The number $\sqrt[4]{2}$ has different minimal polynomials over the number fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}$.

16.5 More on Field Extensions

This section presupposes a basic knowledge of linear algebra. Nevertheless we have chosen to give a self contained treatment, in that the material which is needed and used will be explained. But the treatment is brief, and a reader may well omit the section at the first reading.

Definition 32. A number field L is referred to as an extension of another number field K if $L \supset K$. If all the numbers in L are algebraic over K, then we say that L is an algebraic extension of K.

In particular such a field extension L is a *vector space* over K. This concept will not be given a full explanation here, since we need only one important aspect of it, namely *the order* or *the dimension* of L over K.

Let $\beta_1, \ldots, \beta_m \in L$. We say at these elements are *linearly dependent* over K if there exist elements $a_1, \ldots, a_m \in K$ which are not all zero, such that

$$a_1\beta_1 + \cdots + a_m\beta_m = 0.$$

Otherwise we say that $\beta_1, \ldots, \beta_m$ are linearly *independent* over K.

We have seen an important example of this concept in the previous section: Let $L = K(\alpha)$, where α is algebraic over K, with minimal polynomial $p_0(X)$. Let $d = \deg(p_0(X))$. Then

$$1, \alpha, \alpha^2, \ldots, \alpha^{d-1}$$

are linearly independent, while

$$1, \alpha, \alpha^2, \ldots, \alpha^{d-1}, \alpha^d$$

are linearly dependent over K. For the latter of the two claims, $p_0(\alpha) = 0$ yields the relation

$$a_0 1 + a_1\alpha + a_2\alpha^2 + \cdots + a_{d-1}\alpha^{d-1} + \alpha^d = 0$$

thus $1, \alpha, \alpha^2, \ldots, \alpha^{d-1}$ and α^d are linearly dependent. The former claim, that $1, \alpha, \alpha^2, \ldots, \alpha^{d-1}$ be linearly independent, follows since d is the degree of the minimal polynomial of α over K: No polynomial of lower degree than d with coefficients from K may have α as a root.

In this example we have a further property: Namely that all elements β from L may be written as

$$\beta = b_0 1 + b_1\alpha + b_2\alpha^2 + \cdots + b_{d-1}\alpha^{d-1},$$

where $b_i \in K$. We say that β is a linear combination in the linearly independent elements $1, \alpha, \alpha^2, \ldots, \alpha^{d-1}$ with coefficients from K. Such a set of elements in L is called a *basis* for L over K:

Definition 33. Let the number field L be an extension of the number field K. The elements $\alpha_1, \dots, \alpha_m \in L$ is called a basis for L over K if the following two conditions are satisfied:

1. $\alpha_1, \dots, \alpha_m$ are linearly independent.
2. Every element in L may be written as a linear combination in $\alpha_1, \dots, \alpha_m$ with coefficients from K.

In this definition we may combine the two conditions into one *single condition*, so that we get the following definition of a basis for L over K:

3. Every element in L may be written *uniquely* as a linear combination $\alpha_1, \dots, \alpha_m$ with coefficients from K.

We need the following important result:

Theorem 33. *Let $\alpha_1, \dots, \alpha_n$ and $\beta_1, \dots, \beta_m$ be two bases for L over K. Then $m = n$.*

Proof. We may assume that $m \geq n$, if necessary by interchanging the two bases. To show is that $m = n$. We have:

$$\beta_1 = a_{1,1}\alpha_1 + \cdots + a_{1,n}\alpha_n$$

$$\beta_2 = a_{2,1}\alpha_1 + \cdots + a_{2,n}\alpha_n$$

$$\cdots$$

$$\beta_m = a_{m,1}\alpha_1 + \cdots + a_{m,n}\alpha_n.$$

We prove this by replacing the base $\beta_1, \dots, \beta_m$ by a new one, $\beta'_1, \dots, \beta'_m$, where the number of elements is still m, but which is easier to compare to $\alpha_1, \dots, \alpha_n$. This is repeated until we can read off that $n = m$. We use the method of *Gaussian elimination*, which was introduced in Sect. 6.7.

Step 1. By renumbering $\alpha_1, \dots, \alpha_n$, we may assume that $a_{1,1} \neq 0$. By putting $\beta'_1 = \frac{1}{a_{1,1}}\beta_1$ we get $a_{1,1} = 1$. We now replace β_2 by $\beta'_2 = \beta_2 - a_{2,1}\beta_1$. $\beta_1, \beta'_2, \beta_3, \dots, \beta_m$ still is a base, and α_1 does not occur in β'_2. Thus we have accomplished that $a_{2,1} = 0$. Repeating the procedure we may assume that $a_{3,1} = \cdots = a_{m,1} = 0$. We now proceed to the next step.

Step 2. By renumbering $\alpha_2, \dots, \alpha_n$, we may assume that $a_{2,2}$ is $\neq 0$. Putting $\beta'_2 = \frac{1}{a_{2,2}}\beta_2$ we get $\beta_1, \beta'_2, \dots, \beta_m$ which still is a base, and we may replace β_2 by β'_2. Then $a_{2,2} = 1$. As above we may simplify further, and assume that $a_{3,2} = \cdots = a_{m,2} = 0$.

We continue in this manner. After $r \leq n$ steps we will be left with the situation

$$
\begin{array}{lllllll}
\beta_1 = & \alpha_1 + & a_{1,2}\alpha_2 + & a_{1,3}\alpha_3 + & \cdots + a_{1,r}\alpha_r + \cdots + a_{1,n}\alpha_n \\
\beta_2 = & & \alpha_2 + & a_{2,3}\alpha_3 + & \cdots + a_{2,r}\alpha_r + \cdots + a_{2,n}\alpha_n \\
\beta_3 = & & & \alpha_3 + & \cdots + a_{3,r}\alpha_r + \cdots + a_{3,n}\alpha_n \\
 & & \cdots & & \\
\beta_r = & & & & \alpha_r + \cdots + a_{r,n}\alpha_n \\
 & & \cdots & & \\
\beta_m = & & & & a_{m,r}\alpha_r + \cdots + a_{m,n}\alpha_n
\end{array}
$$

and when finally $r = n$ we have

$$
\begin{array}{lllllll}
\beta_1 = & \alpha_1 + & a_{1,2}\alpha_2 + & a_{1,3}\alpha_3 + & \cdots + a_{1,r}\alpha_r + \cdots + a_{1,n}\alpha_n \\
\beta_2 = & & \alpha_2 + & a_{2,3}\alpha_3 + & \cdots + a_{2,r}\alpha_r + \cdots + a_{2,n}\alpha_n \\
\beta_3 = & & & \alpha_3 + & \cdots + a_{3,r}\alpha_r + \cdots + a_{3,n}\alpha_n \\
 & & \cdots & & \\
\beta_n = & & & & a_{n,n}\alpha_n \\
 & & \cdots & & \\
\beta_m = & & & & a_{m,n}\alpha_n
\end{array}
$$

This is a base for L over K. But that is absurd unless $n = m$, and the claim is proved. □

Definition 34. Let L be a number field which contains the number field K. If L has a finite base $\alpha_1, \ldots, \alpha_n$ over K then we put $[L : K] = n$. Otherwise we write $[L : K] = \infty$.

With this notation we have shown the following important result above:

Theorem 34. *Let α be a number which is algebraic over the number field K. Then $[K(\alpha) : K]$ is equal to the degree of the minimal polynomial for α over K. If α is transcendental over K, then $[K(\alpha) : K] = \infty$.*

We finally prove a very important theorem, which is the key to understanding which constructions we may legally perform using compass and straightedge:

Theorem 35. *Let $M \supseteq L \supseteq K$ be three number fields. Then the following equality holds:*

$$[M : K] = [M : L][L : K].$$

Proof. Let $\alpha_1, \ldots, \alpha_m$ be a base for L over K, and $\beta_1, \ldots, \beta_n$ be a base for M over L. We claim that then

$$
\begin{gathered}
\alpha_1\beta_1, \ldots, \alpha_1\beta_n, \\
\alpha_2\beta_1, \ldots, \alpha_2\beta_n, \\
\cdots \\
\alpha_m\beta_1, \ldots, \alpha_m\beta_n
\end{gathered}
$$

is a base for M over K. This will of course suffice to prove the theorem.

We first show that all the elements in M may be expressed as a linear combination in these elements with coefficients from K: Let $\gamma \in M$. Since $\beta_1, \ldots, \beta_n$ is a base for M over L, we have

$$\gamma = \delta_1\beta_1 + \cdots + \delta_n\beta_n,$$

where $\delta_1, \ldots, \delta_n \in L$. Since $\alpha_1, \ldots, \alpha_m$ is a base for L over K, we have $a_{i,j} \in K$ such that

$$\begin{aligned}
\delta_1 &= a_{1,1}\alpha_1 + \cdots + a_{1,m}\alpha_m, \\
\delta_2 &= a_{2,1}\alpha_1 + \cdots + a_{2,m}\alpha_m, \\
&\cdots \\
\delta_n &= a_{n,1}\alpha_1 + \cdots + a_{n,m}\alpha_m.
\end{aligned}$$

This gives

$$\begin{aligned}
\gamma = {} & (a_{1,1}\alpha_1 + \cdots + a_{1,m}\alpha_m)\beta_1 \\
& + (a_{2,1}\alpha_1 + \cdots + a_{2,m}\alpha_m)\beta_2 \\
& + (a_{3,1}\alpha_1 + \cdots + a_{3,m}\alpha_m)\beta_3 \\
& + \cdots \\
& + (a_{n,1}\alpha_1 + \cdots + a_{n,m}\alpha_m)\beta_n \\
= {} & \Sigma a_{i,j}\alpha_i\beta_j
\end{aligned}$$

We next show that the elements $\alpha_i\beta_j, 1 \leq i \leq m, 1 \leq j \leq n$, are linearly independent. Assume that we have

$$\Sigma a_{i,j}\alpha_i\beta_j = 0,$$

we shall prove that then all $a_{i,j} = 0$. We obtain that

$$\begin{aligned}
0 = {} & (a_{1,1}\alpha_1 + \cdots + a_{1,m}\alpha_m)\beta_1 \\
& + (a_{2,1}\alpha_1 + \cdots + a_{2,m}\alpha_m)\beta_2 \\
& + (a_{3,1}\alpha_1 + \cdots + a_{3,m}\alpha_m)\beta_3 \\
& + \cdots \\
& + (a_{n,1}\alpha_1 + \cdots + a_{n,m}\alpha_m)\beta_n.
\end{aligned}$$

Since $\beta_1, \ldots, \beta_n$ are linearly independent over L, this gives that

$$\begin{aligned}
a_{1,1}\alpha_1 + \cdots + a_{1,m}\alpha_m &= 0, \\
a_{2,1}\alpha_1 + \cdots + a_{2,m}\alpha_m &= 0, \\
&\cdots \\
a_{n,1}\alpha_1 + \cdots + a_{n,m}\alpha_m &= 0.
\end{aligned}$$

But since $\alpha_1, \ldots, \alpha_m$ are linearly independent over K, it follows that all the coefficients $a_{i,j} = 0$.

This completes the proof. □

Example. We shall look at an example, which will be used later. Let $\alpha = \sqrt[3]{2}$, and let $L = \mathbb{Q}(\alpha)$. Then

$$[L : \mathbb{Q}] = 3.$$

Indeed, we have that $\alpha^3 = 2$, so that α is a root of the equation

$$x^3 - 2 = 0.$$

We prove that $p(x) = x^3 - 2$ is *the minimal polynomial* of $\sqrt[3]{2}$ over $\mathbb{Q}$. Assume that this is not the case. Then the minimal polynomial would have to be a proper factor of $p(x)$, and thus be of degree either 1 or 2. In both cases there would exist rational numbers a, b and c such that

$$x^3 - 2 = (x^2 + ax + b)(x + c).$$

Thus the equation would have a rational root, namely $-c$.

We now use Proposition 25: In fact, according to this proposition any rational root of $x^3 - 2 = 0$ would have to be an *integer* dividing the constant term, and thus the only possibilities would be the numbers $\pm 1, \pm 2$. Since none of these are solutions to the equation, the claim is proved.

This example may be viewed as a special case of a general fact, which we formulate in the proposition below:

Proposition 29. *A polynomial $p(x)$ of degree 3 with coefficients from $\mathbb{Q}$ is irreducible if and only if it has no rational root, in other words there is no number $\alpha \in \mathbb{Q}$ such that $p(\alpha) = 0$.*

Proof. If $p(x)$ factors as a product of two polynomials with coefficients from $\mathbb{Q}$, one of the factors is of degree 1. Thus $p(x)$ has a rational root. For the converse, we use the following:

If an arbitrary polynomial $p(x)$ with coefficients from some number fields K has a root α from K, then $x - \alpha$ divides $p(x)$:

$$p(x) = q(x)(x - a).$$

Indeed, we perform the division of $p(x)$ by $x - \alpha$ and get a quotient and a remainder,

$$p(x) = q(x)(x - a) + r,$$

where the remainder r is a number in K. But since α is a root in $p(x)$, we get $p(\alpha) = r = 0$. This completes the proof. □

Chapter 17
Constructions with Straightedge and Compass

17.1 Review of Legal Constructions

The usual meaning of the term *construction* throughout the history of Geometry, is to *draw a figure*, usually in the plane, such that the figure possesses certain properties specified a priori. In doing so one is required to start out from a given set of points, in some cases also certain fixed curves. This is referred to as the *start data* for the construction. One is required to use only certain tools, which have been specified as allowable, and to use them in certain prescribed ways only. Normally one is allowed only a finite number of steps in the construction. If an infinite number of steps is needed, then we speak of an *asymptotic* version of the construction. To carry out such an asymptotic construction is, of course, not humanly possible. But the method may be used to create constructions which approximate the required one arbitrarily well.

Recall from Sect. 3.6 that when we speak of *constructions by straightedge and compass*, or constructions by the *Euclidian tools*, we mean the following: Starting from a finite number of points, a construction of a figure with the required properties should be achieved using straightedge and compass only, in such a way that:

(1) The straightedge may be used to draw a line through two different points which are given or already have been constructed, and this line may be produced arbitrarily in both directions.
(2) The compass may be used to draw a circle with center in a point which is given or already has been constructed, passing through another point which is given or already has been constructed.

In Chap. 3 we have seen how Greek geometers used more powerful tools than the Euclidian ones to solve the classical problems. This included devising mechanical instruments which could perform needed constructions like finding *a double mean proportionality*, required in doubling a cube as explained in Sect. 3.8. These gadgets were scorned by the purists, as representing a despicable *mechanization* of geometry. Such purists were more approving when various curves, including conic sections, were used as start data in constructions which solved these problems. And, of course, Archytas' space-geometric construction was praised as being one of the

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_17,

high points of Greek geometry. The gadgets referred to *are* interesting, and they do indeed form part of our geometric heritage. We have treated some of them in the appropriate sections.

Another direction taken by people interested in constructions has been *not to strengthen* the tools, but *to weaken them*, at least prescribe tools which are *apparently weaker* than the Euclidian ones.

Thus for instance, around 980 the Arabian geometer *Abul Wafa* had the idea of performing constructions by means of a straightedge and a *rusty compass*, that is to say a compass with which one is allowed to draw circles with a radius fixed once and for all. It may be surprising that using only such a deficient compass in addition to the straightedge, one may still perform *all* constructions which are possible by the Euclidian tools. But this is not the end: In fact, in 1822 the French mathematician *Jean Victor Poncelet* (1788–1867) found the remarkable result that with *one circle* of any fixed radius added to the start data, a straightedge suffices to perform all constructions possible by the Euclidian tools! The rusty compass needs to be used only once, and may then be discarded.

Such questions still beckon mathematicians and students. This is manifest by all the papers continuing to appear on similar subjects. Thus for instance the interesting and readable paper [64] by *Peter Y. Woo* of the University of Hong Kong, shows by elementary geometric means that with start data including a fixed parabola, with its focus and directrix, all Euclidian constructions may be performed with a straightedge.

Finally, it has been shown that all points obtainable by Euclidian tools can also be obtained by compass alone. This result is frequently credited to the Italian geometer *Lorenzo Mascheroni* (1750–1800), but had actually been discovered 125 years earlier by the Danish mathematician Georg Mohr (1640–1697).

We shall return to some of these findings when we have developed the powerful algebraic tools which can settle these questions with one stroke.

There is an entirely different notion of *"construction"*, namely that of *Construction by Folding*. Then we draw no line between two points A and B. Instead, we simply *fold the paper*, thus producing the line as the resulting straight indentation in the paper. We return to this in Sect. 17.8.

17.2 Constructible Points

A point is *constructible* if it is one of the points of the start data, or a point of intersection between two constructible lines or circles.

In Sect. 3.6 we showed that we are allowed to use the compass in the following way:

(3) If A, B and C are any three points which are given or have already been constructed, then we may draw a circle through A with radius BC.

This is the assertion that in the presence of the straightedge, the Euclidian compass is equivalent to the modern compass.

As we saw in Sect. 3.9, we may solve the *Verging problem* to which the Trisection Problem may be reduced by the use of a *marked straightedge.* This means that we allow ourselves to move a distance, not only by means of the compass, but also by means of the straightedge, and to *insert* the distance between any lines or circles in the construction, while at the same time having the straightedge pass through, or *verge towards*, a suitable point:

(4) Two points, the distance between which are equal to the distance between two given or already constructed points may be marked on the straightedge and lines may be drawn through a point which is given or already constructed in such a way that the two marked points on the straightedge fall on constructed lines or circles.

In the presence of the Euclidian or the modern compass this is equivalent to the *Rule of a (fixed) Marked Straightedge:* There is a *fixed distance* marked off on the straightedge, with which the operation (4) above is allowed. We return to this in Sect. 17.8.

17.3 What is Possible?

We are now going to determine *which points one may construct* by the procedures (1) and (2) above, or equivalently, by (1) and (3), starting from a given set of points $P_0, P_1, \ldots, P_n$.

We introduce a coordinate system in the plane by taking the origin in P_0, and letting the x-axis pass through P_1. Moreover, we chose the scale so that such that $P_1 = (1, 0)$. Put $P_i = (a_i, b_i)$, for $i = 1, \ldots, n$. Let $Q = (a, b)$ be a point which is constructed in *one operation* from the points $P_0, \ldots, P_n$. If it is operation (1) which has been employed, then Q is the intersection between two lines, $P_i P_j$ produced and $P_k P_\ell$ produced. This means that $x = a$ and $y = b$ is a solution of the following system of equations

$$(a_i - a_j)(y - b_j) = (b_i - b_j)(x - a_j)$$
$$(a_k - a_\ell)(y - b_\ell) = (b_k - b_\ell)(x - a_\ell).$$

Clearly this yields a and b which are *rational expressions* in the given coordinates, thus that a and b may be computed from a_i, a_j, a_k, a_ℓ, b_i, b_j, b_k and b_ℓ by repeated use of addition, subtraction, multiplication and division.

Next assume that it is operation (3) which is employed. Then Q will either be a point of intersection between a line $P_i P_j$ produced and a circle about P_k with radius r equal to the distance between two points P_ℓ and P_m, or between two such circles.

In the former case this means that $x = a$ and $y = b$ are solutions of the system of equations

$$(a_i - a_j)(y - b_j) = (b_i - b_j)(x - a_j)$$
$$(x - a_k)^2 + (y - b_k)^2 = r^2$$

where $r^2 = (a_m - a_\ell)^2 + (b_m - b_\ell)^2$. Here we see that a and b may be expressed by the a_i and b_i, $i = 1, \ldots, n$ by repeated use of the operations of addition, subtraction, multiplication, division *and square root*.

In the latter case Q is a point of intersection between two circles, centered at two of the points and with radii equal to distances between two pairs of points among the $P_0, \ldots, P_n$. Then $x = a$ and $y = b$ are solutions of

$$(x - a_j)^2 + (y - b_j)^2 = r_1^2$$
$$(x - a_k)^2 + (y - b_k)^2 = r_2^2$$

where r_1 and r_2 are distances between two pairs of points as asserted above. This system may be simplified to one in which there is only one equation of degree two: Multiplying out, we obtain

$$x^2 - 2a_j x + y^2 - 2b_j ay = r_1^2 - a_j^2 - b_j^2$$
$$x^2 - 2a_k x + y^2 - 2b_k y = r_2^2 - a_k^2 - b_k^2$$

which is equivalent to the system

$$x^2 - 2a_j x + y^2 - 2b_j y = r_1^2 - a_j^2 - b_j^2$$
$$(a_k - a_j)x + (b_k - b_j)y = \tfrac{1}{2}(r_1^2 - r_2^2 + a_k^2 + b_k^2 - a_j^2 - b_j^2).$$

Again we get that a and b may be computed from $a_i, a_j, a_k, a_\ell, b_i, b_j, b_k$ and b_ℓ by repeated use of addition, subtraction, multiplication, division *and square root*. We have shown the following:

Proposition 30. *If $Q = (a, b)$ may be constructed from $P_0, \ldots, P_n$ in one step by the operations (1) or (3), then a and b may be computed from the coordinates of these points by repeated use of addition, subtraction, multiplication, division and square root.*

We next proceed one step further, and ask which points we may construct in a finite number of steps, at each stage including the previously constructed points among the allowable ones. We shall say that the real number *a is constructible* if the point $(a, 0)$ is constructible. Clearly the point (a, b) is constructible if and only if a and b are constructible real numbers.

Moreover, it is clear that if a is constructible in *one step* from $P_0, \ldots, P_n$ by the operations (1) and (3), then we may express a by $a_1, \ldots, a_n$, $b_1, \ldots, b_n$, using operations $+, -, \cdot, :$, and $\sqrt{}$.

Actually, if a is constructible in one step, then we get at most *simple* $\sqrt{}$*'s*: there are no expressions of the type

$$\sqrt{\alpha + \sqrt{\beta}}.$$

We say that a may be expressed rationally by $a_1, \ldots, a_n, b_1, \ldots, b_n$ and simple $\sqrt{\ }$'s.

But if a were constructible in *two steps*, then we would get such double root-expressions as well: Indeed, then there would be a constructible point $P_{n+1} = (a_{n+1}, b_{n+1})$ such that a would be expressed in $a_1, \ldots, a_n, a_{n+1}, b_1, \ldots, b_n, b_{n+1}$ using operations $+, -, \cdot, :$, and simple $\sqrt{\ }$'s, while a_{n+1} and b_{n+1} would be expressed by $a_1, \ldots, a_n, b_1, \ldots, b_n$ using operations $+, -, \cdot$, and simple $\sqrt{\ }$'s. When the latter expressions are substituted into the former, we obtain a, expressed by $a_1, \ldots, a_n, b_1, \ldots, b_n$ using operations $+, -, \cdot, :$, as well as single *and double* $\sqrt{\ }$'s.

Continuing like this, we obtain that if the number a is constructible from the given start data above, then it may be expressed in $a_1, \ldots, a_n, b_1, \ldots, b_n$ using operations $+, -, \cdot$, and *multiple* $\sqrt{\ }$'s.

We have the following key result. Note that the numbers $a_1, \ldots, a_n, b_1, \ldots, b_n$ given by the start data of course are constructible:

Theorem 36. *a is constructible from the start data $P_0, \ldots, P_n$ given above, with $a_1, \ldots, a_n, b_1, \ldots, b_n$ as above, if and only a may be expressed by*

$$a_1, \ldots, a_n, b_1, \ldots, b_n$$

using operations $+, -, \cdot, :$, and multiple *$\sqrt{\ }$'s.*

Proof. It only remains to prove that the criterion implies constructibility. For this, it suffices to show the following

Proposition 31. *Assume that α and β are constructible. Then so are $\alpha + \beta$, $\alpha - \beta$, $\frac{\alpha}{\beta}$ (if $\beta \neq 0$), $\alpha\beta$ and $\sqrt{\alpha}$ (if $\alpha \geq 0$).*

We shall first see that this proposition suffices to prove the theorem. The method of proof is best illuminated by an example. Assume that $n = 3$ and that a looks like this:

$$a = \frac{a_2 + b_1\sqrt{b_2 - \sqrt{a_3 + b_3}}}{b_2 - b_3\sqrt{b_1 + \sqrt{a_1 - b_1}}}.$$

By the proposition the number $a_4 = a_3 + b_3$ is constructible, and so is $b_4 = a_1 - b_1$. Then we also have that $a_5 = \sqrt{a_4}$ and $b_5 = \sqrt{b_4}$ are constructible. We now have

$$a = \frac{a_2 + b_1\sqrt{b_2 - a_5}}{b_2 - b_3\sqrt{b_1 + b_5}}$$

where all the numbers which occur on the right hand side are constructible. Furthermore, we find that $a_6 = \sqrt{b_2 - a_5}$ and $b_6 = \sqrt{b_1 + b_5}$ are constructible, so that

$$a = \frac{a_2 + b_1 a_6}{b_2 - b_3 b_6}$$

where again the numbers occurring on the right hand side are constructible. The proposition again implies that $a_7 = b_1 a_6$ and $b_7 = b_3 b_6$ are constructible, and we have

$$a = \frac{a_2 + a_7}{b_2 - b_7} = \frac{a_8}{b_8}$$

where the numbers a_i and b_i involved are all constructible. Thus, with a final use of the proposition, a is constructible. This completes the example.

To give a formal proof, we proceed by induction. In order to make the induction work properly, we need to phrase the statement $P(N)$, which is to be proven by induction, a little carefully. We formulate the following:

> P(N): Assume that the real number a may be expressed in terms of constructible numbers by the operations $+, -, \cdot, :$, and multiple $\sqrt{}$'s. Assume that the total number of operations $+, -, \cdot, :$, and $\sqrt{}$'s which are needed is less than or equal to the natural number N. Then a is constructible for all N.

To prove this assertion we proceed by induction on the number N.

If $N = 0$, then a itself is constructible and there is nothing to prove. Assume the claim for all numbers b which may be expressed by $\leq N - 1$ operations in some constructible numbers.

As in the example we now go to one of the innermost expressions in the formula giving a in terms of some constructible numbers, say $a_1, \ldots, a_m$. This innermost expression must be of the simple kind covered by the proposition, otherwise it would not be innermost, hence it is constructible by the proposition. Thus including this number in the set of the $a_1, \ldots, a_m$, labelling it a_{m+1}, we now have a expressed in terms of $a_1, \ldots, a_m, a_{m+1}$ with $\leq N - 1$ operations. This reduction in the number of operations has been achieved at the expense of increasing m, the size of the set of constructible numbers involved in the expression for a, but that is OK: It is covered by our induction assumption, the statement P(N) says nothing about the size of the set of constructible numbers involved. So now we may take a as b above, and the claim follows by the induction assumption.

Proof of the proposition. With the normalizing assumptions we have made, $P_0 = (0, 0)$ in our coordinate system and $P_1 = (0, 1)$. Thus the x-axis as well as the y-axis are constructible lines. In particular 0 and 1 are constructible numbers.

We now refer to Fig. 17.1. We first prove that $\alpha + \beta$ and $\alpha - \beta$ are constructible. This is shown in the upper left construction: We use Rule (3), and draw the circle about O with radius α. It intersects the positive x-axis in the point A. About A we then draw the circle with radius β, finding the points B and C. Then $OC = \alpha + \beta$ and $OB = \alpha - \beta$. It follows in particular that all integers are constructible.

Next, the upper right construction shows that $h = \frac{\alpha}{\beta}$ is constructible: Using Rule (3), we set off A and B on the y-axis, and C on the x-axis, so that $OA = \alpha$, $OB = \beta$ and $OC = 1$. Constructing the line through A parallel to the line BC, we find the point of intersection with the x-axis, D. From the similar triangles $\triangle BOC \sim \triangle AOD$ we then find that

$$\frac{OD}{1} = \frac{\alpha}{\beta}$$

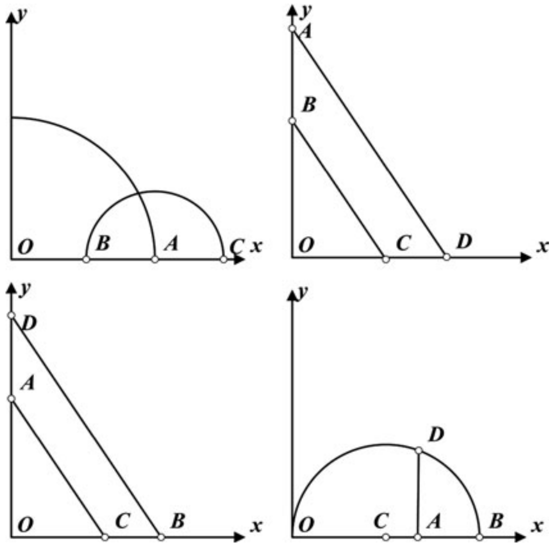


Fig. 17.1 The constructions used in the proof of Proposition 31

hence $OD = h$, so that h is constructible.

Proving that $h = \alpha\beta$ is constructible follows the same lines, the construction is shown in the lower left corner of Fig. 17.1. We now mark A on the y-axis such that $OA = \alpha$, and B and C on the x-axis such that $OB = \beta$ and $OC = 1$. Through B we then draw the parallel to the line AC, it intersects the y-axis in the point D. From the similar triangles we find

$$\frac{OD}{\alpha} = \frac{\beta}{1}$$

and hence $h = OD$ is constructible.

We finally prove that $h = \sqrt{\alpha}$ is constructible. We have

$$h^2 = \alpha \text{ or } \frac{h}{\alpha} = \frac{1}{h}.$$

The construction is given in the lower right corner of Fig. 17.1. We set off α along the x-axis from O, finding A, and 1 from A further along the x-axis finding B. Bisecting OB we find C, about C we draw the circle through O, which of course also passes through B. Now the normal to the x-axis erected at A intersects the circle in D, and we then have the similar triangles $\triangle ODA \sim \triangle DBA$. From this it follows that AD is the mean proportional between OA and AB, whence

$$AD : 1 = OA : AD$$

thus $AD^2 = \alpha$, so $AD = \sqrt{\alpha}$ is constructible. □

The theorem which we have just proved, has the following consequence which is the key to deciding which constructions we can perform by Euclidian tools:

Theorem 37. *Assume that the real number α is constructible, and let $f(z)$ denote the minimal polynomial of α over the number field K generated by the coordinates $a_1, \ldots, a_n, b_1, \ldots, b_n$ of the points of the start data, namely the points $P_1, \ldots, P_n$. Then $f(z)$ is of degree 2^m for a suitable integer $m \geq 0$.*

Proof. By Theorem 36 we may express α by $a_1, \ldots, a_n$, $b_1, \ldots, b_n$ and operations $+$, $-$, $\cdot$ and $:$ as well as $\sqrt{}$. Thus α is contained in a field extension L of K, see Sect. 16.5, which we may obtain by a finite number of field extensions

$$K = K_0 \subset K_1 \subset K_2 \subset \cdots \subset K_{M-1} \subset K_M = L$$

where K_i comes from K_{i-1} by adjoining to K_{i-1} *one square root* of an element from K_{i-1}. By Theorem 35 we therefore obtain that

$$[L : K] = 2^M,$$

since $[K_i : K_{i-1}] = 2$, by Theorem 34. But this means that

$$2^M = [L : K] = [L : K(\alpha)][K(\alpha) : K],$$

and since $[K(\alpha) : K]$ is the degree of the minimal polynomial of α, again by Theorem 34, this degree as well must be a power of 2, and the proof is complete. □

17.4 Trisecting Any Angle

The assignment is to divide an arbitrary angle in three equal parts, using compass and straightedge according to the Rules (1) and (2) from Sect. 17.1, or equivalently using (1) and (3) from Sect. 17.2. It may be explained as follows:

Let u be an arbitrary angle, and let $v = \frac{u}{3}$. The angle u is given, in the form of *three points* P_0, P_1 and P_2. These three points represent our start data, and we choose a coordinate system in the plane $\mathbb{R}$ in such a way that

$$P_0 = (0, 0), P_1 = (1, 0), P_2 = (\alpha, 0),$$

where $\alpha = \cos(u) \in [-1, 1]$. Now we have

$$\alpha = \cos(3v) = 4\cos^3(v) - 3\cos(v),$$

$\cos(v)$ is thus root in the equation

$$4x^3 - 3x - \alpha = 0.$$

This equation has three real roots. In fact this is clear from the way we found it: Namely, letting

$$v_0 = v, v_1 = v_0 + \frac{2\pi}{3}, v_2 = v_0 + \frac{4\pi}{3},$$

we have

$$3v_0 = u, 3v_1 = u + 2\pi, 3v_2 = u + 4\pi,$$

thus the three real roots are

$$\cos(v_0), \cos(v_1), \cos(v_2).$$

It is clear that if we are able to construct one of these roots, then we may construct the remaining two as well. Therefore we have reduced the problem of trisecting u in three equal parts using Euclidian tools to the following:

Problem. Given start data P_0, P_1 and P_2 as above, so $P_0 = (0,0)$, $P_1 = (1,0)$ and $P_2 = (\alpha, 0)$, where $\alpha = \cos(u) \in [-1, 1]$. Then construct, with Euclidian tools, one root of

$$4x^3 - 3x - \alpha = 0.$$

For certain values of u or α this is certainly possible: For example if $u = \frac{\pi}{2}$, such that $\alpha = 0$, then the equation becomes

$$4x^3 - 3x = 0$$

which gives

$$x = 0, x = \pm\frac{1}{2}\sqrt{3},$$

and these roots are all constructible from our start data, which in this case are just

$$P_0 = (0,0), P_1 = (1,0).$$

But for most choices of α the roots will not be constructible from the given start data $P_0 = (0,0), P_1 = (1,0), P_2 = (\alpha, 0)$. Thus for instance, take $\alpha = \frac{1}{4}$. Then the equation becomes

$$4x^3 - 3x - \frac{1}{4}.$$

Putting $z = 4x$, we get the equation

$$p(z) = z^3 - 12z - 4 = 0.$$

It suffices to show that no root of this equation is constructible.

Let z_0 denote a root in the equation, and let $m(z)$ denote the minimal polynomial of z_0 over $\mathbb{Q}$. Let g denote the degree of $m(z)$. Then clearly $g \leq 3$, we shall prove that $g = 3$.

Suppose that $m(z)$ is of degree 2. Then Theorem 29 implies that

$$p(z) = q(z)m(z),$$

where $q(z)$ must be of degree 1. If $g = 1$, then we find in the same manner that

$$p(z) = q(z)m(z),$$

where now $m(z)$ is of degree 1 while $q(z)$ is of degree 2. In either case we see that $p(z)$ must have a *rational root*, namely the root given by the factor of degree 1. But by Proposition 25 we have that such a root must be one of the integers $\pm 4, \pm 2, \pm 1$. But none of these integers are roots in the equation $p(z) = 0$, and thus the claim that $g = 3$ is proved. By the powerful Theorem 37 this of course suffices to settle the question: z_0 *is not constructible*, and thus the first of the famous classical problems is settled in the negative: The trisection of any angle in equal parts by Euclidian tools is impossible.

17.5 Doubling the Cube and Constructing the Regular Heptagon

Having developed the tools needed to settle the trisection problem, it is simple to decide the problem of *Doubling the Cube* by Euclidian tools as well. The problem now consists in deciding whether or not the number $z_0 = \sqrt[3]{2}$ is constructible. This number is a root of the equation

$$p(z) = z^3 - 2 = 0.$$

This polynomial is also the *minimal polynomial* for z_0 over $\mathbb{Q}$: If not, then it would be possible to factorize $p(z)$ as a product of a rational polynomial of degree 1 and another of degree 2. But as we saw in the example at the end of Sect. 16.5 is this not possible.

Thus again, the second one of the classical problems is settled in the negative: A cube may not be doubled by a construction using Euclidian tools.

In Sect. 4.4, we found that the regular heptagon may be constructed with Euclidian tools if and only if a root in the equation

$$X^3 - 2X^2 - X + 1 = 0$$

may be constructed. But since this equation has no rational roots, it follows in the same way as for doubling the cube that this construction is not possible.

17.6 Squaring the Circle

We may decide the question of *Squaring the Circle* as well. A circle of radius r have area $A = \pi r^2$. A square with the same area must, therefore, have side equal to $a = r\sqrt{\pi}$. The problem then is to decide if the number $\alpha = \sqrt{\pi}$ is constructible from the start data $P_0 = (0, 0)$ and $P_1 = (1, 0)$. But it is known that the number π is *transcendental*, in other words it does not satisfy any polynomial equation with coefficients from $\mathbb{Q}$. This result is not so easy to prove, but nevertheless quite within reach of a reasonably complete course in calculus at the college level. We have already quoted this fact, in Sect. 16.4. Then the number $\alpha = \sqrt{\pi}$ can also not be root in an algebraic equation, since that would imply

$$[\mathbb{Q}(\sqrt{\pi}) : \mathbb{Q}] = m < \infty$$

by Theorem 34. On the other hand we have

$$[\mathbb{Q}(\pi) : \mathbb{Q}(\sqrt{\pi})] = 2,$$

so that Theorem 35 implies that

$$[\mathbb{Q}(\pi) : \mathbb{Q}] = 2m < \infty.$$

This of course contradicts that π is a transcendental number, and we have shown that the Squaring of the Circle is impossible by Euclidian tools.

This proof hinges on the fact that the number π is transcendental. This was shown by the German mathematician *Carl Louis Ferdinand von Lindemann* (1852–1939) in 1882, the year before he joined *Hurwitz*[1] and *Hilbert* as professor at the University of *Königsberg*, now the Russian enclave named *Kaliningrad*. Lindemann gave a remarkable proof of the fact that π is transcendental, tying it to the much simpler fact that the number e, base of the natural logarithms, is transcendental, and heavily using complex numbers.

In 1873 the French mathematician *Charles Hermite* (1822–1901) proved that the number e is transcendental. His proof was long and difficult, but soon simpler proofs were found. In fact, in 1893 the journal *Mathematische Annalen* contained three different and simpler proofs of the transcendency of e. One of them was due to Hilbert, another was by Hurwitz. Hilbert also gave a proof of the transcendency of π in the same journal. Hilbert's very elegant proof uses basic calculus, and the computations seem to work by pure magic. We refer to [42] for his proof.

Now, in 1882 von Lindemann had shown that the assertion that e is transcendental may be generalized to the following result:

[1] The German mathematician Adolph Hurwitz (1859–1919) had been a student of Felix Klein, and taught 8 years at Königsberg.

Theorem 38 (Lindemann's Theorem). *Assume that $b_0, \dots, b_n$ are distinct real or complex algebraic numbers, and let $a_0, \dots, a_n$ be real or complex algebraic numbers, not all zero. Then*

$$a_0 e^{b_0} + \cdots + a_n e^{b_n} \neq 0.$$

Again we are not going to give the proof of this theorem, but shall use it to prove the transcendency π. Assume that π *were not* transcendental, in other words that it were algebraic. Then we get a contradiction by taking $a_0 = 1, a_1 = -1$, and $b_0 = i\pi, b_1 = 0$: Indeed, we have the famous formula, due to *Euler*, which we have proved in Sect. 11.4. It ties together the five most important constants in mathematics, namely $0, 1, e, i = \sqrt{-1}$ and π:

$$e^{i\pi} + 1 = 0.$$

17.7 Regular Polygons

The last construction problem which we shall treat, is that of subdividing the circumference of a circle in n equal parts. It amounts to the same as to construct the *regular n-gon*. In Fig. 17.2 we see the subdivision in n equal parts for $n = 3, 4, 5, 6, 8, 10$.

The construction, by Euclidian tools, of regular n-gons were known for several values of n by Greek geometers. We shall treat these constructions as that of *subdividing the circumference of a given circle* in n equal parts. Thus our start data is two points, A and B, the aim is to subdivide the circle about A through B in n equal parts.

Some of these constructions are quite simple, of course. Thus the regular 6-gon is constructed by drawing the circle about A through B, then drawing a circle about B through A, finding C and D, and so on, as shown in Fig. 17.3. The construction leads to a beautiful ornament.

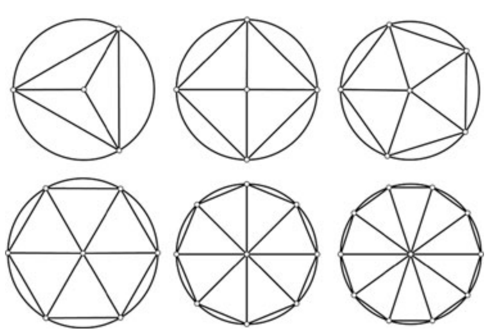


Fig. 17.2 The circle divided into 3,4,5,6,8 and 10 equal parts

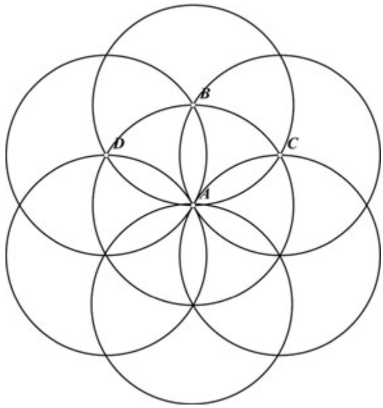


Fig. 17.3 The beautiful ornament of the construction of the regular 6-gon

The regular 3-gon, or triangle, may be obtained by choosing every second point of the regular 6-gon, the regular 4-gon, or the *square* by erecting the normal to AB at A, AB produced and this normal produced then intersect the circle in the points dividing its circumference in four equal parts.

In general it is clear that once we have subdivided the circumference of the circle in n_0 equal parts, then we obtain the subdivision into $2n_0$ equal parts simply by bisecting all the angles coming from the n_0-subdivision. The procedure may be continued, and we find that once the n_0-subdivision has been achieved, then the $n = 2^r n_0$-subdivision follows for all positive integers r.

For the 3-subdivision of the circle we found, however, that the simplest construction is to perform the 6-subdivision first, and then take every second point to get the 3-subdivision.

This is also the case for the construction of subdividing the circumference into 5 equal parts: The simplest and fastest construction is to find the 10-subdivision first.

The construction of the regular pentagon is one of the high points of Euclid's Elements. The construction is given in Book IV, Proposition 11. We shall not examine Euclid's treatment, as before we present the themes in a modern setting. For an explanation of Euclid's arguments, we refer to Hartshorne's beautiful book [25]. However, the construction we give here contains the same ideas as Euclid's version.

The construction hinges on the construction known as *the Golden Section*: To divide the line segment AB by the point G such that

$$AG/GB = AB/AG.$$

If we denote the length of AB by r, and the length AG by x, then we get

$$\frac{x}{r-x} = \frac{r}{x},$$

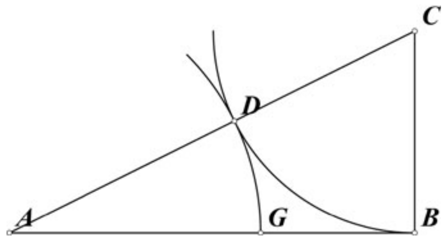


Fig. 17.4 The construction of the Golden Section of *AB*. The length of *AB* is r, and that of *BC* is $\frac{r}{2}$. Thus the length of *AG* is $\frac{r}{2}(\sqrt{5}-1)$, and G is the point of the Golden Section of *AB*

or

$$x^2 + rx - r^2 = 0$$

and solving this equation we obtain

$$x = \frac{r}{2}(-1 + \sqrt{5}).$$

Actually the term *Golden Ratio* or *Golden Mean* usually refers to the ratio $\frac{x}{r-x}$ above, which is

$$\frac{x}{r-x} = \frac{-1+\sqrt{5}}{3-\sqrt{5}} = \frac{1}{2}(1+\sqrt{5}).$$

Now a Golden Section for the segment AB may be constructed by Euclidian tools by a very simple construction. We refer to Fig. 17.4, where the construction of the point G such that $AG/GB = AB/AG$ is explained.

In Fig. 17.5 we show how the construction of the Golden Section is carried out in Euclid's Elements. The square *ABCD* has side $AB = r$. E is the mid point of *DA*, about E a circle through B is drawn, it intersects *DA* produced in F, the square *AGHF* is constructed. Then it is shown, in the Elements, that the latter square has area equal to the rectangle *GBCI*. Thus, in our present day notation, $AG^2 = AB(AB - AG)$, from which it follows that G is the Golden Section of *AB*. Of course we easily verify the claim about the areas today, by computing as we did using Fig. 17.4.

To subdivide the circumference of the circle about B through A in 10 equal parts, we observe that if the side of the inscribed 10-gon is set off from A along *AB*, then we find the Golden Section G of the radius *AB*. Using this, all we have to do is to find the Golden Section of the radius. See Fig. 17.6.

To see that the construction is correct, we look at the triangle *BPQ*. The angle at B is 36°, hence the angles at P and Q are both 72°. Bisecting the angle at P, we find the point R, and now have $\triangle BPQ \sim \triangle PQR$. Putting $PQ = x$, we then have $RB = x$ and thus from the similar triangles that

$$\frac{r}{x} = \frac{x}{r-x}$$

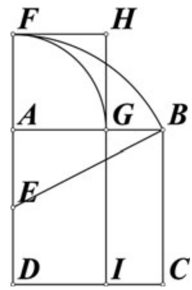


Fig. 17.5 The construction of the Golden Section of *AB* according to Euclid

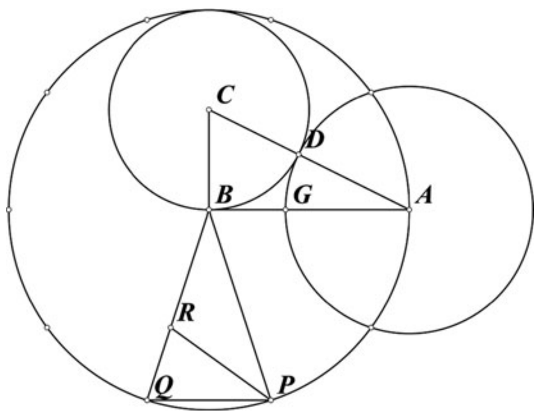


Fig. 17.6 The construction of the regular 10-gon, inscribed in the circle of center B through A: The Golden Section of the radius *BA*, which is of length r, is carried out by erecting *BC*, normal to *BA* at B of length $\frac{1}{2}r$. The circle about C through B intersects *CA* at D, the circle about A through D intersects *AB* at G, the Golden Section. We now find the subdivision of the circumference of the circle about B through A by setting off *AG* as chords around the circle. The triangle BPQ is used in the text to show that the construction is correct

and the claim follows. We note that this assertion is XIII.9, Proposition 9 in Euclid's Elements, Book XIII. It is equivalent to the following formula for s_{10}, the side of the regular 10-gon inscribed in a circle of radius r:

$$s_{10} = \frac{r}{2}(-1 + \sqrt{5}).$$

Moreover, the assertion Euclid XIII.10, which is equivalent to

$$s_5^2 = r^2 + s_{10}^2,$$

may be shown as follows, in modern language: Letting q denote the height from B to $s_{10} = QP$ we find

$$q^2 = r^2 - \frac{r^2}{16}(-1 + \sqrt{5})^2 = \frac{r^2}{16}(10 + 2\sqrt{5})$$

thus

$$q = \frac{r}{4}\sqrt{10 + 2\sqrt{5}}.$$

A simple consideration of similar triangles then yields

$$\frac{\frac{s_5}{2}}{q} = \frac{s_{10}}{r}$$

i.e.,

$$s_5 = \frac{2}{r} q s_{10}$$

which after a short computation yields a formula for the side of the regular pentagon inscribed in a circle of radius r,

$$s_5 = \frac{r}{2}\sqrt{10 - 2\sqrt{5}}.$$

We finally make the following observation: If we may subdivide the circumference of a circle in n_1 and in n_2 equal parts, by separate constructions, and if there is no integer greater than 1 which divides both n_1 and n_2 , then we may subdivide the circumference in $n_1 n_2$ equal parts. In fact, we carry out the two subdivisions of the same circle, starting at the same point A. Since there is no common factor other than 1 in n_1 and n_2, we get altogether $n_1 + n_2 - 1$ points, spaced at different distances around the circle. But now find two adjacent points where this distance is minimal. Draw the cord connecting them. Setting this minimal chord off around the circumference, we get the subdivision in $n_1 n_2$ equal parts. In Fig. 17.7 we carry this out to find the regular $15 = 5 \cdot 3$-gon.

We may sum up what we have learned so far by the following list of numbers n for which the regular n-gon may be constructed:

$$3, 4, 5, 6, 8, 10, 12, 15, 16, 20, \ldots$$

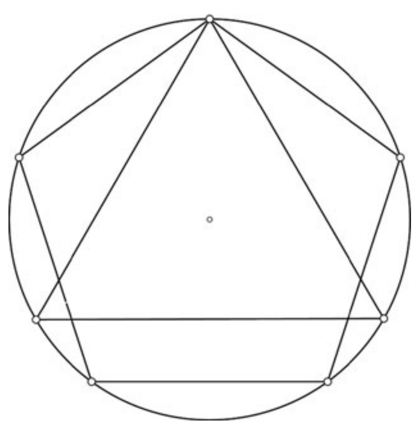


Fig. 17.7 The subdivision of the circumference of a circle in 15 equal parts, thereby constructing the regular 15-gon

The first positive integer not on the list is 7. In Sect. 4.4 we have given Archimedes' construction of the regular 7-gon, the regular *heptagon*, by a *verging construction*. We asserted there that the construction is not possible by legal use of compass and straightedge, in other words by the Euclidian tools. We shall now prove this.

Our start data are the points $P_0 = (0, 0)$ and $P_1 = (1, 0)$. The assignment is to construct the angle $\frac{2\pi}{7}$, which evidently is equivalent to constructing the angle $\frac{\pi}{14}$. Letting $v = \frac{\pi}{14}$, we find $4v + 3v = \frac{\pi}{2}$, thus

$$\cos(4v) = \sin(3v).$$

The formula

$$\sin(3v) = 3\sin(v) - 4\sin^3(v)$$

yields, together with the formula

$$\cos(4v) = 1 - 8\sin^2(v) + 8\sin^4(v)$$

that $a = \sin(v)$ is a root in the equation

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

This polynomial is divisible by $x + 1$, and we get that a is a root in the equation

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

Letting $y = 2x$ we get

$$y^3 - y^2 - 2y + 1 = 0.$$

But this polynomial is *irreducible* by Proposition 25: Indeed, suppose that it could be factored, then one of the factors would have to be linear, thus the polynomial would have a rational root. But by the proposition, the only possibilities are ± 1, and none of them fits the equation. Therefore the minimal polynomial of $b = 2a$ over $\mathbb{Q}$ is

$$p(y) = y^3 - y^2 - 2y + 1$$

and so b is not constructible by Theorem 37, hence a is not constructible and the claim is proven.

In this way one could go on, and resolve the question for one integer n at the time. But *Gauss* realized how all the equations one obtains in this way may be given a unified treatment, and proved the following remarkable result:

Theorem 39. *The regular n-gon may be constructed by Euclidian tools if and only if*

$$n = 2^r p_1 p_2 \cdots p_s,$$

where $p_1 < p_2 < \ldots < p_s$ are prime numbers which are all of the form

$$p = 2^m + 1,$$

and where $r \geq 0$.

We see that $m = 1$ gives $p = 3$, $m = 2$ gives $p = 5$, $m = 3$ gives $p = 9$ which is not a prime number. The regular 9-gon can of course not be constructed with Euclidian tools by the theorem. This also follows by the results we obtained for the *Trisection Problem*: In fact, in the situation of Sect. 17.4, let $u = 60°$. The problem is equivalent to trisecting u. Then $\alpha = \frac{1}{2}$, and the equation from Sect. 17.4 becomes

$$4x^3 - 3x - \frac{1}{2}.$$

Letting $y = 2x$, we get the equation

$$y^3 - 3y - 1.$$

Exactly as in the case of the equation coming from the regular heptagon, we see that this equation has no rational roots, hence is irreducible, and thus its root 2α cannot be constructed by the Euclidian tools, so neither can α.

But $m = 4$ gives $p = 17$, which is the next constructible case. The construction is carried out in Hartshorne's book [25].

So the problem is reduced to the study of prime numbers of the form $2^m + 1$, the so called *Fermat–primes*. Unfortunately we only know of *five* such primes, and we do not even know if the total set of such prime numbers is finite. We do know this, however:

First of all we must have $m = 2^\nu$ for $2^m + 1$ to be a prime. In fact if m has an odd factor, then $2^m + 1$ cannot be prime. For assume that $m = m_1 m_2$, where m_2 is odd. The formula

$$1 + k + k^2 + \cdots + k^{m_2 - 1} = \frac{1 - k^{m_2}}{1 - k},$$

yields for $k = -2^{m_1}$ that

$$1 - 2^{m_1} + 2^{2m_1} - \cdots + 2^{(m_2 - 1)m_1} = \frac{1 + 2^{m_1 m_2}}{1 + 2^{m_1}}$$

since m_2 is odd. This yields

$$2^{m_1 m_2} + 1 = (2^{m_1} + 1)(1 - 2^{m_1} + \cdots + 2^{m_1(m_2 - 1)}).$$

Now put $F_\nu = 2^{2^\nu} + 1$. Then

$$F_0 = 3, F_1 = 5, F_2 = 17, F_3 = 257, F_4 = 65{,}537$$

are all primes. But then

$$F_5 = 4294967297$$

is not prime, as it has the factor 641. After F_4 there are no further Fermat primes known. So the problem of which regular n-gons we can construct by the Euclidian tools is still open. But if a new Fermat prime should be found, the actual construction would be way outsider what is humanly possible, as indeed is the case already for the last known case F_4. But the F_3-gon construction is supposed to have been carried out by enthusiasts.

17.8 Constructions by Folding

There is an amusing way of performing "constructions" by *folding the paper*. The lines obtained are thus given as the *fold* left when the paper is flattened after having been folded according to certain rules. This activity is called *Origami*, and may be carried out in various ways. Here we follow the article [52], but the reader may also find more information in [25].

As with compass and straightedge constructions, we start out from a certain set of points, our *start data*. Then we construct new points by either of the following procedures:

A new point is:

(1) A point P of intersection between two previously constructed different lines ℓ_1 and ℓ_2,

or

(2) A point Q obtained by folding the paper along a previously constructed line ℓ, from a previously constructed or given point P. In other words, a new point Q is constructed by reflecting P in ℓ.

Furthermore,

A new line is:

(3) The line obtained by folding the paper along two given or previously constructed points P and Q. In other words, the line PQ produced in both directions.

or

(4) The line ℓ obtained by folding the paper as follows: Two given or previously constructed points P_1 and P_2 are selected, as well as two previously constructed lines ℓ_1 and ℓ_2. Then the paper is folded in such a way that P_1 falls on ℓ_1 and P_2 falls on ℓ_2.

The four basic rules are illustrated in Fig. 17.8.

As for Rule (4), there may be several such lines, and all of them are constructible. There may also, in special cases, be no such line. If P_1 is on ℓ_2 and P_2 is on ℓ_1, then one allowable fold is along the mid normal of P_1P_2. Thus we have

(5) *The mid normal:* For two given or constructed points P and Q, the fold sending P to Q is allowed. In other words, the mid normal of PQ is constructible.

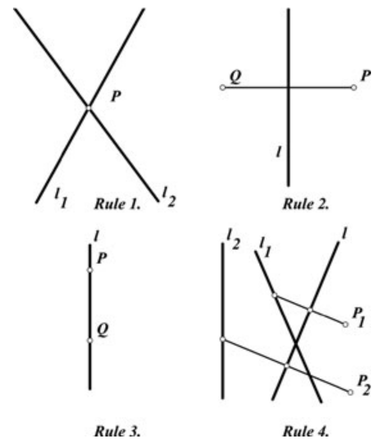


Fig. 17.8 The four basic rules of Paper Folding

Further, we have

(6) *Dropping the normal:* For a given or constructed point P outside a constructed line ℓ, the fold which leaves both P and ℓ fixed. In other words, the normal to ℓ may be dropped from P.

The following rule is some times taken as one of the basic rules:

(7) *The middle line of two lines:* For two different constructed lines ℓ_1 and ℓ_2, it is permitted to fold the paper in such a way that the two lines coincide.

Finally we have the following:

(8) *"Parabolas and circles":* For two points C and F which are given or already have been constructed, and a line d which does not pass through F, the paper may be folded such that C is fixed and F is sent to d.

There are infinitely many such folds, all of them are permitted. The phenomenon is illustrated in Fig. 17.9.

The two allowable folds are *the tangents* passing through C to the parabola with focus F and directrix d. Moreover, the points F' and F'' are the points of intersection between the circle about C passing through F, and the line d. This latter observation is important, since it shows that the folding-rules are at least as powerful as the Euclidian tools in performing constructions.

But the strongest rule is *Rule (4)* The folding lines allowed by this rule are the *common tangents to two different* parabolas. In fact, if we are given a line d and a point F only, then the folds sending F to d are precisely all the tangents to the parabola with focus F and directrix d.

It turns out that Rule (4) is equivalent to working with a *cubic curve*, but we shall not go into the details here. Instead, we refer to the very interesting article [52], already referred to above.[2]

[2] Of course, the article is in Norwegian, but the computation leading to the cubic curve, on p. 175 in the volume of the journal, stands out as clearly legible in any language!

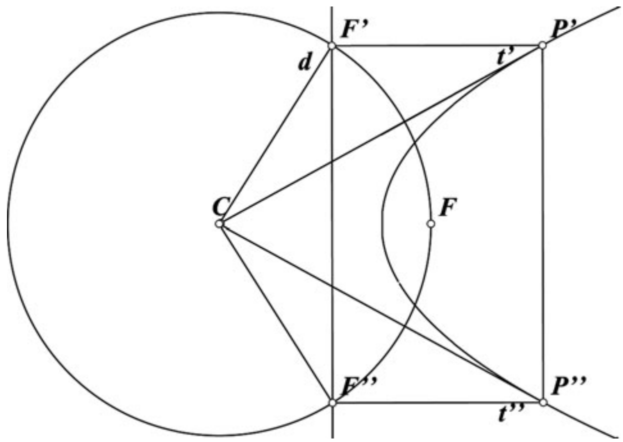


Fig. 17.9 The parabola and the circle provided by Rule 8

The process of construction by Rules (1), (2), (3) and (4) is exactly as powerful as the process of construction with compass and a *marked straightedge*. This is shown in [43, Chap. 10]. Thus any angle may be trisected in equal parts, the regular 7-gon (*heptagon*) may be constructed and the cube may be doubled.

Since we may scale any figure, constructible by Euclidian Tools, up and down as we wish with these tools, we may perform any construction using the straightedge "illegally" to move any distance, by means of a straightedge *on which there is marked off a single, fixed, distance.* Now Nicomedes has achieved this by his Conchoid.

Moreover, adding a single conchoid somewhere as part of the start data, we achieve the same as marking a fixed distance on the straightedge we are using.

Now the folding constructions leads to the appearance of a curve of degree 3. And in fact, it is a conchoid. Hence constructions by folding can achieve all constructions we can carry out by using a marked straightedge. But conversely, the constructions by folding may also be carried out using a marked straightedge. We shall not pursue this issue further, however.

Chapter 18
Fractal Geometry

18.1 Fractals and their Dimensions

Loosely speaking we may say that a *fractal* is an object which is so far from being smooth that its dimension is no longer an integer.

We shall not pursue the issue of defining the concept of *dimension*, however. Instead, we take a relatively naive point of view, and simply examine how the magnification process of selected pieces of a figure works in different dimensions. The concept of dimension which emerges, is a simple form of the concept developed by Hausdorff in [24]. So our point is this: We cut out a piece of our figure, and enlarge it by a scaling factor of s. This makes it multiply by a factor of m: For instance, assume that we do this with a line segment, with a piece of a plane or with a domain in 3-space. The results are summarized in Fig. 18.1.

As we see, for the line segment a scaling up by a factor of 3 of a small piece, yields a total of three copies of the original piece. But for the piece of the plane, a scaling factor of 3 yields altogether nine new copies. And, for a small piece of space, the scaling factor of 3 yields altogether 27 new copies. The simple fact is, that letting d denote the dimension in our usual sense, we have

$$m = s^d.$$

We now take a very bold step, and proclaim this as the definition of *Fractal Dimension*:

Fractal Dimension. Let F be a self-similar figure, which is to say, a figure such that whenever we cut out a certain small piece of it denoted by F', and enlarge it by a scaling factor of s, then we get a total of m identical copies of F'. Then the Fractal Dimension d of F is defined by the relation

$$m = s^d.$$

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_18,

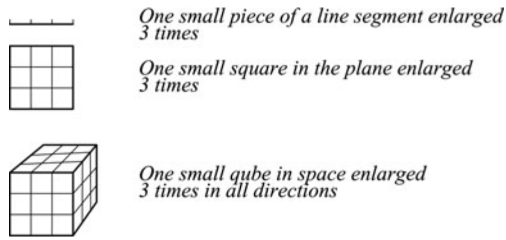


Fig. 18.1 Subdivision and enlargement for three different figures. The simplest case

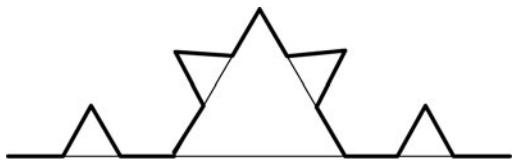


Fig. 18.2 First steps in the construction of the von Koch Snowflake Curve

18.2 The von Koch Snowflake Curve

We now examine the *von Koch Snowflake Curve*, which was introduced in Sect. 7.3. The curve of von Koch may be defined recursively, by a replacement algorithm. We start with a line segment, which is subdivided into three equal parts. The middle part is then replaced by *two* pieces, which together with the middle one, which has been removed, would have formed an equilateral triangle. See Fig. 18.2.

The recursion consists in that the figure obtained after n steps is replaced by a figure in which all line segments are replaced by the four segments, each of length equal to $\frac{1}{3}$ of the original one, as described above.

This *replacement algorithm* is really the key to how many fractals are generated. Here the algorithm is *deterministic*, in that the procedure is uniquely determined. But there are also *stochastic* procedures, which are capable of generating fractal images with great speed.

In the small segment we find *four* smaller copies of the original piece, and an enlargement with a scaling factor of 3 yields the four of them. Thus the dimension d of the von Koch Snowflake Curve is given by

$$4 = 3^d$$

which gives

$$d = \frac{\log(4)}{\log(3)} \approx 1.262.$$

From the construction of the Snowflake Curve we may also get an intuitive understanding of why this curve is continuous, but has no tangents. In fact, the curve may be approximated arbitrarily well by a *polygonal curve*, that is to say, a continuous curve consisting of line segments. It is not hard to come up with an ϵ-δ–type proof that the limit of such curves is a continuous curve. As for tangents, we choose two points P and Q on the Snowflake Curve, and select P_n, Q_n on the n-th polygonal curve C_n in the construction of the Snowflake Curve, such that P_n tends to P, Q_n tends to Q. The line $P_n Q_n$ will then tend to the secant PQ of the Snowflake Curve. As Q approaches P, the secant PQ will tend to the tangent at P, if it exists. Assume that the tangent at P exists, denote it by T. We conclude that whenever P_n and Q_n are different points on C_n, such that P_n *and* Q_n tend to P as n increases, then the line $P_n Q_n$ will tend towards the line T. However, a look at Fig. 18.2 yields ample intuitive evidence that no common limiting position for $P_n Q_n$ for all choices of P_n and Q_n tending to P can exist. Thus the Snowflake Curve has no tangents.

18.3 Fractal Shapes in Nature

Suppose we wish to apply these ideas to *fractal shapes in nature*. We might want to compute, or estimate, the dimension of a shoreline, of a cloud, and so on.

We are then faced with the difficulty that these objects are only *approximately* self similar.

There are several theoretical remarks one could make concerning this definition, at this point. However, we shall not go into the matter here. Except to note the following: It is obvious that a *different selection* G' from F may have the same property as F': Scaling it up by a factor of t would make it multiply by a factor of n. This certainly is possible as long as we do not demand an exact fit with the original, only an approximate equality of appearance. So we could get a different number e such that

$$n = t^e.$$

Nevertheless, people do compute – or "compute" – the *"fractal dimensions"* of approximately self-similar objects occurring in nature.

According to this the typical shoreline would have a dimension of about 1.2. Similarly *clouds* would be objects with an estimated dimension of more than 2 but less than 3, most being deemed to be of dimension around 2.3.

18.4 The Sierpinski Triangles

Another simple *self similar* figure is the *Sierpinski triangles.* This figure is not generated by a replacement algorithm, but rather by an *insertion algorithm*: We start with an equilateral triangle, where we draw three other equilateral triangles with side equal to half the side of the original one. Inside each one of these we repeat the process, and so on. The Sierpinski Triangles is shown in Fig. 18.3.

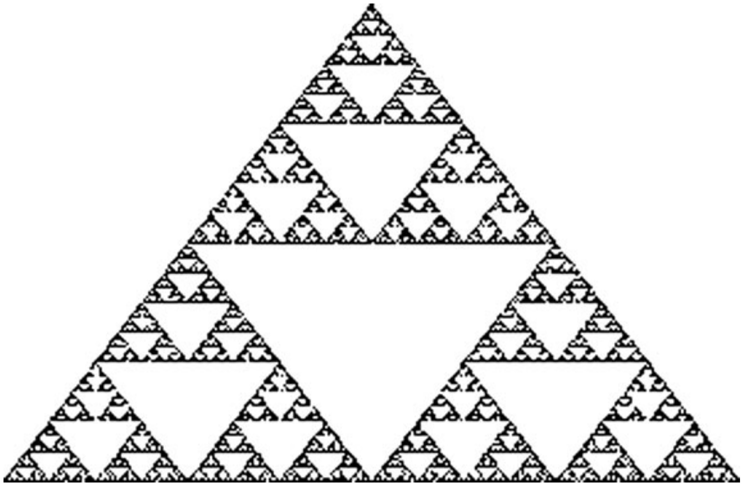


Fig. 18.3 The Sierpinski triangles

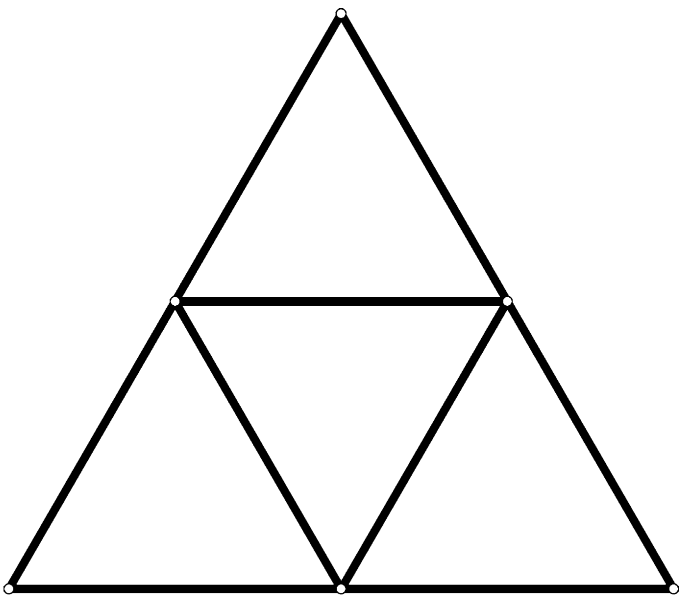


Fig. 18.4 First steps in the construction of the Sierpinski triangles

We find the dimension of the Sierpinski Triangles as follows: In Fig. 18.4 we indicate the enlargement with a scaling factor of 2.

As we see, a scaling factor of 2 yields four copies of the original. Thus

$$4 = 2^d$$

Fig. 18.5 The first four steps in the construction of the Cantor set

which gives

$$d = \frac{\log(4)}{\log(2)} \approx 1.585.$$

The Polish mathematician *Waclav Sierpinski* laid the foundation for an important school of topology in Poland.

18.5 A Cantor Set

We finally treat a figure obtained by an *excision algorithm.* We now start with an interval on a line, say of length 1. First remove the middle third of it. From the remaining two we also remove the middle third, and so on. In the end we obtain a set known as a *Cantor set*. The process is shown in Fig. 18.5.

Enlarging with a scaling factor of 3 we get two copies, thus the dimension of the Cantor Set is $d = \frac{\log(2)}{\log(3)} \approx 0.631$.

The *Cantor sets*, one of which is indicated in Fig. 18.5, form a class of sets with this fractal nature. We have seen in earlier chapters how Georg Cantor has made significant contributions to our understanding of mathematics in general and set theory in particular.

Chapter 19
Catastrophe Theory

19.1 The Cusp Catastrophe: Geometry of a Cubic Surface

We are going to treat, mathematically, the case of a Cusp Catastrophe, already encountered in Sect. 7.1. To start out, we consider a simple cubic surface in affine 3-space $\mathbb{R}^3$. Recall that an algebraic equation of degree 3 of the form

$$X^3 + aX + b = 0,$$

has its roots expressed by a beautiful formula involving cubic roots, usually referred to as *Cardano's Formula*. We shall not give the formula here, but we note that it follows from it that a number called the *discriminant of the equation* plays an important role: The discriminant is defined as

$$\Delta(a,b) = \left(\frac{b}{2}\right)^2 + \left(\frac{a}{3}\right)^3.$$

Let x_1, x_2 and x_3 denote the roots of the equation, real or complex. Then we find that

$$\Delta(a,b) = -\frac{1}{108}(x_1 - x_2)^2(x_1 - x_3)^2(x_2 - x_3)^2,$$

using the relations

$$\begin{aligned} x_1x_2x_3 &= -b \\ x_1x_2 + x_1x_3 + x_2x_2 &= a \\ x_1 + x_2 + x_3 &= 0. \end{aligned}$$

This is a simple exercise with a symbolic calculator or with MAPLE, say.

We have the following result:

Proposition 32. *The equation $X^3 + aX + b = 0$, where a and b are real, has a complex root if and only if $\Delta(a,b) > 0$. Otherwise all its roots are real.*

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_19,

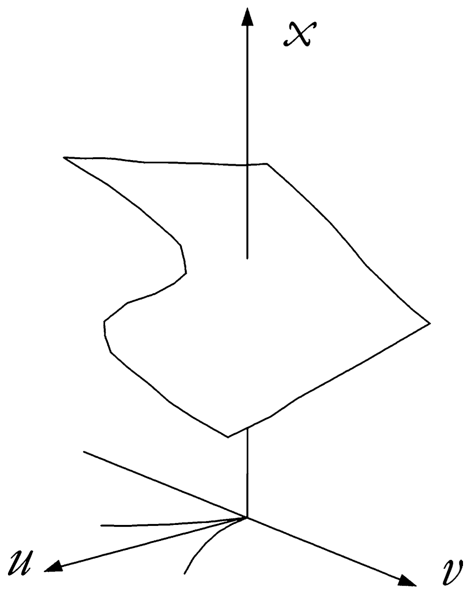


Fig. 19.1 The folded surface and the area of bifurcation above the area inside the semi-cubic parabola. The equation of the surface is $x^3 - ux + v = 0$

Proof. If there are no complex roots[1] then the computation of $\Delta(a, b)$ above shows that it is non-positive.

Conversely, if there is one complex root, say x_1, then the complex conjugate $x_2 = \overline{x_1}$ is also a root, and x_3 must be real. Then $(x_1 - x_2)^2 < 0$ while $(x_1 - x_3)^2(x_2 - x_3)^2 > 0$, thus $\Delta(a, b) > 0$. □

The surface obtained from the cubic equation

$$x^3 - ux + v = 0,$$

is shown in Fig. 19.1. We have plotted the curve given by

$$\Delta(-u, v) = 0, \text{ i.e., by } u^3 = \frac{27}{4}v^2$$

in the (u, v)-plane. The negative sign of the linear term in x is just to get a diagram similar to the ones discussed in Sect. 7.1.

Now we shall address a question which may have puzzled some readers: *Why is it that over the wedge shaped bifurcation area, only two of the points are possible? In other words, why is the middle piece of the fold "prohibited area"?* For this we need some rudiments from *Control Theory.*

---

[1] By a complex number we here understand a proper complex number, that is to say a number which is of the form $x = \alpha + \beta\sqrt{-1}$, where $\beta \neq 0$. Of course the set of complex numbers include the real ones as a subset, strictly speaking.

19.2 Rudiments of Control Theory

We are given a set of free variables $u_1, u_2, \ldots, u_m$ called *control variables* representing a point in a certain domain in $\mathbb{R}^m$, and a set of variables depending on these, referred to as *state variables*, $x_1, x_2, \ldots, x_n$. The way in which the state variables depend on the control variables may be subtle, but in the cases we shall consider here there will be a finite number of possibilities for each of the state variables for any given choice of control variables.

Control theory now deals with the problem of how the state variables change when the control variables are altered along a continuous curve from a certain point in the control space $\mathbb{R}^m$ to another. The corresponding point in the state space $\mathbb{R}^n$ will then move along a track in the state space, a curve but possibly a curve with discontinuities.

Frequently the process by which the control variables determine the possible values of the state variables is that of an *optimization process*. This may be in the simple form of *minimizing* a certain function in all the variables, control and state:

$$V = V(x_1, x_2, \ldots x_m; u_1, u_2, \ldots, u_m).$$

The semicolon instead of the comma in the functional notation serves no other purpose than to distinguish the two types of variables from each other.

The function V may appear in a variety of situations. In economic theory it would typically be a *cost function*, it could be an estimate of *exposure to risk*, in physics it might be the energy stored in some mechanical system, which tends to find an equilibrium where this energy is minimized. In some application to psychology it could be an estimate of *discomfort* suffered by an individual or by a group of people, like the inmates in a *prison.* For the dog analyzed in Zeeman's famous example, one might speculate that the function would be measuring the amount of certain *hormones* in the dog's bloodstream.

We now consider a simple situation, in which there are *two* control variables u and v, and just *one* state variable x. The function to be minimized is then denoted by

$$V = V(x; u, v).$$

For given values of u and v a possible value of x must then give a local minimum, assuming differentiability we therefore have that x must satisfy the equation $V'(x; u, v) = 0$, the derivative with respect to x must vanish. But not all such values of x are possible: If the *second* derivative is *positive*, x will give a local *maximum*, this corresponds to a *non-stable* equilibrium. If, however, the second derivative is *negative*, then we do have a local minimum, which corresponds to a *stable* equilibrium. Should the second derivative also vanish, a further analysis is required.

We may now return to the *Cusp Catastrophe* treated in Sect. 7.2. The point is that even though the surface defining the Cusp Catastrophe is of degree 3, the phenomenon actually arises from a problem related to a polynomial of degree 4:

Indeed, the potential $V = V(x; u, v)$ entering into the situation is of degree 4 in x. After some simplifying considerations one finds that without serious loss of generality one may in this case assume an expression for $V(x; u, v)$ which is linear in u and v, where there is no constant term and where the term with x^3 does not occur. After a convenient scaling, the simplest way the control variables may enter into the situation is when

$$V(x, u, v) = \frac{1}{4}x^4 - \frac{1}{2}x^2 u + xv.$$

Thus (x, u, v) must lie on the surface with equation

$$V'(x, u, v) = x^3 - xu + v = 0,$$

exactly the one we encountered in the cusp catastrophe. Furthermore, we must have

$$v''(x, u, v) = 3x^2 - u \leq 0,$$

and we find the boundary of this area by eliminating x from the equations

$$x^3 - xu + v = 0,$$
$$3x^2 - u = 0,$$

which yields the curve given by the *discriminant* of $x^3 - xu + v$, which is $u^3 = \frac{27}{4}v^2$.

We now understand why the *"middle fold"* of the surface had to be cut out: Here the second derivative is *positive*, so the points there are non-stable equilibria, and therefore x cannot remain at these values.

With this example of reasonably advanced algebra and geometry throwing light on phenomena in everyday life, we conclude this treatment of the ancient field of Geometry, so much part of our cultural heritage.

Chapter 20
General Polyhedra and Tessellations, and Their Groups of Symmetry

In this final chapter we return to polyhedra and tessellations, and study them in light of their groups of symmetry. This also applies to the more general situation of patterns and *their* groups of symmetry. We have to start out with the important groups of symmetries in the Euclidian plane and the Euclidian 3-space. This chapter presupposes more knowledge of linear algebra and group theory than the earlier parts of the book. Good sources and references for some of the material in this section are [9] and [57].

We first recall some general concepts and facts.

20.1 Isometries of $\mathbb{R}^n$

We will be dealing with Euclidian n-space over $\mathbb{R}$, but soon restrict our attention to $n = 2$ and 3. A bijective mapping $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ which preserves (the Euclidian) distance between two points is called an *isometry* of $\mathbb{R}^n$. The set of all isometries of $\mathbb{R}^n$ will be denoted by $E(n)$. The composition of two isometries is evidently again an isometry, and as is easily seen $E(n)$ is a group under composition. More generally the affine transformations also form a group, as we saw in Sect. 12.3, $E(n)$ being a subgroup.

The elements of $E(n)$ are some times called *Euclidian motions*, an important subgroup is $E^+(n)$ of orientation preserving Euclidian motions. $E^+(n)$ is a normal subgroup of $E(n)$ of order 2, the other coset consists of the so called *indirect isometries* or *improper motions*, the elements of $E^+(n)$ being referred to as the *direct isometries,* or the *proper motions*. They are also called the *rigid motions*.

$E^+(n)$ is generated by two classes of transformations, the *translations* and the *rotations*. Recall that a translation is defined by any element $a \in \mathbb{R}^n$ by

$$T_a(x) = x + a \text{ for all } x \in E(n)$$

and a rotation by

$$R_M(x) = x \cdot M$$

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_20,

where M is an $n \times n$ orthogonal matrix of determinant $\det(M) = 1$, x is written as a row-vector and $\cdot$ as before denotes matrix multiplication. This describes a rotation about the point $(0, 0, \ldots, 0) \in \mathbb{R}^n$, a rotation about the point $P = (p_1, p_2, \ldots, p_n)$ is given as

$$R_{P,M} = T_p \circ R_M \circ T_p^{-1}$$

where $p = (p_1, p_2, \ldots, p_n)$ is the n-dimensional vector $\overrightarrow{OP}$.

Moreover $E(n)$ is generated as $E^+(n)$, without the assumption of $\det(M)$ being positive.

Thus any element $f \in E(n)$ can be described as

$$f(x) = x \cdot M + a$$

where $b \in \mathbb{R}^n$ and $M \in O(n)$, the orthogonal matrices. Then

$$f^{-1}(x) = M^{-1} \cdot x - a.$$

The set of translations $T(n)$ form a normal subgroup of $E(n)$, and $E(n)/T(n) \cong O(n)$. Indeed, with f as above we find

$$(f^{-1} \circ T_b \circ f)(x) = x \cdot (M \cdot M^{-1}) + (a + b) \cdot M^{-1} - a = x + c$$

where $c = (a + b) \cdot M^{-1} - a$. This shows normality, and the isomorphism follows by our description of the elements in $E(n)$ given above.

We also have the following observation:

Proposition 33. *Let $F = T_a \circ R_{P,M}$ and $G = T_b \circ R_{Q,N}$ be two transformations in $E^+(n)$. Then*

$$G \circ F = T_c \circ R_{P,MN} \text{ where } c = (p - q + a)N + q - p + b$$

p and q being the coordinate vectors of P and Q, respectively.

In particular the composition of two transformations involving opposite rotations is a translation.

Proof. We have

$$\begin{aligned}(G \circ F)(x) &= G((x - p)M + p + a) = ((x - p)M + p + a - q)N + q + b \\ &= (x - p)MN + (p + a - q)N + q + b \\ &= (x - p)MN + p + (p - q + a)N + q - p + b \\ &= (x - p)MN + p + c = T_c \cdot T_{P,MN}\end{aligned}$$

□

For $n = 2$ we frequently write $R_{P,v}$ instead of $R_{P,M}$ when

$$M = \begin{Bmatrix} \cos(v) \sin(v) \\ -\sin(v) \cos(v) \end{Bmatrix}.$$

The following, some times referred to as the *Additivity Theorem*, is an immediate consequence of Proposition 33:

Proposition 34. *Given the two transformations F and G of $\mathbb{R}^2$ involving rotations by angles φ and ψ about the points $P = (p_1, p_2)$ and $Q = (q_1, q_2) \in \mathbb{R}^2$ as follows*

$$R_{P,v} = T_p \circ R_v \circ T_p^{-1}, R_{Q,u} = T_q \circ R_u \circ T_q^{-1}.$$

Then

$$F \circ G = T_c \circ R_{P,u+v} \textit{ for the } c \textit{ given in Proposition 33.}$$

20.2 Topological Spaces and Topological Groups

Before we proceed, we need to make a digression concerning the notion of a *topological space*. An important class of such spaces is the *metric spaces.* The concept of a metric space was explained in Sect. 10.4.

$\mathbb{R}^n$ is an example of such a space. Here we cannot pursue the field of topology much further than to present some basic terminology frequently used in connection with transformation groups. A topological space is a space X together with a collection of subsets $\mathcal{U}$ referred to as the *open subsets*, such that the following properties hold:

(1) $X, \emptyset \in \mathcal{U}$.
(2) The union of any collection of sets from $\mathcal{U}$ is again an element of $\mathcal{U}$.
(3) The intersection of any *finite* collection of sets from $\mathcal{U}$ is again in $\mathcal{U}$.

A subset F is said to be closed if it is the complement of an open set. It follows that the intersection of any family of closed subsets is again closed, while the union of any finite family of closed sets is closed.

A mapping $f : X \longrightarrow Y$ from one topological space to another is called continuous if for all open subsets $V \subseteq Y$ we have that $f^{-1}(V)$ is open in X. These mappings are the structure preserving maps for topological spaces, in technical terms they are *the morphisms in the category of topological spaces.* The identity map is of this type, and the property of being continuous (being a morphism in the category) is preserved under composition. When a mapping is bijective, and the inverse is also continuous, then it is called a *homeomorphism.* The two spaces are then called homeomorphic, and from a topological point of view they are then equivalent.

There are two extreme cases of topological spaces, namely an *indiscrete space* where only X and $\emptyset$ are open and no others, and the case where for all points $p \in X$ the set $\{p\}$ is both open and closed.

$\mathbb{R}^n$ satisfies the axioms of a topological space with the usual definition that a subset $U \subseteq \mathbb{R}^n$ is open if for all points $p \in U$ there exists ϵ such that

$$B(p, \epsilon) = \{x | \, d(x, p) < \epsilon\} \subseteq U$$

where $d(x, p)$ denotes the usual Euclidian distance, or *metric*, between x and p.

A mapping from one metric space to another $f : X \longrightarrow Y$ such that for all $x_1, x_2 \in X$ we have $d_Y(f(x_1), f(x_2)) = d_X(x_1, x_2)$ is said to be *preserve distance* or to be *distance preserving*. A mapping which preserves distance is easily seen to be continuous, and bijective mappings of metric spaces which are distance preserving are called *isometries.*

The Cartesian product $X \times Y$ of two topological spaces is again a topological space by specifying that a set is open if it is the union of sets of the form $U \times V$ where U is open in X and V is open in Y. It is easily checked that the product topology on $\mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$ is equal to the one given directly by Euclidian distance in $\mathbb{R}^{m+n}$.

$\mathbb{R}^n$ further satisfies the *Hausdorff Axiom*, namely:

(4) For any two points $x \neq y \in X$ there exist disjoint open subsets $U \ni x$ and $V \ni y$.

If $Y \subset X$ is any subset, then Y becomes a topological space by defining all intersections $Y \cap U$ as open in Y whenever U is open in X. We call this topology the one *induced* from X.

A group G is said to act on a set X provided there exists a map

$$\Phi : G \times X \longrightarrow X$$

written

$$(g, x) \mapsto g \cdot x$$

and such that the following properties hold:

$$e_G \cdot g = g, (gh) \cdot x = g \cdot (h \cdot x).$$

Definition 35. When the group G acts on the set X, then the following subset of X

$$\text{Orb}_G(x) = \{g \cdot x \in X | \, g \in G\} \subset X$$

is referred to as the orbit of x under G. For all $x \in X$ the following subgroup of G

$$\text{Stab}_G(x) = \{g \in G | \, g \cdot x = x\} \subset G$$

is referred to as the stabilizer of x under G.

A *topological group* is defined as a group which is a Hausdorff space and is such that the functions

$$f(g) = g^{-1} \text{ and } g(x, y) = xy$$

are continuous from G to G, respectively from $G \times G$ to G. A topological group is G is said to act on a topological space X provided the mapping defining the action, i.e., the mapping

$$\Phi : G \times X \longrightarrow X, (g, x) \mapsto g \cdot x$$

is continuous. In that case so is the mapping $\varphi_g = g \cdot (\,) : X \longrightarrow X, x \mapsto g \cdot x$, as this map is nothing but the composition of the two continuous mappings

$$X \hookrightarrow G \times X \longrightarrow X, x \mapsto (g, x) \mapsto g \cdot x.$$

G is a Hausdorff, thus $x \mapsto (g, x)$ gives a closed embedding $X \hookrightarrow G \times X$, which is a continuous mapping. If X is Hausdorff, in particular if X is a metric space, then all sets $\{x\}$ consisting of a singe point is closed, thus we have shown the

Lemma 3. *The stabilizer* $\mathrm{Stab}_G(x) = \varphi_g^{-1}(x)$ *is a closed subset of* X.

A closed subset of a topological group which is a subgroup is called a topological subgroup.

20.3 Discrete Transformation Groups of Metric Spaces

The important concept of a discrete transformation group will not be needed here in full generality. However, we need the following

Definition 36. A (topological) group G of isometries of a metric space is called a discrete group of transformations if it satisfies the following two conditions:

(i) For all $x \in X$, $\mathrm{Stab}_G(x)$ is finite, and
(ii) For all $x \in X$, $\mathrm{Orb}_G(x)$ is a discrete subspace of X.

Discrete transformation groups are also referred to as discontinuous groups.

The group of transformations $E(n)$ as well as the subgroups we have encountered are topological groups of isometries of metric spaces. In fact, the topology on $E(n)$ is the one induced from $\mathbb{R}^{2n+n}$, as $E(n)$ may be identified with a closed subset of that space in the following manner: Let $T_b \circ R_M \in E(n)$, M being an orthogonal matrix and $b \in \mathbb{R}^n$. Then the tuple (M, b) is an element of $\mathbb{R}^{2n+n}$, and the condition $\det(M)^2 = 1$ yields a closed subset. It is straightforward to check that compositions and inversions of corresponding transformations yield continuous functions between these subspaces of $\mathbb{R}^{2n+n}$ and $\mathbb{R}^{2n+n} \times \mathbb{R}^{2n+n}$. $E(n)$ acts on $\mathbb{R}^n$ by $t \cdot a = t(a)$. All groups of transformations G on a set X can be viewed as acting on X in this way. Since the corresponding map $E(n) \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is continuous, the topological group $E(n)$ as well as its topological subgroups act as topological groups on the metric space $\mathbb{R}^n$.

For a discrete subgroup $G \subset E(n)$ the orbit $\mathrm{Orb}_G(x)$ of any point $x \in \mathbb{R}^n$ has the property that for all $y \in \mathrm{Orb}_G(x)$ there exists $\epsilon > 0$ such that $B(y, \epsilon) \subset \mathbb{R}^n$ contains no other point from $\mathrm{Orb}_G(x)$ than y. Now $y = g(x)$ for some isometry $g \in G$, and since g is an isometry of $\mathbb{R}^n$ we have

$$B(y, \epsilon) = g(B(x, \epsilon)).$$

This easily implies the following simple but useful

Lemma 4. *For a discrete group of isometries $G \subset E(n)$ and $x \in \mathbb{R}^n$ there exists an $\epsilon > 0$ depending only on x and G such that for all $y \in \mathrm{Orb}_G(x)$ the ϵ-neighborhood $B(y, \epsilon)$ contains no other point from $\mathrm{Orb}_G(x)$ than y.*

20.4 Isometries of $\mathbb{R}^2$

In this section we shall restrict ourselves to the case of $n = 2$. The group of proper motions $E^+(2)$ of $\mathbb{R}^2$ is generated by translations T_a by vectors a, and rotations $R_{P,v}$ of angles v about points P, in the notation from the Sect. 20.6 the latter is given by the orthogonal matrix

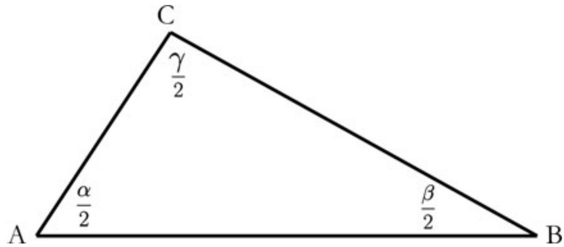
$$M = \begin{Bmatrix} \cos(v) & \sin(v) \\ -\sin(v) & \cos(v) \end{Bmatrix}.$$

To get the full group of Euclidian motions we have to include the reflections F_ℓ in a line ℓ among the generators, we also introduce the notation $F_{\ell,d}$ for a reflection in ℓ followed by a translation in the direction of ℓ by a distance d. For generation we only need one reflection, say the reflection F_x in the x-axis, since if R is a proper motion mapping ℓ onto the x-axis we have $F_\ell = R \circ F_x \circ R^{-1}$.

Proposition 35. *Let $\triangle ABC$ be a triangle where $\angle CAB = \frac{\alpha}{2}$, $\angle ABC = \frac{\beta}{2}$ and $\angle BCA = \frac{\gamma}{2}$.*

Then the composition of the rotations by α about A, of β about B and γ about C, is the identity.

Proof. This proposition is an easy consequence of the Additivity Theorem, Proposition 34. We give an alternative and more geometric proof: Form the composition G of the two reflections in the lines AC and AB, so $G = F_{AB} \circ F_{AC}$. By this transformation A is left invariant while the line AC is rotated clockwise by the angle α. Since R is the composition of two improper motions, it is itself a proper motion, and leaving A invariant shows that it is a rotation, necessarily by an angle α since this is the effect on the line segment AC.



Similarly we decompose the two other rotations and get a total of six reflections in the lines AC, AB, BA, BC, CB and CA, respectively. Thus the end result is a rotation of 2π, leaving A, B and C fixed. So the composition is the identity. □

The proposition has the following immediate

Corollary 7. *The composition of rotations by angles α and β about points A and B is a rotation about the point C constructed as in Proposition 35, by an angle γ such that $-\gamma \equiv \alpha + \beta (\operatorname{mod} 2\pi)$*

We finally list two more obvious observations:

Proposition 36. *If S is an isometry, then:*

(i) $S \circ T_b \circ S^{-1} = T_{S(b)}$.
(ii) $S \circ F_{\ell,d} \circ S^{-1} = F_{S(\ell),d}$.

20.5 Symmetry of Plane Ornaments

Definition 37. A discrete, or discontinuous, group G of isometries in $\mathbb{R}^2$ is referred to as an ornamental group. If the ornamental group G contains rotations and/or reflections but no translations it is called a rosette group. If G contains rotations and/or reflections but only translations in one direction and their inverses, it is referred to as a frieze group. Finally, if G contains rotations and/or reflections and non-parallel translations in $\mathbb{R}^2$ it is called a wallpaper group.

20.5.1 Rosette Groups

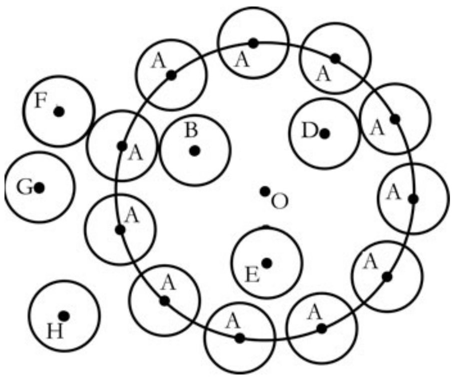
Assume first that G is a rosette group. We prove the following

Lemma 5. *Let G be a rosette group, and denote by $\mathfrak{c}$ the subgroup of rotations in G. Then there exists an integer p such that*

$$\mathfrak{c} = \mathfrak{c}_{\mathfrak{p}} = \left\{S, S^2, \ldots, S^{p-1}, S^p = I\right\}$$

where S is the rotation about a fixed point O by the angle $\frac{2\pi}{p}$.

Proof. In this case all rotations in G are about the same center. Indeed, assume that there were two rotations S_1 and S_2 about different centers O_1 and O_2. Then by Proposition 34, $S_1^{-1}S_2^{-1}S_1S-2$ would be a rotation by an angle 0, but evidently not the identity, thus being a direct isometry it would be a translation. Thus $\mathfrak{c}$ consists of rotations about O only. Consider a point A different from the center of rotation O, then its orbit $\mathrm{Orb}_{\mathfrak{c}}(A)$ under $\mathfrak{c}$ consists of a set of points on a circle C about O of radius $r = OA$. It then follows by Lemma 4 that these points are separated by circle segments of C whose cords are all of length greater than the ϵ given by Lemma 4 which only depends on G and A. Hence there is only a finite number $A_1 = A, A_2, \ldots, A_n$ of points in $Orb_{\mathfrak{c}}(A)$. The points $A = A_1, A_2$ etc are all marked by A in the figure below, some other points in the orbit of A by the whole group G by B,D,E,...



So $\mathfrak{c}$ consists of at most a finite number of rotations $R_1, \ldots, R_m$, about O, say by $\frac{2\pi}{p_i}, i = 1, \ldots, m$. But this implies that $\mathfrak{c}$ is generated by the rotation by $\frac{2\pi}{p}$ where p is the least common multiple of the $p_1, \ldots, p_m$. □

We return to the rosette group G. There are two cases to consider. First assume that G consists of direct isometries only, in other words there are no flips, so $G = \mathfrak{c} = \mathfrak{c}_{\mathfrak{p}}$.Thus $G = \{S, S^2, \ldots, S^{p-1}, S^p = I\}$, the group of p-fold rotations, for some integer p. The angles of rotation are indicated in black on Fig. 20.1, the centers of rotation in blue.

We next consider the case when the group G of transformations contain rotations and reflections. Again, since there are no translations in G, the rotations are all about the same center and are powers of the rotation S by $\frac{2\pi}{p}$ for some integer p. Since the composition of two reflections from G is a direct transformation, it must be a rotation R, and thus the two reflections are reflections in lines ℓ and ℓ', respectively, which correspond under R.

Thus the opposite transformations in G are the transformations $\mathcal{F}_\ell \circ S^i, i = 1, 2, \ldots p$. The p lines of reflection are the images of ℓ by $S^i, i = 1, 2, \ldots, p$, they are the bisectors of the angles corresponding to the rotations in G. This gives the possibilities $\mathfrak{d}_p, p = 1, 2, 3, \ldots, n, \ldots$ for the group in this case, $\mathfrak{d}_p$ being generated by $\mathfrak{c}_p$ and the reflection in one of the bisector lines, say ℓ. This is the group

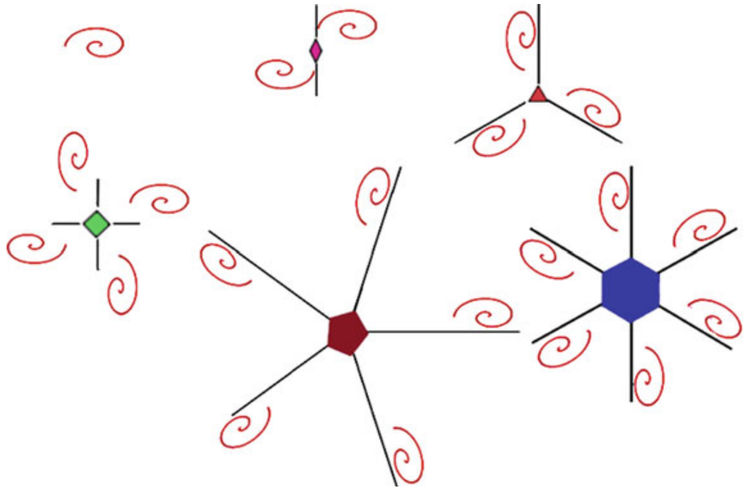


Fig. 20.1 Above rosettes with symmetry $\mathfrak{c}_1, \mathfrak{c}_2, \mathfrak{c}_3$, below rosettes with symmetry $\mathfrak{c}_4, \mathfrak{c}_5, \mathfrak{c}_6$. The centers of rotation are indicated by corresponding polygons

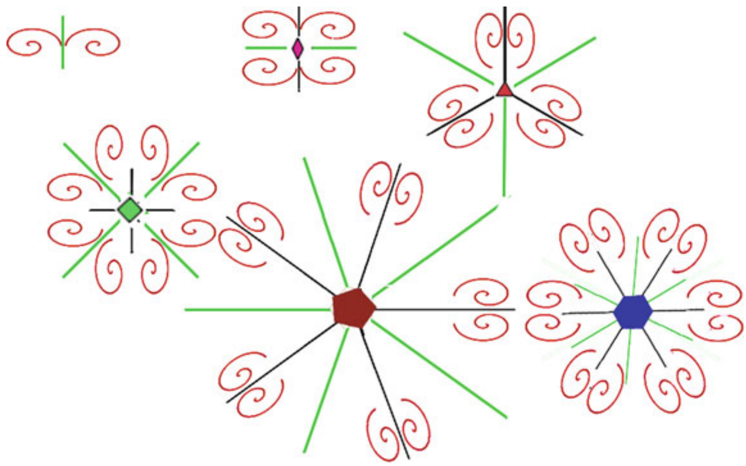


Fig. 20.2 Above rosettes with symmetry $\mathfrak{d}_1, \mathfrak{d}_2, \mathfrak{d}_3$, below rosettes with symmetry $\mathfrak{d}_4, \mathfrak{d}_5, \mathfrak{d}_6$

of symmetries of the regular p-gon. Thus the lines of reflection are the bisectors of the angles indicated in black in the figures. The corresponding rosettes are shown in Fig. 20.2. The centers of rotation are indicated by corresponding polygons, in general they are referred to as *p-gonal centers* for $p = 2$ (diagonal), $p = 3$ (trigonal), etc. The lines of reflection are in green.

This completes the discussion of the rosette groups, which are the only ornamental groups of finite order. We note that in this case $\mathbb{R}^2 - \{O\}$ is divided into a set of connected regions, disjoint except for their boundaries which are rays from

O to infinity. If the ray to the right is included in one of them, we obtain a connected region referred to as a *fundamental domain* or a *unit cell* for G. Its transforms under G will cover $\mathbb{R}^2 - \{O\}$ without overlap and without repetition. For a rosette group we note that the fundamental domain is unbounded.

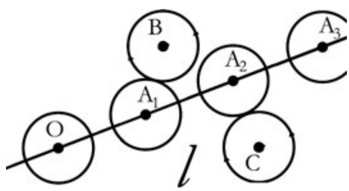
20.5.2 Frieze Groups

We next turn to the *frieze groups,* in other words ornamental groups containing translations, but only *in one direction.* We follow the treatment given in [57]. More details and explanations may be found in that reference.

We first prove the

Lemma 6. *Let G be a frieze group. Then there exists a non-zero vector $a \in \mathbb{R}^2$ such that the subgroup $\mathfrak{F}$ of translations in G is generated by the minimal translation T_a.*

Proof. $\mathfrak{F}$ consists of translations T_b where all the vectors b are some real multiple of a single non-zero vector a. To show is that we may choose a to be of minimal positive length, such that all the vectors b are integral multiples of this a. The proof is similar to that of Lemma 5:



Choose a point $O \in \mathbb{R}^2$ and let ℓ be the line through O in the direction of a. In the figure above the points in the orbit of O under $\mathfrak{F}-\{id\}$, the translations different from the identity, are denoted by $A_1, A_2, \ldots$ Some other points in the orbit under G are denoted by $B, C, \ldots$ Let A_1 be one of the two orbit points under $\mathfrak{F}-\{id\}$ closest to O. Replace a by $\overrightarrow{OA}_1$. Then all $T_b \in \mathfrak{F}$ are translations by an integral multiple of a, otherwise we could find an integer m such that the vector $b' = b - ma$ would give a point $A' = T_{b'}(O)$ closer to O than $T_a(O)$. Thus T_a generates $\mathfrak{F}$. □

A group of translations generated by a single translation as in the above proof is denoted by $\mathfrak{F}_1$. Strictly speaking this refers to the conjugate class of the 1-dimensional translation groups (Fig. 20.3).

We next turn to frieze groups containing translations and rotations, but no flips and no glide transforms. We find that the choice of rotations is quite limited in frieze groups:

Lemma 7. *Let G be a frieze group, and let $R \in G$ be a rotation. Then R is a 2-fold rotation. The diads (centers of 2-fold rotations) of G are obtained by starting with*

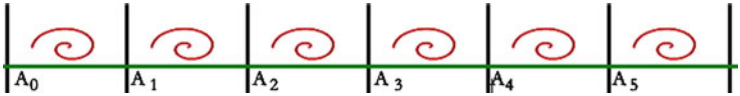


Fig. 20.3 Above we see a frieze with symmetry $\mathfrak{F}_1$. The fundamental domain can be taken as the unlimited strip between any two consecutive vertical black lines in the figure

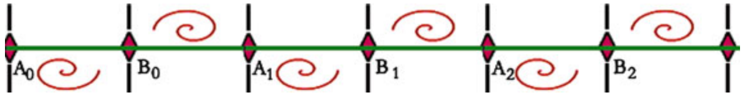


Fig. 20.4 Above we see a frieze with symmetry $\mathfrak{F}_2$. The fundamental domain can be taken as the unlimited half-strip with base $A_i A_{i+1}$ above the green line

some diad A_0, and then let $A_i = T_a^i(A_0)$ for all $i \in \mathbb{Z}$, where T_a is the minimal translation of Lemma 6. Then the A_i, together with the mid points of the segments B_i of $A_i A_{i+1}$, are all the diads of G.

Proof. Let R be the rotation by the angle α. Then evidently the transformation $U = R \circ T_a$ is equal to the translation T_b where b is a turn of the vector a by the angle α. But all translations in G are in the direction a, thus $b = \pm a$, thus α is an integral multiple of π, and the rotation is 2-fold.

Clearly all the A_i are diads. If R is a rotation by π about A_1, then the composition $S = R \circ T_a$ is a rotation by π which maps A_0 to A_1. Thus it is the 2-fold rotation about B_0, thus all the B_i are diads. Finally, let R_P be a 2-fold rotation about the point P, and R_0 the rotation about A_0. Then $T = R_P \circ R_0$ is a translation in G, therefore it maps A_0 to one of the A_n. Since R_0 leaves A_0 invariant, we have $R_P(A_0) = A_n$. Therefor P is the mid point of the segment $A_0 A_n$, thus P is one of the diads we have found above. □

This group is denoted by $\mathfrak{F}_2$ (Fig. 20.4).

We now consider frieze groups containing flips, and have the following

Proposition-Definition 1 *Let G be a frieze group generated by $\mathfrak{F}_1$ or $\mathfrak{F}_2$ and one or more reflections. Then the only possibilities are one reflection R_ℓ in a line ℓ parallel to the direction of the smallest translation $T = T_a$, or reflections $R_{\ell'}$ in lines ℓ' orthogonal to ℓ, or both. Moreover:*

(i) The sets of all compositions of elements from $\mathfrak{F}_1$ and one reflection R_ℓ or $R_{\ell'}$ (and its inverse) form groups of transformations denoted by $\mathfrak{F}_1^1$ and $\mathfrak{F}_1^2$, respectively. The composition of both reflections is a 2-gonal rotation. Hence adjoining both to $\mathfrak{F}_1$ gives the same result as adjoining one of them to $\mathfrak{F}_2$, so this case is covered by (ii).

(ii) If a group G of transformations is generated by $\mathfrak{F}_2$ and a reflection in a line ℓ parallel to the direction of T, then the line must pass through the 2-fold centers

of rotation of $\mathfrak{F}_2$. Then the set of all compositions of transformations from $\mathfrak{F}_2$ and $R = R_\ell$ is a group of transformations, denoted by $\mathfrak{F}_2^1$.

(iii) *If a group G of transformations is generated by $\mathfrak{F}_2$ and a reflection in a line perpendicular to the direction of T_a, then either the line passes through a center of rotation, or else it bisects the segment joining two centers of rotation. Compositions of elements from $\mathfrak{F}_2$ and such reflections yield a group of transformations. In the former case we obtain the group $\mathfrak{F}_2^1$. In the latter case we get a new group which is denoted by $\mathfrak{F}_2^2$.*

Proof. Let S be the reflection in a line n. Then $S^{-1} \cdot T_a \cdot S = T_{S(a)}$, hence n must be either an ℓ or an ℓ' as in the statement of the lemma.

To prove (i), we immediately verify that the following sets of transformations are groups under composition:

$$\{R \cdot T^i \,\big|\, i \in \mathbb{Z}\}, \{R' \cdot T^i \,\big|\, i \in \mathbb{Z}\}.$$

Indeed, recall that a non-empty subset A of a group G is a subgroup if and only if it is closed under the operation $(a, b) \mapsto a \cdot b^{-1}$. Two reflections in two different lines ℓ both parallel to the direction of T_a is impossible because their composition gives a translation not parallel to a. However, any ℓ' orthogonal to a may be replaced by a translation of it by a power of T_a.

To prove (ii), we note that any transformation in G maps a center of rotation to a center of rotation. Since the only centers of rotation of G are those of $\mathfrak{F}_2$, the first part of the claim follows. The last part is shown in the same way as (i).

The first part of (iii) also follows because all the transformations preserve centers of rotation. To show that adjoining a reflection in a line ℓ' perpendicular to ℓ gives $\mathfrak{F}_2^1$, note that the composition of a reflection in ℓ, respectively ℓ', and a 2-gonal rotation yields a reflection in ℓ', respectively ℓ. See Fig. 20.5. Thus adjoining either reflection to $\mathfrak{F}_2$ gives the same result. □

First we see a Frieze with symmetry $\mathfrak{F}_1^1$ (see Fig. 20.5). The fundamental domain can be taken as the unlimited half-strip with base $A_i A_{i+1}$ above the green line.

Next, we see a Frieze with symmetry $\mathfrak{F}_1^2$ (Fig. 20.6). The fundamental domain can be taken as the unlimited half-strip with base $A_i A_{i+1}$ above the green line.

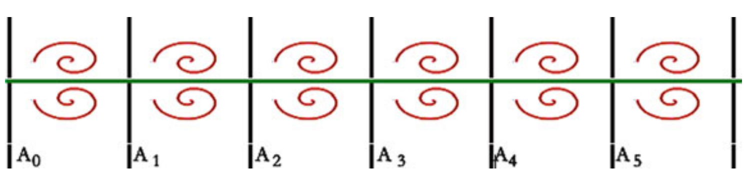


Fig. 20.5 Frieze with symmetry $\mathfrak{F}_1^1$

Fig. 20.6 Frieze with symmetry $\mathfrak{F}_1^2$

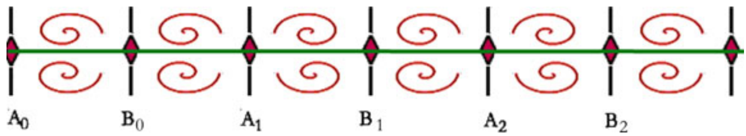


Fig. 20.7 Frieze with symmetry $\mathfrak{F}_2^1$

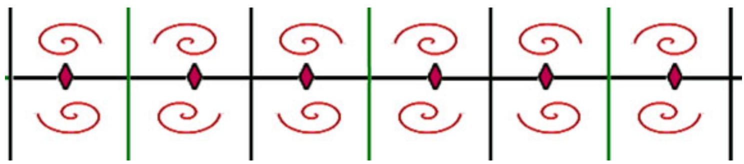


Fig. 20.8 Frieze with symmetry $\mathfrak{F}_2^2$

Below we see a Frieze with symmetry $\mathfrak{F}_2^1$ (Fig. 20.7). The fundamental domain can be taken as the unlimited strip between consecutive black lines. Reflection in the green (and therefore also the black) lines.

The last symmetry type based on the group $\mathfrak{F}_2$ is $\mathfrak{F}_2^2$. Below we see a Frieze with this symmetry (Fig. 20.8).

The final possibility, which is not covered by Proposition-Definition 1, is when G has a glide-reflection among its generators. Let L be such a glide-reflection. Then L^2 is a translation, so $L^2 = T^m = T_{ma}$ for some integer m, and replacing L by L^{-1} if necessary we may assume that

$$L^2 = T^{2n} \text{ or } L^2 = T^{2n-1} \text{ for some integer } n \geq 1.$$

In the former case LT^{-n} is a reflection and thus G is $\mathfrak{F}_1^1$ or $\mathfrak{F}_2^1$. In the latter case $M = L \circ T^{-n+1}$ is a glide reflection such that $M^2 = T$. Replacing L by M we may assume that $L^2 = T$. As in the proof of Proposition-Definition 1 we verify that

$$\{T^n, LT^n | \, n \in \mathbb{Z}\}$$

is a subgroup of $E(2)$, which we denote by $\mathfrak{F}_1^3$.

So far we have only assumed that G, in addition to the glide reflection L, contain translations. Now assume that G has a rotation R. Then it follows that G contains

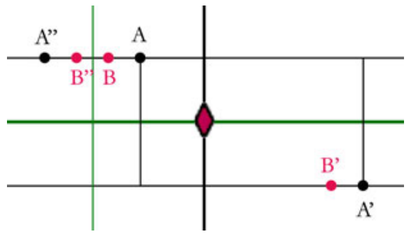


Fig. 20.9 SL is a reflection in a line ℓ perpendicular to the axis of the diads

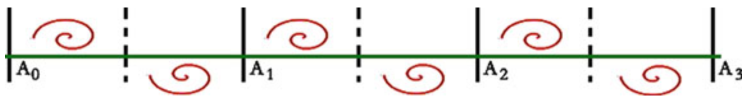


Fig. 20.10 Frieze with symmetry $\mathfrak{F}_1^3$

$\mathfrak{F}_2$ as a subgroup. Then the axis of the glide reflection L coincides with the axis of the diads. Therefore, if S is a 2-gonal rotation in G, then LS is a reflection in a line ℓ perpendicular to the axis of the diads, as illustrated in Fig. 20.9.

Here $L(A) = A'$ and $S(A') = A''$, similar for B, thus the composition is the reflection in the green vertical line. As in the figure we may normalize by letting the glide reflection be in the line $y = 0$, and $a = (\alpha, 0)$. We may also assume that the center of rotation is $(0, 0)$, thus the rotation S is $S(x, y) = (-x, -y)$. The composition SL is then easily seen to be reflection in the line $x = -\frac{\alpha}{2}$. Hence the group is $\mathfrak{F}_2^2$ which we found already.

Below, in Fig. 20.10, we see a Frieze with symmetry $\mathfrak{F}_1^3$.

This completes the treatment of the frieze groups. We finally turn to the wallpaper groups.

20.5.3 Wallpaper Groups

We follow the treatment given in [57]. More details and explanations may be found in that reference. In the cases we have treated so far the transformation groups have fundamental domains which are *unbounded*. We now come to the case when the fundamental domains are bounded, in other words to the *wallpaper groups*. The simplest one among these groups is described in the

Proposition-Definition 2 *(i) A wallpaper group W is an ornamental group which contains non-parallel translations.*

(ii) If W only contains translations it is generated by two translations T_a and T_b where a and b are non-parallel vectors. This group is denoted by W_1.

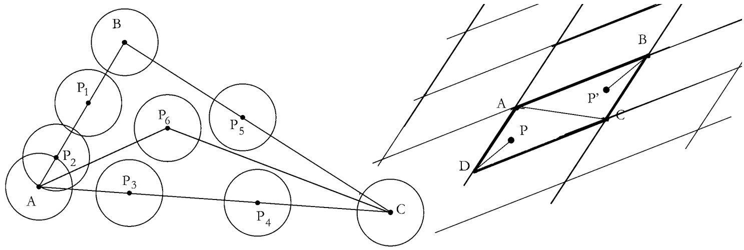


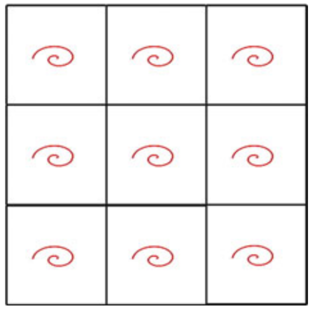
Fig. 20.11 Illustration to the proof

Proof (of (ii)). Pick any point $A \in \mathbb{R}^2$, and let $S = \mathrm{Orb}_G(A)$. By Lemma 4 there is an $\epsilon > 0$ which depends only on A and W such that the ϵ-neighborhood $\mathrm{B}(Q, \epsilon)$ about any $P \in S$ contains no other point of S. Thus the points in S have distance at least ϵ from each other. Moreover, since there are non-parallel translations in W there are two points $B, C \in S$ such that A, B and C are non-collinear. We may assume that there are no other points from S on the closed $\triangle ABC$: In fact, by the lemma there are in any case at most a finite number $m + 3, m > 0$, of such points, say $A, B, C, P_1, \ldots, P_m$. Then by replacing B or C with P_m, we reduce to the case of at most $m - 1 + 3$ points, repeating this we arrive finally at $m = 0$ and the claim follows. The first step in the procedure of finding the minimal configuration of 3 non-collinear points in S is illustrated to the left in Fig. 20.11. Here $m = 6$, and in the next step after B has been replaced by P_6 we have $m = 2$. If then C is replaced by P_3, we are done.

To the right on the figure we let D be the image of A under the transformation (necessarily a translation) in W which maps B to C. The parallelogram $ABCD$ gives rise to a grid, the corners of which form a lattice of points as shown on the figure. These points are all contained in S. But conversely any point from S appears as a point in the lattice. In face, let $P \in S$ be a point not in the lattice. Then there is a translation $T \in W$ which moves P to the closed parallelogram $ABCD$, so we may assume $P \in ABCD$. But then $P \in \triangle ACD$ by the assumption on $\triangle ABC$. Now both P and D are in the orbit of A, thus there is a translation $L \in W$ which maps P to D. But this translation also maps B to a point P' in $\triangle ABC$, see the right hand side of Fig. 20.11. This is against the assumption on $\triangle ABC$, and the claim follows. Let T_a be the translation mapping A to B, and T_b the one mapping C to B. Then all transformations in W are of the form $T_a^i T_b^j$ with $i, j \in \mathbb{Z}$ since no other translation is possible when the lattice is preserved. □

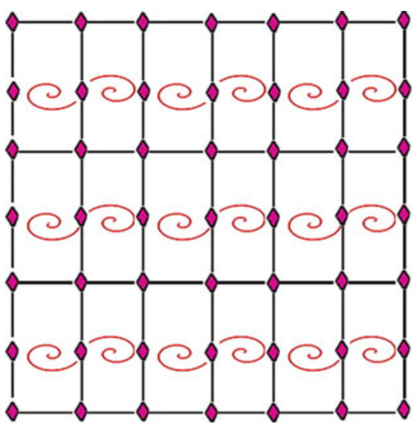
This proposition is the starting point of the complete description of all the types of wallpaper groups, to be given in the following section.

20.5.4 The 17 Types of Wallpaper Groups



Symmetry type W_1: The simplest type of wallpaper groups, with only translations. Another, more modern name for it is $p1$. There are neither reflections, glide reflections, nor translations. The two translation axes may be inclined at any angle to each other.

We next consider the case when the group contains a 2-uple rotation, in other words a half-turn, but no other kind of rotation. Let A be a center of rotation. Then the subgroup of translations of the group maps A to the points of a lattice. As in the proof of Lemma 7 we find that the vertices and the midpoints of the sides, as well as the center, of the resulting parallelograms are all the diads of the group. Thus the group consists of the set of all translations and half-turns which leave this lattice invariant, these transformations form the group. We denote this type by W_2.



Symmetry type W_2: The wallpaper group with only translations and 2-fold rotations. A modern name for it is $p2$. It is obtained by adjoining a 2-fold rotation as generator to $p1$. It has no reflections nor glide reflections. The diads are indicated as shown on Figs. 20.1 and 20.2.

Next, let the point P be a p-gonal center of rotation with $p > 2$. Since the group is discrete, there is a minimal distance form P at which we can find another center of rotation Q, with a q-fold rotation where $q > 2$. Rotation about P followed by rotation about Q by the angles $\frac{2\pi}{p}, \frac{2\pi}{q}$ then by Corollary 7 yield a rotation about some point R of an angle $\frac{2\pi}{r} \equiv -(\frac{2\pi}{p} + \frac{2\pi}{q}) (\mathrm{mod}\, 2\pi)$. Thus

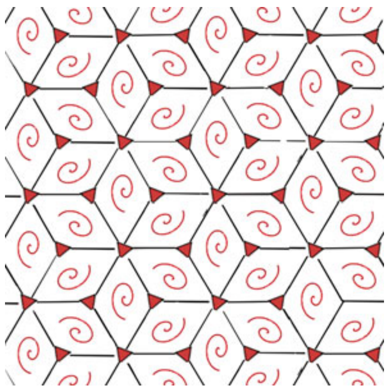
$$\frac{2\pi}{p} + \frac{2\pi}{q} + \frac{2\pi}{r} = 2\pi$$

or

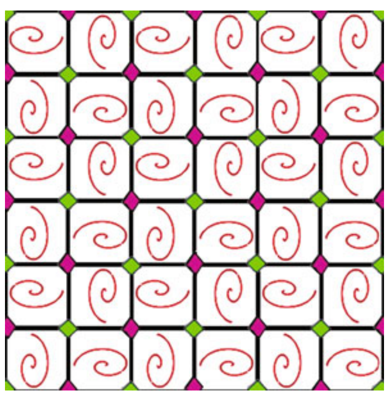
$$\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1.$$

This equality implies that p, q and r may only take the values $2, 3, 4, 6$, which is known as the *Crystallographic Restriction Theorem.* In particular the centers of rotation will have to correspond closely to the three regular tessellations of the plane.

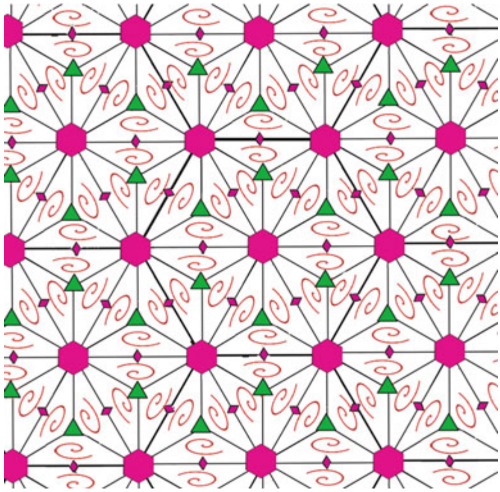
Using this result, we may now list all the remaining direct symmetries of wallpapers:



Symmetry type W_3: The wallpaper group with only translations and 3-fold rotations. A modern name is $p3$. It is the simplest group which contains 3-fold rotations. It has no reflections nor glide reflections. All vertices are trigonal centers of rotation.

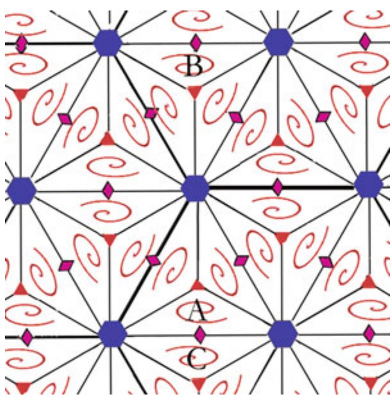


Symmetry type W_4: The wallpaper group with translations and 4-fold rotations. A modern name is $p4$. It also has rotations of order 2, with centers midway between the 4-fold centers. There are no reflections nor glide reflections.



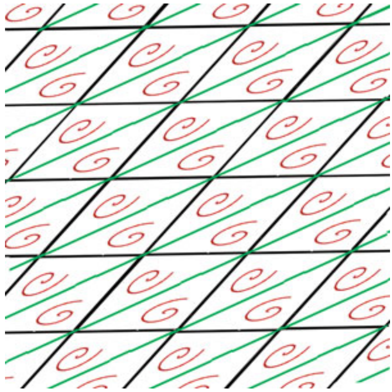
Symmetry type W_6: A modern name is $p6$. In addition to the translations, it contains 6-fold rotations and also 2-fold and 3-fold ones. It has no reflections nor glide reflections.

The five types presented above represent all the crystallographic groups of proper motions in the plane, in other words, all the direct symmetries of wallpapers. The type W_6 has transformation group generated by the rosette group $\mathfrak{d}_6$ of 6-fold rotations, and the two translations given by the vector a from the central hexagon to the neighboring one up left, and the vector b from the central hexagon and up right. This yields the symmetry group W_6, or $p3$, and it illustrates nicely how adding two translations to $\mathfrak{d}_6$ produce 2-fold and 3-fold rotations: For example a 2-fold rotation is obtained as shown below.

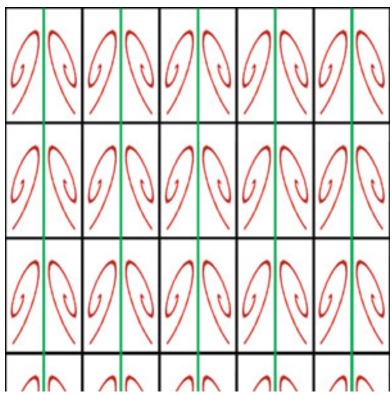


Here $A \mapsto B$ is a 6-gonal rotation to the power 3, then $B \mapsto C$ is by the translation given by the vector $-(a + b)$.

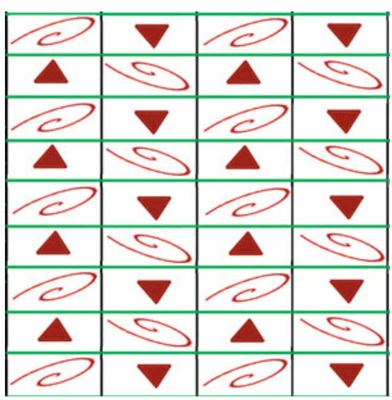
We now attempt to enlarge the groups $W_1 - W_6$ by reflections and/or glide-reflections. Adding a reflection to the generators of W_1 may be done in two different ways, yielding the groups W_1^1 and W_1^2, respectively. Addition of a glide-reflection yields the group W_1^3.



Symmetry type W_1^1: A modern name is *cm*. These wallpaper groups are generated by two translations and a reflection in a line bisecting the angle between the translations. The lattice is rhombic, and fundamental domains are triangles limited by two black and one green line segment.

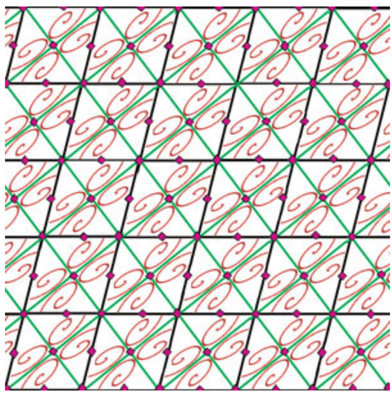


Symmetry type W_1^2: A modern name is *pm*. These wallpaper groups are generated by two orthogonal translations and reflection in a line parallel to one of the translations. The lattice is rectangular, fundamental domains are limited by three black and one green line segment.



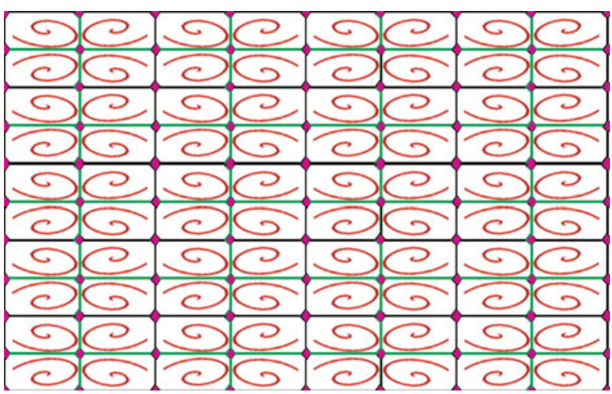
Symmetry type W_1^3: A modern name is *pg*. The wallpaper group generated by a glide reflection and a translation orthogonal to the glide axis. Thus there is a set of parallel glide-axes (green on the figure), the distance between adjacent axes being half the distance of the generating translation. There are neither rotations nor reflections. Fundamental domains are the squares limited by two black and two green line segments.

We next consider the case of enlarging group W_2 by adding reflections and/or glide reflections. We first consider the case when the group contains a reflection in a line. Then the translations will transform the diads into a rhombic lattice of points. The group contains the reflection in a line, and since this reflection preserves the lattice of diads, the line of reflection is a diagonal of a rhombus. By rotation the other diagonal is also a line of reflection, thus there are two perpendicular lines of reflection, and the group W_2^1 is determined.

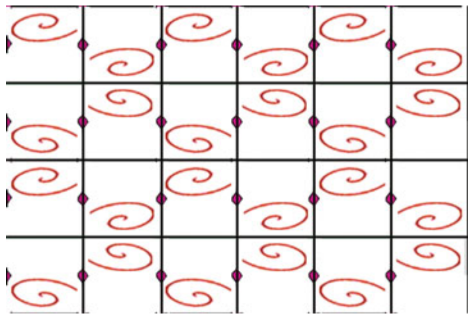


Symmetry type W_2^1: A modern name is *cmm*. Two sets of parallel mirror lines, mutually perpendicular. Running horizontally is a set of parallel glide lines. The group thus has reflections in two perpendicular directions, and a rotation of order two whose center is not on a reflection axis. It also has two rotations whose centers *are* on a reflection axis.

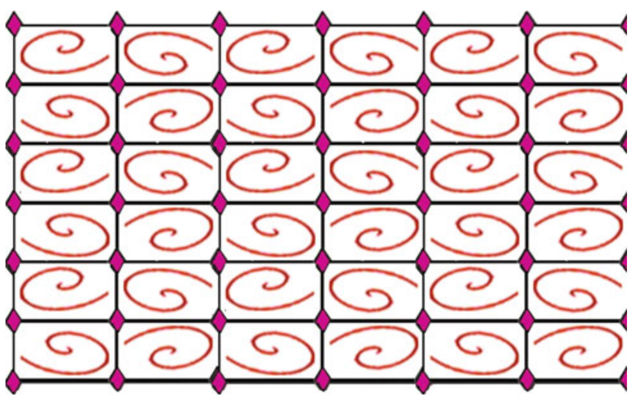
A rhombic lattice of diads only yields the type W_2^1, but a rectangular lattice leads to three more types:



The symmetry type W_2^2, with the modern name *pmm*, is obtained by enlarging a rectangular group of type W_2 by a reflection in a line (green) passing through a diad. By rotation this gives a second, perpendicular, line of reflection, also shown in green. The group contains perpendicular axes of reflection, with 2-fold centers of rotation where the axes intersect.

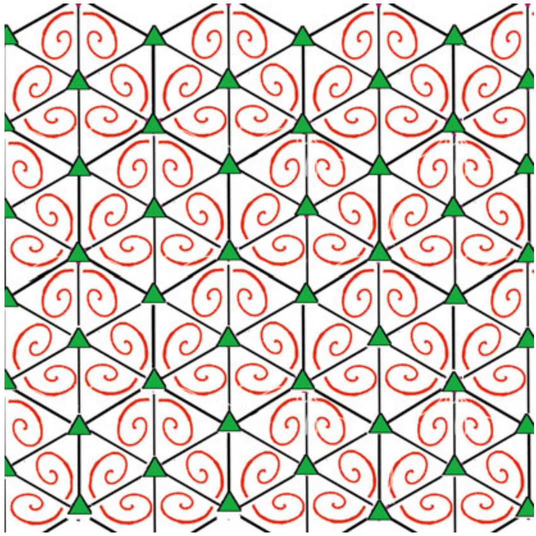


Symmetry type W_2^3, with modern name *pmg*. Here there is a line of reflection which does not contain diads. Thus the line must run halfway between the diads as shown on the figure. There are glide reflections perpendicular to the lines of reflection.

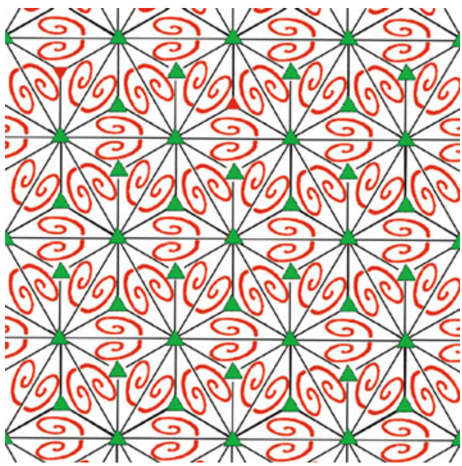


Symmetry type W_2^4, a modern name is *pgg*. A glide reflection in a line passing through a diad leads to this symmetry type with two sets of parallel glide lines and 2-fold rotations.

In order to enlarge W_3 by a reflection, we have two possibilities: Either the line of reflection contains the shorter diagonal of the rhombi in the diagram for W_3 which is shown above, or it contains the longer diagonal. The two possibilities are realized as follows:

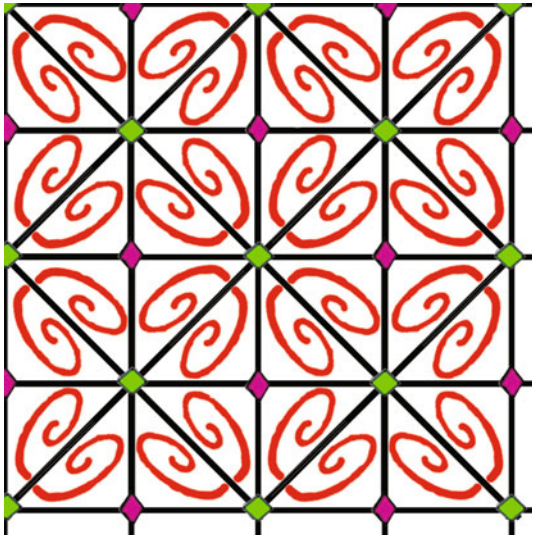


Symmetry type W_3^1. A modern name is $p3m1$. It has 3-fold rotations, all their centers lie on the reflection axes. The lines of reflections, inclined $\frac{\pi}{3}$ to each other, contain the shorter diagonals of the hexagon.

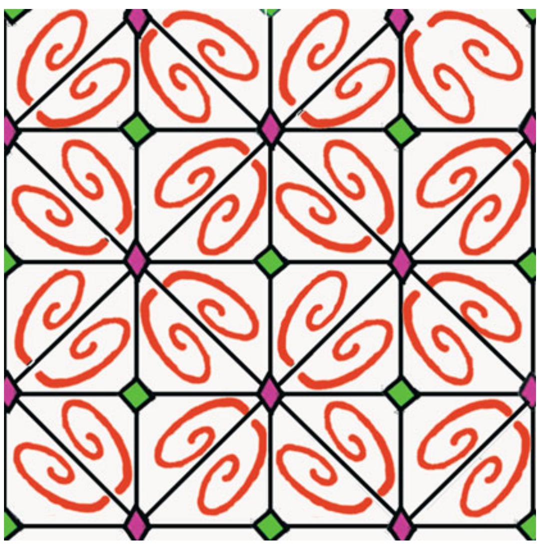


Symmetry type W_3^2. A modern name is $p31m$. It contains reflections, with lines inclined $\frac{\pi}{3}$ to each other and rotations of order 3. The lines of reflection contain the longer diagonals of the hexagon.

Enlarging the symmetry type W_4 by reflections, we get two possibilities.

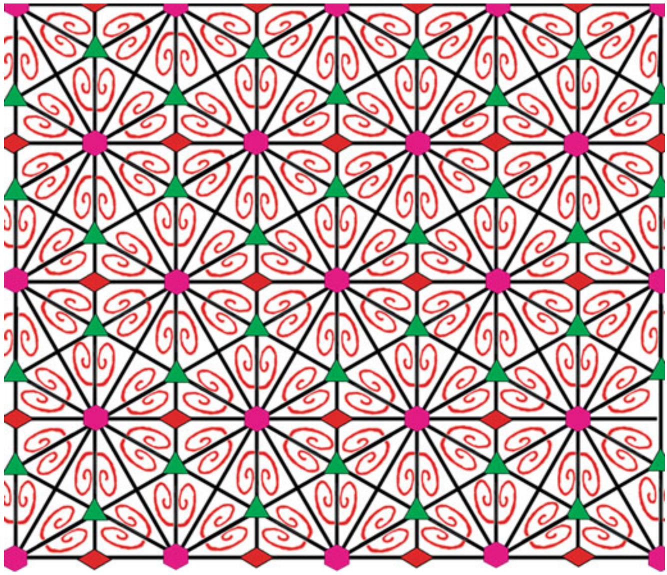


Symmetry type W_4^1. A modern name is $p4m$. The group has both order 2 and order 4 rotations. The four axes of reflection contain a tetrad, i.e., a center of 4-gonal rotation. Every rotation center lies on some reflection axes. This is the symmetry type of the regular tessellation by squares.



Symmetry type W_4^2. A modern name is $p4g$. The group contains reflections and rotations of orders 2 and 4. There are two perpendicular reflection lines passing through each center of order 2 rotation. But the axes of reflection do not contain a tetrad. There are four directions of glide reflections.

The final one of the 17 symmetry types for wallpapers is the following, obtained by enlarging W_6 by a reflection.



Symmetry type W_6^1. W_6 enlarged by a reflection. A modern name is $p6m$.

20.6 Symmetries in Space

Recall from Sect. 20.1 that the isometries of $\mathbb{R}^3$ form a group $E(3)$ of *Euclidian motions* under composition. An important subgroup is $E^+(3)$ of orientation preserving Euclidian motions, $E^+(3)$ is a normal subgroup of $E(3)$ of order 2, the other coset is constituted by the so called *indirect isometries* or *improper motions*, the elements of $E^+(3)$ being referred to as the *direct isometries,* or the *proper motions*. They are also called the *rigid motions*.

$E^+(3)$ is generated by two classes of transformations, the *translations* and the *rotations*. Recall that a translation is defined by any element $a \in \mathbb{R}^3$ by

$$T_a(x) = x + a \text{ for all } x \in E(3)$$

and a rotation by

$$R_M(x) = x \cdot M$$

where M is a 3×3 orthogonal matrix of determinant $\det(M) = 1$, x is written as a row-vector and $\cdot$ denotes matrix multiplication. This describes a rotation about the point $(0, 0, 0) \in \mathbb{R}^3$, a rotation about the point $P = (p_1, p_2, p_3)$ is given as

$$R_{P,M} = T_p \circ R_M \circ T_p^{-1}$$

where $p = (p_1, p_2, p_3)$ is the 3-dimensional vector $\overrightarrow{OP}$.

Moreover $E(3)$ is generated as $E^+(3)$, without the assumption of $\det(M)$ being positive.

Thus any element of $E(3)$ can be described as

$$f(x) = x \cdot M + a$$

where $b \in \mathbb{R}^3$ and $M \in O(3)$, the orthogonal matrices. Then

$$f^{-1}(x) = M^{-1} \cdot x - a.$$

The set of translations $T(3)$ form a normal subgroup of $E(3)$, and $E(3)/T(3) \cong O(3)$. Indeed, with f as above we find

$$(f^{-1} \circ T_b \circ f)(x) = x \cdot (M \cdot M^{-1}) + (a + b) \cdot M^{-1} - a = x + c$$

where $c = (a + b) \cdot M^{-1} - a$. This shows normality, and the isomorphism follows by our description of the elements in $E(3)$ given above.

Also recall from Sect. 20.1 the following observation, given as Proposition 33: Let $F = T_a \cdot R_{P,M}$ and $G = T_b \cdot R_{Q,N}$ be two transformations in $E^+(3)$. Then

$$G \cdot F = T_c \cdot R_{P,MN} \text{ where } c = (p - q + a)N + q - p + b$$

p and q being the coordinate vectors of P and Q, respectively. In particular the composition of two transformations involving opposite rotations is a translation.

In order to study the kinds of symmetry a polyhedron can have, and thus the degrees of regularity of different polyhedra, we need only examine the various types of symmetries in Euclidian 3-space which do not involve translations. Thus we need to examine the 3-dimensional analogues of the rosette groups we encountered in the previous section.

As in the case of symmetries in the plane, we distinguish between the *direct symmetries*, corresponding to physically moving a body, and *indirect symmetries*, which may only be realized by utilizing a mirror in addition to physically moving the body.

Hence the symmetry types to be studied are the *rotational symmetries* as well as *reflection symmetries.*

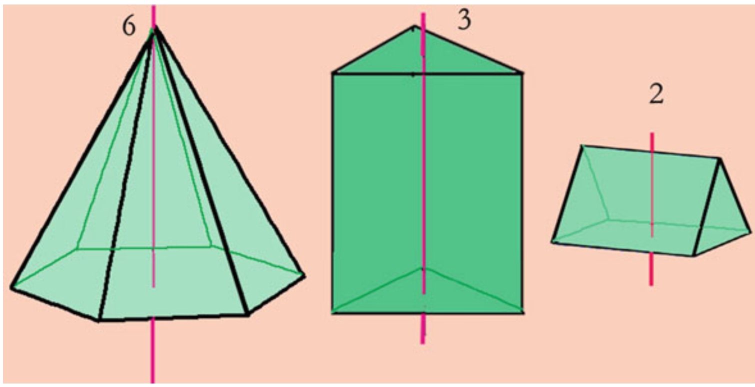
Following Cromwell [9] we now give a brief account of the symmetries in space which a polyhedron can have. For proofs and more details we refer to this source.

20.6.1 Systems of Rotational Symmetries in Space

The various symmetry types are labelled according to a system whereby a capital letter, C, D, T, O or I, indicate the nature of the system of rotations involved. Lower case letters as subscript carry additional information on the actual complete system of symmetries.

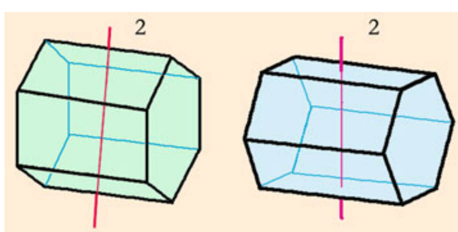
20.6.1.1 Cyclic Symmetry C_n and Dihedral Symmetry D_n

The simplest type of rotational symmetry is *cyclic symmetry*, which we illustrate by the rotations of a hexagonal pyramid through the axis ℓ displayed to the left below.

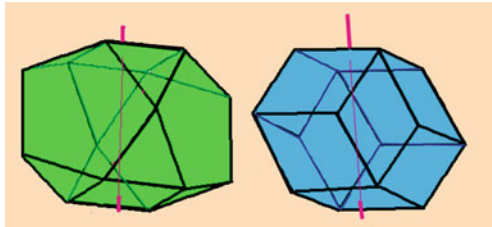


This polyhedron has 6-fold cyclic symmetry, its group of symmetries is denoted by C_6. Analogously C_n is defined for any positive integer n, C_n contains the n rotational symmetries through an axis ℓ by angles $\frac{2\pi}{n}, 2\frac{2\pi}{n}, \ldots, (n-1)\frac{2\pi}{n}, 2\pi$.

To the right we have two triangular prisms with regular base. These prisms also have six rotational symmetries, but with more than one axes of rotation. As we see there is one 3-fold axis and three 2-fold axes. In general an n-gonal prism (with $n > 2$) has one n-fold axis and n 2-fold ones, the former is referred to as the *principal axis* and the latter group as *secondary* ones. This type of symmetry is denoted by D_n. When n is odd, then the secondary axes are equivalent, while the even case yields two classes of secondary axes, as shown on the figure below for an hexagonal prism.



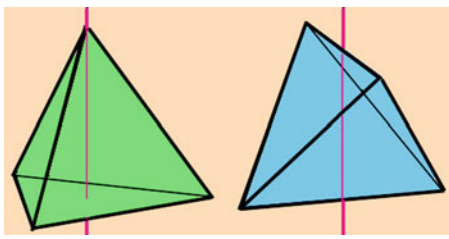
For some polyhedra there are only 2-fold axes of symmetry, and no way of distinguishing one of them as being "principal". This kind of symmetry is labelled as D_2, and exemplified by the two polyhedra shown below.



In addition to the polyhedra with symmetry D_2 there are other cases where there is no preferred axis of rotation. This includes the five regular polyhedra. This kind of symmetry is called *spherical*. The five types of spherical symmetry are listed below.

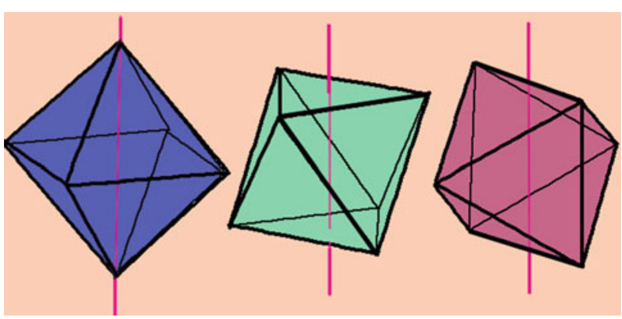
20.6.1.2 Tetrahedral Symmetry *T*

The regular tetrahedron has seven axes of symmetry, four 3-fold and three 2-fold, as shown in the illustration below. A polyhedron with this system of rotational symmetries is said to have *tetrahedral symmetry*. The system is labelled by T.



20.6.1.3 Octahedral Symmetry *O*

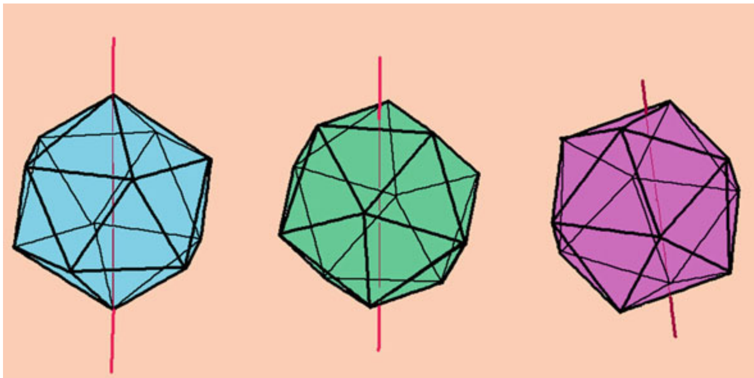
The regular octahedron has three types of rotational axes: Three mutually perpendicular 4-fold axes, passing through opposite vertices. Then four 3-fold axes pass through the centers of opposing faces. Finally six 2-fold axes pass through midpoints of opposing edges. The situation is shown in the illustration below. The system is labelled O.



The last system of direct symmetries a polyhedron can have is the

20.6.1.4 Icosahedral Symmetry I

The regular icosahedron has 2-fold, 3-fold and 5-fold symmetries. There are six 5-fold axes, passing through each pair of opposing vertices, ten 3-fold axes passing through the centers of each pair of opposing faces and 15 2-fold axes passing through the midpoints of opposing edges. The system is labelled I.

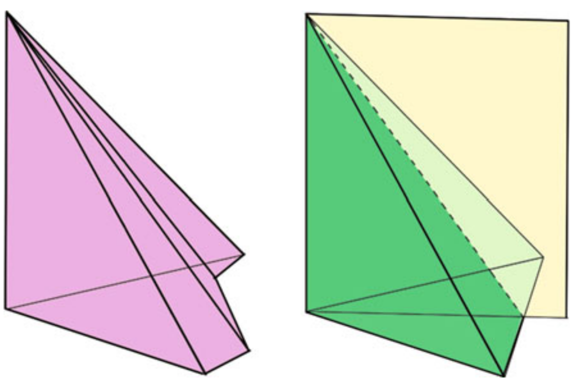


We now turn to the *indirect* symmetries.

20.6.2 Reflection Symmetry

20.6.2.1 Bilateral Symmetry C_s

Below we see two polyhedra, both are skew pyramids. The pyramid to the left has no symmetry except for the identity, it is *asymmetrical*. This type is denoted by C_1, one might say a *one-gonal rotation*. The pyramid to the right also has no rotational axis, but it does allow reflection in the plane shown on the figure. This is called *bilateral symmetry*, and denoted by C_s. Note that only one mirror plane is possible in this case since the composition of reflections in two different planes yields a rotation.

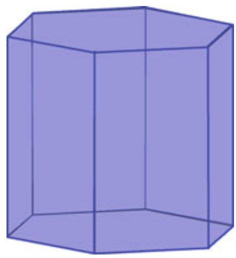


20.6.3 Prismatic Symmetry Types

These symmetry types are illustrated by the symmetries of prisms with certain decorations.

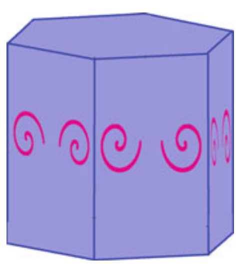
20.6.3.1 Symmetry Type D_{nh}

This is the symmetry type of an unmarked prism, the n indicates n-gonal rotational axis, h indicates the existence of a horizontal mirror plane. There are six axes of 2-fold rotational symmetry.

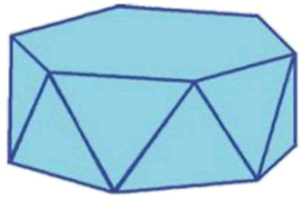


20.6.3.2 Symmetry Type D_{nv}

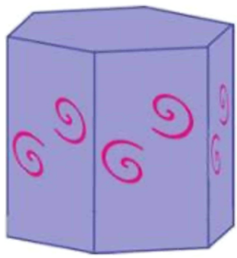
This is the symmetry type of a prism marked as shown, the n indicates n-gonal rotational axis, v indicates the existence of vertical mirror planes. Comparing this to the unmarked hexagonal prism, we see that the horizontal plane is no longer a plane of symmetry, and that the planes passing through opposite edges are also not planes of symmetry. For the unmarked prism there were three 2-fold rotational axes joining the centers of opposite faces, they are not present here while the remaining three 2-fold rotational axes joining midpoints of opposing (vertical) sides are still present. Finally the principal axis is reduced from a 6-gonal axis to a 3-gonal one. The symmetry type of this decorated prism is denoted by D_{3v}.



The hexagonal antiprism has the symmetry type D_{6v}. The rotational symmetries are the same as for the unmarked hexagonal prism, but there is no horizontal mirror plane.

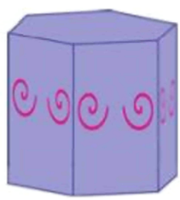


20.6.3.3 Symmetry Type D_n



The symmetry type of n fold rotations is labelled D_n. When a hexagonal prism is decorated as shown above, then all reflectional symmetries disappear but all the rotational symmetries remain. Here the decorated polyhedron has symmetry type D_6.

20.6.3.4 Symmetry Type C_{nv}

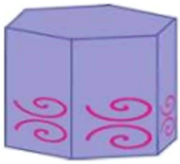


When a hexagonal prism is decorated as shown above, there are no 2-fold axes any more, only an n-fold one. Thus the system is not dihedral, but cyclic. There is no horizontal mirror plane, but there are vertical ones, signified by the v.

20.6.3.5 Symmetry Type C_{nh}

Here the rotational symmetry is only the n fold axis, so the symmetry type is cyclic. There are no vertical mirrors, but there is a horizontal one, signified by the h.

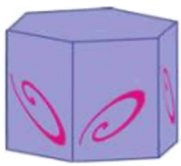
20.6.3.6 Symmetry Type C_n



When the prism is decorated as shown above, only the 6-gonal rotational symmetry remains. This object has symmetry C_6.

20.6.4 Compound Symmetry and the S_{2n} Symmetry Type

This is the analogue of a *glide reflection* from the 2-dimensional case.

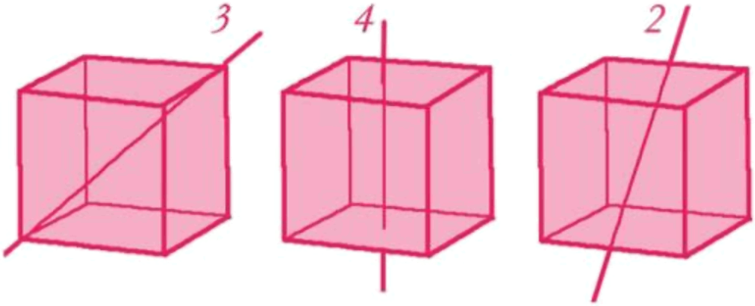


This example has S_{2n} symmetry, here with $n = 3$. For more details we refer to the source [9].

20.6.5 Cubic Symmetry Types

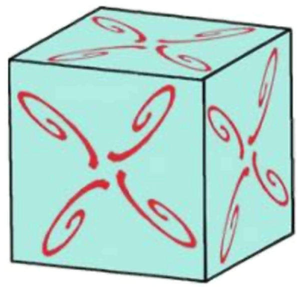
20.6.5.1 Rotational Symmetries of the Cube O

The system of rotational symmetries of a cube is the same as that of the octahedron, namely the system labelled O:



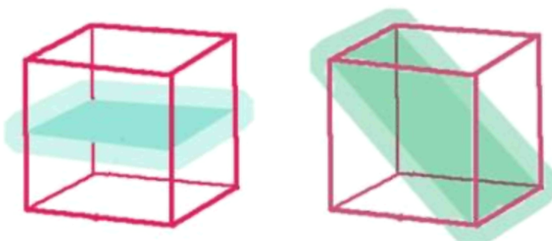
There are four axes of 3 fold rotations, three axes of 4 fold rotations and six axes of 2-fold rotations.

The following decorated cube has only the rotational symmetries O:



The snub cube as well as the octahedron has this symmetry type.

The reflectional symmetries of a cube are given by three mirror planes, each containing two of the 4-fold axes, one of them is indicated to the left, and six mirror planes, each containing two of the 3-fold axes and one of the 2-fold ones, one being indicated to the right below. The cube has a center of inversion. Such a system is labelled O_h. This is the complete set of symmetries of an undecorated cube.



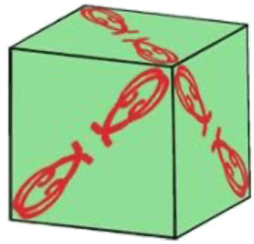
20.6.5.2 Symmetry Type T_h

When the cube is decorated as shown below, only the 3-fold rotational symmetries remain unchanged. The 2-fold axes joining midpoints of opposite edges have disappeared, and the 4-fold axes are reduced to 2-fold ones. This system of rotations is the same as for the tetrahedron.

The three mutually perpendicular mirror planes are still present, but the other ones are gone. There is a point of inversion. The system is labelled T_h.

20.6.5.3 Symmetry Type T_d

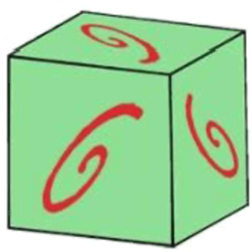
We next consider a cube decorated as shown below.



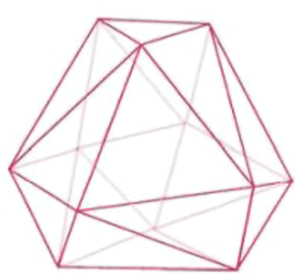
This decorated cube also has the same system of rotations as the regular tetrahedron, but in this case it is the *orthogonal* planes of reflection of the undecorated cube which are gone, and the *skew* ones remain. Here the full system of symmetries is the same as that of the regular tetrahedron, it is labelled T_d.

20.6.5.4 Symmetry Type T

When the cube is decorated as shown below, the only symmetries are the rotational symmetries of the regular tetrahedron. The system is labelled T.



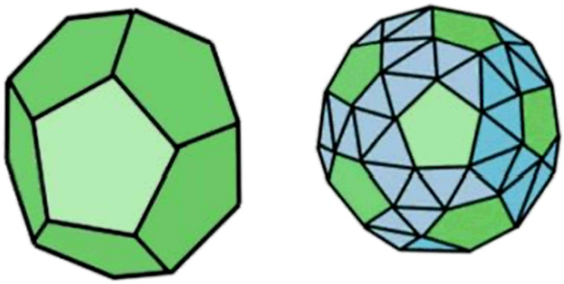
The polyhedron shown below, which is neither regular nor semi-regular, has this system of symmetry. In [9] it is referred to as a "snub tetrahedron", somewhat misleading as snubification of a regular tetrahedron yields an icosahedron. Actually the polyhedron shown below is better understood as a certain deformation of the regular icosahedron:



20.6.5.5 Icosahedral Symmetry

The icosahedral symmetry types are tied to the regular icosahedron, as the name indicates.

Recall that there are 5-fold axes, 3-fold axes and 2-fold axes. There are two types of full icosahedral symmetry types: One type contains only the rotational symmetries, it is labelled I, the other has planes of reflection, it is labelled I_h. The regular dodecahedron has symmetry type I_h, while the snub dodecahedron has type I:



Here we can see clearly how the reflectional symmetries of the dodecahedron to the left are no longer present for the snubification to the right, while all the rotational symmetries remain intact.

20.6.6 The Possible Symmetry Types

We study the rotational symmetries of any polyhedron, not necessarily regular or convex. To each n-fold axis of a polyhedron P there are associated two *poles*, namely the two points where the axis meets the surface of the polyhedron. Two poles of the polyhedron P are said to be equivalent if there is a symmetry of P carrying one into the other. The poles thus fall into separate equivalence classes, see Definition 7 in Sect. 11.1. We have the following observation:

Lemma 8. *Let N be the number of rotational symmetries of the polyhedron P. Assume that P has one or more axes of n-fold rotations. Then there are exactly $\frac{N}{n}$ poles in each equivalence class.*

Proof. [1] Indeed, place a point p close to an n-pole and consider the set of all images of p under the symmetries of P. We then obtain a total of N points on P arranged in groups of n around each equivalent pole. Thus there are $\frac{N}{n}$ such groups, and hence the total number of poles in each class is $\frac{N}{n}$, as claimed. □

We now prove the following important and quite amazing result, following [9]:

Theorem 40. *The rotational symmetries of a polyhedron is either cyclic, dihedral, tetrahedral, octahedral or icosahedral.*

Proof. [2] Using Lemma 8 we find that the total number of non-identity rotations is

$$\frac{1}{2}\sum_{\text{poles}}\frac{N}{n}(n-1)$$

since the number $\frac{N}{n}$ counts each rotation twice, once for each pole on the axis. Clearly we may assume that $n \geq 2$, as the case $N = 1$ is trivial and corresponds to symmetry type C_1. Thus

$$2(N-1) = \sum_{\text{poles}}\frac{N}{n}(n-1)$$

or

$$2 - \frac{2}{N} = \sum_{\text{poles}}(1 - \frac{1}{n})$$

where all $n \geq 2$. We thus have

$$\frac{1}{2} \leq 1 - \frac{1}{N} < 1 \text{ and } \frac{1}{2} \leq 1 - \frac{1}{n} < 1.$$

If there were only one equivalence class of poles, then $\sum_{\text{poles}}(1 - \frac{1}{n}) < 1$ while $2 - \frac{2}{N} \geq 1$, which is impossible. So there are at least two equivalence classes of poles.

The system cannot have four or more equivalence classes of poles, since we would then have

$$\sum_{\text{poles}}\left(1 - \frac{1}{n}\right) \geq \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2,$$

a contradiction.

Now assume that there are two equivalence classes, corresponding to p- and q-fold rotations. From the relation above we find

[1] From [9, p. 297]. See this reference for more details and a very informative illustration.

[2] From [9, pp. 298–300].

$$2 - \frac{2}{N} = \left(1 - \frac{1}{p}\right) + \left(1 - \frac{1}{q}\right)$$

or

$$2 = \frac{N}{p} + \frac{N}{q}$$

which since the two fractions to the right are in fact integers, representing the number of poles in the respective equivalence classes, and consequently these numbers must both be 1. So $N = p = q$. The system has two poles and a single axis, and the symmetry type is C_p.

It remains to treat the case of three equivalence classes of poles, say p-, q- and r-poles. As for the case of two classes we get

$$1 + \frac{2}{N} = \frac{1}{p} + \frac{1}{q} + \frac{1}{r}.$$

Here at least one of p, q and r must be 2, since otherwise

$$1 + \frac{2}{N} \leq \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$$

which is impossible. Assume $r = 2$, and $p \geq q$. Then

$$1 + \frac{2}{N} = \frac{1}{p} + \frac{1}{q} + \frac{1}{2}$$

which after a short calculation yields

$$(p-2)(q-2) = 4\left(1 - \frac{pq}{N}\right) < 4$$

from which follows that

$$(p-2)(q-2) = 0, 1, 2, \text{ or } 3.$$

When $(p-2)(q-2) = 0$ we find $q = 2$, and there are three equivalence classes of poles: Two classes of 2-poles and one of p-poles. This is the dihedral system D_p. The cases $(p-2)(q-2) = 1, 2$ or 3 yield $(p, q) = (3, 3), (4, 3)$ and $(5, 3)$, corresponding to the systems T, O and I. This completes the list of the possible rotational symmetry types. □

It follows from the above that a polyhedron can only have one of the following symmetry types:

$$C_1, C_i, C_s$$

$$C_n, C_{nv}, C_{nh}, D_n, D_{nv}, D_{nh}, S_n$$

$$T, T_d, T_h, O, O_h, I, I_h.$$

The symmetry type may be determined by various schematic procedures, summarizing the information gained by the above. One such procedure is given in [9].

Using the information gained above, the symmetry types of the semi-regular polyhedra we have found are as follows:

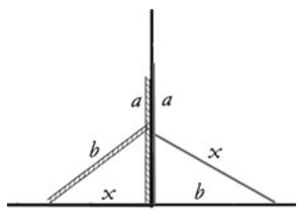
1. Cube O_h
2. Cuboctahedron O_h
3. Dodecahedron I_h
4. Great rhombicosidodecahedron I_h
5. Great rhombicuboctahedron O_h
6. Icosahedron I_h
7. Icosidodecahedron I_h
8. Octahedron O_h
9. Small rhombicuboctahedron T_h
10. Small rhombicosidodecahedron I_h
11. Snub cube O
12. Snub dodecahedron I
13. Tetrahedron T
14. Truncated cube O_h
15. Truncated dodecahedron I_h
16. Truncated icosahedron I_h
17. Truncated tetrahedron T_d
18. Truncated octahedron O_h

A closer examination reveals that they are all vertex transitive. Miller's polyhedron has symmetry group D_{4v}, which has order 16. However, the polyhedron has 24 vertices. Hence it is not vertex transitive.

Chapter 21
Hints and Solutions to Some of the Exercises

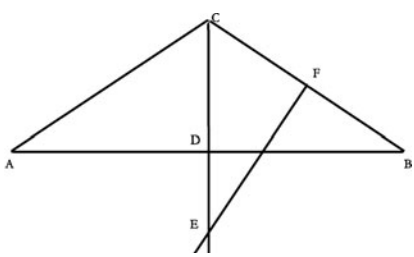
Exercise 2.1: If the radius of the circle is 1, the area of the circumscribed square is 4, while the area of the inscribed square is $(\sqrt{2})^2 = 2$. The average is 3, which is therefore the corresponding value for π.

Exercise 2.2:



(a) To the left the ladder stands upright against a wall, and is then allowed to slide down a known a distance a. It is clear that we must have $0 \leq a \leq b$. Pythagoras yields $x = \sqrt{2ab - a^2}$.
(b) To the right a reed of unknown length stands up against a wall, and then slides down a distance a, while the lower end moves a to the distance b from the wall. Pythagoras yields $x = \frac{(a^2+b^2)}{2a}$.

Exercise 2.3:



Here $AC = BC = b$ and $AB = a$. Moreover$\angle ADC = \angle CFE = 90°$ and $\angle CEF = \angle CBD$ as their legs are pairwise perpendicular. Thus the two right triangles $\triangle CEF$ and $\triangle CBD$ are similar, and we find $CE : CB = CF : CD$. The unknown radius is $r = CE$, and we therefore get $r = b\frac{\frac{b}{2}}{m}$ where $m = CD = \sqrt{b^2 - (\frac{a}{2})^2}$. Thus

A. Holme, *Geometry*, DOI 10.1007/978-3-642-14441-7_21,

$r = \frac{b^2}{\sqrt{4b^2 - a^2}}$. With the values $a = 60$, $b = 50$, or sexagesimally $a = (50), b = (1)(0)$, we get $r = (31) \cdot (15)$.

Exercise 2.4: We have the relations $xy = A, x^3 = B$ and $x^2 + y^2 = d^2$. Substituting $y = \frac{A}{x}$ and $d = \frac{B}{x^3}$ into the last relation we get $x^2 + \frac{A^2}{x^3} = \frac{B^2}{x^6}$. This yields the following equation for $z = x^4$:

$$z^2 + zA^2 = B^2$$

which yields

$$x^4 = z = -\frac{A^2}{2} + \sqrt{\left(\frac{A^2}{2}\right)^2 + B^2} = \frac{\sqrt{A^4 + 4B^2} - A^2}{2}$$

For the first set of values, $x = 4$ and $y = 3$. For the values from the tablet, we only get an approximate answer. To compute it by modern means, we may use a modern calculator or MAPLE, and obtain the following simple worksheet:

```
> z:=((sqrt(A^4+4*B^2)-A^2)/2);
                    z := 1/2 sqrt(A^4 + 4 B^2) - 1/2 A^2
> A:=20*60;
                    A := 1200
> B:=14*60^3+28*60^2+53*60+20;
                    B := 3128000
> z;
                    1/2 sqrt(41211136000000) - 720000
> x:=evalf(root[4](z));
                    x := 39.72289552

> x1:=x-39;
                    x1 := 0.72289552
> x1:=x1*60;
                    x1 := 43.37373120
> x2:=x1-43;
                    x2 := 0.37373120
> x2:=x2*60;
                    x2 := 22.42387200
> x3:=x2-22;
                    x3 := 0.42387200
> x3:=x3*60;
                    x3 := 25.43232000
```

Thus $x \approx (39) \cdot (43)(22)(26)$. We omit the value of y and its sexagesimal form, which can be found similarly.

Exercise 2.5: $A = \frac{1}{4}(a+b)\sqrt{4s^2-(a-b)^2}$. With the numbers given we find $A = (12)\cdot(48)$.

Exercise 2.6: We have the relations $\frac{x+y}{2}h = A$ and $\frac{y}{x} = \frac{a-h}{a}$. Thus $y = \frac{a-h}{a}x$ and the first relation yields $\frac{2a-h}{a}x = \frac{2A}{h}$, thus $x = \frac{2Aa}{(2a-h)h}$. Thus $y = \frac{a-h}{a}\frac{2Aa}{(2a-h)h} = \frac{2A(a-h)}{(2a-h)h}$. Substituting the given numbers yields $x = 20$ and $y = 12$.

Exercise 2.7: Denote the side of the larger square by x, the side of the smaller by y. Then we have the relations

$$x^2 + y^2 = A \text{ and } y = \frac{2}{3}x - 10$$

Exercise 2.8: Letting the sides be x and y we get the relations

$$xy = A \text{ and } 2(x+y) = B.$$

Exercise 2.9: If r is the radius of the circle, then the perimeter of the regular inscribed hexagon is $6r$. If the ratio of the perimeter and circumference in question is ρ, then the circumference of the circle is $6r/\rho$. Now $(0)\cdot(57)(36) = \frac{6r}{2\pi r} = \frac{3r}{\pi r}$, thus since $(0)\cdot(57)(36) = \frac{57}{60} + \frac{36}{3{,}600} = \frac{3{,}456}{3{,}600}$ we find $\pi \approx \frac{3\cdot 3{,}600}{3{,}456} = 3.125(3)\cdot(7)(30)$.

Exercise 2.10: We find that the sides are $a = \sqrt{(\frac{Aq}{p})}$ and $b = \frac{p}{q}\sqrt{(\frac{Aq}{p})}$. Substituting the given values yields $a = 4$, $b = 3$.

Exercise 2.11: Letting the sides be x and $y = mx$ we find $x = \sqrt{(\frac{A}{2m})}$. With the given numbers $x = 4$, $y = 10$.

Exercise 2.12: Obviously incorrect even for trapezoids. The method gives the correct area only for rectangles.

Exercise 2.13: $\pi \approx 3.1605\ldots$

Exercise 3.1: Compute as follows:

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

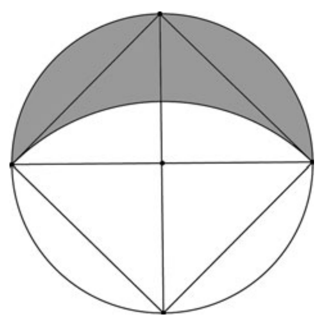
As part of the computation above we found $(\sqrt{2}-1) = \frac{1}{2+(\sqrt{2}-1)}$ and thus we may continue with

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

and so on. Thus $\sqrt{2} = 1; [2]$. The next case:

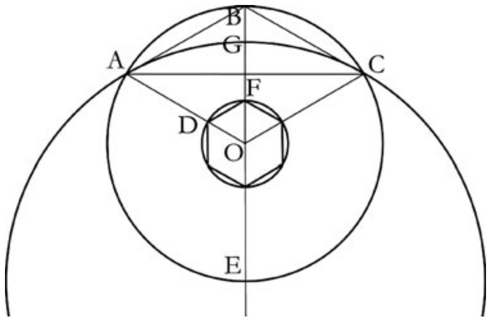
$$\sqrt{3} = 1 + (\sqrt{3} - 1) = 1 + \frac{(\sqrt{3}-1)(\sqrt{3}+1)}{(\sqrt{3}+1)} = 1 + \frac{3-1}{(\sqrt{3}+1)}$$
$$= 1 + \frac{2}{(\sqrt{3}+1)}$$
$$1 + \frac{1}{\frac{\sqrt{3}+1}{2}} = 1 + \frac{1}{1 + \frac{\sqrt{3}+1}{2} - 1} = 1 + \frac{1}{1 + \frac{\sqrt{3}-1}{2}} = 1 + \frac{1}{1 + \frac{(\sqrt{3}-1)(\sqrt{3}+1)}{2(\sqrt{3}+1)}}$$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

Here we see that the process will repeat itself, and we have shown that $\sqrt{3} = 1; [1, 2]$. Going on, we find, for example, $\sqrt{5} = 2; [4]$, $\sqrt{6} = 2; [2, 4]$, $\sqrt{7} = 2; [1, 1, 1, 4]$, $\sqrt{10} = 3; [6]$.



Exercise 3.2: The moon is shown above. If the radius of the half circle is r, then its area is $\frac{1}{2}\pi r^2$. The radius of the quarter circle is $r\sqrt{2}$, and thus that area is $\frac{1}{4}(r\sqrt{2})^2 = \frac{1}{2}\pi r^2$. Thus the area of the moon is equal to the area of the lower triangle, which is $2(\frac{1}{2}r)r = r^2$. Hence the moon in question may be squared.

Exercise 3.3: Referring to the illustration given in the exercise we use the fact found by Hippocrates, that the relation of the areas of two similar circle segments is equal to the relations between the squares on their cords. Therefore, the sum of the areas of the three small segments is equal to the large one, and the area of the moon is therefore equal to the area of the trapezoid, which may easily be squared.



Exercise 3.4: See the figure above. We draw two circles with center O, with radius r and $R = r\sqrt{6}$, respectively. We inscribe a regular hexagon in the small circle and draw a line from O to through the upper left corner of this hexagon, meeting the outer circle in the point A. The points B and C are found similarly. Draw AB, AC and BC. Construct the point E, and draw the circle with center E through A (and thus C). Now $\triangle OAB$ is isosceles, and so is $\triangle ACE$, therefore the circular segment ACA is similar to the circular segment ABA. The ratio between these two segments is equal to the ratio between the squares on the corresponding cords. Now $AC^2 : AB^2 = 3 : 1$. In fact, $AB^2 = AF^2 + FB^2 = (\frac{AC}{2})^2 + (\frac{OB}{2})^2$, and since $OB = AB$, $3AB^2 = AC^2$. Thus the segment $ACGA$ has three times the area as the segment ABA. On the other hand, the relation between the area of the segment AFD and the area of ABA is equal to the ratio between the squares on their diameters, that is to $1 : 6$. Thus the area of the segment $ACGA$ is equal to the sum of the six small segments of the inner circle *and* the segments ABA and BCB. Hence the area of the full and the crescent moon combined is equal to the combined areas of the inscribed hexagon and the $\triangle ACB$. As the latter combination may be squared, so may the former.

In the two previous exercises we saw how crescent moons of different shapes may be squared. It is understandable if Hippocrates and others may have been led to hope that the crescent moon in this exercise might be squarable as well. If so, the circle could be squared. Indeed, the difference between two squares may be squared: If the two squares have sides a and b, then $a^2 - b^2 = (a + b)(a - b)$, the area of a rectangle. It is explained in Sect. 3.3, and also among many other things in Chap. 17, that if $x^2 = \alpha\beta$ where α, β are constructed or given line segments, then x is constructible.

Exercise 4.1: A point (x, y) on the spiral is given by

$$x = \frac{r}{2\pi} v \cos(v)$$
$$y = \frac{r}{2\pi} v \sin(v)$$

where v is the angle which the rotating line forms with the x-axis. This is the spiral in polar coordinates. We further have that

$$\mathrm{d}x = \frac{r}{2\pi}(-v\sin(v) + \cos(v))\mathrm{d}v$$
$$\mathrm{d}y = \frac{r}{2\pi}(v\cos(v) + \sin(v))\mathrm{d}v$$

At the point P we then find the slope of the tangent by substituting $v = 2\pi$ in

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\frac{r}{2\pi}(v\cos(v) + \sin(v))}{\frac{r}{2\pi}(-v\sin(v) + \cos(v))}$$

which yields

$$\left[\frac{\mathrm{d}y}{\mathrm{d}x}\right]_{v=2\pi} = 2\pi$$

The equation of the tangent at P therefore becomes

$$y = 2\pi(x - r)$$

and letting $x = 0$, we find the point of intersection between this line and the y-axis: It is the point $(0, -2\pi r)$, and the claim is proven.

As for the last question, the difficulty with using a similar method to treat the volume of a sphere, is that for the circle all triangles will have the seme height, whereas the individual pyramids will not be of the same height. Although all heights tend to the radius of the sphere in the limit as the subdivision of the sphere's surface is refined in such a way that the all individual pieces tend to zero, we still have a problem. If you try to work this out with ϵ and δ, you will see why.

Exercise 4.3: The area of the "shoemaker's knife" evidently is

$$\frac{\pi}{2}\left(\left(\frac{AB}{2}\right)^2 - \left(\frac{AN}{2}\right)^2 - \left(\frac{NB}{2}\right)^2\right)$$

which since $AB = AN + NB$, i.e., equal to

$$\frac{\pi}{2}\left(\frac{1}{2}AN \cdot NB\right)$$

and NP is the mean proportional between AN and NB this equals

$$\frac{\pi}{4}NP^2$$

in other words, the area of the circle with diameter NP.

As for the salinon we denote the radii by a, b, c, so $AO = a, AD = 2b = CB$ and $DO = c$. Thus the area is

$$A_1 = \frac{\pi}{2}(a^2 - 2b^2 + c^2).$$

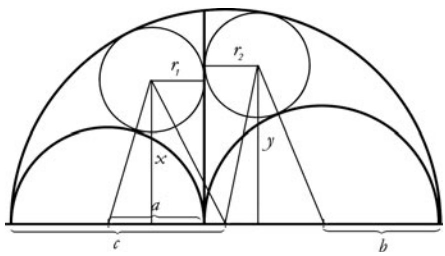
The area of the circle of diameter EF is

$$A_2 = \pi(\frac{a+c}{2})^2 = \frac{\pi}{4}(a^2 + 2ac + c^2).$$

Now $a = 2b + c$ so $b = \frac{a-c}{2}$ thus $b^2 = \frac{1}{4}(a^2 - 2ac + c^2)$. Substituted into the expression for A_1 this yields

$$A_1 = \frac{\pi}{2}(a^2 - \frac{1}{2}(a^2 - 2ac + c^2) + c^2) = \frac{\pi}{4}(a^2 + 2ac + c^2) = A_2.$$

Exercise 4.4: It is frequently asserted that this property of the arbelos is substantially more difficult than the computation of the area. Perhaps this may be so, but it really amounts to nothing more than a careful use of Pythagoras in combinations with an elementary property of tangent circles. Indeed, the following solution was handed in by Ms. *Karen Sofie Ronæss*, one of the students in my class in the spring of 2007 at the University of Bergen. We use notations as indicated below:



To show is that $r_1 = r_2$. By Pythagoras the two right triangle yield the following expressions for x^2:

$$x^2 = (a + r_1)^2 - (a - r_1)^2 = 4ar_1,$$

and

$$x^2 = (c-r_1)^2-(c-2a+r_1)^2 = c^2-2cr_1+r_1^2-(c^2+4a^2+r_1^2-4ac+2cr_1-4ar_1)$$

$$= 4ac - 4cr_1 + 4ar_1 - 4a^2.$$

The only proposition from circle geometry used in this proof, is the simple observation that when two circles are tangent to each other, then the point of tangency and their two centers are collinear. Combined the two expressions above yield

$$4ar_1 = 4ac - 4cr_1 + 4ar_1 - 4a^2$$

i.e.

$$cr_1 = ac - a^2 = ab, \text{ that is to say, } r_1 = \frac{ab}{c}.$$

Reflection in the line with a change of size for the semi-circle and replacing a by b, r_1 by r_2 and x by y, yields free of charge that

$$cr_2 = bc - b^2 = ba \text{ thus } r_2 = \frac{ab}{c}.$$

Hence the claim is proven.

It is interesting that for instance *Howard Eves* writes, in [14, p. 189], having treated some simple properties of the arbelos, that *"The arbelos has many properties not so easily established. For example, it is alleged that Archimedes showed that the circles inscribed in the curvilinear triangles [...] are equal [...]"*. This allegation seems to be easily believable.

Exercise 4.5: Let H be the foot of the normal from O to AC, and denote O_3H by y, OH by x. The circle about O with radius r is tangent to the one about O_1 at the point T_1, to the circle about O_2 at T_2 and the one about O_3 at T_3. Again we use the simple property of the point of tangency being collinear with the two centers, and conclude that

$$O_3O = a + r, O_1O = a + b - r \text{ and } O_2O = b + r.$$

Hence we obtain the following three equalities from Pythagoras:

$$x^2 + y^2 = (a + r)^2$$

$$x^2 + (b - y)^2 = (a + b - r)^2$$
$$x^2 + (a + b - y)^2 = (b + r)^2$$

They give

$$x^2 + y^2 - r^2 = 2ar + a^2$$
$$x^2 + y^2 - r^2 = -2(a + b)r + a^2 + 2ab + 2by$$
$$x^2 + y^2 - r^2 = 2br - a^2 - 2ab + 2(a + b)y$$

which again imply the two relations

$$2ar + a^2 = -2(a + b)r + a^2 + 2ab + 2by$$

$$2ar + a^2 = 2br - a^2 - 2ab + 2(a + b)y$$

thus

$$2ar + a^2 + 2(a + b)r - a^2 - 2ab = 2by$$
$$2ar + a^2 - 2br + a^2 + 2ab = 2(a + b)y$$

or

$$r(2a + b) - ab = by$$
$$r(a - b) + a^2 + ab = (a + b)y.$$

Thus we obtain

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

which yields

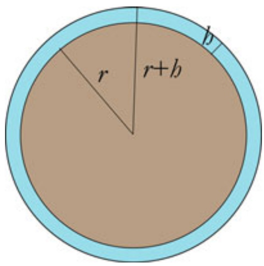
$$r((a + b)(2a + b) - b(a - b)) = ab(a + b) + b(a^2 + ab)$$

that is to say

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

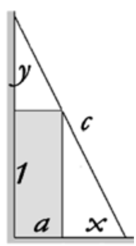
from which the claim follows.

Exercise 4.6: The area of the planet's surface is $A = 4\pi r^2$ while the area of the limit of the atmosphere is $B = 4\pi(r + h)^2$.



The volume V of the atmosphere is the same as the volume of a frustum of a pyramid with height h and bottom and top equal to A and B, respectively. Thus we have $V = \frac{1}{3}h(A + \sqrt{AB} + B)$, which answers the question of the exercise. But moving on, we find$V = \frac{4}{3}\pi h(r^2 + r(r + h) + (r + h)^2)$. Thus we finally get $V = \frac{4}{3}\pi h(3r^2 + 3hr + h^2)$. Normally the height h is insignificant in comparison to r, so a practical approximation is $V \approx 4\pi h(r^2 + hr)$, or perhaps even $V \approx 4\pi hr^2$, the last of which is frequently used without any second thoughts.

Exercise 4.7: We may scale the situation so that $b = 1$ without loss of generality.



If we wish to interpret our final formula for x with some other value of b, then the formula must be interpreted as giving the ratio $\frac{x}{b}$. Moreover in the formula a must be interpreted as $\frac{a}{b}$ and c as $\frac{c}{b}$.

We now introduce the distance from the top of the ladder to the top of the box as a new variable y, as shown in the figure to the left. Then $(y+1)^2+(x+a)^2 = c^2$ by the Pythagorean Theorem. Moreover, the similar right triangles with bases a and x

yield $y = \frac{a}{x}$. Substituting this in the relation and multiplying by x^2 we get the equation $a^2 + 2ax + x^2 + a^2x^2 + 2ax^3 + x^4 = c^2x^2$ and thus

$$x^4 + 2ax^3 + x^2(1 + a^2 - c^2) + 2ax + a^2 = 0.$$

In general the roots of a degree four equation with integral coefficients cannot be constructed by ruler and compass used in the legal fashion, as explained in Sects. 17.2 and 17.3.

However, in the case when $a = b$ the situation is different. By our normalization this yields $a = b = 1$, and thus $y = \frac{1}{x}$, thus Pythagoras yields

$$(x + 1)^2 + \left(\frac{1}{x} + 1\right)^2 = c^2$$

or

$$x^2 + 2x + 1 + \frac{1}{x^2} + 2\frac{1}{x} + 1 = c^2$$

and hence letting $d = x + \frac{1}{x}$ we find the equation $d^2 + 2d - c^2 = 0$, so in this case the line-segment $d = x + y$ is constructible since roots of equations of degree two with constructible coefficients are constructible. But

$$x^2 - dx + 1 = 0,$$

and so x itself is constructible, for the same reason.

It should finally be pointed out, in case the reader has not already observed it, that in the case $a = b$ there are in general *two different* solutions to the problem. In fact, the solution is unique only for $c = 2a\sqrt{2}$. Why?

Exercise 4.9: We have $\alpha + \beta + \gamma = \pi$. To show is that

$$\sin^2(\alpha) = \sin^2(\beta) + \sin^2(\gamma) - 2\sin(\beta)\sin(\gamma)\cos(\alpha).$$

For this, draw an arbitrary line AB, at A and B draw lines forming angles α and β with AB. These lines intersect at a point C, since the angular sum of $\triangle ABC$ is π, $\angle ACB = \gamma$. Let R be the circumradius of $\triangle ABC$, then by the Rule of Sines

$$AB = 2R\sin(\gamma), BC = 2R\sin(\alpha), CA = 2R\sin(\beta).$$

Now by the Law of Cosines, $BC^2 = CA^2 + AB^2 - 2 \cdot CA \cdot AB\cos(\alpha)$, thus

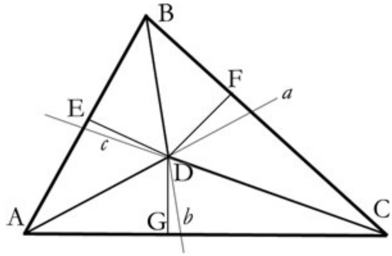
$$4R^2\sin^2(\alpha) = 4R^2\sin^2(\beta) + 4R^2\sin^2(\gamma) - 2 \cdot 4R^2\sin(\beta)\sin(\gamma)\cos(\alpha),$$

hence the claim follows.

Exercise 4.11: The lines n_1, n_2 and n_3 intersect in one point because any two of them intersect in a point which is equidistant from all vertices A, B and C. For the same reason this point is the circumcenter of $\triangle ABC$.

Exercise 4.12: Use the Rule of Sines and $S = \frac{1}{2}cb\sin(A)$.

Exercise 4.13:



All points on the bisector a of $\angle BAC$ are equidistant from the sides AC and AB, similarly the points on the bisector b are equidistant from AB and BC, and the points on c from AC and BC. Thus they intersect in one point, namely the center D of the inscribed circle. Letting R be the radius of the circumscribed circle, the Rule of Sines yields $\frac{\sin(2\alpha)}{BC} = \frac{\sin(2\beta)}{AC} = \frac{\sin(2\gamma)}{AB} = \frac{1}{2R}$. Thus we find $\frac{\sin(\alpha)\cos(\alpha)}{BC} = \frac{\sin(\beta)\cos(\beta)}{AC} = \frac{\sin(\gamma)\cos(\gamma)}{AB} = \frac{1}{4R}$. This yields

$$\frac{1}{4R} = \frac{\sin(\alpha)\cos(\beta)}{BC} = \frac{\sin(\alpha)\cos(\beta)}{BF + FC} = \frac{\sin(\alpha)\cos(\alpha)}{r(\frac{\cos(\beta)}{\sin(\beta)} + \frac{\cos(\gamma)}{\sin(\gamma)})}$$

using the Rule of Sines for $\triangle FBD$ and $\triangle FCD$, and the relation $\sin(\frac{\pi}{2} - v) = \cos(v)$. Thus

$$\frac{1}{4R} = \frac{1}{r}\frac{\sin(\alpha)\cos(\alpha)\sin(\beta)\sin(\gamma)}{\cos(\beta)\sin(\gamma) + \sin(\beta)\cos(\gamma)} = \frac{1}{r}\frac{\sin(\alpha)\cos(\alpha)\sin(\beta)\sin(\gamma)}{\sin(\beta + \gamma)}$$

Since now $\alpha + \beta + \gamma = \frac{\pi}{2}$, we have $\sin(\beta + \gamma) = \cos(\alpha)$, and thus finally

$$\frac{1}{4R} = \frac{1}{r}\sin(\alpha)\sin(\beta)\sin(\gamma)$$

from which the claim follows.

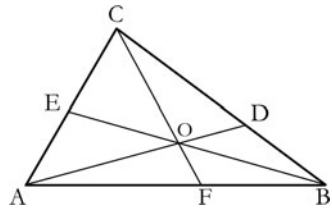
Exercise 4.15: By the Law of Cosines $\cos(C) = \frac{a^2+b^2-c^2}{2ab}$. Thus

$$S = \frac{1}{2}ab\sin(C) = \frac{1}{2}ab\sqrt{1 - \cos^2(C)} = \frac{1}{2}ab\sqrt{1 - (\frac{a^2 + b^2 - c^2}{2ab})^2}$$

$$= \frac{1}{4}\sqrt{4a^2b^2 - (a^2+b^2-c^2)^2} = \frac{1}{4}\sqrt{(2ab - (a^2+b^2 - c^2))(2ab + (a^2+b^2-c^2))}$$
$$= \frac{1}{4}\sqrt{(c^2 - (a - b)^2)((a + b)^2 - c^2)}$$
$$= \frac{1}{4}\sqrt{(c-a+b)(c + a - b)(a + b-c)(a + b+c)}$$

from which the claim follows easily.

Exercise 5.2: Following [61], we have the situation shown in the figure.



First assume that the three lines have the point O in common. The two triangles $\triangle BOD$ and $\triangle COD$ have the same heights, thus $\frac{\text{Area}(\triangle BOD)}{\text{Area}(\triangle COD)} = \frac{BD}{DC}$. Similarly $\frac{\text{Area}(\triangle BAD)}{\text{Area}(\triangle CAD)} = \frac{BD}{DC}$. This implies that $\frac{BD}{DC} = \frac{\text{Area}(\triangle BAD)-\text{Area}(\triangle BOD)}{\text{Area}(\triangle CAD)-\text{Area}(\triangle COD)} = \frac{\text{Area}(\triangle ABO)}{\text{Area}(\triangle CAO)}$. We find two analogous relations, namely $\frac{CE}{EA} = \frac{\text{Area}(\triangle BCO)}{\text{Area}(\triangle ABO)}$ and $\frac{AF}{FB} = \frac{\text{Area}(\triangle CAO)}{\text{Area}(\triangle BCO)}$. Multiplying these three equalities yields $\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = 1$.

To show the converse, assume that $\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = 1$ and let AD and BE intersect at O and CO intersect AB at F'. We then have proved above that $\frac{AF'}{F'B} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = 1$, which together with the assumption yields $\frac{AF'}{F'B} = \frac{AF}{FB}$ thus $\frac{AF'}{F'B} + 1 = \frac{AF}{FB} + 1$ so $\frac{AF'+F'B}{F'B} = \frac{AF+FB}{FB}$ hence $\frac{AB}{F'B} = \frac{AB}{FB}$, therefore $F'B = FB$ so $F' = F$, and the claim is proven.

Exercise 6.1: Indeed, this is quite elementary: Let s denote the product of all the numbers r used in the steps of type (1) during the reduction process from T to $\overline{T}$. If $\overline{T} = T'$ then $(-1)^m as = 1$ where m is the number of times operation (2) has been performed. Since $s \neq 0$, we find $a \neq 0$. Conversely, suppose $a \neq 0$. Since A' has zeroes under the diagonal, we get $\det(A') = a'_{1,1}a'_{2,2} \cdots a'_{n,n} = (-1)^m as$. But since a and s are $\neq 0$, so are all the $a'_{i,i}$. As $\overline{T}$ is on reduced row echelon form, they are all equal to 1 with zeroes above them. So the claim follows. The last part of the exercise is very simple, check the cases $n = 2, 3, 4$ first, then do the general case.

Exercise 6.2: We refer to the figure in the statement of the problem, and let $\angle BAC = 3\alpha$, $\angle ABC = 3\beta$ and $\angle ACB = 3\gamma$. We compute the lengths of s of DF, t of FE and u of ED, verifying that they are equal. Using the Law of Cosines it is enough to compute a_1, a_2, b_1, b_2, c_1 and c_2. Computing in terms of the circumradius R and α, β and γ, the latter three will enter the computations cyclically and thus it suffices to compute any one of them, say $a_1 = AD$, the others may then be written down by symmetry. The important condition tying together α, β, γ is that

$$\alpha + \beta + \gamma = \frac{\pi}{3}$$

The Law of Sines applied to $\triangle ABC$ yields $AB = 2R\sin(3\gamma)$, then the Law of Sines applied to $\triangle ABD$ yields

$$\frac{AD}{\sin(\beta)} = \frac{AB}{\sin(\pi - (\alpha + \beta))} = \frac{2R\sin(3\gamma)}{\sin(\alpha + \beta)} = \frac{2R\sin(3\gamma)}{\sin(\frac{\pi}{3} - \gamma)}$$

so

$$a_1 = \frac{2R\sin(\beta)\sin(3\gamma)}{\sin(\frac{\pi}{3}-\gamma)}$$

By Exercise 11.2 we have $\sin(3\gamma) = 4\sin(\gamma)\sin(\frac{\pi}{3}+\gamma)\sin(\frac{\pi}{3}-\gamma)$, which yields

$$a_1 = \frac{8R\sin(\beta)\sin(\gamma)\sin(\frac{\pi}{3}+\gamma)\sin(\frac{\pi}{3}-\gamma)}{\sin(\frac{\pi}{3}-\gamma)}$$

$$a_1 = 8R\sin(\beta)\sin(\gamma)\sin(\frac{\pi}{3}+\gamma)$$

By symmetry we have

$$a_2 = 8R\sin(\gamma)\sin(\beta)\sin(\frac{\pi}{3}+\beta)$$

and so by the Law of Cosines

$$s^2 = a_1^2 + a_2^2 - 2a_1a_2\cos(\alpha) =$$

$$(8R)^2\sin^2(\beta)\sin^2(\gamma)\{\sin^2(\frac{\pi}{3}+\gamma)+\sin^2(\frac{\pi}{3}+\beta)-2\sin(\frac{\pi}{3}+\gamma)\sin(\frac{\pi}{3}+\beta)\cos(\alpha)\}$$

Now

$$\frac{\pi}{3}+\gamma+\frac{\pi}{3}+\beta+\alpha = \pi,$$

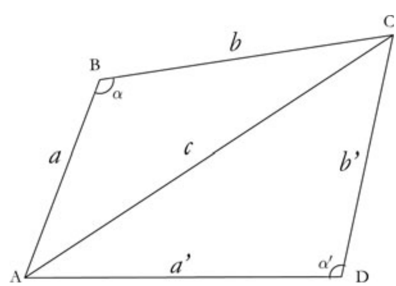
and thus by Exercise 4.9 we have the expression in { } equal to $\sin^2(\alpha)$, thus

$$s = 8R\sin(\alpha)\sin(\beta)\sin(\gamma).$$

By symmetry we have the same expression for t and u, and the proof is complete.

Exercises 6.3: With notation as in the figure we get

$$c^2 = a^2 + b^2 - 2ab\cos(\alpha)$$
$$c^2 = a'^2 + b'^2 - 2a'b'\cos(\alpha').$$



Denote the areas of the two triangles by F and F', then

$$2F = ab\sin(\alpha) \text{ and } 2F' = a'b'\sin(\alpha')$$

and we get the area R of the quadrilateral as

$$R = F + F'$$

We introduce an auxiliary entity

$$G = a^2 + b^2 - (a'^2 + b'^2) = 2ab\cos(\alpha) - 2a'b'\cos(\alpha')$$

The trick is to use G as a catalyst. We compute

$$\begin{aligned}
&G^2 + (4R)^2 \\
&\quad = (2ab\cos(\alpha) - 2a'b'\cos(\alpha'))^2 + (2ab\sin(\alpha) + 2a'b'\sin(\alpha'))^2 \\
&\quad = 4a^2b^2\cos^2(\alpha) - 8aba'b'\cos(\alpha)\cos(\alpha') + 4a'^2b'^2\cos^2(\alpha') \\
&\qquad + 4a^2b^2\sin^2(\alpha) + 8aba'b'\sin(\alpha)\sin(\alpha') + 4a'^2b'^2\sin^2(\alpha') \\
&\quad = 4a^2b^2 + 4a'^2b'^2 - 8aba'b'\cos(\alpha + \alpha')
\end{aligned}$$

Now $\cos(2v) = 2\cos^2(v) - 1$, so that $\cos(\alpha + \alpha') = 2\cos^2(\frac{\alpha+\alpha'}{2}) - 1$ and therefore

$$16R^2 = (2ab + 2a'b')^2 - G^2 - 16aba'b'\cos^2(\frac{\alpha + \alpha'}{2})$$

$$= (2ab + 2a'b' + G)(2ab + 2a'b' - G) - 16aba'b'\cos^2(\frac{\alpha + \alpha'}{2})$$

Further

$$2ab + 2a'b' + G = 2ab + 2a'b' + a^2 + b^2 - (a'^2 + b'^2)$$

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

and in the same way we find

$$2ab + 2a'b' - G = 2ab + 2a'b' - a^2 - b^2 + (a'^2 + b'^2)$$

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

When this is substituted in the expression for $16R^2$ and we divide by 16, we get the so called *Brahmagupta's formula*

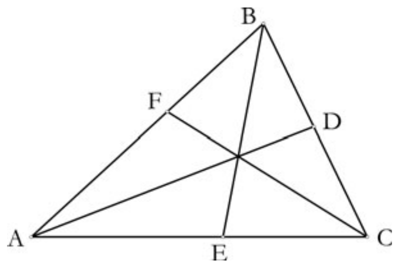
$$R^2 = (s-a)(s-b)(s-a')(s-b') - aba'b'\cos^2(\frac{\alpha+\alpha'}{2})$$

in its general form. Of course $s = \frac{a+b+a'+b'}{2}$.

Exercise 6.4: This is an easy applications of "Ceva's Theorem" from Exercise 5.2.

Exercise 6.5: Left to the reader.

Exercise 6.6: Also by "Ceva's Theorem". We refer to the figure below.



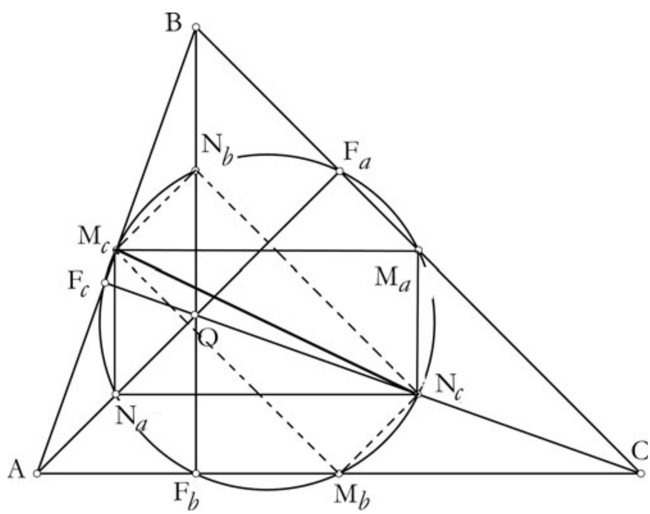
Let the angle at A be 2φ, and denote the opposite sides of A, B, C by a, b, c, respectively. Denote the altitude of $\triangle ABC$ on BC by h. Then

$$\text{Area}(\triangle ABD) = \frac{1}{2}AD \cdot c \cdot \sin(v) = \frac{1}{2}h \cdot BD$$

$$\text{Area}(\triangle ADC) = \frac{1}{2}AD \cdot \sin(v) \cdot b = \frac{1}{2}h \cdot DC$$

From this it follows that $\frac{BD}{DC} = \frac{c}{b}$, and hence by symmetry $\frac{CE}{EA} = \frac{a}{c}$ and $\frac{AF}{FB} = \frac{b}{a}$. Thus the condition in "Ceva's Theorem" holds.

Exercise 6.9: We redraw the figure, with some lines added which will be explained below:



Since $\triangle BM_aM_c \sim \triangle BCA$ we have $M_cM_a \| AC$, similarly $M_cM_b \| BC$ and $M_bM_a \| AB$. By the same argument we also find $M_cN_a \| BO$, so $M_cN_a \| BF_b$, as well as $M_aN_c \| BF_b$. As $BF_b \perp AC$ we therefore finally conclude that the quadrilateral $\square M_cN_aN_cM_a$ is a rectangle, which is therefore inscribed in a circle S, a diameter of which is the diagonal M_cN_c. The symmetric argument yields that $\square M_bN_cN_bM_c$ also is a rectangle, an as it too has M_cN_c as a diagonal, it is inscribed in the same circle S. It now follows that the three foot points F_a, F_b and F_c lie on S. Indeed, since $\angle M_cF_cN_c = \frac{\pi}{2}$, F_c lies on the circle whose diameter is M_cN_c. The symmetric arguments yield the claim for the two other foot points.

Exercise 8.1 and Exercise 8.2: See [27, pp. 240 and 244] for treatments close to Euclid's proofs.

Exercise 9.1: For $m = 2$ we have the matrix

$$\begin{Bmatrix} 1 & 2 \\ 2 & 1 \end{Bmatrix}$$

Since this is not orthogonal to the only alternative, namely

$$\begin{Bmatrix} 2 & 1 \\ 1 & 2 \end{Bmatrix}$$

the claim follows.

Exercise 9.2: We first show how a finite projective plane of order m is used to construct a set of $m-1$ mutually orthogonal Latin squares, and illustrate the procedure for $m = 4$, since this is a power of a prime we know that a projective plane of order m exists.

This plane has altogether $n^2 + n + 1 = 21$ points, and the same number of lines. We refer to Fig. 21.1.

We select a line as the line at infinity, and label it ℓ_∞. On this line we chose two distinct points, V and H, and label the four remaining lines through each of them v_1, v_2, v_3, v_4 and h_1, h_2, h_3, h_4, respectively. $P_{i,j}$ is the point of intersection between h_i and v_j. Besides H and V, the line ℓ_∞ contains three more points, which we label P^1, P^2 and P^3. We start with P^1, besides ℓ_∞ there are four more lines passing through this point, we denote them by ℓ_1, ℓ_2, ℓ_3 and ℓ_4. In the positions where line ℓ_1 meets h_1, h_2, h_3 and h_4 we put the number 1. So this will be in positions $(1, 1)$, $(2, 2)$, $(3, 3)$ and $(4, 4)$. Similarly we determine the positions of 2 from where ℓ_2 meets h_1, h_2, h_3 and h_4, and so on. Obviously we then obtain a Latin square. The second Latin square is obtained by the same procedure by means of the point P^2 and finally a third by P^3. Since lines through P^1, P^2 and P^3 are all distinct, it follows that the Latin squares are mutually orthogonal.

Exercise 9.3: We are given a total of $m-1$ mutually orthogonal $m \times m$ matrices $M_\alpha, \alpha = 1, \dots, m-1$ with entries $1, 2, \dots, m$. We start by introducing m^2 points $P_{i,j}, 1 \le i, j \le m$, and m lines which we denote by $h_1, \dots, h_m$ and

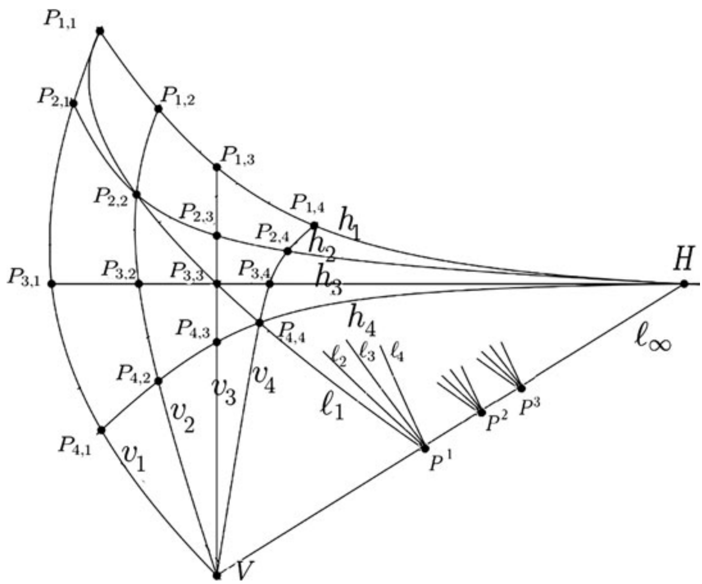


Fig. 21.1 A projective plane of order 4 used to construct a set of four mutually orthogonal Latin squares

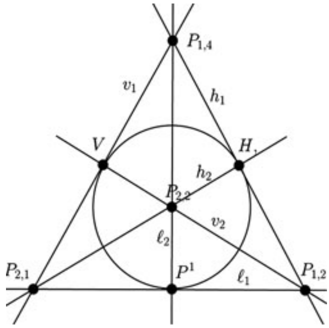
$v_1, \ldots, v_m$. We postulate that the point $P_{i,j}$is incident with the lines h_i and v_j, thus these two lines meet in the unique point $P_{i,j}$. We introduce a line ℓ_∞, incident with none of the points $P_{i,j}$, and introduce two points on it V, incident with all the $v_1, \ldots, v_m$, and H, incident with all the $h_1, \ldots, h_m$. On ℓ_∞ we also introduce $m-1$ more points $P^1, \ldots, P^{m-1}$. The point P^α is associated with the matrix M_α in a manner to be explained below. We finally introduce m lines incident with P^α for $\alpha = 1, \ldots, m-1$, we denote these lines by $\ell_n^\alpha, n = 1, \ldots, m$ and $\alpha = 1, \ldots, m-1$.

The total number of points now is $m^2 + m - 1 + 2 = m^2 + m + 1$, so we have enough points. As for lines, there are $2m + m(m-1) + 1 = m^2 + m + 1$, and we have all the lines we need as well.

It remains to define which points are incident with the lines $\ell_n^\alpha, n = 1, \ldots, m$, $\alpha = 1, \ldots, m-1$. This is where the mutually orthogonal matrices are used. In fact, the line ℓ_n^α is incident with P^α and with the point $P_{i,j}$ provided that the integer n appears in position (i, j) of matrix M_α.

Now it is a straightforward task to verify the axioms for a projective geometry. Of course the assumption of mutual orthogonality for the matrices M_α is essential. We leave this last part to the reader.

Exercise 9.4: The Latin square of course is $\left\{ \begin{matrix} 2 & 1 \\ 1 & 2 \end{matrix} \right\}$ which one should see right away without even labelling the model as we have done on the illustration below.



Exercise 9.5: The labelled model for the projective plane of order 3 is shown in Fig. 21.2.

Using the figure, we may read off the following pair of Latin squares:

$$M_1 = \begin{Bmatrix} 3 & 1 & 2 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{Bmatrix} \text{ and } M_2 = \begin{Bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{Bmatrix}$$

They are evidently orthogonal.

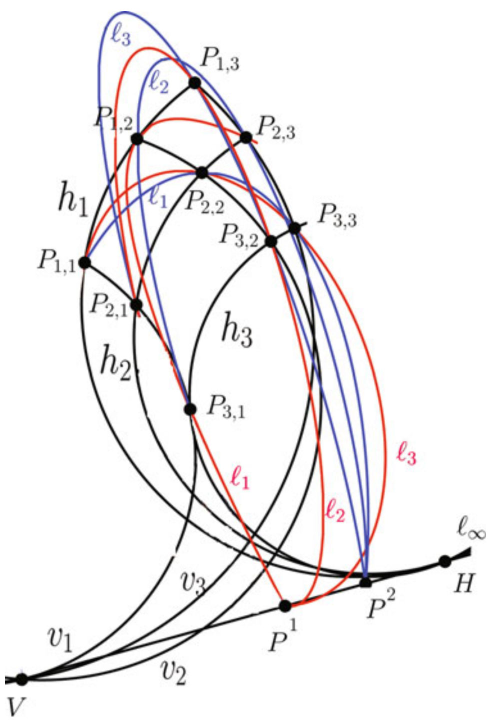


Fig. 21.2 The labelled model for the projective plane of order 3

Exercise 11.2: The first part is straightforward. To prove the last expression for $\sin(3\varphi)$, we write $\sin(3\varphi) = \sin(\varphi)(3 - 4\sin^2(\varphi)) = 4\sin(\varphi)(\frac{3}{4} - \sin^2(\varphi)) = 4\sin(\varphi)(\sin(\frac{\pi}{3} + \sin(\varphi))(\sin(\frac{\pi}{3} - \sin(\varphi))$. Then use the formulas

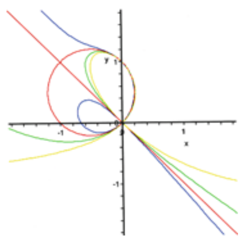
$$\sin(A) + \sin(B) = 2\sin\left(\frac{A+B}{2}\right)\cos\left(\frac{A-B}{2}\right)$$

$$\sin(A) - \sin(B) = 2\sin\left(\frac{A-B}{2}\right)\cos\left(\frac{A+B}{2}\right)$$

Exercise 14.2:

(a) Factorizing gives $P(x, y) = x^3 + x^2y + xy^2 + y^3 + x^2 - y^2 = (x + y)(x^2 + x + y^2 - y)$.

(b) Using the equation on the factorized form we get $\frac{\partial P}{\partial x} = (x^2+x+y^2-y)+(x+y)(2x+1)$ and $\frac{\partial P}{\partial y} = (x^2+x+y^2-y)+(x+y)(2y-1)$. Thus a singular point must lie on the curve given by $(x+y)(2x+1) = (x+y)(2y-1)$, that is on the curve given by $(x+y)(2x+1-(2y-1)) = 0$, thus $(x+y)(x-y+1) = 0$. Now the line $y = x + 1$ has the points (-1,0), (0,1) and $(-\frac{1}{2}, \frac{1}{2})$ in common with the curve, none of them are zeroes of the partial derivatives. Thus the singularities are the two points of intersection between the line $x + y = 0$ and the circle $x^2 + x + y^2 - y = 0$.



(c) Yellow corresponds to $t = 0$, green to $t = 0.5$, red to $t = 1$ and blue to $t = 1.5$. Conjecture: Except for $t = 1$ all the curves for finite values of t are irreducible, and in the irreducible cases they all have one singularity, at the origin. Furthermore all the curves have the same two tangential lines at the origin.

Indeed, the assertion on the tangential lines is easily seen to be true since the initial form at the origin is $x^2 - y^2$ for all t. The statement about the singularities is a consequence of Bézout's theorem on the intersection of curves in $\mathbb{P}^2(\mathbb{C})$, since the degree is 3 and the multiplicity at the origin is 2. See [36] for statement and proof of this important result. When $t \to \infty$ then the curve approaches the triple y-axis. However, this is not readily seen from the plots, as the process is rather slow. Instead we replace t by $u = \frac{1}{t}$, and get the family $x^3+ux^2y+uxy^2+uy^3+ux^2-uy^2 = 0$, where $u = 0$ yields the triple y-axis.

All illustrations in this chapter are by the author.

References

1. Archimedes. *The Method.* At Walters Art Gallery. http://www.thewalters.org/
2. J. F. Cantlon and E. M. Brannon. Animal Arithmetic. In *Encyclopedia of Animal Behavior*. Elsevier, Cambridge, Mass., USA, 2010.
3. J. L. Berggren. Al Kuhi's "Filling a Lacuna in Book II of Archimedes" in the Version of Nasir al Din al Tusi. *Centaurus*, Vol. 38, pp. 140–207, 1996.
4. R. A. Brualdi and H. R. Ryser. *Combinatorial Matrix Theory.* Encyclopedia of Mathematics And Its Applications, Volume 39. Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney 1991.
5. V. Brun. *Alt er Tall* (Everything is Numbers). Universitetsforlaget, Oslo.
6. S. Burris. Gauss and non-Euclidian Geometry. Notes available at www.math.uwaterloo.ca/~snburris/htdocs/geometry.pdf
7. *The Catholic Encyclopedia.* www.newadvent.org/cathen/
8. B.A. Cipra. Computer Search Solves an Old Math Problem. *Science*, Vol. 242, Nr. 4885, pp. 1507–1508. Des. 1988.
9. P. A. Cromwell. *Polyhedra.* Cambridge University Press, Cambridge, United Kingdom, 1997.
10. *Cut The Knot.* www.cut-the-knot.org.
11. M. Dzielska. Hypatia of Alexandria. Harvard University Press, Cambridge, Massachusetts, 1996.
12. W. Durant. *The Story of Civilization.* Simon and Schuster, New York 1950.
13. Encyclopedia Britannica, Americanized. The Werner Company, New York, Chicago 1898.
14. H. Eves. An Introduction to the History of Mathematics. Saunders College Publishing, 1992.
15. J. Friberg. *Amazing Traces of a Babylonian Origin in Greek Mathematics.* World Scientific Publishing Co. Ltd., Singapore 596224.
16. W. Fulton. *Algebraic Curves*. W. A. Benjamin, Inc., 1969.
17. E. Gibbon. http://www.ccel.org/g/gibbon/decline
18. E. Gibbon. *The Decline and Fall of the Roman Empire.* Edited and with a preface by D. A. Saunders. The Vicing Press, New York 1952.
19. C.C. Gillispie. *The Dictionary of Scientific Biography.* 16 volumes, 2 suppl. Charles Scibner's Sons, New York 1979–1990.
20. M. Grant. *Cicero-On the Good Life.* Penguin Books, New York 1971.
21. M. Greenberg. Euclidian and Non-Euclidian Geometries. Development and history. W. H. Freeman and Company, San Fransisco 1972.
22. A. Grothendieck with the collaboration of J. Dieudonné. Éléments de géométire algébrique $I - IV$. *Publications Mathématiques de l'IHES*, 4, 8, 11, 17, 20, 24, 28 and 32, 1961–1967.
23. T.W. Gray and J. Glynn. *Exploring Mathematics with Mathematica.* Addison-Wesley Publishing Company, 1991.
24. F. Hausdorff. *Dimension und äußeres Maß*. Mathematische Annalen 79, pp. 157–179, 1919.
25. R. Hartshorne. *Geometry: Euclid and Beyond*. Springer Graduate Texts in Mathematics. Springer, New York, Berlin, Heidelberg 2000.

26. T.L. Heath. *A History of Greek Mathematics*. New York: Dover 1981, 2 vols., paperback version of 1921.
27. T.L. Heath. *Euclid: The thirteen books of the Elements.* Translated from the text of Heiberg. Translated with introduction and commentary by Sir Thomas L. Heath. 3 Vols. Dover Publications, New York 1956.
28. Herodotus. *The Histories.* Translated by A. de Sélincourt, revised by J. Marincola. 1954 and 1972, Penguin Books, London 1996.
29. D. Hilbert. *Grundlagen der Geometrie*. 1899. Tenth Printing, Teubner Verlag 1968.
30. K. Holing. På gjengrodde stiger. A series of articles in Norwegian in *Nordisk Matematisk Tidsskrift*, 1997, 1998, 2000, 2002.
31. A. Holme. *Innføring i Geometri. Fra Euklid of Mandelbrot.* (Introduction to Geometry. From Euclid to Mandelbrot.) Alma Mater Forlag, Bergen 1996. In Norwegian.
32. A. Holme. *Innføring i lineær algebra med anvendelser.* (Introduction to linear algebra with applications.) Fagbokforlaget, Bergen 1999. In Norwegian.
33. A. Holme. *Matematikkens Historie, bind 1. Fra Babylon of mordet på Hypatia.* (History of Mathematics, volume 1. From Babylon to the Murder of Hypatia.) Fagbokforlaget, Bergen 2001. In Norwegian.
34. A. Holme. *Matematikkens Historie, bind 2. Fra de arabiske vise til Niels Henrik Abel.* (History of Mathematics, volume 2. From the Arabian Sages to Niels Henrik Abel.) Fagbokforlaget, Bergen 2001. In Norwegian.
35. A. Holme. *Da Matematikken ble Til.* (When Mathematics was Born.) Damm, Oslo 2006. In Norwegian.
36. A. Holme. *A Royal Road to Algebraic Geometry.* Work in preparation.
37. Iamblichus. *Life of Pythagoras*. Translated by T. Taylor. London, 1818. Our reference is a summary written anonymously by a student in 1905, reprinted on Kessinger Publishing, LLC.
38. Victor J. Katz. *A History of Mathematics. An Introduction.* Second Edition. Addison-Wesley. Reading Massachussets 1998.
39. Lam, C. W. H: *The Search for a Finite Projective Plane of Order 10* American Mathematical Monthly, Vol. 98, no. 4, pp. 305–318, 1991. Also at www.cecm.sfu.ca/organics/papers/lam/paper/html/node2.html
40. E.W. McLemore, Past Present (we) Present Future (you). *Women in Mathematics Newsletter* Vol. 9, no. 6, pp. 11–15, 1979.
41. *Livius. Articles on ancient history*. www.livius.org
42. W.J. LeVeque. *Fundamentals of Number Theory.* Addison-Wesley, 1977.
43. G. E. Martin. *Geometric Constructions.* Undergraduate Texts in Mathematics. Springer, Berlin, Heidelberg, New York 1998.
44. O. Neugebauer. *The Exact Sciences of Antiquity.* Second edition. Dover Publications, Inc. New York 1969. Reprinted from the original, published by Brown University 1957.
45. O. Neugebauer and A. Sachs. *Mathematical Cuneiform Texts.* Yale University Press, New Haven, Connecticut, USA 1945.
46. Plutarchus. *Lives of Illustrious Men.* 3 volumes. Translated from the Greek by John Dryden and others. Hovendon Company, New York 1880.
47. J. Richter-Gebert and U.H. Kortenkamp. *The Interactive Geometry Software Cinderella.* Springer, Berlin, Heidelberg, New York 1999.
48. G. Robins and C. Shute. *The Rhind Mathematical Papyrus.* Published by the Trustees of British Museum, by British Museum Press, London 1997, reprinted 1990, 1998.
49. R. Rashed. *The development of Arabic Mathematics: Between Arithmetic and Algebra.* Kluwer, Dordrecht, Boston, London 1994.
50. E. Robson. Neither Sherlock Holmes nor Babylon: A reassessment of Plimpton 322. *Historia Mathematica* Vol. 28, pp. 167–206, (2001).
51. E. Robson. Words and Pictures: A reassessment of Plimpton 322. *Amer. Math. Monthly*. Vol. 109, pp. 105–120, (2002).
52. B. Scimemi. Algebra og geometri ved hjelp av papirbretting. (Algebra and geometry by folding paper.) Translated from the Italian with comments by Christoph Kirfel. *Nordisk Matematisk Tidsskrift*, 4, 1998.

53. P.T. Saunders. *An introduction to Catastrophe Theory*. Cambridge University Press, 144 pp., 1980.
54. Robert R. Stoll. *Set Theory and Logic.* W. H. Freeman and Company, San Fransisco and London 1961.
55. R.D.J. Struik. *A Concise History of Mathematics.* Fourth Revised Edition Dover Publications Inc, New York 1987.
56. W.W. Struve. Mathematisher Papyrus des Staatlichen Museums der Schönen Künste. *Part A, Quellen* (1930).
57. L. F. Tóth. *Regular Figures.* International Series of Monographs in Pure and Applied Mathematics. Pergamon Press 1964.
58. B.-L. van der Waerden. *Science Awakening.* English translation by Arnold Dresden, with additions by the author. P. Nordhoof Ltd – Groningen, Holland, 1954.
59. Agnes Scott University, *Women in Mathematics.* http://www.agnesscott.edu/lriddle/women
60. University of St. Andrews, *History of Mathematics.* http://www-history.mcs.st-andrews.ac.uk/history/Mathematicians
61. Wikipedia. The Free Encyclopedia. *wikipedia.org*
62. Wikimedia Commons. Wikimedia Commons is a media file repository making available public domain and freely-licensed educational media content (images, sound and video clips) to all.
63. University of Arizona, Women's study department. *Greek Women Philosophers.* info-center.ccit.arizona.edu
64. P. Y. Woo. Straightedge Constructions, Given a Parabola. *The College Mathematics Journal.* Vol. 31, No. 5, pp. 362–372, 2000.
65. E.C. Zeeman. Catastrophe Theory. *Scientific American*, pp. 65–83, 1976.

Index